L1 sciences éco 2010-2011

Statistique descriptive
Notes de cours

Hélène Boistard
Université Toulouse 1

Table des matières

Chapitre 1

Les données statistiques

1.1 Les variables statistiques - éléments de vocabulaire

On observe un **échantillon** composé de n **individus** appartenant à une même **population** de taille N. Chaque individu de l'échantillon est observé à travers des caractéristiques, caractères ou indicateurs appelés **variables**. Une **série statistique** $\{x_1, x_2, \ldots, x_n\}$ est la suite des valeurs prises par une ou plusieurs variables pour chacun des individus de l'échantillon.
Exemple : un questionnaire est distribué à 20 personnes. Il comporte diverses questions. La population = l'échantillon = les étudiants ayant répondu au questionnaire. Les individus sont les personnes interrogées. Les variables correspondent aux questions posées : l'âge, la taille, la couleur des yeux, etc.
Schéma :

1.2 Les types de variables

1.2.1 Variables qualitatives

Une variable est appelée **qualitative** lorsque les réponses possibles à la question posée, ou les valeurs prises par la variable, ne correspondent pas à une quantité mesurable par un nombre mais appartiennent à un groupe de **catégories**. On les appelle **modalités** de la variable.
Exemple : le sexe, la couleur des yeux, la mention au baccalauréat, la fréquence d'une activité (jamais, rarement, parfois, souvent, très souvent).
On distingue :

- les variables **qualitatives nominales** : il n'y a pas de hiérarchie entre les différentes modalités ; exemple : sexe, couleur des yeux.
- les variables **qualitatives ordinales** : les différentes modalités peuvent être ordonnées de manière naturelle ; exemple : la mention au baccalauréat, la fréquence d'une activité.

Remarque : certaines variables nominales peuvent être désignées par un code numérique, qui n'a pas de valeur de quantité. Exemple : le code postal, le sexe (1=garçon, 2=fille).

1.2.2 Variables quantitatives

Les réponses correspondent à des quantités mesurables et sont données sous forme de nombre.
On distingue :

- les variables quantitatives discrètes : elles prennent leurs valeurs dans un ensemble discret, le plus souvent fini ; exemple : le nombre d'enfants, la pointure du pied.
- les variables quantitatives continues : elles peuvent prendre toutes les valeurs d'un intervalle réel ; exemple : la taille des individus, une note à un examen.

Remarque : l'âge peut être vu et traité comme une variable quantitative discrète ou continue suivant la précision que l'on choisit et le nombre de valeurs qu'il prend au sein de la population. Il peut également exister des variables basées sur l'âge qui sont qualitatives. Si dans un sondage on pose la question "quelle est votre tranche d'âge parmi les possibilités suivantes : - de 25 ans, entre 25 et 40, entre 40 et 60 et + de 60 ans", on peut voir la variable "tranche d'âge" comme une variable qualitative ordinale.

1.3 Les variables qualitatives : tableaux de fréquence et représentation graphique

Exemple : On s'intéresse à la variable "couleur des yeux" sur un groupe de 20 personnes. On code chaque modalité de la manière suivante : M=marron, V=vert, N=noir, B=bleu. On obtient la série statistique suivante :
M, V, M, M, M, N, M, B, M, B.

1.3.1 Tableaux de distribution de fréquences absolues, relatives et cumulées

Exemple : Pour l'exemple précédent, on remplit le tableau suivant :

Couleur des yeux	M	V	N	B	Total
Effectif					
Proportion					

Tableau-type : On choisit une notation pour la variable, par exemple : X. n désigne le nombre d'individus dans l'échantillon. On note C_1, ..., C_k les k modalités de la variable. Pour $1 \leq j \leq k$, on note

- n_j l'effectif associé à la modalité C_j (le nombre d'individus pour lesquels la valeur prise par la variable est C_j),
- $f_j = n_j/n$ la fréquence relative ou proportion associée à cette modalité,

– et si la variable est qualitative **ordinale** : $\Phi_j = f_1 + f_2 + \cdots + f_j$ la fréquence relative cumulée pour cette modalité (avec la convention : $\Phi_0 = 0$). Elle n'a de sens que si la variable est qualitative ordinale et si les modalités C_1, ..., C_k sont ordonnées suivant l'ordre croissant naturel (ou hiérarchique ascendant) qui règne parmi ces modalités.

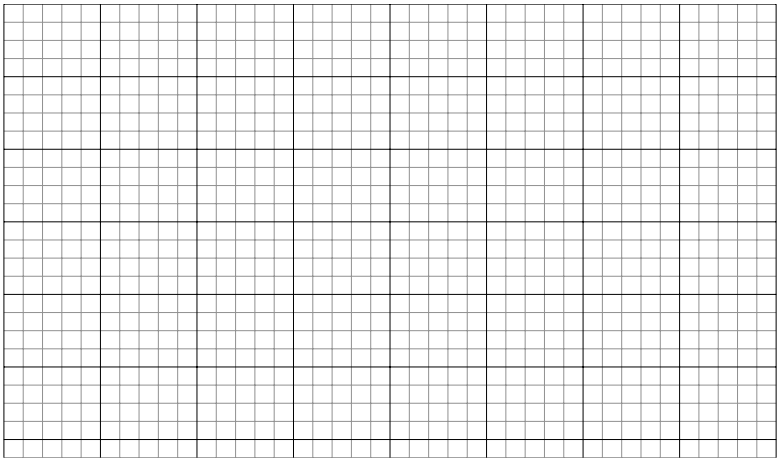
Le tableau suivant est un tableau-type qui permet de résumer les données.

Variable X	C_1	C_2	...	C_k	Total
Fréquence absolue ou effectif	n_1	n_2	...	n_k	n
Fréquence relative ou proportion	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_k = n_k/n$	1
Fréquence relative cumulée*	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	...	$\Phi_k = f_1 + f_2 + \cdots + f_k = 1$	pas de sens

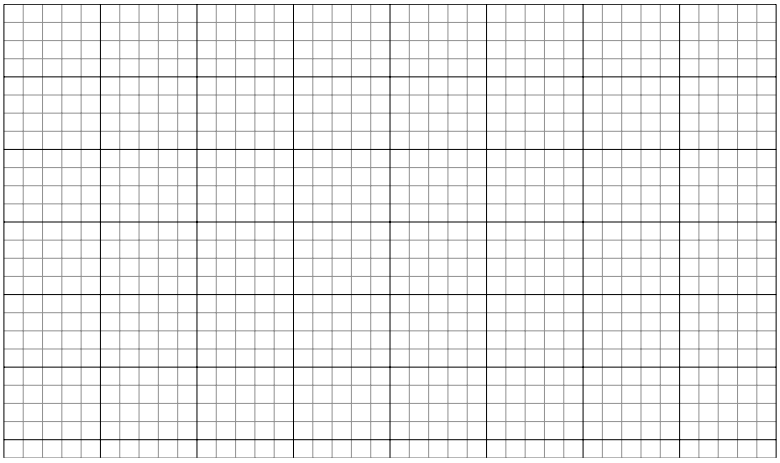
$*$ Attention : uniquement dans le cas de variables qualitatives ordinales.

1.3.2 Représentation graphique : diagrammes en secteurs et diagrammes en tuyaux d'orgue

1. **Diagramme en secteurs** : chaque modalité est représentée par un secteur d'un disque dont l'angle est proportionnel à la fréquence de la modalité (ou au pourcentage), l'angle 360 degrés équivalant à la fréquence relative 1 (ou au pourcentage 100%).
   Exemple :



2. **Diagramme en tuyaux d'orgue** : en abscisse sont disposées les différentes modalités auxquelles on associe des rectangles espacés entre eux, de largeur constante, dont les hauteurs (en ordonnée) sont proportionnelles à l'effectif ou à la fréquence relative de chaque modalité. Préciser le nom des axes, le nom du graphique et la source des informations. Dans le cas d'une variable qualitative ordinale, on peut également construire le diagramme en tuyaux d'orgue des effectifs ou des proportions cumulés.
   Exemple :



Remarque : cette représentation graphique est plus adaptée dans le cas d'une variable qualitative ordinale car elle rend compte de la structure d'ordre entre les modalités, disposées de gauche à droite par ordre croissant. C'est impossible de suggérer une structure d'ordre dans un diagramme en secteurs.

1.4 Les variables quantitatives discrètes

Exemple : pour 20 individus, on a relevé le nombre de fois où chacun a assisté à une séance de cinéma durant le mois d'août 2010. Pour simplifier, on nomme « ciné »la variable « nombre de séances de cinéma pendant le mois d'août ». La variable « ciné »sera notée C. La série statistique est résumée sous la forme du tableau suivant :

C	0	1	2	3	4
Effectif	4	6	7	2	1

1.4.1 Tableaux de distribution de fréquences

Exemple : pour la variable C, on remplit le tableau suivant :

C	0	1	2	3	4
Effectif					
Proportion ou fréquence relative					
Proportion cumulée ou fréquence relative cumulée					

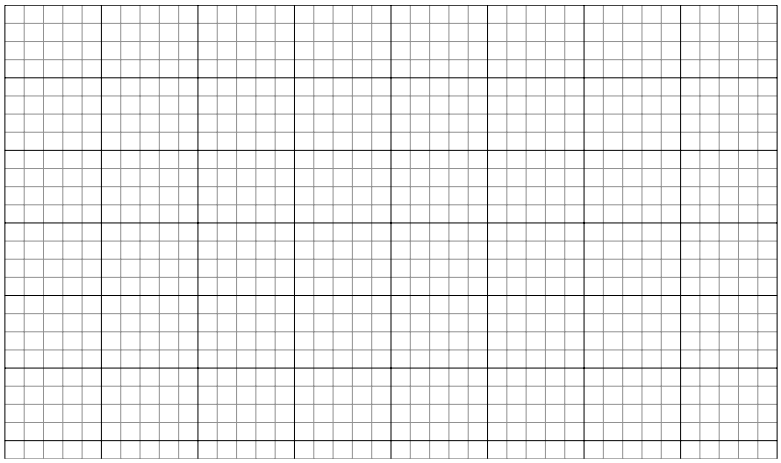
On note $v_1, \ldots, v_k$ les k valeurs différentes que peut prendre la variable (remarque : on n'en rencontrera pas d'exemple dans ce cours, mais une variable discrète peut prendre une infinité de valeurs). Pour $1 \leq j \leq n$, on note n_j l'effectif des individus pour lesquels la variable prend la valeur v_j. On note f_j la fréquence relative ou proportion pour la valeur v_j et $\Phi_j = f_1 + \cdots + f_j$ la j-ème fréquence relative cumulée (avec la convention : $\Phi_0 = 0$). On résume habituellement les données comme dans le tableau-type suivant :

Valeurs prises par la variable	v_1	v_2	$\ldots$	v_k	Total
Fréquence absolue	n_1	n_2	$\ldots$	n_k	n
Fréquence relative	$f_1 = n_1/n$	$f_2 = n_2/n$	$\ldots$	$f_k = n_k/n$	1
Fréquence relative cumulée	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	$\ldots$	$\Phi_k = f_1 + f_2 + \cdots + f_k = 1$	pas de sens

1.4.2 Représentation graphique : diagramme en bâtons

On trace un graphique avec
- sur l'axe des abscisses les différentes valeurs prises par la variable, placées **en respectant une échelle**,
- en ordonnée les fréquences relatives ou les fréquences absolues.
- Pour chaque valeur v_j on construit un bâton vertical à l'abscisse v_j, de hauteur proportionnelle à la fréquence de la valeur v_j.

Exemple : ciné.

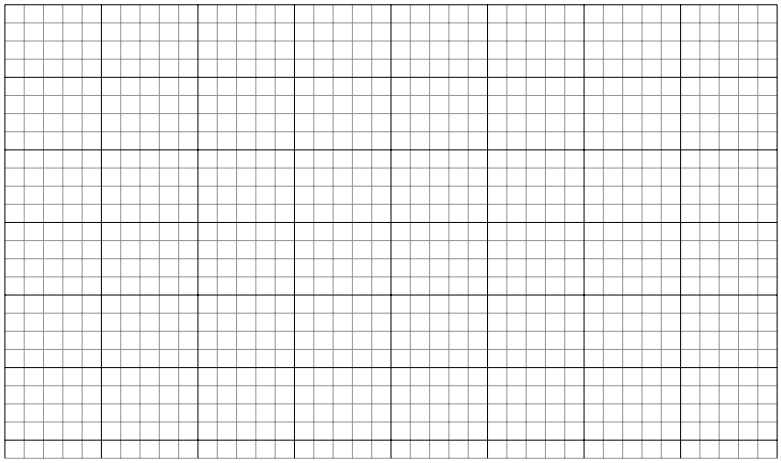


1.4.3 Autre représentation graphique : fonction de répartition empirique

La fonction de répartition empirique permet de décrire la série statistique de manière complète. Elle est définie sur $\mathbb{R}$ et prend ses valeurs dans $[0, 1]$. Pour x dans $\mathbb{R}$, elle est définie par :

$$F(x) = \begin{cases} 0 & \text{si } x < v_1 \\ \Phi_j & \text{si } v_j \leq x < v_{j+1} \\ 1 & \text{si } v_k \leq x. \end{cases}$$

Exemple : ciné.



1.5 Les variables quantitatives continues

Exemple : on s'intéresse à la taille, notée T et exprimée en mètres, de 20 individus. On a obtenu la série statistique suivante :
1,72 ; 1,87 ; 1,66 ; 1,73 ; 1,64 ; 1,77 ; 1,80 ; 1,81 ; 1,60 ; 1,78 ; 1,83 ; 1,75 ; 1,70 ; 1,58 ; 1, 68 ; 1,66 ; 1,93 ; 1,75 ; 1,80 ; 1,85.

1.5.1 Tableaux de distribution de fréquences - fréquences cumulées

Les données brutes de la variable pour chaque individu sont notées $x_1, \ldots, x_n$. Elles peuvent prendre n'importe quelle valeur dans un intervalle de $\mathbb{R}$ et il est très rare d'avoir deux fois la même valeur pour deux individus différents. Il serait donc inutile de tracer un diagramme en bâtons comme dans le cas d'une variable discrète : il consisterait en un amoncellement illisible de bâtons de hauteur $1/n$. On choisit donc de faire un **regroupement en classes**.

Regroupement en classes :
- L'intervalle où la variable prend ses valeurs est divisé en k classes : $[b_0, b_1[$, $[b_1, b_2[$, ..., $[b_{k-1}, b_k[$ (il est possible d'avoir des bornes infinies).
- Pour $1 \leq j \leq k$, on note n_j l'effectif associé à la classe $[b_{j-1}, b_j[$, $f_j = n_j/n$ la fréquence relative associée à cette classe et $\Phi_j = f_1 + \cdots + f_j$ la j-ième fréquence cumulée (avec la convention : $\Phi_0 = 0$).
- On note $a_j = b_j - b_{j-1}$ l'amplitude de la classe $[b_{j-1}, b_j[$.
- On note $d_j = f_j/a_j$ la densité de proportion pour la classe $[b_{j-1}, b_j[$.

Exemple de la taille :

T	[1,50 ; 1,65[	[1,65 ; 1,75[	[1,75 ; 1,85[	[1,85 ; 2,00[
Effectif	3	6	8	3
Proportion				
Proportion cumulée				
Amplitude				
Densité de proportion				

Remarques :
- la densité de proportion permet de comparer les effectifs dans chaque classe en tenant compte de la taille de ces classes (cf. la notion de densité de population en géographie).
- Dans le cas de classes qui ont toutes la même longueur, il n'est pas nécessaire de calculer la densité de proportion, il est suffisant d'étudier les fréquences relatives ou absolues (qui sont directement proportionnelles à la densité de proportion).

Tableau-type :

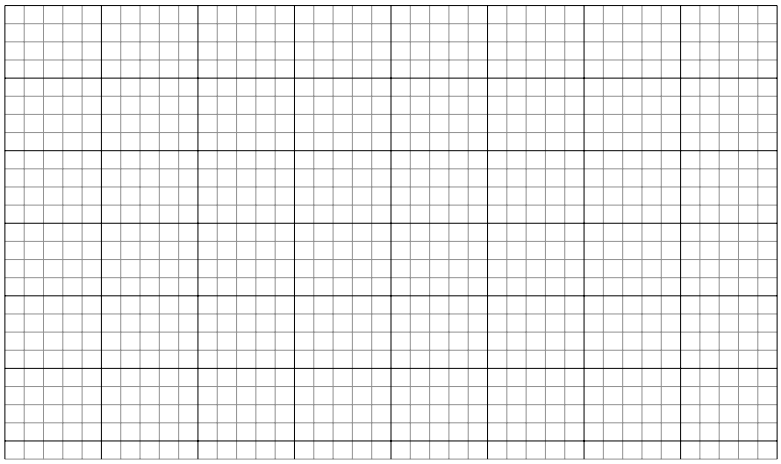
Variable X	$[b_0, b_1[$	$[b_1, b_2[$	...	$[b_{k-1}, b_k[$	Total
Fréq. absolue	n_1	n_2	...	n_k	n
Fréq. relative	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_k = n_k/n$	1
Fréq. relative cumulée	$\Phi_1 = f_1$	$\Phi_2 = f_1 + f_2$	...	$\Phi_k = 1$	
Amplitude	$a_1 = b_1 - b_0$	$a_2 = b_2 - b_1$	...	$a_k = b_k - b_{k-1}$	
Densité de proportion	$d_1 = f_1/a_1$	$d_2 = f_2/a_2$	...	$d_k = f_k/a_k$	

Remarque : Ce tableau contient-il toute l'information apportée par les données brutes ou bien représente-t-il une perte d'information ? Quel est l'intérêt d'un tel tableau ?

1.5.2 Représentation graphique : histogramme et fonction de répartition empirique

Sur l'axe des abscisses sont placées les bornes des classes en respectant une échelle. Pour chaque classe, on élève un rectangle de hauteur proportionnelle à la densité de proportion.

Exemple de la taille T :



Remarque : on représente la **densité de proportion** et non pas les fréquences relatives ou absolues.

Conséquence : l'aire d'un rectangle est proportionnelle à la fréquence (relative ou absolue) de la classe correspondante. En effet, pour le rectangle correspondant à la classe $[b_{j-1}, b_j[$, l'aire est

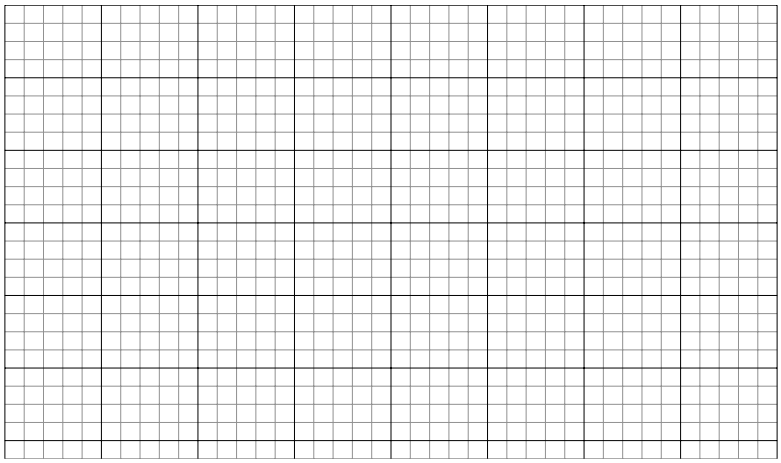
$$(b_j - b_{j-1}) \times d_j = f_j.$$

Approximation de proportions : pour x une valeur dans l'intervalle $[b_{j-1}, b_j[$, on approche la proportion d'individus pour lesquels la variable est inférieure ou égale à x par l'aire de l'histogramme entre les abscisses b_0 et x, notée $F(x)$:

$$F(x) = f_1 + f_2 + \cdots + f_{j-1} + (x - b_{j-1}) \times d_j = \Phi_{j-1} + (x - b_{j-1}) \times d_j.$$

On a ainsi défini une fonction Φ qui vaut 0 sur $]-\infty, b_0[$, et 1 sur $[b_k, +\infty[$. Elle vaut Φ_j en b_j. Sur $[b_{j-1}, b_j[$, c'est une fonction affine de pente d_j. Cette fonction, affine par morceaux, est appelée **fonction de répartition empirique**.

Fonction de répartition empirique de la variable T :



Chapitre 2

Résumés numériques d'une variable quantitative

Dans ce chapitre, X désigne une variable quantitative.

2.1 Paramètres de position

2.1.1 Le mode

Le mode rend compte de l'endroit où les données sont le plus concentrées.

Pour une variable **discrète**, le mode est la ou les valeurs de la variable qui correspond(ent) à l'**effectif maximal** (ou à la fréquence relative maximale).

Pour une variable **continue** regroupée en classes, le mode est la ou les classe(s) de **densité de proportion maximale**.

Exemples : ciné, taille.

2.1.2 La moyenne

On note $\{x_1 \ldots, x_n\}$ la série statistique. La moyenne est définie par :

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Exemple : ciné, taille.

Cas d'une variable discrète : si v_1, ..., v_k sont les k valeurs prises par la variable X, n_j l'effectif et f_j la fréquence relative correspondant à la valeur v_j, on peut réécrire :

$$\bar{x} = \frac{n_1 v_1 + n_2 v_2 + \cdots + n_k v_k}{n} = \frac{1}{n} \sum_{i=1}^{n} n_j v_j = \sum_{i=1}^{n} f_j v_j.$$

Exemple : ciné.

Cas d'une variable continue regroupée en classes : la variable X est regroupée dans les classes $[b_{j-1}, b_j[$ $(1 \leq j \leq n)$, les fréquences relatives associées à ces classes sont notées f_j, $1 \leq j \leq n$. Lorsque les données brutes ne sont plus accessibles et qu'on ne dispose que des données regroupées en classes, on calcule une **moyenne approchée** grâce à des représentants des classes (leurs centres) : $c_j = (b_{j-1} + b_j)/2$, par la formule :

$$\bar{x}_{app} = f_1 c_1 + f_2 c_2 + \cdots + f_k c_k = \sum_{i=1}^{n} f_j c_j.$$

Exemple : calcul d'une moyenne approchée de la variable « taille » à partir du regroupement en classes.

Propriétés de la moyenne : si on fait le changement de variable $Y = aX + b$ (traduction sur les séries statistiques : $y_i = ax_i + b$, $1 \leq i \leq n$), alors

$$\bar{y} = a\bar{x} + b.$$

Exemple : calcul de la taille moyenne en centimètres.

2.1.3 La médiane

"En gros", le calcul de la médiane revient à ranger les observations par ordre croissant et trouver un point au-dessous duquel se situent 50 % des observations et au-dessus duquel se situent 50 % des observations.

a) Cas d'une variable discrète.

- Si n est **impair**, la médiane est la $\frac{n+1}{2}$-ième observation.
- Si n est **pair**, il y a plusieurs façons convenables de définir la médiane. Nous choisirons la suivante : la médiane est la plus petite valeur observée v_j telle que l'effectif cumulé en v_j dépasse $n/2$ (dépasse au sens large : est supérieure ou égale). Autrement dit, c'est la plus petite valeur v_j pour laquelle la proportion cumulée dépasse $1/2$. **Remarque** : cette définition est encore vraie pour n **impair**.

La détermination de la médiane se fait donc à l'aide des effectifs cumulés, des proportions cumulées ou de la fonction de répartition empirique (graphiquement).

Exemple : ciné.

b) Cas d'une variable continue. La médiane est définie comme la solution Q_2 de l'équation :

$$F(Q_2) = 0.5,$$

où F est la fonction de répartition empirique de la variable. On sait que cette solution existe parce que F est continue, et $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$. Si de plus F est strictement croissante, la solution Q_2 est unique. La méthode pratique est la suivante :

1. S'il existe une borne de classe b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exactement 0.5, autrement dit : $F(b_j) = 0.5$, alors **la médiane est ce** b_j.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

   $$F(b_{j-1}) < 0.5 < F(b_j).$$

   Cette classe est la première sur laquelle la fréquence cumulée dépasse 0.5. Pour $x \in [b_{j-1}, b_j[$, $F(x) = \Phi_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

   $$F(Q_2) = \Phi_{j-1} + (Q_2 - b_{j-1}) \times d_j = 0.5$$

   D'où

   $$Q_2 = \frac{0.5 - \Phi_{j-1}}{d_j} + b_{j-1}.$$

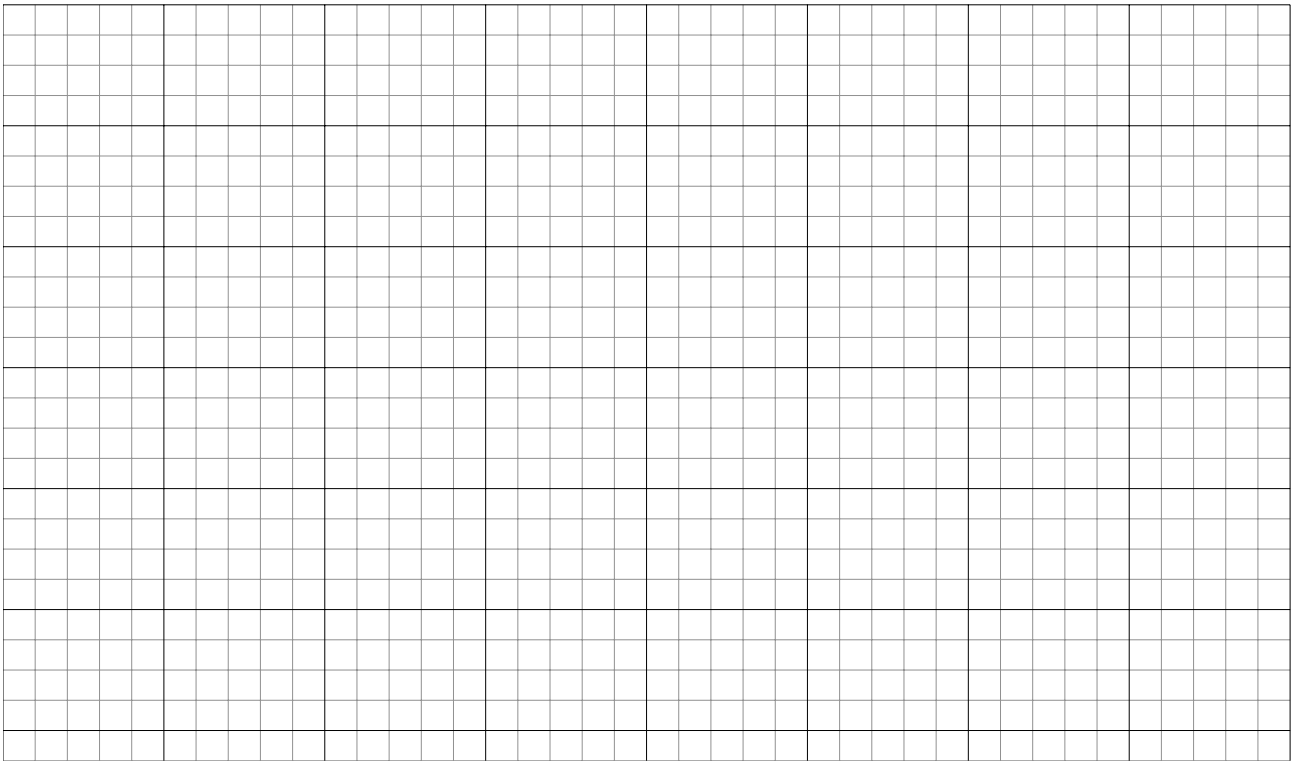
   Ou encore, en termes des b_j et de F :

   $$Q_2 = \frac{0.5 - F(b_{j-1})}{F(b_j) - F(b_{j-1})} \times (b_j - b_{j-1}) + b_{j-1}.$$

Cette méthode peut se traduire graphiquement en utilisant le graphe de la fonction de répartition empirique et le théorème de Thalès.

Exemple : médiane de la variable « taille », regroupée en classes.

Méthode graphique avec la fonction de répartition empirique :



2.1.4 Quantiles

a) Cas d'une variable continue

Soit X une variable quantitative continue, de fonction de répartition empirique F. On suppose qu'on dispose de la répartition en classes des observations.

Le **quantile d'ordre** p de X est la solution notée q_p de :

$$F(q_p) = p.$$

Cela signifie qu'une proportion d'environ p des observations est inférieure à q_p et qu'une proportion d'environ $1-p$ des données est supérieure à q_p.

Quantiles particuliers

- Quartiles : quantiles correspondant aux proportions multiples de 0.25 (un quart). On note Q_1 le premier quartile, qui correspond à $q_{0.25}$, Q_3 le troisième quartile, qui correspond à $q_{0.75}$. La médiane est le deuxième quartile $Q_2 = q_{0.5}$.
- Déciles : quantiles correspondant aux proportions multiples de 0.1 : $q_{0.1}$ (premier décile), $q_{0.2}$ (deuxième décile), etc.
- Percentiles ou centiles : quantiles correspondant aux proportions multiples de 0.01. Par exemple, le 65ème percentile est le quantile $q_{0.65}$.

Calcul du quantile q_p : même méthode que pour le calcul de la médiane.

1. S'il existe une borne de classe b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exactement p, autrement dit : $F(b_j) = p$, alors $q_p = b_j$.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

$$F(b_{j-1}) < p < F(b_j).$$

Cette classe est la première sur laquelle la fréquence cumulée dépasse p. Pour $x \in [b_{j-1}, b_j[$, $F(x) = \Phi_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

$$F(q_p) = \Phi_{j-1} + (q_p - b_{j-1}) \times d_j = p$$

D'où

$$q_p = \frac{p - \Phi_{j-1}}{d_j} + b_{j-1}.$$

Ou encore, en termes des b_j et de F :

$$q_p = \frac{p - F(b_{j-1})}{F(b_j) - F(b_{j-1})} \times (b_j - b_{j-1}) + b_{j-1}.$$

Exemple : troisième quartile de la variable « taille ».

b) Cas d'une variable discrète

Comme pour la médiane, il existe diverses manières de définir les quantiles d'une loi discrète : comme la fonction de répartition empirique n'est pas continue mais a des paliers, elle ne prend pas toutes les valeurs entre 0 et 1. Pour une proportion p fixée, on cherche donc une valeur x telle que $F(x)$ s'approche, en un certain sens, de p. Nous choisissons la définition suivante :

$$q_p = \begin{cases} v_1 & \text{lorsque } 0 < p \leq \Phi_1 = f_1, \\ v_2 & \text{lorsque } \Phi_1 < p \leq \Phi_2, \\ \dots, & \\ v_j & \text{lorsque } \Phi_{j-1} < p \leq \Phi_j, \\ \dots, & \\ v_k & \text{lorsque } p = \Phi_k (= 1). \end{cases}$$

Exemple : troisième quartile de la variable « ciné ».

2.1.5 Utilisation des paramètres de tendance centrale

Robustesse
La médiane est plus **robuste** que la moyenne : une ou plusieurs données erronnées ne font pratiquement, voire pas du tout, changer la médiane, alors qu'elles peuvent affecter considérablement la moyenne.

Assymétrie
La comparaison de la médiane et de la moyenne permet de détecter des assymétries dans les données :

2.2 Paramètres de dispersion

2.2.1 L'étendue

Soit x_{min} la plus petite observation et x_{max} la plus grande. On définit **l'étendue** $e = x_{max} - x_{min}$. Elle a la même unité que l'unité de la variable. Elle n'est pas très informative car elle ne tient pas du tout compte de la répartition des données à l'intérieur de l'intervalle $[x_{min}, x_{max}]$.

Exemple : étendue de la variable « taille ».

2.2.2 L'intervalle inter-quartile

On appelle **intervalle interquartile** l'intervalle $[Q_1, Q_3]$, qui contient environ 50% des observations. La **distance interquartile** $Q_3 - Q_1$ est une mesure de dispersion.

Exemple : intervalle inter-quartile de la variable « taille ».

2.2.3 La variance et l'écart-type

La **variance** est définie par :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

L'expression suivante est plus pratique pour le calcul de la variance :

$$Var(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x})^2.$$

Preuve : en développant le carré dans la définition de la variance.

Pour une **variable quantitative discrète** prenant la valeur v_j un nombre n_j de fois (ou avec la fréquence f_j), pour $1 \leq j \leq k$:

$$\begin{aligned} Var(X) &= \frac{1}{n}\sum_{j=1}^{k} n_j (v_j - \bar{x})^2 = \sum_{j=1}^{k} f_j (v_j - \bar{x})^2 \\ &= \left(\frac{1}{n}\sum_{j=1}^{k} n_j v_j^2\right) - (\bar{x})^2 = \left(\sum_{j=1}^{k} f_j v_j^2\right) - (\bar{x})^2. \end{aligned}$$

Dans le cas d'une variable continue pour laquelle on dispose seulement des **données regroupées en classes**, on peut faire un calcul approché similaire à celui de la moyenne approchée $\bar{x}_{app}$. On calcule une valeur approchée de la variance, notée $Var_{app}(X)$. Toutes les expressions qui suivent sont équivalentes.

$$\begin{aligned} Var_{app}(X) &= \frac{1}{n}\sum_{j=1}^{k} n_j (c_j - \bar{x}_{app})^2 = \sum_{j=1}^{k} f_j (c_j - \bar{x}_{app})^2 \\ &= \left(\frac{1}{n}\sum_{j=1}^{k} n_j c_j^2\right) - (\bar{x}_{app})^2 = \left(\sum_{j=1}^{k} f_j c_j^2\right) - (\bar{x}_{app})^2, \end{aligned}$$

où c_j est le centre de la j-ème classe, dotée de l'effectif n_j (ou de la fréquence relative f_j).

Propriétés de la variance

- La variance est toujours positive ou nulle. Elle est nulle si et seulement si toutes les observations sont identiques :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \Leftrightarrow \forall i,\ x_i - \bar{x} = 0.$$

- L'unité de la variance est l'unité de X au carré.

L'**écart-type** σ_X est défini par :

$$\sigma_X = \sqrt{Var(X)}.$$

Propriété : l'unité de σ_X est l'unité de X.

Exemple : variance et écart-type de la variable « ciné », de la variable « taille ».

2.3 Changement de variable linéaire ou affine - Variable centrée réduite

2.3.1 Changement de variable linéaire ou affine

On considère une variable quantitative X et on lui fait subir une application affine qui la transforme en une variable Y. a et b sont des constantes réelles.

Nouvelle variable Y	Observations y_i	Moyenne de Y	Variance de Y	Ecart-type de Y
$Y = aX$	$y_i = ax_i$	$\bar{y} = a\bar{x}$	$Var(Y) = a^2 Var(X)$	$\sigma_Y = \lvert a \rvert \sigma_X$
$Y = X + b$	$y_i = x_i + b$	$\bar{y} = \bar{x} + b$	$Var(Y) = Var(X)$	$\sigma_Y = \sigma_X$
$Y = aX + b$	$y_i = ax_i + b$	$\bar{y} = a\bar{x} + b$	$Var(Y) = a^2 Var(X)$	$\sigma_Y = \lvert a \rvert \sigma_X$

Exemple :

2.3.2 Variable centrée réduite

On considère une variable X de moyenne $\bar{x}$ et de variance $Var(X)$, d'écart-type $\sigma_X = \sqrt{Var(X)}$. On définit une nouvelle variable

$$Y = \frac{X - \bar{x}}{\sigma_X}.$$

Elle est **sans unité**. Cette variable est appelée variable **centrée réduite associée à X**. En effet, elle est :

- **centrée** : $\bar{y} = \frac{\bar{x} - \bar{x}}{\sigma_X} = 0$.
- **réduite** $Var(Y) = \frac{Var(Y)}{Var(Y)} = 1$.

Quand on transforme une variable en la variable centrée réduite associée, on retire à cette variable toute l'information concernant son échelle ou unité, et sa *localisation*. Il ne reste plus que des informations sur la **forme** de la distribution. Cette transformation permet de comparer plusieurs variables sur le plan de la forme, même si ce sont des variables exprimées dans des échelles différentes ou qui ont des moyennes complètement différentes.

Exemple : Variable centrée réduite associée à la variable « ciné », à la variable « taille ».

Autre utilisation : Etant donné un individu i pour lequel la variable prend la valeur x_i, on peut situer cet individu dans l'ensemble des observations en calculant son écart à la moyenne réduit :

$$\frac{x_i - \bar{x}}{\sigma_X}.$$

Exemple : quel est l'écart à la moyenne, mesuré en écart-types, d'un individu mesurant 177 cm ?

2.4 Boîtes à moustaches

La boîte à moustaches est une représentation graphique qui permet de visualiser les quartiles ainsi que la dispersion des données et de repérer les données extrêmes ou *outliers*. Elle se fait couramment pour les variables quantitatives continues ou pour les variables quantitatives discrètes prenant un grand nombre de valeurs différentes. En revanche, elle n'a pas beaucoup d'intérêt pour une variable discrète prenant peu de valeurs différentes.

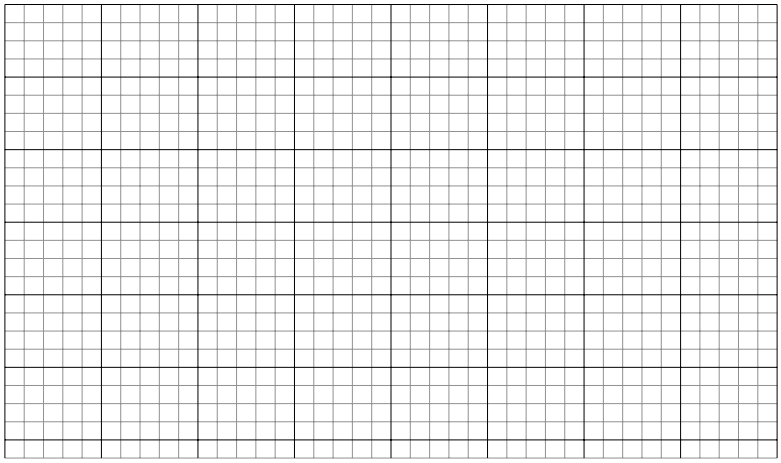
Elle est constituée :

- d'une **boîte** dont les bornes sont les premier et troisième quartile Q_1 et Q_3. A l'intérieur de la boîte figure la médiane Q_2.
- de **moustaches**. On définit tout d'abord deux bornes : $b_- = Q_1 - 1.5(Q_3 - Q_1)$ et $b_+ = Q_3 + 1.5(Q_3 - Q_1)$. On note m_{inf} la plus petite observation supérieure à m_-, et m_{sup} la plus grande observation inférieure à m_+. Soit :

$$\begin{aligned} m_{inf} &= \min\{x_i : x_i \geq b_-\}, \\ m_{sup} &= \max\{x_i : x_i \leq b_+\}, \end{aligned}$$

  La moustache inférieure est le segment $[m_{inf}; Q_1]$. La moustache supérieure, de la même manière, est le segment $[Q_3; m_{sup}]$.
- des **données extrêmes** éventuelles : les observations qui sont en dehors de la boîte et des moustaches, c'est à dire : supérieures à m_+ ou inférieures à m_-. On place ces données une à une quand on en dispose.

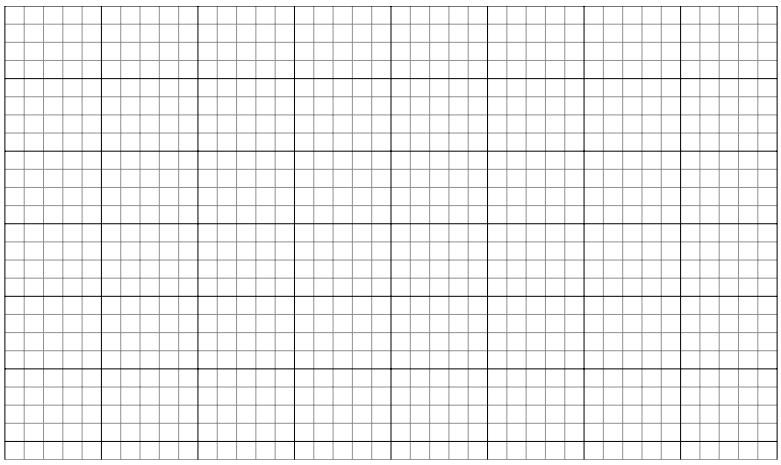
Exemple : Boîte à moustaches de la variable « taille »à partir de la série statistique de 20 observations.



Dans le cas où on ne dispose pas des données brutes mais seulement des données regroupées en classes, on utilise les extrémités b_0 et b_k de la première et de la k-ème classe.

- La limite inférieure m_{inf} de la moustache inférieure est $\max\{m_-, b_0\}$ et la limite supérieure m_{sup} de la moustache supérieure est $\min\{m_+, b_k\}$.
- On ne peut pas placer les données extrêmes, sauf si elles sont fournies en plus.

Exemple : Boîte à moustaches de la variable « taille »à partir des données regroupées.



2.5 Paramètres de forme

2.5.1 Moments d'une distribution

Définition 1. *On appelle* ***moment à l'origine*** *d'ordre $r \in \mathbb{N}$ le paramètre*

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r.$$

Définition 2. *On appelle* ***moment centré*** *d'ordre $r \in \mathbb{N}$ le paramètre*

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r.$$

Remarque : La moyenne est le moment à l'origine d'ordre 1, et la variance le moment centré d'ordre 2, qui s'exprime également en fonction des moments à l'origine d'ordre 1 et 2 :

Exemple : calculer le moment d'ordre 3 de la variable Taille de manière approchée.

2.5.2 Coefficient d'asymétrie de Fisher

Une distribution est parfaitement **symétrique**, si les valeurs qu'elle prend sont également dispersées de part et d'autre de la moyenne. Dans ce cas, son mode, sa moyenne et sa médiane sont confondues et son histogramme admet un axe de symétrie (symétrie par rapport à la valeur de la moyenne).

On peut également mesurer cette asymétrie au moyen d'un moment.

Le moment centré d'ordre 3 est défini par :

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Ce coefficient peut prendre des valeurs positives, négatives ou nulles. Si la distribution est parfaitement symétrique, le moment d'ordre 3 est nul. L'asymétrie peut se mesurer au moyen du coefficient d'asymétrie de Fisher :

$$\gamma_1 = \frac{m_3}{\sigma_X^3}.$$

Remarque : γ_1 est également le moment d'ordre 3 de la distribution centrée réduite associée à X.

Schéma : asymétrie positive, asymétrie négative.

Exemple : calculer le coefficient d'asymétrie de Fisher pour la variable Taille.

2.5.3 Coefficient d'aplatissement

L'aplatissement (ou *kurtosis* en anglais) est mesuré à l'aide du moment d'ordre 4 par le **coefficient d'aplatissement de Pearson**

$$\beta_2 = \frac{m_4}{\sigma_X^4}$$

ou par le coefficient d'aplatissement de Fisher

$$\gamma_2 = \beta_2 - 3 = \frac{m_4}{\sigma_X^4} - 3.$$

Une distribution est dite :
- **mésokurtique** si $\gamma_2 \simeq 0$ (la densité de la loi normale, ou courbe en cloche, est mésokurtique).
- **leptokurtique** si $\gamma_2 > 0$: son histogramme est plus pointu et possède des queues plus longues.
- **platykurtique** si $\gamma_2 < 0$: son histogramme est plus arrondi et possède des queues plus courtes.

Schéma : aplatissement d'une distribution.

2.6 Courbe et indice de Gini

Contexte : on dispose d'observations d'une variable décrivant une quantité correspondant à une richesse, par exemple :
- chiffre d'affaires,
- patrimoine,
- salaires.

On se pose alors des questions sur la répartition égalitaire ou non de ces richesses : les salaires sont-ils répartis de manière égalitaire ? le chiffre d'affaire des entreprises d'un secteur est-il concentré entre les mains d'une poignée d'entreprises au détriment de toutes les autres ? Quelle fortune représentent les 3% de familles les plus riches de France ?

2.6.1 Courbe de concentration de Lorentz

Dans toute la suite, X désigne une variable quantitative regroupée en k classes de centres respectifs $c_1, \ldots, c_k$. Les efectifs des observations dans chaque classe sont $n_1, \ldots, n_k$. L'effectif total est n.

Définition 3. *On appelle* ***masses cumulées*** *du caractère X les quantités définies pour $1 \leq j \leq k$ par :*

$$\sum_{l=1}^{j} n_l c_l.$$

On appelle ***masses cumulées relatives*** *du caractère X les rapports définis pour $1 \leq j \leq k$ par :*

$$M_j = \frac{\sum_{l=1}^{j} n_l c_l}{\sum_{l=1}^{k} n_l c_l}.$$

On appelle ***masse totale*** *le dénominateur :* $\sum_{l=1}^{k} n_l c_l$.

On peut remarquer que la k-ième masse totale M_k est égale à 1.
Remarque : la masse totale intervient dans le calcul approché de la moyenne $\frac{1}{n}\sum_{l=1}^{k} n_l c_l$.

Rappel : les fréquences cumulées Φ_j sont définies par :

$$\begin{aligned}\Phi_j &= f_1 + \cdots + f_j \\ &= \frac{n_1}{n} + \cdots + \frac{n_j}{n} \\ &= \frac{\sum_{l=1}^{j} n_l}{\sum_{l=1}^{k} n_l}.\end{aligned}$$

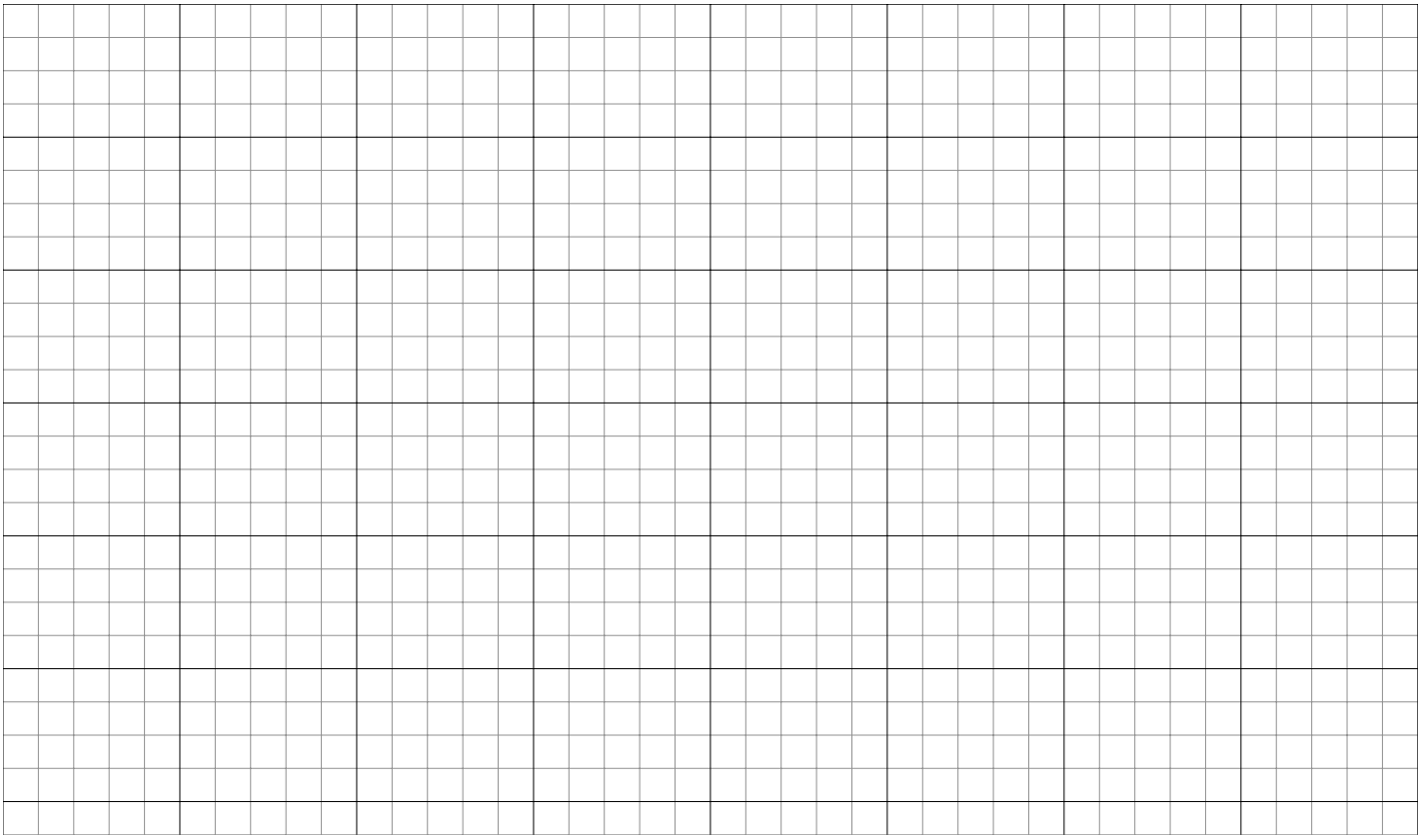
Définition 4. *On appelle* ***courbe de concentration de Gini ou de Lorentz*** *la ligne polygonale qui joint les points O et $P_1, \ldots, P_k$ définis par leurs coordonnées :*

$$O = (0,0), P_1 = (\Phi_1, M_1), P_2 = (\Phi_2, M_2), \ldots, P_k = (\Phi_k, M_k) = (1,1).$$

Remarque : dans le cas d'une variable discrète on définira les mêmes quantités en faisant jouer aux valeurs $v_1, \ldots, v_k$ le rôle des centres de classes.

Exemple : Dans le tableau suivant, on donne la distribution des effectifs du chiffre d'affaire en milliers d'euros d'entreprises ayant participé à une enquête.

Chiffre d'affaires	[0,5[	[5,20[	[20,100[	[100, 500[	[500, 1000[
Effectif n_j	48	16	8	4	4
Centre de classe c_j	2.5	12.5	60	300	750
$n_j c_j$	120	200	480	1200	3000
$\sum_{l=1}^{j} n_l c_l$	120	320	800	2000	5000
M_j	0.024	0.064	0.16	0.4	1
Effectif cumulé $\sum_{l=1}^{j} n_l$	48	64	72	76	80
Fréquence cumulée Φ_j	0.6	0.8	0.9	0.95	1



Propriété : La courbe se situe au-dessous de la diagonale d'équation $y = x$.
Preuve :

Interprétation de la courbe de Gini : plus la courbe est éloignée de la diagonale, plus la série observée est inégalitaire (cf. l'indice de Gini).
Un exemple d'interprétation du tableau : les 10% d'entreprises à plus fort chiffre d'affaires concentrent 84% du chiffre d'affaires total.

2.6.2 Notion de médiale

Soit G la fonction affine par morceaux définie sur $[b_0, b_k[$ de la manière suivante :

$$G(b_0) = 0, G(b_j) = M_j \text{ pour } 1 \leq j \leq k \text{ (et en particulier } G(b_k) = 1),$$

et G est affine sur chaque intervalle $[b_{j-1}, b_j[$, c'est-à-dire : pour $x \in [b_{j-1}, b_j[$,

$$G(x) =$$

Illustration de la formule par le théorème de Thalès :

Définition 5. *La **médiale** de la série, notée Mle, est la solution de*

$$G(Mle) = 0.5$$

Exemple : médiale du chiffre d'affaires des entreprises.

Comparaison avec la médiane et interprétation des deux quantités
Calcul de la médiane : on note F la fonction de répartition empirique de la variable "chiffre d'affaire". $Q_2 \in [0,5[$ car $F(0) = 0 < 0.5 < F(5) = 0.6$. De l'équation de F sur l'intervalle $[0,5[$, on déduit :

$$Q_2 = 0 + (5-0) \times \frac{0.5-0}{0.6-0} \simeq 4.2 \text{ milliers d'euros.}$$

Interprétation : 50% des entreprises ont un chiffre d'affaires inférieur à 4.2 milliers d'euros.

Interprétation de la médiale : la médiale est toujours supérieure à la médiane. C'est-à-dire, on a toujours

$$Mle \geq Q_2.$$

Preuve :

Plus la médiale est éloignée de la médiane, plus la série est inégalitaire. On peut mesurer l'écart relatif entre la médiale et la médiane :

$$\theta = \frac{Mle - Q_2}{Q_2}.$$

Lorsque θ est grand, la série est très inégalitaire (concentrée), si θ est petit, la série est plutôt égalitaire (peu concentrée).

2.6.3 Indice de Gini

Définition 6. *On appelle* ***indice de Gini****, noté I_G, le réel : $I_G = 2S$ où S est la surface de la partie du carré unité délimitée par la courbe de Gini et la diagonale d'équation $y = x$ (cf. représentation graphique de la courbe de Gini).*

Interprétation : on a toujours

$$0 \leq I_G \leq 1.$$

Plus I_G est élevé, plus la série est inégalitaire (la courbe de Gini est d'autant plus éloignée de la diagonale).

Calcul pratique : méthode des trapèzes
$I_G = 2S = 2(0.5 - A) = 1 - 2A$, où A est l'aire située dans le carré unité en-dessous de la courbe de Gini. L'aire A se décompose en une somme d'aires de trapèzes.
Rappel : l'aire d'un trapèze de hauteur h, de petite et grande bases b et B est : $\frac{1}{2}(b+B) \times h$.
On en déduit :

$$I_G = 1 - \sum_{j=1}^{k} (\Phi_j - \Phi_{j-1})(M_j + M_{j-1})$$

avec la convention : $\Phi_0 = 0$, $M_0 = 0$.

Calcul dans l'exemple du chiffre d'affaires d'entreprises

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 7. *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 8. *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

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

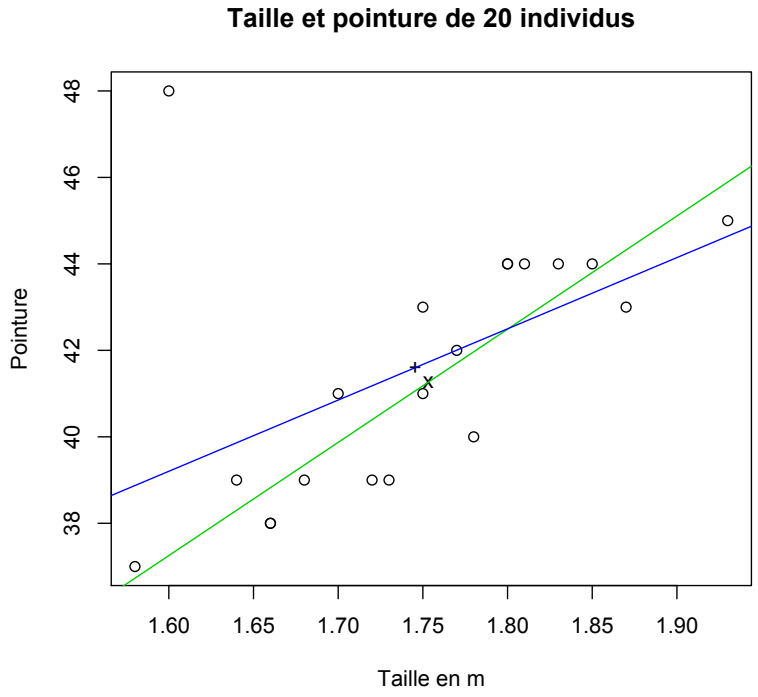
La solution, qui s'obtient en annulant les dérivées partielles de la fonction de (a, b) qui est minimisée dans (3.3), est :

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

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 :

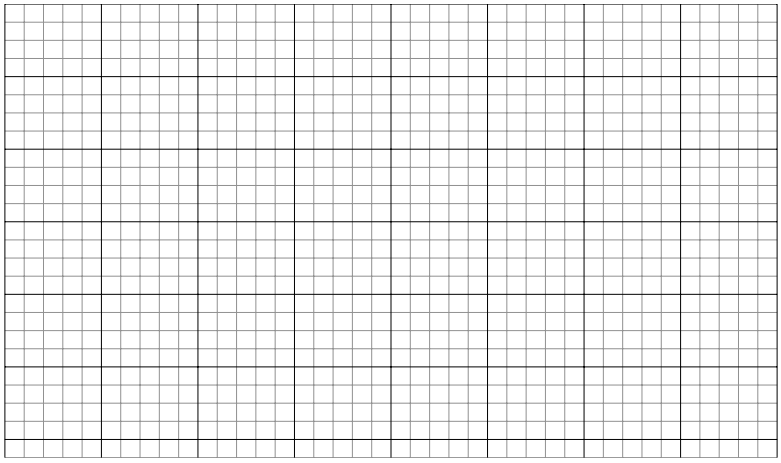
Y / $X_{\|Y}$	D_1	D_2	...	D_J
C_1	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	...	$n_{1J}/n_{.J}$
C_2	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	...	$n_{2J}/n_{.J}$
...	...	...	...	...
C_I	$n_{I1}/n_{.1}$	$n_{I2}/n_{.2}$	...	$n_{IJ}/n_{.J}$
$Total$	1	1	...	1

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

3.2.4 Représentation graphique

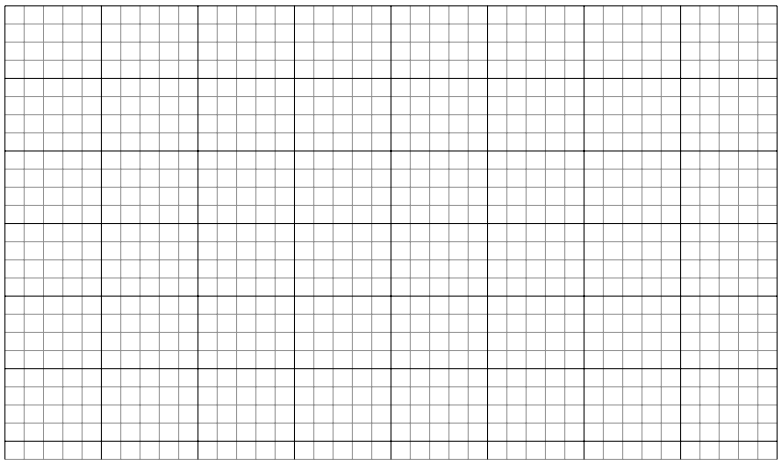
a) Distribution jointe

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 9. *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 10. *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 11. *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 12. *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 :

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

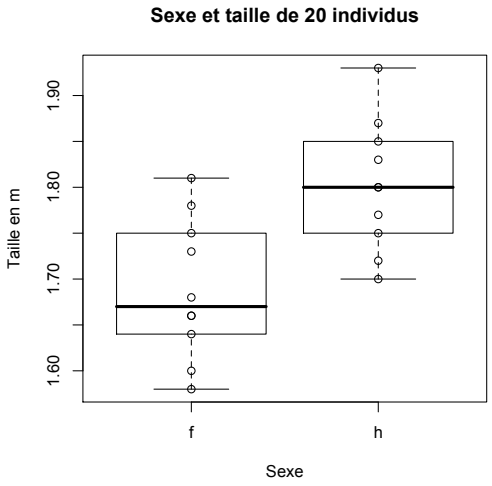
Exemple :

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 13. *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 14. *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 15. *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.