Statistique descriptive

La statistique descriptive est l'ensemble des outils qui permettent de structurer l'information contenue dans les données, de les visualiser et de les résumer.

1 Vocabulaire

1.1 Type de variable

Soit X la variable d'intérêt sur une certaine population.

Exemple. : Etude des conditions de vie des familles résidant à Paris. Cette étude porte notamment sur les variables suivantes :

1. Catégorie socio-professionnelle des plus de 18 ans,
2. Nombre de personnes par ménage,
3. Superficie du logement des familles...

Deux types de variables :
- variables qualitatives (catégorie socio-professionnelle)
- variables quantitatives :
  - soit discrètes si elles ne prennent qu'un nombre fini ou dénombrable de valeurs (nombre de personnes par ménage)
  - soit continues si elles peuvent prendre des valeurs dans $\mathbb{R}$ tout entier, ou dans un intervalle de $\mathbb{R}$. (Superficie du logement)

1.2 Echantillon

La population étant large, on travaille à partir d'un échantillon de n individus représentatifs de la population. Le nombre n est la taille de l'échantillon. La valeur x_i prise par la variable X pour le i-ème individu s'appelle observation i. C'est sur ces valeurs $x_1, \dots, x_n$ appelées observations ou données, que porte l'étude.
En général les données sont très nombreuses. La statistique descriptive est une étape incontournable de toute étude statistique, qui va permettre :

- de visualiser rapidement les données en les résumant de manière synthétique par des représentations graphiques, des tableaux, et par certains indicateurs statistiques
- de détecter d'éventuels problèmes : valeurs manquantes, données atypiques...
- de suggérer des modèles statistiques adaptés aux données.
  On suppose que les observations $x_1, \dots, x_n$ sont des réalisations de n variables aléatoires $X_1, \dots, X_n$ i.i.d de même loi que X (et donc que les données ont bien été recueillis de façon indépendante, dans les mêmes conditions expérimentales ...) On ne connait pas la loi de la variable X. Tout ce dont on dispose, c'est de n réalisations de cette variable. Outre le fait de visualiser rapidement les données, un autre objectif de la statistique descriptive, est justement d'intuiter un modèle statistique adapté aux données : est-ce-que au vu des n observations, la variable X peut être modélisée par exemple par une loi normale ?

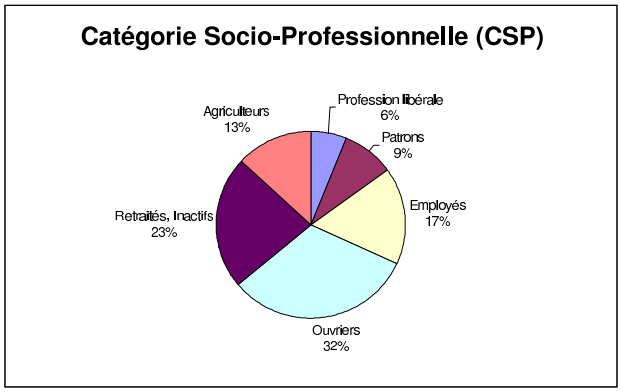
1.3 Effectifs, fréquences

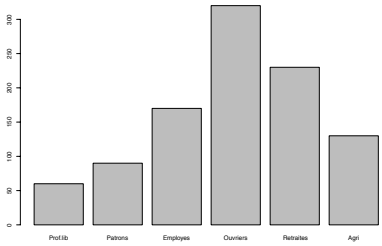
Lorsque la variable X est qualitative ou quantitative discrète, alors en général on a beaucoup de fois la même valeur. On préfère alors parler en terme de modalités et de fréquence.
- modalités : ensemble des valeurs que peut prendre la variable.
- effectif N_i d'une modalité : nombre d'individus pour lesquels la variable X prend cette modalité.
- fréquence d'une modalité : $f_i = \frac{N_i}{n} \in [0, 1]$.

2 Représentation graphique des données

2.1 Variable qualitative

Diagramme circulaire (camembert) ou diagramme en bâtons

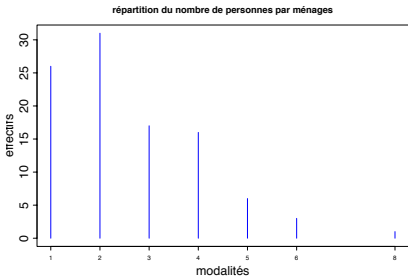




► Commande sous R : `pie` ou `barplot`

2.2 Variable quantitative discrète

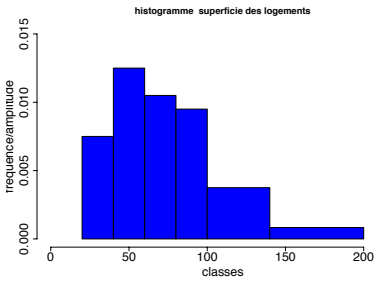
Diagramme en bâtons



► Commande sous R : `barplot`

2.3 Variable quantitative continue

Histogramme :



1. Soit en choisissant nous même les classes qui nous semblent adaptées, soit en laissant le logiciel faire un choix automatique. Le choix du nombre de classes dépend du nombre de données n et de la dispersion des données. Un trop petit nombre de classes fait perdre de l'information sur la série et un trop grand nombre aboutit à de nombreuses classes vides.
2. Chaque classe est représentée par un rectangle dont l'aire (et non la hauteur car les classes ne sont pas forcément d'amplitude égales) est égale à l'effectif ou à la fréquence de la classe.

► Commande sous R : `hist`

3 Résumé numérique d'une variable quantitative

But : résumer l'information contenue dans les données en quelques paramètres, et permettre de comparer plusieurs séries statistiques.
On distingue les paramètres dits de position/de tendance centrale, et ceux dits de dispersion.
Remarque : Tous ces indicateurs numériques n'ont de sens que pour des variables quantitatives (discrètes ou continues) à l'exception du mode que l'on peut définir pour tous les types de variables.

3.1 Paramètres de position

— Mode : valeur de la variable la plus fréquemment observée. Lorsque les données sont regroupées en classe, on parle de classe modale.
  . Le mode peut ne pas être unique. Si le mode est unique on parle de distribution unimodale, s'il est double on parle de distribution bimodale.
  . Une distribution multimodale peut être un indicateur du fait que l'on peut distinguer plusieurs sous-populations par rapport à la variable étudiée.

— Moyenne (empirique) des données $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Attention à ne pas confondre la moyenne empirique (qui est calculée à partir de données) avec l'espérance. L'espérance d'une variable aléatoire est une grandeur théorique, basée sur la loi de la variable aléatoire (si on ne connait pas la loi, on ne peut pas calculer son espérance). La moyenne empirique est calculée à partir d'un échantillon de n observations de la variable. Un autre échantillon de taille n donnera une autre valeur pour la moyenne empirique (proche a-priori, mais différente) de celle calculée sur le premier échantillon. La moyenne empirique est une estimation de l'espérance de la variable aléatoire.
▶ Commande sous R : `mean`

— Médiane (empirique) m_e : valeur qui partage la série d'observations (préalablement rangée par ordre croissant) en deux sous ensembles de même effectif.
▶ Commande sous R : `median`

— Quartiles (empiriques) $Q_1, Q_2 = m_e$ et Q_3 : partagent la série de données en quatre sous ensemble de mêmes effectifs : Q_1 est la valeur de la série telle qu'au moins 25% des observations ont une valeur qui lui est inférieure ou égale. Idem pour Q_3 avec 75%.
▶ Commande sous R : `quantile`

Remarques :

— Ne pas confondre la médiane empirique, et les quartiles empiriques avec la médiane théorique et les quartiles théoriques qui sont à partir de la fonction de répartition, i.e de la loi, de la variable aléatoire. Même si on oublie souvent le terme "empirique", bien comprendre la différence !

— La moyenne a l'inconvénient d'être sensible à des valeurs aberrantes, mais c'est le critère le plus utilisé car se calcule facilement.

— La médiane a justement l'avantage d'être peu sensible aux valeurs extrêmes qui peuvent ne pas être fiables, et elle peut donc être plus pertinente que la moyenne. Mais elle a le gros inconvénient de mal se prêter aux calculs algébriques.

3.2 Boxplot ou boîte à moustaches

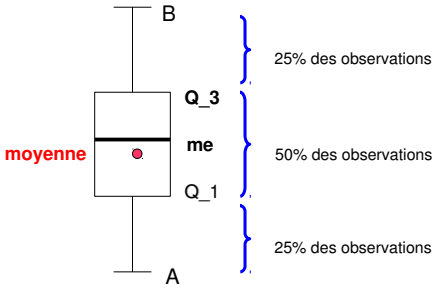
C'est un graphique permettant de résumer un ensemble de données par les paramètres de position.

Sur un axe vertical (ou horizontal), on place les 3 quartiles $Q_1, Q_2 = me$ et Q_3, ainsi que les valeurs extrêmes A et B. Ces valeurs extrêmes A et B peuvent être les valeurs minimales et maximales des données, mais dans la plupart des logiciels elles sont définies par :

$$A = min\{x_i : x_i \geq Q_1 - 1.5(Q_3 - Q_1)\}$$

$$B = max\{x_i : x_i \leq Q_3 + 1.5(Q_3 - Q_1)\}$$

On trace ensuite un rectangle de hauteur $Q_3 - Q_1$ et dont la largeur est proportionnelle à la racine carrée de la taille n de l'échantillon. On peut superposer la valeur de la moyenne des données sur ce boxplot.
▶ Commande sous R : `boxplot`



Interprétation du boxplot :

— On peut ainsi comparer le boxplot d'un ensemble d'observations d'une variable X, avec celui auquel devrait ressembler un ensemble de données issues d'une variable gaussienne (loi symétrique donc médiane $\sim$ moyenne), et en déduire visuellement si la variable considérée X peut être modélisée par une loi gaussienne.

— A et B représentent les limites de l'intervalle en dehors duquel les données sont considérées comme atypiques (on les appelle aussi outliers).

— Ce graphique permet aussi de comparer facilement plusieurs séries de données.

3.3 Paramètres de dispersion

Avec les paramètres de dispersion, on ajoute une information sur la variabilité des observations.

— étendue : $\max(x_i) - \min(x_i)$.

— écart interquartile : $Q_3 - Q_1$.

— variance (empirique) : $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, et l'écart-type (empirique) s_x définie comme sa racine carrée.

► Commande sous R : `var` pour la variance empirique et `sd` pour l'écart-type empirique

Remarques :

. 50% des observations se trouvent dans l'intervalle interquartile $[Q_1; Q_3]$, donc plus l'écart interquartile est petit, plus la distribution est "concentrée". On peut définir l'intervalle interdécile qui lui regroupe 80% des observations autour de la médiane.

4 Ajustement à une loi

On a vu des tests qui nous permettent de vérifier que nos données s'ajustent bien à une loi donnée ! Kolmogorov-Smirnov (loi générale) et Shapiro-Wilk (loi normale).
Un outil de statistique descriptive fréquemment utilisé est le *qqplot*, notamment pour visualiser un ajustement à une loi normale.

qqplot

Cas d'adéquation à une variable aléatoire gaussienne :
Si X est une variable aléatoire gaussienne $\mathcal{N}(\mu, \sigma^2)$, alors

$$\mathbb{P}(X \leq x) = \mathbb{P}(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = \phi(\frac{x-\mu}{\sigma})$$

où ϕ est la fonction de répartiton d'une $\mathcal{N}(0,1)$.

$$\Rightarrow \phi^{-1}(\mathbb{P}(X \leq x)) = \frac{x-\mu}{\sigma}$$

Notons u_i le quantile de la loi normale centrée réduite d'ordre $\mathbb{P}(X \leq x_i)$, ie tel que

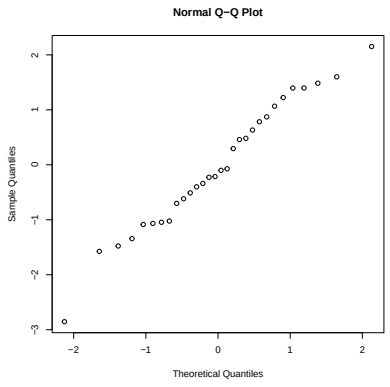
$$\phi(u_i) = \mathbb{P}(X \leq x_i) \Leftrightarrow u_i = \phi^{-1}(\mathbb{P}(X \leq x_i))$$

Donc si la variable X est gaussienne, on a :

$$u_i = \frac{x_i - \mu}{\sigma}$$

On trace le nuage de points (x_i, u_i). Si les données sont gaussiennes, les points devraient être alignés selon la droite d'équation $y = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$. Cette droite s'appelle la droite de Henry.
En pratique, en supposant les données $x_{(1)}, \ldots, x_{(n)}$ rangées par ordre croissant, on a $\mathbb{P}(X \leq x_{(i)}) \approx \frac{i}{n}$.

► Commande sous R : `qqnorm` pour une loi normale et `qqplot` permet de généraliser à d'autres lois que la loi normale.



5 Conclusion sur cette étape d'analyse descriptive

Cette étape des statistique descriptive, même si elle peut sembler rudimentaire du fait des outils plutôt triviaux utilisés, est indispensable ! Il ne faut pas la négliger et chercher trop rapidement à mettre en oeuvre des méthodes statistiques plus sophistiquées.

- La représentation graphique des données et le calcul d'indicateurs statistiques permettent de se faire une première idée d'un jeu de données.
- Cette étape permet de repérer d'éventuels problèmes comme la présence de données manquantes. Dans ce cas, faut-il supprimer les individus ou variables incriminés ? Ou faut-il compléter les valeurs manquantes pas une prévision partielle ? Cela dépend du taux de valeurs manquantes, de leur répartition, et de la robustesse des méthodes que l'on souhaite utiliser...
- Cette étape permet de repérer la présence de données atypiques qui peuvent influencer/fausser la future analyse statistique. Faut-il alors supprimer ces données ou adopter des méthodes statistiques plus robustes aux valeurs extrêmes (typiquement une médiane est plus robuste qu'une moyenne).
- Cette étape (notamment le fait de tracer des histogrammes et boxplot) permet de donner des informations sur la forme de la distribution observée (symétrie, dispersion...). En les comparant avec les distributions de variables aléatoires de loi connue (typiquement la loi Gaussienne...), cela permet de suggérer un modèle adapté aux données.
- Lorsque la distribution des données ne semblent pas symétriques, on peut préférer transformer les données (prendre le log des données typiquement). Cela peut permettre d'améliorer la symétrie de la distribution, voire de se rapprocher d'une distribution gaussienne.

Remarque. Un package pour faire de jolis graphes : ggplot2

Une fois cette étape descriptive réalisée on peut

- utiliser des méthodes de statistique inférentielle pour valider le modèle intuité : par des tests statistiques tels que le test de Kolmogorov-Smirnov, de Shapiro...
- tirer des conclusions des données en utilisant des tests (comparaison de moyenne, indépendance, etc...)

6 Analyse bidimensionnelle

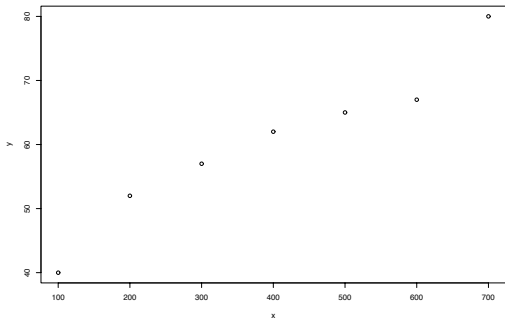
Etude simultanée de deux variables X et Y étudiées sur le même échantillon.

6.1 Deux variables quantitatives

6.1.1 Graphiquement : le nuage de points

Tableau de données. Exemple : n=7 parcelles :
x_i : quantité d'engrais (quintaux) sur la parcelle i
y_i : rendement de blé (kg) sur la parcelle i

x_i	100	200	300	400	500	600	700
y_i	40	52	57	62	65	67	80



Si ce nuage paraît dispersé dans toutes les directions, rien ne laisse penser qu'il existe un lien entre X et Y. Mais si ce nuage paraît concentré autour d'une courbe particulière, on peut penser que X et Y sont d'une certaine façon liées.

▶ Commande sous R : `plot`

6.1.2 Indicateur numérique de liaison

Définition. On définit la covariance (empirique) entre deux séries de données $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ par

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

C'est la version empirique (cad calculée à partir de n observations) de la covariance (théorique) définie par $E\big[(X - E(X))(Y - E(Y))\big]$.

▶ Commande sous R : `cov`

Interprétation : la covariance est un réel positif si, quand X est supérieur à sa moyenne, Y l'est en moyenne aussi, et est un réel négatif quand c'est le contraire. Une covariance positive indique donc que les deux variables X et Y ont tendance à varier dans le même sens et une covariance négative qu'ils ont tendance à varier en sens contraires. L'inconvénient de la covariance est qu'elle dépend des unités choisies pour mesurer X et Y. C'est pourquoi on définit le coefficient de corrélation.

Définition. On définit le coefficient de corrélation linéaire par

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

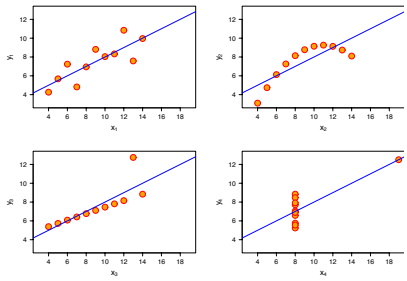
Le coefficient de corrélation ne dépend pas des unités de mesure de X et de Y . Il est symétrique : $r_{xy} = r_{yx}$ et prend ses valeurs entre -1 et $+1$: $-1 \leq r_{xy} \leq 1$

Interprétation du coefficient de corrélation linéaire :

- Si $|r_{xy}| \sim 1$, le nuage de points est presque aligné sur une droite : on dira que les variables aléatoires X et Y sont fortement corrélées linéairement (positivement si $r_{xy} > 0$ et dans ce cas le nuage est presque aligné sur une droite de pente positive, négativement si $r_{xy} < 0$).
- Si $|r_{xy}| \sim 0$, X et Y ne sont pas corrélées linéairement. C'est le cas si X et Y sont indépendantes statistiquement, mais ce n'est pas le seul cas : il peut aussi exister une relation non linéaire entre X et Y.

▶ Commande sous R : `cor`

Attention : il est possible de fabriquer des données pour lesquelles le coefficient de corrélation linéaire entre les deux variables est proche de 1, alors que le nuage de points ne ressemble pas à une droite. Il est donc primordial de tracer le nuage de points et d'examiner si un lien linéaire est envisageable entre les deux variables, avant d'interpréter le coefficient de corrélation linéaire.



Données d'Anscombe. Pour les 4 graphes, $r_{xy} \sim 0.81$

Remarque : Un coefficient de corrélation proche de 1, peut signifier une forte corrélation linéaire entre X et Y mais ne signifie pas pour autant une relation de cause à effet entre X et Y (une troisième variable peut expliquer l'augmentation ou la diminution conjointe de X et Y).

6.1.3 Test de corrélation

Une fois le nuage de points réalisés (ce qui permet d'avoir une idée du type de relation que l'on cherche, d'avoir pu supprimer des données atypiques...), on peut vouloir tester la corrélation entre 2 variables. Plusieurs tests existent :

- Le test de corrélation de Pearson est basé sur le coefficient de corrélation linéaire. Il n'est applicable que si les variables X et Y sont gaussiennes. On rappelle que l'absence d'une relation linéaire en signifie pas l'absence de toute relation !
  Ici on va tester :
  H_0 Il n'y a pas de corrélation linéaire entre X et Y
  H_1 Il y a une corrélation linéaire entre X et Y
- Les tests de corrélation de Spearman et de Kendall (en général moins performant que que le premier, sauf en cas de beaucoup d'ex-aequo) sont des tests de corrélation non-paramétriques basé sur le rang. Ils peuvent donc être utilisé lorsque X ou Y ne sont pas gaussiennes. Ils permettent aussi de détecter l'existence de relations monotones (croissante ou décroissante) quelle que soit leur forme précise (linéaire mais pas seulement !). Ce coefficient est donc utile lorsque l'analyse du nuage de point révèle une forme curviligne dans une relation qui semble mal s'ajuster à une droite. Ici on va tester :
  H_0 Il n'y a pas de corrélation entre X et Y
  H_1 Il y a une corrélation entre X et Y

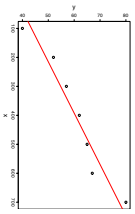
▶ Commande sous R : `cor.test`

6.1.4 Régression

Objectif : On cherche à "expliquer" Y en fonction de X, c'est-à-dire à exprimer une dépendance du type $Y = f(X)$. Cette fois, on n'a pas donc une relation symétrique : on veut tester si X influe sur Y ! L'allure du nuage de points permet de choisir le type de dépendance f : linéaire $y = f(x) = ax + b$; polynomiale, $y = f(x) = ax^2 + bx + c$; exponentielle...
Bien sûr, la relation n'est pas exacte, mais on cherche à faire passer une courbe simple (droite ou autre) au plus près du nuage de points. Cette courbe pourra être utilisée pour indiquer une "tendance" et faire certaines prévisions.

Idée de la régression linéaire : On cherche la droite d'équation $y = a + bx$ la plus "proche" des n points (x_i, y_i).

▶ Commande sous R : `lm`

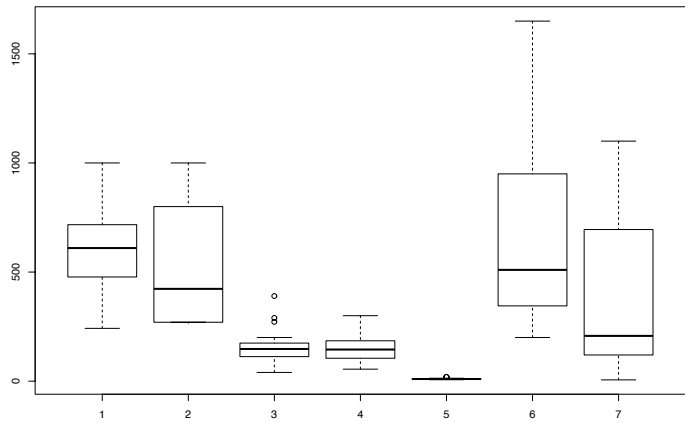


On ne peut en effet pas trouver une droite qui passe par les n points. On cherche donc à ce que l'écart entre l'observation y_i et la valeur donnée par la droite $y = a+bx$ soit le plus petit possible, et cela pour tous les points en même temps.
→ Méthode des moindres carrés, voir cours Modèle linéaire.

Autres types de régression : le nuage de points peut laisser apparaître une autre forme de dépendance entre X et Y qu'une relation linéaire (cad autre qu'une droite). Par exemple, si le nuage des points est proche d'une courbe d'équation $y = ae^{bx}$, on effectue une régression linéaire de $ln(Y)$ par X. De même si le nuage est proche d'une courbe d'équation $y = aln(x) + b$, on effectue une régression linéaire de Y par $ln(X)$.

6.2 Une variable quantitative et une variable qualitative

On peut représenter un boxplot par modalité de la variable qualitative (ou quantitative discrète). L'analyse inférentielle correspondante sera vue au 2e semestre dans le cours de Modèle Linéaire.



Boxplot de la variable poids pour 7 espèces de poisson.

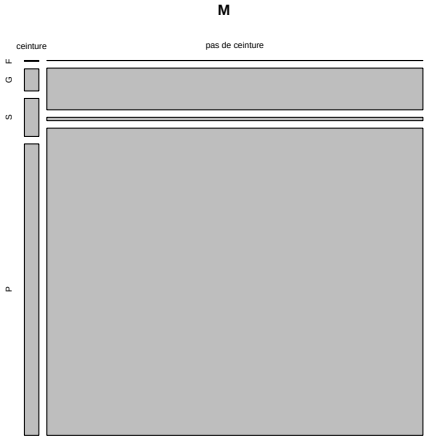
6.3 Deux variables qualitatives

Soient X et Y deux variables aléatoires qualitatives prenant respectivement les modalités $\{a_1, \ldots, a_k\}$ et $\{b_1, \ldots, b_m\}$. La distribution jointe de X et de Y est donnée dans un tableau à double entrée appelée table de contingence.
Exemple : Etude de la liaison entre le port de la ceinture de sécurité et le degré de gravité des blessures. Données sur $n = 10779$ conducteurs ayant eu un accident.

X : degré de gravité des blessures ($k = 4$ modalités)
Y : port ou non de la ceinture ($m = 2$ modalités)
N_{ij} : nombre d'individus avec modalités a_i et b_j
$N_{i\bullet}$: nombre d'individus avec modalité a_i
$N_{\bullet,j}$: nombre d'individus avec modalité bj

	ceinture	Pas ceinture	Total
Blessures Fatales	$N_{11} = 1$	$N_{12} = 43$	$N_{1\bullet}$
Graves	$N_{21} = 4$	$N_{22} = 98$	$N_{2\bullet}$
sérieuses	25	330	$N_{3\bullet}$
Pas ou peu de blessures	1229	9049	$N_{4\bullet}$
Total	$N_{\bullet 1}$	$N_{\bullet 2}$	1

► Commande sous R pour représenter une matrice de contingence : `mosaiicplot`



6.4 Vers le cas multi-dimensionel

Lorsqu'on étudie p variables, on peut représenter p^2 sous-graphes sous forme d'un "tableau" de taille $p \times p$. Chaque sous-graphe correspond au croisement entre 2 variables. Voir TP.

Matrice des covariances et des corrélations

Lorsqu'on étudie p variables quantitatives, on peut calculer les variances de toutes les variables, et les covariances 2 à 2. L'ensemble de ces valeurs est alors disposé dans une matrice (p, p) symétrique, comportant les variances sur la diagonale, et les covariances à l'extérieur de la diagonale. Cette matrice s'appelle la matrice des variances-covariances (ou matrice des covariances). Si on ne dispose pas de la loi des variables, on peut obtenir la matrice des variances-covariances empiriques dans laquelle sont données les variances empiriques et les covariances empiriques. On note cette matrice S.

On peut également construire la matrice des corrélations : c'est une matrice (p, p) symétrique, avec des 1 sur la diagonale, et les corrélations entre les variables à l'extérieur de la diagonale. A nouveau, si la loi des variables est inconnue, on a la version empirique de la matrice des corrélations. on la note R.

► Commande sous R : `heatmap`

Vous verrez au S8 comment on peut se ramener à des représentations graphiques facilement interprétables lorsqu'on veut étudier plus de 2 variables. Il s'agit de faire de la réduction de dimension : trouver de nouvelles variables "résumant" les variables initiales. Voir le cours d'ACP.