

Tableaux croisés et diagrammes en mosaïque, pour visualiser les probabilités marginales et conditionnelles

Monique Le Guen

CNRS- MATISSE¹

Résumé

Cet article s'inscrit dans une démarche de sensibilisation aux différentes facettes de la Statistique. La visualisation de l'information par des méthodes graphiques lorsqu'elle s'appuie sur les Nouvelles Technologies de l'Information et de la Communication, apparaît comme une voie prometteuse vers une meilleure compréhension des concepts abstraits de la Statistique.

Notre propos est axé sur l'aspect visuel des tableaux croisés représentés par des diagrammes en mosaïque. Après avoir replacé les graphiques en Statistique, nous présentons sur un exemple les tableaux croisés à double entrée. Cet exemple nous permet d'introduire le vocabulaire et les différents éléments statistiques, effectifs, probabilités marginales, probabilités conditionnelles, repérables sur un tableau croisé.

Nous montrons l'apport selon les situations, des représentations visuelles offertes par les diagrammes en barres, les diagrammes en bandes et les diagrammes en mosaïque. Nous terminons sur les prolongements en cours de développement autour des diagrammes en mosaïque, et les logiciels interactifs. Les références citées en fin d'article donnent des liens vers des articles et des logiciels accessibles par internet.

Mots Clés

Visualisation, NTIC, tableaux croisés, probabilités marginales, probabilités conditionnelles, diagrammes en barres, diagrammes en bandes, diagrammes en mosaïque.

Sommaire

INTRODUCTION.....	2
LA PLACE DES GRAPHIQUES EN STATISTIQUE	2
LES TABLEAUX CROISES A DOUBLE ENTREE	5
LES STATISTIQUES D'UN TABLEAU CROISE.....	6
<i>Les Notations</i>	6
<i>Probabilités Marginales - Distributions Marginales</i>	7
<i>Probabilités Conditionnelles - Distributions Conditionnelles</i>	7
<i>Dépendance-Indépendance</i>	7
REPRESENTATIONS GRAPHIQUES.....	9
<i>Les Diagrammes en Barres (Bar Chart)</i>	9
<i>Les Diagrammes en Bandes</i>	10
<i>Les Diagrammes en Mosaïque (Mosaic Plot)</i>	10
PROLONGEMENTS DES DIAGRAMMES EN MOSAÏQUE.....	13
<i>Représentation en surface des résidus standardisés</i>	13
<i>Représentation en surface des écarts à l'Indépendance</i>	14
CONCLUSION.....	15
RÉFÉRENCES	16

¹ MATISSE-CNRS UMR8595, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris.

Introduction

Cet article s'inscrit dans une démarche de sensibilisation aux différentes facettes de la Statistique. La visualisation de l'information par des méthodes graphiques lorsqu'elle s'appuie sur les Nouvelles Technologies de l'Information et de la Communication, apparaît comme une voie prometteuse vers une meilleure compréhension des concepts abstraits de la Statistique.

L'interactivité entre l'homme et la machine, tout comme l'interactivité entre les graphiques, apportent une aide certaine à l'apprenant. Des gains en efficacité sont attendus.

Ce domaine des NTIC est un champ de recherche très jeune, donc en grande évolution. Les réalisations appliquées à l'éducatif sont très récentes. Tout est à inventer, à concevoir et à réaliser. Quelques équipes universitaires américaines, anglaises, australiennes, allemandes sont en avance dans ce domaine.

Enseigner la Statistique et la méthodologie d'Analyse de Données au niveau théorique comme au niveau pratique, est un véritable défi qui mobilise la communauté des statisticiens au niveau international. Les travaux de HARTIGAN, KLEINER, FRIENDLY, UNWIN, HOFFMAN et bien d'autres, sur les représentations graphiques des variables catégorielles, apportent un nouveau regard et de nouvelles facilités.

Ce document est axé sur l'aspect visuel des tableaux croisés représentés par des diagrammes en mosaïque. Après avoir replacé les graphiques en Statistique au chapitre 1, nous présentons au chapitre 2, les tableaux croisés à double entrée, à partir d'un exemple. Cet exemple nous permet d'introduire au chapitre 3 les différentes statistiques repérables sur un tableau croisé. Au chapitre 4, nous montrons selon les situations, l'apport des représentations visuelles offertes par les diagrammes en barres, les diagrammes en bandes et les diagrammes en mosaïque. Nous terminons sur les prolongements en cours de développement autour des diagrammes en mosaïque, et les logiciels interactifs. Les références citées en fin d'article donnent des liens vers des articles et des logiciels accessibles par internet.

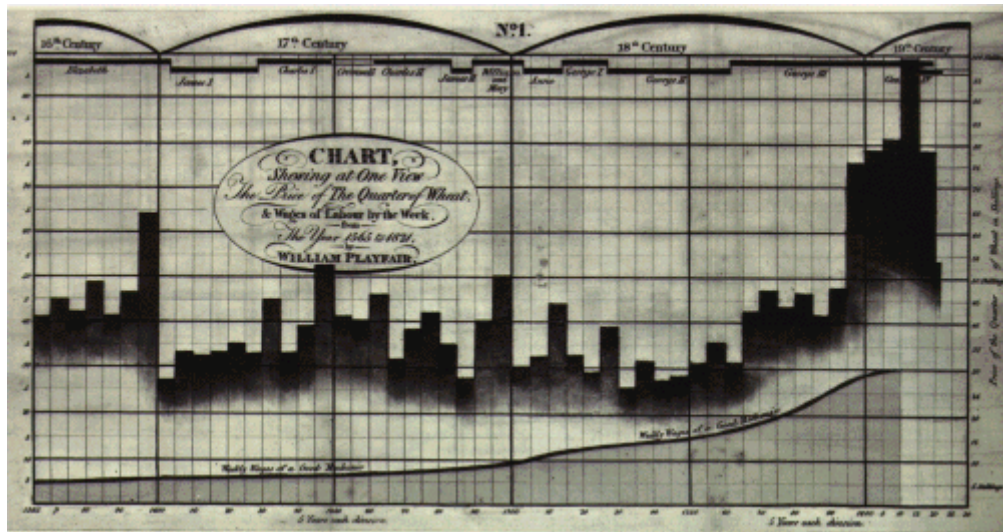
La Place des Graphiques en Statistique

En Mathématique le lien entre l'algèbre et la géométrie ne s'est répandu qu'au XVII^{ème} siècle avec Descartes (1596-1650), inventeur des graphiques cartésiens.

Dans le domaine de la Statistique il fallut attendre l'économiste écossais WILLIAM PLAYFAIR (1759-1823) qui eut l'idée d'allier la statistique quantitative à une représentation visuelle de l'information. Il est l'inventeur des diagrammes en barres (Bar Chart) et des diagrammes en secteurs (Pie Chart). Il est également le concepteur et le réalisateur « *manuel* » de beaux graphiques esthétiques comme le montre le graphique 1 paru en 1821. WAINER (1997) publiera ce graphique, également accessible par Internet, dans son ouvrage « *Visual Revelation* ».

Ce graphique rassemble trois sources d'information (variables) qui permettent des analyses conjointes. L'histogramme du haut schématise les salaires hebdomadaires et la courbe inférieure, le prix du blé, sur une période de 260 années. En haut du graphique sont mentionnés les siècles et les périodes de règnes, des reines et des rois britanniques. On y voit

l'ébauche d'une analyse de séries chronologiques avec la recherche d'une relation de corrélation.



Graphique 1 : réalisé en 1821 par WILLIAM PLAYFAIR « *Chart Shewing at Once View, The price of the quarter of Wheat , & Wages of Labour by the week, from the year 1565 to 1821* », ce graphique montre la relation entre les salaires hebdomadaires (histogramme du haut) et le prix du blé (courbe inférieure) sur une période de 260 années. En haut du graphique sont mentionnés les siècles et les périodes de règnes, des Reines et des Rois britanniques.

<http://www.math.yorku.ca/SCS/Gallery/images/playfair-wheat1.gif>

Les graphiques portant sur l'information quantitative sont de nos jours largement utilisés aussi bien dans la phase d'analyse des données que dans celle de communication. VALOIS J. P. (1999, 2000) propose dans deux articles récents : *Une Typologie des Graphiques Statistiques* et *Approche Graphique en Analyse de Données*, une typologie de classement.

A l'opposé les graphiques portant sur l'information qualitative sont peu diversifiés, peu connus, et donc peu utilisés. Cependant par le biais des nouveaux moyens de production automatique, des recherches innovantes s'intensifient et leur diffusion s'accélère grâce à Internet².

Depuis 20 ans, des noms poétiques: comme les Spine Plot, Mosaic Plot, Sieve Diagram, Fourfold Displays, sont apparus dans la littérature anglo-saxonne qui traite des données catégorielles (*Categorical Data*). L'article en ligne de FRIENDLY "Visualizing Categorical Data: Data, Stories, and Pictures", en dresse une rétrospective.

Les graphiques visualisant les contributions significatives à la statistique du χ^2 dans un tableau croisé ont débuté dans les années 70 avec les travaux de SNEE R. (1974). En 1981 HARTIGAN & KLEINER proposèrent les diagrammes en mosaïques, qui sont des représentations en surface de tableaux croisés à 2 entrées.

Ces diagrammes furent ensuite développés et étendus aux tableaux à n-entrées par FRIENDLY M. (1994, 1995), psychologue et statisticien de l'Université de Toronto (Canada). Des macros

² Si vous voulez accéder à une galerie de graphiques statistiques du 16^{ème} siècle à nos jours, allez voir ce site <http://www.math.yorku.ca/SCS/Gallery/>

écrites en SAS, sont disponibles sur son site Internet. UNWIN A. & HOFMANN H. (2001) informaticiens et statisticiens de l'Université de Augsburg (Allemagne) ont prolongé ces recherches en développant le logiciel MANET destiné au traitement visuel et interactif des variables catégorielles.

Les diagrammes en mosaïque sont peu répandus. Ils sont adaptés à la lecture des tableaux croisés. Leur lecture n'est pas triviale, elle nécessite un apprentissage. La généralisation proposée par FRIENDLY permet de faire le lien avec les modèles Log-linéaires, la régression logistique, et même l'Analyse des Correspondances. Aussi nous proposons dans cet article, une prise de connaissance de ces notions qui devraient conduire le praticien à une meilleure compréhension des modèles log-linéaires.

Traditionnellement l'analyse des tableaux croisés depuis PEARSON repose sur la statistique du χ^2 . Le χ^2 donne un diagnostic global sur la situation de dépendance/indépendance entre les 2 variables. Lorsque le χ^2 est significatif, l'analyste de données souhaite connaître quelles sont les lignes ou les colonnes qui sont responsables des associations.

On peut faire un parallèle entre la statistique du χ^2 pour les tableaux croisés, et la statistique F de Fisher-Snedecor en analyse de variance. Le χ^2 tout comme le F ne renseignent pas sur les sous groupes responsables des différences. L'analyste de données doit ensuite s'intéresser aux comparaisons 2 à 2.

Pour les données catégorielles, les diagrammes en mosaïque vont faciliter ces comparaisons.

Les Tableaux Croisés à Double Entrée

Soit un échantillon de 124 élèves. On relève pour chaque élève la couleur des yeux et la couleur des cheveux, (variables YEUX et CHEVEUX). Le *tableau de contingence* cf. tableau 1, aussi appelé tableau croisé, répartit l'effectif total (124) selon les croisements 2 à 2 des modalités des 2 variables.

Ce tableau permet de présenter le vocabulaire : modalités, effectifs marginaux, ligne et colonne marginales, pourcentages et probabilités, distributions marginales, distributions conditionnelles.

Modalités

Les entêtes de ligne sont les *modalités* (BLEU, MARRON, VERT) de la variable YEUX, les entêtes de colonnes sont les *modalités* (BLOND, BRUN, NOIR, ROUX) de la variable CHEVEUX.

Eléments marginaux

La ligne Total et la colonne Total donnent les *effectifs marginaux*. La ligne marginale donne la distribution (tri à plat) de la variable CHEVEUX sans distinction de la couleur des yeux. La colonne marginale donne la distribution (tri à plat) de la variable YEUX sans distinction de la couleur des cheveux.

YEUX	CHEVEUX				
Frequency	BLOND	BRUN	NOIR	ROUX	Total
BLEU	25	9	3	7	44
MARRON	7	13	8	5	33
VERT	13	17	10	7	47
Total	45	39	21	19	124

Tableau 1 : *Tableau de contingence : croise la couleur des yeux avec la couleur des cheveux.*
Source des données : SCHWARTZ D., (1963).

Pourcentages et Probabilités

A partir des effectifs en marge du tableau croisé on calcule les pourcentages en lignes et en colonnes. On fait de même pour les distributions marginales. Si l'échantillon est représentatif les pourcentages observés sont des estimations des probabilités³, cf. tableau 2.

³ Pour passer de la probabilité estimée sur l'échantillon à la probabilité au niveau de la population il faudrait adjoindre un intervalle de confiance. Mais ce n'est pas notre propos pour l'instant.

YEUX	CHEVEUX				Total	
	BLOND	BRUN	NOIR	ROUX		
	Frequency					
	Percent					
	Row Pct					
Col Pct						
BLEU	25 20.16 56.82 55.56	9 7.26 20.45 23.08	3 2.42 6.82 14.29	7 5.65 15.91 36.84	44 35.48	
MARRON	7 5.65 21.21 15.56	13 10.48 39.39 33.33	8 6.45 24.24 38.10	5 4.03 15.15 26.32	33 26.61	
VERT	13 10.48 27.66 28.89	17 13.71 36.17 43.59	10 8.06 21.28 47.62	7 5.65 14.89 36.84	47 37.90	
Total	45 36.29	39 31.45	21 16.94	19 15.32	124 100.00	

Distribution marginale des lignes

Tableau 2 : *Tableau statistique donnant les effectifs et les pourcentages*

Dans chaque cellule du tableau on repère 4 lignes :

- La 1^{ère} ligne donne l'effectif ,
- La 2^{ème} ligne donne le pourcentage par rapport à l'effectif total
- La 3^{ème} ligne donne le pourcentage en ligne
- La 4^{ème} ligne donne le pourcentage, en colonne

Les Statistiques d'un Tableau Croisé

Les Notations

Afin de faciliter la compréhension des calculs et des statistiques élaborés sur un tableau croisé, une convention dans la notation s'est instaurée au niveau international. Elle est résumée dans le Tableau 3 ci-dessous.

X	<i>Blond</i>	<i>Brun</i>	...	<i>j</i>	...	Total
Y	1	2				
1- Bleu						n_{1+}
2- Marron						n_{2+}
<i>i</i>				n_{ij}		n_{i+}
...						
Total	n_{+1}			n_{+j}		n_{++}

Tableau n°3 : *Notations utilisées dans les tableaux de contingence*

On désigne par :

i : indice du n° de ligne (i varie de 1 à I modalités en ligne)

j : indice de n° de colonne (j varie de 1 à J modalités en colonne)

n_{ij} : effectif de la cellule (i,j)

n_{i+} : effectif marginal de la ligne i définit par $n_{i+} = \sum_j n_{ij}$

n_{+j} : effectif marginal de la colonne j défini par $n_{+j} = \sum_i n_{ij}$

n_{++} : Total global du tableau $n_{++} = \sum_i \sum_j n_{ij}$

On notera que dans un tableau croisé, constitué à partir de données issues d'un échantillon représentatif, les variables X et Y jouent le même rôle. Il n'en serait pas de même si on s'intéressait à des données d'enquêtes prospectives ou rétrospectives. Dans ces deux cas le sens de lecture du tableau importerait.

Probabilités Marginales - Distributions Marginales

Généralisons sous forme symbolique.

Si X désigne la variable aléatoire en colonne et x_j la modalité de rang j de cette variable, la *probabilité marginale* de X se note :

$$\Pr(X = x_j) = p_j = \frac{n_{+j}}{n_{++}}$$

Si Y désigne la variable aléatoire en ligne et y_i la modalité de rang i de cette variable, la *probabilité marginale* de Y se note :

$$\Pr(Y = y_i) = p_i = \frac{n_{i+}}{n_{++}}$$

Par commodité, les modalités de rang i, respectivement j seront notées par la suite *modalité i*, respectivement *modalité j*.

La suite des probabilités marginales en lignes (respectivement en colonnes) définissent la *distribution marginale* en lignes (respectivement en colonnes).

Probabilités Conditionnelles - Distributions Conditionnelles

Les probabilités conditionnelles d'une modalité i de la 1^{ère} variable sachant une modalité j de la 2^{ème} variable (et respectivement une modalité j de la 2^{ème} variable sachant une modalité i de la 1^{ère} variable) se notent :

$$\Pr(i | j) = p_{i|j} = \frac{n_{ij}}{n_{+j}} \quad \text{et} \quad \Pr(j | i) = p_{j|i} = \frac{n_{ij}}{n_{i+}}$$

Ces probabilités seront utilisées pour construire les diagrammes en mosaïques.

Dépendance-Indépendance

Introduisons maintenant les notions d'association, de liaison ou de dépendance/indépendance entre 2 variables. Ces notions portent sur le même concept statistique : les probabilités conditionnelles sont-elles identiques ?

Situation d'indépendance

Si la variable X (couleur des cheveux) est indépendante de la variable Y (couleur des yeux) alors la probabilité d'avoir à la fois la couleur x_i des yeux et la couleur y_j des cheveux ne dépend que du produit des probabilités marginales.

Probabilité sous indépendance :

$$\begin{aligned} \Pr(X = x_i \cap Y = y_j) &= \Pr(X = x_i) * \Pr(Y = y_j) \\ &= p_{ij} = p_i * p_j = \frac{n_{i+}}{n_{++}} * \frac{n_{+j}}{n_{++}} \end{aligned}$$

Cette situation correspond à des probabilités conditionnelles qui seraient égales quelque soit la strate, c'est à dire des profils (pourcentages en ligne) identiques. Dans ce cas on dit qu'il n'y a pas d'association entre la variable X et la variable Y.

L'effectif *théorique* ou effectif *attendu*, d'une cellule, sous l'hypothèse d'indépendance s'obtient en multipliant cette probabilité par l'effectif total, cf. Tableau 4.

$$\text{Effectif Théorique} = \frac{n_{i+} * n_{+j}}{n_{++}} \quad \text{formule 1}$$

FRIENDLY présente l'effectif théorique de la cellule (i,j), comme un effectif issu d'un **modèle**, ici le modèle d'indépendance, et il le note m_{ij} .

L'écart entre l'effectif observé n_{ij} et l'effectif modélisé m_{ij} est un résidu.

Résidu : $r_{ij} = (n_{ij} - m_{ij})$

On calcule le résidu standardisé de Pearson $d_{ij} = \frac{(n_{ij} - m_{ij})}{\sqrt{m_{ij}}}$ qui correspond à la contribution

de la cellule (i,j) à la Statistique du χ^2 .

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - (n_{i+} * n_{+j} / n_{++}))^2}{(n_{i+} * n_{+j} / n_{++})}$$

PEARSON a démontré que l'on peut rejeter l'hypothèse d'indépendance si $|d_{ij}| \geq 2$ avec un niveau de significativité $p < 0.05$.

Ces résidus standardisés peuvent être représentés sur des diagrammes en mosaïques, qui permettront de visualiser les écarts à l'indépendance.

Situation de dépendance

Si les deux variables sont en situation de dépendance, la probabilité dans une cellule est égale au produit de la probabilité marginale et de la probabilité conditionnelle.

$$\Pr(X \cap Y) = \Pr(X) * \Pr(Y / X)$$

ou

$$\Pr(X \cap Y) = \Pr(Y) * \Pr(X / Y)$$

Prenons un exemple à partir des données du tableau 2 repris ci-dessous.

En résumé :

Parler d'hypothèse d'indépendance en statistique est formellement la même chose que de dire : les probabilités conditionnelles, de la couleur des cheveux sont les mêmes pour toutes les modalités (strates) de la couleur des yeux et réciproquement.

Représentations Graphiques

Les Diagrammes en Barres (Bar Chart)

Le diagramme en barres ou diagramme en bâtons, est aux variables nominales ce que l'histogramme est aux variables quantitatives. Dans un diagramme en barres la **largeur** de chaque barre est **fixe**, la **hauteur** de chaque barre correspond à l'**effectif** de chaque modalité, cf. figure 1. Traditionnellement les barres sont disjointes pour indiquer l'absence de continuité entre les modalités (catégories), mais dans le module SAS/INSIGHT utilisé dans ce document, celles-ci sont accolées⁴.

Les diagrammes en barres permettent de **comparer visuellement les effectifs** de chaque modalité d'une variable, en associant la **hauteur** des barres aux effectifs.

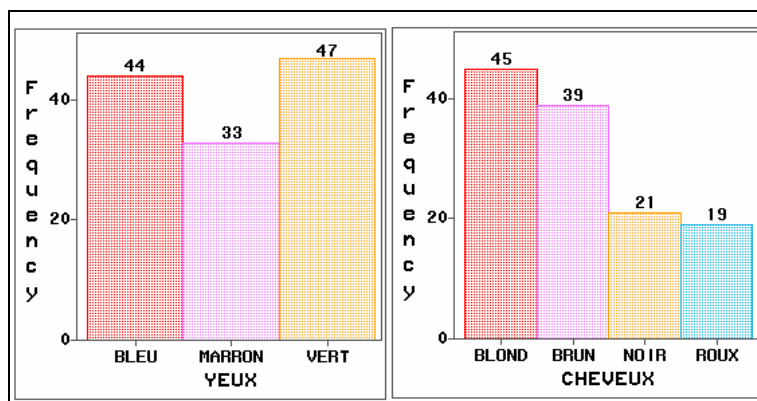


Figure 1 : Diagrammes en Barres des variables Yeux et Cheveux.

⁴ Les graphiques ont été réalisés avec le module SAS/INSIGHT du logiciel SAS®. Selon les logiciels les représentations les barres de ces diagrammes sont disjointes ou accolées. La première solution est la plus rationnelle.

A l'écran, les barres apparaissent en couleur. Les couleurs sont imposées par le logiciel SAS et elles ne peuvent être modifiées. L'ordonnancement des modalités sur l'axe est donné par leur ordre alpha-numérique. Les barres ne peuvent pas être déplacées au moyen de la souris. Toutes ces limitations font de ces diagrammes produits par SAS/INSIGHT un moyen rudimentaire, mais ils restent d'une grande utilité grâce à l'interactivité entre les graphiques.

Les Diagrammes en Bandes

Dans la figure 2, les mêmes distributions marginales, des Yeux et des Cheveux, exprimées en % sont visualisées sous forme de diagrammes en bandes, appelés **Spine Plots** dans la littérature anglo-saxonne. Dans SAS/INSIGHT ce sont des diagrammes en mosaïque à 1 dimension.

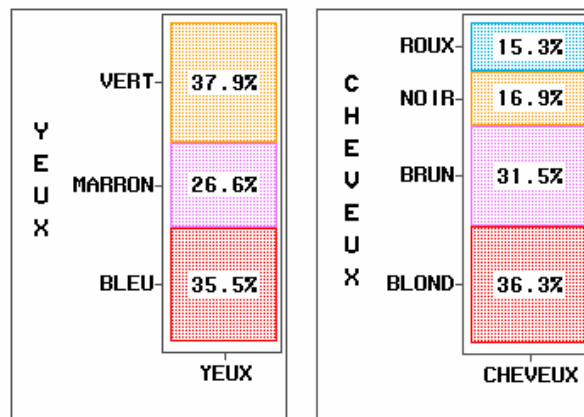


Figure 2 : Diagrammes en bandes des variables Yeux et Cheveux.

Chaque bande verticale représente 100% de l'échantillon, et chaque mosaïque est proportionnelle au pourcentage de la modalité dans l'échantillon.

Un diagramme en mosaïque à une dimension est une représentation de la distribution marginale du tableau croisé.

Par construction les hauteurs des mosaïques sont également proportionnelles aux effectifs. Dans cette forme de représentation, les comparaisons visuelles sont difficiles à faire. La discrimination entre 31.5 et 36.3 ne saute pas aux yeux, alors qu'elle serait évidente si on plaçait les 2 mosaïques côte à côte sur la même base horizontale, comme pour les diagrammes en barres.

Nous avons présentés les diagrammes en bandes car ils seront utiles pour comprendre la construction des diagrammes en mosaïque à 2 dimensions.

Les Diagrammes en Mosaïque (Mosaic Plot)

Un diagramme en mosaïque à 2 dimensions est une visualisation d'un tableau de contingence. Le graphique symbolise, les effectifs d'un tableau de contingence, par des mosaïques, dont la surface est proportionnelle aux effectifs des cellules du tableau, voir figure 3 : *Diagrammes en mosaïque du tableau croisé.*

La construction et la lecture de ce diagramme ne sont pas triviales, elles nécessitent un apprentissage.

Construction d'un diagramme en mosaïque.

Pour construire ce graphique partons, pour fixer les esprits d'un carré de taille 100*100. Sur l'axe horizontal la distribution marginale de la variable CHEVEUX est utilisée pour déterminer les largeurs des mosaïques, voir figure 3.

La **largeur** de chaque bande est donc proportionnelle à $p_{+j} = \frac{n_{+j}}{n_{++}}$

Pour chaque modalité de la variable CHEVEUX, on répartit les modalités de la variable YEUX (probabilités conditionnelles).

La **hauteur** dans chaque bande est donc proportionnelle à $p_{ij} = \frac{n_{ij}}{n_{+j}}$

La surface d'une mosaïque, produit de la largeur par la hauteur représente bien la fréquence relative observée par rapport à l'effectif total : $\frac{n_{ij}}{n_{++}}$.

$$p_{ij} = \frac{n_{+j}}{n_{++}} * \frac{n_{ij}}{n_{+j}} = \frac{n_{ij}}{n_{++}}$$

La surface d'une mosaïque est proportionnelle à l'effectif observé dans la cellule d'un tableau croisé. Cette représentation en surface montre non seulement l'effectif mais la manière dont il se compose en terme de produit.

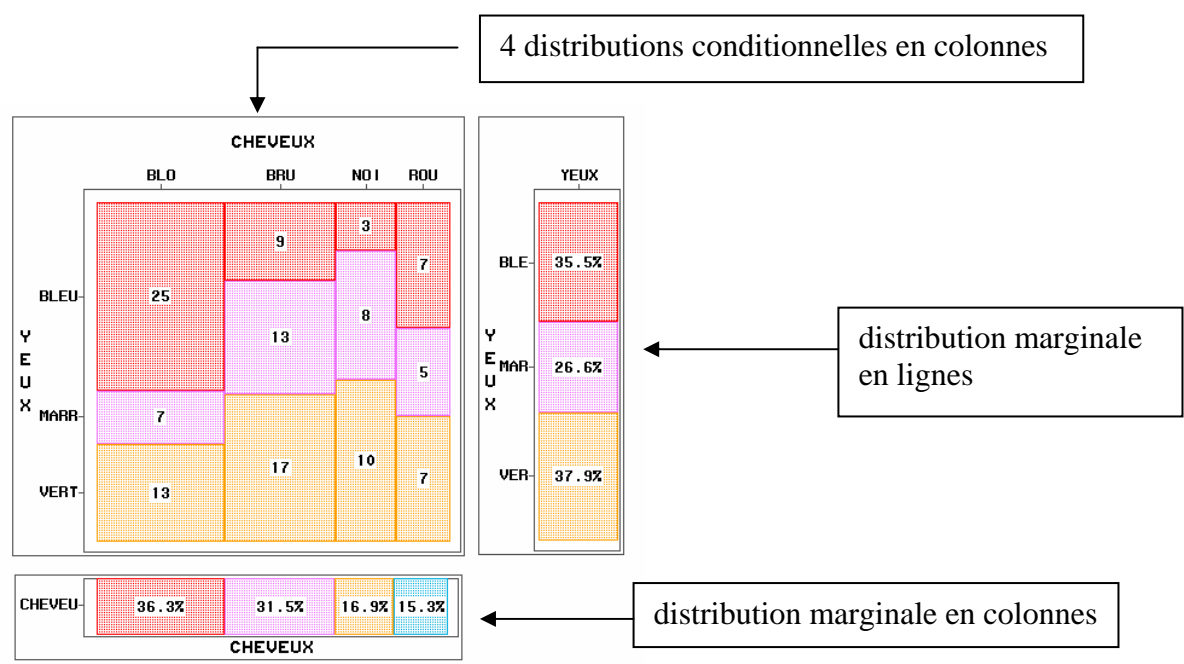


Figure 3 : Diagramme en mosaïque du tableau croisé, Avec ajout des deux distributions marginales YEUX en lignes et CHEVEUX en colonnes.

La distribution marginale en lignes sert de référence visuelle.

Les yeux bleus sont sur-représentés parmi les blonds et sous-représentés parmi les bruns. Les yeux verts sont sur-représentés parmi les cheveux noirs. Tandis que le profil des roux est identique à celui de l'échantillon total.

Toutes les comparaisons sont possibles au niveau visuel.

On s'intéresse maintenant aux effectifs qu'on devrait obtenir si les 2 variables étaient indépendantes, cf. Tableau 4 *Effectifs observés et effectifs théoriques*, obtenus à partir de la formule 1.

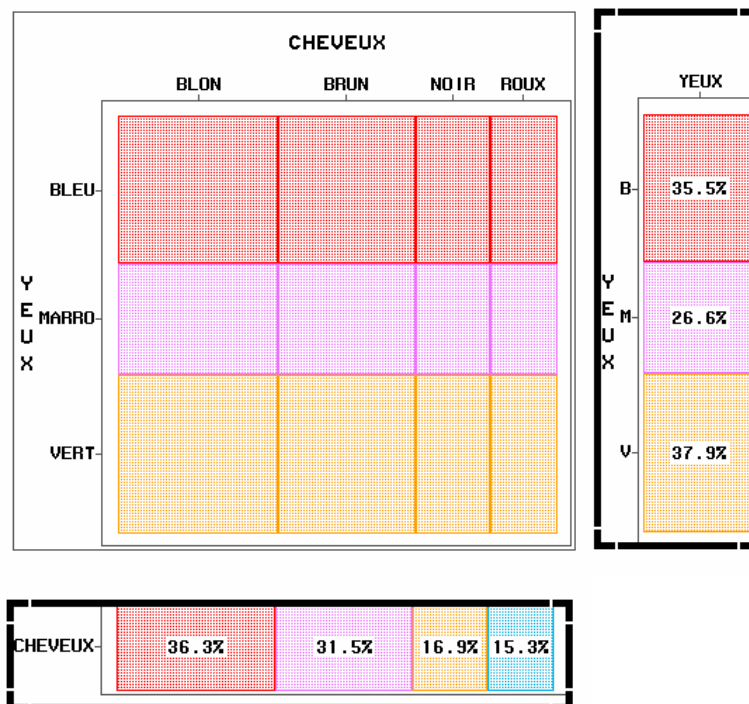


Figure 4 : Diagramme en mosaïque des effectifs théoriques calculés à partir des marges. C'est une visualisation d'une situation d'indépendance entre les variables YEUX et CHEVEUX.

La figure 4 montre les effectifs théoriques calculés sous hypothèse d'indépendance. S'il y avait indépendance entre les 2 variables YEUX et CHEVEUX les distributions conditionnelles seraient égales aux distributions marginales, et les mosaïques seraient toutes alignées. Les anglo-saxons parlent de modèle **baseline**. Ce vocabulaire imagé est bien adapté à ce diagramme.

Par comparaison, la désorganisation des mosaïques dans la figure 3 *Diagramme en mosaïque du tableau croisé* indique les écarts entre les effectifs observés et les effectifs théoriques. Ces diagrammes permettent d'avoir une vue globale et une vue locale des inadéquations des données observées, au modèle d'indépendance postulé.

Avec cette représentation visuelle les notions abstraites de dépendance/indépendance prennent un sens concret. Cette image est facilement mémorisable par l'apprenant et son accessibilité en mémoire est plus rapide. L'image sert de lien, aux sens des réseaux de neurones naturels, vers le vocabulaire et les formules.

Prolongements des Diagrammes en Mosaïque

Représentation en surface des résidus standardisés

FRIENDLY a prolongé ces travaux en représentant sur des diagrammes en mosaïque⁵, les résidus standardisés (écarts standardisés entre les effectifs observés n_{ij} et les effectifs théoriques m_{ij}).

$$d_{ij} = \frac{(n_{ij} - m_{ij})}{\sqrt{m_{ij}}}$$

d_{ij} correspond à la contribution de la cellule (i,j), à la Statistique du χ^2 .

Pour améliorer la lecture et le décodage visuel, les catégories (modalités) sont réordonnées selon les coordonnées factorielles résultant d'une analyse factorielle des correspondances. De plus les couleurs des mosaïques sont choisies selon des classes des résidus standardisés, ce qui permet une lecture très rapide.

Classes de résidus standardisés : <-4 , -4 à -2 , -2 à 0 , 0 à $+2$, $+2$ à $+4$, $>+4$.

Les sous-représentations sont en dégradés de couleur rouge et les sur-représentations sont en dégradés de couleur bleue.

La figure 5 ci-dessous, empruntée à FRIENDLY, montre la représentation des résidus standardisés sur un tableau des données analogue à notre exemple : croisement de la couleur des cheveux avec la couleur des yeux.

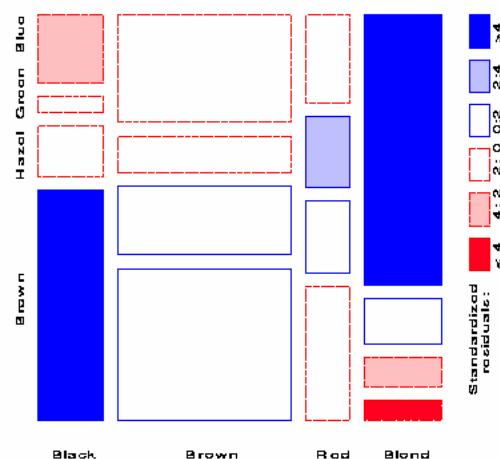


Figure 5 : Diagramme en mosaïque des résidus standardisés sur un tableau qui croise la couleur des cheveux avec la couleur des yeux.

Source FRIENDLY URL : <http://www.math.yorku.ca/SCS/Papers/drew/>

⁵ Voir l'article en ligne de FRIENDLY M. "Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data", <http://www.math.yorku.ca/SCS/Papers/drew/>

FRIENDLY a également étendu cette technique à l'analyses des tableaux à n entrées (n-ways). Il utilise ces représentations comme un outil de diagnostic pour tester différents modèles log-linéaires. Les écarts sont « portraitisés » et leur analyse permet parfois de suggérer des termes à ajouter dans le modèle pour améliorer l'ajustement.

Ces prolongements ne sont pas développés dans les logiciels standards. SAS/Insight se limitent à la représentation des tableaux croisés à double entrée. Il n'est pas possible de représenter les écarts.

Représentation en surface des écarts à l'Indépendance

En France, il existe aussi des recherches sur les représentations graphiques des tableaux croisés. Ph. CIBOIS, Professeur de Sociologie à l'Université Versailles St Quentin et membre du Laboratoire Printemps, a développé le logiciel Tri-Deux, qui permet de visualiser les écarts à l'indépendance par des représentations en surface. Son cours ainsi que son logiciel gratuit sont disponibles sur le site du Printemps⁶.

Nous empruntons à Ph. CIBOIS, la figure 6 ci dessous réalisée avec son logiciel Tri-Deux. Les données sont issues d'une enquête sur la pratique religieuse (4 modalités : Catholique pratiquant, Catholique non pratiquant, de tradition Catholique, Sans religion) et le choix politique (3 modalités : Gauche, Centre, Droite).

La référence horizontale représente la situation d'indépendance. La largeur des surfaces est comme pour les diagrammes en mosaïque, proportionnelle aux probabilités marginales du choix politique. Les hauteurs sont proportionnelles aux valeurs des écarts. Les écarts à l'indépendance sont « portraitisés » en surfaces noires pour les sur-représentations (valeurs positives), et surfaces blanches pour les sous-représentations (valeurs négatives).

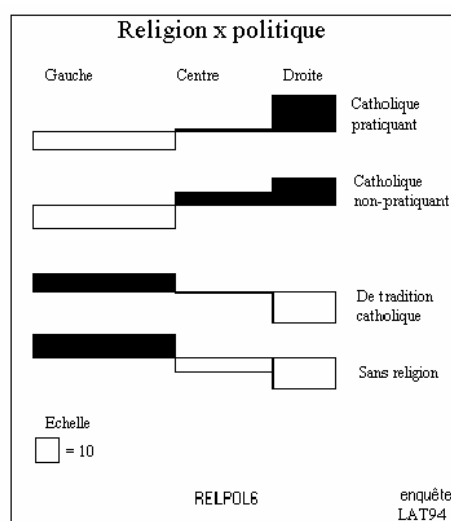


Figure 6 : Représentation en surface des écarts à l'indépendance.

Source CIBOIS URL : <http://www.printemps.uvsq.fr/cours.htm>

⁶ Laboratoire Printemps URL : <http://www.printemps.uvsq.fr/>

Conclusion

Le diagramme en mosaïque est aux données catégorielles, ce que le diagramme de dispersion (diagramme cartésien, nuage de points) est aux données quantitatives. Il permet de visualiser les dépendances entre les variables et les écarts par rapport à un modèle théorique.

Les diagrammes en mosaïque remplissent deux fonctions, d'une part ils facilitent l'apprentissage de la notion d'indépendance, d'autre part, ils permettent à l'analyste de données de mieux utiliser les modèles théoriques.

Du côté de l'enseignement nous avons montré le lien qui existe entre les tableaux croisés et les diagrammes en mosaïque. La représentation en surface des effectifs observés permet de comprendre comment interviennent et se composent les probabilités marginales et les probabilités conditionnelles pour un tableau à double entrée. Les présentations simultanées, des éléments statistiques constitutifs d'un tableau croisé : probabilités marginales et probabilités conditionnelles, leurs représentations sous forme visuelle, et les formules mathématiques associées devraient permettre un apprentissage plus opérationnel et une mémorisation facilitée de ces notions abstraites complexes. Cette approche permet de concrétiser.

Du côté de l'analyste de données visualiser les attractions et les répulsions d'un tableau croisé par les écarts standardisés ou par des écarts à l'indépendance, permet un diagnostic rapide des écarts entre les données observées et le modèle.

Nous n'avons fait qu'évoquer la généralisation de cette approche aux modèles log-linéaire (FRIENDLY). Les écarts ou résidus entre les données observées et les valeurs estimées par différents modèles d'indépendance (modèles log-linéaires), peuvent être visualisés par des graphiques en mosaïque un peu plus complexes dont la lecture est facilitée par les logiciels exploratoires interactifs (cf. MANET). Il faudra apprendre à les utiliser et à les lire.

Le travail de l'analyste de données est de rechercher des indices, à la manière d'un détective, en s'aidant de la visualisation, en testant des modèles et en étudiant les résidus qui eux portent les informations qui vont le guider. C'est la structure des mosaïques (*patterns*) qui permet de suggérer des hypothèses par le biais des comparaisons visuelles.

Les analyses et les diagnostics déduits de ces représentations visuelles sont dans l'esprit de l'analyse exploratoire des données de TUKEY (1962, 1969, 1977) : visualisation, graphiques de diagnostics à partir des résidus, émergence de nouvelles hypothèses, améliorations du modèle.

Références

CIBOIS PH. (1984), “*L’analyse des données en Sociologie*”, Paris, PUF, coll. « Le Sociologue ».

CIBOIS PH. (1994), “*L’analyse Factorielle*”, Paris, PUF, Que Sais-je? n° 2095, 4^{ème} édition.

CIBOIS PH. cours de DEUG sur les écarts à l’indépendance sur le site du Laboratoire Printemps:<http://www.printemps.uvsq.fr/cours.htm>

Logiciel Tri-Deux :

<http://www.printemps.uvsq.fr/> sélectionner logiciels puis Trideux.

DESCARTES : “The marriage of Geometry and Algebra”

<http://www.geocities.com/CapeCanaveral/Lab/4661/Descartes-1.html>

<http://scidiv.bcc.ctc.edu/Math/Descartes.html>

<http://www.ensc.sfu.ca/people/grad/brassard/personal/THESIS/node20.html>

FRIENDLY M. (1991), “*Statistical Graphics for Multivariate Data*”, SAS SUGI 16 Conference”, April.

<http://www.math.yorku.ca/SCS/sugi/sugi16-paper.html>

FRIENDLY M. (1992) “*Mosaic displays for log-linear models*”, American Statistical Association, Proceedings of the Statistical Graphics Section, 1992, pp. 61-68.

FRIENDLY M. (1992) “*Graphical Methods for Categorical Data*”, SAS SUGI 17 Conference, April.

<http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html>

FRIENDLY M. (1994), “*Mosaic Displays for Multi-way Contingency tables*”, JASA March 1994, Vol. 89, n°425, pp.190-200.

FRIENDLY M., (1995), “*Conceptual and Visual Models for Categorical Data*”, The American Statistician, May 1995, vol. 49, n°2, pp. 153-160.

FRIENDLY M., (1995), “*Graphical Methods for Categorical Data*”, Programmes écrits en SAS/GRAPH et SAS/IML.

<http://www.math.yorku.ca/SCS/Courses/grcat/grcprog.html>

FRIENDLY M. “*Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data*”.

<http://www.math.yorku.ca/SCS/Papers/drew/>

FRIENDLY M. “*Visualizing Categorical Data: Data, Stories, and Pictures*”,

<http://www.math.yorku.ca/SCS/vcd/vcdstory.pdf>

HARTIGAN, J. A., AND KLEINER, B. (1981), “*Mosaics for contingency tables*”, In W. F. Eddy (Ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. New York: Springer-Verlag, pp. 268-273.

HOFMANN H. (1997), “*Simpson on Board the Titanic ? Interactive Methods for dealing with multivariate Categorical Data*”, Statistical Computing & Statistical Graphics”, Newsletter vol. 9 n°2.

<http://cm.bell-labs.com/cm/ms/who/cocteau/newsletter/issues/v92/v92.pdf>

HOFMANN H., UNWIN A., (1999) “*Graphical Methods for Categorical Data Analysis*”, June 9, 1999.

<http://www1.math.uni-augsburg.de/~hofmann/Interface99.pdf>

HOFMANN H. (1997), “*Visualisation in Data Mining – Screening Multivariate Categorical Data*”.

<http://www.stat.fi/isi99/proceedings/arkisto/varasto/hofm0335.pdf>

NOVI M. (1998), “*Pourcentages et tableaux statistiques*” », Que Sais-je? n° 3337.

PLAYFAIR W., Graphique « *Chart Shewing at Once View, The price of the quarter of Wheat , & Wages of Labour by the week, from the year 1565 to 1821* »

<http://www.wmich.edu/ssc/about.html>

SFDS site de la Société Française de Statistique, groupe Enseignement de la Statistique

http://www.sfds.asso.fr/groupe/c_grou01.htm

SNEE, R. D. (1974), “*Graphical Display of Two way Contingency Tables*”, The American Statistician, February 1974, vol. 28, n° 1, pp9-12.

SCHWARTZ D., (1963), “*Méthodes statistiques à l’usage des médecins et des biologistes*”, Flammarion.

TUKEY J.W. , (1962) “*The Future of Data Analysis*”, Annals of Mathematical Statistics, 33, pp. 1-67.

TUKEY J.W., (1969), “*Analyzing Data : Sanctification or Detective Work*”, American Psychologist, 24, pp. 83-91., 1969F

TUKEY J.W. (1977), “*Exploratory Data Analysis*”, Addison-Wesley.

UNWIN A. (2001) “*Graphical Methods, Analytic*”, Encyclopedia of the Social and Behavioral Sciences.

<http://www1.math.uni-augsburg.de/~unwin/AntonyArts/sbs201146.pdf>

UNWIN A. (2001) “*Patterns of Data Analysis?*”, Journal of the Korean Statistical Society, 30(2), pp. 219-230.

<http://www1.math.uni-augsburg.de/~unwin/AntonyArts/Patterns2001.pdf>

UNWIN A. (2001) “*Statistification or Mystification? The need for Statistical Thought in visual Data Mining*”, ECML/PKDD meeting in Freiburg, September 2001.

<http://www1.math.uni-augsburg.de/~unwin/AntonyArts/UnwinFreiburg01.pdf>

UNWIN A. et HOFMANN H., (2001), MANET : logiciel pour l’exploration de données.

<http://www1.math.uni-augsburg.de/Manet>

VALOIS J. P. (1999) “*Une Typologie des Graphiques Statistiques*”, S.F.d.S., / XXXI^e Journées de Statistique, Grenoble, 17-21 Mai 1999.

VALOIS J. P. (2000) “*Approche Graphique en Analyse de Données*”, Journal de la Société française de Statistique, Tome 141, n°4, pp5-40.

Remerciements

Je remercie mes collègues et amis JOSIANE CONFAIS (Université Pierre et Marie Curie-ISUP) et YVETTE HOUZEL (Université Paris1-MATISSE) pour leur apport et leurs conseils quant à la réalisation de ce document.

Je remercie également ANNIE MORIN, Directrice de l’IREM de Rennes et Responsable de la Revue « Statistiquement Votre » accessible sur le site de la SFDS, qui m’a permis de publier la première version de cet article.

Article publié dans le Bulletin de Méthodologie Sociologique (référence à citer) :

LE GUEN M. (2003) « *Tableaux croisés et Diagrammes en mosaïque, Pour visualiser les probabilités marginales et conditionnelles* », Bulletin de Méthologie Sociologique, n°77, January 2003.