Outils statistiques

Notes de cours.

Clotilde Fermanian – Françoise Lucas

Année 2010 – 2011

L2-L3

Université Paris 12 –Val de Marne.

Avertissement : Ce texte constitue des notes qui couvrent ce qui a été fait en cours. Mais les Exemples n'y sont pas développés. Il faut donc s'appuyer en complément sur des notes manuscrites ou des exemples tirés de manuels ou des travaux dirigés.

Bibiographie : [1] Statistique théorique et appliquée, Pierre Dagnelie, Editions de boeck.

Chapitre 1

Collecte de données - Expérimentation

(cf. notes de cours de F. Lucas)

Chapitre 2

Statistique descriptive à une dimension

(C. Fermanian)

2.1 Introduction

La statistique descriptive a pour but de présenter les données sur une forme telle qu'on puisse en prendre connaissance et les exploiter facilement. Elle peut concerner une seule variable ou une seule caractéristique d'une variable à la fois ; on parle alors de *statistique descriptive à une dimension*. Elle peut aussi s'attacher à deux (ou plusieurs) variables, on parle alors de *statistique descriptive à deux (ou plusieurs) dimensions*.

Pour décrire ces données, on va utiliser plusieurs moyens. Des tableaux statistiques permettent de présenter les données sus formes de *distribution en fréquences*. Différents types de diagramme permettent d'obtenir des *représentations graphiques* qui donnent une appréhension visuelle rapide des données. Enfin, certaines valeurs typiques sont attachées aux donnés et donnent un 'condensé' d'information : calculer ces paramètres constitue la *réduction des données*.

2.2 Les distributions en fréquence

2.2.1 Fréquences

La forme la plus élémentaire de présentation de données statistiques consiste en l'énumération des observations

$$x_1, x_2, x_3, \cdots, x_n.$$

Cette liste peut-être ou non ordonnée. Par ailleurs, la même valeur peut apparaître plusieurs fois. On peut alors présenter les données sous la forme d'une *distribution de fréquences* : on ne fait figurer qu'une seule fois la même valeur mais on spécifie combien de fois elle apparait. On retient alors une liste de la forme

$$x_1, x_2, \cdots, x_p; n_1, n_2, \cdots, n_p.$$

Les valeurs $x_1, \cdots, x_p$ sont généralement rangées par ordre croissant et on sait que la donnée x_i apparait n_i fois. On a donc

$$p \leq n \text{ et } \sum_{i=1}^{p} n_i = n.$$

On peut aussi exprimer les fréquences en valeurs relatives par-rapport à l'effectif total. On parle alors de la *fréquence relative* n'_i

$$n'_i = \frac{n_i}{n}.$$

On a alors

$$\sum_{i=1}^{p} n'_i = 1.$$

On peut exprimer les fréquences relatives en pourcentage

$$n'_i\% = 100 \cdot \frac{n_i}{n}.$$

On utilise aussi la notion de *fréquences cumulées*. La fréquence absolue cumulée $N'(x_k)$ associée à la donnée x_k est le nombre d'observation correspondant à une donnée inférieure ou égale à x_k :

$$N'(x_k) = \sum_{i=1}^{k} n_i = n_1 + \cdots + n_k.$$

La *fréquence relative cumulée* est son expression en valeur relative

$$\frac{N'(x_k)}{n} = n'_1 + \cdots + n'_k.$$

Exemple : Distribution de fréquences du nombre de pieds d'asphodèles observées dans 512 carrés de 1 m^2 (tiré de la référence [1]).

2.2.2 Les distributions groupées

Quand le nombre de valeurs observées est élevé, on condense les tableaux statistiques en groupant les observations en *classes*. On obtient ainsi des distributions de fréquences groupées en classes ou *distributions groupées*. Chacune des classes est caractérisée par les *valeurs extrêmes* qu'elle peut contenir. L'écart entre les limites des classes est appelé *amplitude* ou *intervalle de classe*. La *fréquence d'une classe* est le nombre d'observations qui y sont contenues.

Exemple : Distribution de fréquences du poids des feuilles de 1 000 plantes de chicorée witloof (exemple tiré de la référence [1]).

2.3 Les représentations graphiques

2.3.1 Diagrammes de fréquence non cumulées

Les *diagrammes en bâtons* sont éablis en traçant parallèlement à l'axe des ordonnées, en face de chaque valeur observée x_i, un segment de longueur égale à la fréquence de cette valeur. Ce type de graphique est particulièrement adapté au cas des distributions non groupées.
Les *polygônes de fréquence* sont construits en joignant par une ligne brisée les extrémités des segments voisinss des diagrammes en bâtons.
Les *histogrammes* se composent de rectangles dont les intervalles de classe sont les bases et les fréquences les hauteurs. Ce type de graphique est adapté au cas des distributions groupées.
Pour chaque type de représentation graphique, les échelles des abcisses et des ordonnées sont choisies de manière à mettre en valeur les caractéristiques essentielles des distributions.

Exemples : 1- Diagramme en bâtons et polygone de fréquence donnant le nombre de pieds d'asphodèles observés dans 512 carrés de $1\,m^2$.
2- Histogramme donnant le poids des feuilles de 1 000 plantes de chicorée witloof.

2.3.2 Diagrammes de fréquence cumulées

Les distributions de fréquence cumulées peuvent être représentées graphiquement par des polygones de fréquences ou des histogrammes. Au dessus du point x_i de l'axe des abcisses se trouve un point dont

l'ordonnée indique en valeur absolue ou relative, la fréquence des observations inférieures ou égales à l'abcisse considérée. Les *polygones de fréquence cumulées* sont construits différemment selon le type de distribution.

Pour les distributions non-groupées, le polygone est construit en escalier : on dessine des segments de droites verticaux de longueur proportionnelle aux fréquences mais en les décalant progressivement vers le haut de telle sorte que l'origine de chacun d'eux soit située à hauteur de l'extrémité du précédent. On joint ensuite ces différents segments verticaux par des segments horizontaux.

Pour les distributions groupées, on joint par une ligne brisée les points obtenus en portant en face des limites supérieures des classes, des ordonnées égales aux fréquences cumulées, absolues ou relative. Dans le cas des fréquences relatives, la fonction obtenue est appelée *fonction cumulative de fréquences* ou *fonction de distribution*. Elle est croissante et prend la valeur 1 en x_p.

Exemples : Polygone de fréquences cumulées pour les deux exemples précédant.

Remarque : On rencontrera fréquemment des distributions en cloche ou des distributions avec deux ou plusieurs cloches. Les valeurs ont tendance à se regrouper autour de l'une d'entre elles (distribution à une cloche) ou autour de deux ou plusieurs valeurs (distribution à deux ou plusieurs cloches).

2.3.3 Autres types de représentation graphique

(Non abordé cette année, faute de temps)

Les *boxplots* : L'ensemble des observations, classées par ordre croissant, est subdivisé en quatre groupes de même effectif ou d'effectifs quasi égaux. Deux rectangles contigus (les 'boîtes') sont affectés aux deux groupes intermédiaires et deux lignes (les 'moustaches') sont affectées, de part et d'autre de ces rectangles, aux deux groupes extrêmes.

Les *diagrammes circulaires* ou *camemberts* permettent de représenter les distributions en fréquence dans des cercles : les aires des différents secteurs sont proportionnelles aux fréquences. Ce type de diagramme est adapté aux donnés qualitatives.

L'utilisation d'*échelles non-linéaires* est adapté dans certains cas, échelles logarithmiques par exemple.

2.4 La réduction des données

Le calcul de certains paramètres permet de caractériser de façon simple les séries statistiques observées. Les *paramètres de position* servent à caractériser l'ordre de grandeur des observations. Les *paramètres de dispersion* permettent de chiffrer la variabilité des valeurs observées autour d'un des paramètres de position.

2.4.1 Les paramètres de position

1- La *moyenne arithmétique* que l'on appelle généralement *moyenne* est la somme des valeurs observés divisée par le nombre d'observations :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Comme chaque valeur x_i doit être prise en considération autant de fois qu'elle a été observée, cette expression devient pour les distributions en fréquence

$$\overline{x} = \sum_{i=1}^{p}(n_i x_i).$$

Dans le cas des distributions non groupées, les deux expressions sont rigoureusement équivalentes. Par contre, pour les distributions groupés, on commet en général une certaine erreur, en remplaçant chacune des valeurs réellement observées par le point central de la classe correspondante.

Propriétés :
- Si $y_i = a + bx_i$, alors $\overline{y} = a + b\overline{x}$.
- Si $y_i = x_i - \overline{x}$ alors $\overline{y} = 0$.

2- La *médiane* $\tilde{x}$ est un paramètre de position tel que la moitié des observations lui sont inférieures (ou égales) et la moitié supérieures (ou égales).
Pour les séries statistiques et les distributions non groupées, quand le nombre d'observations est impair, la médiane est l'observation de rang $\frac{n+1}{2}$

$$\tilde{x} = x_{\frac{n+1}{2}} \text{ si } n \text{ est impair.}$$

Quand n est pair, tout nombre compris entre $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$ répond à la définition. On prend comme valeur de la médiane la moyenne entre ces deux observations

$$\tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \text{ si } n \text{ est pair.}$$

Dans le cas des distributions non groupées, la médiane peut être déterminée graphiquement en utilisant les diagrammes de fréquences cumulées :

$$N'(\tilde{x}) = \frac{1}{2}.$$

3- De façon analogue, on définit les *quartiles* q_1, q_2 et q_3 d'une distribution de fréquence par

$$N'(q_1) = \frac{1}{4}, \quad N'(q_2) = \frac{1}{2}, \quad N'(q_3) = \frac{3}{4}.$$

Les trois quartiles divisent l'ensemble des observations en quatre sous-ensembles de même effectif, le deuxième quartile étant confondu avec la médiane. Les quartiles se calculent de la même manière que la médiane. Des problèmes peuvent se poser quand l'effectif n'est pas un nombre pair.

4- On appelle *mode* ou *valeur dominante* d'une distribution non groupée la ou les valeurs observées de fréquence maximum. On appelle *classe(s) modale(s)* d'une distribution groupée la ou les classe(s) de fréquence maximum si l'intervalle de classe n'est pas constant. On dit qu'une distribution est *unimodale* si elle ne possède qu'un maximum de fréquence, *plurimodale* s'il y en a plusieurs.

2.4.2 Les paramètres de dispersion

La *variance* s^2 d'une série statistique ou d'une distribution de fréquence est la moyenne arithmétique des carrés des écarts par rapport à la moyenne

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ ou } \frac{1}{n}\sum_{i=1}^{p}\left(n_i(x_i - \overline{x})^2\right).$$

Les deux définitions sont équivalentes dans le cas des distributions non groupées. Par contre, comme pour la moyenne, on commet une certaine erreur dans le cas des distributions groupées.

L'*écart-type* s est la racine carrée de la variance et le *coefficient de variation* cv est obtenu en exprimant l'écart type en valeur relative ou en pourcentage de la moyenne (quand celle-ci est positive) :

$$cv = \frac{s}{\overline{x}} \text{ ou } 100 \cdot \frac{s}{\overline{x}}.$$

Propriétés :

- La variance, l'écart-type et le coefficient de variation sont nuls si et seulement si tous les écarts $x_i - \overline{x}$ sont égaux à 0. Toutes les valeurs sont alors égales entre elles.
- La variance et l'écart type sont invariants par changement d'origine : si $y_i = a + bx_i$,

$$s_y = |b| s_x, \quad cv_y = cv_x.$$

En effet, on a alors $\overline{y} = a + b\overline{x}$ et

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n} \left((a + bx_i) - (a + b\overline{x})\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(b(x_i - \overline{x})\right)^2 \\ &= \frac{b^2}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= b^2 s_x^2 \end{aligned}$$

L'*écart moyen absolu* ou *écart moyen* est la moyenne des valeurs absolues des écarts par rapport à la moyenne

$$e_m = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}| \text{ ou } \frac{1}{n}\sum_{i=1}^{p} \left(n_i |x_i - \overline{x}|\right).$$

On appelle *amplitude* l'écart entre les valeurs extrêmes d'une série d'observations classées par ordre croissant :

$$w = x_n - x_1.$$

Ce paramètre n'est pas défini exactement pour les distributions groupées, les valeurs extrêmes n'étant plus connues avec exactitude après le groupement en classe. On peut montrer que

$$s \leq \frac{w}{2}.$$

La détermination de l'amplitude peut donc permettre de vérifier l'ordre de grandeur de la variance.

L'*écart interquantile* est la différence $q_3 - q_1$. Cet intervalle englobe la moitié ou approximativement la moitié des observations qui se situent au centre de la distribution.

2.5 Exécution des calculs, différents types d'erreur

Les *erreurs d'approximation* ou *d'arrondi* sont liées au caractère approché ou arrondi de la majorité des nombres impliqués dans les calculs. Le but est de conserver à tout moment le nombre de chiffres le plus adéquat pour assurer une précision suffisante des résultats sans compliquer outre mesure le travail. Il y a un équilibre à assurer entre une perte d'information liée à un arrondi excessif au cours de résultats intermédiaires et une complexification dangereuse des calculs impliquée par la conservation de trop de décimales.

Il est donc important de différencier *valeurs exactes* et *valeurs approchées* : les fréquences observées et la plupart des constantes intervenant dans les calculs sont des valeurs connues de manière exacte tandis que les résultats de mesure et les nombres arrondis ne sont en général que des valeurs approchées.

La précision des valeurs approchées peut être caractérisée soit par leur nombre de *décimales exactes*, soit par leur nombre de *chiffres significatifs*.

Les chiffres qui, dans une valeur approchée, servent uniquememnt à indiquer l'ordre de grandeur du nombre envisagé sont dits *non significatifs*. Les autres chiffres sont considérés comme *significatifs*.

Exemple : Les chiffres non significatifs sont soulignés :

$$5,802 - 2,307 - \underline{0},70 - \underline{0,002}1.$$

On remarquera que les valeurs approchées $0,7$, $0,70$ et $0,700$ ne représentent pas exactement la même chose. Ces nombres représentent des valeurs comprises respectivement entre $0,65$ et $0,75$, $0,695$ et $0,705$, $0,6995$ et $0,7005$.

Quelques règles simples :

Pour *les sommes et les différences*, le dernier chiffre significatif du résultat est celui qui correspond vers la droite au dernier chiffre significatif du terme qui possède (vers la droite également) le moins de chiffres significatifs :

$$103,2 + 8,753 - 92,39 = 19,563$$

Le résultat correctement arrondi est $19,6$.
Pour les *produits et les quotients*, le résultat possède autant de chiffres significatifs que le facteur qui en possède le moins :

$$2,1 \times 0,0129 \times 11,2 = 0,303408$$

le résultat correctement arrondi est $0,30$ puisqu'un des trois facteurs du produit ne possède que deux chiffres significatifs.
Enfin, notons qu'il est toujours opportun de vérifier l'ordre de grandeur des résultats obtenus.

Chapitre 3

Statistique descriptive à deux dimensions

(C. Fermanian)

3.1 Introduction

La statistique descriptive à deux dimensions a pour objet de mettre en évidence les relations qui existent entre deux séries d'observations considérées simultanément.

3.2 Distribution de fréquence à deux dimensions

Les observations relatives à deux variables se présentent sous la forme d'une *série statistique double* c'est-à-dire de la suite de n couples de valeurs observées (x_i, y_i) rangées dans l'ordre croissant de l'une des deux variables

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{array}$$

Comme dans le cas unidimensionnel, on condense les données en *distribution de fréquence*. On note

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_p \\ y_1 & y_2 & \cdots & y_q \end{array}$$

les valeurs distinctes. On construit un *tableau à double entrée* dont les p lignes donnent les valeurs de x, les q colonnes, celles de y et l'on met dans la cellule correspondant au couple (x_i, y_j) le nombre $n_{i,j}$ correspondant au nombre d'observations de (x_i, y_j). L'ensemble

des valeurs x_i et y_j d'une part et des fréquences $n_{i,j}$ constitue une *distribution de fréquences à deux dimensions.*

On peut aussi grouper les observations en une *distribution groupée* en réunissant en classe les valeurs observées. Les symboles x_i et y_j représentent alors les points centraux des classes et l'on désigne par Δx et Δy les intervalles de classe pour x et y respectivement.

Exemple : Charge en matière en suspension et en carbone organique total dans les eaux usées arrivant à une centrale d'épuration (données communiquées par F. Lucas).

On peut également calculer des *fréquences relatives*

$$n'_{ij} = \frac{n_{ij}}{n}.$$

Dans le cas des distributions de fréquence à deux variables, on introduit une nouvelle notion : les *distributions marginales* et les *distributions conditionnelles.*

3.2.1 Distributions marginales

On obtient les fréquences marginales $n_{i\cdot}$ et $n_{\cdot j}$ en calculant les totaux relatifs aux différentes lignes ou colonnes

$$n_{i\cdot} = \sum_{j=1}^{q} n_{ij} \text{ et } n_{\cdot j} = \sum_{i=1}^{p} n_{ij}.$$

Ces fréquences sont reliées par les relations

$$\sum_{i=1}^{p} n_{i\cdot} = \sum_{j=1}^{q} n_{\cdot j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j} = n.$$

Les *fréquences marginales relatives* correspondantes sont

$$n'_{i\cdot} = \frac{n_{i\cdot}}{n} \text{ et } n'_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

Ces fréquences sont telles que

$$n'_{i\cdot} = \sum_{j=1}^{q} n'_{ij}, \quad n'_{\cdot j} = \sum_{i=1}^{q} n'_{ij}, \quad \sum_{i=1}^{p} n'_{i\cdot} = \sum_{j=1}^{q} n'_{\cdot j} = 1.$$

3.2.2 Distributions conditionnelles

Non traité cette année

En considérant une ligne particulière du tableau à double entrée, on définit par l'ensemble des valeurs $y_1, \cdots, y_q$ et les fréquences $n_{i1}, \cdots, n_{iq}$ une distribution à une dimension appelée *distribution conditionnelle* de y sous la condition $x = x_i$.
Les fréquences relatives associées sont appelées *fréquences conditionnelles*. On appelle fréquence de y sous la condition $x = x_i$

$$n'_{j|i} = \frac{n_{ij}}{n_i} = \frac{n'_{ij}}{n'_i}.$$

De même, en considérant la j-ième colonne, on définit la fréquence de x sous la condition $y = y_j$

$$n'_{i|j} = \frac{n_{ij}}{n_j} = \frac{n'_{ij}}{n'_j}.$$

On vérifie que

$$\sum_{j=1}^{q} n'_{j|i} = 1 \text{ et } \sum_{i=1}^{p} n'_{i|j} = 1.$$

3.3 Représentation graphique

3.3.1 Diagramme de dispersion ou nuage de points

On représente la série à deux variables sous forme de diagramme de dispersion ou nuage de points en faisant figurer les n points de coordonnée $(x_1, y_1), \cdots, (x_n, y_n)$. On peut aussi faire figurer des boxplots sur ces diagrammes.

Exemple : Diagramme correspondant à l'exemple précédent.

3.3.2 Représentation des distributions de fréquences à deux dimensions

On utilise des figures en trois dimensions.
Les *diagrammes en bâtons* sont établis en traçant perpendiculairement au plan (x, y), en chaque point (x_i, y_j) un segment de longueur égale à n_{ij} ou n'_{ij}.

Les *stéréogrammes* sont composés de parallélépipèdes rectangles juxtaposés dont les bases correspondent à chacune des cellules du tableau à double entrée et dont les hauteurs sont égales aux fréquences absolues ou relatives.

Figure : (schématique...)

3.4 Réduction des données

Les paramètres utilisés pour caractériser les séries statistiques doubles sont de deux types.

- Les uns ne concernent qu'une variable à la fois, ils servent à caractériser les distributions marginales ou conditionnelles.
- Les autres servent à décrire les relations existant entre les deux séries d'observation.

Pour caractériser les distributions marginales ou conditionnelles, on utilise les paramètres des distributions à une variable.
On définit les *moyennes marginales*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ ou } \frac{1}{n}\sum_{i=1}^{p}(n_{i\cdot}x_i),$$
$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \text{ ou } \frac{1}{n}\sum_{j=1}^{q}(n_{\cdot j}y_j).$$

les *variances marginales*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \text{ ou } \frac{1}{n}\sum_{i=1}^{p}\left[n_{i\cdot}(x_i-\overline{x})^2\right],$$
$$s_y^2 = \frac{1}{n}\sum_{j=1}^{n}(y_i-\overline{y})^2 \text{ ou } \frac{1}{n}\sum_{j=1}^{q}\left[n_{\cdot j}(y_j-\overline{y})^2\right],$$

les *moyennes conditionnelles*

$$\overline{x}_j = \frac{1}{n_{\cdot j}}\sum_{i=1}^{p}(n_{ij}x_i) \text{ et } \overline{y}_i = \frac{1}{n_{i\cdot}}\sum_{j=1}^{q}(n_{ij}y_j),$$

et les *variances conditionnelles*

$$s_{x|j}^2 = \frac{1}{n_{\cdot j}}\sum_{i=1}^{p}\left[n_{ij}(x_i-\overline{x}_j)^2\right] \text{ et } s_{y|i}^2 = \frac{1}{n_{i\cdot}}\sum_{j=1}^{q}\left[n_{ij}(y_j-\overline{y}_i)^2\right].$$

L'étude simultanée des deux séries d'observation se fait grâce aux outils détaillés dans la fin de ce paragraphe : la covariance et le coefficient de corrélation.

3.4.1 Covariance

La *covariance* des deux séries d'observation x et y est définie par

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\overline{x})(y_i-\overline{y})\right] \quad \text{ou} \quad \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}\left[n_{ij}(x_i-\overline{x})(y_j-\overline{y})\right].$$

La covariance est positive lorsqu'à des valeurs élevées des x_i correspondent des valeurs élevées des y_i. Réciproquement, la covariance est négative lorsqu'à des valeurs élevées des x_i correspondent des valeurs faibles des y_i. Elle est donc positive ou négative selon que le nuage de points a une allure croissante ou décroissante.

Propriétés :
- Si $x' = a + bx$ et $y' = c + dy$ alors

$$\mathrm{cov}(x',y') = bd\,\mathrm{cov}(x,y).$$

- La covariance est inférieure ou égale en valeur absolue au produit des écarts-types :

$$|\mathrm{cov}(x,y)| \leq s_x\, s_y.$$

Preuve : On regarde la quantité

$$P(b) = \frac{1}{n}\sum_{i-=1}^{n}\left[b(x_i-\overline{x})-(y_i-\overline{y})\right]^2.$$

Cette quantité est un polynôme du second degré en b

$$P(b) = b^2 s_x^2 - 2b\mathrm{cov}(x,y) + s_y^2.$$

Ce polynôme a un signe constant, il a donc un discriminant négatif, d'où

$$4\,\mathrm{cov}(x,y)^2 - 4s_x^2 s_y^2 \leq 0.$$

- Si $\mathrm{cov}(x,y) = s_x\, s_y$, alors tous les points observés se trouvent sur une même droite

$$y-\overline{y} = b_{yx}(x-\overline{x}) \text{ avec } b_{yx} = \frac{\mathrm{cov}(x,y)}{s_x}.$$

Preuve : Si $\mathrm{cov}(x, y) = s_x\, s_y$, alors le discriminant du polynôme $P(b)$ est nul. Ce polynôme a alors une unique racine

$$b_{yx} = \frac{\mathrm{cov}(x, y)}{s_x}$$

et le fait que $P(b_{yx}) = 0$ implique que pour tout i,

$$b_{yx}(x_i - \overline{x}) - (y_i - \overline{y}) = 0,$$

ce qui signife que tous les points (x_i, y_i) sont sur la droite $y - \overline{y} = b_{yx}(x - \overline{x})$.

3.4.2 Coefficient de corrélation

Le *coefficient de corrélation* est défini par

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y}.$$

Ce coefficient est toujours compris entre -1 et 1 et a le même signe que la covariance. Il ne peut être égal à ± 1 que si les points sont situés sur une même droite non parallèle aux axes. Il s'interprète comme suit

- $r = 1$ quand toutes les points se trouvent sur une même droite croissante,
- $r \sim 1$ quand toutes les points se trouvent à proximité d'une même droite croissante,
- $0 < r < 1$ quand le nuage de points est allongé parallèlement à une droite croissante,
- $r = 0$ ou $r \sim 0$ quand le nuage de points est allongé prallèlement à l'un des axes de coordonnées ou de forme arrondi,
- $-1 < r < 0$ quand le nuage de points est allongé parallèlement à une droite décroissante,
- $r \sim -1$ quand toutes les points se trouvent à proximité d'une même droite décroissante,
- $r = -1$ quand toutes les points se trouvent sur une même droite décroissante.

Figures : Schéma correspondant à chacune de ces situations.

Propriété : Si $x' = a + bx$ et $y' = c + dy$ alors $r = r'$.

Pour conclure, remarquons qu'il ne faut pas perdre de vue que l'existence d'une corrélation entre deux séries d'observation n'implique

pas nécessairement une relation de cause à effet. La corrélation peut être due au fait que les deux variables sont soumises à des influences communes.

3.4.3 Régression linéaire au sens des moindre carrés

Quand le nuage de points a une forme générale linéaire, on peut tenter de préciser la relation qui lie les variables x et y par la recherche d'une droite qui s'ajuste au mieux aux valeurs observées. La *méthode des moindre carrés* permet de trouver une telle droite qui minimise la somme des carrés des écarts entre les points observés et les points correspondants de la droite.

Si l'équation de la droite est

$$y = ax + b$$

la somme des carrés des écarts à minimiser est

$$\Sigma = \sum_{i=1}^{n} \left(y_i - y(x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2 .$$

On cherche a et b tel que cette quantité soit minimale, il faut donc annuler les dérivées partielles par-rapport à a et b

$$\partial_a \Sigma = \partial_b \Sigma = 0.$$

On trouve

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \ \text{ et } \ \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0,$$

soit

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \ \text{ et } \ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i y_i).$$

La première équation donne

$$\overline{y} = a + b\overline{x}$$

ce qui implique que la droite de régression passe par le point moyen $(\overline{x}, \overline{y})$.
En multipliant la première équation par $\overline{x}$ et en la soustrayant à la seconde, on obtient

$$b = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\mathrm{cov}(x,y)}{s_x^2}.$$

En effet, en développant les produits $(x_i - \overline{x})(y_i - \overline{y})$, on démontre

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right].$$

La droite de régression de y en x a donc pour équation

$$y = \frac{\mathrm{cov}(x,y)}{s_x^2}(x - \overline{x}) + \overline{y}.$$

On appelle *coefficient de régression* de y en x la quantité

$$b_{yx} = \frac{\mathrm{cov}(x,y)}{s_x^2}.$$

On peut aussi calculer la droite de régression de x en y

$$x = b_{xy}(y - \overline{y}) + \overline{x}$$

où

$$b_{xy} = \frac{\mathrm{cov}(x,y)}{s_y^2}.$$

Ces deux droites se coupent au point moyen $(\overline{x}, \overline{y})$ et forment entre elles un angle d'autant plus petit que la valeur absolue du coefficient de corrélation est proche de 1.

Chapitre 4

Probabilités mathématiques et distributions théoriques

4.1 Notion de probabilité

La notion de probabilité est liée aux notions d'expérience et d'événement aléatoires. Une *expérience* est dite *aléatoire* quand on ne peut pas en prévoir exactement le résultat parce que tous les facteurs dont dépendent ce résultat ne sont pas controlés. Un *événement aléatoire* est un événement qui peut éventuellement se réaliser au cours d'une expérience aléatoire.

Quand une expérience alátoire a été répétée un certain nombre de fois n, on peut déterminer le nombre de réalisations de l'événement A qui y est associé. On connait alors sa fréquence abolue n_A et on peut calculer sa fréquence relative

$$n'_A = \frac{n_A}{n}.$$

Si l'expérience est réalisée un grand nombre de fois dans des conditions uniformes, on constate que la fréquence relative a tendance à se stabiliser. On peut alors postuler, pour tout événement aléatoire qui remplit ces conditions, l'existence d'un nombre fixe dont la fréquence relative a tendance à s'approcher. Ce nombre est par définition la *probabilité mathématique* de l'événement considéré. La probabilité ainsi définie est une forme idéalisée de la fréquence relative.

4.2 Propriétés mathématiques de la probabilité

La notion de probabilité n'est pas définie de façon suffisante par son seul postulat d'existence. Aussi doit-on lui attribuer un certain

nombre de propriétés sous forme d'axiomes. Ceux-ci peuvent être compris par analogie avec certaines propriétés de la fréquence relative.

1- La probabilité de tout événement aléatoire A est comprise entre 0 et 1 :

$$0 \leq P(A) \leq 1.$$

2- Si deux événements A et B associés à une même expérience aléatoire ne peuvent pas se produire simultanément, alors

$$P(A \text{ ou } B) = P(A) + P(B).$$

De tels événements sont dits *exclusifs*.
Si $A_1, \cdots, A_m$ sont m événements exclusifs

$$P(A_1 \text{ ou} \cdots \text{ou } A_m) = P(A_1) + \cdots + P(A_m).$$

Ces propriétés impliquent que dans le cas de deux événements A et B non nécessairement exclusifs

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ et} B).$$

En effet

$$P(A \text{ ou } B) = P(A \text{ sans } B) + P(B \text{ sans } A) + P(A \text{ et } B)$$

avec

$$P(A) = P(A \text{ sans } B) + P(A \text{ et } B) \text{ et } P(B) = P(B \text{ sans } A) + P(A \text{ et } B).$$

4.3 Probabilité conditionnelle et indépendance stochastique

Non traité cette année

Par analogie avec les propriétés des fréquences conditionnelles, on définit la *probabilité conditionnelle* de l'événement A sous la condition B par

$$P(A|B) = \frac{P(A \text{ et } B)}{P(B)}.$$

On a donc la propriété

$$P(A) = P(A|B)P(B).$$

On dira que deux événements sont *stochastiquement indépendants* si

$$P(A|B) = P(A|\text{non } B) = P(A)$$

ou non B désigne la non-réalisation de B. Lorsque cette condition n'est pas réalisée, on dit que ces événements sont dépendants.

4.4 Notion de variable aléatoire et distributions discontinues

4.4.1 Définitions

Une variable aléatoire X est une variable associée à une expérience aléatoire et servant à caractériser le résultat de cette expérience. Elle est dite discontinue ou discrète si elle varie de façon discontinue. A chacune des valeurs x que peut prendre la variable X, on associe une probabilité $P(x)$

$$P(x) = P(X = x).$$

Nous considérerons des variables aléatoires prenant des valeurs entières positives. L'ensemble des valeurs admissibles x et des probabilités correspondantes $P(x)$ constitue une *distribution de probabilité* ou *distribution théorique discontinue*. La relation existant entre x et $P(x)$ est appelée loi de probabilité. La distribution cumulée des probabilités donne naissance à la *fonction de distribution*

$$F(x) = P(X \leq x).$$

Les distributions théoriques discontinues et leurs fonctions de répartition ont des propriétés analogues à celles des distributions non groupées exprimées en fréquences relatives et de leurs fonctions de distribution :

$$\sum_{x=0}^{\infty} P(x) = 1,$$

$$0 \leq F(x) \leq 1, \quad F(x) = 0 \ \text{ pour } \ x < 0 \ \text{ et } \ F(\infty) = 1.$$

4.4.2 Paramètres d'une variable aléatoire

On appelle *espérance mathématique* ou *valeur moyenne* d'une variable aléatoire la quantité

$$E(X) = \sum_{x=0}^{+\infty} xP(X = x).$$

Cette valeur correspond à la *valeur attendue* ou valeur la plus probable de la variable aléatoire.

Propriétés : **1-** $E(aX + b) = aE(X) + b$.
2- $E(X + Y) = E(X) + E(Y)$.

On appelle *valeur médiane* de la variable aléatoire X le nombre $\tilde{m}$ tel que

$$F(\tilde{m}) = \frac{1}{2}.$$

Cette définition est ambiguë car la fonction F peut être discontinue et $\frac{1}{2}$ peut ne pas être une valeur prise.

On appelle *variance* de la variable aléatoire X la quantité

$$\sigma^2 = \sum_{x=0}^{\infty} \left[(x - m)^2 P(x)\right]$$

où $m = E(X)$. Le nombre σ est l'*écart-type* et le nombre $CV = \frac{\sigma}{m}$ le *coefficient de variation*.

Propriété : $\sigma(aX + b) = |a|\sigma(X)$.

On remarquera que toutes ces définitions sont calquées sur les formules relatives aux séries statistiques. On les obtient en remplaçant les fréquences relatives par les probabilités.

4.4.3 Exemples

Distributions binomiales : On considère un ensemble de n expériences aléatoires identiques et stochastiquement équivalentes, à chacune desquelles sont associés deux événement exclusifs A et B. Par expériences identiques, on veut dire que les probabilités de A et B ne varient pas d'une expérience à l'autre et sont telles que

$$P(A) = p \text{ et } P(B) = q = 1 - p.$$

Ce schéma d'expérience appelé *Schéma de Bernouilli* est réalisé par exemple par le jet de n pièces de monnaie identiques. L'événement A désigne le fait que la pièce tombe sur pile et l'événement B que la pièce tombe sur face. Cela concerne aussi le prélèvement dans une population de n personnes possédant chacun l'un ou l'autre de deux caractères opposés.

On considère la variable aléatoire X correspondant au nombre de réalisations de l'événement A au cours des n expériences. La variable X prend ses valeurs entre 0 et n. La probabilité d'avoir x réalisations de A et $n-x$ réalisations de B est

$$p^x q^{n-x}$$

Par aillleurs il y a C_n^x façons d'avoir x réalisations de A, C_n^x est le coefficient binômial

$$C_n^x = \frac{x!(n-x)!}{n!}.$$

On a donc

$$P(X = x) = \mathrm{C}_n^x p^x (1-p)^{n-x}.$$

Cette loi s'appelle la *loi binomiale* et on dit que X est une variable binomiale par référence à la formule du binôme qui donne le dévelopement de $(p+q)^n$. Comme $p+q=1$, on a bien

$$\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{n} P(x) = \sum_{x=0}^{n} C_n^x p^x q^{n-x} = 1.$$

Les paramètres d'une variable aléatoire binomiale sont

$$m = np =, \quad \sigma^2 = npq.$$

Les preuves de ces formules seront vues en exercice en TD.
Il faut remarquer qu'une variable aléatoire binomiale est complètement déterminée par sa moyenne et sa variance puisque

$$p = 1 - \frac{\sigma^2}{m} \text{ et } n = \frac{m^2}{m - \sigma^2}.$$

Distributions de Poisson. Une variable aléatoire X suit une *distribution de Poisson* si on a

$$P(X = x) = \mathrm{e}^{-m} \frac{m^x}{x!}.$$

Ces distributions sont caractérisées par un seul paramètre m. On peut voir cette distribution comme un cas limite de distribution binomiale lorsque $p \to 0$ et $n \to \infty$ en conservant $np = m$. On peut alors monter que

$$C_n^x p^x q^{n-x} \to \mathrm{e}^{-m} \frac{m^x}{x!}.$$

Ce résultat est le théorème de Poisson.
Les paramètres des distributions de Poisson sont

$$E(X) = \sigma^2 = m.$$

4.5 Variables aléatoires et distributions continues

4.5.1 Définitions et paramètres

Une variable aléatoire pouvant valoir n'importe quel nombre réel est dite continue. On s'intéresse alors à la probabilité d'observer une valeur dans un certain intervalle près d'une valeur x :

$$P(x < X < x + \delta x).$$

Cette probabilité tend en général vers 0 lorsque Δx tend vers 0 : la probabilité d'obtenir exactement une valeur donnée est généralement nulle même si cet événement n'est pas impossible. La notion de distribution n'a donc pas de sens pour des valeurs aléatoires continues. En revanche, la notion de fonction de répartition reste pertinente et on note

$$F(x) = P(X \leq x).$$

Si F est dérivable, la fonction $f(x)$ définie par

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}$$

est la *densité de probabilité* associée à la variable aléatoire X. On a

$$F(x) = \int_{-\infty}^{x} F(t)dt.$$

On a donc

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

On appelle *espérance mathématique* de X la quantité

$$E(X) = \int_{-\infty}^{+\infty} tf(t)dt.$$

La *médiane* $\tilde{m}$ est définie par

$$F(\tilde{m}) = \frac{1}{2}.$$

La *variance* par

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - m)^2 f(x)dx$$

où $m = E(X)$. L'*écart-type* est le nombre σ et le *coefficient de variation* est $CV(X) = \frac{\sigma}{m}$.

On remarquera que l'intégrale joue pour les variables continues le rôle de la somme pour les variables discontinues. Par ailleurs, ces paramètres ont les mêmes propriétés que dans le cas des distributions discontinues.

4.5.2 Exemple : les distributions normales

On appelle *distribution normale* de paramètres σ et m toute distribution continue de densité de probabilité

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-m)^2}.$$

Une variable aléatoire admettant une telle densité de probabilité est dite *normale*. On peut vérifier que m et σ sont respectivement la moyenne et l'écart type de cette distribution. Lorsque $m = 0$ et $\sigma = 1$, on parle de distribution normale réduite.

Figure : Tracé de $f(x)$ et $F(x)$.

Propriétés : **1-** On remarque que $F(m) = \frac{1}{2}$. On a donc $\tilde{m} = m$.
2- Si X est une variable aléatoire normale de moyenne m_X et d'écart-type σ_X, alors $Y = aX + b$ aussi avec pour paramètres

$$m_Y = am_X + b \text{ et } \sigma_Y = |a|\sigma_X.$$

4.6 L'indépendance stochastique des variables aléatoires

Paragraphe non traité cette année

Par extension de la notion d'indépendance de deux événements, on dira que deux variables aléatoires discontinues X et Y sont indépendantes si

$$P(X = x \text{ et } Y = y) = P(X = x)P(Y = y).$$

Pour des variables aléatoires continues X et Y, on définit

$$F(x, y) = P(X \leq x \text{ et } Y \leq y)$$

et on définit la fonction

$$f(x, y) = \lim_{\Delta x, \Delta y \to 0} \frac{F(x + \Delta x, y + \Delta y) - F(x, y)}{\Delta x \, \Delta y} = \frac{\partial^2 F}{\partial x \partial y}.$$

On dit que les variables sont indépendantes lorsque

$$f(x,y) = f_X(x) f_Y(y)$$

où f_X et f_Y sont respectivement les densités de probabilité de X et de Y.
On a alors les propriétés suivantes

Propriétés : **1-** Si X et Y sont indépendantes

$$E(XY) = E(X)E(Y) \text{ et } \sigma^2_{XY} = \sigma^2_X \sigma^2_Y + m^2_y \sigma^2_X + m^2_X \sigma^2_X.$$
$$CV_{XY} = \sqrt{CV^2_X CV^2_Y + CV^2_X + CV^2_Y}.$$

2- La somme ou la différence de plusieurs variables aléatoires normales indépendantes est une variable aléatoire normale.

Chapitre 5

Tests d'hypothèses

(cf. notes de cours de F. Lucas)