STATISTIQUES

Licence pluridisciplinaire
Université d'Angers
2007/08

D. Schaub

Chapitre 1

Introduction

Inventées dans le cadre d'études démographiques, les statistiques ont conservé un vocabulaire proche : population, individu, caractère... Leur but est de décrire une population, estimer des paramètres, tester des hypothèses.

Un premier exemple va nous permettre de cerner le type de problèmes que l'on peut rencontrer. En 1936, on essaya de prévoir les résultats de l'élection présidentielle américaine. Pour ce faire, on constitua un échantillon à l'aide de listes du genre annuaire téléphonique, membres de clubs, etc... L'échantillon ainsi constitué comptait plus de républicains que dans la réalité et, pire même, un quart seulement des personnes interrogées répondirent au questionnaire, et là encore, ceux qui répondirent étaient majoritairement républicains. La conclusion fut : les républicains seront largement vainqueurs. Or, comme on le sait, Roosevelt -démocrate- fut très largement élu.

Pour éviter ce type d'erreur, on a amélioré la façon de concevoir un échantillon. Ainsi, dans l'exemple ci-dessus, chaque électeur doit avoir les mêmes chances de figurer dans l'échantillon. La meilleure méthode est de tirer l'échantillon *au hasard*. Le procédé est souvent un peu plus complexe car il peut, ainsi réalisé brutalement, se révéler lent et coûteux (on peut réaliser des échantillonnages en cascade : par exemple, tirer des villes au hasard et, dans chaque ville sélectionnée, des électeurs au hasard).

Bien évidemment aussi, il faut que l'échantillon compte suffisamment de personnes. En effet, si l'on tire 10 personnes au hasard, il se pourrait très bien que 8 sur 10 soient républicains ; de même que dans un lancer de pièce - non truquée ! -, vous avez certainement pu expérimenter que les chances d'obtenir sur 10 tirages 8 ou 9 neuf fois "pile" existent réellement.

En fait, si l'on dénote par P la proportion vérifiant une certaine propriété dans l'échantillon et π la proportion dans la population totale, P constitue une *approximation* de π, on écrira que $\pi = P \pm m$ où m désigne une marge d'erreur qu'il s'agira d'évaluer. Deux questions se posent : quelle est la taille de m ? quelle est le degré de confiance de la relation ?

Nous montrerons que, pour un échantillon simple choisi au hasard, de taille n, alors, à 95 %,

$$\pi = P \pm 1,96\sqrt{\frac{P(1-P}{n}}.$$

Exemple : Appliquons cette formule au cas d'un sondage de Gallup : en 1998, sur 1500 électeurs, 840 proclamaient leur intention de voter Bush (père) et 660 Dukakis. On désire calculer la proportion d'électeurs qui votera Bush et on veut en être sûr à 95 %. En calculant à l'aide de la

formule précédente, on obtient $P = \frac{840}{1500} = 0,56$, d'où

$$\pi = 0,56 \pm 1,96\sqrt{\frac{0,56 \times 0,44}{1500}} = 0,56 \pm 0,03.$$

Ainsi, Bush devrait être élu, avec une chance de 95 %, avec un score compris entre 53% et 59%. Il fut élu avec $53,9\%$.

Plusieurs causes peuvent fausser les résultats d'un sondage : la taille trop petite, un "mauvais" hasard (parfois d'ailleurs le hasard est impraticable - exemple : pour expérimenter un traitement médical, on ne va pas l'administrer au hasard à toute une population, on peut aussi chercher à l'éviter pour des raisons de coûts : dans de tels cas, il faut voir comment minimiser les erreurs), et aussi le temps passé entre le sondage et le moment final, sans compter les événements qui peuvent se produire. De plus, comme ci-dessus, le résultat n'est pas assuré, il n'est "vrai qu'à 95 %", autrement dit, il y a 5% de chances que ce résultat soit erroné.

Pour finir, disons qu'il y a deux objectifs principaux que peuvent satisfaire les statistiques : *construire des intervalles de confiance*, mais aussi *tester des hypothèse*. On a vu un exemple du premier cas. Pour ce qui est du second, il s'agit de faire une hypothèse concernant une population dans son ensemble et de vérifier, sur un échantillon, la validité de cette hypothèse.

Chapitre 2

Statistiques descriptives

2.1 Notions de base

Définition 2.1.1 *L'ensemble des objets (personnes, plantes,...) étudié s'appelle* ***population****. On le désigne par* $\mathbf{\Omega}$*. Un élément de la population est un* ***individu****, une partie de la population est appelée un* ***échantillon*** *et le nombre d'individus dans la population ou dans un échantillon est appelé* ***taille*** *de la population ou de l'échantillon.*

A chaque individu est associé un certain nombre de **caractères** : valeur, couleur, numéro, réponse... L'association d'un caractère à un individu est décrite par une **variable** X. Cela s'écrit en mathématique : $X : \Omega \to E\,;\ \omega \mapsto X(\omega)$. L'ensemble des valeurs prises par $X(\omega)$ s'appelle ensemble des **modalités**.

Il existe 2 types de variables (ou de caractères) :

• des caractères **quantitatifs continus** (surface habitable, taille d'un individu,...) ou **discrets** (ou discontinus) (nombre de pièces, nombre de personnes,...) qui prennent des valeurs numériques. Dans le premier cas, ces valeurs appartiennent à un intervalle réel, dans le deuxième, ce sont des valeurs isolées (par exemple des valeurs entières).

• des caractères **qualitatifs**. Il y en a de 2 ordres possibles : soit *nominatifs* (type de logement,...), soit *ordinaux* (année de naissance, rang de classement,...). Dans le premier cas, on ne peut ni les ordonner naturellement, ni faire d'opérations, dans le deuxième cas, on ne peut faire d'opérations avec eux, mais on peut les ordonner naturellement.

L'**effectif** ou **fréquence absolue** associé à la valeur $X(\omega) = x$ d'un caractère est le nombre d'éléments de $X^{-1}(x)$. Par exemple, Ω est une classe de 30 élèves, X est la variable qui donne l'âge, entre 15 et 18 ans par exemple, alors l'effectif de 16 est le nombre d'élèves ayant 16 ans. La **fréquence relative** d'une valeur est le rapport de la fréquence absolue sur le nombre d'individus de l'échantillon (ou population).

2.2 Tableaux et graphes

2.2.1 Tableau de fréquences à un caractère

Un tel tableau établit une correspondance entre 2 séries de nombres : les valeurs du paramètre étudié et les effectifs (ou fréquences) correspondants.

Exemple de série quantitative discrète Dans un immeuble de 64 familles, le caractère étudié est le nombre d'enfants par famille. Cela peut se traduire par le tableau suivant :

Nb d'enfants	0	1	2	3	4	5	Total
Nb de familles	16	18	14	11	3	2	64
Fréquence (rel)	0.250	0.281	0.218	0.172	0.047	0.031	1

Exemple de série qualitative Dans le même exemple, on considère le caractère "profession du père" auquel on attribue un code arbitraire.

Profession	Code	Effectif	Fréquence
Ouvriers et employés	1	24	0.375
Cadres moyens et supérieurs	2	9	0.140
Commerçants	3	10	0.156
Fonctionnaires	4	15	0.234
Professions libérales	5	6	0.093

La correspondance valeur-effectif (ou fréquence) définit une **fonction de distribution**.

2.2.2 Tableaux de fréquences cumulées

On peut aussi présenter les résultats ci-dessus sous forme cumulée, par valeurs inférieures ou effectifs cumulés croissants ($<$) :

Nb d'enfants	Effectifs cumulés croissants
Moins de 1 enfants	16 familles
Moins de 2 enfants	16+18=34
Moins de 3 enfants	34+14=48
Moins de 4 enfants	48+11=59
Moins de 5 enfants	59+3= 62
Moins de 6 enfants	62+2= 64

ou par valeurs supérieures ou effectifs cumulés décroissants (≥ 0) :

Nb d'enfants	Effectifs cumulés décroissants
0 enfant ou plus	64 familles
1 enfant ou plus	64-16=48
2 enfant ou plus	48-18=30
3 enfant ou plus	30-14=16
4 enfant ou plus	16-11= 5
5 enfant ou plus	5-3= 2

La correspondance valeur-effectif cumulé définit une fonction dite de **répartition**.

2.2.3 Tableau de fréquence à 2 caractères

On peut représenter 2 caractères simultanément dans un tableau à double entrée. Ainsi, avec le même échantillon que ci-dessus, on considère le caractère X = nombre de personnes vivant dans un appartement, et Y = nombre de pièces par appartement. A l'intersection de la ligne X et de la colonne Y, on trouve le nombre de fois où l'on trouve X personnes dans un appartement de Y pièces.

$Y \setminus X$	2	3	4	5	6	Total nb appart.
2	8	5	2	0	0	15
3	5	7	4	2	0	18
4	3	6	8	9	5	31
Total nb fam.	16	18	14	11	5	64

On peut ainsi remarquer que, dans 9 cas, on trouve 5 personnes dans un appartement de 4 pièces.

Remarque : une question importante lorsque l'on étudie plusieurs caractères est de savoir s'ils sont **corrélés** ou non. Ce que nous étudierons plus loin.

2.2.4 Cas d'une série quantitative continue

Pour rendre compte d'une série statistique continue, il est nécessaire de regrouper les valeurs en intervalles (**classes**) successifs. Les extrémités de ces intervalles sont appelés les **limites de classe**. Dans chaque classe, on remplace les valeurs du caractère par une valeur unique, celle du milieu de l'intervalle ou **centre de classe**.

Exemple : On considère 80 personnes et on s'intéresse à leur taille. On constitue des classes d'intervalle 0.05 m.

Classes	Limites	Centres de classe	Effectifs	Eff. cumulés croissants	Eff. cumulés décroissants
	1.545				80
1.55-1.59		1.57	3		
	1.595			3	77
1.60-1.64		1.62	12		
	1.645			15	65
1.65-1.69		1.67	18		
	1.695			33	47
1.70-1.74		1.72	25		
	1.745			58	22
1.75-1.79		1.77	15		
	1.795			73	7
1.80-1.84		1.82	5		
	1.845			78	2
1.85-1.89		1.87	2		
	1.895			80	

2.3 Représentations graphiques

Avantage : meilleure vue d'ensemble.

2.3.1 Cas sans regroupement en classes

On utilise le **diagramme en bâtons** : les valeurs du caractère sont portées en abscisses, les effectifs ou fréquences correspondantes en ordonnées. En joignant les sommets des bâtons, on obtient le **polygone des fréquences**.

Exemple : en reprenant les exemples ci-dessus, on obtient, par exemple (faire le dessin).
On peut aussi faire le diagramme des fréquences cumulées ou **diagramme intégral** aussi bien croissantes que décroissantes. Reprendre l'exemple ci-dessus.

2.3.2 Cas de regroupement en classes

Dans ce cas, on utilise l'**histogramme** qui constitue une généralisation du diagramme en bâtons.

Premier cas : les classes sont égales (où chaque classe est représentée par un rectangle de base l'intervalle de la classe et de hauteur l'effectif. Le polygone des fréquences est obtenu en joignant les points dont les abscisses sont les milieux des classes et les ordonnées les fréquences.

Là encore, on peut aussi tracer le polygone des effectifs cumulés en remarquant que celui-ci joint les sommets des rectangles (et non les points correspondants aux milieux des classes).

Remarque : l'aire d'un histogramme est proportionnelle au produit de l'intervalle de classe par l'effectif total. S'il s'agit des fréquences (relatives) et en prenant l'intervalle de classe comme unité, l'aire totale est 1.

Deuxième cas : les classes sont inégales. Dans ce cas, si l'on veut encore que l'aire soit proportionnelle à l'effectif, on représente une classe dont l'étendue est n fois l'intervalle de classe fondamental avec une ordonnée égale à l'effectif de cette classe divisé par n.

2.4 Analyse d'une distribution de fréquences

Définition 2.4.1 *On appelle fonction de distribution d'une série statistique l'application qui à une valeur du caractère associe l'effectif ou la fréquence (relative). Son graphe est précisément l'histogramme.*

La fonction de répartition est l'application qui à une valeur du caractère associe soit la somme des effectifs ayant moins que cette valeur, soit la somme des effectifs ayant cette valeur du caractère ou plus. Les graphes sont respectivement les diagrammes intégral croissant et intégral décroissant.

2.4.1 Paramètres de position

Le premier paramètre auquel on peut s'intéresser est le *mode* qui est simplement la valeur la plus fréquente du caractère.

Soit $x_1, \ldots, x_n$ une suite finie de nombres, la *moyenne arithmétique* est le rapport $\overline{X} = (x_1 + \cdots + x_n)/n$.

Remarquons que, si on a $x_1 = x_2 = \cdots = x_n = a$, alors $\overline{X} = a$. Dans cet ordre d'idée, on peut ainsi noter que si la valeur x_i apparaît n_i fois dans la somme précédente, alors

$$\overline{X} = \frac{n_1 x_1 + \cdots + n_k x_k}{n} \text{ où } n = n_1 + \cdots + n_k.$$

Comme les fréquences relatives $f_i = n_i/n$, on pourra aussi écrire

$$\overline{X} = f_1 x_1 + \cdots + f_k x_k.$$

Ainsi dans l'exemple déjà vu, les variables $x_1, \ldots, x_n$ sont les valeurs du caractère "nombre d'enfant" et les n_i sont les effectifs correspondants.

Nb d'enfants	0	1	2	3	4	5	Total
Nb de familles	16	18	14	11	3	2	64
Fréquence (rel)	0.250	0.281	0.218	0.172	0.047	0.031	1

$$\overline{X} = \frac{(16 \times 0) + (18 \times 1) + (14 \times 2) + (11 \times 3) + (3 \times 4) + (2 \times 5)}{64} \cong 1.58 \text{ enfant ou encore}$$

$$\overline{X} = (0.25 \times 0) + (0.281 \times 1) + (0.218 \times 2) + (0.172 \times 3) + (0.047 \times 4) + (0.031 \times 5) \cong 1.58.$$

Lorsque les valeurs sont groupées en classes, on prend pour x_i les centres des classes.

Remarques : 1. on peut s'intéresser à d'autres types de moyennes qui peuvent s'avérer plus *significatives* en fonction des variations des valeurs du caractère : par exemple, la moyenne géométrique, la moyenne harmonique, la moyenne quadratique.

2. On peut faciliter le calcul de la moyenne en changeant d'origine et/ou d'échelle (voir TD.).

Définition 2.4.2 *La médiane est la valeur du caractère M telle qu'il y ait autant d'individus pour lesquels le caractère est inférieur à M que d'individus dont le caractère est supérieur à M.*

Exemple : reprenons le tableau 2.2.4. On détermine la médiane en trouvant d'abord la classe contenant la médiane, puis en faisant une interpolation linéaire dans cette classe. La moitié de l'effectif global est 40. Dans la colonne des effectifs cumulés croissants, on constate que 33 personnes ont une taille inférieure ou égale à $1,695m$. L'interpolation linéaire sur l'intervalle $1,695 - 1,745$, dont l'effectif est de 25 personnes donne pour la médiane

$$M = 1,695 + \frac{0,05 \times (40 - 33)}{58 - 33} = 1,695 + 0,014 = 1,709m.$$

(on considère que la croissance à l'intérieur d'un intervalle est linéaire, par conséquent, la médiane est l'abscisse du point d'intersection entre le segment joignant les deux sommets consécutifs du polygone des fréquences cumulées croissantes et la droite horizontale d'ordonnée la moitié de l'effectif total).

On peut aussi remarquer (ce qui permet une détermination graphique) que la droite verticale qui passe par la médiane coupe l'histogramme en deux parties d'aires égales. Mais elle correspond aussi au point d'intersection de la courbe des effectifs cumulés croissants et de celle des effectifs cumulés décroissants, ou encore, comme ci-dessus à l'intersection de l'une de ces deux avec l'horizontale représentant l'effectif moitié.

Quant à savoir quel, du mode, de la médiane et de la moyenne, est celui qui décrit le mieux une série statistique, cela dépend beaucoup des distributions étudiées. Ainsi, une étude américaine sur les salaires de 1975 a montré que le mode était 0 (du fait que beaucoup de gens n'avaient presqu'aucun revenu), la médiane (élément plutôt stable) est d'environ 8000 $, qui est un bon indicateur puisque 50 % des revenus sont supérieurs et 50 % inférieurs, alors que la moyenne - qui peut être affectée par des variations extrêmes concernant peu d'individus -, est de 10000 $ environ.

Définition 2.4.3 *Le k-ième percentile est la valeur C du caractère pour telle que l'ensemble des individus dont le caractère est au plus C représente les k % de l'effectif global.*

On a, en particulier, les déciles, les quartiles et la médiane. Le calcul des percentiles est analogue à celui de la médiane.

2.4.2 Paramètres de dispersion

La caractéristique la plus "évidente" concernant la dispersion est *l'étendue*, c'est-à-dire l'écart entre la plus grande et la plus petite valeur obtenue. Bien entendu, cette caractéristique n'est pas très stable : il suffit d'une observation très différente pour la modifier considérablement.

Une autre caractéristique souvent utilisée est l'écart interquartile : on mesure l'écart entre le premier et le dernier quartile. On comprend que celui-ci a une meilleure stabilité. Mais d'autres caractéristiques se révèlent plus intéressantes.

i. Variance. Ecart-type

Définition 2.4.4 *La variance d'une série de valeurs est la moyenne arithmétique des carrés des écarts de ces valeurs par rapport à leur moyenne arithmétique :*

$$V = (1/n)\sum_{i=1}^{n} n_i(x_i - \overline{X})^2.$$

Comme la variance est de l'ordre du caractère au carré, il faut en prendre la racine carré pour obtenir un paramètre caractéristique : c'est l'écart-type $\sigma = \sqrt{V}$. On peut aussi remarquer que $V = \overline{X^2} - \overline{X}^2$ (en effet : $V = (1/n)\sum_{i=1}^{n} n_i(x_i - \overline{X})^2 = (1/n)\sum_{i=1}^{n} n_i(x_i^2 - 2x_i\overline{X} + \overline{X}^2) = (1/n)\sum_i n_i x_i^2 - 2\overline{X}(1/n)\sum_i n_i x_i + \overline{X}^2 = \overline{X^2} - 2\overline{X}^2 + \overline{X}^2$, d'où le résultat).

Considérons le tableau suivant (voir plus haut) concernant la taille d'un groupe d'individus, avec une moyenne de $\overline{X} = 1,707m$:

Centres de classe	n_i	$(X_i - \overline{X})^2$	$n_i(X_i - \overline{X})^2$
1,57	3	0,01877	0,05631
1,62	12	0,00757	0,09084
1,67	18	0,00137	0,02466
1,72	25	0,00017	0,00425
1,77	15	0,00397	0,05955
1,82	5	0,01277	0,06385
1,87	2	0,02657	0,05314

On trouve alors $V = (1/80) \times 0,35260 = 0,0044$ et $\sigma = \sqrt{V} = 0,066m$.

ii. Moments

On appelle **moment d'ordre q par rapport à** x_0, la moyenne arithmétique des puissances q-ième des déviations des valeurs du caractère par rapport à x_0 :

$$m_q = (1/n)\sum_{i=1}^{n} n_i(x_i - x_0)^q.$$

On note tout de suite 2 cas particuliers : si $x_0 = 0$ et $q = 1$, m_1 est la moyenne arithmétique $\overline{X}$; si $x_0 = \overline{X}$ et $q = 2$, le moment m_2 n'est autre que la variance $V(X)$.

Ces moments améliorent la connaissance de la distribution, mais généralement, on se limite à $q = 1$ et $q = 2$.

Chapitre 3

Corrélation linéaire, lois classiques

3.1 Problème de la corrélation

La dépendance la plus simple entre deux VA est la relation $aX + bY + c = 0$, avec comme cas particuliers, $Y = \alpha X + \beta$ et $X = \gamma Y + \delta$. Une telle relation n'est peut-être pas vérifiée, tout au moins de manière approchée, pour **tous** les couples (x_i, y_i), mais, en pratique, elle peut l'être pour la plupart. Il s'agit là de corrélation *linéaire*. On pourrait de même chercher si les points (x_i, y_i) sont majoritairement proches d'une autre courbe qu'une droite. Cela traduirait aussi un phénomène de corrélation, mais plus compliqué.

Il s'agit de détecter si une telle relation existe (et la déterminer) et savoir l'interpréter. Dans ce dernier cas, il n'y a pas de théorie pour une telle interprétation. Par exemple, on a remarqué une bonne corrélation entre "l'accroissement du nombre de familles équipées d'un téléviseur" et "l'augmentation du nombre de maladies mentales" durant la même période. Il y a peu de gens pour oser affirmer que cela démontre un lien de cause à effet ou de dépendance entre les deux. Par contre, on croira plus volontiers à une corrélation entre "nombre de navires de passage" et "nombre d'oiseaux mazoutés".

3.1.1 Méthode des moindres carrés

Etant données 2 variables aléatoires X et Y, prenant des valeurs $x_i, y_i, i = 1, \dots n$ il s'agit de rechercher la droite qui approche *au mieux* l'ensemble des points (x_i, y_i). Que signifie "au mieux" ? En fait, c'est la droite qui minimise la somme des distances entre les (x_i, y_i) et la droite $y = ax + b$ (ou $x = ay + b$).

On peut bien sûr calculer cette somme (moyennant le fait de savoir calculer la distance d'un point à une droite) en fonction des coefficients définissant la droite, puis par un calcul de minima, trouver les valeurs qui minimisent cette somme. En réalité, et parce que ce calcul est plus simple, on préfère minimiser la somme des (valeurs absolues des) différences d'ordonnées, c'est-à-dire $S(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$.

Les minimaux ou maximaux d'une fonction de a et b doivent vérifier l'annulation des dérivées partielles $\partial S/\partial a = 0$ et $\partial S/\partial b = 0$. Il s'agit donc de résoudre le système

$$\left\{ \begin{array}{rcl} -\sum x_i y_i + a \sum (x_i)^2 + b \sum x_i & = & 0 \\ -\sum y_i + a \sum x_i + nb & = & 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} a & = & \dfrac{\sum x_i y_i - (1/n)(\sum x_i)(\sum y_i)}{\sum x_i^2 - (1/n)(\sum x_i)^2} \\ nb & = & \sum y_i - a \sum x_i \end{array} \right.$$

En fait, utilisant la définition de moyenne, et remarquant que $b = \overline{Y} - a\overline{X}$, on s'aperçoit

que

$$a = \frac{\sum x_i y_i - n(\overline{X})\,(\overline{Y})}{\sum x_i^2 - n(\overline{X})^2} = \frac{\overline{XY} - \overline{X} \times \overline{Y}}{V(X)}.$$

Au passage, on remarque que la droite de régression passe par le point $(\overline{X}, \overline{Y})$, ce qui est la moindre des choses.

En utilisant la définition de a, en développant et en remarquant que $\sum (x_i - \overline{X})^2 = \sum x_i^2 - n\overline{X}^2$, on constate que le développement de $S(a,b)$ est une somme de 2 termes :

$$S(a,b) = \sum_i (y_i - \overline{Y})^2 - a^2 \sum_i (x_i - \overline{X})^2.$$

Le premier caractérise la dispersion des données en l'absence de relation entre X et Y, le deuxième conduisant à une diminution de cette dispersion lorsque l'on tient compte de la relation (a). Le rapport, $r(X,Y)$, de ces 2 est le **coefficient de corrélation** (qui mesure la précision de l'ajustement), on a :

$$r^2(X,Y) = \frac{a^2 \sum_i (x_i - \overline{X})^2}{\sum_i (y_i - \overline{Y})^2}.$$

On vérifie (calcul) que le coefficient de corrélation r est toujours compris entre -1 et 1. Lorsque $r = -1$, cela traduit une relation linéaire parfaite avec $a < 0$, lorsque $r = 1$, une relation linéaire avec $a > 0$. Une valeur nulle ou voisine de 0 signifie l'absence de relation linéaire entre X et Y.

Remarque : il faut se garder de croire que le coefficient de corrélation mesure une relation de causalité entre X et Y, même lorsque r est voisin de 1.

Il peut être utile d'introduire une autre expression, la **covariance** donnée par

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) = \overline{XY} - \overline{X} \times \overline{Y}. \tag{3.1}$$

On constate alors que

$$r(X,Y) = \frac{\sum x_i y_i - n\overline{X}\,\overline{Y}}{\sum x_i^2 - n\overline{X}^2} \times \frac{\sqrt{\sum (x_i - \overline{X})^2}}{\sqrt{\sum (y_i - \overline{Y})^2}} = \frac{\sum x_i y_i - n\overline{X}\,\overline{Y}}{\sqrt{\sum (x_i - \overline{X})^2}\sqrt{\sum (y_i - \overline{Y})^2}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Exemple : Lors d'un examen, 12 candidats ont obtenu les notes suivantes dans 2 matières A et B différentes :

Candidats	1	2	3	4	5	6	7	8	9	10	11	12
Matière A	3	4	4	5	5	6	6	7	7	8	8	9
Matière B	3	3	5	4	5	5	6	5	6	6	8	7

Pour faire les calculs, il est plus facile de reporter toutes les valeurs dans un tableau et d'utiliser les formes "développées" de a et de r, à savoir :

$$a = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \text{ et } r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}.$$

Ainsi, on peut utilement dresser le tableau suivant :

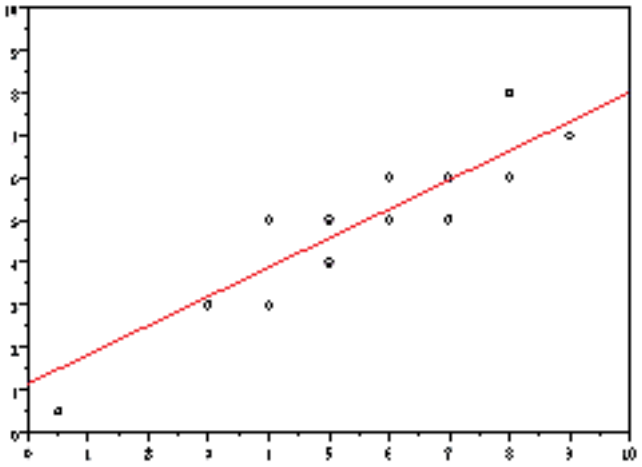

FIG. 3.1 – Droite de régression

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
3	3	9	9	9
4	3	16	9	12
4	5	16	25	20
5	4	25	16	20
5	5	25	25	25
6	5	36	25	30
6	6	36	36	36
7	5	49	25	35
7	6	49	36	42
8	6	64	36	48
8	8	64	64	64
9	7	81	49	63
$\sum$ =72	$\sum$ =63	$\sum$ =470	$\sum$ =355	$\sum$ =404

On a donc

$$r = \frac{404 - \frac{72 \times 63}{12}}{\sqrt{(470 - \frac{72^2}{12})(355 - \frac{63^2}{12})}} = 0,856$$

qui est relativement proche de 1. Il existe donc une corrélation forte entre les 2 séries de notes. La droite de régression (voir figure) est $y = ax + b$ où $a \cong 0,684$, $b \cong 1,146$, qui d'ailleurs passe bien par le point $(\overline{X}, \overline{Y}) = (6; 5.25)$.

3.2 Lois classiques

Rappelons que, pour une *variable aléatoire* discrète ou continue, on définit **l'espérance mathématique** comme étant :

$- > E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$ pour une variable prenant n valeurs ;

$- > E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i)$ pour une v.a. discrète infinie (il faut alors vérifier que l'expression - la "somme de la série" - $\sum_{i=1}^{\infty} |x_i| P(X = x_i)$ a un sens) ;

$- > E(X) = \int_{-\infty}^{+\infty} x f(x) dx$ pour une v.a. continue dont la densité de probabilité est f (là encore seulement lorsque cette expression a un sens).

On définit de même les **moments centrés d'ordre** $k \in \mathbb{N}$ comme :
$-> \mu_k(X) = \sum_{i=1}^{n}(x_i - E(X))^k P(X = x_i)$ pour une v.a. finie ;
$-> \mu_k(X) = \sum_{i=1}^{\infty}(x_i - E(X))^k P(X = x_i)$ pour une v.a. discrète infinie (à condition que la somme ait un sens) ;
$-> \mu_k(X) = \int_{-\infty}^{+\infty}(x - E(X))^k f(x)dx$ pour une v.a. continue (là encore sous réserve que cela ait un sens).

Remarquons que dans tous les cas, on peut écrire : $\mu_k(X) = E[(X - E(X))^k]$.

Cas particulier : si $k = 2$, le moment centré d'ordre 2, $\mu_2(X)$ est encore appelé **variance** de X : $V(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2$. L'**écart-type** est la racine carrée $\sigma = \sqrt{V(X)}$.

3.2.1 Loi binomiale ou loi de Bernoulli

Une variable aléatoire dénombrable X, à valeurs dans $\mathbb{N}$, suit une loi *binomiale de paramètres n et p* si, pour tout $k \in \mathbb{N}$, $P(X = k) = \binom{n}{p} p^k q^{n-k}$ où $q = 1 - p$. On la note $B(n, p)$.

On rencontre cette loi chaque fois qu'il s'agit de déterminer la probabilité de réaliser k fois un événement A dans une série de n expériences aléatoires. Chaque événement ayant la probabilité p de se réaliser et $q = 1 - p$ de ne pas avoir lieu.

Exercice : On pourra à titre d'exercice tracer les diagrammes de $B(8; 0.1)$, $B(8; 0.2)$ jusqu'à $B(8; 0.5)$ et consulter la table de cette loi.

Caractéristiques : $E(X) = np$ et $V(X) = npq$.

3.2.2 Loi hypergéométrique

Une VA dénombrable X suit une loi *hypergéométrique de paramètres N, $K \leq N$ et $n \leq N$* si

$$P(X = k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}.$$

On rencontre ce type de loi dans les tirages *sans remise*. Autrement dit, lorsque l'on dispose d'une population de N individus dont K possèdent un caractère donné et qu'on prélève au hasard un échantillon de n individus.

Caractéristiques : $E(X) = nK/N$ et $V(X) = np\frac{N-K}{N}\frac{N-n}{N-1}$.

En fait, cette loi est assez peu utilisée car elle peut être remplacée par la loi binomiale $B(n, p)$ où $p = K/N$ dès que N est grand par rapport à n, càd. si le "prélèvement" est insignifiant par rapport à la population totale. En pratique, quand $n/N < 0, 1$.

3.2.3 Loi de Poisson

Une v.a. dénombrable X à valeurs dans $\mathbb{N}$ suit une *loi de Poisson de paramètre λ* si, pour tout $n \in \mathbb{N}$, $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$.

Une telle loi intervient lorsque l'événement est rare sur un grand nombre d'observations. Là encore, on pourra étudier le diagramme en bâtons d'une telle loi et consulter une table donnant les valeurs de $P(X = k)$ pour des valeurs de λ entre 0 et 20 (domaine d'utilisation courante).

Caractéristiques : $E(X) = \lambda$ et $V(X) = \lambda$.

Relation de récurrence :

$$P(X = k + 1) = P(X = k) \cdot \frac{\lambda}{k + 1}$$

il suffit donc de connaître $P(X = 0) = e^{-\lambda}$ pour calculer de proche en proche.

3.2.4 Loi normale ou loi de Gauss

Une v.a. réelle continue X suit une *loi normale* si sa densité de probabilité est

$$\forall x \in \mathbb{R},\ f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

où m et $\sigma > 0$ sont deux constantes. On la note $\mathcal{N}(m, \sigma)$.

En effectuant le changement de variable $t = \frac{x-m}{\sigma}$, on définit une nouvelle densité de probabilité

$$\rho(t) = \frac{1}{\sqrt{2\pi}}\ e^{-t^2/2}.$$

Ce changement correspond à un changement d'échelle et à une translation sur l'axe des abscisses. On obtient ainsi une densité ρ indépendante de m et σ, ce qui permet d'utiliser la même courbe pour des v.a. de lois normales de différents paramètres.

La fonction $\rho(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ a une courbe représentative simple ("courbe de Gauss" ou "courbe en cloche"). Elle atteint son maximum en 0, est symétrique par rapport à Oy et tend rapidement vers 0 à l'infini.

La v.a. T qui a pour densité de probabilité ρ suit donc une loi normale $\mathcal{N}(0, 1)$. Son espérance $E(T) = \int_{-\infty}^{+\infty} t\rho(t)dt = 0$ par symétrie par rapport à O. Sa variance, qu'on peut calculer en passant par un calcul d'intégrale double est $V(T) = 1$. Plus généralement, l'espérance et l'écart-type de la v.a. X de loi $\mathcal{N}(m, \sigma)$ sont m et σ.

La fonction de répartition $F(x)$ d'une variable aléatoire X de densité de probabilité $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$ est alors

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du,$$

et par changement de variable, on remarque que $F(x) = \int_{-\infty}^{t} \rho(t)dt$ où $t = (x - m)/\sigma$.

La fonction $H(t) = \int_{-\infty}^{t} \rho(u)du$ peut être calculée pour différentes valeurs de t et les résultats répertoriés dans une table - pour faire l'intégration, on passe par le calcul d'une intégrale double - (pour l'utilisation de tables : voir TD). On peut encore remarquer que $H(0) = 0,5$ et donc $H(t) = H(0) + G(t)$ où $G(t) = \int_0^t \rho(u)du$: ce sont, en fait, les valeurs de G qu'on trouve dans une table.

On ainsi $P(a \leq t \leq b) = H(b) - H(a)$ pour tout intervalle $[a, b]$, et, par symétrie par rapport à l'origine, lorsque $a < 0$, $P(a \leq t \leq b) = G(b) + G(-a)$. En passant à la variable $T = (X - m)/\sigma$, cela devient

$$P(a \leq X \leq b) = P(\frac{a-m}{\sigma} \leq T \leq \frac{b-m}{\sigma}).$$

Exemple : Soit X une VA de type $\mathcal{N}(5, 4)$, on veut calculer $P(3 \leq X \leq 8)$. On a

$$P(3 \leq X \leq 8) = P(\frac{3-5}{2} \leq T \leq \frac{8-5}{2}) = P(-1 \leq T \leq 1,5)$$
$$= G(1) + G(1,5) = 0,3413 + 0,4332 = 0,7745.$$

Plus généralement, on peut ainsi remarquer, en utilisant la table, que

$$P(m - \sigma \leq X \leq m + \sigma) = P(-1 \leq T \leq 1) = 2G(1) \cong 0,683,$$
$$P(m - 1,96\sigma \leq X \leq m + 1,96\sigma) = P(-1,96 \leq T \leq 1,96) = 2G(1,96) \cong 0,95,$$
$$P(m - 2,6\sigma \leq X \leq m + 2,6\sigma) = P(-2,6 \leq T \leq 2,6) = 2G(2,6) \cong 0,99.$$

3.2.5 Somme de variables aléatoires

On dit que deux v.a., définies sur le même espace, sont *indépendantes* si tout événement défini par X seul est indépendant de tout événement défini par Y seul. Cela se traduit, dans le cas discret, par

$$P(X = i \text{ et } Y = j) = P(X = i) \times P(Y = j),$$

pour tout couple d'entiers (i, j), et, dans le cas continu, cela se traduit au niveau des fonctions de répartition

$$H(x, y) = F(x)G(y),$$

pour tout couple $(x, y) \in \mathbb{R}^2$.

On peut aussi rappeler la définition de probabilité conditionnelle, ce qui peut d'écrire pour deux v.a. X, Y par

$$P(X = k|Y = \ell) = \frac{P(X = k \text{ et } Y = \ell)}{P(Y = \ell)}$$

et revoir utilement la formule de Bayes.

On donnera sans preuve les résultats suivants :

Soit $Z = X_1 + \cdots + X_n$ la somme de n variables aléatoires $X_1, \ldots, X_n$ définies sur le même espace. Alors

$$E(Z) = E(X_1) + \cdots + E(X_n).$$

La variance de $X + Y$ est

$$V(X + Y) = V(X) + V(Y) + 2cov(X, Y)$$

où $cov(X, Y) = E(XY) - E(X)E(Y)$ (rappelez-vous la formule donnant la covariance de 2 séries statistiques : $cov(X, Y) = \overline{XY} - \overline{X} \cdot \overline{Y}$ cf. équation 3.1).

D'où si X et Y sont *indépendantes*, ce qui se traduit aussi par $cov(X, Y) = 0$, on a

$$V(X + Y) = V(X) + V(Y)$$

qui se généralise au cas de n variables indépendantes.

Un cas particulièrement intéressant est le cas d'un *échantillon aléatoire*, càd. n variables indépendantes de même loi, $X_1, \ldots, X_n$. Si on note μ et σ leur espérance et variance (commune), on peut considérer la moyenne arithmétique

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

alors $E(Y) = \mu$ et $V(Y) = \sigma^2/n$.

Un élément intéressant est la propriété de stabilité : ainsi, si X et Y sont deux v.a. indépendantes de même loi, il se peut que leur somme suive encore la même loi. C'est le cas pour :

Si $(X_1) \sim B(n_1, p)$ et $X_2 \sim B(n_2, p)$, alors $X_1 + X_2 \sim B(n_1 + n_2, p)$.
Si $(X_1) \sim P(\lambda_1)$ et $X_2 \sim P(\lambda_2)$, alors $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$.
Si $(X_1) \sim N(\mu_1, \sigma_1)$ et $X_2 \sim N(\mu_2, \sigma_2)$, alors $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

3.2.6 Approximation de la loi binomiale

Enonçons encore deux théorèmes d'approximation de la loi binomiale.

Théorème 3.2.1 *Soit (X_n) une suite de variables aléatoires tel que X_n suit la loi binomiale $B(n, p_n)$, pour tout n avec $\lim_{n\to\infty}(np_n) = \lambda$, $\lambda > 0$. Alors (X_n) converge en loi vers une variable aléatoire discrète X qui suit la loi de Poisson $P(\lambda)$, autrement dit $\lim_{n\to\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.*

Utilisation pratique : Si n est assez grand (≥ 30) et p assez petit ($p \leq 0,1$ avec $np \leq 10$), on peut remplacer $B(n, p)$ par $P(np)$.

Théorème 3.2.2 *Soit X une v.a. qui suit une loi binomiale $B(n, p)$. Pour n assez grand et p pas trop voisin de 0 et de 1, X suit à peu près la loi normale $N(np, \sqrt{npq})$ (de même espérance et de même écart-type).*

En pratique : on utilise ce résultat lorsque $n \geq 30$, $np \geq 5$ et $nq \geq 5$.

Exemple : On sait qu'en moyenne 5% des articles produits par un certain procédé de fabrication sont défectueux. Quelle est la probabilité qu'il y ait plus de 60 pièces défectueuses dans un lot de 1000 choisi au hasard ?

En admettant que la taille du lot est petite par rapport à la production totale, on peut utiliser le modèle "avec remise" ; autrement dit, en posant X = nombre de pièces défectueuses, $n = 1000$ et $p = 0,05$,

$$P(X = k) = C_n^k p^k q^{n-k} \text{ d'où } P(X \leq 60) = \sum_{k=0}^{60} C_n^k p^k q^{n-k}.$$

Mais n est "grand" ($n = 1000$, $np = 1000 \times 0,05 = 50$, $nq = 1000 \times 0,95 = 950$), on peut donc admettre que X suit une loi normale $N(np, \sqrt{npq}) = N(50, \sqrt{47,5})$. Il en résulte, si T désigne la loi normale centrée réduite, que $P(X \leq 60) \cong P(T \leq (60-50)/6,9) = 0,5+G(1,45) \cong 0,5 + 0,4265 = 0,9265$ où la valeur $G(1,45)$ a été lue dans la table.
Remarque : pour qu'une "machine" fasse un calcul raisonnable d'une somme binomiale, il faut des valeurs plus petites. Ainsi, pour pouvoir effectuer des calculs dans l'exemple ci-dessus, sans approcher par la loi normale, il faut prendre par exemple $n = 100$ et remplacer 60 par 10 (si ce dernier nombre est > 20, on trouve toujours 1). En *Maple*, on obtient par

sum(binomial(100,k)*0.05$\hat{}$k*0.95$\hat{}$(100-k),k=0..10) $\cong$ 0.9885275899

ou en appliquant directement des fonctions statistiques

stats[statevalf,dcdf,binomilad[100,0.05]](10)

qui donne le même résultat, mais en faisant le calcul pour la loi normale

stats[statevalf,cdf,normald]((10-5),sqrt(4.75) $\cong$ 0.9891092687.

Chapitre 4

Estimation et notions de tests

4.1 Introduction

On s'intéresse à l'étude d'un caractère dans une population P à laquelle on n'a pas accès. Si on extrait plusieurs échantillons représentatifs de taille n fixée, les différences entre les résultats obtenus sont dues à des *fluctuations d'échantillonnage*. A partir d'un échantillon, on n'a donc pas de certitudes mais des *estimations* de paramètres.

L'échantillonnage est dit *non-exhaustif* si le tirage des n individus constituant l'échantillon a lieu avec remise. Il est *exhaustif* s'il est réalisé sans remise. En fait, le plus souvent, la taille d'un échantillon est faible par rapport à la population totale, alors, on peut assimiler le cas exhaustif au cas non-exhaustif. C'est aussi le cas où les théorèmes sont les plus simples.

On distingue deux types de problèmes statistiques :

Les *problèmes d'estimation* dans lesquels il s'agit d'estimer la valeur inconnue d'un paramètre tel que la moyenne d'une variable aléatoire. A partir d'une série d'observations, on cherche à déterminer un nombre dont on a de bonnes raisons de penser qu'il est proche de la valeur inconnue du paramètre. Souvent on se propose aussi de déterminer un intervalle qui a de fortes chances de contenir la paramètre inconnu.

Les *problèmes de tests* dans lesquels on cherche à établir si, au vu des valeurs observées lors d'une série d'essais, il faut accepter ou rejeter une hypothèse sur un paramètre statistique ou sur une forme de loi de probabilité.

4.2 L'échantillonnage

Que peut-on attendre d'un échantillon issu d'une population connue ?

Définition 4.2.1 *On appelle échantillon aléatoire simple un échantillon où chaque individu de la population a la même chance d'être choisi chaque fois que l'on fait une observation. Autrement dit, dans un échantillon aléatoire, chaque observation a la même distribution de probabilité $p(x)$ que la population.*

Exemple : Reprenons l'exemple d'une taille de population de 80 personnes. Nous avons trouvé que la moyenne de cette population était $\mu = 1,7075$ et la variance est de $V = 0,0044$ et l'écart-type $\sigma = 0.06$. Cette moyenne et cet écart-type sont ceux de la population entière des 80 personnes, mais ce sont aussi la moyenne et l'écart-type de la variable aléatoire qui donne le résultat d'un tirage au hasard. Dans un tirage au hasard, la probabilité de tomber sur quelqu'un de taille $1m62$ est de $0.15 = 12/80$.

4.3 Estimation

L'estimation se préoccupe de la représentativité de la population totale par **un** échantillon. Il s'agit d'attribuer une valeur à un paramètre inconnu de la population. On peut chercher à attribuer à ce paramètre une valeur unique (**estimation ponctuelle**) ou un intervalle susceptible de contenir sa valeur (**estimation par un intervalle de confiance**).

Soit X une v.a. dont la loi de probabilité dépend d'un paramètre inconnu θ.

Définition 4.3.1 *Une suite de v.a. $X_1, \ldots, X_n$ indépendantes et obéissant à la même distribution que X, est appelée échantillon aléatoire de taille n de la variable X.*

Remarquons que le choix de n v.a. **indépendantes** permet de modéliser le caractère aléatoire de l'"échantillon.

Toute fonction des variables d'un échantillon aléatoire $T(\phi(X_1, X_2, \ldots, X_n))$ est dite une *statistique*. T est donc une variable aléatoire définie par rapport à la même expérience. L'exemple le plus connu est la *moyenne de l'échantillon* $\overline{X} = (X_1 + \cdots + X_n)/n$.

Une telle fonction T est appelée *estimateur* du paramètre θ si la valeur de T sert à estimer θ. Ainsi, par exemple, $\overline{X}$ est un estimateur de la moyenne $\mu = E(X)$.

On exige, en général, d'un estimateur d'être
$->$ *sans biais*, càd. $E(T) = \theta$ (on définit le biais b par $b = E(T) - \theta$) ;
$->$ *efficace* : plus $Var(T)$ est petit, mieux c'est ;
$->$ *convergent* : $E(T) \to \theta$ et $Var(T) \to 0$ lorsque n augmente indéfiniment.

4.3.1 Estimation ponctuelle de la moyenne

Soit X une variable aléatoire définie sur une population avec $E(X) = \mu$ et $V(X) = \sigma^2$ et l'on veut estimer la moyenne μ et la variance σ^2.

Théorème 4.3.1 *La moyenne de l'échantillon $\overline{X} = (X_1 + \cdots + X_n)/n$ est un estimateur sans biais et convergent de μ ($E(\overline{X}) = \mu$ et $Var(\overline{X}) = \frac{\sigma^2}{n} \to 0$ si n croît indéfiniment).*

Remarquons que $\overline{X}$ est une v.a. dont on peut vouloir trouver la distribution de probabilité $P(\overline{X})$. Pour cela, on peut, soit réaliser un grand nombre de n-échantillons aléatoires, soit utiliser les théorèmes d'approximation du chapitre précédent. On remarque que **quelle que soit la distribution de la population, lorsque la taille de l'échantillon est suffisamment grande** (≥ 10 ou 20 est en général suffisant), **la distribution de $\overline{X}$ suit une loi approximativement normale**.

Remarquons encore que $\overline{X}$ varie autour de la moyenne μ de la population avec un écart-type égal à $\sigma/\sqrt{n}$ (cf. 3.2.5).

Il en va de même pour la proportion : **dans un échantillon aléatoire de taille n, la proportion P de l'échantillon varie autour de la proportion π de la population avec un écart-type $\sigma_P = \sqrt{\pi(1-\pi)/n}$.**

Remarques : Si, dans le théorème précédent, la variance σ^2 est également inconnue, on utilise pour l'estimer

$$S^2 = \sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{n-1}$$

appelée variance de l'échantillon. On vérifie alors que $E(S^2) = \sigma^2$ et que S^2 est convergent.

Par contre, si la moyenne μ est connue, $S^{*2} = \sum_{k=1}^{n}(X_k - \mu)/n$ est un estimateur plus efficace (on vérifie que $Var(S^{*2}) < Var(S^2)$).

4.3.2 Estimation par intervalle de confiance

On a vu ci-dessus que $\overline{X}$ est un bon estimateur de la moyenne μ de la population, mais $\overline{X}$ ne coïncide qu'en moyenne avec μ, en réalité la moyenne $\overline{X}$ d'un échantillon donné est toujours un peu plus grande ou un peu plus petite que μ. C'est pourquoi, il peut être utile de construire un *intervalle de confiance* de la forme :

$$\mu = \overline{X} \pm \text{ marge d'erreur.}$$

Une question importante est : quelle est l'importance de cette marge d'erreur et avec quelle certitude μ appartient-il à l'intervalle ?

Considérons deux échantillons de la même v.a.X : l'un de taille 10, l'autre de taille 100. On suppose que dans celui de taille 10, 3 individus ont la caractéristique A et dans celui de taille 100, 30 ont cette caractéristique. On veut estimer la fraction π d'individus de la population totale qui a cette caractéristique A. Dans les deux cas, l'estimation $p = 0,3$ est la même. Mais il est intuitivement évident qu'on peut avoir plus de confiance dans le deuxième échantillon que dans le premier.

Définition 4.3.2 *Un intervalle de confiance pour un paramètre inconnu θ est la donnée de deux statistiques C_1, C_2 (fonctions de l'échantillon aléatoire) telles que toute réalisation $[c_1, c_2]$ a précisément $1-\alpha$ % chances de contenir μ.*

Le quantité $1-\alpha$ est appelée *niveau de confiance* ou *seuil de confiance* ; ses valeurs sont le plus souvent 0.90, 0.95, 0.99, 0.999.

Considérons une v.a. X suivant une loi normale de variance σ^2 connue. On sait alors que la moyenne de l'échantillon suit une loi de type $N(\mu, \sigma/\sqrt{n})$ et qu'elle constitue un estimateur sans biais de la moyenne inconnue $\mu = E(X)$.

Soit $z_{\alpha/2}$ le nombre positif défini par $H(z_{\alpha/2}) = 1-\alpha/2$, où H est la fonction de répartition normale réduite (voir 3.2.4).

La statistique $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ est du type $N(0,1)$, d'où l'on déduit que

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha,$$

ou encore que

$$P(|\overline{X} - \mu| \leq z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$

Cela signifie qu'avec une probabilité égale à $1-\alpha$, la valeur inconnue de μ est recouverte par l'intervalle aléatoire

$$[\overline{X} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}]$$

qui constitue donc un intervalle de confiance de μ au seuil $1-\alpha$.

Exemple : On sait que la résistance X d'un certain type d'équipements électriques est distribuée (approximativement) suivant une loi normale d'écart-type $\sigma = 0,12$ ohm. Un échantillon de taille 64 a donné comme moyenne empirique la valeur $\bar{x} = 5,34$ ohms. L'intervalle de confiance de la moyenne μ, au niveau 95% est donné par

$$\left[5,34 - \frac{1,96 \times 0,12}{8} \ , \ 5,34 + \frac{1,96 \times 0,12}{8}\right] = [5,31 \ , \ 5,37],$$

sachant que, pour un seuil de confiance de $0,95$, $z_{\alpha/2} = 1,96$, comme on peut le lire dans une table de la loi normale.

Remarques : 1. On procède d'une manière différente, lorsque la variance est inconnue : on considère la statistique $T = (\overline{X} - \mu)/(S/\sqrt{n})$ qui suit une loi *de Student*. Puis, on procède de même façon en utilisant les tables de cette loi.

2. Lorsque la taille de l'échantillon est élevée ($n \geq 50$), un théorème ("central limite") permet d'affirmer que $\overline{X}$ suit une loi proche d'une loi de type $N(\mu, \sigma/\sqrt{n})$.

4.4 Tests statistiques

Le manque de temps nous oblige malheureusement à nous limiter à des considérations générales. Le lecteur peut utilement se référer aux ouvrages cités en référence pour approfondir la question.

Il s'agit de méthodes permettant de décider si un échantillon "empirique" $\{x_1, \ldots, x_n\}$ est compatible avec une hypothèse donnée relative au type d'une loi de probabilité. Ainsi, par exemple, nous nous proposons de tester si une v.a. X obéit à une distribution donnée.

Les résultats d'essais effectués permettent souvent d'émettre une *hypothèse* relative soit à un paramètre inconnu, soit à la loi de probabilité, soit à des liaisons entre v.a. (par exemple : un dé n'est pas truqué, deux machines fonctionnent indépendamment, la durée de vie d'un équipement électrique suit une li exponentielle,..). Il s'agit alors de décider si une telle hypothèse, notée ne général H_0, peut être considérée (avec toujours le risque de se tromper) comme vraie ou fausse.

Il existe des tests d'ajustement (pour vérifier une loi de distribution), des tests paramétriques (pour vérifier des hypothèses relatives à un paramètre d'une loi de probabilité), des tests d'indépendance d'événements ou de v.a. Le test d'ajustement le plus connu est le *test du khi-deux* qui compare les "distances" (ou plutôt leurs carrés) entre les fréquences obtenues par expériences et les fréquences théoriques.

Bibliographie

[1] *A.I.D.E.P, C.I.E.F.O.P.*, Probabilités et statistiques, Dunod (1970).

[2] *E. Amzallag, N. Piccioli, F. Bry*, Introduction à la statistiqueLes nombres, Hermann (1993).

[3] *A. Ruegg*, Probabilités et statistique, Presses polytechniques et universitaires romandes (1994).

[4] *T.H. et R.J. Wonnacott*, Statistique, 4ième édition, Economica, (1972-1995).

Loi normale : la fonction de répartition de la loi normale centrée réduite : $P(0 \leq T \leq t) = G(t) = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{t^2}{2}} dt$

t	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1983	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4397	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4523	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4998	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4983	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4950	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4994	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4999	0,4998	0,4999	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
4,0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Loi binômiale : $P(X = k) = C_n^k p^k q^{n-k}$

n	k			p		
		0,10	0,20	0,30	0,40	0,50
2	0	0,8100	0,6400	0,4900	0,3600	0,2500
	1	0,1800	0,3200	0,4200	0,4800	0,5000
	2	0,0100	0,0400	0,0900	0,1600	0,2500
3	0	0,7290	0,5120	0,3430	0,2160	0,1250
	1	0,2430	0,3840	0,4410	0,4320	0,3750
	2	0,0270	0,0960	0,1890	0,2880	0,3750
	3	0,0010	0,0080	0,0270	0,0640	0,1250
4	0	0,6561	0,4096	0,2401	0,1296	0,0625
	1	0,2916	0,4096	0,4116	0,3456	0,2500
	2	0,0486	0,1536	0,2646	0,3456	0,3750
	3	0,0036	0,0256	0,0750	0,1536	0,2500
	4	0,0001	0,0016	0,0081	0,0256	0,0625
5	0	0,5905	0,3277	0,1681	0,0778	0,0312
	1	0,3280	0,4096	0,3602	0,2592	0,1562
	2	0,0729	0,2048	0,3087	0,3456	0,3125
	3	0,0081	0,0512	0,1323	0,2304	0,3125
	4	0,0004	0,0064	0,0284	0,0768	0,1562
	5	0,0000	0,0003	0,0024	0,0102	0,0312
6	0	0,5314	0,2621	0,1176	0,0467	0,0156
	1	0,3543	0,3932	0,3025	0,1866	0,0938
	2	0,0984	0,2458	0,3241	0,3110	0,2344
	3	0,0146	0,0819	0,1852	0,2765	0,3125
	4	0,0012	0,0154	0,0595	0,1382	0,2344
	5	0,0001	0,0015	0,0102	0,0369	0,0938
	6	0,0000	0,0001	0,0007	0,0041	0,0156
7	0	0,4783	0,2097	0,0824	0,0280	0,0078
	1	0,3720	0,3670	0,2471	0,1306	0,0547
	2	0,1240	0,2753	0,3177	0,2613	0,1641
	3	0,0230	0,1147	0,2269	0,2903	0,2734
	4	0,0026	0,0287	0,0972	0,1935	0,2734
	5	0,0002	0,0043	0,0250	0,0774	0,1641
	6	0,0000	0,0004	0,0036	0,0172	0,0547
	7	0,0000	0,0000	0,0062	0,0016	0,0078

n	k			p		
8	0	0,4305	0,1678	0,0576	0,0168	0,0039
	1	0,3826	0,3355	0,1977	0,0896	0,0312
	2	0,1488	0,2936	0,2965	0,2090	0,1094
	3	0,0331	0,1468	0,2541	0,2787	0,2188
	4	0,0046	0,0459	0,1361	0,2322	0,2734
	5	0,0004	0,0092	0,0467	0,1239	0,2188
	6	0,0000	0,0011	0,0100	0,0413	0,1094
	7	0,0000	0,0001	0,0012	0,0079	0,0312
	8	0,0000	0,0000	0,0001	0,0007	0,0039
9	0	0,3874	0,1342	0,0404	0,0101	0,0020
	1	0,3874	0,3020	0,1556	0,0605	0,0176
	2	0,1722	0,3020	0,2668	0,1612	0,0703
	3	0,0446	0,1762	0,2668	0,2508	0,1641
	4	0,0074	0,0661	0,1715	0,2508	0,2461
	5	0,0008	0,0165	0,0735	0,1672	0,2461
	6	0,0001	0,0028	0,0210	0,0743	0,1641
	7	0,0000	0,0003	0,0039	0,0212	0,0703
	8	0,0000	0,0000	0,0004	0,0035	0,0176
	9	0,0000	0,0000	0,0000	0,0003	0,0020
10	0	0,3487	0,1074	0,0282	0,0060	0,001
	1	0,3874	0,2684	0,1211	0,0403	0,0098
	2	0,1937	0,3020	0,2335	0,1209	0,0439
	3	0,0574	0,2013	0,2668	0,2150	0,1172
	4	0,0112	0,0881	0,2001	0,2508	0,2051
	5	0,0015	0,0264	0,1029	0,2007	0,2461
	6	0,0001	0,0055	0,0368	0,1115	0,2051
	7	0,0000	0,0008	0,0090	0,0425	0,1172
	8	0,0000	0,0001	0,0014	0,0106	0,0439
	9	0,0000	0,0000	0,0001	0,0016	0,0098
	10	0,0000	0,0000	0,0000	0,0001	0,0010

Table des matières