Collecte des données, analyse univariée et bivariée

Introduction

Dans l'époque actuelle où le savoir produire ne suffit plus pour assurer la place d'une entreprise ou d'un produit sur le marché à cause d'un environnement très concurrentiel, l'approche marketing semble offrir la clé du succès.
L'idée centrale du marketing est la concentration sur le consommateur(ou le client) et l'adaptation à ses attentes. Le marketing c'est un *état d'esprit*, c'est un *art*, mais c'est aussi une *science*. Traditionnellement on fait la distinction entre deux phases du marketing, la *phase analytique* ou la recherche marketing (Marketing Research) et la *phase opérationnelle* qui regroupe le marketing opérationnel et stratégique.
La *recherche marketing* développe un ensemble de techniques pour l'étude du consommateur, de la clientèle et du marché. Elle détecte et enregistre les besoins et motivations des clients, qu'elle déduit souvent à partir des attitudes exprimées ou des comportements observés. Elle dispose de tout un arsenal cohérent de méthodes pour arriver à évaluer les marchés et leurs attentes, pour identifier la perception (positionnement) des produits et des services par rapport aux attentes et aux concurrents, pour segmenter la clientèle afin d'isoler et cibler les groupes que l'entreprise avec ses moyens et ses compétences peut le mieux servir.

Figure 1. Le marketing et le consommateur

Le marketing opérationnel et le marketing stratégique utilisent les informations fournies par la recherche marketing pour développer et appliquer une politique de produit, de prix, de communication et de distribution adaptée aux attentes de la clientèle ciblée.

L'objectif de ce cours est de présenter les *techniques de recueil, traitement et analyse des données* intégrées au *cadre conceptuel de la recherche marketing*. Ces techniques sont illustrées par de nombreux *exemples et études de cas*. La solution de la plupart des traitements des données qu'elles impliquent est présentée sur *tableur* (Excel), qui est l'instrument microinformatique le plus utilisé dans les entreprises. Un *logiciel d'analyse des données* multidimensionnelles écrit par l'auteur est mis à disposition. Il intègre sur le tableur les méthodes d'analyse multivariée les plus importantes pour la pratique de la recherche marketing.

La typologie des études de marché

Toute recherche commence par une phase **exploratoire**. Les entretiens individuels et/ou de groupe sont des moyens de découverte de la problématique du marché (ou domaine) étudié. Avec la problématique éclairée on passe à la phase **descriptive**, à l'aide de techniques *d'observation* et de *communication*. Les techniques d'observation permettent d'enregistrer le comportement d'une manière inobtusive. Mais dans la plupart des cas l'information est obtenue par enquête sur échantillon à l'aide d'un questionnaire. Dans cette catégorie on trouve les enquêtes ponctuelles, les enquêtes omnibus les panels. Par fois l'expérimentation à travers l'étude **causale** est aussi exigée. Les tests de concept, les copy tests, tests de produit, les marchés test servent à accroître la qualité de la décision managériale.

La collecte des données et les sources d'informations

Pour obtenir des informations il y a le choix entre des données préexistantes (**information secondaire**) publiées ou non publiées, externes ou internes à l'entreprise qui sont souvent peu coûteuses mais aussi peu adaptées aux objectifs de l'étude et des données recueillies directement en observant ou en interrogeant les clients (**information primaire**).
Les *données internes* sur les ventes, les clients, les actions marketing passées, les coûts, sur la distribution et les acheteurs organisés souvent en base de données, sont des outils redoutables (voir essentiels) pour cibler la clientèle. Le marketing direct, la vente par correspondance, le marketing des bases de données illustrent la force que peuvent apporter les données internes à une entreprise.
Les *données externes* sont soit publiques (en libre accès ou à des coûts modiques) soit commerciales. Parmi les données *publiques* on trouve des informations publiées ou des informations accessibles on-line dans des bases de données spécialisées ou (et maintenant de plus en plus) sur Internet.
Les informations *commerciales* sont offertes par des services d'informations standardisées pour le marketing (des profils clientèle, mesure des ventes et des parts de marché, mesure d'exposition à des medias et de leur impact et plus rarement des études à la carte).

Les informations primaires

Les données à obtenir par une étude standard. Les modèles conceptuels du *comportement du consommateur* que ce soit le consommateur individuel ou le consommateur industriel sont un guide essentiel pour spécifier les informations nécessaires.
D'une manière simplifiée on suppose que le consommateur est modelé par des facteurs d'environnement (de son environnement), que par rapport à un produit ou service il est soumis à un ensemble de stimuli qui viennent de sources incontrôlables par l'entreprise (l'environnement) mais

aussi à des stimuli générés par l'entreprise ou bien par le produit lui-même.

Parmi les **facteurs d'environnement** qui déterminent le comportement du consommateur on trouve les caractéristiques socio-demographiques (âge, sexe, occupation, état civil, niveau d'éducation, classe ou catégorie sociale, cycle de vie familial) mais aussi les caractéristiques psycho-sociales (personnalité, style de vie). Tous ces éléments sont importants pour la segmentation du marché.

La **réaction aux stimuli** générés par le marché est progressive, les individus ne deviennent pas brusquement des acheteurs d'une marque (produit/service), ils évoluent en apprenant et traversent plusieurs phases. Il s'agit d'une étape cognitive (information, apprentissage, connaissance), d'une étape affective (ou se forment les attitudes et les préférences et ou les motivations ont un rôle à jouer) et étape conative (d'action, achat). L'ordre de ces étapes diffère en fonction des situations d'achat (premier achat, achat répété, produit impliquant ou non-impliquant etc.). Pour l'entreprise il est important de savoir dans quelle phase se trouve le client par rapport au produit offert, afin de préparer et doser les stimuli à administrer et trouver un mix-marketing adapté pour faire avancer les sujets visés vers l'état d'acheteur ou d'acheteur fidèle. C'est pour cela que le recueil direct d'informations auprès de la clientèle devra permettre de déceler progressivement par rapport au produit, le degré de connaissance, les attitudes, les intentions d'achat, les motivations, les habitudes et comportements d'achat.

Moyens pour obtenir les informations primaires

Les données primaires sont obtenues par observation ou par communication directe. La communication suppose souvent la rédaction d'un questionnaire et c'est le seul moyen pour obtenir des informations sur les attitudes, connaissances, motivations et intentions. L'observation est plus adaptée pour procurer des données comportementales.

Le questionnaire : administration, collecte, dépouillement

Le questionnaire est l'instrument le plus fréquemment utilisé pour le recueil direct des informations. Après une rédaction attentive qui essaye de couvrir tout les informations à obtenir, suivi par des phases de prétest et correction, le questionnaire est administré (par voie postale, téléphone ou télématique) à un échantillon représentatif de la population ciblée. Les questionnaires remplis sont collectés et dépouillés. Toutes les réponses sont codifiées et enregistrées dans des tableaux de données (appelés parfois bordereaux de dépouillement) où chaque ligne représente un questionnaire, comme dans la Figure 2.

(a) exemple de questionnaire :

Questionnaire No. 1 (1)

Courte présentation des objectifs de l'étude....

1) Laquelle des marques suivantes préférez-vous? (2)
Marque A ___ 1
Marque B ___ 2
Marque C _x_ 3
Autres ___ 4

2) Qui vous a signalé cette marque ?

	OUI 1	NON 2	
la Pub	_x_	___	(3)
la presse spécialisée	___	_x_	(4)
des amis	_x_	___	(5)
autres sources	___	_x_	(6)

3) Indiquez votre degré d'accord sur la possession des attributs suivants par les marques A,B et C: (1 = pas du tout d'accord, 2 = pas d'accord, 3 = d'accord 4 = tout à fait d'accord)

	Marque A 1 2 3 4	Marque B 1 2 3 4	Marque C 1 2 3 4
Pas cher	_ _ _ x(7)	x _ _ _(11)	_ _ x _(15)
Bonne qualité	_ x _ _(8)	_ _ _ x(12)	_ _ x _(16)
Innovant	x _ _ _(9)	_ _ x _(13)	_ x _ _(17)
Commode	_ _ x _(10)	_ _ x _(14)	_ _ x _(18)

4) Etes vous ? (19)
Femme _x_ 1
Homme ___ 2

5) Quel est votre âge ?
_ _ (20)

(b) exemple de tableau de dépouillement d'enquête

```
00122112344242414312 2
00211112143113332313 22
00311112144144144143 42
.....
59942112112322111134 21
```

Figure 2. Le questionnaire: précodage, codage et enregistrement des reponses

Le questionnaire présenté en Figure 2 est un exemple réduit et simplifié. Un vrai questionnaire (le cas CAMIP) est donné en annexe ainsi qu'une collection des types de questions qui tient compte d'un modèle hiérarchique du comportement du consommateur.

La codification des questions

Type de question	Nombres de colonnes (variables) par question
Questions **nominales**:	
Q. au choix multiple exclusif	une
Q. au choix multiple non exclusif	autant de colonnes que le nombre de modalités (codées 0 ou 1)
Q. ouvertes (autres ...)	une et on augmente le nombre de modalités en fonction des réponses
questions. **ordinales**	plusieures modalités (tant de colonnes que propositions possibles)
questions. **quantitatives**	une; si échelle d'attitude une colonne pour chaque item

Tableau 1.1 - Règles d'enregistrement des réponses aux questionnaires dans les tableaux de données d'enquête

Il est possible d'enregistrer les réponses au questionnaire directement sur l'ordinateur. Les questionnaires automatisés sur ordinateur ont l'avantage d'éliminer en grande partie les fautes de frappe, ils codent et enregistrent les données automatiquement dans des tableaux. Un exemple de questionnaire automatisé sur tableur est donné en annexe. Sur le Web de l'Internet le mécanisme des formulaires (Forms) facilite la saisie des réponses à l'aide d'objets d'interface adaptés (zones de texte, zone d'entrée, cases d'options, cases à cocher, listes à sélection unique ou multiple etc.). Les informations saisies sont captées sur le serveur de celui qui mène l'enquête et enregistrées dans une base de données ou autrement traitées par des programmes adaptés, qui profitent d'un autre standard qui s'est imposé sur Internet le CGI (Common Gateway Interface). En profitant des deux standards les Forms et les CGI il est relativement facile de mener des enquêtes interactives en temps réel et à distance. Des informations supplémentaires concernant la préparation des enquêtes sur Internet se trouvent dans l'annexe.

Nature des informations

Types d'informations[1]

Informations quantitatives ➔ résultent de l'observation des comportements des intervenants sur le marché, elles sont relativement objectives et contrôlables et répondent aux questions "combien ?" et "comment ?" (ex. notoriété des produits, possession des produits, quantités achetées, fréquences d'achat, lieux d'achat, modes d'information sur les produits etc.)
Informations qualitatives ➔ s'intéressent aux facteurs qui déterminent les comportements et sont de nature subjective. Leur recueil et interprétation sont en général plus difficiles (ex. les motivations, perceptions, les opinions et attitudes, les préférences etc.)

La mesure et les échelles de mesure

En marketing le questionnaire sert aussi comme instrument de mesure.
La **mesure** est définie: comme étant formée de "règles pour attribuer des nombres à des objets, de telle sorte qu'elles représentent des quantités d'attributs"[2] (1. on mesure des attributs des objets, non pas les objets eux-mêmes; 2. la manière dans laquelle les nombres sont attribués n'est pas spécifiée, les nombres jouent un rôle de symbole, ce ne sont pas les propriétés intrinsèques des nombres, qui sont prises en compte mais les propriétés des attributs qu'ils représentent). Une autre définition considère la mesure "comme le moyen d'obtenir des symboles représentant les propriétés des personnes, d'objets, d'événements ou d'états, de telle sorte que ces symboles aient les mêmes relations entre eux que les choses représentées"[3].
Les **échelles nominales**: ont comme seule propriété *l'identité*, la plus simple des propriétés d'une échelle de nombres. L'identité exprime l'appartenance des objets étudiés à une catégorie. Elles sont les plus pauvres en informations.
Les **échelles ordinales** ont une seconde propriété des échelles de nombres, l'*ordre* , qui est ajouté à la première. Elles permettent de classer les objets selon un ordre donné.
Elles apparaissent dans les mesures de préférences sous la forme d'échelles de classement, comparaison par paires, tiercé des préférences etc..
Les **échelles d'intervalle** ajoutent une troisième propriété des échelles de nombres, le fait que *l'intervalle* entre les nombres a un sens, ce qui signifie que les *différences* peuvent être comparées. (exemple: les échelles de température Celsius et Farenheit)
Les **échelles de ratio** se différentient par rapport aux échelles d'intervalle par le fait qu'elles possèdent un *zéro absolu*.

Le tableau de données: documentation, vérification, transformation des données.

Dans la Figure 3. est présentée une organisation possible sur tableur des données collectées par enquête.

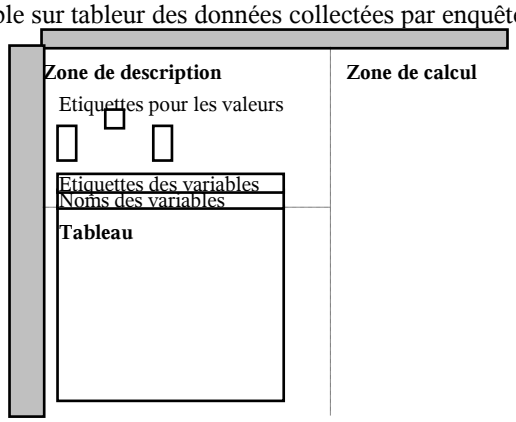

Figure 3. Organisation des données sur une feuille de calcul (sur tableur)

Chaque colonne du tableau représente une variable (voir par exemple le tableau de la Figure 2.) . Pour faciliter la sélection des variables pour les éventuels traitements, elles doivent avoir des noms.

[1]Cf. J-P. Vedrine "Le traitement des données en marketing ", Ed. d'Organisation 1991, p.14-20.
[2]Jum C. Nunnally, "Psychometric Theory, 2nd ed.", New York: McGraw-Hill, 1978), p.3.
[3]P.E. Green, D. S. Tull et G. Album, "Research for marketing decisions", Prentice -Hall, 5-e édition, 1988, p.242

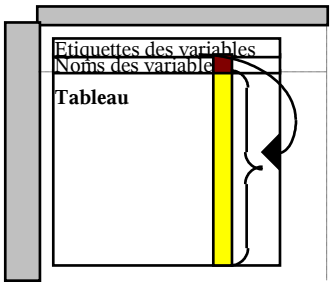

Figure 4. La place et le nom d'une variable dans le tableau de données (tableur)

Analyses préliminaires des données: tableaux de fréquence et statistiques descriptives

Les premiers traitements appliqués aux données cherchent à résumer et à mettre en évidence les éventuelles erreurs et/ou données aberrantes qui ont échappé à la correction au moment de la saisie. Il s'agit des statistiques descriptives et de tableaux de fréquence (simples et croisés). Une grande partie des études marketing (90%) ne vont pas plus loin dans les analyses pour décrire le marché.

Tableaux de fréquence

Les tableaux de fréquence montrent le nombre des cas enregistré dans différents intervalles ou catégories. Les tableaux de fréquences simples (tris à plat) comptent les fréquences pour une seul variable. Ils sont souvent representé sous forme d'**histogrammes**.

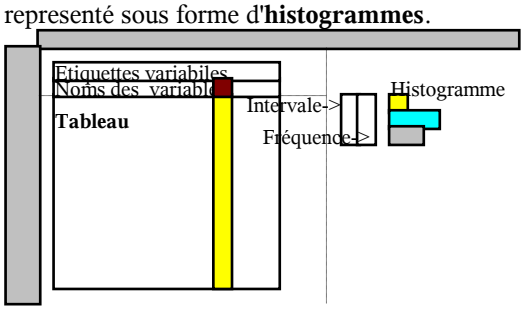

Figure 5. Le tableau de fréquence simple (ou tris à plat). Sur tableur il y a la fonction FREQUENCE(nom variable, matrice d'intervalles) qui le calcule.

Voila comment on calcule sur un tableur la frequence et la frequence cumulé des reponses à la première question de la mini-enquête présenté en Figure 3.

	D	E	F	G	H
1		Interv.	Freq.	Freq.Cum	<--Formule
2	Marque A	1	3	3	=F3
3	Marque B	2	4	7	=G3+F4 ↓
4	Marque C	3	7	14	..
5	Marque D	4	2	16	..

Tableau 1.2 - Calcul de la distribution et répartition de fréquence sur tableur

A partir de la distribution de fréquence et la répartition des fréquences (ou la fréquence cumulée) on calcule un grand nombre de statistiques descriptives. La Figure 6. montre graphiquement comment des statistiques comme la moyenne, la médiane, les quartiles sont déduites a partir d'une distribution de fréquence. Elle montre aussi quel est l'intérêt d'utiliser la fréquence cumulée pour identifier certaines de ces mesures. Pour mieux reconnaître l'allure de la distribution de fréquence empirique on trace sur l'histogramme le **polygone de fréquence** en réunissant les points qui se trouvent au milieux sur le coté haut des barres.

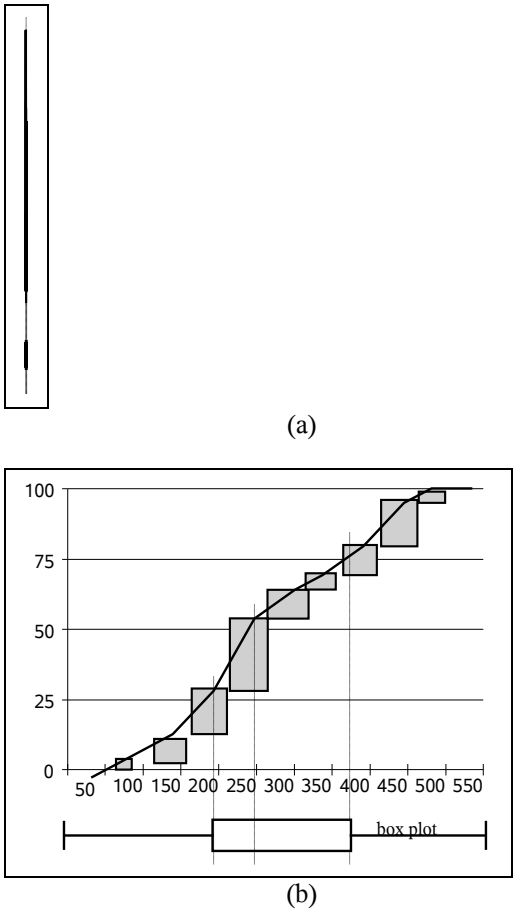

Figure 6. Anatomie d'une distribution de fréquence (a) et d'une répartition de fréquence (b) . Le lien entre un histogramme et un box plot

Les tableaux croisés (tris croisés) comptent la fréquence des cas ayant de caractéristiques (intervalles) communes sur deux ou plusieurs variables. Ils résument l'information contenue dans les variables juxtaposés et permettent d'analyser l'association et les liens entre les variables. Les tableaux croisés de deux variables seront traités au chapitre sur les analyses bivariés et les tableaux croisés multiples apparaîtront dans les chapitres qui portent sur l'analyse de données multidimensionnelle.

Statistiques descriptives

Mesures de position (de tendance centrale)

Moyenne (arithmétique) : la moyenne d'une série statistique de n observations est le quotient de leur somme par leur nombre.

$$\bar{X}_k = \Sigma X_{ik}/n \qquad (1)$$

Médiane: la valeur d'une série statistique pour laquelle le nombre d'observations inférieures ou supérieures à cette valeur sont égal. La détermination de la médiane nécessite le classement de la série par ordre de grandeur (croissante ou décroissante). S'il y a n = 2p + 1 observations celui de rang p+1 sera la médiane. Si par contre n = 2p observations toute valeur comprise entre celle de rang p et celle de rang p+1 peut convenir comme médiane. Mesures appariées **quantile**, **quartiles**(quantile d'ordre 4, Q1, Q2 et Q3) et **déciles**, centile.

Mode: le mode et la valeur la plus souvent rencontrée d'une série statistique

Mesures de dispersion

Les indicateurs de dispersion

ont pour objet de mesurer la plus ou moins grande concentration des valeurs autour de leur tendance centrale.

Etendue (Range)

La différence entre la valeur maximum et minimum d'une série.

Etendue

est l'intervalle qui sépare les deux valeurs extrêmes

Ecart moyen

est la moyenne des valeurs absolues des écarts par rapport à la moyenne;

Dispersion ou variation totale

est la somme des carrés des écarts par rapport à la moyenne :

$$DISP(k) = \Sigma (X_{ik} - \bar{X}_k)^2 \qquad (2)$$

Variance

est la moyenne des carrés des écarts par rapport à la moyenne :

$$VAR(k) = \sigma(k)^2 = \Sigma (X_{ik} - k)^2/n = DlSP(k)/n(3)$$

Si la moyenne a été obtenue sur échantillon, un degré de liberté a été consommé pour le calcul de cette moyenne et la variance devient

$$VAR(k) = s(k)^2 = \Sigma (X_{ik} - k)^2/(n-1) \qquad (4)$$

Ecart type

est la racine carrée de la variance : dans le cas d'un calcul sur échantillon :

$$ET(k) = s(k) = [\Sigma (X_{ik} - k)^2/(n-1)]^{1/2} \qquad (5)$$

Données qualitatives

Des données qualitatives apparaissent chaque fois que la personne interrogée a le choix entre plusieurs modalités qui lui sont proposées explicitement ou implicitement comme dans le cas d'une question ouverte avec post-codification.

La fréquence relative (proportion) d'une modalité joue rôle de moyenne

Pour chaque individu, la réponse correspond à un code ou éventuellement à plusieurs si le choix est multiple (voir questionnaire CAMIP). Sur l'ensemble de la population enquêté, on dénombrera alors le nombre de fois qu'un code j donné est apparu pour la variable k étudiée : ceci indique la fréquence absolue $\mathbf{N_{jk}}$ de la modalité. Si cette fréquence absolue est rapportée aux N personnes considérées, on obtient la fréquence relative de la modalité

$$pjk = Njk/N \qquad (6)$$

Variance et écart type d'une proportion

Pour une modalité donnée, la fréquence relative joue un rôle similaire à celui de la moyenne pour les variables quantitatives. Des Indicateurs de dispersion sont également disponibles. Dans la mesure ou un individu a choisi ou non une modalité donnée, on a affaire à un processus binomial. Il est donc possible d'associer une variance et un écart type à chaque modalité d'une variable qualitative :

$$VAR(jk) = (p_{jk})\,(1 - p_{jk})/N \qquad (7)$$

$$ET(jk) = [(p_{jk})\,(1 - p_{jk})/N]^{1/2} \qquad (8)$$

On constate que ces indicateurs sont d'autant plus faibles que p_{jk} est proche de 1 ou de 0. Dans les deux cas, cela signifie que les réponses sont très concentrées, soit sur la modalité j, soit sur l'ensemble des autres modalités.

Données ordinales

Les données ordinales sont plus difficiles à présenter que les autres catégories de données. Comme on l'a vu, il s'agit de données concernant des rangs de préférence ou de similarité.

On notera que la notion de rang moyen n'a pas de signification, le passage d'un rang au suivant ne correspondant généralement pas a une variation d'intensité de préférence constante[4].

Si l'on prend le cas des préférences, pour chaque individu, on disposera d'un classement des m items proposés. Sur l'ensemble de la population interrogée, il sera ainsi possible de comptabiliser :

fonction de répartition des rangs

- le nombre de fois qu'un item donné a été classé en 1^{re} position, en 2^{e}.., en m^{e} ;

matrice de préférences.

- le nombre de fois qu'un item donné a été classé avant un autre item;

Analyses univariées[5]

[4]Certaines procédures d'agrégation des rangs existent. On citera, pour le cas des préférences, l'échelle de Thurstone (cf. P.E.Green, D.S. Stull et G. Albaum, *Resesearch for marketing decision,* Prentice-Hall, 1988 et pour le cas des similarités, la procédure de Coombs (cf. P.E. Green et V. Rao, *Applied rnultidimesional scaling,* Holt, Rinehart, Winston, 1972, p. 193 et ss).

[5]Cf. J-P. Vedrine "Le traitement des données en marketing ", Ed. d'Organisation 1991, p.25-39.

Analyse des données quantitatives

Les intervalles de confiance

Dans la plupart des cas, une enquête ne portera que sur un échantillon extrait de la population étudiée. On aura alors à déduire des résultats obtenus sur échantillon les valeurs, c'est-à-dire celles qui seraient disponibles si l'ensemble de la population était connue.

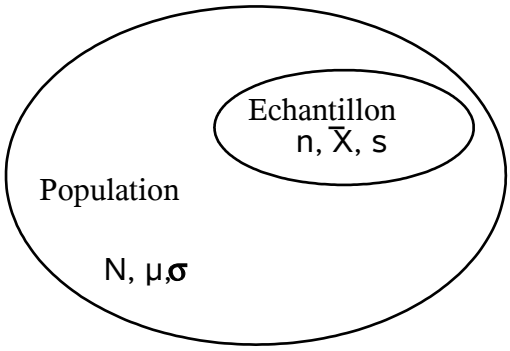

Figure 10: Caractéristiques de la population totale et de l'échantillon

Figure 7 - Population, échantillons et distribution d'échantillonnage

Théorème Central Limit

Quand on tire des échantillons de dimension **n** d'une population à moyenne **μ** et variance σ^2 pour des **n** grands la **moyenne des échantillons** sera distribuée approximativement normalement avec une moyenne égale à **μ** et une variance **σ** égale à σ^2/n

Comme **σ** est inconnu on l'estime à partir de s:

$$\sigma \approx s = \sqrt{\quad} \qquad (9)$$

Pour des échantillons exhaustifs quand n/N < 1/7

$$\qquad (10)$$

Distribution normale d'une variable centrée et réduite Z (m=0 et ET = 1)

-3,5	-3	-2,5	-2	-1,5	-1	-0,5	0	0,5	1	1,5	2	2,5	3	3,5
0,00	0,00	0,02	0,05	0,13	0,24	0,35	0,40	0,35	0,24	0,13	0,05	0,02	0,00	0,00
0,00	0,00	0,01	0,02	0,07	0,16	0,31	0,50	0,69	0,83	0,93	0,97	0,99	1,00	1,00

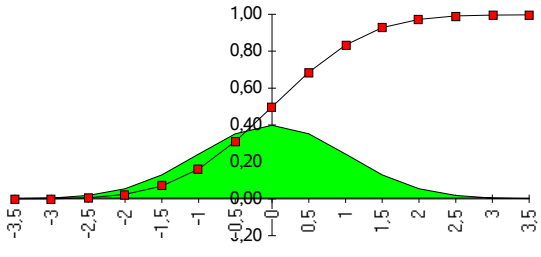

Figure 9 - La distribution normale

a) Détermination de l'intervalle de confiance

La population totale est de taille **N**; la valeur vraie de la moyenne de la variable analysée est **μ**, et son écart type **σ**. Ces deux valeurs **μ** et **σ** sont inconnues, mais sur l'échantillon de taille **n**, une moyenne $\bar{X}$ et un écart-type **s** ont été repérés (cf. graphique 1). Il s'agit de déduire **μ** et **σ** de ces valeurs $\bar{X}$ et **s**.

Cette déduction suit des règles simples issues de la théorie des sondages, dans la mesure où les hypothèses suivantes sont respectées :

- les éléments de l'échantillon ont été sélectionnés de manière aléatoire;
- l'échantillon est non exhaustif (n/N < 1/7)
- l'échantillon comprend au moins 30 individus.

Dans ces conditions, on montre[6] que les moyennes d'échantillon suivent une loi normale de moyenne **μ** et d'écart type **σ**, avec :

σ = **σ** /

Comme **σ** est inconnu, il est estimé à partir de **s** :

σ ≈ **s** = $[\Sigma(X - X)^2/(n - 1)]^{1/2}$

Si l'on désire travailler avec un seuil de confiance 1-α, un intervalle de confiance pour la moyenne **μ** est obtenu à l'aide de l'expression:

$$\mu = \pm z_{\alpha/2} \cdot \sigma \qquad (11)$$

où $z_{\alpha/2}$ est la valeur lue dans la table de la loi normale réduite pour une probabilité (1 - α/2). Il y a ainsi une probabilité (1 - α) que la valeur recherchée se situe dans cette fourchette.

Exemple:

L'association des étudiants d'une université envisage d'ouvrir un ciné-club; afin d'en évaluer la fréquentation, elle a réalisé une enquête par sondage sur un échantillon de 400 individus.

Une moyenne de fréquentation de 10 séances par an et par individu a été obtenue avec un écart type s = 20. Au seuil de confiance (1 - α) = 95 % l'intervalle de confiance est:

μ = 10 ± (1,96) (20/) = 10 ± 1,96

Il a 95 chances sur 100 de se situer dans la fourchette [11,96: 8,04]. Si l'université comprend 5000 étudiants, une fréquentation globale de 50 000 places peut être attendue en moyenne; la fréquentation globale a 95 % de chances de se situer dans l'intervalle [59800: 40200].

b) Cas particuliers

[6]Pour la démonstration mathématique, voir V. Giard, *Statistique appliquée à la gestion,* Economica, 1985, p. 176 et ss.

Pour une démonstration empirique voir annexe

1- Dans le cas d'un *échantillon exhaustif*, c'est-à-dire avec n > N/7, l'écart type σ, des moyennes d'échantillons doit être corrigé par le facteur d'exhaustivité [(N - n)/(N - I)]1/2. L'intervalle de confiance devient alors :

$$\mu = \pm z_{\alpha/2} \cdot \sigma_{\bar{x}} \cdot [(N - n)/(N - 1)]^{1/2} \quad (12)$$

On remarque que si n est faible par rapport à N, (N - n)/(N - 1) est proche de 1. Au contraire, si n est grand par rapport à N, (N - n)/(N - 1) est proche de 0; à la limite, pour n = N, μ = .

Exemple:

Dans l'exemple précédent, supposons que l'université considérée ne comporte que 2 000 étudiants au total. L'échantillon de 400 personnes prélevé par l'association des étudiants doit donc être qualifié d'exhaustif, et il faut utiliser le facteur de correction, égal ici à [(2 000 - 400)/(2 000 - 1)] = 0,80.

Au seuil de confiance (1 - α) = 95 %, l'intervalle de confiance devient:

$\mu = 10 \pm (1{,}96)\ (20/)\ (0{,}80)^{1/2} = 10 \pm 1{,}74$

μ a 95 chances sur 100 de se situer dans la fourchette [11,74: 8,26].

2 - Dans le cas d'un *petit échantillon*, avec n < 30, et lorsque σ est estimé, les moyennes d'échantillons ne sont plus répartis autour de la moyenne vraie selon une loi normale, mais selon une loi de Student à n - 1 degrés de liberté. Dans la formule (1l), $z_{\alpha/2}$ est alors remplacé par $t_{\alpha/2}$, lu sur la table de Student pour n -1 degrés de liberté et un seuil de confiance (1 - α).

Exemple:

Au lieu d'utiliser un échantillon de 400 personnes, L'association des étudiants s'est limitée à 21 interviews. La moyenne d'échantillon (15) suit une loi de Student à 20 degrés de liberté. Dans la mesure où l'écart type repéré sur l'échantillon s'élève à 20, au seuil de confiance de 95 %, t = 2,086 et l'intervalle de confiance devient alors:

$\mu = 15 \pm (2{,}086)\ (20/) = 15 \pm 9{,}10$

à 95 chances sur 100 de se situer dans la fourchette [5,89: 24,10].

Les tests d'hypothèse

a) Position du problème

(La valeur de la moyenne trouvée sur échantillon aura souvent à être mise en relation avec une valeur a *priori* μ_o.) On peut faire des hypothèses concernant la relation entre la moyenne de la population et une telle valeur apriori. Une idée simple est à la base du teste d'hypothèses: une Hypothèse peut être rejetée mais elle ne peut jamais être acceptée, par ce que des preuve ultérieures peuvent montrer le contraire. (exemple: l'homme qui à un comportement d'homme pauvre est-il vraiment pauvre...)

Hypothèse nulle H_0

On appellera **Hypothèse nulle H_0** l'hypothèse selon laquelle la situation vraie est différente ou plus défavorable que celle qui est matérialisée par cette valeur a priori. L'hypothèse nulle doit être choisie de telle manière que son rejet permet "d'accepter" la conclusion désirée. L'**hypothèse alternative** est H_a. Par le biais d'un test d'hypothèse il s'agira d'évaluer dans quelle mesure Ho peut être rejetée.

Test unilatéral et bilatéral

On parlera de **test unilatéral** quand il s'agira de vérifier que la moyenne vraie est plus forte (test dit "à droite"), ou plus faible (test dit "à gauche") que μ_o. On aura affaire à un **test bilatéral** quand il s'agira de démontrer que la moyenne vraie est différente de μ_o.

Exemple:

Les intentions d'achat X d'un produit nouveau découlant d'une enquête par sondage auprès des utilisateurs potentiels doivent être comparées avec le seuil de rentabilité de ce produit μ_o, et il faut vérifier l'hypothèse selon laquelle ce seuil de rentabilité sera bien dépassé. L'hypothèse H_O s'énonce ici de la façon suivante: *"la situation du marché est telle que le seuil de rentabilité ne sera pas atteint*" et H_1,: "*le seuil de rentabilité sera dépassé*". Le test d'hypothèse nécessaire est alors un *test unilatéral* à droite.

Si > μ_o, ce peut être dû au fait que la vraie moyenne est réellement supérieure à μ_o. Ce peut être également dû au fait que la vraie moyenne est inférieure à μ_o mais que le hasard a fait porter le sondage sur un échantillon particulièrement favorable. Il est évident que plus (- μ_o) est grand, moins le risque de se trouver dans cette deuxième situation est fort.

b) Réalisation d'un test unilatéral à droite

1) $H_o : \mu < \mu_o$

Dans le problème posé, H_o est associée à la situation $\mu < \mu_o$. Une première façon de procéder consiste à déterminer la probabilité - dénommée **probabilité critique p.c.** - avec laquelle H_o est conforme aux résultats lus sur échantillon.

Le graphique 2 résume les termes du problème : Si la moyenne vraie était μ, la probabilité d'obtenir sur échantillon une valeur supérieure ou égale à serait donnée par la surface lue sous la courbe au-delà de la valeur [7].

[7]Dans ce test, on ne s'occupe pas des situations où la moyenne X serait inférieure à μo. Dans ces situations, la probabilité critique apparaît encore plus forte.

Figure 2: Test unilatéral à droite

Dans la mesure où le sondage est aléatoire, non exhaustif et porte sur un effectif supérieur à 30, cette probabilité est calculée à partir d'une table de la loi normale réduite[8] :

$$Z = (\ - \mu_o)/\ \sigma \text{ et p.c.} = P\ (Z \geq\ Z) \qquad (13)$$

Le fait de rejeter l'hypothèse nulle est associée à un risque égal à p.c. Plus cette probabilité critique est faible et moins il y a de risque à rejeter Ho.

2) valeur seuil X^*

X*, telle que tout résultat de sondage X supérieur à X* permette de rejeter l'hypothèse nulle avec moins de chances de se tromper.

La valeur seuil X* est obtenue à l'aide de l'expression suivante, issue de la formule [11] :

$$X^* = \mu o + Z\alpha\ \sigma \qquad (14)$$

La règle est alors la suivante:

- Si $< X^*$: acceptation de Ho
- Si $\geq X^*$: rejet de Ho

3) tests unilatéraux à gauche

Les tests unilatéraux à gauche s'effectuent de la même façon; la probabilité critique est la surface sous la courbe au-dessous de la valeur X trouvée sur échantillon. La valeur-seuil X* est calculée à partir de la relation

$$X^* = \mu o - Z\alpha\ \sigma \qquad (15)$$

Les tests bilatéraux nécessiteront l'évaluation de deux valeurs-seuil: une X* à droite de µo et une X** à gauche, par utilisation simultanée des formules (14) et (15).

4. Exemple:

Le seuil de rentabilité d'un produit industriel nouveau s'élève à 50 en moyenne par entreprise appartenant au marché potentiel Sur un échantillon de 100 entreprises, une intention d'achat moyenne de 62 a été repérée, avec un écart-type de 60.

Z_{62} = **(62 - 50) / (60/ 1100) = 2**

p.c. = P(Z > Zx) = p(Z > 2) = 0,023

Avec un seuil de risque a = 5 %, l'hypothèse nulle est rejetée. En **fait Ho peut être *rejetée* dès que l'on *trouve, sur* échantillon une valeur au moins égale à:**

= 50 + (1,64) . (60/ $100^{1/2}$) = 59,84

c) Les risques associés au test d'hypothèse

La procédure qui vient d'être exposée ne s'intéresse qu'à une seule catégorie de risque, celui de rejeter Ho alors qu'elle est vraie. C'est le risque α ,risque de première espèce ou encore de risque de type 1. Il sera souvent nécessaire de prendre également en considération le risque d'accepter à tort Ho : c'est le risque β, risque de seconde espèce ou encore de type II.

Le tableau 3 reproduit les résultats possibles d'un test d'hypothèse :

	DÉCISION	
SITUATION REELLE	**Ho n'est pas rejetée**	**Ho est rejetée**
Ho est vraie	**Décision correcte Seuil de confiance:** *Proba* = 1 - α	**Décision incorrecte Erreur de type I** *Proba* = α
Ho n'est pas vraie	**Décision incorrecte Erreur de type II** *Proba* =β	**Décision correcte Puissance du test:** *Proba = 1 - β*

Tableau 1.3.: Résultats d'un test hypothèse

Il est bien évident que, pour une taille d'échantillon donnée, le risque α et le risque β évoluent de façon opposée. Réduire le risque a demande de choisir une valeur-seuil X* plus forte, mais ceci s'accompagne d'une augmentation du risque β, puisqu'il y aura plus de chances d'accepter à tort l'hypothèse nulle [9].

Exemple:

Avec les données de l'exemple précédent, on a vu que le risque α était limité à 5 % si l'on choisissait une valeur-seuil de 59,84.

Supposons que la véritable valeur des ventes moyennes par entreprise soit 62. **Avec une vraie moyenne de 62 et un écart type des moyennes d'échantillon de** 6 ($60/100^{1/2}$ = 6), il y **a une probabilité de** 35,94 % **de sélectionner un échantillon dont la moyenne observée sera inférieure ou égale à** 59,84. **En effet:**

Z = (59,84 - 62)/6 = - 0,36 **et P(Z ≤** 0,36) = 0,3594

Il y a donc ici une probabilité de 35,94 % **d'accepter à tort l'hypothèse nulle. C'est la valeur du risque β.**

[8]Sinon, il faudra utiliser les règles indiquées pour les cas particuliers dans le paragraphe précédent.

[9]I1 est possible de déterminer des tailles d'échantillons nécessaires pour permettre à la fois au risque de type I et au risque de type 11 d'être inférieurs à une valeur prédéterminée. Cf. P. E. Green, D. S. Tull et G. Albaum, *Research for rnarketing decisions,* Prentice-Hall, 1988, p. *345.*

Si l'entreprise veut se prémunir plus fortement contre le rejet à tort de Ho avec un risque a de 2,5 % seulement, elle sera amenée à choisir une valeur seuil plus forte, égale à:
$X^* = 50 + 1,96(6) = 61,76$
La sélection d'une telle valeur-seuil augmente le risque β. Dans l'hypothèse où la moyenne vraie est 62:
$Z = (61,76 - 62)/6 = - 0,04$ et $P(Z \leq -0,04) = 0,484$

Le risque β est ici de 48,4 %.

Analyse des données qualitatives

Les intervalles de confiance et tests d'hypothèse

Dans le cas de variables qualitatives, la problématique de la prévision des valeurs réelles se pose dans les mêmes termes que pour les variables quantitatives, mais maintenant, il s'agit de fréquences d'apparition de modalités et non plus de moyennes.

A - La population totale est de taille N; la valeur vraie de la fréquence de la modalité analysée est π. Sur l'échantillon de taille n, une proportion p a été trouvée.

On montre [10] que les proportions lues sur les échantillons suivent une loi normale de moyenne π et d'écart type $\sigma_p = [p(1 - p)/n]^{1/2}$.

Au seuil de risque α, l'intervalle de confiance est obtenu par l'expression :

$$\pi = p \pm z_{\alpha/2}.[p(1 - p)/n]^{1/2} \qquad (15)$$

Généralement, on prendra, pour calculer l'écart type des proportions, p = 50 %, qui correspond au cas le plus défavorable et non la proportion observée.

Exemple:
Dans le cadre d'une étude de notoriété, 25 % des personnes interrogées ont déclaré connaître la marque M. Un échantillon aléatoire non exhaustif de 1000 individus a été utilisé. L'écart type des proportions est alors:
$[0,5(0,5)/1000]^{1/2} = 0,0158.$
Il y a 95 chances sur 100 que le véritable taux de notoriété se situe dans la fourchette:
$\pi = 0,25 \pm 1,96.(0,0158)$
Il doit être ainsi compris entre 0,25 - 0,03 = 0,22 et 0,25 + 0,03 = 0,28, c'est-à-dire entre 22 % et 28 %.

B - En ce qui concerne les test d'hypothèse, les mêmes procédures que pour les variables quantitatives sont employées. C'est la formule (15) qui servira désormais dans le calcul des probabilités critiques et des valeurs-seuil.

Exemple:
Le taux de notoriété de la marque M dont il était question dans l'exemple précédent a été mesuré à la suite d'une campagne publicitaire. Le taux de notoriété précédemment connu s'élevait à 21%. Peut-on en conclure que la publicité a fait augmenter de façon significative la connaissance de la marque ?

L'hypothèse nulle correspond ici au cas où la publicité n'a eu aucun effet sur la notoriété de la marque et donc que la proportion vraie est toujours au niveau ancien de 21%. Le rejet éventuel de l'hypothèse nulle demande de calculer la probabilité critique p.c., définie ici comme la probabilité d'obtenir une proportion observée sur échantillon au moins égale à 25 % dans une population où la proportion vraie est 21%.
La proportion observée p correspond, en valeur centrée réduite
$Zp = (0,25 - 0,21)/0,0158 = 2,53$ probabilité critique est donc: $P(Z > Zp) = 0,57$ %

Les tests de conformité avec une distribution théorique

Les résultats du dépouillement d'une question qualitative se présentent comme une distribution de fréquences d'apparition des différentes modalités de la variable concernée.

Cette distribution peut être comparée à une distribution a priori, dite distribution théorique. Comme dans les tests d'hypothèses vus plus haut, deux hypothèses sont alors testées:

- Ho la distribution observée n'est pas significativement différente de la distribution théorique.
- Hl: la distribution observée est significativement différente de la distribution théorique.

a) Application du test du Khi-Deux

La loi du Khi-Deux (χ^2) donne la répartition des écarts entre les fréquences absolues théoriques et les fréquences absolues observées, sous hypothèse nulle.

On mesure le χ^2 par :

$$\chi^2 = \Sigma\, [Nj - \theta j]^2 / \theta j \qquad (16)$$

où **Nj** = fréquence absolue observée pour la modalité j;
θ j = fréquence absolue théorique pour la modalité j.

Cette valeur calculée du χ^2 est comparée avec la valeur lue sur la table du χ^2, pour **m - 1** degrés de liberté lorsque la variable qualitative comporte m modalités [11], et pour un seuil de confiance donné 1 - α. Si la valeur calculée du χ^2 est supérieure à la valeur de la table, Ho peut être rejetée avec un risque inférieur à α.

Le tableau 1.4. reproduit une application du test du χ^2 pour le traitement des résultats d'une étude sur les clients d'une ligne aérienne. Il s'agit ici de vérifier si l'échantillon interrogé respecte bien les proportions

[10] Pour la démonstration, voir V. Giard, *Statistique appliquée à la gestion,* Economica, 1985, p.177 et ss.

[11] Lorsque certains paramètres (moyenne, écart type) de la distribution a priori des fréquences sont déterminés à l'aide des résultats des observations, on enlève un nombre de degrés supplémentaires égal au nombre de paramètres déduits des observations.

connues des passagers eu égard à leur qualité d'abonné ou non. Le χ^2 calculé apparaissant plus faible que le χ^2 lu sur la table (5,99 pour 2 degrés de liberté au seuil de 5 %), les différences constatées ne sont pas significatives.

Tableau 1.4.: Application du test du Khi-Deux

Modalité	Effectif observé Nj	Effectif théorique Θj	N_j-$Θ_j$;	$(N_j-Θ_j)^2$	$(N_j-Θ_j)^2$ /Θ
Abonné année	80	90	- 10	100	1,11
Abonné mois	50	45	5	25	0 56
Autre	100	95	5	25	0 26
Total	230	230			χ^2= 1,93

Nombre de degrés de liberté: 2
Valeur du Khi-Deux au seuil de 5 %: 5,99

b) Application du test de Kolmogorov-Smirnov

La qualité de l'ajustement d'une fonction de répartition observée à une fonction de répartition a priori peut également être évaluée à l'aide du test de Kolmogorov-Smirnov.

On aura recours à un test chaque fois que les modalités de la variable qualitative considérée sont ordonnables, mais aussi lorsque les effectifs des différentes classes sont trop faibles pour autoriser l'utilisation du test du χ^2 [12].

Le test demande de calculer des fréquence relatives observées cumulées $F_O(j)$ et des fréquences relatives cumulées théoriques $F_\theta(j)$: Fo(j) et $F_\theta(j)$ représentent respectivement les pourcentages des effectifs observés et théoriques enregistrés jusqu'à la modalité j. Pour chaque modalité la valeur $|Fo(j) - F_\theta(j)|$ est calculée. Un indicateur D est alors établi, tel que :

$$D = Max_j |Fo(j) - F\theta(j)| \qquad (17)$$

Cette valeur est comparée à celle lue sur une table du D de Kolmogorov-Smirnov pour un seuil de confiance donné. A un seuil de risque de 5%, et pour des effectifs totaux supérieurs à 35, D est approximativement égal à 1,36/

Le tableau 1.5. donne une application de ce test à l'étude sur les clients d'une ligne aérienne. Le D calculé est plus faible que le D de la table au seuil de 5% : les différences ne sont pas significatives, comme on l'avait déjà constaté avec le test du χ^2.

Tableau 1.5.: Application du test de Kolmogorov-Smirnov

Modalité	Fréquence relative observée	Cumulé Fo(j)	Fréquence relative théorique	Cumulé Fθ(j)	|Fo(J)-Fθ(J)|
Abonné année	21,74 %	56,52 %	19,57 %	58,70 %	2,17 %
Abonné mois	43,48 %	100,00 %	41,30 %	100,00 %	0,00 %
Total	100,00 %		100,00 %		

Valeur calculée de D = 0,0435
Valeur de D au seuil de 5 %: 0,089

Analyses bivariées[13]

Introduction

Les traitements bivariés ont pour objet de mettre en évidence les relations éventuelles qui existent entre deux variables analysées simultanément.

Dans la plupart des cas l'analyste cherchera à expliquer une des deux variables - dite variable à expliquer (Y) - à l'aide de l'autre variable - dite variable explicative (X). Expliquer une variable à l'aide d'une autre revient à repérer dans quelle mesure les différentes valeurs que peut prendre la variable explicative ont une conséquence sur les valeurs prises par la variable à expliquer.

Exemple:
Le fait de changer de conditionnement a-t-il un effet sur le niveau des ventes d'un produit donné ? Le conditionnement joue ici le rôle de variable explicative et le niveau des ventes. de variable à expliquer.
Le fait de posséder un four à micro-ondes dépend-il de l'âge ou de la taille de la famille? La possession ou non du four à micro-ondes est la variable à expliquer; l'âge ou la taille de la famille sont des variables explicatives.

Comme dans le cas des traitements univariés, le mode d'analyse utilisable va dépendre de la nature des variables étudiées: quantitatives, ordinales ou nominales. Ces analyses seront à nouveau présentées ici dans le cadre des études marketing par questionnaires: il s'agira donc du traitement des tri-croisés.

Traitement des tris croises et nature des donnés

Présentation des tris croisés

Dans la mesure où une question peut relever fondamentalement de trois niveaux de mesure différents, on comptera neuf types de croisements possibles entre les questions Qi et Qj Les plus fréquentes sont présentés ci-dessous.

2 - **Qi quantitatif x Qj quantitatif:** étude des relations entre deux séries de n chiffres s'il y a n questionnaires. Ces deux séries de chiffres n'apparaissent généralement pas explicitement. Leurs relations sont matérialisées par différents indicateurs.

[12] L'application du teste du χ^2 demande que les effectifs associés à chaque classe ne soient pas inférieurs à 5 pour plus de 20 % des modalités.

[13] Cf. J-P. Vedrine "Le traitement des données en marketing ", Ed. d'Organisation 1991, p.40-54.

Exemple:

Dans le questionnaire CAMIP, étude des relations entre le proportion des achats par catalogue (question 1) et le revenu de la personne (question 11). L'appartenance à une catégorie de revenu plus élevé entraîne-t-elle une plus grande proportion des achats par catalogue ?

2 - **Qj nominal x Qj nominal** : croisement le plus fréquent qui se traduit par la création d'un **tableau de contingence** où, en ligne, figurant les modalités de la variable à expliquer et en colonne, celles de la variable explicative. Lorsque le tableau de contingence est traduit en pourcentages de colonne, pour chaque modalité de la variable explicative apparaîtra une distribution de fréquences des modalités de la variable à expliquer.

Exemple:

Croisement entre la question 22 sur la préférence pour un type de magasin ou une modalité d'achat et la question 26: le fait de préférer d'acheter par catalogue ou par une autre modalité, dépende-t-elle de la situation civile du répondant?

3 - **Qi ordinal X Qj ordinal** : mise en correspondance de deux classements au niveau de chaque individu interrogé ou sur l'ensemble de l'échantillon si une procédure d'agrégation des rangs a été utilisé.

Exemple:

Croisement entre l'ordre de préférence pour les différentes catégories d'équipement de bureau (micro-ordinateurs, imprimantes, scanners..) exprimé par chaque répondant et un classement a *priori* correspondant à l'importance accordé à chaque catégorie d'équipement dans le catalogue (exprimé par exemple en nombre de pages).

4 - Qi **quantitatif X Qj nominal** : correspond à un tri-à-plat de la variable quantitative pour chacune des modalités de la variable nominale qui joue le rôle de variable explicative.

Exemple:

Croisement entre une question ouverte concernant le nombre d'objets achetés par catalogue et la question 26. La situation civile influence-t-elle le nombre d'objets achetés par catalogue?

5 - **Qi ordinal X Qj nominal** : repérage des rangs donnés à la question Qi pour différentes classes d'une variable Qj nominale explicative.

Exemple:

Croisement entre un classement de préférence et la question 26. Le fait de relever d'un statut familial donné entraîne-t-il des préférences pour une catégorie de produit bureautiques ?

Analyse des tris croisés

Le tableau 2.1. donne les principaux tests utilisables dans l'analyse des tris croisés. Seuls les tests correspondant aux croisements les plus fréquents y sont indiqués.

Tableau 2.1 .: Tests des tris croisés selon la nature des variables

Qj \ Qi	NOMINAL	ORDINAL	QUANTITATIF
NOMINAL	Test du Khi-Deux Test de Kolmogorov-Smirnov Test de comparaison de fréquences	Test de Kruskal-Wallis Test de Wilcoxon du rang et du signe Test de Wilcoxon de la somme des rangs	Test F (ANOVA) Test de comparaison de moyennes
ORDINAL		Corrélation des rangs de Spearman Test de Kendall	
QUANTITATIF			Coefficient de corrélation de Pearson Test de comparaison de moyennes

En colonne figurant les variables à expliquer et en ligne les variables explicatives.

Traitement des variables quantitatives

Le croisement de deux variables quantitatives peut être effectué dans des circonstances très variées :

- comparaison des résultats obtenus pour une variable sur deux ou plusieurs populations indépendantes;
- comparaison des résultats obtenus pour une variable sur deux échantillons appariés;
- comparaison des résultats obtenus pour deux variables quantitatives différentes pour la même population.

Une variable et deux populations indépendantes

a) Le test de comparaison de moyennes

En ce qui concerne le cas où une seule variable est étudiée, le **test de comparaison de moyennes** est la statistique classique lorsque deux populations sont concernées.

L'analyste dispose des données suivantes :

- deux populations A et B respectivement d'effectifs Na et Nb;
- la moyenne de la variable étudiée est a dans la population A et b dans la population B.
- la variance de la variable analysée est s_a^2 pour A et s_b^2 pour B.

Dans la mesure où l'on estime que Xa et Xb suivent une loi normale, respectivement de moyenne μ_a et μ_b et de variance σ_a^2 et σ_b^2, on montre que la différence

$D = Xa - Xb$

suit également une loi normale de moyenne $\mu_a - \mu_b$ et de variance:

$$\sigma_D^2 = [\sigma_a^2 /Na + \sigma_b^2 /Nb] \approx [s_a^2 /Na + s_b^2 /Nb]$$

L'intervalle de confiance de la différence de moyenne au risque a est donné par :

$$\mu_a-\mu_b = a-b \pm z_{\alpha/2}[s_a^2/Na+s_b^2/Nb]^{1/2} \qquad (18)$$

L'hypothèse nulle Ho correspond au cas où la différence D = µa - µb de moyennes est nulle. Sous Ho, la variable réduite devient :

$$Z\ (a-b)/[s_a^2/Na+s_b^2/Nb]^{1/2} \qquad (19)$$

La valeur z ainsi calculée doit être comparée avec la valeur lue dans la table normale réduite pour le seuil de confiance désiré et compte tenu du caractère unilatéral ou bilatéral du t(est. Pour un test bilatéral par exemple, Ho sera rejeté au seuil de risque de 5 % si **|z| > 1,96.**

Exemple:

Supposons que dans le cadre de l'étude CAMIP, on a ajouté une question sur la nombre de disquettes d'ordinateur achetées par an qu'on croise avec la question 27(le fait d'être homme ou femme) a fait apparaître les résultats suivants:

- les hommes (A): Na = **155; nombre moyen de disquettes = 10, avec Sa^2 = 64;**

- les femmes (B): Nb = **75; Xb = 3, Sb = 25.**

La variance des différences de moyennes est donnée par:

64/155 + 25/75 = 0,74.

L'écart type de D est alors = 0,86. L'hypothèse nulle pour laquelle il n'existe pas de différence dans les quantités de disquettes achetées par les homme et par les femmes peut être rejetée, puisque z = (10 - 3)/0,86 est supérieur à 1 96

b) Autres tests

1) test **de Student**

- Pour des petits échantillons (Na et Nb < 30), on utilisera le test **de Student.** Dans la mesure où la variance des Xa et des Xb est estimée la variance de la distribution des différences de moyennes est approchée par l'expression :

$$\sigma_D^2 = [(Na-1)s_a^2 + (Nb-1)\ s_b^2]/(Na + Nb - 2)(1/Na + 1/Nb) \qquad (20)$$

La différence D suit alors une loi de Student à (Na + Nb - 2) degrés de liberté.

2) (test en **F et Kruskal - Wallis**)

- Quand plus de deux populations sont concernés, on aura recours au **test** F. Si l'hypothèse de normalité évoquée plus haut n'est pas respecté on pourra employer le **test de Kruskal-Wallis.**

Une variable et deux échantillons appariées

Dans le cas d'échantillons appariés, à chaque individu d'un premier groupe est associé un individu du second groupe (le groupe-témoin) offrant les mêmes caractéristiques.

Pour chaque couple i de deux individus appariés, une différence $D_i = X_{ai} - X_{bi}$ est calculée. Sur l'ensemble n des couples étudiés, la différence moyenne est donnée par = $\Sigma D_i/n$ et la variance des différences est alors

$$s_D^2 = \Sigma(Di -)^2/(n - 1)$$

On montre que D est distribué selon une loi normale de moyenne µa - µb et de variance

$$s^2 = [\Sigma(Di -)^2/(n - 1)]/n. \qquad (21)$$

Deux variables et la même population

Les relations entre deux variables quantitatives sur la même population sont généralement analysées à l'aide du **coefficient de corrélation** de Pearson qui sera étudié au cours du chapitre portant sur la régression linéaire

Traitement des variables nominales

Analyse des tableaux de contingence

a) Test du caractère significatif de la relation entre les variables

Dans le chapitre précédent on a vu une application du test du Khi-Deux pour l'évaluation de la qualité de l'ajustement d'une distribution de fréquences observées à une distribution théorique.

De façon plus générale, ce test est employé pour analyser les tableaux de contingence et repérer le caractère statistiquement significatif de l'association entre deux variables nominales.

La statistique du χ^2 est donnée, en ce qui concerne les tableaux de contingence, par la formule suivante :

$$\chi^2 = \Sigma\Sigma(N_{ij} - \Theta_{ij})^2 /\Theta_{ij} \qquad (22)$$

où $\mathbf{N_{ij}}$ = nombre d'observations dans la case ij;

$\mathbf{\Theta ij}$ = nombre théorique associé à la case ij = (total de la ligne i)·(Total de la colonne j)/Nombre total d'observations.

Pour un tableau comportant **C** colonnes et **L** lignes, la valeur ainsi calculée est comparée à la valeur critique χ^2 lue sur la table du Khi-Deux pour un seuil de confiance 1-α, et pour un nombre de degrés de liberté égal à (C - 1)(L - 1) [14].

La table du Khi-Deux donne la distribution de probabilité des valeurs de χ^2 obtenues dans un tableau lorsque l'hypothèse nulle est vraie, c'est-à-dire dans le cas d'indépendance entre les deux variables étudiées. Par exemple, au seuil de 5 % et pour deux degrés de liberté, le χ^2 lu sur la table vaut 5,99 : ceci veut dire que sous Ho il n'y a que 5 % de tableaux à deux degrés de liberté pour lesquels on pourrait calculer un x2 supérieur ou égal à 5,99. Si le χ^2 calculé est plus fort, il y a donc moins de cinq chances sur cent de se tromper en rejetant Ho.

[14]Lorsque le nombre de degrés de liberté est supérieur à 30, on peut approximer la distribution de χ^2 par une loi normale dont la moyenne est le nombre de degrés de liberté et la variance deux fois ce nombre.

Exemple:

Le tableau 2.2.a donné les résultats d'un éventuel résultat du croisement des questions 3 (le fait d'avoir commandé) et 27 (être homme où femme). Le tableau 2.2. fournit les valeurs théoriques": par exemple, pour la case 1, la valeur 53,91 = (80)(155)/ 230. Les différences (Nij-Θij;) apparaissent sur le tableau 2.2.c, et le χ^2 associé à chaque case, sur le tableau 2.2.d. Au total, le χ^2 calculé s'élève à 28,64: il dépasse le %2 critique (5,99). La qualité d'abonné est donc liée significativement au motif du voyage.

Tableau 2.2.: Application du Khi-Deux à un tableau de contingence

	Homme	Femme	Total
Jamais	70	10	80
l'année dernière	35	15	50
cette année	50	50	100
Total	155	75	230

a) Croisement question Q3 et Q27: valeurs observées

Jamais	53,91	26,09	80
l'année dernière	33,70	16,30	50
cette année	67,39	32,61	100
Total	155,00	75,00	230

b) Valeurs théoriques

Jamais	16,09	- 16,09	0,00
l'année dernière	1,30	- 1,30	0,00
cette année	-1 7,39	1 7,39	0,00
Total	0,00	0,00	0 00

Différences entre valeurs observées et valeurs théoriques

	Homme	Femme	Total
Jamais	4,80	9,92	14,72
l'année dernière	0,05	010	015
cette année	4,49	9,28	13,76
Total	9,34	19,30	28,64

d) Croisement questions Q3 et Q27: calcul du Khi-Deux
Khi-Deux calculé: 28,64 **nombre ddl:** 2
Khi-Deux critique: 5,99 **risque:** 5 %

Le test est fourni couramment par les logiciels; il souffre cependant de certaines limitations:
- l'effectif sur lequel porte le tableau doit être suffisamment important on ne doit pas trouver plus de 20 % de cases avec un effectif inférieur à 5 [15];
- le χ^2 est calculé à partir des valeurs absolues; il est donc très sensible à la taille des effectifs considérés;

le χ^2 permet de repérer le caractère significatif de la relation entre les deux variables, mais pas l'intensité de cette relation.

b) Test du degré d'association entre les variables

Plusieurs coefficients peuvent repérer le degré d'association entre les deux variables étudiées. Pour les tableaux de contingence de taille 2 x 2, on peut déduire du χ^2, un coefficient **d'association Φ**, tel que :

$$\Phi = [\chi^2/n]^{1/2} \qquad (23)$$

Φ présente l'avantage d'être indépendant de la taille de l'échantillon, et de varier entre 0 et 1 [16].
Pour les tableaux plus grands, c'est un coefficient de contingence C **qui sera utilisé** :

$$C = [\chi^2/(\chi^2 + n)]^{1/2} \qquad (24)$$

Plus C est élevé et plus forte est l'association entre les deux variables concernées. Le minimum de ce coefficient est 0 (indépendance totale des variables avec $\chi^2 = 0$). Par contre, le maximum ne peut jamais s'élever jusqu'à 1. Dans le cas d'un tableau 2 x 2, on montre, par exemple, que le maximum est de = 0,707.

Test de comparaison de fréquences

Lorsque l'on considère deux modalités de la variable explicative, les effectifs associés à une modalité donnée de la variable à expliquer peuvent être traduits en fréquence relative. Il s'agit de vérifier dans quelle mesure la différence de fréquences observées est significative.
Soient Pa la fréquence relative associée à la modalité A (effectif Na) de la variable explicative et Pb celle qui est associée à la modalité B (effectif Nb)
La variance des différences de proportions est donnée par $\sigma_D^2 = [\pi_a(1-\pi_a)/Na + \pi_b(1-\pi_b)/Nb]$ où π_a et π_b sont les proportions réelles. Dans la mesure où les véritables fréquences π_a et π_b ne sont pas connues, on utilisera comme estimateur de σ_D^2 l'expression Pc(1-Pc)[1/Na + 1/Nb] où Pc est la fréquence moyenne observée sur l'ensemble des deux groupes avec Pc = (Na Pa + Nb Pb)/(Na + Nb).
Les intervalles de confiance et le test des différences de proportions s'obtiennent dans les mêmes conditions qu'en ce qui concerne les moyennes.

Exemples:

[15] Si cette condition n'est pas respectée, on regroupera des modalités afin d'aboutir à des cases suffisamment remplies; on pourra également utiliser le test de Kolmogorov-Smirnov, s'il n'y a que deux colonnes dans le tableau.

[16] $\Phi^2 = [\chi^2/n]$ est quelquefois appelé le **lien**; d'autre part, à partir de Φ^2 on peut calculer le **coefficient de Tschuprov** $T = \Phi^2/[(C-1)(L-1)]$.

Avec **le croisement des questions 3 et 27 du questionnaire CAMIP, il apparaît que sur 155 hommes, 70 n'ont jamais commandé, soit 45,16 % et seulement 10 sur 75 femmes , soit 13,33 %.**

La fréquence moyenne pondérée est donc p = [155(0,4516) + 75(0,1333)]/230= 0,3478.

La variance des différences de proportions est alors σ_D^2 **= [0,3478(1 - 0,3478)] [1/155 + 1/75] = 0,00449 et**

σ_D **= 0,067**

Dans ces conditions l'intervalle de confiance des différences de proportions au seuil de risque de 5 % est donné par:

$\pi_a - \pi_b$ = 0 4516 - 0,1333 ± 1 96 (0,067) = [0,4496 0,187]

La valeur 0 ne figurant pas dans cet intervalle de confiance, la différence de proportions apparaît significative.

Traitement des donnes ordinales

Relations entre deux variables ordinales

a) Test de corrélation des rangs de Spearman

Ce test permet de repérer le caractère significatif de la relation qui existe entre deux classements. Il est également utilisé pour montrer la relation éventuelle qui existe entre deux variables quantitatives.

Soit n le nombre d'items à classer; Xi est le rang de l'item i dans un premier classement et Yi son rang dans un second. Dj = |Xi - Yi| est la différence de rangs observés entre les deux classements. Le coefficient de corrélation des rangs de Spearman a pour expression :

$$R_S = 1 - 6 \Sigma D_i^2/[n(n^2 - 1)] \qquad (25)$$

Plus R_S est proche de 1 et plus les deux classements sont proches; à la limite, ils sont complètement identiques si Rs = 1. Au contraire, plus R_S est proche de 0 et plus les deux classements sont indépendants.

La signification statistique de Rs obtenue peut être testée à partir de la relation:

$$t = Rs[n - 2]^{1/2}/[1 - R_S^2]^{1/2} \qquad (26)$$

qui suit une loi de Student à n - *2* degrés de liberté [17]

Exemple:

Croisement entre l'ordre de préférence pour les différentes catégories d'équipement de bureau (micro-ordinateurs, imprimantes, scanners..) exprimé par chaque répondant et un classement a *priori* correspondant à l'importance accordé à chaque catégorie d'équipement dans le catalogue (exprimé par exemple en nombre de pages).

Le tableau 2.3. montre un exemple de calcul du coefficient de Spearman avec le traitement d'une question concernant le classement d'intérêt d'achat pour les catégories de produits (A - **micro-ordinateurs, B - imprimantes, C - scanners et D fax)** par l'individu 124**.** On constate une corrélation R. = 1 - (6)(38)/(5)(25-1)= - 0,9 entre les choix de la direction et ceux de l'individu n° 128. Cette corrélation est significative puisque t = (-0,9)( )/ (= - 3,576. Cette valeur est supérieure au t de Student pour 3 degrés de liberté et un seuil de confiance de 95 % (3,182).

L'application de cette procédure à l'ensemble de l'échantillon permettrait de repérer quels sont les individus qui manifestent des préférences conformes à l'importance donnée dans le catalogue à ces catégories .

Tableau 2.3.: Application du coefficient de Spearman

Propositions	**Classement individu n° 124**	**Classement CAMIP**	**|Dj|**	**|Dj|^2**
A	2	5	3	9
C	3	3	0	0
D	5	1	4	16
E	4	2	2	4
				$\Sigma\|D_i\|^2$ =38

R_S= -0,9

t = -3,576

b) Tau de Kendall

Le tau de Kendall est un autre indicateur du caractère éventuellement significatif de la relation qui existe entre deux classements.

Les n items sont tout d'abord rangés dans l'ordre conforme au premier classement. Puis pour tout couple d'items dont l'ordre dans le premier classement est concordant avec celui du second on attribue un score de + 1; dans le cas de discordance on attribue un score de - 1. S est la somme de ces scores sur l'ensemble des n(n - 1)/2 comparaisons possibles.

Le tau de Kendall est alors :

$$\tau = 2 S/(n^2 - n) \qquad (27)$$

Le test de ce coefficient s'effectue à l'aide de tables. Pour n > 10 une approximation par la loi normale est envisageable [18].

Exemple:

Le tableau 2.4. donne une application du τ **de Kendall pour l'analyse des résultats de la question**

[17]Des tables peuvent également être utilisées pour tester la signification du coefficient t,. Cf. celle fournie par S. Siegel dans son ouvrage, Nonparapetric statistics for the behavioral sciences, Mc Graw-Hill, 1956.

[18]Cf. M G Kendall, Rank correlation methods, Griffin, London, 1970.

précédente pour l'étude CAMIP. On constate un t de - 0,8 entre les choix de la direction et ceux de l'individu n° 124. En effet sur un ensemble de 10 comparaisons possibles, une seule est concordante et neuf sont discordantes.

Tableau 2.4.: Application du tau de Kendall

Propositions	Classement Individu n° 128	Classement Catalogue CAMIP
B	1	4
A	2	5
C	3	3
E	4	2
D	5	

	Classe-ments concor-dants	Classements discordants	Score
Paires Bj	BA	BC, BE, BD	- 2
Paires A		AC, AD, AE	- 3
Paires C,		CD, CE	- 2
Paire Ej		ED	-1
			Total S = - 8

Nb. combinaisons = n(n - 1)/2= 10
Tau de Kendall = - 0.8

ELEMENTS D'ANALYSE DES DONNEES D'ENQUETE SUR ECHANTILLON

SOMMAIRE:

M. Calciu - Université Lille I