STATISTIQUE

Gilles Mauffrey Bureau 5-W2 Tel 7261

OBJECTIFS

L'objectif de ce cours est de permettre aux étudiants d'HEC en Formation Fondamentale de comprendre les principes de base des méthodes statistiques utilisées dans les entreprises et d'en appréhender les limites.

La **Statistique** est une discipline regroupant des techniques et méthodes permettant de

- collecter des données
- les organiser pour décrire et résumer
- expliquer et modéliser leur évolution et leur variabilité

Les données utiles une entreprise étant de plus en plus nombreuses, il est nécessaire de les recueillir par des méthodes "fiables", pour pouvoir en tirer des conclusions sur l'ensemble de la population :

- Sondages
- Inférence statistique : Estimation et Tests

Les traitements statistiques interviennent sous deux formes :

- La *structuration des données* lorsqu'il s'agit de décrire, en les visualisant ou en les résumant, une masse importante de données, ceci afin de mieux comprendre un phénomène étudié; outils de première approche d'un domaine ou outils de synthèse, ces méthodes sont fréquemment utilisées en marketing et en sciences humaines.
- La *modélisation* lorsqu'il s'agit de construire un modèle mathématique d'une réalité observée; on rencontre fréquemment une telle démarche en Finance, en Production, en Comptabilité ou en Economie.

L'analyse factorielle, l'analyse des proximités, des préférences, la typologie correspondent plus au premier type de méthodes. La régression multiple, la régression logistique, la segmentation, l'analyse discriminante, l'analyse des séries chronologiques correspondent au second type d'approche statistique. Ces différents thèmes sont enseignés dans la formation fondamentale et dans des électifs plus spécialisés.

CONTENU DU COURS

Les thèmes suivants seront traités lors de ce cours

Statistiques descriptive

- Concepts de base : population, unités statistiques, variables statistiques
- Etude d'une variable : Résumés statistiques (tendance centrale, dispersion), représentations graphiques (histogrammes, boites à moustaches, courbe Q-Q)
- Etude de deux variables : liaison(étude graphique), tableaux croisés, indicateurs de liaison entre deux variables (covariance, corrélation)

Inférence statistique

- Sondage, échantillon
- Estimation d'un paramètre : estimation ponctuelle, précision, estimation par intervalle, taille d'un échantillon
- Test statistiques : les hypothèses et les erreurs, tests de comparaison bilatéral et unilatéral.

La régression linéaire

- Notion de modèle statistique
- Hypothèses du modèle de la régression linéaire
- Estimation des coefficients
- Tests du modèle (Fisher global, Student et Fisher partiel)
- Construction et validation d'un modèle

METHODES PEDAGOGIQUES

La présentation des différentes méthodes s'effectue à deux niveaux :

Théorique
Il est important de connaître le modèle mathématique formant l'hypothèse de travail. La diversité des origines des étudiants HEC nous impose de limiter au minimum, et donc à l'essentiel, l'étude des bases mathématiques des modèles. Précisons cependant que cette restriction ne nous paraît pas être un handicap à l'utilisation des méthodes quantitatives en gestion. L'objectif du cours étant plus de permettre à de futurs gestionnaires de dialoguer avec des spécialistes que de former des experts.

Pratique
Chaque méthode est illustrée par des exercices préparés par les étudiants, nous utiliserons le logiciel **SPSS** (Statistical Package for Social Sciences), disponible sur le Campus. Deux séances de travaux pratiques seront consacrées à la pratique de ce logiciel. Un cas final sera à remettre par groupe de 5 étudiants au maximum.
Il est recommandé aux étudiants d'installer ce logiciel sur leur ordinateur (se renseigner auprès des Moyens Informatiques du Campus).

CONTROLE DES CONNAISSANCES

Il est organisé de la manière suivante :

- Un projet informatique SPSS à préparer par groupe de 5 étudiants au plus (Easton agency 30%)
- Un test final individuel (70%)

Il est nécessaire d'avoir une moyenne entre la note au projet et la note au test individuel au moins égale à 10/20 pour obtenir la validation du cours.

PROJET STATISTIQUE

Pour le projet SPSS à remettre, il est impératif de rédiger un rapport professionnel, *tant au niveau de la forme que du fond*. Votre travail doit être soigné et approfondi. Un des objectifs du cours est l'apprentissage du logiciel SPSS. Il est donc obligatoire de travailler le cas avec ce logiciel.

Site WEB

Les documents du cours et les fichiers de données sont disponibles sur le site :

www.hec.fr/mauffrey

à la rubrique Statistique.

Table des matières

1. STATISTIQUES DESCRIPTIVES

Nous présenterons ici le vocabulaire de la statistique et les éléments de base de la statistique descriptive à une et deux variables.

1.1.Vocabulaire de la statistique

Population

La population P est l'ensemble des éléments (objets, personnes ….) satisfaisant à une définition commune auxquels on s'intéresse au cours d'une étude.

Chaque élément de la population est appelé unité statistique ou individu.

On notera N la taille de cette population (cette taille n'est pas toujours connue avec exactitude)

Exemples :

1 – Ensemble des Français se connectant au moins une heure par jour à Internet.

2 – Ensemble des comptes clients d'une entreprise

3 – Ensemble des consommateurs achetant des produits frais en hypermarché.

Variables

Une variable statistique X est une application qui à chaque individu ou unité statistique associe une valeur prise dans un ensemble E . Cette valeur peut être numérique ou non.

Suivant la nature de l'ensemble E , on distingue trois types de variables statistiques :

- Les variables quantitatives associées à une caractéristique mesurable de la population, dans ce cas l'ensemble E est un sous ensemble de l'ensemble des nombre réels, par exemple l'âge, le montant d'une facture, le temps de connexion etc…
- Les variables qualitatives qui permettent d'organiser la population en classe, par exemple la profession, le fait d'acheter sur internet, la marque du produit acheté, la satisfaction du consommateur, les tranches d'âge etc… On fait parfois la distinction entre les variables qualitatives nominales où les classes sont sans hiérarchie (CSP, département,…) et les variables qualitatives ordinales pour les quelles les classes adjacentes peuvent être regroupées (tranches d'âge, degré de satisfaction..).

La valeur prise par la variable X pour l'individu i sera notée x_i .

Paramètre

Un paramètre θ est une valeur numérique associée à une population P et une variable X . La valeur de ce paramètre est calculée à partir des N valeurs prises par la variable X :

$$\theta = f(x_1, x_2, \ldots, x_N)$$

Pour connaître la valeur d'un paramètre, il faut donc connaitre chacune des valeurs prises par la variable.

Exemples :

- Temps moyen passé sur les sites de recherche

- Pourcentage d'internautes faisant des achats sur Internet
- Moyenne et écart-type des comptes clients
- Coefficients de corrélation entre deux variables
- Coefficient d'une variable dans une équation de régression….

Remarque : Dans ces deux derniers cas la variable x est en fait un couple ou un n-uple de variables.

1.2.Collecte données – Tableau statistique

Les données peuvent être internes à l'entreprise ou externes. Il est quelque fois possible d'obtenir les informations sur l'ensemble de la population à partir d'une **base de données**, par exemple.

La plupart du temps, il ne sera pas possible, pour des raisons de coût si la population est très nombreuse ou simplement de connaissance parfaite de la population, de faire un recueil exhaustif de l'ensemble des valeurs prises par les variables que l'on veut étudier. On recueillera alors des données soit par **sondage** soit sur un **panel**. On traitera donc alors une sous population appelé échantillon.

Dans la suite nous considérerons la variable X restreinte à la sous population.

Il faudra ensuite organiser et traiter ces données. Pour cela les données sont regroupées dans un tableau statistique où les colonnes représentent les variables et les lignes les individus, l'intersection d'une ligne i et d'une colonne j donnant la valeur de la variable j pour l'individu i. Exemple de tableau utilisé sous SPSS :

	Kms	Revision
1	25500	Révision
2	25700	Révision
3	21700	Pas de révision
4	27300	Révision
5	29900	Révision
6	21600	Révision
7	20200	Révision
8	14800	Révision
9	19800	Révision
10	29800	Révision
11	22500	Révision
12	27000	Pas de révision

1.3.Statistiques descriptives d'une variable

Pour une variable, les statistiques descriptives se composent de résumés numériques et de graphiques, nous ne donnerons ici que les éléments essentiels.

Variable qualitative

Une variable qualitative partageant la population (ou la sous population) en classes, le résumé que l'on va obtenir est constitué de l'effectif ce ces classes et de leur pourcentage par rapport à la population (ou sous population) totale.

Dans le cas d'une variable qualitative ordinale, les pourcentages cumulés peuvent avoir un sens si l'on regroupe des catégories voisines (par exemple tranches d'âges ou degré de satisfaction).

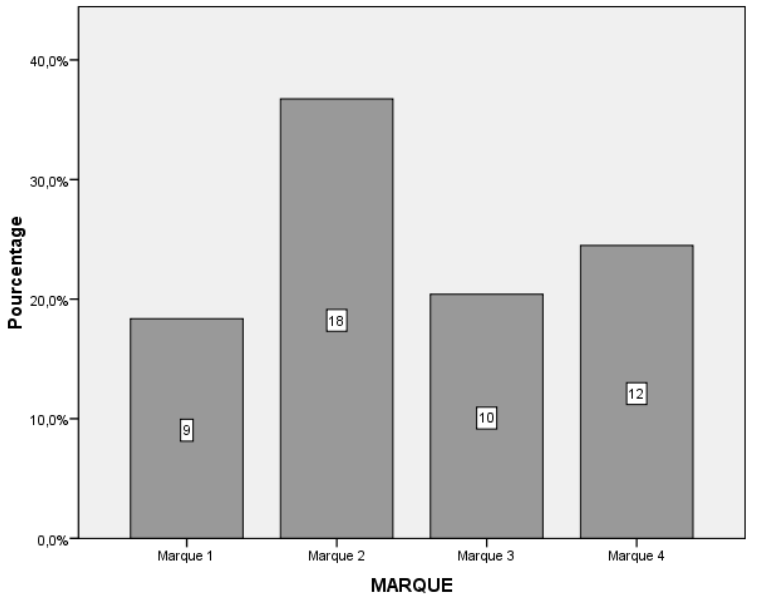
Voici un exemple de résumé fourni par SPSS, pour la variable qualitative Marque du fichier Pfrais.sav :

MARQUE

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Marque 1	9	18,4	18,4	18,4
	Marque 2	18	36,7	36,7	55,1
	Marque 3	10	20,4	20,4	75,5
	Marque 4	12	24,5	24,5	100,0
	Total	49	100,0	100,0	

La colonne pourcentage valide est le pourcentage calculé sur les individus ayant renseigné cette variable.

La représentation associée est le diagramme en bâtons, qui se distingue de l'histogramme par le fait que les rectangles représentant les effectifs ou les pourcentages sont disjoints :



Ici apparait dans chaque rectangle l'effectif de la classe.

Variable quantitative

Le résumé pour une variable qualitative est plus complet, car il doit éventuellement donner des indications sur la loi de probabilité sous-jacente à ces données, en statistique en effet de

nombreuses méthodes supposent des hypothèses sur cette loi. Nous ne verrons ici qu'une partie de ces indicateurs. Nous noterons N la taille de la population ou sous population et X la variable quantitative.

Indicateur de position centrale

Deux indicateurs sont particulièrement utilisés :

- La moyenne : $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$, cette valeur est celle qui est associée à la métrique euclidienne habituelle. La moyenne μ est la valeur la plus proche de toutes les observations pour cette métrique, c'est-à-dire que pour cette valeur la fonction : $d^2(y) = \sum_{i=1}^{N}(x_i - y)^2$ est minimum. Le principal défaut de cet indicateur, comme il est facile de le voir, est sa sensibilité aux valeurs extrêmes, une erreur de saisie peut la modifier profondément.
- La médiane m est la valeur qui partage l'ensemble des données en deux parties égales : 50% des observations sont inférieures ou égales à cette valeur m et 50% sont supérieures à m . Cette valeur est associé à la métrique définie par la valeur absolue, c'est cette valeur m qui minimise la fonction $\partial(y) = \sum|x_i - y|$. Cette valeur est beaucoup moins sensible aux valeurs extrêmes.

Indicateurs de dispersion

L'indicateur de dispersion le plus simple est donné par la valeur la plus petite et la valeur la plus grande. La différence entre ces deux valeurs s'appelle l'étendue :

$$etendue = \max - \min .$$

Les autres indicateurs de dispersion sont liés aux indicateurs de position centrale.

- A la moyenne est associé l'écart-type qui est la racine carré de la distance moyenne au carré, appelée variance :
$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \text{ et l'écart - type } \sigma = \sqrt{V}$$
- A la médiane on pourrait associer de façon "naturelle" l'écart absolu moyen défini par
$$e = \frac{1}{N}\sum_{i=1}^{N}|x_i - m|$$
mais on préfère utiliser les quartiles, déciles ou centiles qui partagent respectivement les données en quatre, dix ou cent parties ayant le même nombre d'éléments. L'intervalle interquartile est la différence entre le premier et le troisième quartile.

Statistiques descriptives

Voici un exemple (fichier Forfait.sav) de résumé fourni par SPSS (l'étendue étant nommée plage ou intervalle) :

Statistiques

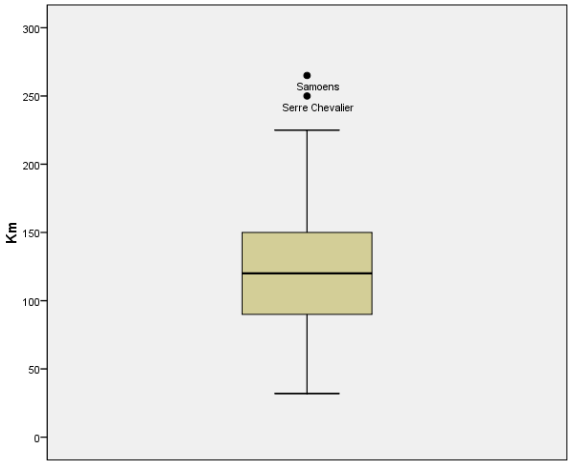
Km

N	Valide	42
	Manquante	0
Moyenne		128,10
Médiane		120,00
Ecart-type		54,134
Variance		2930,479
Intervalle		233
Minimum		32
Maximum		265
Centiles	25	89,50
	50	120,00
	75	152,50

Les représentations associées aux variables qualitatives permettent de visualiser ces résumés et de se faire une idée de la distribution théorique que l'on pourrait associer à cette variable, dans les cas les plus fréquents on cherchera à voir si cette distribution peut suivre une loi normale. En dehors des histogrammes bien connus, nous présenterons ici les boites à moustaches (Box Plot) et les diagrammes Q-Q (Q-Q Plot).

Boite à moustaches

Une boite à moustache est une représentation associée au résumé médiane-quartiles, la boite (rectangle) représente le premier et le troisième quartile avec un trait pour la médiane, les moustaches (traits verticaux) représentent (aux données exceptionnelles près –outliers) le minimum et le maximum. Ces moustaches sont limitées à 1,5 fois la distance interquartile.



Ici deux stations ont un domaine skiable "anormalement" étendu, mais pour le reste la boite est assez symétrique et l'hypothèse de normalité pour la lois sous jacente ne paraît pas absurde.

Diagramme Q-Q

L'idée d'un diagramme Q-Q est de comparer les percentiles des observations avec les percentiles d'une loi théorique. Nous ne traiterons que le cas de la loi normale centrée réduite, le cas général étant facilement compréhensible.

Dans un premier temps les données sont réduites, c'est-à-dire que l'on soustrait la moyenne aux observations et on divise par l'écart-type, la nouvelle variable est donc définie par :

$$X_1 = \frac{X - \mu}{\sigma}$$

Les N données sont ensuite ordonnées par ordre croissant, la valeur de la première observation est alors comparée au percentile $\frac{0,5}{N}$ de la loi normale centrée réduite, la seconde au percentile $\frac{1,5}{N}$ etc.. la dernière au percentile $\frac{N-0,5}{N}$. On représente alors graphique cette comparaison en mettant en abscisse les valeurs observées et en ordonnées les valeurs théoriques. Si l'ajustement à la loi normale était parfait les points seraient alignés sur la diagonale.

Sur notre exemple on obtient le graphique suivant :

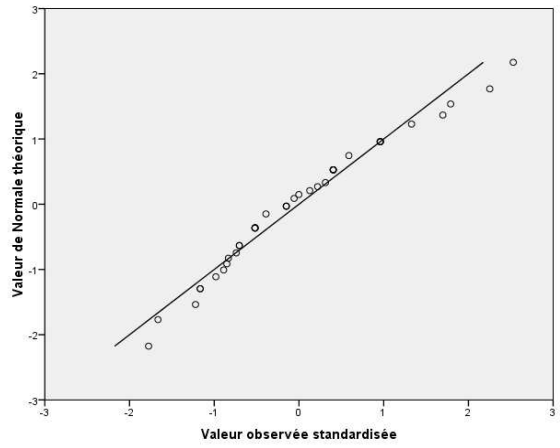


Figure 1 : Toutes les données

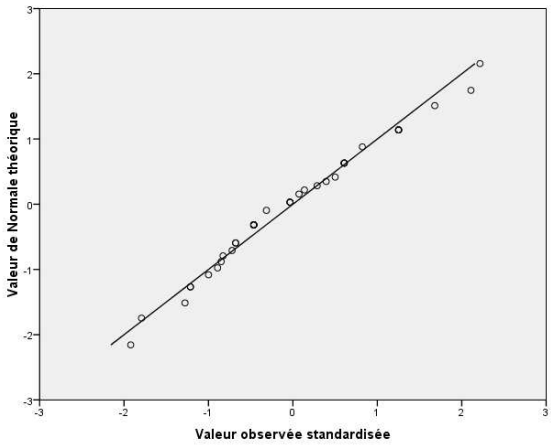


Figure 2 : Sans Serre Chevalier et Samoens

L'ajustement est correct, bien que l'on retrouve les valeurs extrêmes en queue de distribution (Figure 1) mais bien meilleurs après élimination des valeurs éloignées (Figure 2)

1.4.Statistiques descriptives d'un couple de variables

L'objectif de l'étude descriptive d'un couple de variables statistiques est de mettre en évidence une relation éventuelle entre ces deux variables.

Variables quantitatives

L'indicateur de liaison entre deux variables quantitative est la corrélation. Cet indicateur est calculé à partir de la covariance :

$$\mathbf{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

où μ_X et μ_Y désignent respectivement les moyennes des variables X et Y . Pour se débarrasser des effets d'échelle, on divise par les écart-type des variables (ce qui revient à prendre la covariance des variables centrées réduites) :

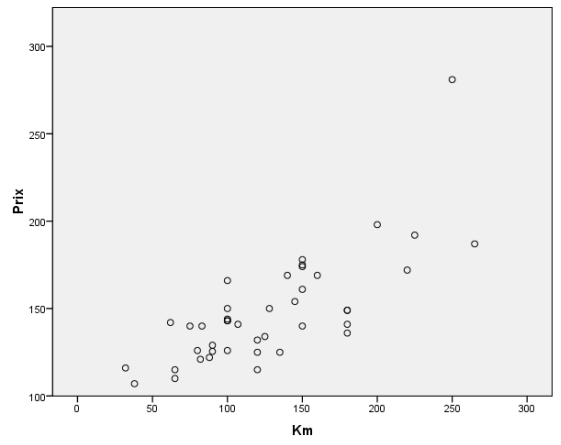
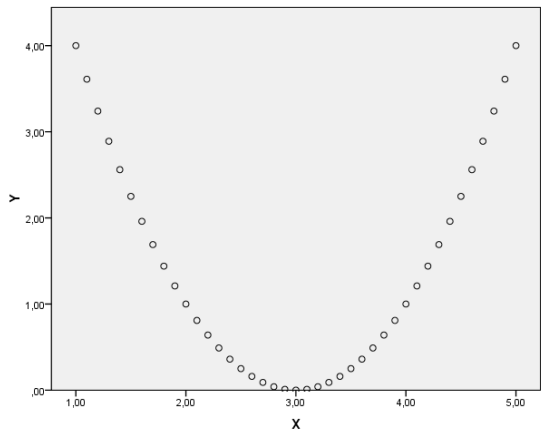
$$\rho(X,Y) = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

Cette corrélation est toujours comprise entre **-1 et 1**. La liaison entre les variables est d'autant plus forte que la valeur absolue est proche de 1.

Une corrélation positive indique une variation moyenne dans le même sens des deux variables, une corrélation négative une variation moyenne en sens inverse.

Remarque : cette corrélation n'est un indicateur que d'une liaison linéaire entre les variables (cf infra). Une corrélation nulle n'indique pas une absence de liaison entre les variables.

La représentation graphique associée est le diagramme cartésien :

	
Corrélation = 0,756	Corrélation = 0

Une variable qualitative et une variable quantitative

Ici on donnera pour chaque modalité de la variable qualitative, les indicateurs de tendance centrale et de dispersion de la variable quantitative restreinte à cette modalité.

Par exemple (fichier Pib.sav) pour les pays de l'Union Européenne, nous avons relever le PIB en $, et la période d'adhésion avec les modalités :

- 1 - adhésion avant 1980
- 2 – adhésion entre 1980 et 2000
- 3 – adhésion après 2000

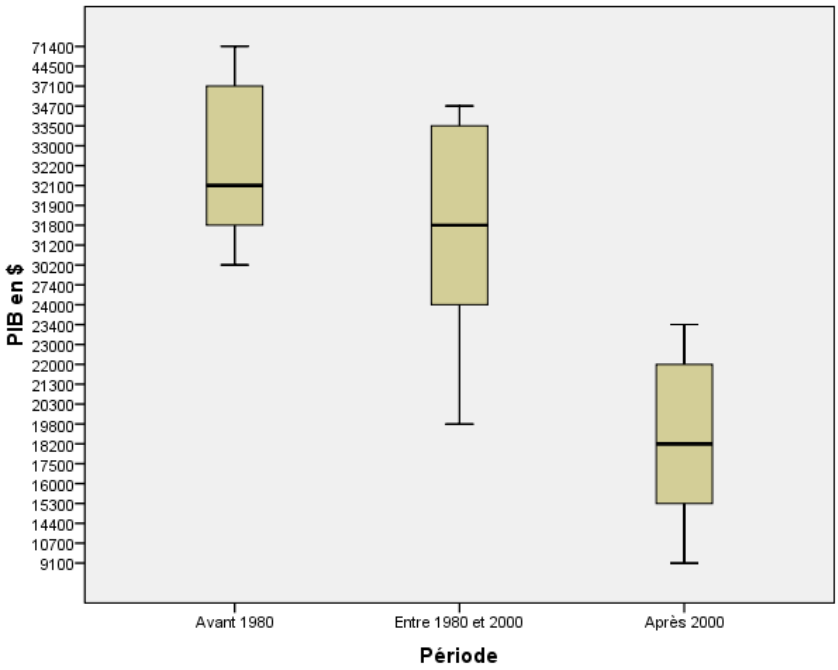
Statistiques descriptives

Le résumé donné par SPSS (après) simplification est le suivant :

Descriptives

	Période		Statistique	Erreur standard
PIB en $	1	Moyenne	38133,33	4411,160
		Médiane	32100,00	
		Variance	1,751E8	
		Ecart-type	13233,480	
		Intervalle interquartile	9300	
	2	Moyenne	28600,00	2409,841
		Médiane	29800,00	
		Variance	3,484E7	
		Ecart-type	5902,881	
		Intervalle interquartile	10850	
	3	Moyenne	17600,00	1352,495
		Médiane	17850,00	
		Variance	2,195E7	
		Ecart-type	4685,180	
		Intervalle interquartile	7200	

On constate que les moyennes et médianes sont très différentes pour la période postérieure à 2000, ce que l'on peut vérifier en demandant un graphique de boîte à moustaches :



Variables qualitatives

On testera ici l'"indépendance" de deux variables qualitatives. Comme en probabilité, mais ici les variables statistiques ne sont pas des variables aléatoires, on dira que deux variables sont indépendantes si les répartitions de la variables X selon les modalité de la variable Y sont les mêmes quelque soit la modalité de X prise en compte (et bien sur réciproquement si les répartition de la variable Y selon les modalités de la variable X sont les mêmes quelque soit la modalité de Y prise en compte). Comme les effectifs de chaque modalité ne sont pas identiques pour que cette définition est un sens il faut raisonner en fréquence, on doit donc avoir en cas d'indépendance (en notant $f_{i,j}$ la fréquence dans la population de la présence simultanée des modalités i **et** j :

$$f_{i,j} = f_i \times f_j \text{ soit en effectifs } N_{i,j} = \frac{N_i \times N_j}{N}$$

Comme résumé numérique on donnera le tableau croisé, en mettant en ligne les modalités de X et en colonne les modalités de Y, chaque cellule du tableau contenant l'effectif réel (constaté) ainsi que l'effectif calculé en cas d'indépendance noté effectif théorique.

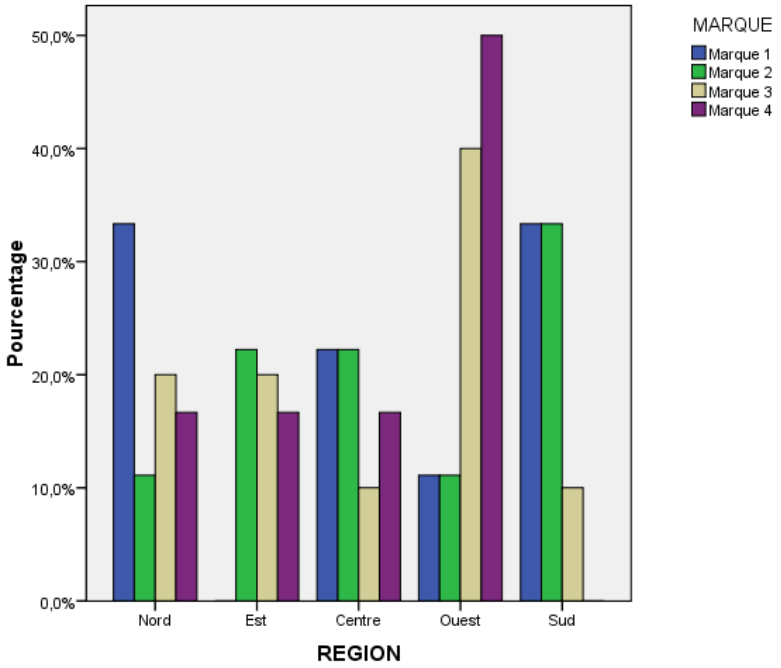
Exemple (fichier pfrais.sav) relation entre marque et région :

Tableau croisé MARQUE * REGION

			REGION					Total
			Nord	Est	Centre	Ouest	Sud	
MARQUE	Marque 1	Effectif	3	0	2	1	3	9
		Effectif théorique	1,7	1,5	1,7	2,4	1,8	9,0
	Marque 2	Effectif	2	4	4	2	6	18
		Effectif théorique	3,3	2,9	3,3	4,8	3,7	18,0
	Marque 3	Effectif	2	2	1	4	1	10
		Effectif théorique	1,8	1,6	1,8	2,7	2,0	10,0
	Marque 4	Effectif	2	2	2	6	0	12
		Effectif théorique	2,2	2,0	2,2	3,2	2,4	12,0
Total		Effectif	9	8	9	13	10	49
		Effectif théorique	9,0	8,0	9,0	13,0	10,0	49,0

Remarquons qu'un tel tableau est difficile à interpréter puisque les écarts se répercutent sur plusieurs cellules (cf test du Khi-2).

On peut associer à un tel tableau un diagramme en bâtons en mettant les pourcentages en ordonnée, en cas d'indépendance stricte tous les blocs seront alors identiques. Avec l'exemple précédent on obtient le graphique suivant :



2. SONDAGE-ESTIMATION

2.1.Un exemple.

Monsieur Martin, chef de produit d'une voiture de moyenne gamme, lancée depuis trois ans, veut savoir si la promotion qu'il a mis en place pour les révisions annuelles a eu un impact sur les clients.

D'ordinaire 60% des clients font leurs révisions annuelles chez les concessionnaires, il aimerait avoir une idée de la proportion des utilisateurs du modèle qui ont fait leur révision chez un garagiste du réseau ; malheureusement son budget ne lui permet de faire des interviews de tous les clients ayant acheté un véhicule depuis plus d'un an (au nombre de 42 612 pour les deux années) et il ne pourra demander à un institut de marketing téléphonique que d'interroger 500 personnes.

Monsieur Martin se demande comment va procéder l'institut et quelle est la fiabilité du résultat obtenu, non pas sur les 500 personnes mais sur l'ensemble des clients. Il aimerait par la même occasion savoir quel kilométrage parcourt environ un client type par an pour pouvoir affiner son offre.

Posons le problème de Monsieur Martin en termes statistiques. Monsieur Martin s'intéresse à une population précise, les personnes ayant acheté une voiture du modèle donné depuis plus d'un an, et l'ayant gardé ; en fait pour le kilométrage la population n'est pas la même, c'est seulement les clients ayant cette voiture depuis plus d'un an. Nous noterons P cette population.

Sur cette population deux variables statistiques concernent Monsieur Martin, une variable qualitative à savoir le lieu où le client a fait sa dernière révision variable que nous noterons X, une variable quantitative le nombre de kilomètres parcourus en 1 an que nous noterons Y.

Présentation mathématique

Nous noterons N la taille de la population.

La variable qualitative X, étant à deux modalités (révision chez le concessionnaire ou non), peut être considérée comme une variable à valeurs dans {0 ;1}, 1 signifiant que la révision est faite chez le concessionnaire :

$$P \xrightarrow{X} \{0;1\}$$

Le paramètre qui nous intéresse, le pourcentage de clients faisant leur révision chez le concessionnaire, peut s'exprimer facilement en fonction de cette variable :

$$\boldsymbol{p = \frac{1}{N}\sum_{i=1}^{N} X(i)}$$

c'est en effet la moyenne de la variable X sur l'ensemble de la population, il suffit en effet de compter les clients qui vont chez un concessionnaire, c'est à dire ceux pour lesquels X prend la valeur 1, et de diviser par la taille de la population.

Pour la variable Y qui est numérique nous pouvons la considérer comme une application de la population P dans l'ensemble des nombres réels $\mathbf{R}$

$$P \xrightarrow{Y} \mathbf{R}$$

Estimation

Les paramètres qui peuvent être intéressants sur cette variable sont la moyenne et la variance (ou sa racine carrée l'écart type) de cette variable :

$$\mu=\frac{1}{N}\sum_{i=1}^{N}Y(i)$$

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y(i)-\mu)^2}$$

L'écart type donne une indication sur la dispersion des valeurs prises par la variable Y, mais jouera aussi un rôle sur les moyennes prises sur les échantillons, comme nous le verrons plus loin.

Paramètres de la population.

Dans le fichier Martin.sav, vous trouverez le tableau statistique relatif à ces populations et à ces variables, nous connaissons ces données, mais malheureusement pour lui Monsieur Martin n'y a pas accès.

Ce fichier contient 42540 données, la première colonne contient le nombre de kilomètre parcouru dans l'année, la deuxième colonne le fait que le client aie fait sa révision chez un concessionnaire ou non.

Nous pouvons obtenir des résultats exacts sur la population (analyse descriptive de SPSS) pour les deux variables qui nous intéressent (mais Monsieur Martin lui ne les aura pas) :

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Kms	42540	8600	41600	25005,16	3978,211
Revision	42540	0	1	,76	,428
N valide (listwise)	42540				

Remarquons tout d'abord que Monsieur Martin fait une première erreur, il croit connaître le nombre des clients, mais en fait un certain nombre d'entre eux ont revendu ou cassé leur voiture et le fichier client ne peut pas être réellement à jour ; cela peut le conduire à sous estimer le coût de son enquête car pour obtenir 500 réponses (même en supposant que toute personne interrogée veut bien répondre), il faudra contacter plus de 500 personnes. C'est pour cela que le fichier de données fourni ne contient que 42540 clients (cellule nommée Taille).

Les données relatives au kilométrage se trouvent dans la première colonne, celles relatives à la révision dans la deuxième, et pour les données concernant la révision, nous avons noté 1 le fait de faire la révision chez un concessionnaire, 0 sinon ; avec des étiquettes affichant respectivement Oui ou Non.

2.2.Constitution d'un échantillon

Pour qu'un échantillon puisse nous donner un résultat fiable, il semble naturel qu'il soit représentatif de la population, c'est à dire qu'il soit une image fidèle de la diversité des individus constituant la population.

Pour atteindre cet objectif il est possible de procéder de différentes façons, nous ne parlerons ici que de deux méthodes les plus fréquemment utilisés, les sondages par quotas et les sondages aléatoires, nous illustrerons ce dernier concept avec le fichier de données.

La méthode de sondage par quotas, méthode utilisée par exemple dans les enquêtes d'opinion, repose sur une constitution raisonnée de l'échantillon. En partant du fait que les variables qui vont être analysées dépendent d'autres caractères connus de la population (par exemple la catégorie socioprofessionnelle) on tâchera de respecter dans l'échantillon les mêmes proportions de chacune des catégories dans la population entière. Ensuite on chargera chaque enquêteur d'interroger un nombre donné d'individu de chaque catégorie, l'avantage de cette méthode est qu'elle est moins coûteuse que la méthode aléatoire indiquée ci-dessous, l'inconvénient est que l'on ne connaît pas exactement la précision des résultats obtenus. On peut cependant utiliser les résultats des sondages aléatoires pour avoir une idée de la précision. Remarquons qu'il ne faut pas confondre cette méthode avec la méthode des sondages aléatoires stratifiés, qui permet sous certaines conditions de diminuer de façon significative la taille des échantillons pour une précision donnée ; cette dernière méthode est une méthode aléatoire et permet d'évaluer la précision des résultats.

La méthode de sondage aléatoire permet de constituer des échantillons qui ont une forte probabilité de reconstituer la diversité de la population originelle. Pour cela on procède à un tirage aléatoire uniforme dans la population initiale, c'est à dire que chaque individu de la population a la même probabilité d'être le kième élément de l'échantillon, c'est à dire que l'on transforme la population statistique en un ensemble probabilisé, les variables statistiques devenant alors des variables aléatoires ; nous renvoyons le lecteur intéressé à l'annexe pour la suite de l'illustration mathématique du sondage aléatoire simple. On peut alors procéder soit par tirage sans remise dans la population soit par tirage avec remise, nous supposerons toujours que le tirage effectué est avec remise, ce qui n'est pas trop contraignant si la taille de l'échantillon est faible par rapport à la taille de la population, ce qui est généralement le cas.

Remarquons dès maintenant qu'il est malheureusement possible de « tomber » sur des échantillons aberrants et que donc la notion de précision sera sûrement liée à l'élimination de ces échantillons, donc à un pari sur le fait de ne pas avoir tiré ce type d'échantillon.

Pour pouvoir réaliser ce type de sondage, il est nécessaire de connaître explicitement toute la population, ce qui n'est pas toujours le cas. On numérote les individus de la population de 1 à N, et on effectue, grâce à des nombres aléatoires, un tirage au hasard dans cet intervalle ; on va ensuite « interroger » (dans certains cas consulter, factures, stocks) les individus tirés au hasard. Quand les individus ont des localisations très réparties géographiquement, il est possible, pour diminuer les coûts du sondage de procéder à un tirage hiérarchisé (choix d'une commune proportionnellement à son nombre d'habitants, puis choix d'un quartier etc..).

L'échantillon ainsi tiré s'appelle l'échantillon individu, en lui-même cet échantillon n'a que peu d'intérêt, ce sont les valeurs prises par les variables étudiées qui nous intéresse, c'est ce que l'on appelle l'échantillon image.

Estimation

Présentation mathématique

Le tirage aléatoire simple consiste, tout d'abord, à munir la population *P* d'une loi de probabilité uniforme, c'est à dire que chaque individu a la même probabilité $\frac{1}{N}$ d'être tiré.

Les deux variables statistiques deviennent alors des variables aléatoires, précisons les deux cas que nous trouvons ici.

La variable qualitative X, ne prend que deux valeurs 0 et 1, la valeur 1 ne peut être prise que par les clients allant faire leur révision chez le concessionnaire, c'est à dire que cette valeur à une probabilité *p* d'être tirée, on a donc à faire à une variable de Bernouilli de paramètre *p*, dont l'espérance est *p* et l'écart-type $\sqrt{p(1-p)}$.

La variable quantitative Y, prend un grand nombre de valeurs distinctes, on peut la considérer comme une variable aléatoire continue, très fréquemment on fera l'hypothèse que cette variable quantitative peut être considérée comme une approximation d'une variable suivant une loi normale de paramètre μ et σ : $N(\mu,\sigma)$.

Dans le cas de tirage avec remise, un échantillon individu est un élément de P^n, un échantillon image pour les valeurs de la révision est un élément de $\{0;1\}^n$, pour le kilométrage un élément de $\mathbf{R}^n$ (on pourrait donc considérer l'échantillon image comme un élément de $\{0;1\}^n \times \mathbf{R}^n$). En appelant X_1 (respectivement Y_1) la valeur prise par *X* (respectivement *Y*) pour le premier individu de l'échantillon, et de même pour les autres individus de l'échantillon, on peut mettre en évidence un **n_uple de variables aléatoires indépendantes** qui permettent de passer de l'échantillon individu à l'échantillon image :

$$P^n \xrightarrow{(X_1, X_2, \ldots, X_n)} \{0;1\}^n \text{ ou}$$

$$P^n \xrightarrow{(Y_1, Y_2, \ldots, Y_n)} \mathbf{R}^n$$

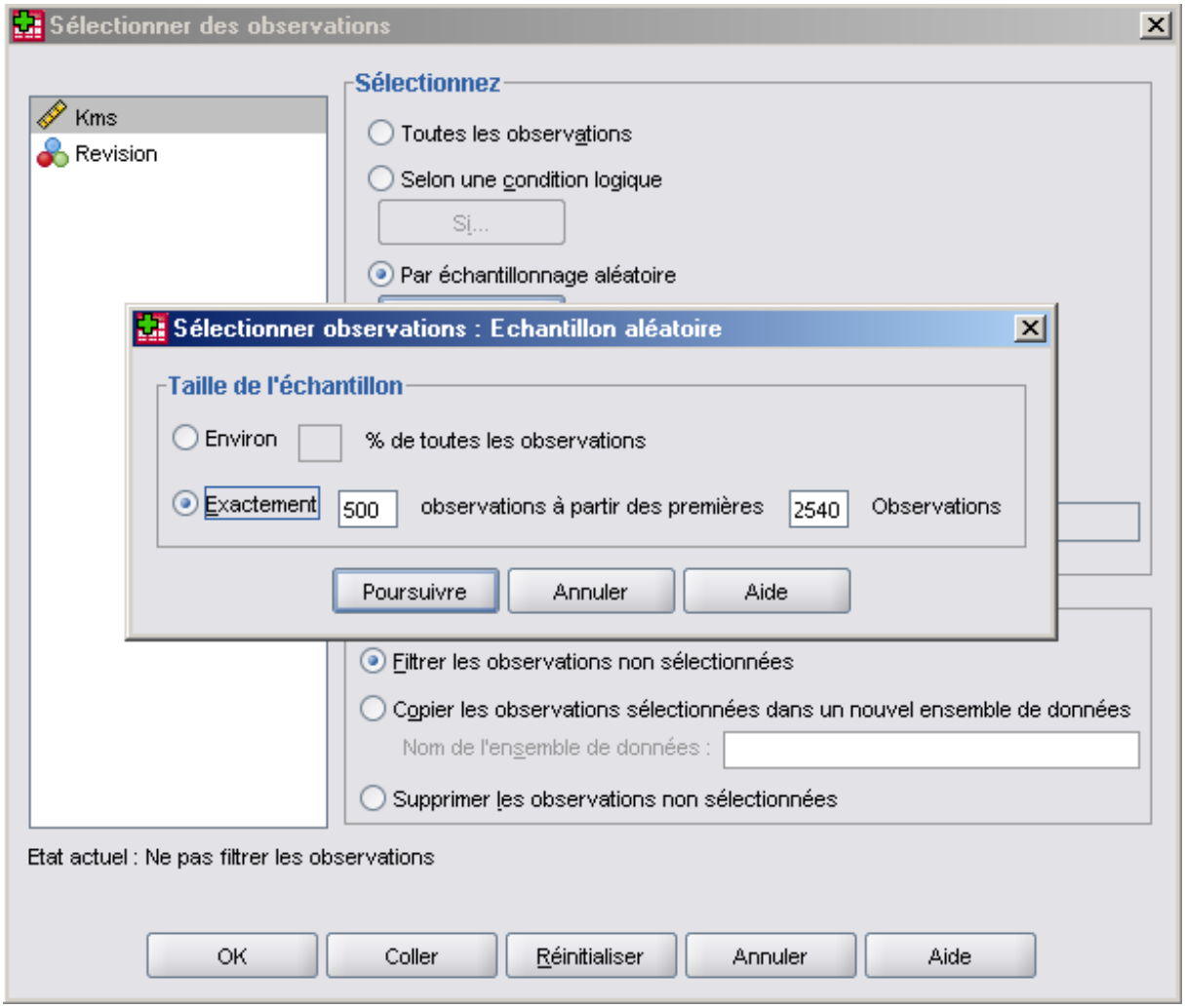
Illustration de cette procédure avec SPSS.

Nous conseillons au lecteur de refaire lui-même le travail.

Avec SPSS nous allons construire de façon aléatoire un échantillon image de taille 500.

- Construction de l'échantillon image

  Dans le menu données, sélectionner des observations, nous choisissons échantillonage aléatoire et une taille de 500 sur l'ensemble des données :



Il apparait une variable filter_$ qui indique si l'observation est choisie dans l'échantillon (valeur 1) ou non. Les observations écartées sont barrées. On obtient ainsi 500 observations qui pourront être utilisées pour l'analyse.

2.3.Estimation – Estimateur

Généralités

Une fois que notre échantillon est obtenu, il nous faut prévoir les résultats sur l'ensemble de la population, c'est à dire extrapoler des valeurs calculées sur l'échantillon comme valeurs des paramètres sur la population. Bien évidemment, cette valeur calculée sur l'échantillon va dépendre de l'échantillon que nous aurons tiré, nous appellerons estimation (ou estimation ponctuelle) cette valeur. Cette estimation est donc le résultat de l'application d'une formule, d'une fonction sur l'échantillon, cette fonction s'appelle l'estimateur.

Aspects mathématiques

Soit donc X une variable statistique définie sur une population P (ici soit la variable X caractéristique de la révision, soit la variable Y liée au kilométrage), soit θ un paramètre de cette variable. On appelle estimateur du paramètre θ sur un échantillon de taille n, une application notée Θ_n :

$$P^n \xrightarrow{\Theta_n} \mathbf{R}$$

et on appellera estimation la valeur prise par cette fonction sur un échantillon particulier. D'un point de vue mathématique, l'estimation n'a en soi que peu d'intérêt, alors que pour l'utilisateur c'est le plus important ; mais ce sont les propriétés de l'estimateur qui sont intéressantes et qui vont garantir la fiabilité de l'estimation.

Les deux propriétés intéressantes pour un estimateur sont :

- Etre non biaisé, c'est à dire que les valeurs prises par l'estimation se répartissent autour de la vraie valeur du paramètre, et ne sont pas systématiquement trop grandes ou trop petites, mathématiquement ceci s'exprimera par $E(\Theta_n)=\theta$, pour tout *n*.
- Etre consistant, ceci signifie que plus la taille de l'échantillon est grande, meilleur est l'estimation, c'est à dire qu'elle a moins de « chances » d'être éloignée de la vraie valeur, ceci se traduit mathématiquement par le fait que la variance de l'estimateur diminue quand la taille *n* de l'échantillon augmente, de façon plus précise on dira que l'estimateur est convergent (dans le cas d'un estimateur non biaisé) si $\lim_{n\to\infty} Var(\Theta_n)=0$.

Estimation de la moyenne ou d'une proportion

Intuitivement, puisque l'échantillon est représentatif de la population, pour estimer la moyenne du kilométrage ou le pourcentage de clients faisant leur révision chez un concessionnaire, il suffira de prendre les mêmes caractéristiques sur l'échantillon. C'est à dire que nous prendrons comme estimation du kilométrage moyen sur la population, la moyenne du kilométrage sur l'échantillon et comme estimation de la proportion sur la population, la proportion de clients faisant leur révision chez un concessionnaire dans l'échantillon.

Suivant les conventions statistiques habituelles, nous noterons $\hat{p}$ l'estimation de la proportion *p* sur l'échantillon de taille n, et nous noterons $\bar{y}_n$ l'estimation de la moyenne du kilométrage sur ce même échantillon. Remarquons qu'il serait plus cohérent de noter $\bar{x}_n$ plutôt que $\hat{p}$ l'estimation de la proportion puisque c'est en fait l'estimation de la moyenne de la variable X.

Propriété mathématique de l'estimateur de la moyenne

Nous ne traiterons ici que le cas de la moyenne, puisque comme il vient d'être noté la proportion en est un cas particulier pour une variable indicatrice (à valeur {0 ;1}).

L'estimateur de la moyenne d'une variable statistique *X* sur un échantillon de taille *n* sera noté $\bar{X}_n$ est défini en fonction de l'échantillon aléatoire $(X_1, X_2, ..., X_n)$ par :

$$\bar{X}_n=\frac{X_1+X_2+...+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Puisque les variables X_i sont toutes de même loi et que l'espérance mathématique est linéaire, il vient immédiatement :

$$E(\bar{X}_n)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=E(X)$$

ce qui signifie que l'estimateur de la moyenne est non biaisé.

D'autre part comme les variables X_i sont de plus indépendantes, nous avons :

$$Var\left(\overline{X}_n\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var\left(\overline{X}_i\right)=\frac{nVar(X)}{n^2}=\frac{Var(X)}{n}$$

ce qui montre que l'estimateur de la moyenne est convergent, en augmentant la taille de l'échantillon, les estimations sont généralement plus proches de la vraie valeur ; nous préciserons plus loin cette notion de "généralement plus proche".

Estimation de la variance

Il peut sembler naturel d'estimer la variance de la population par la variance de l'échantillon ; cependant comme dans ce cas on ne centrerait pas les observations par rapport à la « vraie » moyenne (celle de la population) mais par rapport à la moyenne de l'échantillon, on aura certainement un biais, on aura même certainement tendance à sous estimer la valeur réelle de la variance de la population. Il est facile de démontrer (voir ci-dessous) qu'un estimateur non biaisé de la variance est donné par la formule :

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$$

c'est à dire qu'au lieu de diviser la somme des carrés par *n*, taille de l'échantillon, il faut diviser cette somme par *n-1*. L'estimation est alors :

- Pour une variable quantitative $s_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2$
- Pour une variable indicatrice, comme dans le cas de l'estimation de la proportion de clients faisant leur révision chez un concessionnaire $s_n^2=\frac{n}{n-1}\hat{p}(1-\hat{p})$

Et pour l'écart type on prendra comme estimateur, la racine carré de l'estimateur de la variance ; il faut noter que cet estimateur est biaisé, mais contrairement à la variance on ne sait pas déterminer pas son biais et donc le "débiaiser". Il est cependant asymptotiquement sans biais, ce qui signifie que le biais tend vers 0, donc diminue quand la taille de l'échantillon augmente.

Propriétés mathématiques de l'estimateur de la Variance

Partant de l'"estimateur naturel" de la variance, c'est à dire la variance de l'échantillon, nous allons montrer que c'est un estimateur biaisé, mais que l'on peut calculer ce biais.

Soit donc $V_n=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$ la variable aléatoire qui permet de calculer la variance de l'échantillon.

Comme les variables Xi et $\overline{X}n$ ont même moyenne μ, nous pouvons écrire que

$$E\left(\left(X_i-\overline{X}_n\right)^2\right)=E\left(\left(X_i-\mu-\left(\overline{X}_n-\mu\right)\right)^2\right)=Var\left(X_i\right)+Var\left(\overline{X}_n\right)-2Cov\left(X_i,\overline{X}_n\right)$$

En notant σ^2 la variance commune des Xi nous avons vu que $Var\left(\overline{X}_n\right)=\frac{1}{n}\sigma^2$, il ne nous reste plus qu'à calculer la covariance de Xi et $\overline{X}n$. Comme Xi et X_j sont indépendants pour $i\neq j$, cette covariance est en fait égale à la covariance de Xi et $\frac{Xi}{n}$, c'est à dire $\frac{1}{n}\sigma^2$. On en déduit donc :

$$E\left(\left(X_i-\overline{X}_n\right)^2\right)=\sigma^2+\frac{1}{n}\sigma^2-\frac{2}{n}\sigma^2=\left(1-\frac{1}{n}\right)\sigma^2 \text{ d'où } \mathrm{E}\left(V_n\right)=\frac{1}{n}\left(\sum_{i=1}^{n}\left(1-\frac{1}{n}\right)\sigma^2\right)=\frac{n-1}{n}\sigma^2$$

L'estimateur V_n est donc biaisé, puisque son espérance n'est pas égale au paramètre σ^2, de plus comme $\frac{n-1}{n}$ est strictement inférieur à 1, cet estimateur sous estime la vraie variance. En revanche, il est facile d'obtenir un estimateur non biaisé en prenant :

$$S_n^2=\frac{n}{n-1}V_n=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2$$

On peut de plus monter que cet estimateur est convergent (à condition que les moments d'ordre inférieur ou égal à 4 existent), mais cette démonstration beaucoup plus lourde est laissée au lecteur.

Illustration du comportement de l'estimateur de la moyenne.

Nous avons tiré des échantillons de taille 100, 200, 300, 400 et 500. Pour chacun de ces échantillons, vous trouverez dans le tableau ci-dessous : la moyenne estimée, l'écart-type estimé, l'estimation de l'écart-type de l'estimateur de la moyenne :

Estimation

Taille		N	Moyenne		Ecart type
		Statistique	Statistique	Erreur std	Statistique
100	Kms	100	25162,00	435,648	4356,483
	Revision	100	,80	,040	,402
	N valide (listwise)	100			
200	Kms	200	24996,50	285,776	4041,486
	Revision	200	,79	,029	,408
	N valide (listwise)	200			
300	Kms	300	25063,67	209,248	3624,274
	Revision	300	,78	,024	,417
	N valide (listwise)	300			
400	Kms	400	25015,50	188,294	3765,875
	Revision	400	,79	,020	,406
	N valide (listwise)	400			
500	Kms	500	24964,60	178,842	3999,019
	Revision	500	,77	,019	,421
	N valide (listwise)	500			

A la lecture de ce tableau on constate que si les valeurs ponctuelles de la moyenne et de l'écart-type estimés sur les échantillons ne "s'améliorent pas", en revanche l'écart-type de l'estimateur de la moyenne diminue, c'est à dire que sa précision augmente.

2.4.Estimation par intervalle, précision d'un sondage

Les estimations obtenues pour un paramètre à partir d'un échantillon de même taille sont très variables, il nous faut donc associer à ces estimations une précision qui nous permettra dans un certain sens d'encadrer la vraie valeur du paramètre. Cette notion de précision est plus délicate que celle des mesures en physique, dire qu'un pain pèse 400g à 5g près, cela signifie que le poids du pain est compris de façon certaine entre 395 et 405g. Il n'est pas possible en

statistique d'obtenir cette même notion, nous allons donc introduire une autre notion de précision, associée à un degré de confiance.

Nous nous intéresserons ici qu'au cas de la moyenne ou du pourcentage, mais ce que nous dirons est généralisable à d'autres paramètres.

Tout d'abord, une mauvaise nouvelle : dans la mesure ou nous effectuons des tirages avec remise, nous ne pouvons pas espérer diminuer l'étendue des valeurs obtenues, en effet il est toujours théoriquement possible de tirer un échantillon constitué n fois de l'individu présentant la plus petite (ou la plus grande valeur), il donc inutile d'espérer pouvoir majorer de façon certaine l'erreur commise lors d'un sondage. En revanche dans la mesure, où l'écart type de l'estimateur tend vers 0 quand la taille de l'échantillon augmente, les valeurs extrêmes vont avoir des probabilités de plus en plus faibles d'apparaître, et donc ne seront observées que dans des échantillons de plus en plus exceptionnels. C'est cette notion que nous allons formaliser en étudiant la loi de l'estimateur du pourcentage et de la moyenne.

Généralités : Précision de l'estimation au degré de confiance 1-α

On appellera intervalle de l'estimation au degré de confiance 1-α (α étant un nombre plus petit que 1), l'intervalle dans lequel se trouvent les valeurs l'estimation, quand on a décidé de négliger les échantillons les plus extrêmes ayant la probabilité α d'apparaître.

C'est à dire que l'on fait un pari, on pense que l'on aura la « chance » de ne pas tirer un de ces échantillons extrêmes, et 1-α représente la probabilité que l'on a de gagner ce pari ; α représente le risque d'erreur (ou la malchance). Notons bien que nous ne saurons jamais si oui ou non ce pari a été gagné.

Formellement, nous pouvons écrire : la précision e au degré de confiance 1-a, est définie par :

$$\Pr\left(\left|\overline{X}n-\mu\right|\leq\varepsilon\right)=1-\alpha$$

$\overline{X}n$ étant l'estimateur du paramètre μ. On voit donc sur cette formule qu'il nous faut connaître la loi de l'estimateur $\overline{X}n$ pour pouvoir déterminer ε en fonction de α et de *n*.

Quelques remarques générales :

- Pour *n* fixé, quand α augmente ε diminue, il faudra donc faire un arbitrage (pour un coût donné) entre la précision que l'on désire et le risque que l'on a de perdre son pari.
- En se fixant α et ε, on peut déterminer une taille d'échantillon convenable permettant d'atteindre une précision voulue avec un risque donné, puisque la variance de $\overline{X}n$ tend vers 0. Toutefois, il faudra dans ce cas arbitrer avec le budget disponible.
- Une fois la taille de l'échantillon fixée, la formule ci-dessus peut être inversée et nous obtenons, un intervalle d'estimation qui est un intervalle aléatoire $\left[\overline{X}n-\varepsilon;\overline{X}n+\varepsilon\right]$, dans le quel la vraie valeur du paramètre a une probabilité $1-\alpha$ de se trouver. En remplaçant la variable aléatoire par sa valeur observée sur m'échantillon réellement tiré, on dira souvent, par un raccourci un peu brutal, qu'il y a une probabilité $1-\alpha$ que le paramètre soit dans l'intervalle $\left[\overline{x}n-\varepsilon;\overline{x}n+\varepsilon\right]$, ce qui n'a aucun sens puis que toutes les valeurs sont certaines et que l'on n'a plus alors de loi de probabilité.

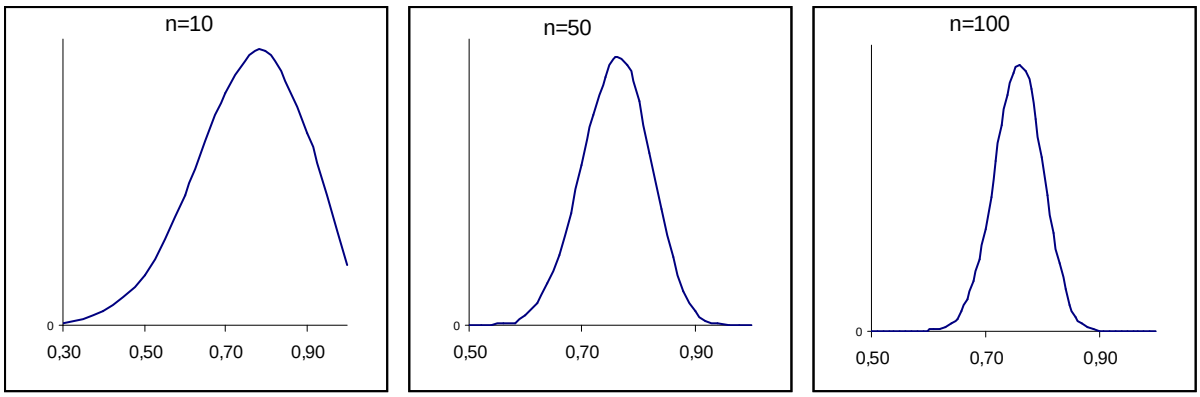
Cas du pourcentage

Loi de probabilité de $\overline{X}n$

La loi de *X* sur la population initiale est, comme nous l'avons vu (0), une loi de Bernouilli de paramètre *p*.

Il est possible dans ce cas de déterminer exactement la loi de l'estimateur du pourcentage, puisque nous avons à faire la moyenne de n variables indépendantes de Bernouilli. La variable $n\overline{X}_n$ est donc la somme de n variables de Bernouilli indépendantes, et suit donc une loi binomiale bien connue. Il est donc possible de définir la loi de $\overline{X}_n$ en fonction du paramètre *p* (pourcentage à estimer).

Cependant comment faire pour donner la précision d'une estimation quand on ne connaît pas la vraie valeur ? Comme dans la pratique la taille des échantillons est généralement beaucoup plus grande que 10 (les sondages d'opinion se font sur des échantillons d'au moins 500 personnes, le plus souvent un millier), nous allons pouvoir répondre à cette question en regardant l'évolution de la loi de $\overline{X}n$ en fonction de n. On obtient les graphiques suivants :



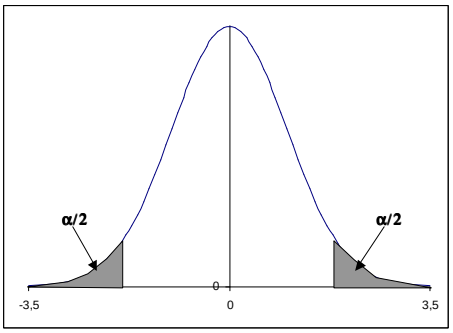
On obtient rapidement une loi de probabilité caractéristique : en forme de cloche, symétrique autour de la valeur moyenne, on reconnaît la loi de Gauss ou loi normale. C'est une simple illustration du théorème de la limite centrée, sur ce cas particulier la variable aléatoire $\frac{\overline{X}n-E(\overline{X}n)}{\sigma(\overline{X}n)}$ tend, quand n tend vers l'infini, en loi vers la loi normale centrée réduite. On peut en pratique considérer que la limite est atteinte pour n>30, on pourra donc assimiler la loi de $\overline{X}n$ à une loi normale de moyenne $E(\overline{X}n)=E(X)=p$, et d'écart-type $\sigma(\overline{X}n)=\sqrt{Var(\overline{X}n)}=\sqrt{\frac{Var(X)}{n}}$.

Nous pouvons maintenant utiliser ce résultat pour donner une estimation par intervalle à un degré de confiance donné.

Calcul de la précision

Nous noterons z_α le fractile d'ordre α de la loi normale centrée réduite, c'est à dire le nombre défini par :

$$\Pr(Z<z_\alpha)=\alpha \quad \text{où} \quad Z\rightarrow N(0,1)$$



Comme $\overline{X}_n$ suit une loi normale, en la centrant et réduisant, on en déduit que $Z=\frac{\overline{X}_n-p}{\sigma(\overline{X}_n)}$ suit une loi normale centrée réduite. La définition de la précision et du degré de confiance peut donc se réécrire de la façon suivante :

$$\Pr\left(|Z|<\frac{\varepsilon}{\sigma(\overline{X}_n)}\right)=1-\alpha \quad \text{soit encore} \quad \Pr\left(\frac{-\varepsilon}{\sigma(\overline{X}_n)}<Z<\frac{\varepsilon}{\sigma(\overline{X}_n)}\right)=1-\alpha$$

Comme la loi normale centrée réduite est symétrique, cette probabilité s'exprime aussi :

$$\Pr\left(\frac{-\varepsilon}{\sigma(\overline{X}_n)}<Z<\frac{\varepsilon}{\sigma(\overline{X}_n)}\right)=1-2\Pr\left(Z\geq\frac{\varepsilon}{\sigma(\overline{X}_n)}\right) \quad \text{donc} \quad \Pr\left(Z\geq\frac{\varepsilon}{\sigma(\overline{X}_n)}\right)=\alpha/2 \quad \text{ou} \quad \Pr\left(Z<\frac{\varepsilon}{\sigma(\overline{X}_n)}\right)=1-\alpha/2$$

on obtient alors l'expression de la précision en fonction du fractile d'ordre $1-\alpha/2$ lu sur une table de la loi normale inverse :

$$\varepsilon=z_{1-\alpha/2}*\sigma(\overline{X}_n)=z_{1-\alpha/2}*\sqrt{\frac{p(1-p)}{n}}\,.$$

Malheureusement $\sigma(\overline{X}_n)$ dépend du paramètre que l'on veut estimer (le pourcentage), et n'est donc pas connu. L'usage veut que l'on remplace cette valeur inconnue par son estimation sur l'échantillon avec la correction que nous avons signalée :

$$\varepsilon=z_{1-\alpha/2}*\hat{\sigma}(\overline{X}_n)=z_{1-\alpha/2}*\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\,.$$

L'estimation par intervalle au degré de confiance 1-α, est alors le suivant :

$$\left[\hat{p}-z_{1-\alpha/2}*\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}};\hat{p}+z_{1-\alpha/2}*\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right]$$

Le tableau suivant donne les valeurs des fractiles de la loi normale pour les degrés de confiance les plus souvent utilisés :

Dégré de confiance(1-α)	**1-α/2**	**Fractile ($z_{1-\alpha/2}$)**
90%	0,95	1,645
95%	0,975	1,960
99%	0,995	2,576

Sur notre échantillon de taille 500, nous obtenons alors les résultats suivants pour les intervalles de confiances :

Dégré de confiance(1-α)	Intervalle de confiance
90%	[0.739 ; 0.801]
95%	[0.733 ; 0.807]
99%	[0.721 ; 0.819]

Remarque : Les intervalles de confiance ont des valeurs fixes, donc la valeur réelle est dans cet intervalle ou n'y est pas, la "confiance" indique simplement que si l'on répétait le calcul de ces intervalles sur un nombre très grand d'échantillons, 95% des intervalles ainsi calculés contiendrait la "vraie valeur", donc 5% ne la contiendrait pas!

Détermination d'une taille d'échantillon

La formule donnant la précision peut être utilisée aussi, pour déterminer la taille d'échantillon nécessaire pour obtenir une précision voulue à un degré de confiance donné. Nous allons distinguer deux cas, suivant que l'on possède ou non une première estimation du pourcentage.

Détermination d'une taille à priori

Dans ce cas nous allons partir de la formule exacte de la précision :

$$\varepsilon = z_{1-\alpha/2} * \sigma\left(\overline{X}_n\right) = z_{1-\alpha/2} * \sqrt{\frac{p(1-p)}{n}}$$

Pour un niveau donné du degré de confiance, il est facile de déterminer la taille d'échantillon n permettant d'obtenir une précision ε donnée : $n \geq \frac{\left(z_{1-\alpha/2}\right)^2 p(1-p)}{\varepsilon^2}$, et ceci doit être vérifié

pour toute valeur de p sur la population, puisque nous n'avons aucune connaissance à priori sur cette proportion. Or quand $0 \leq p \leq 1$ la quantité $p(1-p)$ reste toujours inférieure ou égale à $1/4$[1]. En conclusion la taille nécessaire pour obtenir une précision donnée e, à un degré de confiance a, sans information à priori sur le pourcentage est donnée par la formule :

$$n = \text{EntierSup}\left(\frac{\left(z_{1-\alpha/2}\right)^2}{4\varepsilon^2}\right)$$

EntierSup(x) désignant le plus petit entier supérieur ou égal à x.

Remarquons que cette formule peut être toujours appliquée, elle seule assurera d'obtenir la précision voulue, mais bien évidemment elle conduira à des tailles importantes d'échantillons pas toujours nécessaires mais toujours coûteuses. Nous illustrerons ceci au paragraphe suivant.

Détermination de la taille après pré échantillonnage

Si nous disposons d'une estimation du pourcentage nous pouvons espérer diminuer la taille de l'échantillon nécessaire, en prenant comme valeur probable de la proportion, la dernière valeur estimée. On utilisera alors la formule approchée de la précision à un degré de confiance donnée. Avec les mêmes notations qu'au paragraphe précédent nous obtenons :

[1] Comme il est facile de le voir par dérivation, ou en remarquant que la surface maximale d'un rectangle de périmètre donné (ici 2) correspond au carré.

$$n=\text{EntierSup}\left(\frac{\left(z_{1-\alpha/2}\right)^2 \hat{p}(1-\hat{p})}{\varepsilon^2}\right)+1$$

La seule différence avec le calcul théorique (c'est à dire utilisant la "vraie" valeur *p*, est le +1 final, qui est souvent négligeable dans la pratique.

Dans les deux cas nous pouvons constater que la précision coûte cher en statistique, en effet la taille de l'échantillon varie comme l'inverse du carré de l'estimation, donc pour diviser par 2 la précision (donc l'imprécision), il faut multiplier par 4 la taille de l'échantillon.

Calculs sur notre exemple

Nous allons calculer de deux façons la taille de l'échantillons nécessaire pour avoir une précision de 3% avec un degré de confiance de 95%.

a) Calcul à priori (avant tout sondage)

Nous prendrons ici un pourcentage "pessimiste" de 0,5 :

$$n = entier\sup\left(1{,}96 * \left(\sqrt{0{,}5 * (1-0{,}5)}\right)/0{,}03\right)^2 = 1068$$

La taille à priori nécessaire est donc de 1068 individus

b) Calcul a posteriori (après échantillonnage de taille 500)

Nous prendrons ici le pourcentage estimé $\hat{p} = 0{,}77$

$$n = entier\sup\left(1{,}96 * \left(\sqrt{0{,}77 * (1-0{,}77)}\right)/0{,}03\right)^2 + 1 = 757$$

La taille de l'échantillon est alors nettement plus petite, il suffirait d'ajouter 250 individus environ pour espérer atteindre la précision voulue.

Remarquons enfin, que dans tous les cas il est nécessaire après avoir fait le sondage de recalculer la précision obtenue, qui ne peut qu'être meilleure (inférieure) si l'on utilise la première méthode de majoration, mais qui peut être supérieure à la valeur désirée dans le cas de la seconde méthode, si la nouvelle valeur estimée est plus proche de 50% que celle qui a servi à la détermination de la taille de l'échantillon.

Cas de la moyenne

Sur la population nous avons une variable aléatoire numérique *Y* qui a une moyenne notée μ et un écart type noté σ.

L'estimateur de la moyenne que nous avons utilisé au paragraphe 0 noté $\overline{Y_n}$ (de moyenne m et d'écart type $\sigma/\sqrt{n}$) a la même propriété asymptotique que l'estimateur du pourcentage, c'est à dire qu'il vérifie le théorème de la limite centrée : $Z_n = \frac{\overline{Y_n}-\mu}{\sigma/\sqrt{n}}$ tend en loi vers la loi normale centrée réduite $N(0,1)$. Cependant la vitesse de cette convergence peut dépendre de façon très significative de la forme de la loi initiale de *Y*, très souvent il est fait l'hypothèse que cette loi est proche d'une loi normale, ce qui assure une convergence rapide. Dans le cas où la variable Y suivrait exactement une loi normale, la variable Z_n précédemment définie suit toujours une loi normale.

Cas où la variance est connue

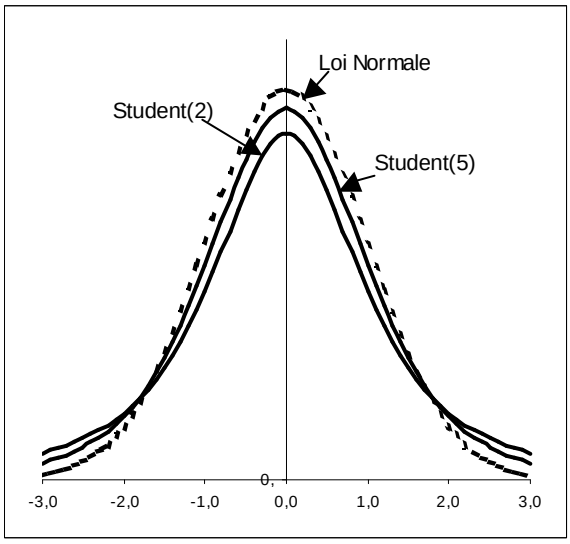
Dans le cas où la variance σ est connue, ce qui est très rare en pratique, on peut utiliser le théorème central limite, pour des échantillons de taille suffisante (n>30, si la loi de Y ne semble pas trop « anormale »). La précision, au degré de confiance α, est alors donnée par :

$$\varepsilon = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{1-\alpha/2}$ désignant le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Cas où la variance est inconnue

Dans ce cas, il nous faut ajouter une hypothèse sur loi de Y. L'hypothèse de normalité de *Y* permet de connaître exactement la loi de la variable aléatoire $T_n = \frac{\overline{Y}_n - \mu}{\sqrt{S_n^2/n}}$ (σ est remplacé par l'estimateur de l'écart type), cette loi est la loi de Student[2] à n-1 degrés de liberté. Cette loi est une loi symétrique comme la loi normale centrée réduite, cependant les queues de distribution sont plus épaisses que celles de la loi normale, ce qui veut dire qu'il y a une probabilité plus forte d'obtenir des échantillons dont la moyenne est éloignée de la moyenne de la population ; toutefois quand n augmente la loi de Student à n degrés de libertés se rapproche de la loi normale centrée réduite qui en est la limite quand $n \to \infty$. En pratique, quand $n > 500$, on pourra sans problème utiliser la loi normale plutôt que la loi de Student.



On obtient alors comme intervalle d'estimation aléatoire au degré de confiance, l'intervalle dont les bornes sont des variables aléatoires :

$$\left[\overline{Y}_n - t^{n-1}_{1-\alpha/2}\sqrt{S_n^2/n}\,;\overline{Y}_n + t^{n-1}_{1-\alpha/2}\sqrt{S_n^2/n}\right]$$

où $t^{n-1}_{1-\alpha/n}$ désigne le fractile d'ordre $1-\alpha/2$ de la loi de Student à n-1 degrés de liberté. Les valeurs de ces fractiles sont lus dans les tables statistiques.

[2] Voir l'annexe pour quelques indications sur cette loi.

Si l'on construit tous les intervalles de cette forme en remplaçant les variables par leurs valeurs prises sur les échantillons (ou du moins un très grand nombre), il y en aura une proportion α qui contiendra la valeur μ du paramètre, et donc 1-α qui ne contiendra pas la valeur μ. On retrouve la notion de pari que nous avons exposée au début de ce paragraphe.

En pratique, on remplacera les variables aléatoires par leurs valeurs, et on dira que l'on a une probabilité de 1-α, que la moyenne se trouve dans l'intervalle $\left[\bar{y}_n - t^{n-1}_{1-\alpha/2}\hat{\sigma}/\sqrt{n}; \bar{y}_n + t^{n-1}_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\right]$,

$\hat{\sigma}$ étant l'estimation de l'écart type.

La précision au degré de confiance α est donc donnée par la formule :

$$\boxed{\varepsilon = t^{n-1}_{1-\alpha/2}\hat{\sigma}/\sqrt{n}}$$

Application au cas de notre exemple

Avec un degré de confiance de 0,95 et un nombre de degrés de liberté de 500-1=499, nous obtenons :

$$\varepsilon = 1{,}965 * 3999/\sqrt{500} \approx 351 km$$

D'où l'intervalle de confiance pour le kilométrage moyen parcouru :

$$[24965 - 351; 24965 + 351] = [24614; 25316]$$

Détermination de la taille d'un échantillon

Comme il a été vu pour le cas d'une proportion, les formules que nous venons de voir permettent aussi, une fois le degré de confiance fixé et une valeur de la précision donnée, de déterminer la taille nécessaire de l'échantillon. Nous ne traiterons ici que le cas où l'écart type de la variable est inconnu, signalant au passage le cas de l'écart type connu.

Remarquons tout d'abord, qu'il est dans ce cas toujours nécessaire d'avoir procéder à un pré sondage, de façon à obtenir une première estimation de l'écart type. Ce pré sondage se fait généralement sur un échantillon d'individus dont le nombre est compris entre 20 et 50. C'est à partir de cette première estimation de l'écart type que sera évaluée la taille de la population nécessaire à l'obtention d'une précision donnée.

Si nous voulons, comme pour le cas d'une proportion, déterminer la taille à partir de la formule de la précision nous obtenons, pour une précision e donnée et un degré de confiance a, le résultat suivant :

$$n = \left(t^{n-1}_{1-\alpha/2}\hat{\sigma}/\varepsilon\right)^2$$

il apparaît un problème, car le fractile de la loi de Student dépend du nombre de degré de libertés, c'est à dire de la taille de l'échantillon. Nous avons donc une équation implicite que nous ne savons pas résoudre analytiquement ; il est possible cependant de la résoudre par approximation de deux façons différentes.

Cas des grands échantillons

D'après ce qui a été dit plus haut quand n est grand, la loi de Student à n degrés de libertés peut être confondue avec la loi normale centrée réduite. La formule établie ci dessus est dans ce cas exploitable et nous obtenons :

$$n=\left(u_{1-\alpha/2}\,\hat{\sigma}/\varepsilon\right)^2$$

où $u_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Cette formule s'applique pour toute taille d'échantillon si on dispose de la valeur de l'écart type sur la population.

Application à notre exemple :

En partant du sondage réalisé sur 100 individus quelle taille d'échantillon est-elle nécessaire pour atteindre une précision de 200 km avec un degré de confiance de 0,95 ?

Seule la valeur de l'écart-type estimé de ce premier sondage nous importe :

$$\hat{\sigma} = 4357$$

D'où le calcul de n :

$$n = (1{,}96 * 4357 / 200)^2 = 1823$$

Dans ce cas évidemment, il faudra vérifier sur l'échantillon final que la précision est bien atteinte, d'autant plus que l'estimation de l'écart-type peut-être très volatile.

Cas général

Si l'on ne veut pas utiliser l'approximation par une loi normale, il faut alors utiliser des méthodes itératives pour déterminer la taille de l'échantillon, mais les résultats trouvés différent peu de l'approximation normale, dont on pourra se contenter en la majorant éventuellement si la valeur trouvée est faible.

2.5.Annexe 1 : La loi de Student

William Sealey Gosset (1876-1937) était chimiste à la brasserie Guiness à Dublin, puis ensuite à Londres. C'est pour le contrôle de qualité qu'il fut conduit à s'intéresser à l'échantillonnage et surtout aux petits échantillons. Il publia ses travaux sous le nom de Student. C'est lui qui mit en évidence la loi qui porte son nom et qui permet de faire des tests sur la moyenne d'une variable quantitative.

Gosset étudia la fonction de répartition de la variable (dite variable de Student à n degrés de liberté) $T = X \Big/ \sqrt{\frac{Z}{n}}$, X étant une variable aléatoire normale centrée réduite et Z une variable aléatoire suivant une loi du khi-deux[3] à n degrés de liberté, X et Z étant de plus indépendantes.

[3] Une loi du khi-deux à n degrés de liberté est la loi suivie par la somme des carrés de n lois normales centrées réduites indépendantes

Dans le cas de l'estimation la variable X est l'estimateur de la moyenne $\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$ qui est bien une variable aléatoire normale centrée réduite, et la variable $Z=\frac{(n-1)S_n^2}{\sigma^2}$ qui suit une loi du khi-deux à n-1 degrés de libertés. Le nombre de degrés de libertés est *n*-1 car les *n* variables $Y_i - \bar{Y}_n$ sont liées par la relation $\sum_{i=1}^{n} Y_i - \bar{Y}_n = 0$; la forme quadratique $(n-1)S_n^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}_n\right)^2$ est donc de rang n-1, ce qui détermine le nombre de degré de liberté de la loi du khi-deux.

La distribution de la loi de Student à ν degrés de liberté est donnée par la formule :

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

où la loi $\Gamma(t) = \int_0^\infty e^{-u}u^{t-1}dt$ est la fonction Gamma. Remarquons que cette distribution peut être étendue aux valeurs non entières de ν.

2.6.Annexe 2 : Intervalle de confiance de la variance

Bien que moins utilisé que pour la moyenne, il est possible de déterminer un intervalle de confiance pour la variance d'une variable quantitative, si l'on fait l'hypothèse que cette variable suit une loi normale. Dans ce cas $Z=\frac{(n-1)S_n^2}{\sigma^2}$ suit une loi du khi-deux à n-1 degrés de libertés, en notant χ_1 le fractile d'ordre $\alpha/2$ de cette loi, et χ_2 le fractile d'ordre 1- $\alpha/2$, on a : $pr(\chi_1 < Z < \chi_2) = \alpha$, on en déduit l'intervalle de confiance pour σ^2 : $\left[\frac{(n-1)s_n^2}{\chi_2}; \frac{(n-1)s_n^2}{\chi_1}\right]$. Notons que cet intervalle n'est pas centré autour de l'estimation s_n^2, mais est centré en probabilité : c'est à dire que l'on élimine « autant » d'échantillons sous estimant la variance que d'échantillons surestimant cette variance. La notion de précision n'a donc pas ici le sens physique habituel comme pour la moyenne.

En prenant les racines carrées des bornes on en déduira un intervalle de confiance pour l'écart type.

3. EXERCICES ESTIMATION

3.1.: RadioLook

RadioLook est une radio privée émettant sur Grenoble et sa région depuis deux ans. Après un an de fonctionnement, une enquête faite auprès de 1200 grenoblois a donné les résultats suivants:

- 240 personnes ont déclaré écouter régulièrement la station
- parmi ces 240 personnes, 30 ont un statut d'étudiant.

Précisons que sur les 1200 personnes interrogées, 100 étaient des étudiants. Actuellement, la direction commerciale veut mener une enquête auprès des étudiants. Elle désire connaître de façon précise, la proportion d'étudiants écoutant régulièrement RADIO-LOOK et envisage donc un deuxième sondage.

1. Préciser la population, la variable de description et le paramètre faisant l'objet de l'étude.
2. Exploiter le sondage fait auprès de 1200 grenoblois pour obtenir une première estimation (ponctuelle et par intervalle) du paramètre défini en 1.
3. Combien de personnes faut-il interroger au cours de la seconde enquête, si le degré de confiance (ou seuil) retenu est de 0.95 et la précision (absolue) désirée 3%.
4. A l'issue du deuxième sondage, il a été constaté 368 auditeurs. Donner une estimation et un intervalle de confiance du paramètre faisant l'objet de l'étude (avec un degré de confiance de 0.95).
5. Peut-on affirmer que l'audience du segment étudiant a augmenté d'une enquête à l'autre?

3.2.La société ABC

La société ABC se propose de lancer un nouveau produit dans l'ensemble des 25000 magasins distribuant sa marque. Elle veut évaluer la capacité de production hebdomadaire nécessaire, pour cela elle a choisi un marché test de 400 magasins. Les résultats obtenus sur cet échantillon sont les suivants :

Moyenne des ventes par magasin et par semaine : 800 unités
Ecart-type estimé des ventes : 360 unités

1°) Donner une estimation ponctuelle, puis un intervalle de confiance à 0,95 du volume total espéré des ventes.

2°) Quelle taille d'échantillon aurait été nécessaire pour atteindre une précision de 200000 unités sur les ventes totales ?

3.3.Une foire au vin

Un magasin organise une foire au vin pour dynamiser ses ventes.
Avant la foire, la fréquentation moyenne s'établit à 1500 clients jour, avec un panier moyen de 23 articles (écart type 7) et un CA moyen par client de 175 € (écart-type 46 €) mesurés sur un échantillon de 200 clients.
Pendant la foire, la fréquentation moyenne s'établit à 2500 clients jour, avec un panier moyen de 18 articles (écart type 8) et un CA moyen par client de 130 € (écart-type 34 €) mesurés sur un échantillon de 200 clients.

Analysez ces résultats d'une part en terme de CA global jour, d'autre part en terme de panier moyen (nombre d'articles et CA par client). On procèdera à l'aide d'intervalles de confiance de 95%.

3.4.Une société d'études …

Une société d'études a proposé à un de ses clients une étude semi quantitative (mixte de question fermées et ouvertes) pour investiguer l'image de la marque auprès des quatre principaux segments de clientèle.

Sachant que les populations des segments sont respectivement :

- S1 540 000
- S2 310 000
- S3 115 000
- S4 430 000

La société d'études a proposé d'interroger :

- Dans le segment 1 un échantillon aléatoire de 540 clients.
- Dans le segment 2 un échantillon aléatoire de 310 clients.
- Dans le segment 3 un échantillon aléatoire de 115 clients.
- Dans le segment 4 un échantillon aléatoire de 430 clients.

Madame Renard, directrice du marketing, s'enquiert auprès de la société d'études de la pertinence de ce choix, et demande quelle précision on peut attendre de la mesure dans chacun des segments.

Pour permettre à la société d'études de lui répondre, elle lui fourni un ordre de grandeur du taux moyen de clients ayant une bonne image de la marque (c'est cette mesure qui l'intéresse), environ 65% (taux constaté lors de la précédente étude d'image de marque).

Elle indique qu'elle souhaite disposer d'une précision semblable dans chacun des segments.

1. Calculez à partir de ces informations la précision des mesures par segment.
2. Ce résultat correspond-il aux souhaits de madame Renard ?
3. Si ce n'est pas le cas quelle stratégie aurait fallu adopter pour répartir au mieux les 1400 questionnaires que madame Renard est prête à financer pour cette étude. Quelle précision obtient-on alors dans chaque segment ?
4. Pouvez-vous expliquez la démarche de la société d'étude ?. Savez vous comment s'appelle le type de sondage proposé par cette société ?

3.5.La société UVJM

La société UVJM a un *compte clients* composé de 7 000 clients. L'auditeur, chargé de la vérification du compte, désire estimer le montant moyen d'une créance à l'aide d'un sondage aléatoire. Le montant de la créance due par un client est le solde positif de son compte. Un échantillon constitué de 25 comptes a été prélevé parmi les 5 000 comptes ayant un solde positif. Chaque compte a été vérifié et son solde réévalué. Cette opération de révision comptable est donnée dans le tableau suivant :

Estimation

	N	Moyenne	Ecart type	Variance
Solde	25	164,820	63,7349	4062,132
N valide (listwise)	25			

1. Préciser la population, la variable de description et le paramètre faisant l'objet de l'étude.
2. Donner les estimations ponctuelles de la moyenne et de l'écart type du montant des créances
3. Etablir un intervalle de confiance de la moyenne des soldes positifs avec un niveau de confiance de 0.95.
4. Le niveau de confiance étant égal à 0.95, quelle taille d'échantillon faut-il envisager pour obtenir une précision de 8 € (la précision est égale à la demi-longueur de l'intervalle de confiance).
5. Un sondage complémentaire permettant d'obtenir un échantillon de taille égale à celle établie en 2 a été mené. Les résultats sont les suivants :

	N	Moyenne	Ecart type	Variance
Solde	219	156,5958	70,14788	4920,726
N valide (listwise)	219			

En fusionnant les deux échantillons, donnez une estimation du montant total des créances et un intervalle de confiance avec un niveau de confiance de 0.95.

3.6.La société de contrôle et de régulation (d'après J. Obadia)

La société de contrôle et régulation est une entreprise fabriquant des matériels électroniques en moyennes séries : appareils de contrôle, de régulation et de mesure. Elle travaille essentiellement sur catalogue et sur devis. L'auditeur responsable du contrôle de la comptabilité de l'entreprise a décidé d'effectuer un sondage pour déterminer la valeur réelle du stock des pièces détachées (petites pièces mécaniques, composantes électroniques, sous-ensembles achetées à l'extérieur, etc...).Ce stock fait l'objet d'un inventaire permanent assuré par l'ordinateur à partir des bordereaux d'entrée (livraison fournisseurs) et des bons de sortie émis par la production.

La diversité des articles constitutifs du stock des pièces détachées a conduit à distinguer :

- les articles de faible valeur regroupant essentiellement les petites pièces mécaniques dont le coût unitaire est inférieur à un euro.
- les articles de valeur moyenne qui regroupent l'essentiel des composants électroniques dont les coûts unitaires sont compris entre un et dix euros.
- les articles considérés comme coûteux et dont le coût unitaire dépasse dix euros et qui sont suivis un à un.

Ces trois catégories se trouvent dans des magasins différents et sont gérées séparément. L'ordinateur peut fournir à tout moment, une liste des valeurs stockées. Pour chaque référence, il est possible de disposer des informations suivantes:

- le numéro de la référence ou code - article : u
- le nombre d'articles N(u) comptabilisés dans le stock sous cette référence
- le coût unitaire auquel ces articles sont valorisés : C(u)
- la valeur stockée correspondante dite *valeur comptable*: Y(u) = N(u)*C(u)

Au jour du contrôle, les chiffres comptables relatifs aux trois catégories sont donnés par l'annexe 1. La catégorie des articles les plus coûteux, a été contrôlée en totalité; la première catégorie a été contrôlée à l'aide d'un sondage portant sur 100 références.

L'annexe 3 donne les résultats de ces deux contrôles. Le contrôle de la seconde catégorie doit être réalisé. Il s'agit donc d'estimer, pour cette catégorie, la valeur réelle du stock. Les erreurs sur les quantités et les coûts étant globalement prises en compte dans la valeur, on ne se préoccupera pas des quantités et des coûts unitaires séparément mais du produit des deux. Si l'estimation de la valeur constitue l'objectif principal du sondage, l'auditeur souhaite également déterminer la proportion des valeurs erronées.

Vous êtes chargé par l'auditeur d'établir un plan de sondage de la deuxième catégorie de pièces détachées.

Un plan de sondage doit indiquer :

- la population, les variables et les paramètres
- le nombre de références constituant l'échantillon
- le mode de sélection de ces unités
- comment, en utilisant les observations ou valeurs constatées faites sur les unités prélevées, établir les estimations des paramètres
- la précision du sondage

Pour établir ce plan de sondage vous disposez des informations fournies par un échantillon préliminaire concernant la variable X = "valeur réelle des références". L'analyse de cette information pourra se faire suivant les deux points ci-dessous.

Examen de l'information apportée par l'échantillon préliminaire sur la variable X = "valeur réelle des références"

1) Déduire une estimation de la valeur totale réelle du stock et la précision de cette estimation
2) On constatera que la précision obtenue n'est pas suffisante. Quelle est la taille de l'échantillon permettant d'obtenir une précision satisfaisante égale à 0,5% de la valeur comptable du stock. Conclusion.

Examen de l'information apportée par l'échantillon préliminaire sur la variable D = X - Y écart entre la valeur réelle et valeur comptable du stock.

1) Donner une estimation de l'écart entre valeur totale réelle et valeur totale comptable du stock. Quelle est la précision de cette estimation?
2) Utiliser les résultats du point a) pour calculer une estimation de la valeur totale réelle du stock et sa précision
3) Quelle est la taille de l'échantillon permettant d'obtenir la précision fixée au point 1.

Estimation

Annexe 1

Données comptables relatives aux trois catégories

Coûts Unitaires	*Nombre de références*	*Valeur totale en stock*
Moins de 1 €	2140	231843
De 1 à 10 €	1500	3366495
Plus de 10 €	180	625380
Total	3520	4223728

Annexe 2

Sondage préliminaire

Taille de l'échantillon : 50 références

Variable	*Moyenne*	*Variance*	*Ecart-type*
Val. Comptable	2315.83	604281	777.35
Val.Réelle	2304.1	568128	753.74
Ecart	-11.73	12170.1	110.32

Nombre de références pour lesquelles l'écart D = X - Y n'est pas nul : 6

Annexe 3

Résultats des contrôles des catégories 1 et 3

Catégorie d'articles de valeur élevée

Le contrôle complet des 180 références a montré que la valeur totale réelle était de 612 750 €.

Catégorie d'articles de faibles valeurs

Un sondage portant sur 100 références a donné les résultats suivants:

Valeur totale : 228 660 €

Précision du sondage :

- degré de confiance : 0.95
- seuil de précision : 4540 €

3.7.La société de contrôle et de régulation – Deuxième partie

Un deuxième sondage a permis de constituer un échantillon de 321 références. Ce deuxième échantillon a été fusionné avec l'échantillon préliminaire de taille 50 (cf. partie I) pour constituer un échantillon de 371 références et vous est donné dans le classeur CasSCR.xls.

Les résultats vous sont donnés dans le tableau suivant :

Statistiques descriptives

	N	Moyenne	Ecart type	Variance
Valeur Comptable (Y)	371	2225,76	767,631	589256,992
Valeur réelle (X)	371	2222,54	770,019	592928,536
Différence (D)	371	-3,23	74,734	5585,100
N valide (listwise)	371			

Le pourcentage d'erreurs est de 14%.

1. Utiliser les résultats de ce deuxième sondage pour obtenir une estimation de la valeur réelle des références de la deuxième catégorie. En déduire une estimation de la valeur réelle de tout le stock et la précision obtenue.
2. Pensez-vous que l'approximation normale soit justifiée pour la variable D=X-Y ? Justifiez économiquement ce fait.
3. Donner une estimation par intervalle du pourcentage d'erreur dans la seconde catégorie.

4. TESTS D'HYPOTHESE

4.1.Un exemple

Monsieur Dupond, directeur commercial d'une chaîne de magasins de distribution, veut tester un nouveau type de promotion sur les produits à forte fréquence d'achat, le client reçoit des coupons en fonction des achats effectués et du montant de la facture. D'ordinaire dans la chaîne de magasin le taux de retour des coupons est de 40% (c'est à dire que 40% des coupons distribués sont utilisés), le nouveau type sera considéré comme plus efficace si le taux de retour est supérieur à ce taux. Dans un magasin considéré comme représentatif de la chaîne, Monsieur Dupond installe son nouveau système, au terme de trois semaines d'essais sur 1000 coupons distribués 452 ont été réutilisés. Monsieur Dupond se demande si ce pourcentage (45,2%) est significatif d'une augmentation du taux de retour ou si la différence observée n'est imputable qu'aux incertitudes d'échantillonnage.

4.2.Généralités

Soit une variable X statistique définie sur une population P, et θ un paramètre lié à cette variable, nous appellerons hypothèse sur ce paramètre le fait de limiter les valeurs prises par ce paramètre à une partie non vide et non totale de l'ensemble des valeurs possibles noté A_0, le complémentaire de cet ensemble noté A_1 sera associée à l'hypothèse alternative. La première hypothèse est appelée hypothèse nulle.

Sur l'exemple précédent, la population est l'ensemble des coupons distribués pour les produits à forte fréquence d'achat, la variable X est la variable indicatrice de l'utilisation du coupon, le paramètre θ est le pourcentage de coupons utilisés. L'ensemble des valeurs possibles est l'intervalle [40%, 100%], puisque le directeur commercial n'envisage pas que sa méthode de distribution puisse être moins efficace que les autres méthodes. Une hypothèse ici serait par exemple que la nouvelle méthode ne soit pas plus efficace, c'est à dire que $\theta = \theta_0 = 40\%$ (ensemble noté $A_0 = \{40\%\}$), une autre hypothèse serait par exemple que la promotion personnalisée soit réellement plus efficace, c'est à dire que $\theta > \theta_0 = 40\%$ (ensemble noté $A_1 =]40\%;100\%]$).

Il arrive souvent que les ensembles associés aux hypothèses soient plus complexes que ceux présentés en exemple, nous le verrons plus loin lors des tests portant sur deux échantillons, ou lors de la régression par exemple.

L'objectif des tests d'hypothèse est de déterminer une règle de décision permettant de rejeter une hypothèse à partir de l'examen d'un échantillon. Comme nous l'avons vu au chapitre sur l'estimation, on ne peut pas prétendre prendre une telle décision sans risque d'erreur, ce risque est lié à la probabilité d'apparition d'échantillons exceptionnels (statistiquement aberrants).

Nous allons donc formaliser cette démarche. Nous noterons H_0 l'hypothèse $\theta \in A_0$, cette hypothèse est appelée hypothèse nulle, et H_1 l'hypothèse $\theta \in A_1$, appelée hypothèse alternative (nous reviendrons plus loin sur le choix de l'hypothèse nulle).

L'application d'une règle de décision peut conduire à l'un des quatre cas suivants :

		Etat Réel (Valeur de θ)	
		$\theta \in A_0$	$\theta \in A_1$
Choix (A partir de l'échantillon)	H_0	Pas d'erreur	Erreur de type II
	H_1	Erreur de type I	Pas d'erreur

A chaque erreur peut être associée une probabilité appelée risque :

- Le risque de première espèce noté α est la probabilité de l'erreur de type I c'est à dire le fait de choisir l'hypothèse H1, alors que le "vrai" paramètre appartient au sous-ensemble A_0 ; on dira plus simplement la probabilité du choix de H_1 alors que H_0 est vraie.
- Le risque de seconde espèce noté β est la probabilité de l'erreur de type II, c'est à dire le choix de H_0 alors que H_1 est vraie.

La définition d'une règle de décision se fait par la définition d'un ensemble $R \subset A_1$, appelé zone de rejet, tel que pour toute estimation du paramètre se trouvant dans cet ensemble on est conduit à rejeter l'hypothèse H_0, c'est à dire à accepter l'hypothèse H_1. La détermination de la zone de rejet se fait en fixant le risque de première espèce : le risque de première espèce est en effet défini à partir de cette région par : $prob(estimateur(paramètre) \in R / paramètre \in A_0)$.

Une autre façon de procéder est de déterminer la probabilité (appelée niveau de signification ou significativité du test) d'obtenir un échantillon conduisant au résultat observé (appelée *niveau de signification du test*), sous l'hypothèse H_0, si cette probabilité est inférieure au risque de première espèce, on rejettera alors l'hypothèse H_0. Ces deux procédures sont équivalentes, toutefois il est possible dans certains cas de définir la région de rejet avant même d'avoir procéder au sondage, ce qui bien sûr n'est pas possible pour le niveau de signification.

Remarquons que les hypothèses ne sont pas traitées de façon symétrique, on veut être assuré que l'hypothèse H_0 n'a qu'une probabilité très faible d'être vérifiée, donc, en fait, on cherche à se convaincre de l'hypothèse H_1. En général quand on rejettera H_0, on sera assuré d'avoir une faible probabilité de se tromper, en revanche, si on est conduit par le test à ne pas rejeter l'hypothèse nulle, il est possible que la probabilité de se tromper soit très grande, comme nous le verrons dans les cas traités dans ce chapitre.

4.3.Comparaison d'un pourcentage à un standard

Dans ce cas la variable est une variable indicatrice d'une caractéristique de la population, c'est à dire, en termes probabilistes, une variable de Bernouilli, le paramètre à estimer est l'espérance de cette variable, c'est à dire le pourcentage d'individus présentant la caractéristique dans la population. Dans tous les cas l'ensemble A_0 est réduit à un seul élément $\{p_0\}$, l'ensemble A_1 étant l'un des trois ensembles suivants

- $A_1 =]p_0;1]$ c'est à dire le test $H_0 : p = p_0$ contre $H_1 : p > p_0$, ce test est dit unilatéral à droite, la région de rejet est de la forme $R = [c;1]$ avec $c > p_0$: il faut que la valeur observée sur l'échantillon soit significativement supérieure à p_0 pour que

l'on soit convaincu de l'hypothèse H_1. C'est le cas de notre exemple avec p_0=40%.

- $A_1 = [0;p_0[$ c'est à dire le test H_0 : $p=p_0$ contre H_1 : $p<p_0$, ce test est dit unilatéral à gauche, la région de rejet est de la forme $R=[0;c]$ avec $c<p_0$: il faut que la valeur observée sur l'échantillon soit significativement inférieure à p_0 pour que l'on soit convaincu de l'hypothèse H_1.

- $A_1 = [0;p_0[\cup]p_0;1]$ c'est à dire le test H_0 : $p=p_0$ contre H_1 : $p\neq p_0$, ce test est dit bilatéral, la région de rejet est de la forme $R=[0;p_0-c[\cup]p_0+c]$ avec $c>0$: il faut que la valeur observée sur l'échantillon soit significativement différente de p_0 pour que l'on soit convaincu de l'hypothèse H_1. Dans ce cas il est d'usage de choisir la zone de rejet symétrique par rapport à p_0, comme l'est l'ensemble A_1, toutefois comme nous le verrons plus loin, un autre choix pourrait être fait.

Nous allons maintenant voir comment sont déterminées les valeurs critiques bornes ouvertes de la zone de rejet, pour cela revenons sur l'hypothèse H_0, et analysons les conséquences de cette hypothèse sur la loi de l'estimateur du pourcentage.

Loi de l'estimateur $\overline{X}_n$ sous l'hypothèse H_0

Sous l'hypothèse H_0 la loi de la variable X définie sur la population est parfaitement connue, c'est une loi de Bernouilli de paramètre p_0, valeur de p sous l'hypothèse retenue. Pour un échantillon de taille n, la loi de $\overline{X}_n$ peut donc en être déduite soit de façon exacte, pour les petites valeurs de n, soit de façon asymptotique pour les grandes valeurs de n.

De façon exacte, la variable $n\overline{X}_n$ somme de n variables de Bernouilli indépendantes suit une loi binomiale de paramètres n et p_0, on peut donc en déduire la loi de $\overline{X}_n$.

Pour les grandes valeurs de n, on pourra se contenter de l'approximation normale:

$\overline{X}_n \longrightarrow \boldsymbol{N}\left(p_0,\sqrt{p_0(1-p_0)/n}\right)$ (voir chapitre sur l'estimation).

Pour déterminer les régions de rejet de l'hypothèse, on éliminera les échantillons les plus improbables correspondant à des valeurs d'estimation dans le sous-ensemble, c'est à dire des échantillons donnant des valeurs exceptionnellement grandes dans le cas de test unilatéral à droite, exceptionnellement petites dans le cas de test unilatéral à gauche ou exceptionnellement éloignées de p_0 dans le cas de test bilatéral.

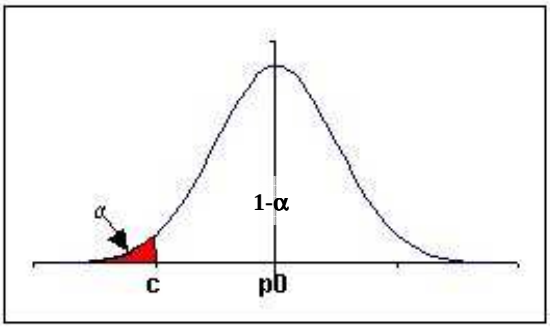
Remarquons que cette loi ne fait pas intervenir des résultats obtenus par sondage, il est donc possible ici de définir la zone de rejet avant même de procéder au sondage. C'est ce que nous allons faire pour les trois cas décrits plus hauts.

Tests unilatéraux

Nous traiterons simultanément les deux cas gauche et droite :

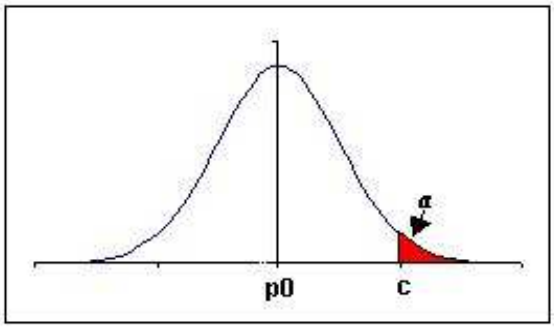
$H_0 : p=p_0$

$H_1 : p<p_0$



$H_0 : p=p_0$

$H_1 : p>p_0$



Cas des grands échantillons, approximation normale

Nous allons ici utiliser, la convergence de la loi de $\overline{X}_n$ vers la loi normale Comme la variable $Z = \dfrac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}}$ suit une loi normale standard (centrée réduite), il est facile de déterminer dans les deux cas la valeur critique c. Cette variable Z est appelée statistique associée au test.

Nous avons ici :

$\dfrac{c-p_0}{\sqrt{p_0(1-p_0)/n}}=z_\alpha(<0)$ donc

$\boxed{c=p_0+z_\alpha*\sqrt{p_0(1-p_0)/n}}$ qui est bien strictement inférieur à p_0.

On en déduit la règle suivante: si la valeur observée sur l'échantillon est inférieure à *c*, on rejettera l'hypothèse H_1 avec un risque d'erreur de α, on dira que la valeur observée est significativement inférieure à p_0 avec un risque inférieur à α.

Nous avons ici :

$\dfrac{c-p_0}{\sqrt{p_0(1-p_0)/n}}=z_{1-\alpha}(>0)$

$\boxed{c=p_0+z_{1-\alpha}*\sqrt{p_0(1-p_0)/n}}$ qui est bien strictement supérieur à p_0.

On en déduit la règle suivante: si la valeur observée sur l'échantillon est supérieure à *c*, on rejettera l'hypothèse H_1 avec un risque d'erreur de α, on dira que la valeur observée est significativement supérieure à p_0 avec un risque inférieur à α.

Niveau de signification du test

Comme nous l'avons signalé, une autre méthode consiste à déterminer le niveau de signification du test, c'est à dire la probabilité d'obtenir un échantillon conduisant à une valeur plus intérieure à l'ensemble A_1 que celle obtenue par sondage; cette valeur sera notée $\hat{p}$. Nous noterons *ns* ce niveau de signification, il représente le risque maximum que l'on prend en rejetant l'hypothèse H_0.

Pour le test unilatéral gauche, le niveau de signification est défini par :

$$ns = prob(\overline{X}_n < \hat{p}, \text{ sous } H_0)$$

ou encore en centrant et réduisant, et en prenant le complémentaire :

$$ns = prob\left(N(0,1) < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

Pour le test unilatéral gauche, le niveau de signification est défini par :

$$ns = prob(\overline{X}_n > \hat{p}, \text{ sous } H_0)$$

ou encore en centrant et réduisant, et en prenant le complémentaire :

$$1 - ns = prob\left(N(0,1) < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right)$$

La règle de décision est, dans tous les cas, la suivante : *si le niveau de signification est inférieur à a, on rejette l'hypothèse* H_0.

Courbe de puissance du test

Pour terminer nous allons nous intéresser au risque de seconde espèce β, ce risque dépend bien sûr de la valeur prise par le paramètre dans le sous-ensemble A_1, on a donc en fait une fonction de la valeur du paramètre *p*, plus le paramètre est loin de la valeur p_0, plus faible est le risque de seconde espèce, en revanche si la valeur de *p* est très proche de p_0, le risque de seconde espèce sera proche de 1-α, la vitesse de décroissance de la fonction en s'écartant de p_0 est donc un indicateur du pouvoir discriminant du test. (Les courbes présentées ci-dessous sont dans le fichier PropPuissance.xls)

Ici l'ensemble $A_1 = [0; p_0[$, traçons la courbe de puissance du test pour p_0=40% et n=100.

Pour une valeur donnée du risque de première espèce α, la valeur critique c est calculée.

Pour une valeur donnée de $p < p_0$, le risque de seconde espèce représente la probabilité de choisir à tort l'hypothèse H_0, c'est à dire que la valeur estimée de la proportion est supérieure à *c*. Si la proportion dans la population est *p*, $\overline{X}_n$ suit approximativement une loi normale $\boldsymbol{N}\left(p, \sqrt{p(1-p)/n}\right)$, le risque de seconde espèce est alors donné par :

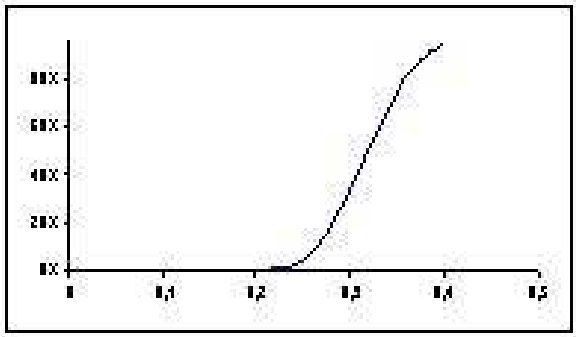
$$\beta = prob(\overline{X}_n > c) = prob\left(N(0,1) > \frac{c - p}{\sqrt{p(1-p)/n}} \right)$$

Ici l'ensemble $A_1 =]p_0; 1]$, traçons la courbe de puissance du test pour p_0=40% et n=100.

Pour une valeur donnée du risque de première espèce α, la valeur critique c est calculée.

Pour une valeur donnée de $p > p_0$, le risque de seconde espèce représente la probabilité de choisir à tort l'hypothèse H_0, c'est à dire que la valeur estimée de la proportion est inférieure à *c*. Si la proportion dans la population est *p*, $\overline{X}_n$ suit approximativement une loi normale $\boldsymbol{N}\left(p, \sqrt{p(1-p)/n}\right)$, le risque de seconde espèce est alors donné par :

$$\beta = prob(\overline{X}_n > c) = prob\left(N(0,1) < \frac{c - p}{\sqrt{p(1-p)/n}} \right)$$

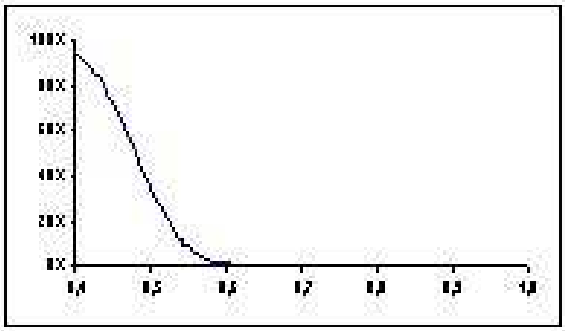


Remarque : le test

$$H_0 : p \geq p_0$$

$$\text{contre } H_1 : p < p_0$$

se ramène à ce test unilatéral



De même le test

$$H_0 : p \leq p_0$$

$$\text{contre } H_1 : p > p_0$$

se ramène à ce test unilatéral

Test bilatéral

Faire le test

$$H_0 : p = p_0$$

$$\text{contre } H_1 : p \neq p_0$$

au risque de première espèce α, revient à faire deux tests unilatéraux :

$H_0 : p = p_0$ $H_1 : p < p_0$ au risque α_1	et	$H_0 : p = p_0$ $H_1 : p > p_0$ au risque α_2

Avec $\alpha_1+\alpha_2=\alpha$, l'usage est de prendre $\alpha_1=\alpha_2=\alpha/2$.

La détermination des valeurs critiques c_1 et c_2 se fait comme nous l'avons vu précédemment, ces deux valeurs sont, avec la convention $\alpha_1=\alpha_2=\alpha/2$, symétriques par rapport à p_0. La règle de décision est alors la suivante :

Si sur l'échantillon la valeur du pourcentage observée est extérieure à l'intervalle $[c_1;c_2]$, *on rejettera l'hypothèse* H_0 *avec un risque d'erreur inférieur à* α, *sinon on conservera l'hypothèse* H_0 *mais sans connaître le risque d'erreur.*

Détermination du niveau de signification

La détermination du niveau de signification est particulière dans ce cas, elle ne peut se faire qu'avec la convention signalée, c'est à dire $\alpha_1=\alpha_2=\alpha/2$.

Soit $\hat{p}$ la valeur du pourcentage observé sur l'échantillon, dans le cas de test bilatéral, le niveau de signification est par définition :

$$\text{si } H_0 \text{ est vraie } prob\left(\left|\overline{X}_n - p_0\right| > \left|\hat{p} - p_0\right|\right),$$

c'est à dire la probabilité pour un échantillon tiré sous l'hypothèse H_0 de donner un écart (absolu) par rapport à la vraie valeur p_0 supérieur à l'écart (absolu) constaté lors du sondage.

Compte tenu de la symétrie de la loi normale, approximation de la loi de $\overline{X}_n$, le niveau de signification est donné par l'équation :

$$ns = 2 * prob\left(\overline{X}_n - p_0 > \left|\hat{p} - p_0\right|\right)$$

soit après centrage et réduction :

$$ns = 2 * prob\left(N(0,1) > \frac{\left|\hat{p} - p_0\right|}{\sqrt{p_0(1-p_0)/n}}\right) = 2 * \left(1 - prob\left(N(0,1) < \frac{\left|\hat{p} - p_0\right|}{\sqrt{p_0(1-p_0)/n}}\right)\right)$$

La règle de décision dans ce cas est toujours la même : *si le niveau de signification du test est inférieur à α, on rejette l'hypothèse H_0.*

Courbe de puissance du test

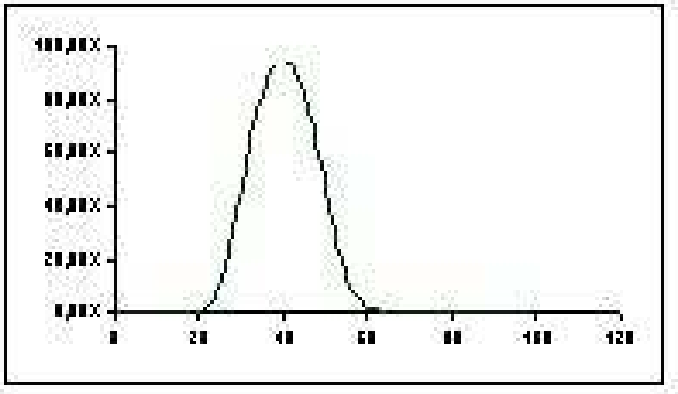
La courbe de puissance du test est symétrique par rapport à p_0, elle n'est pas exactement obtenue comme "recollement" des deux courbes définies précédemment pour les tests unilatéraux. Indiquons rapidement comment on peut construire cette courbe. Ici l'ensemble $A_1 = [0; p_0[\cup]p_0; 1]$, pour une valeur donnée du risque de première espèce α, les valeurs critique c_1 et c_2 sont calculées.

Pour une valeur donnée de $p \neq p_0$, le risque de seconde espèce représente la probabilité de choisir à tort l'hypothèse H_0, c'est à dire que la valeur estimée de la proportion est intérieure à l'intervalle $[c_1; c_2]$. Si la proportion dans la population est p, $\overline{X}_n$ suit approximativement une loi normale $\boldsymbol{N}\left(p, \sqrt{p(1-p)/n}\right)$, le risque de seconde espèce est alors donné par :

$$\beta = prob\left(c_1 \leq \overline{X}_n \leq c_2\right) = prob\left(\frac{c_1 - p}{\sqrt{p(1-p)/n}} \leq N(0,1) \leq \frac{c_2 - p}{\sqrt{p(1-p)/n}}\right) \text{ ou encore}$$

$$\beta = prob\left(N(0,1) \leq \frac{c_2 - p}{\sqrt{p(1-p)/n}}\right) - prob\left(N(0,1) \leq \frac{c_1 - p}{\sqrt{p(1-p)/n}}\right)$$

En utilisant cette définition, on obtient alors la courbe suivante avec $p_0 = 40\%$ et $n = 100$:



4.4.Application à notre exemple.

La population est l'ensemble des clients achetant le produit en promotion, la variable *X* est la variable booléenne indicatrice du renvoi du coupon. Le paramètre p est le pourcentage de coupons renvoyés, l'estimateur sur un échantillon de taille n est la moyenne c'est-à-dire la variable $\overline{X}_n$.

Ici l'hypothèse nulle est H_0:p_0=40%. La taille de l'échantillon est n=1000. L'hypothèse alternative sera :

$H_1 : p > p_0$

Analyse des erreurs.

L'erreur de première espèce consiste à penser que le pourcentage de coupons retournés a augmenté alors que ce n'est pas le cas, la décision associé sera de continuer à tort le nouveau système de coupons qui est sans doute plus onéreux..

L'erreur de seconde espèce sera de penser que le pourcentage n'a pas augmenter alors qu'en fait c'est le cas, donc de continuer l'ancien système, ce qui entrainera un manque à gagner éventuel (qui ne sera pas décelable toutefois).

C'est donc bien l'erreur de première espèce qu'il nous faut contrôler. Le pourcentage de retour doit être suffisamment supérieur à 40% pour que l'on accepte l'hypothèse H_1. C'est le risque de première espèce qui va nous aider à préciser ce "suffisamment", ou le degré de significativité qui va nous assurer d'avoir atteint un pourcentage suffisant.

Détermination a priori de la région critique.

Dans ce cas, il faut se fixer un risque de première espèce, par exemple 5%.

En appliquant la formule du paragraphe 3.2.1, nous obtenons :

$$c = p_0 + z_{1-\alpha} * \sqrt{p_0(1-p_0)/n} = 0{,}4 + 1{,}65 * \sqrt{0{,}4 * 0{,}6/1000} = 0{,}425 = 42{,}5\%$$

D'où la règle de décision suivante (avant tout sonadage) :

Règle : *Si dans un échantillon de taille 1000, on observe plus de 42,5% de retour de coupons, on conclura que la proportion de retour à augmenté et ce avec un risque inférieur à 5%.*

Comme le pourcentage observé est de 45,2%, on conclura que le pourcentage a augmenté.

Calcul de degré de significativité (après échantillonnage)

Après avoir réalisé le sondage, il est possible de déterminer la probabilité d'oberver un tel pourcentage pour un échantillon de taille 1000 sous l'hypothèse H_0, c'est le degré de significativité ou signifiance.

En utilisant la formule du paragraphe 3.2.2 :

$$1 - ns = prob\left(N(0,1) < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right) = prob\left(N(0,1) < \frac{0{,}452 - 0{,}4}{\sqrt{0{,}4 * 0{,}6/1000}} \right) = prob\left((N(0,1) < 3{,}36 \right)$$

En utilisant la table de la loi normale nous obtenons :

$1 - ns = 0{,}9996$ soit ns=0,0004 qui représente la probabilité de recueillir un tel échantillon sous l'hypothèse H_0. On conclura donc que le pourcentage a réellement augmenté.

4.5.Comparaison d'une moyenne à un standard

Un exemple

Monsieur Durlan, nouveau chef de produit chez Nesnone, envisage le lancement (dans les supermarchés) d'un nouveau petit déjeuner biologique. D'après le service économique le produit ne sera rentable que si les ventes moyennes hebdomadaires par magasin dépassent 320 unités. Monsieur Durlan a obtenu de 332 magasins qu'ils présentent ce nouveau produit, au bout de quatre semaines, il vient de recevoir les résultats. Quelle décision doit-il prendre ?

Avant de consulter les résultats de l'échantillon, formalisons sous forme de test d'hypothèse le problème de décision de Monsieur Durlan :

La population que l'on étudie est l'ensemble des supermarchés, la variable statistique est une variable numérique qui à chaque magasin associe les ventes hebdomadaires du produit. Le paramètre μ est la moyenne de ces ventes sur l'ensemble de la population.

Ce paramètre peut prendre des valeurs sur l'intervalle $[0,+\infty[$, ce qui intéresse M. Durlan c'est de placer le paramètre μ par rapport à la valeur (seuil de rentabilité) 320. Nous allons montrer sur cet exemple comment définir les hypothèses en fonction du contexte économique.

Nous avons deux hypothèses candidate au rôle de l'hypothèse H_1, hypothèse que l'on cherche à valider par le test puisque la région de rejet de H_0 est déterminée par le risque de première espèce a. Notons les H_A et H_B :

$$H_A \ : \ \mu > 320$$
$$H_B \ : \ \mu < 320$$

Analysons dans chacun des cas l'erreur de type I correspondant au choix de cette hypothèse comme hypothèse H_1 :

Cas A : Dans ce cas l'hypothèse $H_0 \ : \ \mu \leq 320$, l'erreur de type I (choix de H_1, alors que H_0 est "vraie") revient à croire que le produit va être rentable alors qu'en réalité il ne le sera pas, cette erreur conduira à une perte qui sera tangible, et facilement constatée par le supérieur hiérarchique de M. Durlan. En revanche l'erreur de type II conduirait à ne pas saisir l'opportunité de lancer un nouveau produit rentable, ce qui en fait ne pourra jamais être directement constaté. Poser le test ainsi revient à dire que l'on veut vraiment être convaincu de la rentabilité du produit (observer sur l'échantillon une valeur significativement plus grande que 320) pour se décider à le lancer.

Cas B : Dans ce cas l'hypothèse $H_0 \ : \ \mu \geq 320$, l'erreur de type I (choix de H_1, alors que H_0 est "vraie") revient à croire que le produit va n'est pas rentable alors qu'en réalité il le sera, cette erreur conduira à ne pas lancer le produit, ne sera pas constatée par le supérieur hiérarchique de M. Durlan, mais pourrait à long terme être catastrophique pour l'entreprise si ce type de produit prend une importance très grande sur le marché des petits déjeuners. En revanche l'erreur de type II conduirait lancer un produit non rentable et le risque associé ne sera pas maîtrisé. Poser le test ainsi revient à dire que l'on veut vraiment être convaincu de la non-rentabilitéé du produit (observer sur l'échantillon une valeur significativement plus petite que 320) pour se décider à ne pas le lancer.

Suivant l'importance stratégique du produit et la fragilité de la position de M. Durlan on sera conduit à privilégier l'une des deux approches. Comme ici M. Durlan est un jeune chef de produit, il ne veut pas commencer sa carrière par un lancement raté, il privilégiera le cas A, il voudra contrôler le risque associé à l'erreur constatable par son supérieur. La valeur du risque de première espèce dépend des conséquences économiques ou sociales de l'erreur, c'est un

arbitrage entre l'erreur de première espèce contrôlée et l'erreur de seconde espèce non contrôlée. Généralement il prend une des trois valeurs 10%, 5% ou 1%, plus sa valeur est faible, plus on laisse de "place" à l'erreur de seconde espèce.

Enfin comme dans le cas des proportions on peut toujours se ramener pour l'hypothèse nulle à une hypothèse simple du type :

$$H_0 \ : \ \mu = \mu_0$$

Notons enfin qu'il est d'usage en statistique de supposer que la variable quantitative étudiée est distribuée sur la population (munie d'une loi de probabilité équiprobable) suivant une loi normale.

Comme dans le cas d'une proportion nous traiterons les trois cas de tests possibles, mais plus succinctement dans la mesure ou seule les lois changent.

Statistique utilisée sous l'hypothèse H_0

Sous l'hypothèse H_0 la loi de la variable X définie sur la population est supposée normale de moyenne $\mu=\mu_0$ et d'écart type σ, nous supposerons cet écart type inconnu, le cas où il est connu est peu différent il suffit se supposer la taille de l'échantillon suffisante pour que la loi de Student se confonde avec la loi normale, ou que l'hypothèse de normalité puisse être abandonnée.

Comme pour l'estimation nous utiliserons la statistique, dont la loi est connue sous H_0:

$$T_n = \frac{\overline{Y}_n - \mu_0}{\sqrt{S_n^2 / n}} \xrightarrow{suit} \text{Loi Student à n - 1 degrés de liberté}$$

Pour déterminer les régions de rejet de l'hypothèse, on éliminera les échantillons les plus improbables correspondant à des valeurs d'estimation dans le sous-ensemble A_1, c'est à dire des échantillons donnant des valeurs exceptionnellement grandes dans le cas de test unilatéral à droite, exceptionnellement petites dans le cas de test unilatéral à gauche ou exceptionnellement éloignées de μ_0 dans le cas de test bilatéral.

Remarquons qu'ici cette loi fait intervenir des résultats obtenus par sondage, il est donc impossible ici de définir la zone de rejet avant même de procéder au sondage. Il nous est nécessaire d'avoir une estimation de l'écart type de la variable, en revanche l'estimation de la moyenne n'est nécessaire que pour l'application de la règle de décision.

Les résultats obtenus sur le sondage commandé par M. Durlan sont les suivants :

Taille de l'échantillon : **332**

Moyenne des ventes par magasin : 326

Ecart type des ventes : **51,82**

Sont notées en gras les valeurs qui nous serviront à construire la région de rejet.

Tests unilatéraux

Nous traiterons simultanément les deux cas gauche et droite :

$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

Cas de la loi normale, détermination exacte avec la loi de Student

En utilisant la variable T_n, définie plus haut, il est facile de déterminer la valeur de c à l'aide de la fonction LOI.STUDENT.INVERSE(probabilité; degrés de liberté) qui donne la plus valeur pour laquelle la variable suivant la loi de Student est supérieure en *valeur absolue* à cette valeur a une probabilité donnée, c'est à dire :

$prob(|T_n| > t_q^n) = q$, T_n désignant une variable suivant une loi de Student à n degrés de liberté.

Attention la fonction est toujours bilatérale, donc pour les tests unilatéraux il faudra mettre comme valeur de la probabilité le double du risque de première espèce.

Nous avons ici :

$\frac{c-\mu_0}{\hat{\sigma}/\sqrt{n}} = -t_{2\alpha}^{n-1}$ où $\hat{\sigma}$ est l'estimation de l'écart type de *X* donc

$\boxed{c = \mu_0 - t_{2\alpha}^{n-1} * \hat{\sigma}/\sqrt{n}}$ qui est bien strictement inférieur à μ_0.

On en déduit la règle suivante: si la valeur observée sur l'échantillon est inférieure à *c*, on rejettera l'hypothèse H_1 avec un risque d'erreur de α au maximum, on dira que la valeur observée est significativement inférieure à μ_0 avec un risque inférieur à α.

Nous avons ici :

$\frac{c-\mu_0}{\hat{\sigma}/\sqrt{n}} = t_{2\alpha}^{n-1}$, avec les mêmes notations

$\boxed{c = \mu_0 + t_{2\alpha}^{n-1} * \hat{\sigma}/\sqrt{n}}$ qui est bien strictement supérieur à μ_0.

On en déduit la règle suivante: si la valeur observée sur l'échantillon est supérieure à *c*, on rejettera l'hypothèse H_1 avec un risque d'erreur de α au maximum, on dira que la valeur observée est significativement supérieure à μ_0 avec un risque inférieur à α.

Niveau de signification du test

Comme nous l'avons signalé, une autre méthode consiste à déterminer le niveau de signification du test, c'est à dire la probabilité d'obtenir un échantillon conduisant à une valeur plus intérieure à l'ensemble A_1 que celle obtenue par sondage; valeur qui sera notée $\bar{x}_n$.

Nous noterons *ns* ce niveau de signification, il représente le risque maximum que l'on prend en rejetant l'hypothèse H_0.

Pour le test unilatéral gauche, le niveau de signification est défini par :

$$\boxed{ns = prob\left(\frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}} < \frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}, sous H_0\right)}$$

C'est à dire la valeur de la fonction de

Pour le test unilatéral gauche, le niveau de signification est défini par :

$$\boxed{ns = prob\left(\frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}} > \frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}, sous H_0\right)}$$

C'est à dire 1 - la valeur de la fonction de

répartition de la loi de Student à (n-1) degrés de liberté, pour la valeur (standardisée) :	répartition de la loi de Student à (n-1) degrés de liberté, pour la valeur (standardisée) :
$\frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}$	$\frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}$

La règle de décision est, dans tous les cas, la suivante : *si le niveau de signification est inférieur à α, on rejette l'hypothèse H_0.*

Test bilatéral

Faire le test

$$H_0 : \mu = \mu_0$$

$$\text{contre } H_1 : \mu \neq \mu_0$$

au risque de première espèce α, revient à faire deux tests unilatéraux :

$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$ au risque α_1	et	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ au risque α_2

Avec $\alpha_1+\alpha_2=\alpha$, l'usage est de prendre $\alpha_1=\alpha_2=\alpha/2$. Remarquons que dans le cas du test sur la moyenne cette convention et sans doute à l'origine des fonctions de Student généralement tabulées.

La détermination des valeurs critiques c_1 et c_2 se fait comme nous l'avons vu précédemment, ces deux valeurs sont, avec la convention $\alpha_1=\alpha_2=\alpha/2$, symétriques par rapport à μ_0. La règle de décision est alors la suivante :

Si sur l'échantillon la valeur du pourcentage observée est extérieure à l'intervalle $[c_1;c_2]$, on rejettera l'hypothèse H_0 avec un risque d'erreur inférieur à α, sinon on conservera l'hypothèse H_0 mais sans connaître le risque d'erreur.

Détermination du niveau de signification

La détermination du niveau de signification est particulière dans ce cas, elle ne peut se faire qu'avec la convention signalée, c'est à dire $\alpha_1=\alpha_2=\alpha/2$.

Soit $\bar{x}_n$ la valeur de la moyenne observée sur l'échantillon, dans le cas de test bilatéral, le niveau de signification est par définition :

$$\text{Sous l'hypothèse } H_0 \quad ns = prob\left(\left|\frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}}\right| < \left|\frac{\bar{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right),$$

c'est à dire la probabilité pour un échantillon tiré sous l'hypothèse H_0 de donner un écart (standardisé absolu) par rapport à la vraie valeur $\bar{x}_n$ supérieur à l'écart (standardisé absolu) constaté lors du sondage.

La règle de décision dans ce cas est toujours la même : *si le niveau de signification du test est inférieur à α, on rejette l'hypothèse H_0.*

Application à notre exemple

Le test a été posé au paragraphe 5.1.

Remarquons que nous ne pouvons pas, comme dans le cas de pourcentage, mettre en place notre règle de décision (numériquement) sans avoirs les résultats de l'échantillon.

Seuil de rejet de l'hypothèse H_0

En utilisant la formule du paragraphe 5.3.1, avec n=332 et un risque de première espèce de 0,10=10%

$$c = \mu_0 + t_{2\alpha}^{n-1} * \hat{\sigma}/\sqrt{n} = 320 + 1{,}649 * \hat{\sigma}/18{,}22 = 320 + 0{,}09 * \hat{\sigma}$$

D'où la règle de décision :

Règle : *si sur un échantillon de taille 332, on observe une valeur de la moyenne des ventes supérieure à 320+0,09*l'écart-type de l'échantillon, on pourra conclure que les ventes sont supérieures en moyenne à 320, avec une erreur de première espèce inférieure à 0,05.*

Ici l'écart-type observé sur l'échantillon est $\hat{\sigma} = 51{,}82$, le seuil critique est donc $c = 320 + 0{,}09 * 51{,}82 \approx 324{,}7$. Comme la moyenne observée est supérieure à cette valeur, on peut en conclure que les ventes moyennes sont bien supérieures à 320 et qu'il convient de lancer le nouveau produit.

Degré de significativité

Ce degré nous donne, rappelons le, la probabilité de tirer un échantillon ayant les caractéristiques observées, sous l'hypothèse H_0. Pour le calculer il nous suffit d'appliquer la formule du paragraphe 5.3.2 :

$$ns = prob\left(Sudent(331) > \frac{326 - 320}{51{,}82/\sqrt{332}} \right) = prob\left(Student(331) > 2{,}11\right) = 0{,}018$$

Il y a donc moins de "2 chances sur 100" d'observer un tel échantillon sous l'hypothèse H_0. On décidera donc de lancer le nouveau produit.

4.6.Comparaison de deux pourcentages

Reprenons l'exemple de Monsieur Dupond, il a conclu que sa nouvelle politique de distribution de coupons était plus efficace que l'ancienne. Il serait intéressé par savoir si le comportement des clients est différent suivant date d'achat : semaine ou week-end. Le détail de l'enquête est le suivant :

Semaine		**Week-End**	
Taille échantillon	600	Taille échantillon	400
Nbre de retours	264	Nbre de retours	188
Pourcentage	44%	Pourcentage	47%

Les pourcentages constatés sur l'échantillon sont évidemment différents (44% pour la semaine et 47% pour le week-end), mais cela peut être du aux aléas de l'échantillonnage et non pas à

un comportement différent entre la clientèle de semaine et la clientèle de week-end, ce que voudrait détecter M Martin.

Formalisation du problème

Nous pouvons ici présenter la formalisation de deux façons différentes, soit comme la comparaison de pourcentages sur deux populations, soit comme l'étude d'une liaison entre deux variables indicatrices définies sur une même population (cas particulier de la liaison de deux variables qualitatives présentée en annexe).

Formalisation sous forme de deux populations

La première population est l'ensemble des coupons distribués en semaine que nous noterons P_1, la seconde est l'ensemble des coupons distribués en week-end notée P_2. Sur chacune de ces populations nous définissons une variable indicatrice booléenne, notées respectivement X_1 et X_2, qui correspond au retour du coupon.

$$P_i \xrightarrow{X_i} \{0,1\} \quad \text{pour } i = 1,2$$

en désignant par p_1 et p_2 les pourcentages respectifs, c'est à dire les moyennes sur l'ensemble des variables X_1 et X_2 sur chacune des populations l'hypothèse nulle s'exprime alors sous la forme :

$$H_0 \quad p_1 = p_2$$

l'hypothèse alternative dans le cas de M Dupond est simplement la différence entre les deux valeurs (test bilatéral), mais pourrait être un pourcentage supérieur à l'autre (test unilatéral) :

$$H_1 \quad p_1 \neq p_2 \qquad \text{ou} \quad p_1 < p_2$$

Formalisation à l'aide de deux variables

Dans ce cas la population P unique est l'ensemble des coupons distribués, quelque soit le jour de la semaine, la variable X est toujours la variable indicatrice du retour ou non du coupon, et nous allons introduire une nouvelle variable indicatrice Y de la date de distribution du coupon : cette variable vaut 1 si le coupon est distribué en semaine et 0 s'il l'est le week-end. Le problème de M Dupond se résume à savoir si ces deux variables sont indépendantes, une fois la population munie d'une loi de probabilité uniforme.

En effet, le pourcentage p_1 représente la probabilité conditionnelle, pour que le coupon soit retourné sachant qu'il a été distribué en semaine, de même p_2 est la probabilité conditionnelle pour que le coupon soit retourné sachant qu'il a été distribué le week-end.

L'hypothèse H_0 revient alors à écrire :

$$p_1 = prob(X = 0/Y = 0) = prob(X = 0/Y = 1) = p_2$$

et comme X est une variable de Bernouilli (donc ne prenant que deux valeurs 0 et 1) on a aussi :

$$1 - p_1 = prob(X = 1/Y = 0) = prob(X = 1/Y = 1) = 1 - p_2$$

Ce qui est bien la définition de l'indépendance des deux variables.

L'hypothèse alternative dans le cas bilatéral est simplement la supposition d'une liaison entre les deux variables sans en indiquer le sens, le cas unilatéral étant l'existence d'une corrélation de signe donné.

Remarque : On retrouve aussi l'interprétation des deux hypothèses (nulle et alternative) sous la forme de moyenne, c'est à dire d'espérance en remarquant que p1 et p2 sont aussi les espérances conditionnelles de X sachant Y=0 ou Y=1; on peut aussi retrouver l'interprétation en terme de population en prenant respectivement les images réciproques $Y^{-1}(0) = P_1$ et $Y^{-1}(1) = P_2$.

Dans la suite nous utiliserons la formalisation en termes de deux populations, la deuxième formalisation sera généralisée aux variables qualitatives (du moins pour le test bilatéral) lors du test du Khi2 de contingence.

Statistique associée au test

L'hypothèse nulle peut aussi s'écrire

$$H_0 \quad p_1 - p_2 = 0$$

Sur un échantillon de taille n_1 tiré de la population P_1, le paramètre p_1 aura pour estimateur $\overline{X}^1_{n_1}$, de même pour un échantillon de taille n_2 tiré de la population P_2, l'estimateur du paramètre p_2 sera $\overline{X}^2_{n_2}$; la statistique utilisée sera donc la variable aléatoire $Z = \overline{X}^1_{n_1} - \overline{X}^2_{n_2}$.
Pour n_1 et n_2 suffisamment grands, nous connaissons une approximation normale des lois estimateurs, comme les échantillons sont tirés de façon indépendante dans chacune des populations nous connaissons la loi (approchée) de la variable Z :

$$Z \longrightarrow N(\mu, \sigma) \quad \text{avec } \mu = p_1 - p_2 \quad \text{et} \quad \sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

sous l'hypothèse H_0, en désignant par p la valeur commune de p_1 et p_2, nous aurons donc :

$$\mu = 0 \text{ et } \sigma^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Même si l'hypothèse H0 est vérifiée dans les populations, les estimations obtenues pour p_1 et p_2 seront différentes, quelle estimation devons nous considérer comme estimation commune? Dans la mesure ou l'estimateur du pourcentage est un estimateur convergent, plus la taille de l'échantillon est grande meilleure est la précision de l'estimation, la meilleure estimation sera donc obtenue en "regroupant" les deux échantillons en un seul échantillon de taille $n=n_1+n_2$ et cette estimation sera $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. C'est cette valeur que nous utiliserons comme pour calculer une approximation de l'écart type de la loi de la statistique Z.

Test bilatéral

Dans ce cas l'hypothèse alternative est $H_1 \quad p_1 \neq p_2$, comme pour le test contre un standard, nous éliminerons de l'hypothèse H_0, les échantillons conduisant (sous cette hypothèse) à un écart en valeur absolue entre les moyennes des échantillons trop improbable, c'est à dire dont la probabilité est inférieure au risque de première espèce fixé.

Détermination de la valeur critique

La valeur critique au-delà de laquelle on rejettera l'hypothèse H_0 est donc définie par la valeur c telle que :

$prob(|Z| > c / H_0) = \alpha$ soit encore en tenant compte de la symétrie de la loi normale $prob(Z < c / H_0) = 1 - \alpha/2$. La valeur critique c correspond donc au fractile d'ordre $1 - \alpha/2$ de la loi normale de moyenne 0 et d'écart type σ défini au paragraphe précédent. On peut bien évidemment se ramener au cas de la loi normale centrée réduite, en notant $z_{1-\alpha/2}$ le fractile de la loi normale centrée réduite, on a alors :

$c = z_{1-\alpha/2}\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ où *p* désigne la valeur commune de p_1 et p_2

Dans les applications la valeur *p* est bien sure inconnue, il n'est donc pas possible de déterminer la valeur critique avant de connaître les résultats du sondage ; on remplacera alors cette valeur par l'estimation $\hat{p}$ obtenue en "regroupant" les deux échantillons.

La règle de décision est alors la suivante, si sur les échantillons l'écart absolu observé est supérieur à *c*, alors l'hypothèse H_0 est rejetée au risque d'erreur α ; sinon on conservera l'hypothèse H_0 sans toutefois connaître le risque d'erreur.

Application :

En supposant que M. Martin veut simplement constater une différence entre la semaine et le week-end, nous allons faire un test bilatéral. Nous avons dans notre exemple :

$n_1 = 600$, $n_2 = 400$

$\hat{p}_1 = 0{,}44$, $\hat{p}_2 = 0{,}47$ donc $\hat{p} = (0{,}44 * 600 + 0{,}47 * 400)/1000 = 0{,}452$

D'où la valeur critique au risque de première espèce de 0,05=5%:

$c = 1{,}96 * \sqrt{0{,}452 * 0{,}548\left(\frac{1}{600}+\frac{1}{400}\right)} = 0{,}0630$.

Comme cette valeur est supérieure à la différence observée, nous ne pouvons pas rejeter l'hypothèse nulle, et nous en concluons que la différence observée est due aux aléas de l'échantillonnage.

Calcul du niveau de signification

Le niveau de signification est dans ce cas la probabilité, sous l'hypothèse H0, d'observer un écart entre les deux estimateurs qui soit en valeur absolu au moins égal à l'écart absolu observé sur les échantillons :

$$ns = prob(|Z| >= |\hat{p}_1 - \hat{p}_2|) = (1 - prob(Z < |\hat{p}_1 - \hat{p}_2|)) * 2$$

Puisque la loi normale suivie par *Z* est de moyenne nulle sous l'hypothèse H_0. En normalisant cette loi (c'est-à-dire en divisant par son écart-type), nous pouvons écrire :

$$ns = \left(1 - prob\left(N(0,1) < \left|\hat{p}_1 - \hat{p}_2\right| / \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)\right) * 2$$

Si ce niveau de signification est inférieur au risque de première espèce α, l'hypothèse H_0 est alors rejetée.

Application :

En supposant que M. Martin veut simplement constater une différence entre la semaine et le week-end, nous allons faire un test bilatéral. Nous avons dans notre exemple :

$n_1 = 600$, $n_2 = 400$

$\hat{p}_1 = 0{,}44$, $\hat{p}_2 = 0{,}47$ donc $\hat{p} = (0{,}44 * 600 + 0{,}47 * 400)/1000 = 0{,}452$

D'où le degré de significativité :

$ns = (1 - prob(N(0,1) < 0{,}03 / 0{,}0321)) * 2 = (1 - prob(N(0,1) < 0{,}934)) * 2 = 0{,}35$.

Ce qui signifie que dans 35% d'échantillons ainsi constitués, on pourrait observer une différence supérieure à 3% sous l'hypothèse nulle. Notre échantillon n'est pas assez "exceptionnel" pour que l'on puisse rejeter cette hypothèse. Nous considérerons donc qu'il n'y a pas de différence entre la semaine et le week-end.

Test unilatéral

Dans ce cas l'hypothèse alternative est $H_1 \quad p_1 > p_2$, il est inutile de distinguer ici le test droit du test gauche puisque cela revient simplement à changer les indices, comme pour le test contre un standard, nous éliminerons de l'hypothèse H_0, les échantillons conduisant (sous cette hypothèse) à un écart entre les moyennes des échantillons trop improbable, c'est à dire dont la probabilité est inférieure au risque de première espèce fixé.

Détermination de la valeur critique

La valeur critique au-delà de laquelle on rejettera l'hypothèse H_0 est donc définie par la valeur c telle que :

$prob(Z > c / H_0) = \alpha$ soit encore en prenant le complémentaire $prob(Z < c / H_0) = 1 - \alpha$. La valeur critique c correspond donc au fractile d'ordre $1 - \alpha$ de la loi normale de moyenne 0 et d'écart type σ défini au paragraphe précédent. On peut bien évidemment se ramener au cas de la loi normale centrée réduite, en notant $z_{1-\alpha}$ le fractile de la loi normale centrée réduite, on a alors :

$$c = z_{1-\alpha}\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$ où p désigne la valeur commune de p_1 et p_2.

Dans les applications la valeur p est bien sure inconnue, il n'est donc pas possible de déterminer la valeur critique avant de connaître les résultats du sondage ; on remplacera alors cette valeur par l'estimation $\hat{p}$ obtenue en "regroupant" les deux échantillons (voir plus haut).

La règle de décision est alors la suivante, si sur les échantillons l'écart observé ($\hat{p}_1 - \hat{p}_2$) est supérieur à *c*, alors l'hypothèse H_0 est rejetée au risque d'erreur α ; sinon on conservera l'hypothèse H_0 sans toutefois connaître le risque d'erreur.

Calcul du niveau de signification

Le niveau de signification est dans ce cas la probabilité, sous l'hypothèse H0, d'observer un écart entre les deux estimateurs qui soit en valeur absolu au moins égal à l'écart absolu observé sur les échantillons :

$$ns = prob(Z >= \hat{p}_1 - \hat{p}_2) = (1 - prob(Z < \hat{p}_1 - \hat{p}_2))$$

Ou encore en utilisant la loi normale centrée réduite, ici il suffit simplement de réduire, puisque sous l'hypothèse H_0, la loi de *Z* est déjà centrée :

$$ns = 1 - prob\left(N(0,1) < \frac{\hat{p}_1 - \hat{p}_2}{\sigma} \right) \quad \text{avec} \quad \sigma = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

p étant la valeur commune de p_1 et p_2, sous l'hypothèse H0 ; cette valeur est inconnue est sera bien entendu remplacée par son estimation dans les applications numériques.

Si ce niveau de signification est inférieur au risque de première espèce α, l'hypothèse H_0 est alors rejetée.

5. EXERCICES SUR LES TESTS D'HYPOTHESE

Sauf indication contraire, on prendra pour tous les exercices pour risque de première espèce les deux valeurs 5% et 1%.

5.1.Taux de phosphate

Un fabricant de lessive affirme que le taux de phosphates contenu dans les lessives de sa marque est inférieur à 6% du poids total. Un institut de consommation a fait analyser un échantillon de 150 paquets dont les résultats sont donnés dans le fichier "phosphates.sav", dont l'analyse vous est donnée dans le tableau suivant :

	N	Moyenne	Ecart type
Taux	150	5,89%	1,03%
N valide (listwise)	150		

Questions

1. Définissez la population, la variable et le paramètre concernés par l'analyse.
2. Formulez sous forme de test le problème de l'institut de consommation.
3. Quelle conclusion tirez-vous de l'analyse de l'échantillon?

5.2.AntiSmoke

Un laboratoire pharmaceutique envisage de lancer sur le marché un nouveau "patch" anti-tabac "Antismoke", que s'il assure au moins 25% de réussite, c'est à dire qu'au moins 25% des utilisateurs ne doivent pas recommencer à fumer après un mois de traitement.

Des essais ont été faits sur un panel de 100 fumeurs et les résultats sont donnés dans le fichier "tabac.sav", la reprise=1 indique que le fumeur a rechuté avant la fin du mois sinon il est indiqué 0.

Reprise

Sexe	Moyenne	N	Ecart-type
F	66%	41	48%
H	71%	59	45,7%
Total	69%	100	46,5%

Questions

1. Définissez la population, la variable et le paramètre concernés par l'analyse.
2. Formulez le test du laboratoire
3. Le laboratoire doit-il lancer son produit?
4. Peut-on faire une différence sur l'efficacité du médicament selon le sexe de la personne?

5.3.Le groupe de presse AES

Le groupe de presse AES (Avenir et Société) est spécialisé dans l'édition de livres et de revues scientifiques. L'une de ces revues Sciences du Futur, est diffusée exclusivement par abonnement. La direction commerciale désire prospecter le segment de clientèle des professions médicales par des offres d'abonnement à des tarifs préférentiels. Pour cela elle envisage d'acquérir le fichier des abonnés de la revue médicale CADUCOR.

CADUCOR annonce que l'expérience passée montre qu'entre 8 à 12 % environ des médecins du fichier répondent positivement aux offres qui leur sont faites par correspondance (abonnements, livres, objets etc...). Après un calcul de rentabilité, AES estime que le fichier peut se révéler intéressant s'il présente un taux de réponse supérieur à 10%.

Questions

1. Préciser la population, la variable de description et le paramètre faisant l'objet de l'étude.
2. Formuler le problème sous forme d'un test. Donner la forme générale de la région de rejet de l'hypothèse H_0. Donner une interprétation des deux types d'erreur.
3. AES désire contrôler l'erreur de type I en fixant le risque associé à a = 0.05. Préciser la région de rejet du test si la taille de l'échantillon retenue est de 400.
4. Une proposition d'abonnement a été envoyée à 400 médecins; 58 d'entre eux ont répondu favorablement.

   D'après ce résultat AES doit-il acheter le fichier CADUCOR ?

5.4.Contrôle de qualité

Un fabricant de coque de téléphones portables veut tester la solidité de sa fabrication, effectuée sur deux machines. Il prélève 50 éléments au hasard sur la chaîne de fabrication et les soumet à un essai de chocs. Une machine frappe sur la coque jusqu'à rupture de celle ci ; un bon modèle doit résister à plus de 260 chocs.

Les données résultant du test vous sont fournies dans le fichier "quali.sav" :

	N	Moyenne	Ecart type	Variance
Nombre de chocs	50	267,16	24,408	595,770
N valide (listwise)	50			

Questions

1. Définissez la population, la variable et le paramètre concernés par l'analyse.
2. Formulez le test du fabricant
3. Le produit vous paraît satisfaisant au point de vue résistance?

5.5.Rola-Cola contre Moka-Cola

Monsieur Poulain responsable des études du service Marketing de Rola-Cola vient de recevoir les résultats d'un test de goût dont l'objectif est de déterminer laquelle des deux marques - Rola-Cola ou Moka-Cola - était préférée des consommateurs de boisson à base de cola . Rappelons que Moka-Cola est le principal concurrent de Rola-Cola.

Pour cela 200 consommateurs de boisson à base de cola furent sélectionnés pour participer à un test de goût dit "en aveugle". Chaque participant fut invité à goûter les deux boissons servies dans des verres "anonymes" marqués respectivement des seules lettres A et B. Les marques d'origine des deux boissons étaient donc cachées au participant mais connues des organisateurs.

Questions

1. Sachant que sur 200 participants, 112 ont déclaré préférer Rola-Cola faire un test statistique permettant de rejeter ou d'accepter l'hypothèse que la boisson Rola-Cola est préférée à Moka-Cola. Après avoir formulé les deux hypothèses du test en, on précisera la région de rejet et la procédure permettant de conclure. On prendra un risque de type I de 0.05.
2. Pour éviter que l'ordre dans lequel les deux boissons furent présentées n'affecte les préférences émises, les participants furent partagés en deux groupes égaux; le premier goûta Rola-Cola avant Moka-Cola et le second opéra en sens inverse. Les résultats obtenus furent les suivants :

	Groupe1	Groupe2
	Rola –Cola avant Moka-Cola	Moka-Cola Avant Rola-Cola
Nombre de participants	100	100
Nombre de participants préférant Rola-Cola	54	58

Ces résultats permettent-ils de retenir l'hypothèse que l'ordre de présentation des deux boissons n'a effectivement aucune influence sur les préférences déclarées pour Rola-Cola ?

5.6.La société SVC

La société SVC vend par correspondance des CD-Audio. Pour cela elle procède par publipostage dans lequel on trouve une description du CD proposé, accompagnée d'une offre promotionnelle (remise ou cadeau en cas d'achat). Le publipostage est envoyé aux 120000 personnes figurant dans le fichier clients de la société.

En 1996, la cinquième symphonie de Beethoven fût proposée avec une remise de 10 % en cas d'achat sous huitaine une fois reçu le publipostage. Elle fût vendue à 18 000 exemplaires.

La direction Marketing désire renouveler l'opération avec la neuvième symphonie de Beethoven. Elle hésite entre deux formules :

La formule F1 déjà utilisée pour promouvoir la cinquième symphonie.

La formule F2 offrant un mini dictionnaire de termes musicaux en cas d'achat.

Il a été décidé de tester ces deux formules en recourant à deux sondages dans le fichier des 120 000 clients : la formule F1 étant proposée à un premier échantillon et la formule F2 à un second différent du premier. L'objectif des ces deux sondages est d'estimer la proportion d'acheteurs suivant chacune des deux formules avec un seuil de précision de 1% [4]. La taille retenue pour chaque échantillon est de 4 900.

[4] Le seuil de précision est la demi-longueur de l'intervalle de confiance. Il s'agit d'un seuil de précision absolue.

Les deux sondages ont donné les résultats suivants :

	Formule F1	Formule F2
Nombre d'acheteurs	801	914

Questions

1. Vérifier que la taille de l'échantillon retenue correspond bien à l'objectif de précision de 1%.
2. La direction marketing en se fondant sur les résultats du tableau 1 pense que la neuvième symphonie pourrait se vendre à un nombre d'exemplaires supérieur à celui de la cinquième. Confirmer ou infirmer cette hypothèse.
3. Des deux formules F1 ou F2 laquelle faut-il retenir ?
4. Donner les nombres minimum et maximum de CD de la neuvième susceptibles d'être vendus.

Remarque : pour traiter ces questions on utilisera

un degré de confiance de 0.95

un risque de type I égal à 0.05

5.7.Télémara

L'hebdomadaire Télémara souhaite effectuer une opération de recrutement sur fichier externe. A cet effet, madame Beller, responsable des abonnements, décide de contacter plusieurs fournisseurs et de réaliser des tests sur les fichiers proposés avant de choisir ceux qu'elle va acheter.

Les trois fournisseurs contactés sont :

- Un opérateur de câble : 670 000 abonnés.
- Une méga base de consommation, sous segment de foyers regardant la télévision au moins deux heures par jour : 450 000 foyers.
- Un fichier client de Vépéciste sous segment des clients ayant acheté dans les 18 derniers mois une télévision ou un magnétoscope 320 000 clients.

Le coût d'envoi du message est de 1 €, le prix de location de l'adresse de 0,4 €, la marge sur abonnement peut être estimée à 28 €. Malgré le bénéfice secondaire apporté par l'augmentation d'audience (impact sur le revenu publicitaire) et un taux de renouvellement d'abonnement d'environ 50%, madame Beller estime qu'il lui faut financer ses coûts de recrutement sur la première année, et conserver une marge nette d'au moins 4 €.

Madame Beller propose à chacun des fournisseurs un test sur 5 000 adresses.
Les résultats obtenus sont les suivants :

- Câble : 350 abonnements
- Méga base : 330 abonnements
- VPC : 260 abonnements

Questions

1. Compte tenu de ces informations quel est le taux d'abonnement minimum qui doit être observé sur les fichiers achetés ?

2. Quels fichiers, madame Beller peut-elle acheter, en pouvant affirmer, avec un risque de 5%, que le taux d'abonnement dépassera le minimum fixé.
3. Madame Beller peut-elle dire, au risque 5%, que le fichier du câble est meilleur que celui de la méga base ?
4. Pour les fichiers sélectionnés, pouvez vous donner un intervalle de confiance à 95% de la marge globale attendue lors de la généralisation.

5.8.La société Votre Santé

La société *Votre Santé* est une entreprise de vente par correspondance de produits de beauté dits « *naturels* ». Elle gère un fichier de 350 000 clients et propose chaque mois une offre promotionnelle accompagnée d'un cadeau. Le taux de réponse à cette offre est généralement de 15%, la marge moyenne par réponse de 68 €. Mlle C. Claire, nouvellement en charge de ce fichier, a retenu comme cadeau un abonnement gratuit de six mois, au mensuel « *Votre beauté Madame* ». Elle pense que cela pourrait augmenter le taux de réponse à la prochaine offre ; toutefois cette proposition ne serait rentable que si le taux de réponse dépassait les 17,5% (avec la même marge moyenne évidemment). Elle envisage de tester la réalité de ces hypothèses sur un échantillon de clientes. La précision voulue pour son estimation est de l'ordre de 2%.

Questions

1. Quelle taille d'échantillon doit-elle choisir afin d'atteindre la précision voulue (avec un degré de confiance de 0,95) ?
2. Les résultats d'un sondage sur un échantillon de 1225 clientes vous sont donnés en annexe 1.
3. Donner une estimation par intervalle au degré de confiance 0,95 du pourcentage π de réponses positives attendu à l'offre.
4. Mlle C. Claire se propose de procéder au test d'hypothèse suivant :

   $$\mathbf{H_0\ \pi <= 17{,}5\%}$$

   $$\mathbf{H_1\ \pi > 17{,}5\%}$$

   Expliquer pourquoi elle envisage ce test. Indiquer et déterminer la région de rejet associé à ce test (risque de type I égal à 0,05). Que concluez-vous ?
5. Mlle C. Claire pense que les nouveaux clients (inscrits depuis moins de 6 mois) ont un taux de réponse supérieur aux anciens. Confirmer ou infirmer cette hypothèse.
6. Il s'agit dans cette question de déterminer un intervalle de confiance au degré de confiance 0,95 de la marge de la campagne promotionnelle.

   Peut-on considérer que la marge moyenne attendue de cette campagne sera la même que pour les campagnes précédentes. On posera cette alternative sous forme de test et on prendra un risque de première espèce de 0,05

   En déduire une estimation par intervalle de la marge totale attendue.

Annexe 1 Résultats du sondage

Taille de l'échantillon : 1225 individus

	Total	Anciens Clients
Nombre d'individus	1225	850
Nombre de réponses	258	193

Résultats sur la marge

Marge totale	Marge Moyenne	Ecart-type de la marge
17028€	66 €	33 €

5.9.La société Bricoplus

La société Bricoplus a lancé pendant un mois une campagne publicitaire avec bons de réduction dans la presse régionale. Le montant moyen d'une commande avant la campagne était de 60 €. Le coût de la campagne a été de 200 K€. A la fin du mois elle a reçu 20000 commandes (avec ou sans bon de réduction). Avant de traiter l'ensemble des commandes, la société voudrait avoir une estimation du succès de cette campagne. Pour cela elle étudie un échantillon de 900 commandes prises au hasard. Les résultats de cet échantillon sont donnés dans le tableau suivant :

Origine	Avec Bon	Sans Bon	Total
Nombre	473	427	900
Valeur moyenne			64 €
Ecart-type(Valeur)			40 €

Questions

1°) Peut-on considérer qu'il y a autant de commandes provenant de la campagne publicitaire (avec bon de réduction) que de commandes "ordinaires" (sans bon de réduction) ? (On prendra un risque de première espèce de 0,05)

2°) Le montant moyen des commandes a-t-il augmenté avec la campagne ? (On prendra un risque de première espèce de 0,05)

3°) Donner une estimation ponctuelle et un intervalle de confiance à 0,95 du chiffre d'affaires du mois.

4°) Le directeur financier doute de la performance de cette campagne en termes de rentabilité, il envisage même une diminution de profit. Sachant que le Chiffre d'affaires mensuel avant la campagne était d'environ 900000 € et que le taux de marge par produit est de 50%, poser sous forme de test la conjecture du directeur financier. Qu'en concluez-vous ?

5.10. Une enquête de satisfaction

Une enquête de satisfaction sur les utilisateurs d'une voiture urbaine a montré que sur 1000 personnes interrogées 640 se déclarait satisfaits du service après vente du constructeur.

Donner un intervalle de confiance au degré de confiance 0,95 du pourcentage de personnes satisfaites

Peut-on considérer que plus de 60% des utilisateurs de ce service après vente sont satisfaits.

La répartition des personnes satisfaites par tranche d'âge est la suivante :

Tranche d'âge	18-35 ans	Plus de 35 ans
Nombre de personnes interrogées	600	400
Satisfaits	350	290

Question

Peut-on conclure que chez les moins de 35 ans le taux de satisfaction est significativement plus élevé que chez les plus de 35 ans(on prendra un risque de première espèce de 0,05) ?

5.11. Exercice 11 : La Société Sogec (d'après J. Obadia)

La Société SOGEC, filiale de la banque HERVA est spécialisée dans le crédit à la consommation. En 1998, le montant des crédits accordés à ses clients était de 2 4120 000 F et la provision pour créances douteuses estimée à 1 206 000 F. Jusqu'en 1997, cette provision était calculée après un examen exhaustif de tous les comptes clients, permettant de mettre en évidence les créances douteuses (une créance étant déclarée douteuse lorsqu'il est constaté deux échéances non payées sur les quatre dernières dues).

En 1998, le chef comptable abandonne cette procédure, présentant l'argument suivant : « *Lorsque l'on examine les données des dix dernières années, on constate que la proportion de créances douteuses varie, suivant les années entre 3% et 6%. Aussi afin d'éviter un travail long et fastidieux à mon service (3 employés mobilisés pendant 45 jours), il est préférable d'estimer la proportion de créances douteuses à 5% et d'appliquer ce taux au montant global des crédits accordés pendant l'année. Cela suppose bien sûr que la valeur moyenne des créances douteuses soit égale à la valeur moyenne de l'ensemble des créances. Ce qui a été le cas ces dernières années* ».

M. Allais, chargé par la maison mère du contrôle des données comptables de la Société SOGEC, demande à M. Salmain de réaliser un sondage. Ce sondage devrait permettre, après examen d'un échantillon de comptes clients, de vérifier les deux hypothèses sur lesquelles repose la procédure adoptée par le chef comptable. M. Salmain considéra que l'estimation du pourcentage des créances douteuses établie à partir de ce sondage n'était pas suffisamment précise (avec un degré de confiance de 0.95). Il procéda à un autre sondage, permettant d'obtenir une précision de l'ordre de 4% (toujours avec un degré de confiance de 0.95). Les résultats de ce deuxième sondage sont donnés en annexe. M. Salmain avait en main tous les éléments pour estimer la valeur des créances douteuses.

1 Lorsqu'il présente la nouvelle procédure qu'il a adoptée, le chef comptable précise :« Cela suppose bien sûr que la valeur moyenne des créances douteuses soit égale à la valeur moyenne de l'ensemble des créances **».** Expliquez pourquoi ?

2 **Examen des résultats du premier sondage**

2.1 Le premier sondage permet d'établir une estimation de π proportion des créances douteuses. Donner cette estimation. Quelle est la précision ε obtenue si l'on adopte un degré de confiance α égal à 0.95 ?

2.2 En déduire un intervalle de confiance. M. Salmain considère l'estimation des pourcentages des créances douteuses peu précise. Pourquoi ?

3 Examen des résultats du second sondage

3.1 La taille de l'échantillon retenue est de **323**. Justifier ce choix.

3.2 Donner la région de rejet de l'hypothèse du chef comptable concernant la proportion π de créances douteuses :

$$H_0 : \pi \leq 0.05$$
$$H_1 : \pi > 0.05$$

Le risque de type I , α , est fixé à 0.05.

3.3 Quelle conclusion concernant la valeur de π retenue par le chef comptable faut-il adopter ?

3.4 Etablir un intervalle de confiance du paramètre μ_d, moyenne des créances douteuses.

3.5 Tester l'hypothèse du chef comptable concernant la valeur moyenne μ_d des créances douteuses pour l'année 1992 :

$H_0 : \mu_d = 402$

Justifier la formulation de l'hypothèse H_0. Préciser l'hypothèse H_1. Conclusion ? (le risque de premier type I α fixé à 0.05).

3.6 Etablir un intervalle de confiance du paramètre π (degré de confiance α égal à 0.95).

3.7 Déduire des questions 5 et 6, une estimation de la valeur totale des créances douteuses. Quelle est la précision obtenue ? En déduire un intervalle de confiance. (degré de confiance α égal à 0.95).

Annexe

Résultats du premier sondage

Taille de la population sondée	60 000
Nombre de créances examinées	50
Nombre de créances douteuses dans l'échantillon	8

Résultats du deuxième sondage

Taille de la population sondée	60 000
Nombre de créances examinées	323
Nombre de créances douteuses dans l'échantillon	43
Valeur moyenne des créances douteuses dans l'échantillon	408
Estimation de l'écart-type de la valeur des créances douteuses	92

NB : Pour réaliser le second sondage, il a été tenu compte des cinquante créances examinées au cours du premier sondage.

6. ANNEXE : TEST DU KHI-DEUX

Nous allons présenter ici le test du Khi-deux étant donné son importance en marketing, bien qu'il ne soit pas au programme du cours.

Le test de contingence du Khi deux a pour objectif de mettre en évidence un lien éventuel entre deux variables qualitatives. Nous allons l'illustrer sur un exemple : le fabricant de shampoing DIP, veut déterminer quels sont les critères de choix d'un shampoing suivant les catégories d'âges, de façon plus précise il veut savoir si ces critères différent suivant les tranches d'âges. Après une enquête auprès d'un échantillon de 535 consommateurs, il a été constitué un fichier de données où sont relevés le principal critère de choix, l'âge et le lieu d'achat habituel du consommateur.

6.1.Formalisation du problème

La population *E* est constituée de l'ensemble des consommateurs de shampoing, sur cette population sont définies plusieurs variables qualitatives, dont les deux variables qui nous intéressent notées X et Y concernant le choix et la tranche d'âge.

La variable "choix" est une variable qualitative à $m = 4$ modalités notées a_i pour $1 \le i \le m$:

$$E \xrightarrow{X} \{ \textit{distribution, marque, odeur, texture} \}.$$

La variable "âge" est une variable qualitative à $p = 3$ modalités notées b_j pour $1 \le j \le p$:

$$E \xrightarrow{Y} \{ < 25, 25 - 45, > 45 \}$$

L'hypothèse nulle, que l'on cherche à rejeter est l'indépendance des deux variables, l'hypothèse alternative est la liaison entre les deux variables sans toutefois préciser de quel type est cette liaison.

L'hypothèse nulle peut se formuler de la façon suivante :

$$H_0 \quad \forall i \in [1,m] \forall j \in [1,p] \quad prob(X = a_i, Y = b_j) = prob(X = a_i) * prob(Y = b_j)$$

Les probabilités correspondent aux fréquences observées sur la population toute entière, puisque la loi mise pour l'échantillonnage équiprobable est la loi uniforme.

6.2.Tableaux croisés ou de contingence (observé et théorique)

Sur un échantillon de taille n, nous utiliserons les notations suivantes :

n_{ij} désigne le nombre d'individus de l'échantillon possédant la modalité a_i pour la variable *X* et la modalité b_j pour la variable *Y*. $\frac{n_{ij}}{n}$ est donc l'estimation de $prob(X = a_i, Y = b_j)$.

$n_{\bullet j} = \sum_{i=1}^{m} n_{ij}$ désigne le nombre d'individus de l'échantillon la modalité b_j pour la variable *Y*.

$\frac{n_{\bullet j}}{n}$ est donc l'estimation de $prob(Y = b_j)$.

$n_{i\bullet} = \sum_{j=1}^{p} n_{ij}$ désigne le nombre d'individus de l'échantillon la modalité a_i pour la variable X

$\frac{n_{i\bullet}}{n}$ est donc l'estimation de $prob(X = a_i)$.

On regroupe ces éléments dans un tableau, appelé tableau croisé ou tableau de contingence des deux variables, les éléments $n_{\bullet j}$ et $n_{i\bullet}$ s'appellent les marges du tableau. On a donc la présentation suivante :

Y / X		b_j		Total
	…..		…..	
a_i	…..	n_{ij}	…..	$n_{i\bullet}$
	…..		…..	
Total		$n_{\bullet j}$		n

Sous l'hypothèse H_0, on peut construire le tableau théorique que l'on devrait obtenir si l'indépendance était parfaitement respectée sur l'échantillon ; on suppose que l'échantillon parfait a les mêmes marges que l'échantillon observé. Nous noterons e_{ij} les effectifs théoriques correspondant à l'indépendance. Nous aurons alors les relations suivantes :

$$\forall i \in [1,m] \forall j \in [1,p] \quad \frac{e_{ij}}{n} = \frac{n_{i\bullet}}{n} * \frac{n_{\bullet j}}{n} \quad \text{soit} \quad e_{ij} = \frac{n_{i\bullet} * n_{\bullet j}}{n}$$

On pourra donc construire le tableau théorique correspondant à l'hypothèse H_0 :

Y / X		b_j		Total
	…..		…..	
a_i	…..	e_{ij}	…..	$n_{i\bullet}$
	…..		…..	
Total		$n_{\bullet j}$		n

Seules les cellules grisées différent du tableau de contingence observé sur l'échantillon, si ces deux tableaux sont suffisamment différents nous rejetterons l'hypothèse H_0. Il nous faut donc définir une distance entre tableau et connaître la loi de cette distance sous l'hypothèse nulle, pour appliquer la même démarche que dans les tests précédents.

6.3.Distance du Chi2 – Test

Pour mesurer la distance entre deux tableaux A et B à *m* lignes et *p* colonnes, l'idée naturelle est de prendre la distance euclidienne dans $\mathbf{R}^{mp}$, c'est à dire :

$$d(A,B)^2 = \sum_{i,j=1,1}^{m,p} (a_{ij} - b_{ij})^2$$

cependant dans notre démarche, cette distance ne correspond pas exactement à ce que nous recherchons. En effet, les deux tableaux (observé et théorique) ne jouent pas des rôles symétriques, nous voulons calculer la distance du tableau observé au tableau théorique puisque nous nous plaçons sous l'hypothèse H_0. Il est donc naturel d'accepter un écart plus grand pour une case du tableau théorique présentant un effectif plus grand, on va donc tenir compte dans la distance des effectifs théoriques attendus, et nous utiliserons comme distance, la distance, dite distance du Chi2, définie par $\hat{d}^2 = \sum_{i,j=1}^{m,p} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ où n_{ij} désigne, comme au paragraphe précédent, l'effectif observé et e_{ij} l'effectif théorique.

Une fois les marges fixées, les valeurs e_{ij} sont des constantes et sous l'hypothèse H_0, pour les échantillons présentant les marges données, seuls l'effectif n_{ij} change suivant la loi d'une variable aléatoire N_{ij}, nous pouvons donc considérer la distance D comme une variable aléatoire (statistique) définie par $D^2 = \sum_{i,j=1}^{m,p} \frac{(N_{ij} - e_{ij})^2}{e_{ij}}$, les variables aléatoires N_{ij} ne sont pas indépendantes, car elles doivent respecter les contraintes :

$$\text{pour tout } j \quad \sum_{i=1}^{m} N_{ij} = \sum_{i=1}^{m} e_{ij} = n_{\bullet j}$$

$$\text{pour tout } i \quad \sum_{j=1}^{p} N_{ij} = \sum_{j=1}^{p} e_{ij} = n_{i\bullet}$$

ce qui revient à dire que seules *(m-1)*(p-1)* d'entre elles sont indépendantes, comme on peut le voir quand on veut remplir "au hasard" un tableau à *m* lignes et *p* colonnes en respectant des marges données à l'avance.

On peut alors démontrer le résultat suivant : **quand n tend vers l'infini (et si aucun e_{ij} n'est borné), la variable D^2 tend en loi vers une loi du Chi2 à *(m-1)*(p-1)* degrés de liberté**.

Remarque : la condition imposée sur les e_{ij} est à rapprocher du cas de convergence d'une loi binomiale vers une loi de Poisson.

L'hypothèse H_0 est rejetée si la distance entre le tableau théorique et le tableau observé est trop grande, c'est à dire si la probabilité d'observer sous l'hypothèse H_0 une telle distance est inférieure au risque de première espèce α donné.

La valeur critique *c* de rejet de l'hypothèse H_0 est donc déterminée en fonction du risque α assumée par la formule $prob\left(\chi^2_{(m-1)(p-1)} > c\right) = \alpha$. On voit que la valeur critique peut être fixée avant tirage de l'échantillon. La règle de décision est alors la suivante : si la valeur de la statistique $\hat{d}^2$ observée sur l'échantillon est supérieure à *c*, l'hypothèse H_0 est rejetée et on conclut à une liaison entre les deux variables, ceci avec un risque d'erreur inférieur à.

On peut aussi raisonner en terme de niveau de significativité, en calculant la valeur de la statistique $\hat{d}^2$ sur l'échantillon, le degré de significativité (ou niveau de signification ou significience) est alors défini par

$prob\left(\chi^2_{(m-1)(p-1)} > \hat{d}^2\right) = ns$, la règle de décision consiste à rejeter l'hypothèse H_0 si le niveau de significativité est inférieur à α, dans ce cas le risque d'erreur est inférieur ou égal à *ns*.

6.4.Utilisation de SPSS

Nous allons illustrer ce test avec le fichier "DIP.sav". Le lecteur intéressé pourra refaire les calculs "à la main", nous utiliserons ici le logiciel SPSS, menu "analyse descriptive : tableau croisé".

Nous avons choisis dans l'option cellules de faire figurer les effectifs réels et théoriques et dans l'option test le test du chi2.

Les résultats sont les suivants :

Tableau croisé Age * Choix

			Choix				Total
			Distribution	Marque	Odeur	Texture	
Age	<25	Effectif	63	28	76	12	179
		Effectif théorique	68,3	34,1	51,2	25,4	179,0
	>65	Effectif	50	66	25	33	174
		Effectif théorique	66,3	33,2	49,8	24,7	174,0
	25-45	Effectif	91	8	52	31	182
		Effectif théorique	69,4	34,7	52,0	25,9	182,0
Total		Effectif	204	102	153	76	535
		Effectif théorique	204,0	102,0	153,0	76,0	535,0

La ligne Effectif contient l'effectif réel, l'effectif théoriques est calculé avec les formules du paragraphe 2 ; par exemple :

$$68{,}3 = \frac{179 * 204}{535}$$

La valeur du Khi-deux est calculée suivant la formule du paragraphe 3, nous n'interpréterons la rapport de vraisemblance qui sort du domaine de ce poly.

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	100,517[a]	6	,000
Rapport de vraisemblance	105,040	6	,000
Nombre d'observations valides	535		

a. 0 cellules (.0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 24.72.

La signification asymptotique correspond à ce que nous avons nommé degré de significativité, nous pouvons conclure ici (presque certainement) que les critères de choix dépendent effectivement de l'âge de l'acheteur.

6.5.Exercice : La société LOCVIDEO (fichier Videos.sav)

La société LOCVIDEO est une entreprise de location de vidéos du Sud-Est de la France, il est principalement implanté dans la région Lyonnaise, Grenobloise et Marseillaise. Jusqu'à présent l'approvisionnement des points de ventes se faisait de la même façon quelle que soit la ville, au bout d'un an d'existence la direction se demande si elle ne devrait modifier sa politique. Vous disposez d'un échantillon de la consommation de 1192 clients sur un mois pour faire vos recommandations.

1. Y a t-il une relation entre le premier ou le second choix de location et la ville?
2. Y a t il une relation entre le sexe et le choix des vidéos?
3. Pouvez-vous classer les trois régions en fonction de leur consommation : quelle est la ville qui consomme le plus de vidéos?

7. LA REGRESSION LINEAIRE

7.1.Un exemple (fichier Pubradio.sav)

Une entreprise de produits de grande consommation désire mesurer l'efficacité des campagnes de publicité et promotion pour différents médias. Spécialement trois types de médias sont utilisés régionalement, la presse, la radio et la distribution d'extraits de catalogue gratuits. Un échantillon de 22 villes de même grandeur a été choisi, villes pour lesquelles différents budgets de publicité ont été attribués aux trois. Après une période d'un mois, les ventes du produit (en milliers d'euros) ont été enregistrées ainsi que les dépenses publicitaires.

Ville	Ventes (000€)	Radio (000€)	Journaux (000€)	Gratuits (00€)	Ville	Ventes (000€)	Radio (000€)	Journaux (000€)	Gratuits (00€)
1	894	0	19	9	12	1452	19	19	17
2	1032	0	19	3	13	960	23	0	16
3	804	9	9	7	14	840	23	0	15
4	576	9	9	11	15	1224	26	9	10
5	840	13	13	12	16	1224	26	9	12
6	894	13	13	8	17	1296	29	13	14
7	858	16	16	11	18	1320	29	13	12
8	1086	16	16	17	19	1404	33	16	21
9	810	19	9	15	20	1602	33	16	19
10	906	19	9	10	21	1722	33	19	20
11	1500	19	19	15	22	1584	33	19	15

La direction commerciale peut-elle utiliser ces données pour prévoir les ventes en fonction des budgets dépensés?

7.2.La notion de modèle en statistique

Un modèle statistique met en relation une variable dite variable dépendante ou *variable à expliquer* et des variables dites indépendantes ou *variables explicatives*. Le vocabulaire dépendant, indépendant est plutôt anglo-saxon, la terminologie française correspond à la notion de variables explicatives et à expliquer ; les deux terminologies sont sujettes à caution, dans la mesure où les variables explicatives ne sont pas forcément indépendantes au sens probabiliste (sur la population munie de la loi uniforme), mais ne sont pas non plus cause des variations de la variable à expliquer. Dans la suite nous conserverons la terminologie française, variable à expliquer, variables explicatives. Les variations des variables explicatives sont simplement supposées influencer les variations de la variable à expliquer, le fait d'en être la cause ne peut être prouvé statistiquement, mais résultera d'un raisonnement économique ou autre, étranger à la statistique.

Un tel modèle statistique doit permettre :

- D'établir une relation analytique ou structurelle entre la variable à expliquer et les variables explicatives (généralement à partir d'un échantillon).
- D'analyser l'influence simultanée et/ou individuelle des variables explicatives sur la variable à expliquer. Dans certains cas d'éliminer des variables qui ne

s'avéreraient pas influentes ou de préciser les liens de causalité supposés par ailleurs.

- De prévoir la valeur espérée de la variable à expliquer si les valeurs des variables explicatives sont connues, et de préciser un intervalle de confiance pour cette prévision.

Dans la suite nous noterons toujours *Y* la variable à expliquer et $(X_k)_{k=1,p}$ les variables explicatives (au nombre de p) ; si la variable explicative est unique nous la noterons *X* sans indice. Toutes ces variables sont définies sur une même population ***P***.

Exemples :

- Dans notre exemple ***P*** : population des villes où sont distribués les produits pendant une période donnée

  Y = ventes mensuelles des produits en milliers d'euros

  X_1 = budget mensuel publicitaire radios locales en milliers d'euros

  X_2 = budget mensuel publicitaire presse locale en milliers d'euros

  X_3 = budget mensuel publicitaire pour les gratuits en milliers d'euros

  L'objectif est alors de prévoir les ventes mensuelles en fonction des budgets attribués aux deux médias.

- ***P*** : population des ménages en France pendant une période donnée

  Y = consommation d'un ménage pendant cette période

  X = revenu du ménage pendant cette période

  Ou encore

  Y = consommation d'un ménage pendant cette période

  X = revenu du ménage pendant cette période

  L'objectif pourrait alors être de prévoir l'impact d'une politique de revenus sur la consommation ou l'épargne.

- ***P*** : population des appartements d'un quartier de Paris à une période donnée

  Y = prix d'un appartement

  X_1 = surface de l'appartement

  X_2 = l'existence d'un parking

  Etc..

- ***P*** : population des zones géographiques de représentation médicale pendant une période donnée

  Y = nombre trimestriel de prescriptions d'un médicament

  X_1 = durée moyenne de la visite

  X_2 = nombre d'échantillons distribués

  X_3 = nombre de visites par médecins

  Etc..

Relation déterministe/statistique

Une variable Y est dite en relation déterministe avec des variables $(X_k)_{k=1,p}$ s'il existe une fonction *f* bien définie telle que : $Y = f(X_1, X_2, \ldots, X_p)$. Ce type de relation associe une et seule valeur *y* à *Y* pour des valeurs $x = (x_k)_{1 \le k \le p}$ des variables $X = (X_k)_{k=1,p}$. Un tel modèle appliqué au deuxième exemple du prix d'un appartement signifierait par exemple que tous les appartements de 100m^2 avec un parking ont le même prix de vente. Ceci n'est évidemment pas réaliste, dans un même quartier des appartements de même surface sont à des prix différents, ceci est du à des éléments tangibles tels que l'orientation, l'étage, la présence d'un gardien…, ou à des éléments plus subjectifs regroupés souvent sous le terme de charme.

L'exemple précédent montre que pour une valeur donnée des variables explicatives ne correspond pas une seule valeur de Y, mais tout un ensemble de valeur de Y, qui bien sur s'appliqueront à différents individus de la population pour lesquels les variables explicatives ont les mêmes valeurs : un appartement donné aura toujours un prix et un seul, mais le fait de connaître sa surface et la présence ou non d'un parking ne suffiront pour que l'on connaisse de façon déterministe son prix.

On exprimera cette notion en disant que les variables explicatives déterminent une loi de probabilité de la variable à expliquer *Y*, cette loi sera notée Y_x. Les paramètres de la loi de Y_x seront des fonctions déterministes de la variable $X = (X_k)_{k=1,p}$, en particulier la moyenne sera notée μ_x et sera l'espérance de Y conditionnée par la valeur prise par les variables explicatives :

$$\mu_x = E(Y / X = x)$$

on peut alors écrire sans perdre de généralité que

$$Y_x = \mu(x) + \varepsilon_x$$

où ε_x est une variable aléatoire de moyenne nulle (obtenue après centrage de la variable Y_x) et dont les autres paramètres dépendent théoriquement de la valeur *x* prise par les variables explicatives.

Ainsi sur le prix d'un appartement on aurait pour un appartement de 100 m2 avec parking (cette dernière variable valant 1 pour l'existence d'un parking 0 sinon) :

$$Y_{100,1} = \mu(100,1) + \varepsilon_{100,1}$$

se décompose en deux parties, une partie déterministe qui donnera le prix moyen d'un tel appartement et une partie aléatoire écart entre le prix moyen et le prix de l'appartement, qui prend en compte les autres éléments pouvant intervenir dans la fixation du prix. On écrira souvent de manière abusive, le modèle sous la forme :

$$Y = f(X) + E_X$$

La modélisation statistique consiste à spécifier la nature de la fonction déterministe de la moyenne, et les relations définissant les paramètres de la variable aléatoire e_x en fonction des valeurs de *x*. C'est à dire de se fixer à priori une certaine famille de fonction dépendant de paramètres qu'il faudra estimer à partir de données d'un échantillon, il faudra aussi à l'aide de tests valider la forme prédéfinie des différentes fonctions.

Exemple sur le prix d'un appartement

Il est possible pour ce problème d'envisager trois modélisations :

1. La présence d'un parking n'influence pas le prix de l'appartement dans ce cas seule la surface est un élément déterminant du prix, la fonction déterministe définissant la moyenne est une fonction d'une seule variable :

   $$f(X_1, X_2) = a + bX_1 \text{ d'où } Y = a + bX_1 + \mathrm{E}_X$$

   pour une valeur donnée de la surface x_1, nous aurons alors

   $$Y_{x_1,x_2} = a + bx_1 + \varepsilon_{x_1}$$

   b représente le prix du mètre carré dans le quartier (a serait en quelque sorte le coût d'entrée dans le quartier)

2. La présence d'un parking est un coût fixe donc augmente de façon constante le prix de l'appartement dans ce cas la fonction déterministe définissant la moyenne est une fonction de deux variables :

   $$f(X_1, X_2) = a + bX_1 + cX_2 \text{ d'où } Y = a + bX_1 + cX_2 + \mathrm{E}_X$$

   pour des valeurs données x_1 et x_2, nous aurons alors

   $$Y_{x_1,x_2} = a + bx_1 + cx_2 + \varepsilon_{x_1,x_2}$$

   b représente le prix du mètre carré dans le quartier et c représente le prix d'un parking dans le quartier (a serait en quelque sorte le coût d'entrée dans le quartier).

3. On peut aussi envisager que la présence d'un parking influe aussi sur le prix du mètre carré, auquel cas nous aurions la fonction déterministe suivante :

   $$f(X_1,0) = a + bX_1 \text{ en l'absence de parking}$$

   $$f(X_1,1) = a' + b' X_1 \text{ en présence d'un parking}$$

   en notant $a'=a+c$ et $b'=b+d$ nous pouvons réécrire ces deux équations sous la forme unique suivante :

   $$f(X1, X2) = a + bX_1 + cX_2 + dX_1X_2$$

   ou encore en notant X_3 la variable définie par $X_3=X_1X_2$, nous avons un modèle linéaire à trois variables explicatives :

   $$Y = a + bX_1 + cX_2 + dX_3 + \mathrm{E}_X$$

   pour des valeurs données x_1 et x_2 ($x_3=x_1x_2$), nous aurons alors

   $$Y_{x_1,x_2} = a + bx_1 + cx_2 + dx_3 + \varepsilon_{x_1,x_2}$$

A partir d'un échantillon d'appartement, la modélisation statistique nous permettra d'estimer les coefficients et de tester la validité de chacun des modèles sur l'ensemble de la population. La modélisation fait donc appel aux deux techniques que nous avons présentées précédemment l'estimation et les tests d'hypothèse.

7.3.Le modèle de régression linéaire

Nous allons ici faire des hypothèses tant sur la partie déterministe, fonctionnelle de la moyenne conditionnée, que sur la partie aléatoire ; ces conditions vont nous permettre d'avoir des outils pour estimer les éléments du modèle appelé modèle de régression linéaire.

Hypothèse déterministe du modèle de régression linéaire

La première hypothèse du modèle de régression linéaire consiste à modéliser l'espérance mathématique conditionnelle par une fonction linéaire (ou plus exactement une fonction affine) :

$$\mu(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Remarque : si l'on ajoute la variable "artificielle" X_0 égale à 1 sur toute la population (donc x_0 vaut toujours 1), la formule peut alors s'écrire :

$$\mu(x_0, x_1, x_2, \ldots, x_p) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \sum_{k=0}^{k=p} \beta_k x_k$$

ce qui justifie le nom de linéaire.

Dans le cas d'une seule variable explicative, la régression est dite simple dans tous les autres cas la régression est dite multiple.

Les coefficients $(\beta_k)_{1 \le k \le p}$ sont appelés coefficients de la régression et sont évidemment inconnus, ce sont des coefficients valables sur toute la population, si l'un d'entre eux β_j est nul cela veut dire que la variable associée X_j n'a pas d'influence marginale linéaire sur les variations de la variable Y, mais cela ne veut pas dire que la variable X_j n'a pas d'influence sur les variations de Y, cette influence peut être d'autre nature (logarithmique, exponentielle etc…) ou peut être cachée par des corrélations entre variables explicatives, la part explicative de la variable X_j étant déjà prise en compte par d'autres variables. La variable aléatoire conditionnée par les valeurs $(x_1, \ldots, x_p)$ s'écrit alors :

$$Y_{x_1,\ldots,x_p} = \sum_{k=0}^{k=p} \beta_k x_k + \varepsilon_{x_1,\ldots,x_p}$$

ce qui peut s'écrire de manière abusive, sans rappeler les valeurs spécifiques des variables explicatives :

$$Y = \sum_{k=0}^{k=p} \beta_k X_k + \mathrm{E}_X$$

E_X désignant une famille de variables aléatoires dont les paramètres dépendent des valeurs prises par les variables explicatives $(X_k)_{1 \le k \le p}$. C'est sur cette dernière famille de loi que vont porter les autres hypothèses du modèle de régression linéaire.

Hypothèses probabilistes du modèle de régression linéaire.

Trois hypothèses sont formulées sur la famille de variables aléatoires E_X, ces hypothèses sont nécessaires soit pour l'estimation des paramètres soit pour les tests du modèle.

– Homoscédasticité : La première hypothèse porte sur la variance des lois de la famille E_X, on suppose que cette variance est constante, indépendante de la valeur prise par les différentes variables explicatives. L'écart type associé sera noté σ. Il est important dans la pratique de comprendre ce que cela signifie, par exemple pour le prix d'un appartement, cela voudrait dire que la dispersion des prix est la même pour les appartements de 20m^2 ou pour les appartements de 150m^2. Cette condition peut conduire parfois à limiter la

population pour qu'elle soit réalisée, on pourrait par exemple se limiter aux appartements dont la surface est comprise entre 60 et 120m^2.

- Indépendance : on suppose que les variables $\varepsilon_{x_1,\ldots,x_k}$ et $\varepsilon_{x_1',\ldots,x_k'}$ sont indépendantes, quelles que soient les valeurs $(x_1,\ldots,x_p),(x_1',\ldots,x_p')$; cette hypothèse est particulièrement importante lorsque l'on traite des données indexées par le temps. Par exemple cela signifie qu'un mois de surconsommation n'a pas plus de "chances" d'être suivie d'un mois de sous consommation qu'un autre (pas d'effet de stockage).
- Normalité : on suppose enfin (et ceci pour les tests particulièrement) que toutes les variables aléatoires de la famille E_X sont normales, donc suivent une loi normale de moyenne nulle et d'écart type σ.

Compte tenu de ces trois hypothèses, on pourra alors par abus de langage utiliser une notation générique unique en confondant toutes les lois de la famille E_X en une seule, et le modèle sera alors noté :

$$Y = \sum_{k=0}^{k=p} \beta_k X_k + \varepsilon \quad où \quad \varepsilon \to N(0,\sigma)$$

En définitive un modèle de régression linéaire comporte $p+2$ paramètres à estimer, les $p+1$ coefficients de régression $(\beta_0,\beta_1,\ldots,\beta_p)$ et l'écart type σ de la partie aléatoire.

Estimation des paramètres du modèle

Nous présenterons sous forme géométrique la méthode d'estimation des coefficients, le lecteur peu amateur de mathématiques peut ignorer cette section, puisque les valeurs des estimations seront données par SPSS et l'utilisateur n'aura pas à les retrouver, ces formules ne seront d'ailleurs données qu'en annexe, nous nous limiterons ici à une interprétation géométrique, permettant de mieux comprendre les notions de degrés de liberté attachés au modèle.

Les paramètres du modèle sont estimés à partir d'un échantillon de taille n, sur lequel sont relevées les valeurs des variables explicatives et de la variable à expliquer. On obtient ainsi un tableau de données :

y_1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	x_{i1}	x_{i2}	$\cdots$	x_{ik}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Si le modèle de régression linéaire est valide, nous devons avoir les *n* relations suivantes entre les valeurs prises par la variable à expliquer *Y* et les variables explicatives $(X_k)_{1\le k\le p}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i$$

où e_i, appelée valeur résiduelle, correspond à la réalisation de la variable aléatoire ε pour la i^{ème} observation.

Critère des moindres carrés

Les valeurs résiduelles dépendent des valeurs des paramètres du modèle $(\beta_0, \beta_1, \ldots, \beta_p)$, plus l'amplitude de cette valeur est grande, moins bien l'observation est représentée par le modèle, il est donc naturel de penser que si le modèle de régression est bien adapté aux données sur l'ensemble des observations les valeurs résiduelles ne sont pas, en valeur absolue, trop élevées, cette démarche est à rapprocher, bien que différente mais liée (voir plus loin), de la méthode du maximum de vraisemblance en estimation.

On cherchera donc des valeurs des coefficients de régression telles que l'ensemble des amplitudes des valeurs résiduelles soit le plus faible possible, pour des raisons historiques de commodité de calcul analytiques on utilisera la somme des carrés pour mesurer cet ensemble. Le critère des moindres consiste donc à déterminer les valeurs des coefficients qui minimisent :

$$h(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2$$

Ces valeurs seront notées $(b_0, b_1, \cdots, b_p)$, nous aurons alors :

$$h(b_0, b_1, \ldots, b_p) = \min h(\beta_0, \beta_1, \ldots, \beta_p)$$

Ce minimum peut être déterminé en résolvant le système de *p+1* équations à *p+1* inconnues obtenu en, dérivant la fonction h à chacun des *p+1* coefficients (on suppose que ce système d'équations à une solution unique, ce que nous interpréterons géométriquement au paragraphe suivant).

Nous noterons dans la suite $\hat{y}_i$ l'estimation de la moyenne correspondant à la variable aléatoire de la i^{ème} observation :

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

et $\hat{e}_i$ l'estimation de la i^{ème} valeur résiduelle : $\hat{e}_i = y_i - \hat{y}_i$

Interprétation géométrique du critère des moindres carrés

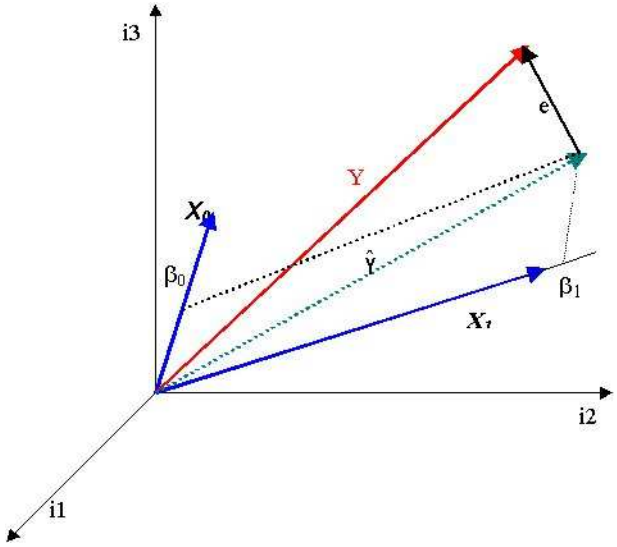
Nous allons interpréter géométriquement la méthode des moindres carrés, ce qui nous permettra d'expliciter certaines propriétés des estimations et estimateurs associés. Pour cela nous allons nous placer dans l'espace des individus, c'est à dire que nous allons considérer un espace vectoriel à n dimensions, chaque dimension étant associée à un individu de l'échantillon. Par exemple pour un échantillon de taille 3 nous aurons un espace de dimension 3, c'est ce que nous utiliserons pour les représentations graphiques.

Dans cet espace nous pouvons associer à chaque variable (plus exactement à chaque échantillon image de chaque variable) un vecteur, que nous noterons avec des lettres majuscules :

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} \quad \ldots \quad X_p = \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}$$ plus les deux autres vecteurs $$X_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

les n relations écrites au paragraphe précédent donnent une seule relation vectorielle :

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + E$$



Le vecteur $\beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ appartient au plan $\prod$ engendré par les vecteurs $(X_0, X_1, \ldots, X_p)$ que nous supposerons indépendants (ce qui revient à considérer que le système d'équations évoqué au paragraphe précédent a une solution unique), quelles que soient les valeurs des β_k, d'autre part le critère des moindres carrés s'interprète comme la norme (au carré) du vecteur *E*. Pour satisfaire le minimum de la norme de ce vecteur, il faut donc projeter *Y* sur le plan $\prod$. Les estimations des coefficients de la régression sont donc les coordonnées du vecteur $\hat{Y}$ projection de *Y* sur le plan $\prod$. Le vecteur E est alors orthogonal à ce plan (donc à tous les vecteurs de ce plan).

Propriétés des estimations des moindres carrés

1. La somme des résidus est égale à 0. En effet le vecteur $\hat{E}$ correspond au minimum de la norme, critère des moindres carrés, est perpendiculaire au vecteur X_0, dont toutes les coordonnées sont égales à 1, donc le produit scalaire de ces deux vecteurs est nul :

$$\left\langle \hat{E}, X_0 \right\rangle = 0 = \sum_{i=1}^{n} \hat{e}_i \cdot 1 = \sum_{i=1}^{n} \hat{e}_i$$

2. Les estimations des moyennes $\hat{y}_i$ ont même moyenne que les observations y_i. En effet :

$$\sum_{i=1}^{n} \hat{e}_i = 0 = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i \quad \textit{donc} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$

3. Le centre de gravité du nuage de points est dans le plan (sur la droite) de régression, c'est à dire que l'on a la relation suivante :

$$\bar{y} = b_0 + b_1 \bar{x}_1 + \ldots + b_p \bar{x}_p$$

où $\bar{y}, \bar{x}_1, \ldots, \bar{x}_p$ désignent les moyennes des variables sur l'échantillon. Ceci résulte immédiatement de la somme nulle des résidus.

4. Le vecteur $\hat{Y}$ des estimations est dans le plan $\prod$, donc orthogonal au vecteur $\hat{E}$ on a donc la relation suivante :

$\left\langle \hat{Y}, \hat{E} \right\rangle = \sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0$ *ou encore* $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})\hat{e}_i = \sum_{i=1}^{n} \hat{y}_i \hat{e}_i - \bar{y}\sum_{i=1}^{n} \hat{e}_i = 0$ car la somme des résidus est nulle.

5. On a la décomposition suivante, appelée décomposition des carrés :

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ce qui résulte de la propriété 4 et du fait que $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$. Cette décomposition peut s'interpréter de la façon suivante :

- La somme du côté gauche est indicatrice de la dispersion totale initiale, elle est appelée Somme des Carrés Totale :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- La première somme du côté gauche, représente la dispersion due aux variables explicative, ce que le modèle permet d'expliquer, elle est appelée somme des carrés reconstituée par le modèle de régression, ou plus simplement Somme des Carrés Expliquée :

$$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- La dernière somme donne une indication de la dispersion autour du plan de régression, c'est à dire de la dispersion non expliquée par le modèle, elle est appelée Somme des Carrés Résiduelle :

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

En conséquence la décomposition des carrés s'exprime de la façon suivante :

$$SCT = SCE + SCR$$

Cette décomposition exprime que la variabilité des valeurs observées $(y_i)_{1 \le i \le n}$ mesurée par *SCT* est la somme des variabilités des valeurs $(\hat{y}_i)_{1 \le i \le n}$ reconstituées par le modèle de régression mesurée par *SCE*, et de la variabilité des résidus mesurée par *SCR*. En conséquence comme *SCT* est constant, on peut être tenté de dire qu'il faut rendre *SCE* le plus grand possible ; il faut toutefois faire attention que seul l'échantillon est reconstitué et que nous sommes concernés par l'ensemble de la population, et que cette "optimisation" ne doit pas être obtenue à n'importe quel prix.

6. L'estimation de la variance commune des variables aléatoires ε, est donnée par :

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n - p - 1}$$

Dans la mesure où l'estimation se fait à partir d'un échantillon de taille *n*, il ne peut y avoir plus de *n-1* variables explicatives, ceci résulte de la dimension de l'espace des individus. Mais de façon plus précise, quelles que soient les n-1 variables choisies

(qu'elles soient économiquement explicatives ou pas) on arrivera toujours à une somme des carrés résiduelle nulle.

- La somme des carrés totale est donc prise dans un espace à n-1 degrés de libertés.
- La somme des carrés expliquée se trouve dans l'espace des variables explicatives, dans un espace de dimension *p*, car il ne faut pas prendre en compte le vecteur constant X_0.
- La somme des carrés résiduelle est dans un espace orthogonal à l'espace des variables explicatives et à X_0, donc dans un espace de dimension *n-p-1*. Pour avoir la moyenne sur un axe de la somme des carrés, qui représentera une estimation de la dispersion moyenne inexpliquée donc de la variance de ε, il faut donc diviser la norme carrée de *E* par la dimension de l'espace dans lequel il se trouve.

4. On peut enfin démontrer les résultats suivants sur les estimateurs obtenus par la méthode des moindres carrés :
   - Les estimateurs des coefficients de régression sont des combinaisons linéaires des observations de la variable à expliquer. Ils suivent donc une loi normale.
   - Les estimateurs des coefficients de régression et de la variance de ε, sont sans biais et convergents.
   - Les estimateurs des coefficients de régression sont les meilleurs estimateurs non biaisés, linéaires, c'est à dire que ce sont parmi les estimateurs linéaires non biaisés ceux qui ont la variance minimum.
   - Les estimateurs des coefficients de régressions par la méthode des moindres carrés sont les même que ceux obtenus par la méthode du maximum de vraisemblance. Ce n'est pas le cas pour l'estimation de σ.

   Certains de ces résultats seront démontrés en annexe, sinon on pourra consulter

Indices de qualité d'un modèle de régression

Dans la mesure où nous travaillons sur un échantillon et non sur la population toute entière, il nous faut disposer d'indicateur, permettant de savoir avec quelle confiance on peut étendre les résultats à la population entière, et avec quelle fiabilité on peut faire des prévisions, à partir de valeurs connues des variables explicatives. Comme nous l'avons vu au paragraphe précédent il est toujours possible de réduire l'incertitude à zéro, sur l'échantillon mais cela n'a aucun intérêt pour la population, c'est un simple effet de saturation mathématique.

Les logiciels statistiques donnent toujours la même structure à un listing de régression linéaire. Cette présentation est faite sous trois chapitres : indicateurs résumés, validité globale, validité marginale.

Résumés de la régression

Cette rubrique contient trois éléments : le coefficient de détermination, le coefficient de corrélation multiple, l'écart type des résidus.

Le coefficient de détermination R^2

Le coefficient de détermination est le pourcentage de la somme des carrés totale expliqué par le modèle. Il est défini par le rapport :

$$R^2 = \frac{SCE}{SCT}$$

très souvent, mais par excès de langage on dit que R^2 représente le pourcentage de variance expliqué par le modèle. L'excès est double, en effet les sommes des carrés (totale et expliquée) ne sont pas des variances, ensuite le rapport ne porte que sur l'échantillon. Plus ce rapport est proche de 1, meilleure est la reconstitution de la variabilité de la variable à expliquer sur l'échantillon. Comme nous l'avons vu au paragraphe précédent, en prenant n-1 variables explicatives quelconques on reconstituera toujours à 100% la variabilité de l'échantillon.

Cet indicateur est donc un indicateur biaisé, il augmentera de façon systématique avec le nombre de variables explicatives. Sans qu'il y ait de règle rationnelle donnant le nombre de variables explicatives maximum pour un nombre donné d'observations, en pratique il est recommandé de prendre au moins 5 à 6 observations par variable explicative.

Enfin plus que la valeur du R^2, ce qui est intéressant, c'est la variation de cette valeur par ajout de variable, si cette variation est trop faible la variable (ou les variables) ajoutée(s) sont sans intérêt pour le modèle, comme nous le verrons plus loin.

Le coefficient de détermination est un indicateur intrinsèque d'adéquation linéaire, un mauvais R^2 n'est pas le signe d'une non influence des variables explicatives choisies, mais le signe d'une absence de liaison linéaire. Si des raisons économiques poussent à croire à une influence des variables explicatives choisies, il faudra alors peut-être utiliser des transformations non linéaires.

Enfin pour terminer, coefficient de détermination, ne peut en aucun cas servir à choisir une régression parmi plusieurs régressions n'ayant pas le même nombre de variables.

Remarque : certains logiciels utilisent, pour diminuer le biais du au nombre de variables explicatives, un coefficient de détermination corrigé (ou ajusté):

$$R^2C = 1-(n-1)\left(1-R^2\right)/(n-p-1)$$

Le coefficient de corrélation multiple *R*

Ce coefficient est simplement la racine du coefficient de détermination, mais il s'interprète comme la corrélation entre la série des valeurs observée $(y_i)_{1\le i\le n}$ et la série des valeurs calculées par le modèle $(\hat{y}_i)_{1\le i\le n}$. Plus ce coefficient est proche de 1, meilleure est la reconstitution des données par le modèle.

Estimation de l'écart type des résidus

Aussi appelée Erreur type de la régression, cet indicateur donne une idée de la dispersion des valeurs autour de la valeur moyenne estimée par la partie déterministe du modèle. Plus cette estimation est faible meilleure est la prévision que l'on pourra faire à partir du modèle. Comme nous l'avons plus haut cette valeur est donnée par la formule :

$$s^2 = \frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-p-1} = \frac{SCR}{n-p-1}$$

Bien que liée au coefficient de détermination, cette valeur n'en a pas les défauts, en effet le dénominateur corrige l'effet de l'augmentation des variables, cette quantité n'est d'ailleurs pas définie dans le cas de modèle saturé pour l'échantillon, c'est à dire à *p=n-1* variables.

Entre deux modèles on aura tendance à choisir celui dont l'erreur type est la plus petite.

Validité globale du modèle

La question posée ici est la suivante : les données observées permettent-elles d'inférer (sur la population) qu'aucune des variables explicatives $(X_k)_{1\le k\le p}$ n'a d'influence sur les variations de la variable *Y*. Ou en prenant la contraposée de cette proposition, peut penser qu'au moins unes des variables $(X_k)_{1\le k\le p}$ a une influence significative (au niveau de la population) sur les variations de *Y*. Comme d'habitude, quand nous parlons d'influence, nous sous-entendons le terme linéaire.

Si aucune des variables $(X_k)_{1\le k\le p}$ n'avait d'influence sur les variations de Y, ceci signifierait que seul resterait le terme aléatoire autour de la moyenne de la population, le modèle serait alors :

$$Y = \beta_0 + \varepsilon \quad où \quad \beta_0 = \mu \text{ moyenne de } Y \text{ sur la population}$$

Nous pouvons donc poser notre problème sous forme de test d'hypothèse, l'hypothèse nulle correspondant à la non influence des variables $(X_k)_{1\le k\le p}$.

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 \quad : \text{il existe au moins un indice } k \text{ tel que } \beta_k \neq 0$$

La région du rejet de l'hypothèse H0 est basée sur la statistique dite du "Fisher global". L'idée du test est de comparer l'apport explicatif moyen des variables choisies par l'analyste avec le pouvoir explicatif moyen de variables complémentaires totalement arbitraires (correspondant aux résidus). Pour cela on va donc faire le rapport entre la diminution de la somme des carrés due en moyenne à chaque variable explicative et la diminution moyenne résiduelle, c'est à dire l'estimation de l'écart type des résidus. Si ce rapport n'est pas suffisamment grand (significativement plus grand que 1), ceci signifiera que les variables explicatives n'ont pas de pouvoir explicatif plus important que les variables résiduelles et n'ont donc pas à en être distinguées. On utilisera donc la statistique :

$$F_c = \frac{\dfrac{SCE}{p}}{\dfrac{SCR}{n-p-1}} = \frac{CME}{CMR}$$

CME désigne le carré moyen expliqué, c'est à dire la somme des carrés expliquée par le modèle, divisée par la dimension de l'espace explicatif (*p = l*e nombre de variables explicatives), *CMR* désigne le carré moyen résiduel, c'est à dire la somme des carrés résiduelle divisée par la dimension de l'espace résiduel (*n-p-1*). La région critique de rejet de l'hypothèse H_0, sera de la forme $[f_\alpha, +\infty[$, f_α étant déterminé en fonction du risque de première espèce par $prob(F_c \ge f_\alpha) = \alpha$.

Pour pouvoir poursuivre la procédure de test, il nous faut connaître la loi de F_c sous l'hypothèse nulle, c'est ici qu'intervient l'hypothèse de normalité de la variable ε. Sous l'hypothèse H_0, la statistique F_c suit une loi dite de Fisher-Snedecor à (*p,n-p-1*) degré de libertés. On peut alors déterminer f_α soit à l'aide de tables. En pratique, on calcule la valeur f_c de la statistique F_c sur l'échantillon, puis on détermine le niveau de signification $ns = prob(FS(p, n-p-1) > f_c)$ du test correspondant à cette valeur, si ce niveau est inférieur à α on rejette l'hypothèse. Le test est présenté de façon classique, dans un tableau nommé Analyse de la Variance :

Source de variation	Degrés de liberté	Somme des carrés	Carré Moyen	f_c	Niveau de signification
Régression	*p*	SCE	$CME = \frac{SCE}{p}$	$f_c = \frac{CME}{CMR}$	*ns*
Résiduelle	*n-p-1*	SCR	$CMR = \frac{SCR}{n-p-1}$		
Totale	*n-1*	SCT			

Validité marginale de chaque variable du modèle

L'objectif est ici de savoir si le modèle n'est pas surdéfini, c'est à dire qu'aucune des variables explicatives du modèle n'a un l'apport marginal dans l'explication des variations de *Y* nul. Ceci revient à dire qu'il faut vérifier que pour chacune des variables individuellement (les autres étant supposées rester dans la régression) le coefficient β n'est pas nul. Le test se pose de la façon suivante, pour une variable explicative X_k et une seule, les autres variables étant supposées dans le modèle :

$$H_0 \quad : \quad \beta_k = 0$$
$$H_1 \quad : \quad \beta_k \neq 0$$

Evidemment l'estimation b_k du coefficient n'est pas nul, mais est la valeur prise par un estimateur sans biais B_k, sur l'échantillon de taille *n*. Cet estimateur suit une loi normale (si les résidus suivent une loi normale), dont l'écart type est inconnu, mais peut être estimé par un estimateur $S(B_k)$, la statistique utilisée pour le test sera alors :

$$T_c = \frac{B_k}{S(B_k)}$$

qui sous l'hypothèse H_0 suit une loi de Student à *(n-p-1)* degrés de liberté.

L'hypothèse nulle sera rejetée si la valeur observée de la statistique est significativement différente de 0, c'est à dire si l'estimation du coefficient est assez éloignée de 0, compte tenu de l'incertitude de cette estimation (incertitude exprimée par l'écart type). La région critique de rejet de l'hypothèse H_0.est de la forme $]-\infty,-t] \bigcup [t,+\infty[$, la valeur de *t* est déterminée en fonction du risque de première espèce α, de façon précise *t* est le fractile d'ordre $1-\alpha/2$ de la loi de Student à *n-p-1* degrés de liberté.

Tous les logiciels statistiques préfèrent donner le niveau *ns* de signification, c'est à dire en notant t_c la valeur de la statistique T_c observée sur l'échantillon :

$$ns = prob(|Student(n-p-1)| > |t_c|) = 2\,prob(Student(n-p-1) > |t_c|)$$

si ce niveau de signification est inférieur à α, on rejette l'hypothèse H_0.

Les éléments nécessaires à cette validation marginale sont toujours présentés, dans les logiciels statistiques, dans un tableau donnant les coefficients du modèle. Ce tableau à la forme suivante :

Variable	Coefficient	Ecart type (du coefficient)	t_c	Niveau de signification

X_1	b_1	$s(B_1)$	$\frac{b_1}{s(B_1)}$	ns_1
⋮	⋮	⋮	⋮	⋮
X_p	b_p	$s(B_p)$	$\frac{b_p}{s(B_p)}$	ns_p
Constante	b_0	$s(B_0)$	$\frac{b_0}{s(B_0)}$	ns_0

Remarques :

1. Si plusieurs variables explicatives ne conduisent pas au rejet de l'hypothèse nulle, ceci ne permet pas de penser que tous leurs coefficients sont nuls, c'est à dire qu'aucune d'entre elles n'est influente sur les variations de Y. En effet, la non influence d'une variable peut résulter de corrélation entre les variables explicatives, ôter alors unes de variables non influentes significativement peut rendre les autres significativement influentes. Ne jamais oublier que ce test porte sur une variable vis à vis de toutes les autres.
2. Si la constante n'est pas significative (et elle seule), il est possible d'essayer un modèle sans constante, en forçant à 0 sa valeur. Dans ce cas il faut modifier en conséquence les degrés de liberté des résidus qui ne sont plus *n-p-1* mais *n-p*.

7.4.Utilisation de SPSS pour la régression

Le listing produit par SPSS contient les éléments suivants :

Le premier tableau indique quelles ont été les variables entrées dans la régression, généralement celles indiquées par l'utilisateur sauf en cas de multicolinéarité, rendant impossible le calcul.

Variables introduites/supprimées

Modèle	Variables introduites	Variables supprimées	Méthode
1	Gratuits, Journaux, Radio[a]	.	Entrée

a. Toutes variables requises saisies.

Vient ensuite un tableau donnant l'écart-type de la régression et le coefficient de corrélation multiple. L'écart-type de la régression est nommé "Erreur standard de l'estimation".

Comme la valeur du R^2 augmente structurellement avec le nombre de variable, SPPS fournit un R^2 ajusté comme indiqué plus haut.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation

1	,916[a]	,839	,813	138,034

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

Le troisième tableau correspond à la validation globale du modèle, avec la décomposition de la somme totale des carrés entre somme des carrés "expliqués" ici sur la ligne intitulée "Régression" et la somme des carrés résiduels. La colonne suivante contient les degrés de liberté associés à cette décomposition, puis vient le F calculé et enfin la significativité de ce F, c'est-à-dire la probabilité (en respectant les hypothèses de la régression) d'observer une telle valeur du F si les variables n'avaient aucune influence linéaire, c'est-à-dire si tous les coefficients étaient nuls.

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1793129,948	3	597709,983	31,370	,000[a]
	Résidu	342959,506	18	19053,306		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

b. Variable dépendante : Ventes

Le dernier tableau donne les coefficients des différentes variables explicatives ainsi que le terme constant. L'erreur standard est l'estimation de l'écart de l'estimateur des coefficients. Les coefficients standardisés sont les coefficients dans le cas où toutes les variables (à expliquer et explicatives) seraient centrées réduites.
Comme pour le F, ici vous est donnée la significativité du t calculé, c'est-à-dire la probabilité d'observer une telle valeur du t sous l'hypothèse H_0 c'est-à-dire si l'apport marginal de la variable était nulle (cf Gratuits).

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	238,458	112,242		2,124	,048
	Radio	23,850	4,524	,749	5,272	,000
	Journaux	32,629	5,369	,585	6,078	,000
	Gratuits	-,619	10,228	-,009	-,060	,952

a. Variable dépendante : Ventes

Nous remarquons sur ce listing que la variable Gratuits, n'est marginalement pas significative, ceci est peut-être du à une corrélation entre les variables explicatives, nous reviendrons plus loin sur cette question. Il est d'ailleurs rassurant de constater que cette variable n'est

statistiquement pas significative, car son coefficient négatif, signifiait qu'une fois les budgets publicitaires Radio et Journaux fixés, le fait de distribuer des extraits de catalogue gratuit faisait diminuer les ventes!

Il faudrait donc faire une autre régression en supprimant cette variable.

7.5.Pratique de la régression - Analyse d'un listing de régression – Choix d'un modèle

Avant de tester un modèle de régression, il est utile de vérifier graphiquement que les hypothèses du modèle de régression linéaire, ne sont pas violées de façon évidente. Une fois cette vérification faite et les changements de variables éventuels effectués, on peut procéder à l'élaboration de plusieurs modèles, et obtenir différents listings de régression.

L'analyse d'un listing de régression consiste à déterminer si un modèle est acceptable statistiquement et économiquement. Le problème ne se pose que si la régression est faite sur un échantillon, et si on envisage d'étendre les résultats à l'ensemble de la population.

Analyse préalable des données – Changement de variables

Généralement on se contente d'une représentation graphique des données, en mettant en abscisse les différentes variables explicatives et en ordonnées la variable à expliquer. On pourra obtenir différents types de graphiques :

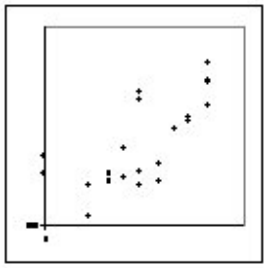


figure 1

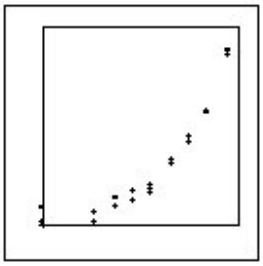


figure 2

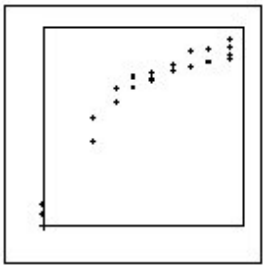


figure 3

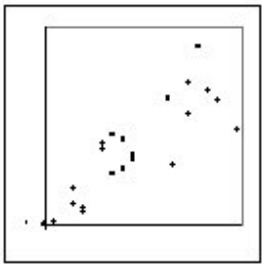


figure 4

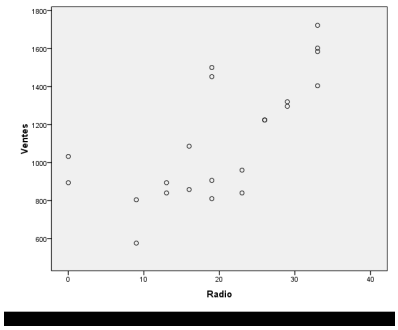
Les figures 2, 3, 4 montrent des distributions de données qui ne satisfont les hypothèses du modèle de régression linéaire. Sur la figure 1, en revanche, rien ne semble à priori contrarier ces hypothèses (sauf éventuellement la normalité, mais il faut d'abord estimer le modèle) : les données semblent bien être réparties autour d'une droite (hypothèse de linéarité) et l'épaisseur du nuage de point paraît à peu près constante, sans être systématiquement d'un côté ou de l'autre de la tendance linéaire.

Les figures 2 et 3 indique clairement une allure non linéaire de la moyenne des *y* pour une abscisse *x* donnée, on pourra dans les deux cas essayer une transformation puissance d'exposant supérieur à 1 pour la figure 2(par exemple x^2) et inférieure à 1 pour la figure 3 (par exemple $\sqrt{x}$). Les cas les plus accentués (les plus loin du linéaire) étant représentés par la fonction exponentielle pour la figure 2 et la fonction logarithmique pour la figure 3.

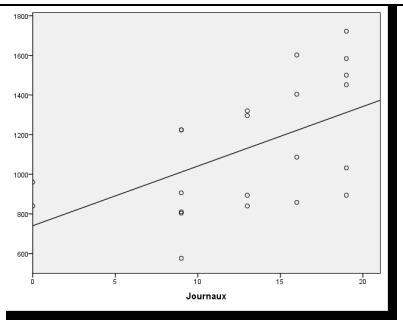
La figure 4 ne met en cause fondamentalement, la linéarité de la moyenne, mais elle montre clairement que la dispersion autour de cette moyenne n'est pas constante, les données ne respectent pas l'hypothèse d' homoscédasticité des résidus, on peut penser ici que la dispersion est proportionnelle à une puissance (ou au logarithme) de la variable explicative X_k représentée en abscisse. On pourra alors utiliser le changement de variable pour la variable à expliquer Y/X^a ou $Y/\ln(X)$.

Toutes ces transformations, simples à réaliser avec SPSS, doivent être validées par un nouveau graphique (faisant intervenir ou non la droite de régression) et aussi par le calcul des corrélations simples éventuellement.

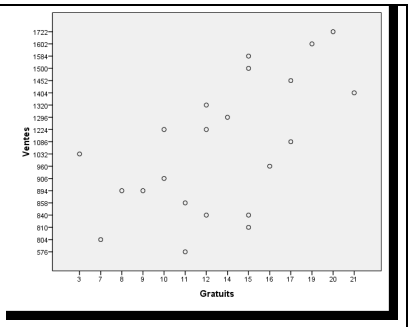
Application à notre exemple, les trois graphiques sont les suivants :



Radio/Ventes



Journaux/Ventes



Gratuits/Ventes

Les graphiques n'infirment pas les hypothèses du modèle de régression, ce qui est confirmé en calculant les corrélations simples entre la variable à expliquer et les variables explicatives :

Corrélations

		Ventes	Radio	Journaux	Gratuits
Ventes	Corrélation de Pearson	1	,707**	,539**	,589**
	Sig. (bilatérale)		,000	,010	,004
	N	22	22	22	22

**. La corrélation est significative au niveau 0.01 (bilatéral).

Validation d'un modèle

La partie résumé ne fournit que des indications générales sur le modèle sans permettre de valider ou non statistiquement le modèle, elle est surtout utile quand on veut choisir parmi plusieurs modèles.

Validation statistique

La validation statistique se fait en fonction d'un risque de première espèce fixé, généralement 5% ou 1%.

La première validation est la validation globale, cette validation se fait à l'aide du tableau d'analyse de la variance. Il suffit de vérifier que le niveau de signification de la statistique de Fisher est inférieur au risque de première espèce. Si ce n'est pas le cas, l'ensemble des variables explicatives est à rejeter, au moins sans transformation nouvelle, l'analyse s'arrête là. Si le modèle est globalement accepté, il faut ensuite passer à la validation marginale. Sur notre exemple le niveau de signification est quasi nul, très inférieur à 1%, donc nous validons globalement notre modèle.

La validation marginale se fait à l'aide du tableau du modèle, pour que le modèle soit statistiquement acceptable, il faut que le niveau de signification de chacun des t_c soit inférieur au risque de première espèce. Si ce n'est pas le cas, il est nécessaire d'ôter au moins une des

variables explicatives prises en compte, généralement on enlèvera une et une seule des variables dont l'apport marginal est non significatif.

Sur notre exemple, seule la variable Gratuits n'est pas marginalement significative nous pouvons alors tester un modèle sans cette variable. Le tableau du modèle est alors le suivant :

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	235,168	95,577		2,461	,024
	Radio	23,646	2,935	,742	8,058	,000
	Journaux	32,571	5,140	,584	6,337	,000

a. Variable dépendante : Ventes

Cette fois toutes les variables sont marginalement significatives et le modèle est donc acceptable statistiquement.

Validation économique

Une fois le modèle accepté statistiquement, il est bon de vérifier que les signes des coefficients sont cohérents avec ce que l'analyste attendait ; sinon des raisons de cette incohérence sont à rechercher économiquement et non pas statistiquement.

Sur notre exemple, le modèle valide statistiquement est cohérent d'un point de vue économique, les deux coefficients sont positifs, comme il est naturel de le supposer : la publicité fait augmenter les ventes. Le modèle nous permet d'ailleurs de quantifier cet effet, à budget Radio fixé, 1000€ de publicité dans les journaux font augmenter les ventes de 32 500€ environ, et à budget Journaux fixé 1000€ de publicité à la Radio fait augmenter les ventes de 23 600€ environ.

Remarque : en comparant les deux listings de régression, on obtient les résumés suivants :

Modèle	R2	Erreur Standard
3 variables	0,83945	138,034
2 variables	0,83941	134,37

Comme nous l'avions dit le coefficient de détermination est plus grand dans le modèle à trois variables que dans le modèle à deux, ce qui est purement mathématique, mais ne garantit en rien une meilleure adéquation du modèle aux données; En revanche l'erreur type, estimation de l'écart type des résidus est nettement plus faible pour le modèle à 2 variables que pour le modèle à 3 variables, ce qui confirme bien l'inutilité de l'une des variables.

Analyse des résidus

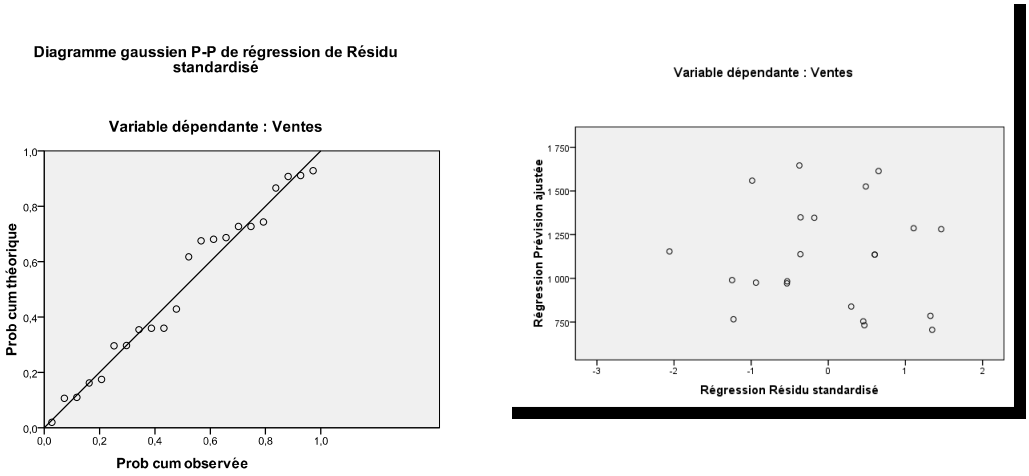
Quand un modèle est satisfaisant statistiquement et économiquement, il nous reste à vérifier que les hypothèses faites sur les résidus, la normalité, l'indépendance et l' homoscédasticité.

L'indépendance n'est facilement vérifiable que lorsque les variables sont temporelles, dans ce cas le plus simple est de représenter sur un graphique cartésien le résidu en t en fonction du résidu en t-1 (on peut aussi utiliser la statistique de Durbin-Watson).

On appelle résidu standardisés, les résidus divisés par l'erreur standard. L'option "enregistrer" du menu régression linéaire de SPSS permet de créer des variables associées aux résidus standardisés ou non, ainsi que les valeurs prédites (moyennes) par le modèle et l'intervalle de confiance de cette moyenne.

Normalité et homoscédasticité des résidus

Pour vérifier l'indépendance, on pourra utiliser le graphique normal ou un histogramme, pour l'homoscédasticité, plutôt que de faire un graphique avec chacune des variables explicatives, il est plus simple de faire un graphique des résidus (ou résidus standardisés) en fonction des estimations $(\hat{y}_i)_{1\leq i\leq n}$ ce qui résume l'ensemble des graphiques. Sur le modèle retenu pour l'exemple, les deux graphiques sont les suivants :



Su le graphique de gauche, les points sont bien alignés sur la diagonale, il n'y a pas lieu de remettre en cause la normalité des résidus, sur le graphique de gauche on ne remarque aucune forme particulière du nuage, qui est bien "équilibré" autour de l'axe des abscisses, l'homoscédasticité ne semble pas non plus à remettre en cause.

Choix d'un modèle de régression

En pratique, il est fréquent de se trouver face à plusieurs modèles satisfaisant tant statistiquement qu'économiquement, se pose alors le problème du choix du modèle. Nous avons vu que le coefficient n'était pas un bon indicateur pour choisir entre différents modèles, quand le nombre de variables explicatives n'est pas le même pour tous les modèles.

L'indicateur qui nous semble le plus approprié pour choisir un modèle est l'erreur type de régression, elle donne une indication non biaisée sur la dispersion autour de la valeur moyenne calculée par la partie déterministe du modèle. Il est toutefois important de distinguer entre un modèle descriptif et un modèle prédictif, si le modèle est uniquement descriptif (pour valider une théorie par exemple), le modèle de moindre erreur type s'impose, c'est celui qui fournira le plus d'indications sur les variations de la variable à expliquer. En revanche, si le modèle est à usage prédictif, il sera important alors de prendre aussi en compte la facilité qu'aura le décideur à prévoir la valeur des variables explicatives, on aura alors tendance à privilégier un modèle ne faisant intervenir que des variables explicatives sous le contrôle du décideur.

7.6.Les variables qualitatives dans le modèle de régression

Très souvent l'étude des variations d'une variable à expliquer peut se faire à l'aide de variables quantitatives, par exemple les ventes d'un produit de grande consommation dans une population de points de points de ventes peuvent s'expliquer par la région, le type de magasin; le type de promotion du produit etc.. Nous prendrons l'exemple dont les données sont dans le classeur Enseignes.xls : un fabricant distribue des produits de jardinage sous trois enseignes de magasin (codées de 1 à 3) et dans quatre régions différentes (codées de 1 à 4). Il a recueilli les résultats de 25 magasins et voudrait déterminer si l'enseigne et/ou la région ont une influence significative sur les ventes :

Ventes (100€)	Enseigne	Région	Ventes (100€)	Enseigne	Région
266	2	3	103	1	1
179	3	4	261	3	3
178	3	2	360	2	2
112	1	1	324	2	2
117	1	1	463	2	4
107	1	1	260	1	1
265	3	4	215	3	3
146	1	1	384	2	2
279	2	4	121	1	1
171	1	1	125	3	1
233	1	1	214	1	4
365	3	3	144	1	2

Il est donc nécessaire de coder convenablement ces variables pour pouvoir les utiliser dans notre modèle de régression. Il nous faudra ensuite pourvoir décider si une variable qualitative a une réelle influence sur les variations de la variable à expliquer.

Le codage d'une variable qualitative – Les indicatrices.

Une variable qualitative organise les unités statistiques en catégories identifiées par une modalité, qu'il est d'usage de coder numériquement de *1* à *m*, *m* étant le nombre de modalités. Il n'est pas possible d'utiliser directement ce codage, supposons en effet que ce soit le cas, nous aurions alors le modèle théorique suivant (en ne faisant intervenir que cette variable) :

$$Y_x = \beta_0 + \beta_1 x + \varepsilon \text{ où x prend les valeurs } 1,2,\ldots,m.$$

Ce qui impliquerait donc, en notant μ_i la moyenne de la variable *Y* restreinte à la sous population présentant la modalité *i,* :

$$\mu_1 = \beta_0 + \beta_1, \mu_2 = \beta_0 + 2\beta_1, \ldots, \mu_i = \beta_0 + i\beta_1, \ldots, \mu_m = \beta_0 + m\beta_1$$

ce qui signifie que les modalités sont ordonnées de telle façon que ces moyennes soient croissantes (si β_1 est positif) ou décroissantes (si β_1 est négatif), et que de plus la différence entre deux moyennes pour de modalités consécutives est constante ($=\beta_1$). Clairement ces hypothèses ont peu de chances de se réaliser dans la pratique, il nous faut donc coder différemment les variables explicatives qualitatives. Nous devons isoler les influences de chaque modalité sur les variations de la variable à expliquer, il est alors naturel d'introduire des variables indicatrices de chacune des modalités, c'est à dire pour chaque modalité une variable prenant la valeur 1 si l'individu statistique présente cette modalité, 0 sinon.

Donc si X_1 est une variable qualitative présentant *m* modalités on introduira *m* variables indicatrices :

$$\text{pour} 1 \leq j \leq m \quad X_{1j} = 1 \quad \text{si} \quad \mathrm{X}_1 = m \quad , \quad X_{1j} = 0 \quad \text{sinon}$$

Toutefois ce codage n'est pas encore parfait dans la mesure où les variables ainsi créées ne sont pas indépendantes, mais sont liées par la relation :

$$\sum_{j=1}^{m} X_{1j} = 1$$

ce qui signifie qu'un individu statistique présente une modalité et une seule. Un modèle de régression incluant les m variables ne peut donc être déterminé, puisqu'il suffirait de remplacer l'une des variables par l'opposé de la somme des autres pour avoir un modèle équivalent. Il nous faudra donc éliminer l'une quelconque de ces variables pour obtenir un modèle déterminable. Si toutes les variables incluses dans le modèle prennent la valeur 0, ceci signifie que l'individu pris en compte présente la modalité associée à la variable absente de la régression.

Création des indicatrices sous SPSS

La création des indicatrices se fait sous SPSS en utilisant le menu Transformer/Créer de nouvelles variables. Il n'est bien sur utile de créer que *m*-1 indicatrices. Nous avons créer ici les variables Enseigne1, Enseigne2, Région1, Région2, Région3.

Interprétation des coefficients du modèle

Nous allons nous placer par le cas d'une seule variable explicative qualitative à m modalités *X*, représentées par m-1 variables indicatrices $\left(X_j\right)_{1\leq j\leq m-1}$ dans la régression, le modèle est alors le suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{m-1} X_{m-1} + \varepsilon$$

Les seules valeurs possibles pour X_j sont 1 ou 0, mais une seule des variables au plus est non nulle, si toutes les variables sont nulles, ce qui correspond à l'appartenance à la modalité absente *m* par exemple, la moyenne $\mu_m = \beta_0$, si seule la variable indicatrice X_1 est non nulle la moyenne correspondante est $\mu_1 = \beta_0 + \beta_1$, de manière générale si seule la variable X_j est non nulle la moyenne correspondant à cette modalité est $\mu_j = \beta_0 + \beta_j$. Aux coefficients de la régression on peut donc associer :

- Pour le coefficient constant : la moyenne de la variable *Y* restreinte à la sous population présentant la modalité absente. Cette modalité sera la modalité de référence.
- Pour les autres coefficients : la différence des moyennes entre variable *Y* restreinte à la sous population présentant la modalité *j* et la variable *Y* restreinte à la sous population présentant la modalité absente.

Le test partiel de Student revient donc à vérifier que les moyennes entre une modalité et la modalité absente sont différentes. On a donc une généralisation du test de comparaison de deux moyennes, vu dans le chapitre précédent. Notons cependant que l'hypothèse d'homoscédasticité des résidus revient à ne faire le test qu'en supposant les variances égales sur chacune des sous populations.

L'estimation b_0 est simplement la moyenne des valeurs de *Y* pour les individus de l'échantillon présentant la modalité absente, de même l'estimation $b_0 + b_j$ est la moyenne des valeurs de *Y* pour les individus de l'échantillon présentant la modalité *j*.

Sur notre exemple nous obtenons le tableau du modèle suivant :

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	226,857	25,295		8,968	,000
	Enseigne1	-69,766	32,357	-,353	-2,156	,043
	Enseigne2	119,143	37,233	,524	3,200	,004

a. Variable dépendante : Ventes

La modalité de référence est la modalité 3, les estimations des moyennes des ventes dans les magasins par enseigne sont les suivantes

- Enseigne 3 (constante de la régression b_0) : 226,86*100€=22 686€.
- Enseigne 1 $(b_0 + b_1)$: (226,86-69,77)*100€ = 157,09*100€=15 709€
- Enseigne 1 $(b_0 + b_2)$: (226,86+119,14)*100€ = 346,10*100€=34 610€

Comme tous les t_c sont significatifs au risque de première espèce de 5%, on peut donc considérer qu'il y a une différence significative entre les enseignes, qui seront classées dans l'ordre croissant des ventes : Enseigne 1, Enseigne 3, Enseigne 2.

Test de l'influence d'une variable qualitative

Si nous introduisons dans le modèle précédent les variables indicatrices de la région (des trois premières régions) nous obtenons le tableau du modèle suivant :

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	235,559	37,096		6,350	,000
	Enseigne1	-21,465	45,861	-,109	-,468	,645
	Enseigne2	121,836	40,856	,536	2,982	,008
	Région1	-66,740	47,968	-,334	-1,391	,181
	Région2	-26,367	43,623	-,109	-,604	,553
	Région3	10,732	47,196	,041	,227	,823

a. Variable dépendante : Ventes

Il y a dans le modèle, plusieurs variables indicatrices non significatives marginalement. Nous pourrions éliminer les unes après les autres les variables non significatives marginalement, mais en faisant cela nous ne tiendrions pas compte du fait que les variables ont une signification "par bloc".

Principe du test

Comme nous l'avons fait pour une variable quantitative il serait en fait plus intéressant de pouvoir tester l'influence marginale d'une variable qualitative quand d'autres variables sont dans la régression. Le problème est ici différent dans la mesure où nous serons conduits à tester l'influence marginale d'un groupe de variables (les variables indicatrices associées à la variable qualitative) et non plus d'une seule variable. Nous nous intéresserons ici au test de l'influence d'un groupe de *m* variables explicatives parmi *p*, que ces variables correspondent à une variable qualitative ou non.

Pour simplifier les notations, et sans rien perdre de la généralité du propos, nous supposons que le groupe de *m* variables dont nous voulons tester l'influence marginale sont les *m* dernières $X_{p-m+1}, X_{p-m+2}, \ldots, X_p$. Le test se posera alors de la façon suivante :

$$\begin{aligned} H_0 &: \beta_{p-m+1} = \beta_{p-m+2} = \ldots = \beta_p \\ H_1 &: \exists j \in [1,m] \quad \beta_{p-j} \neq 0 \end{aligned}$$

Nous serons donc conduits à comparer deux modèles :

- Le modèle dit complet, comprenant les *p* variables explicatives. Nous noterons respectivement *SCEC* et *SCRC* la somme des carrés expliquée et la somme des carrés résiduel de ce modèle et R_C^2 son coefficient de détermination. *SCT* désignera la somme des carrés totale qui est la même pour tous les modèles.
- Le modèle dit partiel ne comprenant que les *p-m* premières variables explicatives. Nous noterons *SCEP* la somme des carrés expliquée de ce modèle, R_P^2 son coefficient de détermination.

Le principe du test sera identique à celui du test global : si les *m* variables explicatives supplémentaires ne sont pas plus intéressantes que les variables associées à la partie résiduelle du modèle complet, autant les laisser dans cette partie. Pour juger de l'apport des m variables explicatives supplémentaires, il suffit de prendre comme indicateur la diminution de la somme des carrés due à leur introduction dans le modèle ; pour pouvoir le comparer aux résidus on utilisera en fait la diminution moyenne par variable introduite dans le modèle. La statistique que nous utiliserons, appelée statistique de Fisher Partiel, sera alors :

$FP = \dfrac{(SCEC - SCEP)/m}{SCRC/(n-p-1)}$ en divisant numérateur et dénominateur par SCT on obtient une définition équivalente souvent utilisée dans la littérature statistique $FP = \dfrac{(R_C^2 - R_P^2)/m}{(1-R_C^2)/(n-p-1)}$.

Sous l'hypothèse nulle cette statistique suit une loi de Fisher-Snedecor à (*m,n-p-1*) degrés de liberté, comme pour la statistique F globale, on rejette l'hypothèse H_0 si la valeur observée est

suffisamment grande, la valeur critique F_α est déterminée en fonction du risque de première espèce α par la formule $prob(FS(m, n-p-1) > F_\alpha) = \alpha$. Nous utiliserons le niveau de signification définie en fonction de la valeur observée pour la statistique sur l'échantillon FP_c : $ns = prob(FS(m, n-p-1) > FP_c)$. Si ce niveau est inférieur à α, l'hypothèse H_0 est rejetée.

Remarques :

- Dans le cas particulier $m = p$, on retrouve le test global de la régression.
- Dans le cas $m = 1$, on retrouve le test marginal sous une autre forme, on peut en effet démontrer les deux résultats suivant : $t_c^2 = FP_c$ et la loi de Fisher-Snedecor à (1,*n-p*-1) degrés de liberté est égale au carré de la loi de Student à *n-p*-1 degrés de liberté.

Tableau d'analyse de la variance

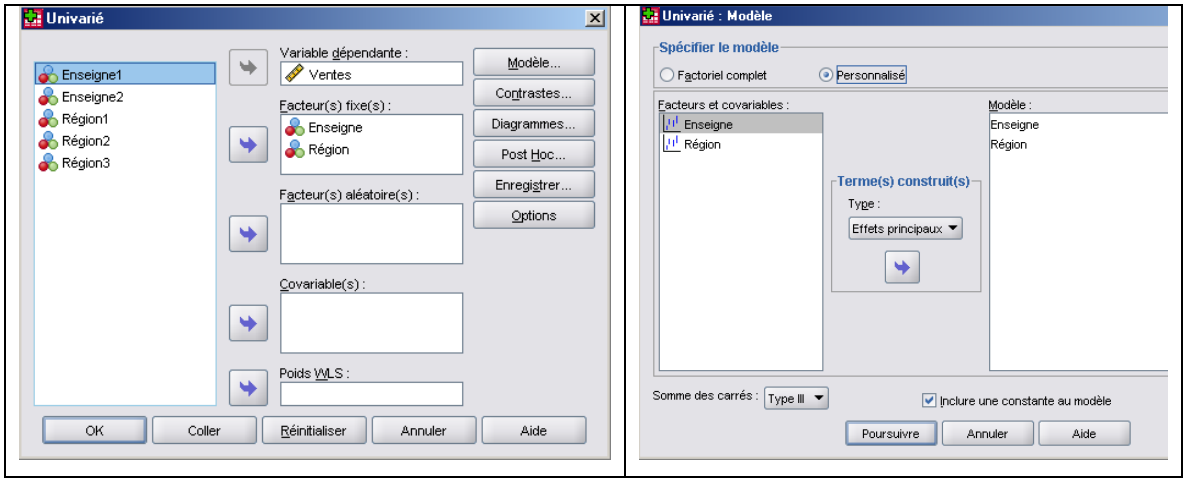
Il est d'usage de présenter le résultat du test par un tableau, permettant l'analyse marginale de deux groupes de variables. Supposons que les *p* variables explicatives soient divisées en deux groupes G_m et G_{p-m} de variables contenant respectivement *m* et *p-m* variables. Nous noterons SCE_m la somme des carrés expliquée par le groupe de *m* variables et SCE_{p-m} celle du groupe de *p-m* variables. Le tableau dit d'analyse de la variance se présente sous la forme suivante :

Source	Somme des Carrés	DL	Carré Moyen	F	*ns* =Prob >F
Complet	$SCEC$	p	$\frac{SCEC}{p} = SME$	$f_g = \frac{SME}{SCRM}$	$prob(F_{p,n-p-1} > f_g)$
G_m	$SCEC\text{-}SCE_{p-m} = S_m$	m	$\frac{S_m}{m} = SM_m$	$f_m^p = \frac{SM_m}{SCRM}$	$prob(F_{m,n-p-1} > f_m^p)$
G_{p-m}	$SCEC\text{-}SCE_m = S_{p-m}$	$p-m$	$\frac{S_{p-m}}{p-m} = SM_{p-m}$	$f_{p-m}^p = \frac{SM_{p-m}}{SCRM}$	$prob(F_{p-m,n-p-1} > f_{p-m}^p)$
Résidus	$SCRC$	$n-p-1$	$\frac{SCRC}{n-p-1} = SCRM$		
Totale	SCT	$n-1$			

La première ligne du tableau correspond à l'analyse de la variance du modèle complet, elle permet de tester l'influence globale des variables explicatives, les deux lignes suivantes permettent de tester l'influence marginale de chacun des groupes de variables G_m et G_{p-m}. Si l'un des deux niveaux de signification est supérieur à a, ce groupe de variables peut être ôté de la régression.

Mise en œuvre sous SPSS

Sous SPSS, on utilisera dans le menu Analyse, le sous menu Modèle Linéaire Général/univarié. La variable dépendante est la variable à expliquer, les facteurs fixes sont les variables qualitatives non recodées, les cofacteurs sont les variables quantitatives explicatives (ici aucune). Dans l'option modèle on ne demandera pas d'interaction :



Ce qui donne les valeurs (modèle corrigé prenant en compte enseigne et région) :

Tests des effets inter-sujets

Variable dépendante:Ventes

Source	Somme des carrés de type III	ddl	Moyenne des carrés	D	Sig.
Modèle corrigé	150023,457[a]	5	30004,691	6,536	,001
Enseigne	53141,374	2	26570,687	5,788	,011
Région	11427,890	3	3809,297	,830	,495
Erreur	82627,876	18	4590,438		
Total corrigé	232651,333	23			

a. R deux = ,645 (R deux ajusté = ,546)

On constate sur ce tableau que la variable Région n'a aucun apport marginal significatif, puisque son niveau de signification est de 50% environ, très largement supérieur au risque habituel de 5%.

Comme nous avons vu plus haut que le modèle Ventes/Enseigne était valable statistiquement nous ne garderons que la variable qualitative Enseigne.

7.7.La régression pas à pas

Pour un nombre donné *p* de variables explicatives candidates pour un modèle de régression linéaire, le nombre de modèle possible est égal au nombre de parties non vides d'un ensemble à *p* éléments soit 2^p-1, pour *p*=5 cela fait déjà 31 modèles possibles, parmi lesquels il faudra choisir un ou plusieurs modèles statistiquement et économiquement valable. Il serait donc utile d'avoir une méthode systématique permettant d'obtenir un bon modèle.

Principe de la méthode

Dans la mesure où il n'existe pas de critère rationnel permettant de dire si un modèle est meilleur qu'un autre, il n'est pas ici question d'optimisation, mais simplement d'obtenir un modèle valable statistiquement. Les méthodes pour atteindre ces résultats sont des méthodes pas à pas reposant sur la statistique t de Student, à chaque étape on introduit la variable la plus

marginalement significative ou on retire la variable la moins significative. Nous n'exposerons ici que la méthode la plus "naturelle", la procédure descendante ou "backward".

La méthode retire à chaque étape une variable du modèle construit à l'étape précédente. Au début de l'algorithme les *p* variables sont présentes dans le modèle. Un seuil de sortie α est fixé qui correspond à la valeur maximale du niveau de signification d'une variable pour qu'elle soit conservée dans la régression (ou ce qui revient au même une valeur minimale de t_c).

A l'étape k, si toutes les variables du modèle ont un niveau de signification supérieur à α, la méthode s'arrête et le modèle est conservé ; sinon parmi les variables qui ont un niveau de signification inférieur à α, on élimine la variable ayant le plus grand niveau de signification et on itère la procédure.

La procédure s'arrêtera donc lorsque l'une des deux conditions suivante sera vérifiée :

- Toutes les variables sont retirées du modèle
- Les variables présentes dans le modèle ont toutes un niveau de signification supérieur à α.

Bien évidemment, le modèle final dépend de la valeur du seuil retenu, plus ce seuil est faible, moins il restera de variables dans le modèle final.

Cette procédure n'est en rien optimale, elle ne remet jamais en cause l'élimination d'une variable. Or il est possible qu'une variable qui a été sortie du modèle au cours des premières étapes, du fait de sa corrélation à d'autres variables du modèle, se trouve finalement avoir un apport marginal significatif par rapport au modèle final, dans la mesure où certaines des variables corrélées ont été éliminées après elle.

Un exemple

Nous avons déjà vu une illustration de cette méthode au paragraphe 0 pour le premier exemple, il était possible de pratiquer cette procédure car les données étaient bien disposées pour l'élimination de la variable non significative, qui ne séparait l'ensemble des variables explicatives. Nous allons illustrer cette méthode sur le deuxième exemple, les ventes en fonction des enseignes et des régions, en prenant un risque de première espèce α=5%. Dans la boîte de dialogue régression de SPSS, nous choisissons dans le bloc de variables explicatives (ou indépendantes) la méthode descendante.

Le listing produit est composé des éléments suivants :

La liste de la variable éliminée à chaque étape :

Variables introduites/supprimées[b]

Modèle	Variables introduites	Variables supprimées	Méthode
1	Région3, Enseigne2, Région2, Enseigne1, Région1[a]	.	Entrée
2	.	Région3	Elimination descendante (critère : Probabilité de F pour éliminer >= ,100).
3	.	Enseigne1	Elimination descendante (critère : Probabilité de F pour éliminer >= ,100).
4	.	Région2	Elimination descendante (critère : Probabilité de F pour éliminer >= ,100).

a. Toutes variables requises saisies.

b. Variable dépendante : Ventes

Les caractéristiques de chaque modèle (on peut remarquer que le R^2 diminue, mais ni le R^2 ajusté, ni l'erreur standard) :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,803[a]	,645	,546	67,753
2	,802[b]	,644	,569	66,040
3	,799[c]	,638	,584	64,883
4	,789[d]	,623	,587	64,642

a. Valeurs prédites : (constantes), Région3, Enseigne2, Région2, Enseigne1, Région1

b. Valeurs prédites : (constantes), Enseigne2, Région2, Enseigne1, Région1

c. Valeurs prédites : (constantes), Enseigne2, Région2, Région1

d. Valeurs prédites : (constantes), Enseigne2, Région1

Les différents tableaux d'analyse de la variance :

ANOVA[e]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	150023,457	5	30004,691	6,536	,001[a]
	Résidu	82627,876	18	4590,438		
	Total	232651,333	23			
2	Régression	149786,081	4	37446,520	8,586	,000[b]
	Résidu	82865,253	19	4361,329		
	Total	232651,333	23			
3	Régression	148454,479	3	49484,826	11,755	,000[c]
	Résidu	84196,854	20	4209,843		
	Total	232651,333	23			
4	Régression	144900,958	2	72450,479	17,339	,000[d]
	Résidu	87750,375	21	4178,589		
	Total	232651,333	23			

a. Valeurs prédites : (constantes), Région3, Enseigne2, Région2, Enseigne1, Région1

b. Valeurs prédites : (constantes), Enseigne2, Région2, Enseigne1, Région1

c. Valeurs prédites : (constantes), Enseigne2, Région2, Région1

d. Valeurs prédites : (constantes), Enseigne2, Région1

e. Variable dépendante : Ventes

Enfin les différents modèles, où il est possible de retrouver la démarche "backward" :

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	235,559	37,096		6,350	,000
	Enseigne1	-21,465	45,861	-,109	-,468	,645
	Enseigne2	121,836	40,856	,536	2,982	,008
	Région1	-66,740	47,968	-,334	-1,391	,181
	Région2	-26,367	43,623	-,109	-,604	,553
	Région3	10,732	47,196	,041	,227	,823
2	(Constante)	241,224	26,791		9,004	,000
	Enseigne1	-23,975	43,389	-,121	-,553	,587
	Enseigne2	119,987	39,026	,528	3,074	,006
	Région1	-70,147	44,416	-,351	-1,579	,131
	Région2	-30,421	38,808	-,125	-,784	,443
3	(Constante)	236,229	24,778		9,534	,000
	Enseigne2	126,979	36,271	,558	3,501	,002
	Région1	-86,729	32,170	-,434	-2,696	,014
	Région2	-34,417	37,460	-,142	-,919	,369
4	(Constante)	227,625	22,854		9,960	,000
	Enseigne2	118,375	34,911	,521	3,391	,003
	Région1	-78,125	30,662	-,391	-2,548	,019

a. Variable dépendante : Ventes

Le dernier modèle est à la fois valide globalement et marginalement, il est donc acceptable statistiquement.

Remarques :

- Le modèle obtenu par régression pas à pas backward n'est pas le même que celui obtenu par analyse du F partiel.
- La variable explicative Région1 n'était pas significative dans les deux premières étapes du processus, ceci était du à une forte corrélation entre cette variable et la variable Enseigne1, c'est ce qui explique le résultat final : les enseignes sont en fait un facteur explicatif des variations des ventes. Si la région apparaît ici c'est uniquement du à un biais qui est la sur représentation de l'enseigne 1 dans la région1.

Un dernier tableau donne pour chaque régression, la validité éventuelle de chacune des variables qui ont été exclues, si elle était introduit dans le modèle de cette étape. La tolérance est une indication de colinéarité entre la variable hors régression et l'ensemble des variables dans la régression :

Variables exclues[d]

Modèle		Bêta dans	t	Sig.	Corrélation partielle	Statistiques de colinéarité
						Tolérance
2	Région3	,041[a]	,227	,823	,054	,618
3	Région3	,061[b]	,358	,724	,082	,656
	Enseigne1	-,121[b]	-,553	,587	-,126	,389
4	Région3	,109[c]	,726	,476	,160	,813
	Enseigne1	-,153[c]	-,718	,481	-,159	,403
	Région2	-,142[c]	-,919	,369	-,201	,758

a. Valeurs prédites dans le modèle : (constantes), Enseigne2, Région2, Enseigne1, Région1

b. Valeurs prédites dans le modèle : (constantes), Enseigne2, Région2, Région1

c. Valeurs prédites dans le modèle : (constantes), Enseigne2, Région1

d. Variable dépendante : Ventes

8. Exercices de regression linéaire

8.1. Régression simple : Prix des forfaits de ski (Forfait.sav)

On veut étudier le prix des forfaits en fonction de l'étendue en kms du domaine skiable. On a relevé un échantillon de 42 stations :

Station	Kms	Prix	Station	Kms	Prix
Auron	135 km	125 €	Morillon	145 km	154 €
Ax les thermes	75 km	140 €	Morzine	107 km	141 €
Chatel	83 km	140 €	Orcières Merlette	100 km	126 €
Isola 2000	120 km	125 €	Pra loup	180 km	136 €
La Clusaz	128 km	150 €	Praz sur arly	120 km	115 €
La Joue du Loup	100 km	143 €	Risoul	180 km	149 €
La Mongie	100 km	150 €	Saint Jean d'Arves	90 km	129 €
La Norma	65 km	115 €	Saint Lary soulan	100 km	166 €
La Plagne	225 km	192 €	Saint Sorlin	120 km	132 €
La Rosière	150 km	161 €	Samoens	265 km	187 €
La Tania	150 km	174 €	Serre Chevalier	250 km	281 €
Le Corbier	90 km	125.5	Superdévoluy	100 km	143 €
Le grand bornand	82 km	121 €	Val cenis	80 km	126 €
Les 2 Alpes	220 km	172 €	Val d'Allos	180 km	141 €
les Arcs	200 km	198 €	Val thorens	140 km	169 €
les Menuires	160 km	169 €	Valfréjus	65 km	110 €
Les Orres	88 km	122 €	Valloire/Valmeinier	150 km	140 €
les saisies	62 km	142 €	Valmorel	150 km	175 €
Méribel	150 km	178 €	Vars	180 km	149 €
Molines	38 km	107 €	Vaujany	32 km	116 €
Montgenevre	100 km	144 €	Villard de Lans	125 km	134 €

Analyse de l'ensemble du fichier

En utilisant les annexes 1, 2, 3 :

1) L'hypothèse d'une liaison linéaire vous semble-t-elle réaliste ?
2) Analyser les résultats de la régression simple, et interpréter les coefficients.
3) Analyser les résidus standards.

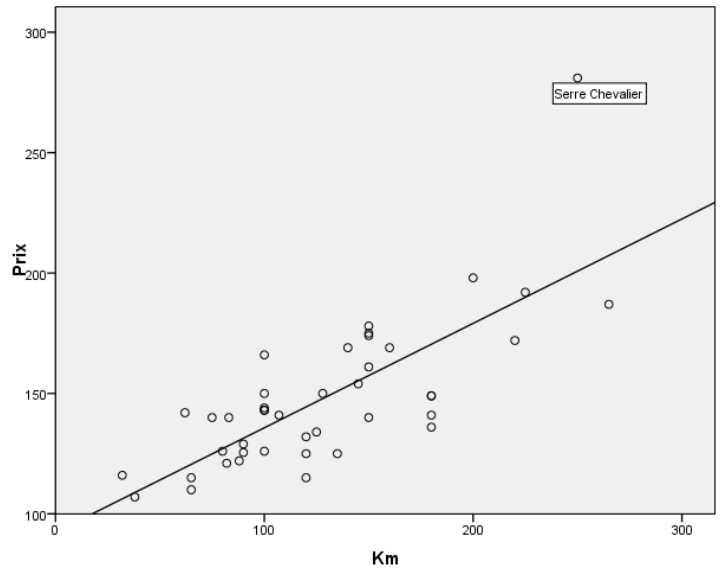
Analyse sans Serre Chevalier

Un statisticien propose de faire l'analyse sans la station Serre Chevalier (pourquoi ?)

Les résultats vous sont donnés en annexe 4.

1) Analyser les résultats et interpréter les résultats. Quels serait la prévision ponctuelle pour Serre Chevalier ?
2) Donner un intervalle de confiance des coefficients (au degré de confiance de 0,95). Qu'en concluez-vous ?

Annexe 1 – Graphique et corrélation



Corrélations		
	Prix	Km
Prix	1	,756
Km	,756	1

Annexe 2 – Résultats de la régression

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,756	,571	,560	20,594

ANOVA

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	22596,541	1	22596,541	53,280	,000
	Résidu	16964,418	40	424,110		
	Total	39560,958	41			

Coefficients

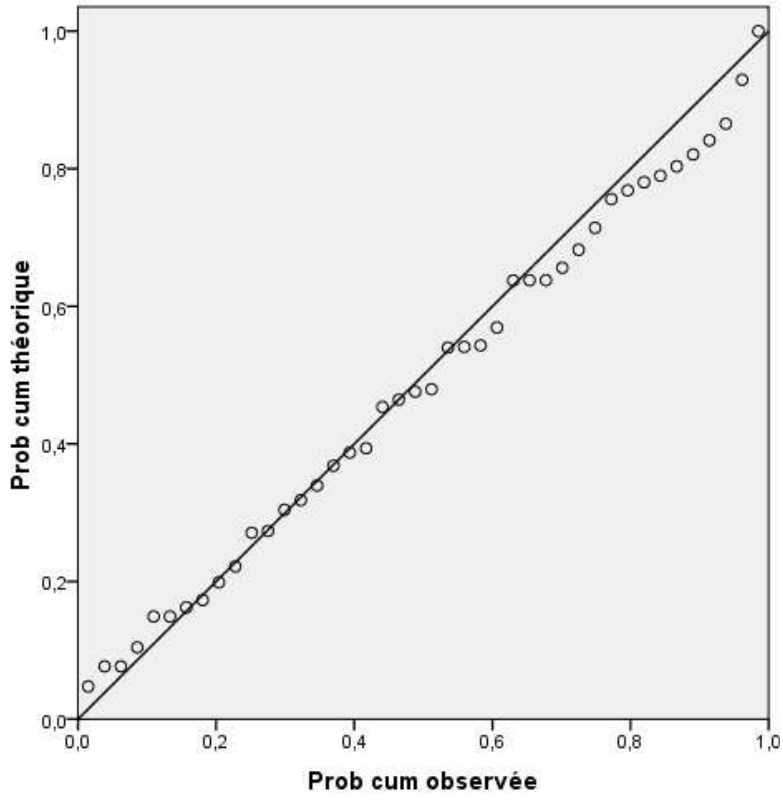
Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	92,366	8,247		11,200	,000
	Km	,434	,059	,756	7,299	,000

Annexe 3 - Résidus

Station	Résidus	Standards	Station	Résidus	Standards

Auron	-25,91106	-1,25819	Morillon	-1,24776	-0,06059
Ax les thermes	15,10917	0,73367	Morzine	2,23171	0,10837
Chatel	11,6398	0,56521	Orcières Merlette	-9,73259	-0,47259
Isola 2000	-19,406	-0,94232	Pra loup	-34,42623	-1,67167
La Clusaz	2,12464	0,10317	Praz sur arly	-29,406	-1,4279
La Joue du Loup	7,26741	0,35289	Risoul	-21,42623	-1,04041
La Mongie	14,26741	0,6928	Saint Jean d'Arves	-2,39589	-0,11634
La Norma	-5,55413	-0,2697	Saint Lary soulan	30,26741	1,46972
La Plagne	2,05861	0,09996	Saint Sorlin	-12,406	-0,60241
La Rosière	3,58389	0,17403	Samoens	-20,28821	-0,98515
La Tania	16,58389	0,80528	Serre Chevalier	80,21685	3,89517
Le Corbier	-5,89589	-0,28629	Superdévoluy	7,26741	0,35289
Le grand bornand	-6,92653	-0,33634	Val cenis	-1,05919	-0,05143
Les 2 Alpes	-15,77304	-0,76591	Val d'Allos	-29,42623	-1,42888
les Arcs	18,90037	0,91776	Val thorens	15,92059	0,77307
les Menuires	7,24718	0,35191	Valfréjus	-10,55413	-0,51249
Les Orres	-8,52855	-0,41413	Valloire/Valmeinier	-17,41611	-0,84569
les saisies	22,74688	1,10454	Valmorel	17,58389	0,85384
Méribel	20,58389	0,99951	Vars	-21,42623	-1,04041
Molines	-1,84503	-0,08959	Vaujany	9,75699	0,47378
Montgenevre	8,26741	0,40145	Villard de Lans	-12,57435	-0,61059

Diagramme gaussien des résidus standardisés



Annexe 4 – Régressions sans Serre Chevalier

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,749	,560	,549	15,539

ANOVA

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	12001,429	1	12001,429	49,707	,000
	Résidu	9416,376	39	241,446		
	Total	21417,805	40			

Coefficients

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	102,354	6,474		15,810	,000
	Km	,338	,048	,749	7,050	,000

8.2. L'entreprise Elec (Elec.sav)

L'entreprise Elec vend du matériel électrique et souhaite évaluer l'importance relative de l'influence de ses vendeurs et des prix sur ses ventes. Pour faire cette évaluation, l'entreprise a réparti ses clients en un certain nombre de zones géographiques. Pour chacune de ces zones, les variables suivantes ont été mesurées :

- Les ventes
- Le nombre de vendeurs pour la zone
- La moyenne des prix facturés par l'entreprise dans cette zone
- La moyenne des prix facturés par la concurrence dans cette zone
- L'indice des prix dans cette zone; l'indice 100 étant l'inde de la France métropolitaine.

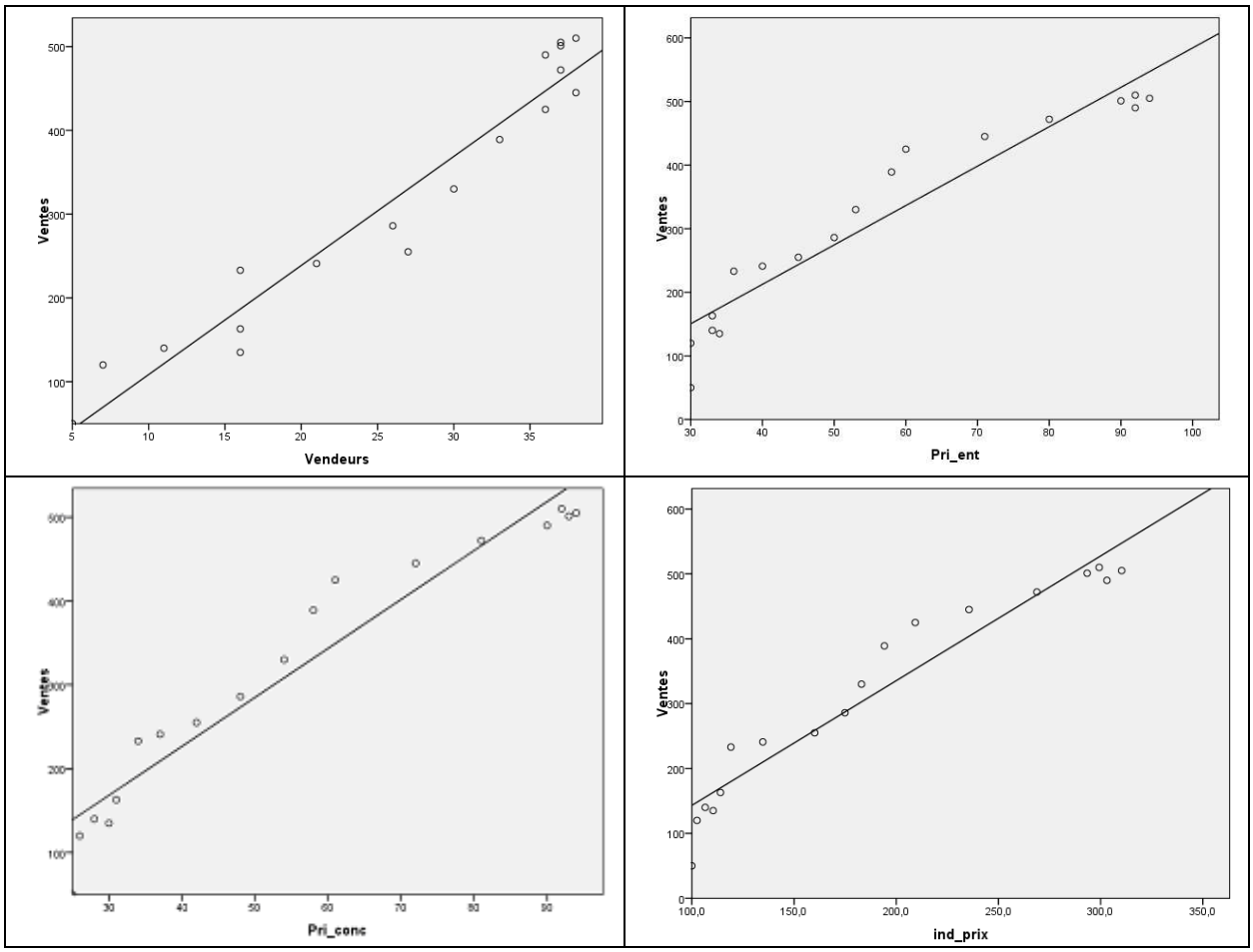
Les données ont été recueillies sur 18 zones. On prendra pour toutes les questions α=0,01 comme risque de première espèce.

1) Analyser la matrice de corrélations données en annexe.
2) Quelle est la meilleure variable explicative, prise isolément? Quel est le meilleur couple de variables explicatives?
3) Que pensez vous des résultats de la régression complète (avec les quatre variables explicatives)? En particulier, le modèle est-il valide? Quelles sont les variables dont la significativité peut être affirmée ?

4) Comment expliquez-vous que certaines variables significatives dans un modèle à deux variables ne le soient plus dans le modèle à quatre variables?
5)
   a) Comment est choisie la première variable explicative de la méthode "Stepwise"?
   b) Et la seconde?
   c) Pourquoi le modèle "Stepwise" ne comporte-t-il que deux variables?
6) Quel est le meilleur modèle possible, en fonction des variables explicativesdisponibles?

Annexe 1 – Tableau des corrélations et graphiques

	Ventes	Vendeurs	Pri_ent	Pri_conc	ind_prix
Ventes	1	,969**	,954**	,966**	,962**
Vendeurs	,969**	1	,889**	,906**	,906**
Pri_ent	,954**	,889**	1	,998**	,998**
Pri_conc	,966**	,906**	,998**	1	,998**
ind_prix	,962**	,906**	,998**	,998**	1



Régressions à une seule variable

Variable explicative : Vendeurs

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,969[a]	,939	,935	39,670

a. Valeurs prédites : (constantes), Vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	384514,013	1	384514,013	244,332	,000[a]
	Résidu	25179,765	16	1573,735		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Vendeurs

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-21,634	23,544		-,919	,372
	Vendeurs	13,018	,833	,969	15,631	,000

a. Variable dépendante : Ventes

Variable explicative : prix de l'entreprise

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,954[a]	,910	,905	47,904

a. Valeurs prédites : (constantes), Pri_ent

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	372976,916	1	372976,916	162,531	,000[a]
	Résidu	36716,861	16	2294,804		

	Total	409693,778	17			

a. Valeurs prédites : (constantes), Pri_ent

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-35,263	29,785		-1,184	,254
	Pri_ent	6,195	,486	,954	12,749	,000

a. Variable dépendante : Ventes

Variable explicative : Prix de la concurrence

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,966[a]	,933	,929	41,501

a. Valeurs prédites : (constantes), Pri_conc

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	382135,883	1	382135,883	221,867	,000[a]
	Résidu	27557,895	16	1722,368		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Pri_conc

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-6,783	23,783		-,285	,779
	Pri_conc	5,835	,392	,966	14,895	,000

a. Variable dépendante : Ventes

Variable explicative : indice des prix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,962[a]	,926	,921	43,525

a. Valeurs prédites : (constantes), ind_prix

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	379382,724	1	379382,724	200,261	,000[a]
	Résidu	30311,053	16	1894,441		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), ind_prix

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-49,024	27,767		-1,766	,097
	ind_prix	1,922	,136	,962	14,151	,000

a. Variable dépendante : Ventes

Annexe 3 : Régressions à 2 variables explicatives

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,990[a]	,980	,977	23,545

a. Valeurs prédites : (constantes), Pri_ent, Vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	401378,056	2	200689,028	362,005	,000[a]
	Résidu	8315,721	15	554,381		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Pri_ent, Vendeurs

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-47,480	14,739		-3,222	,006
	Vendeurs	7,726	1,079	,575	7,158	,000
	Pri_ent	2,876	,522	,443	5,515	,000

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,991[a]	,982	,979	22,258

a. Valeurs prédites : (constantes), Pri_conc, Vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	402262,635	2	201131,317	405,990	,000[a]
	Résidu	7431,143	15	495,410		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Pri_conc, Vendeurs

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-30,728	13,297		-2,311	,035
	Vendeurs	7,034	1,104	,523	6,374	,000
	Pri_conc	2,970	,496	,492	5,985	,000

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,989[a]	,978	,976	24,245

a. Valeurs prédites : (constantes), ind_prix, Vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	400876,703	2	200438,351	340,995	,000[a]
	Résidu	8817,075	15	587,805		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), ind_prix, Vendeurs

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-51,655	15,473		-3,338	,004
	Vendeurs	7,271	1,202	,541	6,047	,000
	ind_prix	,943	,179	,472	5,276	,000

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,979[a]	,958	,952	34,056

a. Valeurs prédites : (constantes), Pri_conc, Pri_ent

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	392296,763	2	196148,382	169,122	,000[a]
	Résidu	17397,015	15	1159,801		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Pri_conc, Pri_ent

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	80,894	35,473		2,280	,038
	Pri_ent	-16,464	5,562	-2,536	-2,960	,010
	Pri_conc	21,128	5,177	3,497	4,081	,001

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,968[a]	,937	,929	41,440

a. Valeurs prédites : (constantes), ind_prix, Pri_ent

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	383934,960	2	191967,480	111,787	,000[a]
	Résidu	25758,818	15	1717,255		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), ind_prix, Pri_ent

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-67,250	28,709		-2,342	,033
	Pri_ent	-11,291	6,935	-1,739	-1,628	,124
	ind_prix	5,390	2,134	2,698	2,526	,023

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,966[a]	,933	,924	42,743

a. Valeurs prédites : (constantes), ind_prix, Pri_conc

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	382288,695	2	191144,348	104,622	,000[a]
	Résidu	27405,082	15	1827,005		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), ind_prix, Pri_conc

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	6,387	51,710		,124	,903
	Pri_conc	7,567	6,000	1,252	1,261	,227
	ind_prix	-,574	1,983	-,287	-,289	,776

a. Variable dépendante : Ventes

8.3.Les stylos Runild (Runild.sav)

Dans le cadre d'une étude sur l'efficacité commerciale de l'entreprise Le responsable des études a recueilli les informations suivantes :

- La distribution des produits est organisée en 40 zones géographiques
- Chaque zone est attribuée en exclusivité à un grossiste assisté par une équipe de représentants commerciaux. Le nombre de ces représentants est décidé par le grossiste et peut varier d'une zone à l'autre.

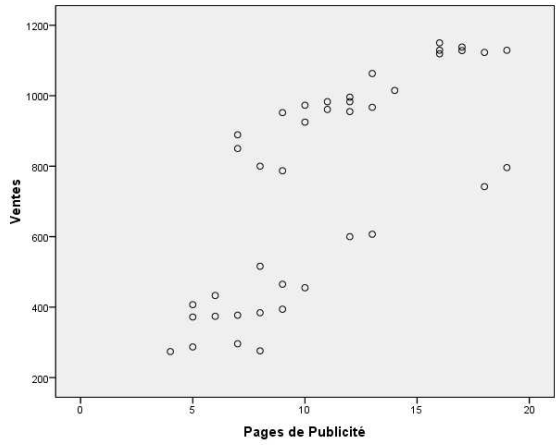
Chaque trimestre les grossistes sont évalués sur une échelle de 1 à 4. La valeur 4 indiquant que le grossiste est jugé très bon, la valeur 1 un grossiste jugé très mauvais. Dans chaque zone la publicité est faite essentiellement par la presse locale et la distribution à domicile. Le classeur Runild.xls donne pour les 40 zones géographiques :

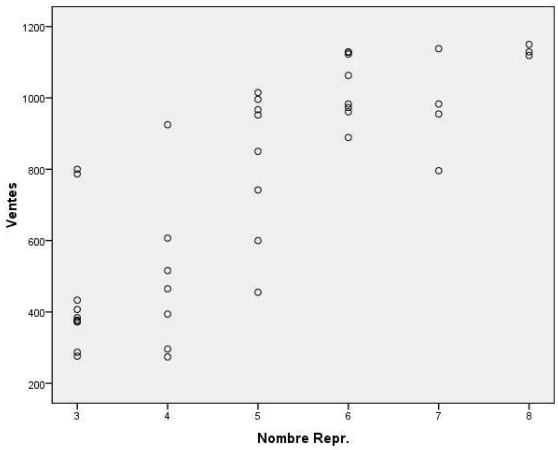
- Le volume des ventes mensuelles
- Le nombre mensuel de page de publicité
- Le nombre de représentants de l'équipe commerciale
- La note de qualité attribuée au grossiste

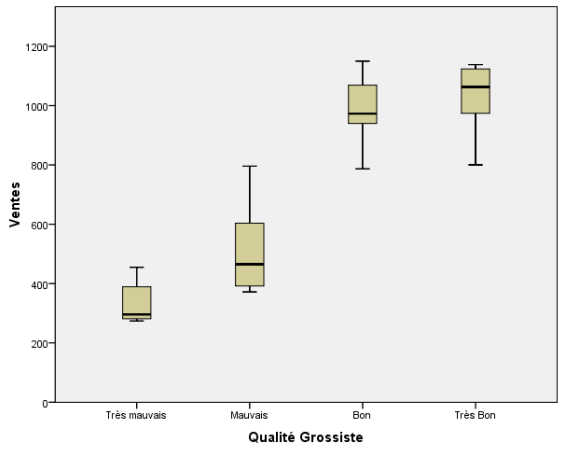
La régression linéaire

1) Etude des ventes en fonction des deux variables publicité et nombre de représentant.
   a) Le modèle de régression linéaire vous semble-t-il adapté?
   b) Quelle est l'influence de chacune des variables prise séparément sur les variations des ventes?
   c) Le modèle à deux variables est-il valide statistiquement et économiquement?
   d) Sachant que le coût mensuel moyen d'un représentant est de 2000€ et le coût moyen d'une page de publicité de 650€, pour quelle marge unitaire sur le produit est-il plus intéressant d'embaucher un représentant ou de faire une page de publicité supplémentaire.
2) Etude des ventes en fonction de la qualité du grossiste
   a) Le chargé d'étude considère que la note de qualité est une variable quantitative et procède à une régression simple sur cette variable. Analyser les résultats obtenus.
   b) Le directeur commercial n'est pas d'accord, il pense que l'on doit considérer cette variable comme qualitative à quatre modalités. Il demande de procéder à une étude en prenant la modalité 4 comme modalité de référence. Analyser les résultats. En prenant un risque a de 0,05 peut considérer que les modalités 3 et 4 sont différentes? Qu'en conclure?
   c) Quel modèle explicatif des variations des ventes en fonction de la qualité du grossiste vous paraît le mieux adapté?
3) Analyser le modèle construit avec les trois variables.

Représentations graphiques







Régression à 1 variable quantitative

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,744[a]	,554	,542	207,907

a. Valeurs prédites : (constantes), Pages de Publicité

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2040677,624	1	2040677,624	47,210	,000[a]
	Résidu	1642561,876	38	43225,313		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Pages de Publicité

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	172,902	90,432		1,912	,063
	Pages de Publicité	53,105	7,729	,744	6,871	,000

a. Variable dépendante : Ventes

La régression linéaire

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,801[a]	,642	,633	186,182

a. Valeurs prédites : (constantes), Nombre Repr.

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2366018,184	1	2366018,184	68,256	,000[a]
	Résidu	1317221,316	38	34663,719		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Nombre Repr.

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-17,775	97,684		-,182	,857
	Nombre Repr.	155,460	18,817	,801	8,262	,000

a. Variable dépendante : Ventes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,872[a]	,761	,755	152,220

a. Valeurs prédites : (constantes), Qualité Grossiste

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2802742,540	1	2802742,540	120,959	,000[a]
	Résidu	880496,960	38	23170,973		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Qualité Grossiste

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	58,305	67,489		,864	,393
	Qualité Grossiste	271,939	24,726	,872	10,998	,000

a. Variable dépendante : Ventes

Regression à 2 variables (Publicité, Représentants)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,828[a]	,685	,668	177,091

a. Valeurs prédites : (constantes), Nombre Repr., Pages de Publicité

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2522878,407	2	1261439,204	40,223	,000[a]
	Résidu	1160361,093	37	31361,111		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Nombre Repr., Pages de Publicité

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-32,463	93,146		-,349	,729
	Pages de Publicité	22,734	10,165	,319	2,236	,031
	Nombre Repr.	108,366	27,636	,559	3,921	,000

a. Variable dépendante : Ventes

Régression qualité du grossiste (qualitatif, toutes lesmodalités)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,928[a]	,862	,850	118,876

a. Valeurs prédites : (constantes), Bon, Très Mauvais, Mauvais

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	3174504,930	3	1058168,310	74,880	,000[a]
	Résidu	508734,570	36	14131,516		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Bon, Très Mauvais, Mauvais

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	1028,000	44,931		22,880	,000
	Très Mauvais	-690,000	63,542	-,864	-10,859	,000
	Mauvais	-510,818	57,476	-,752	-8,888	,000
	Bon	-40,067	54,414	-,064	-,736	,466

a. Variable dépendante : Ventes

Régression qualité du grossiste (3 modalités)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,927[a]	,860	,852	118,138

a. Valeurs prédites : (constantes), Mauvais, Très Mauvais

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	3166843,091	2	1583421,545	113,453	,000[a]
	Résidu	516396,409	37	13956,660		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Mauvais, Très Mauvais

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	1000,682	25,187		39,730	,000
	Très Mauvais	-662,682	51,266	-,830	-12,926	,000
	Mauvais	-483,500	43,625	-,711	-11,083	,000

a. Variable dépendante : Ventes

Régression avec 3 variables (qualité grossiste qualitative)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,994[a]	,989	,987	34,761

a. Valeurs prédites : (constantes), Mauvais, Pages de Publicité, Très Mauvais, Nombre Repr.

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	3640948,817	4	910237,204	753,317	,000[a]
	Résidu	42290,683	35	1208,305		
	Total	3683239,500	39			

a. Valeurs prédites : (constantes), Mauvais, Pages de Publicité, Très Mauvais, Nombre Repr.

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	568,102	27,004		21,038	,000
	Pages de Publicité	22,880	2,053	,321	11,146	,000
	Nombre Repr.	24,835	6,215	,128	3,996	,000
	Très Mauvais	-489,047	17,856	-,612	-27,388	,000
	Mauvais	-374,904	15,340	-,552	-24,440	,000

a. Variable dépendante : Ventes

8.4.Produits frais (fichier pfrais.xls)

On a mis à votre disposition les données concernant 49 points de ventes (constituant un échantillon représentatif) pour faire une étude sur les ventes de yaourt de différentes marques. Une unité statistique étant constituée d'une marque vendue dans un magasin.
Les données recueillies concernent les variables suivantes :

- Chiffre d'affaires du produit en KF
- Budget publicitaire régional du magasin en KF
- Distribution en valeur (DV)[5] pour la marque dans la zone de chalandise concernée (entre 0 et 1)
- Prix moyen du Kg de produit dans le magasin pour la marque concernée en F
- Marque du produit (codée de 1 à 4)
- Région du magasin (codée de 1 à 5)

Votre objectif est de déterminer un modèle explicatif du Chiffre d'affaires.

Etude des variables quantitatives

Dans un premier temps, on n'utilisera que les trois variables explicatives quantitatives (Publicité, DV, Prix moyen). Après avoir effectué les 4 régressions linéaires de la variable Ventes (Chiffre d'affaires) en fonction d'au moins deux des variables explicatives, répondre aux questions suivantes.

Analyse du modèle à 3 variables

Quelle est la validité statistique et économique du modèle ?

Analyse des modèles à deux variables

Analyser rapidement les modèles à 2 variables explicatives. Quelles remarques pouvez-vous faire ? Quel est le meilleur modèle à 2 variables ? Utiliser ce modèle pour faire une estimation du chiffre d'affaires espéré avec les données suivantes :

- Budget Publicitaire 100KF
- DV de 0,95
- Prix moyen du Kg : 8F

Choix d'un modèle

Quel est pour vous le meilleur modèle ne faisant intervenir que les variables explicatives quantitatives ? ?

Etude des variables qualitatives

Ici ne sont prises en compte que les variables qualitatives Marque et Région. Effectuer les trois régressions, ainsi que le tableau d'analyse de la variance (test de Fisher partiel).

Etude de chacune des variables individuellement

1- Rappeler comment est traitée en régression une variable qualitative à k modalités.
2- La marque a-t-elle une influence significative sur le chiffre d'affaires ? Classer les marques en fonction du chiffre d'affaires moyen.

[5] La DV est égale au rapport des CA des magasins offrant la marque divisée par la somme des CA de tous les magasins de la zone. La DV donne une idée de la représentation, pondérée par l'importance des magasins, de la marque dans la zone de chalandise.

3- La région a-t-elle une influence significative sur le chiffre d'affaires ? Classer les régions en fonction du chiffre d'affaires moyen.

Etude des deux variables qualitatives simultanément

1- Quelle est la validité statistique du modèle obtenue ?
2- Analyser le tableau de l'analyse de la variance, conservez-vous les deux variables explicatives ?
3- Quel modèle à variable(s) explicative(s) qualitative(s) conseillez-vous ?

Etude avec l'ensemble des variables

En conservant les variables qualitatives et quantitatives jugées satisfaisantes aux deux questions précédentes, effectuer une régression comprenant ces trois variables.

4- Que pensez-vous de la validité du modèle obtenu ?
5- Quel est le modèle retenu finalement ?
6- Comment pouvez-vous expliquer la non-validitéé d'une des variables explicatives (statistiquement et économiquement) ?
7- Utiliser ce modèle pour donner le chiffre d'affaires espéré pour un produit et un magasin présentant les caractéristiques suivantes :
   - Budget Publicitaire 100KF
   - DV de 0,95
   - Prix moyen du Kg : 8F
   - Marque 3

Conclusion :

Quel modèle vous semble-t-il le plus adapté pour l'explication et la prévision du chiffre d'affaires ?