HEC

Gilles Mauffrey

METHODES QUANTITATIVES AVEC EXCEL

Programmation linéaire, programmation dynamique, simulation, statistique élémentaire

LA MODELISATION

1 Modèle et typologie des modèles

1.1 La notion de modèle

Un modèle est d'après le dictionnaire Robert :

1. Ce qui sert ou doit servir d'objet d'imitation pour faire ou reproduire quelque chose
2. Personne, fait, objet possédant au plus haut point certaines qualités ou caractéristiques qui en font le représentant d'une catégorie
3. Objet de même forme qu'un objet plus grand mais exécuté en réduction
4. Représentation simplifiée d'un processus, d'un système

La notion de modèle qui nous utiliserons ici est en fait un mix des définitions 2, 3 et 4. Nous nous attacherons à donner une représentation schématisée, mais en contrôlant la simplification, de la réalité et nous serons conduits à utiliser parfois des modèles mathématiques préexistants. Pour nous un modèle sera une représentation simplifiée de la réalité dans au moins l'un des deux buts suivants :

- mieux comprendre la réalité
- aider à la prise de décision en fournissant des solutions acceptables aussi bonnes que possible.

1.2 Les composants d'un modèle

On est conduit à modéliser quand on se trouve confronté à un problème dont il n'existe pas de solutions évidentes (soit heuristiques, soit parce qu'on a déjà été confronté à ce type de problème).

Le problème concerne l'entreprise ou une partie de l'entreprise que nous appellerons système (par exemple une unité de production, les caisses d'un supermarché, etc..) ; ce système est sous contrôle d'un décideur (ou d'un groupe de décideurs) qui peut en modifier le comportement par des actions (ou décisions). Ce système est en relation avec des éléments extérieurs non directement contrôlés par le décideur que nous appellerons environnement. Remarquons que les décisions du décideur peuvent avoir des conséquences sur l'environnement (par exemple un fort budget publicitaire peut accroître à la fois la part de marché et la taille du marché).

Enfin certaines caractéristiques du système et de l'environnement peuvent être considérées comme primordiales pour le décideur et servir à comparer entre elles les décisions, nous parlerons alors de conséquences des actions. Bien évidemment ces conséquences sont fonction des objectifs que s'est fixé (ou qui ont été fixés au) le décideur.

1.2.1 Les variables de décisions

Les variables de décisions servent à décrire les actions envisagées. Elles peuvent prendre leurs valeurs sur ensemble fini (par exemple nombre de caisses à ouvrir) ou considéré comme infini (par exemple budget consacré à un média). Elles peuvent être simultanées (par exemple quantités à produire un mois) ou séquentielle s'étalant dans le temps (par exemple faire une étude de marché, puis décider de la taille de la capacité de production).

1.2.2 L'environnement et le système

Pour décrire l'environnement et le système que nous noterons E/S, nous utilisons deux éléments :

- Les paramètres structurels : ce sont des constantes qui ne vont pas être modifiées par les décisions du décideur, ces paramètres structurels sont dépendants des hypothèses simplificatrices qui ont été prises pour construire le modèle et de l'horizon de modélisation que l'on s'est fixé (prix de vente d'un produit, salaire d'une caissière, etc..). Certains paramètres structurels peuvent être définis par une loi de probabilité (par exemple nombre de clients arrivant à une station service pendant un intervalle de temps donné).
- Les variables d'état du système : vont permettre de faire une « photographie » de l'environnement et du système sous l'effet des décisions, ce sont des fonctions à la fois des paramètres structurels et des décisions envisagées. Par exemple :
    - les capacités de production utilisées dépendent des quantités à produire(décision) et des données technologiques de production(paramètres),
    - le budget publicitaire dépensé, le nombre de contacts publicitaires dépendent des spots publicitaires (décisions) , du coût des spots et des audiences des émissions(paramètres),
    - le nombre de clients dans une file d'attente, le nombre de caisses inoccupées dépendent du nombre de caisses ouvertes (décision) et du rythme d'arrivées à la caisse et du temps de service(paramètres).

  Ces variables d'état sont des variables aléatoires si les paramètres dont elles dépendent sont des lois de probabilité.
- Les relations de fonctionnement du système, qui expriment le respect des contraintes d'évolution du système. Ce peut être des équations ou inéquations (respect d'une demande, d'une capacité de production, d'un budget par exemple) ou des relations temporelles (évolution d'une file d'attente toutes les minutes). Ces relations définissent le modèle de fonctionnement du système.

1.2.3 Les conséquences

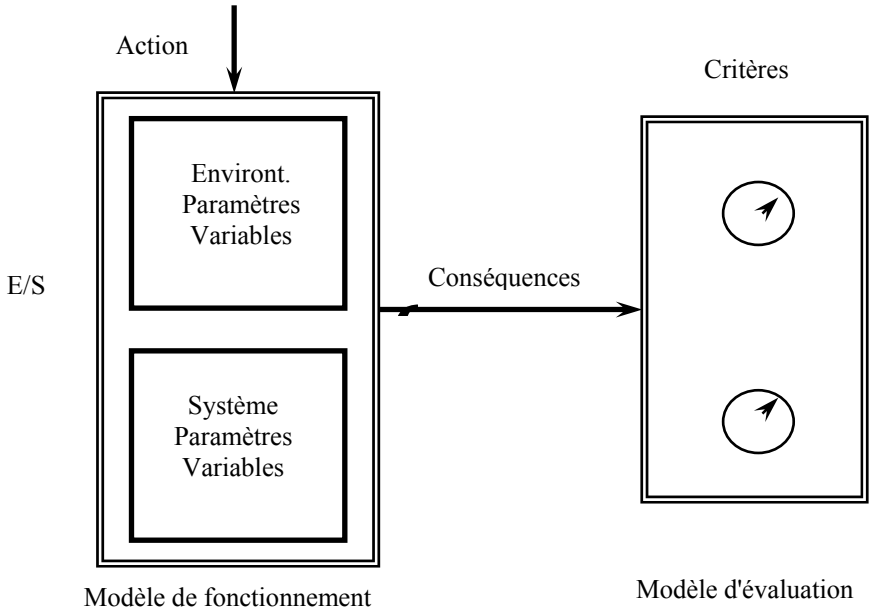
Les conséquences sont des variables d'état privilégiées qui vont permettre de comparer ou de sélectionner les décisions : par exemple le profit réalisé grâce à une production ou le temps moyen d'attente d'un client. Ces conséquences sont évaluées par un modèle d'évaluation.

Le modèle d'évaluation peut consister en une simple optimisation (maximisation ou minimisation) : par exemple marge maximale d'une production, risque minimal d'un portefeuille, minimiser le temps moyen d'attente, dans ce cas la variable d'état privilégiée comme conséquence doit être unique et se nomme fonction économique (ou fonction objectif).

Il peut aussi être constitué de plusieurs compteurs qui déterminent les plages dans lesquelles doivent se trouver les conséquences : par exemple moins de 95% des clients doivent attendre plus de 5 minutes aux caisses et le taux d'occupation des caisses doit au moins être de 80%.

Dans ce cas le modèle d'évaluation permet d'éliminer les décisions qui n'atteignent pas ces objectifs

En conséquence, la structure d'un modèle suivra le schéma suivant :



1.3 Typologie des modèles

Suivant les éléments connus, on peut dégager la typologie suivante :

1.3.1 Modèles descriptifs (E/S) :

Il s'agit de modèles généralement statistiques qui ont pour objet de faire connaître les paramètres structurels du modèle ou les formules définissant les variables d'état du système.

On répond ici aux questions "Quel est mon environnement, comment fonctionne le système ?"

Les méthodes statistiques utilisées vont de l'estimation simple à l'analyse des données ou aux méthodes de prévision.

1.3.2 Modèles de simulation (Calcul des conséquences) (E/S, Action) :

On connaît ici les paramètres structurels et les variables d'état de l'environnement et du système et l'on veut évaluer les conséquences des différentes actions envisagées (donc en nombre fini) sans pour autant chercher à identifier "la meilleure".

Ce choix est laissé au décideur, le modèle peut fournir évidemment plusieurs conséquences (multicritère).

On répond ici à la question "Que se passe-t-il si... ?"

La méthode privilégiée ici est la méthode de simulation, soit avec des langages dédiés, soit sur tableur ou à l'aide de langages "classiques" tels que C, FORTRAN, PASCAL, BASIC.

1.3.3 Modèles d'optimisation (E/S, Action, Critères) :

On connaît ici les paramètres structurels et les variables d'état de l'environnement et du système. On connaît les actions envisagées ainsi que le critère d'évaluation des conséquences. On veut déterminer la meilleure action possible.

Evidemment, le critère de choix est unique (limitation des méthodes mathématiques).

On répond ici à la question "Que faire ?" Les méthodes utilisées sont très variées : elles sont mathématiques ou font appel à la simulation ou à des heuristiques.

Nous nous intéresserons dans ce cours uniquement aux modèles d'optimisation ou de simulation. Dans ce cas la modélisation peut être considérée comme une méthodologie d'aide à la décision stratégique, qui a pour objectif de permettre une allocation efficace des ressources en vue de la réalisation d'objectifs. En voici quelques exemples :

- Déterminer le nombre de guichets à ouvrir pendant une période donnée pour éviter une attente trop longue des clients et une inactivité trop importante des guichetiers
- Déterminer une bonne utilisation d'un budget publicitaire pour atteindre le plus grand nombre de clients potentiels
- Déterminer la composition d'un portefeuille pour atteindre une rentabilité maximale avec risque maximum donné
- Déterminer une production qui conduise à une marge maximum compte tenu des ressources disponibles et des demandes connues

2 La démarche de modélisation

La démarche de modélisation peut s'articuler autour de trois phases :

2.1 Analyse descriptive

1.Fixer les limites géographiques, physiques et aussi temporelles du système étudié et de son environnement. Quels sont les paramètres structurels décrivant ce système ?

2.Enumérer les actions envisagées ou le type d'action envisagée.

3.Déterminer les variables d'état, c'est à dire les éléments qui permettent de "photographier" le système à un moment donné sous l'effet des actions.

4.Choisir la façon dont le fonctionnement du système sera décrit : satisfaction de contraintes structurelles, évolution temporelle.

5.Identifier les conséquences qui serviront à évaluer les actions (variables d'état privilégiées).

6.Sélectionner éventuellement les critères permettant de comparer les actions.

2.2 Mise en équation

1.Nommer la (ou les variables) associée(s) aux actions.

2.Ecrires les relations définissant les variables d'état.

3.Ecrire les relations décrivant le fonctionnement du système, relations entre les variables d'état et les paramètres structurels et les décisions.

4.Identifier les relations définissant les conséquences et exprimer les critères.

2.3 Résolution du modèle

On peut soit utiliser un logiciel spécifique, par exemple un logiciel de programmation linéaire, soit utiliser un progiciel standard du type tableur. Dans ce dernier cas, il faudra veiller à respecter la structuration du modèle, c'est à dire à affecter des zones bien délimitées et séparées aux différents composants du modèle :

- Paramètres structurels
- Variables de décision
- Variables d'état et relations de fonctionnement
- Conséquences évaluées par des critères

Il faut bien noter que les solutions trouvées sont les solutions du modèle et non du problème originel ; il reste au décideur à transcrire ces solutions dans le monde réel en réintégrant éventuellement certains éléments non pris en compte dans le modèle. L'adéquation des solutions trouvées au problème réel dépend bien évidemment de la pertinence du modèle et ceci relève plus d'un art que d'une science.

Le processus de modélisation fait donc appel à trois ressources principales :

- Les données de l'entreprise et l'environnement, recueillies dans le système d'information de l'entreprise (paramètres structurels)
- Les connaissances d'un expert sur le métier et l'environnement (relations de fonctionnement, conséquences)
- Des modèles mathématiques ou des outils de simulation tels qu'un tableur (résolution).

EXERCICE DE MODELISATION

L'entreprise Clairgaz

L'entreprise Clairgaz met en bouteille et distribue des bouteilles de gaz. La mise en bouteille s'effectue dans trois usines notées 1, 2, 3 qui livre 5 dépôts régionaux, notés A,B, C,D, E. Les capacités de production mensuelle (en milliers de bouteilles) de chacune des usines et les demandes mensuelles de chacun des dépôts sont les suivants :

Usine	Production
1	40
2	80
3	120

Dépôt	Demande
A	20
B	10
C	30
D	80
E	100

Les bouteilles doivent être livrées de chaque dépôt à chaque usine, on peut en première approximation considérer que le coût unitaire de transport est proportionnel à la distance, c'est d'ailleurs ainsi que se fait la facturation interne, les coûts de transport étant affectés aux dépôts et donc pris en compte lors de l'évaluation annuelle des directeurs de dépôts. L'annexe 1 vous donnent les valeurs de ces coûts unitaires. On remarquera que le dépôt C et l'usine 2 ont une même localisation.

Actuellement la politique de livraison résulte de négociations entre les directeurs de dépôts et d'usine, cette politique vous est donnée en annexe 2. La direction générale trouve les coûts totaux de transport actuellement trop élevés, et pense qu'il serait possible de les diminuer de façon significative pour les deux années à venir, où il n'est pas envisagé de modifications importante de la demande. Il est fait appel à vous pour étudier ce problème.

Question 1

Analyser le problème de la direction générale :

Quels sont le système, les paramètres structurels, les décisions, les variables d'état, la conséquence ?

Question 2

Ecrire les équations correspondant.

Question 3

Que pensez-vous des réactions possibles des différents intervenant : direction générale, directeurs de dépôt et d'usine; comment y remédier?

Question 4

Pouvez vous proposer une méthode heuristique de résolution?

Annexe 1

Coût de transport unitaire d'usine à dépôt (en €) :

	Dépôts				
Usines	**A**	**B**	**C**	**D**	**E**
1	**7**	**10**	**5**	**4**	**12**
2	**3**	**2**	**0**	**9**	**1**
3	**8**	**13**	**11**	**6**	**14**

Annexe 2

Politique actuelle d'approvisionnement des dépôts

	Dépôts				
Usines	**A**	**B**	**C**	**D**	**E**
1					**40**
2			**30**		**50**
3	**20**	**10**		**80**	**10**

Soit un coût total de 1 440K€

Eléments de Recherche Opérationnelle

3 Un Premier Exemple

Une entreprise fabrique deux produits A et B avec deux matières premières M et P, et une machine T1. Les consommations, les temps de fabrication et les marges réalisées pour chaque produit ; ainsi que les quantités disponibles pour le mois à venir sont donnés dans le tableau suivant :

	Produit A	**Produit B**	**Disponible**
Matière Première M	12	14	1500
Matière Première P	8	4	600
Temps de fabrication	3 H	1 H	210 H
Marge Bénéficiaire	300	250	

3.1 Formalisation du problème

3.1.1 Analyse descriptive :

Le système est constitué de l'unité de production de l'entreprise durant le mois suivant.

Les paramètres structurels sont les données technologiques de production, les disponibilités en matières premières et temps machine et les marges bénéficiaires unitaires.

Les variables d'action sont les quantités respectives de produit A et B à fabriquer le mois suivant

Les variables d'état sont les quantités de matières premières utilisées, le temps machine utilisé et la marge dégagée

Les relations de fonctionnement du système consistent à s'assurer que l'utilisation des ressources reste inférieure à la disponibilité.

La conséquence privilégiée et la marge dégagée par la production décidée, le critère consiste à maximiser cette marge

On a donc affaire à un problème d'optimisation.

3.1.2 Mise en équations du problème

Définition des variables d'action : notons X1 et X2 les quantités respectives de produit A et B à fabriquer durant le mois. On peut considérer que ces quantités sont des nombres réels, la partie fractionnaire correspondant à des produits encours. Ces deux variables sont évidemment positives ou nulles.

Calcul des variables d'état :

- Utilisation de la matière première M : 12*X1 + 14*X2
- Utilisation de la matière première P : 8*X2 + 4*X2
- Utilisation de la machine T : 3*X1 + 1*X2
- Marge bénéficiaire dégagée : 300*X1 + 250*X2

Equations de fonctionnement du système (Contraintes) : (X1>=0 ; X2>=0)

12*X1 +14*X2 <= 1500

8*X1 + 4*X2 <= 600

3*X1 + 1*X2 <= 210

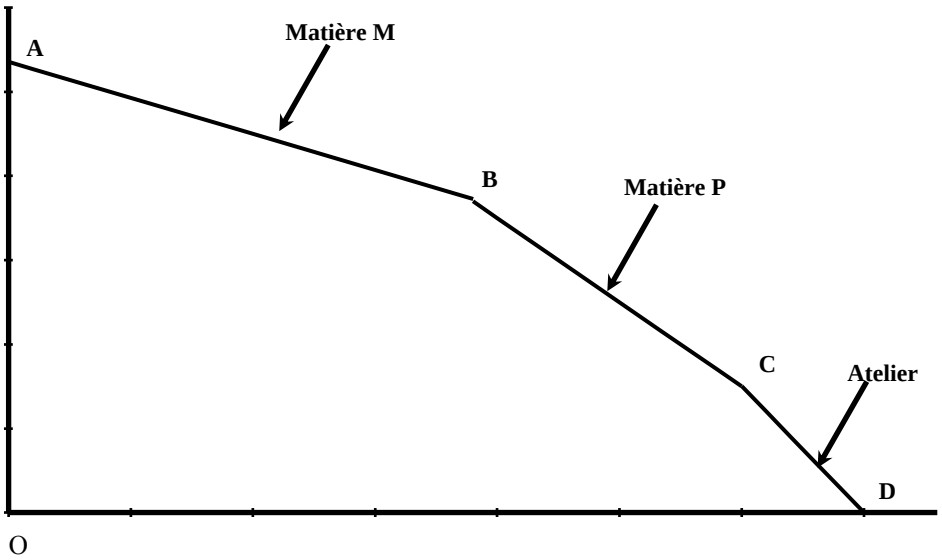
Objectif (fonction économique) et critère :

Maximiser f(X1, X2) = 300*X1 + 250*X2

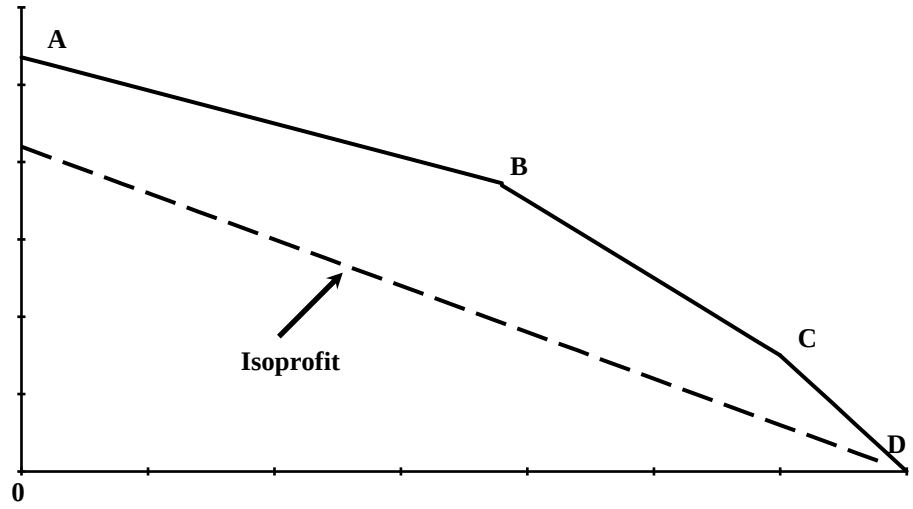
3.1.3 Résolution graphique du problème

Comme il n'intervient ici que 2 variables on peut donner une représentation graphique du problème :

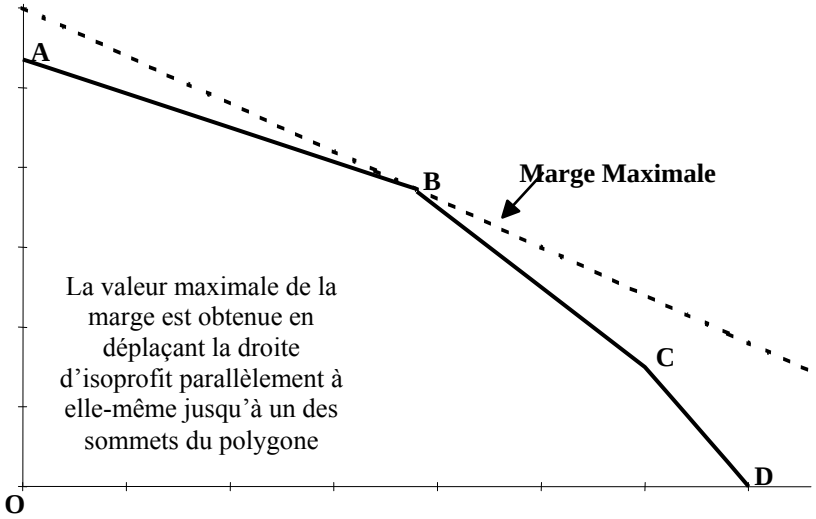
Construction de la surface correspondant aux contraintes : Chaque contrainte partage le plan en deux demi-plans dont un seul correspond à la contraintes. De plus, comme les variables sont positives on se limite qu quadrant supérieur droit. On obtient ainsi l'intérieur d'un polygone convexe appelé Ensemble des solutions réalisables ou admissibles.



Représentation de la fonction économique : Pour une valeur donnée k de la marge l'ensemble des productions conduisant à cette marge se trouvent sur la droite d'équation 300 X1+ 250 X2 = k appelée droite d'isoprofit ; seuls les points de cette droite intérieurs au polygone correspondent à des productions compatibles avec la structure de production actuelle.



Résolution graphique du problème : Toutes les droites d'isoprofit sont parallèles entre elles, il nous faut donc déterminer une droite qui soit parallèle à une direction donnée, qui soit le plus éloignée possible de l'origine tout en coupant l'ensemble des solutions réalisables. Cette droite



intuitivement va passer par l'un des sommets du polygone.

4 Définition d'un programme linéaire

Les caractéristiques d'une situation pouvant conduire à la formalisation sous forme de programme linéaire sont illustrées par l'exemple précédent :

Les actions sont en nombre non fini, (dénombrable, voire même continu), elles ne peuvent prendre que des valeurs positives.

Les paramètres structurels sont connus de façon certaine et déterministe (sans loi de probabilité).

Les variables d'état sont linéaires (ou au moins les relations de fonctionnement sont linéarisables).

Les relations de fonctionnement s'expriment sous la forme d'inégalités.

La conséquence privilégiée est unique et le critère est un critère d'optimisation (maximum ou minimum).

D'un point de vue mathématique, il s'agit de maximiser une fonction linéaire de variables réelles positives ou nulles sous une conjonction de contraintes d'inégalité dont la partie droite (dépendant des variables) est linéaire. C'est cette linéarité qui va permettre de dégager des propriétés mathématiques assez simples de l'ensemble des solutions réalisables et de l'optimum, et de mettre en place un algorithme de résolution du problème.

Remarques :

1.On peut bien évidemment minimiser une fonction linéaire puisque cela revient à maximiser l'opposé de la fonction

2.On peut aussi envisager des contraintes d'égalité puisqu'une contrainte du type

$$f(x, y, z,...)=b$$

est équivalente à la conjonction des deux contraintes :

f(x, y, z,...)<=b

f(x, y, z,...)>=b

5 Propriétés mathématiques d'un programme linéaire

Remarque : ce paragraphe n'est pas nécessaire à la compréhension du reste du document.

Un programme linéaire peut être défini sous la forme générale suivante :

Maximiser une fonction linéaire de **n** variables :

$$c_1X_1+c_2X_2+. +c_nX_n$$

sous **p** contraintes (inéquations inférieures ou égales) dont la partie gauche est une fonction linéaire des **n** variables et la partie droite est constante :

$$a_{i1}X_1+ a_{i2}X_2 +.....+a_{in}X_n <=b_i \qquad i \text{ variant de 1 à p}$$

toutes les variables X_i étant positives ou nulles. Soit donc n+p inéquations.

Les variables X_1, X_2, X_3, …X_n sont appelées variables naturelles.

5.1 Ensembles convexes

L'ensemble des solutions réalisables est un ensemble convexe. C'est à dire si M et P sont deux points de cet ensemble, tout point du segment [MP] est aussi une solution réalisable. Soit pour tout réel t dans [0 ; 1] et tous points M et P dans le convexe C le point Q=tP+(1-t)M (barycentre de M(1-t), P(t)) est dans C.

Point extrémal d'un convexe : Un point E d'un convexe C est dit extrémal s'il n'est pas à l'intérieur d'un segment ; c'est à dire si

la relation E=tP+(1-t)M entraîne t=0 ou t=1 (i.e. E=P ou E=M)

Exemples :

Pour une boule les points extrémaux sont les points de la sphère. Pour un disque, les points du cercle.

Pour un polyèdre (ou polygone en dimension 2) les points extrémaux sont les sommets

Remarque : Dans le cas d'un programme linéaire, l'ensemble des solutions est un polyèdre convexe, appelé simplexe, les points extrémaux sont donc les sommets qui correspondent à la saturation (transformation en équation) de n des n+p inéquations.

Il y a donc au plus C_n^{n+p} points extrémaux.

5.2 Fonction linéaire sur un convexe

Un programme linéaire se présente donc comme un cas particulier de maximisation d'une fonction linéaire sur un convexe. Nous confondrons dans la suite le point M et le vecteur OM. Une fonction linéaire f vérifie la propriété :

Pour tous réels a et b f(aP+bM)=af(P)+bf(M) donc en particulier pour tout point Q du segment [MP] on a min(f(M), f(P))<=f(Q) <= max(f(M), f(P)), on en déduit le

Premier théorème : *Si la fonction présente un maximum sur le convexe, ce maximum est atteint en au moins un point extrémal* (raisonnement par l'absurde).

En conséquence, il nous suffira de chercher le maximum sur les sommets du convexe des solutions réalisables ; toutefois ces sommets peuvent être très nombreux dans la pratique, il nous faut donc trouver une méthode qui permette de sélectionner les sommets à explorer. Le théorème suivant va nous y aider :

Deuxième théorème : *Pour une fonction linéaire définie sur un convexe tout optimum local est global.*

Démonstration : soit A un optimum local (c'est à dire qu'au voisinage de ce point la fonction prend des valeurs inférieures ou égales à f(A)), supposons qu'il existe dans le convexe C un point B tel que f(B)>f(A). Le segment AB est dans le convexe C, donc pour tout t dans l'intervalle ouvert $]0;1[$ le point M = tA+(1-t)B est dans C et on a f(M)=t f(A)+(1-t) f(B)>f(A). Donc en tout point du segment ouvert (AB) la fonction f prend des valeurs supérieures à f(A), ce qui est contraire à l'hypothèse de maximum local.

En conclusion :

> *Nous pouvons donc explorer les sommets de proche en proche (c'est à dire passer d'un sommet à un sommet voisin), et vérifier localement que le maximum est atteint. C'est la démarche de la méthode du simplexe.*

6 Algorithme du simplexe

Dans ce chapitre nous supposerons toujours que le second membre des contraintes (partie constante) est positif ; nous distinguerons donc les contraintes inférieures ou égales des contraintes supérieures ou égales.

6.1 Variables d'écart - Variables de surplus

Considérons une contrainte inférieure ou égale (par exemple ressource utilisée <= ressource disponible) :

$$a1X1+a2X2+....+anXn <= b$$

il est possible de remplacer cette inéquation par une équation en faisant intervenir une variable ***positive ou nulle*** **e** :

$$a1X1+a2X2+....+anXn +\mathbf{e} = b$$

cette variable qui peut représenter l'écart entre le disponible et l'utilisé est appelée variable d'écart (slack variable).

Pour une contrainte supérieure ou égale (par exemple satisfaction d'une demande minimale) :

$$a1X1+a2X2+....+anXn >= b$$

on se ramènera à une équation en soustrayant une variable positive ou nulle s :

$$a1X1+a2X2+....+anXn -s= b$$

cette variable qui peut représenter le surplus de production par rapport au minimum imposé est appelée variable de surplus (surplus variable).

Sur l'exemple de présentation les contraintes s'écrivent alors :

12*X1	+14*X2	+ e1	= 1500	(Matière première M)
8*X1	+ 4*X2	+e2	= 600	(Matière première P)
3*X1	+ 1*X2	+e3	= 210	(Atelier)

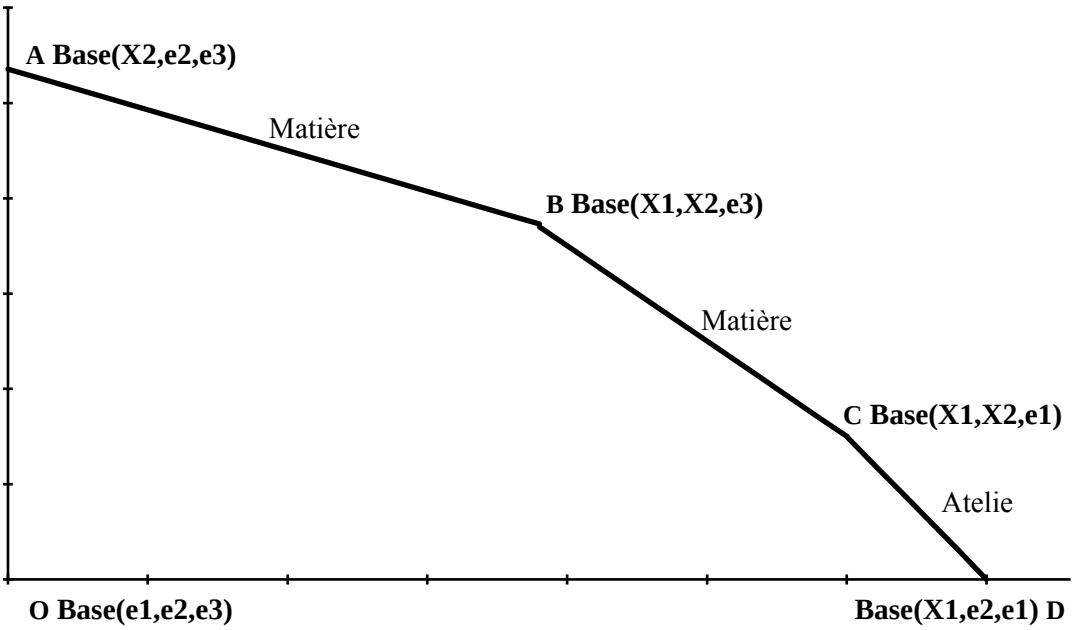
6.2 Variables de base - Variables hors base

Le problème qui faisait intervenir n variables naturelles (définies par la formalisation) et p contraintes inférieures ou supérieures, devient maintenant un problème à n+p variables et p contraintes d'égalité.

Chaque point extrémal du simplexe des solutions réalisables correspond à la saturation de ***p*** contraintes (explicites ou implicites : positivité des variables naturelles). On pourra donc associer à un sommet du simplexe une partition des ***n+p*** variables : ***n*** variables nulles et ***p*** variables solution du système de ***p équations à p inconnues***.

Les ***p*** variables qui servent à résoudre le système s'appellent les variables de base, les ***n*** autres variables sont les variables hors base.

Sur l'exemple de présentation nous avons pour chaque sommet les variables de base :



Remarquons que passer d'un sommet à un sommet voisin revient simplement à échanger une variable de base avec une variable hors base, puisque entre deux sommets voisins seul un hyperplan saturé est modifié.

6.3 Principe de l'algorithme

A partir des remarques précédentes, la démarche va consister à se déplacer d'un sommet en un sommet voisin, et à vérifier si on peut améliorer localement la fonction économique (en effet nous savons que tout optimum local est global). Précisons cette démarche :

1. Trouver un sommet initial ; si on ne peut en trouver il n'y a pas de solution.

2. Exprimer grâce aux contraintes la fonction économique en fonction des variables hors base (il suffit de résoudre le système en fonction des variables de base, les variables hors base étant considérées comme des paramètres)
3. Voir si l'introduction d'une variable hors base améliore la fonction économique (existe-t-il un coefficient strictement positif pour les variables hors base de la fonction économique ?), si ce n'est pas le cas on a atteint l'optimum, sinon choisir la meilleure candidate localement (la variable hors base dont le coefficient positif est le plus grand).
4. Déterminer la valeur maximale prise par cette nouvelle variable, si cette valeur est infinie la solution est aussi infinie ; sinon un sommet améliorant la fonction économique est trouvé, retourner alors en 2.

Cet algorithme converge (avec une modification pour éviter le cyclage quand la valeur maximale en 4 est 0).

Remarque : la première étape peut être délicate s'il existe des contraintes >=, en revanche elle est très simple dans le cas où les seules contraintes sont des contraintes <= en effet dans ce cas l'origine est toujours dans le simplexe (ce qui correspond à la base constituée de toutes les variables d'écart). C'est sur un exemple de ce type que nous illustrerons l'algorithme.

7 Exemple de l'algorithme du simplexe

Nous allons prendre comme exemple l'exemple d'introduction, qui est un problème de maximisation sous contraintes inférieures ou égales. Dans ce cas l'étape 1 de l'algorithme est très simple puisque l'origine appartient toujours à l'ensemble des solutions réalisables.

7.1 Etape 0 : Etat Initial

Ecriture du problème

MAX 300X1 + 250X2

12*X1 +14*X2 + e1 = 1500

8*X1 + 4*X2 +e2 = 600

3*X1 + 1*X2 +e3 = 210

Nous sommes en O, les variables de base sont (e1,e2,e3), les variables hors base (X1,X2). La valeur de la fonction économique est égale à son terme constant 0, et son expression ne fait intervenir que les variables hors base (X1 et X2) ; d'autre part la solution en ce point est donnée par le système de contrainte : e1=1500, e2=600, e3=210.

Nous ne sommes pas à l'optimum car les coefficients des variables hors base sont positifs : on peut améliorer la fonction économique qui vaut actuellement 0. Il nous faut donc passer à un sommet voisin, c'est à dire échanger une variable hors base et une variable de base.

Choix de la variable entrant dans la base : c'est la variable X1 car son coefficient est le plus grand, c'est donc celle qui localement améliore le plus la fonction économique.

Choix de la variable sortant de la base : les trois variables e1, e2, e3 sont candidates, il nous faut voir quelle est la valeur maximale possible de X1 sans qu'aucune autre variable ne soit négative (ne pas oublier que X2 reste nulle). Examinons les 3 équations :

Si X1 remplace e1, X1 prend la valeur 1500/12=125

Si X1 remplace e2, X1 prend la valeur 600/8 = 75

Si X1 remplace e3, X1 prend la valeur 210/3=70

La valeur maximale que peut prendre X1 est donc le minimum de ces 3 valeurs, c'est à dire 70 (sinon on devrait donner des valeurs négatives à e2 ou e1). X1 remplace donc e3. La troisième contrainte qui caractérise l'échange s'appelle la *contrainte pivot*, elle va nous servir à réécrire le système :

1. Remplacer X1 par 70-1/3 e3 - 1/3 X2 dans la fonction économique et les deux premières contraintes
2. Réécrire la troisième contrainte de façon à mettre en évidence les variables de base (sous matrice identité) comme dans l'état initial.

7.2 Etape 1

En utilisant la relation définie précédemment nous obtenons la formulation équivalente suivante :

MAX -100e3 + 150X2 +21000

-4e3 + 10*X2 + e1 = 660

-(8/3)e3 +(4/3)X2 +e2 = 40

(1/3)e3 +(1/3)X2 +X1 =70

Nous sommes au point D, les variables de base sont (e1, e2, X1) les variables hors base (e3, X2). La valeur de la fonction économique est le terme constant 21000 (car e3 et X2 sont hors base donc valent 0), elle est obtenue avec les valeurs lues dans le système de contraintes : e1=660, e2= 40, X1=70. Toutefois cette valeur n'est pas optimale car il reste un coefficient strictement positif, donc la fonction économique peut s'améliorer localement (les variables ne peuvent qu'être positives).

Choix de la variable entrant dans la base : c'est la variable X2 car son coefficient est le seul positif.

Choix de la variable sortant de la base : les trois variables e1, e2, X1 sont candidates, il nous faut voir quelle est la valeur maximale possible de X2 sans qu'aucune autre variable ne soit négative (ne pas oublier que e3 reste nulle). Examinons les 3équations :

Si X2 remplace e1, X2 prend la valeur 660/10=66

Si X2 remplace e2, X2 prend la valeur 40/(4/3) = 30

Si X2 remplace e3, X2 prend la valeur 70/(1/3)=210

La valeur maximale que peut prendre X2 est donc le minimum de ces 3 valeurs, c'est à dire 30 (sinon on devrait donner des valeurs négatives à e1 ou X1). X2 remplace donc e2. La deuxième contrainte est la contrainte pivot, elle va nous servir à réécrire le système :

1. Remplacer X2 par 30-3/4e2+2e3 dans la fonction économique et la première et la dernière contrainte

2. Réécrire la deuxième contrainte de façon à mettre en évidence les variables de base (sous matrice identité) comme dans le système initial.

7.3 Deuxième étape

En utilisant la relation définie précédemment nous obtenons la formulation équivalente suivante :

MAX 200e3 - 112,5e2 +25500

16e3	- 7,5 e2	+ e1		= 360
-2e3	+(3/4)e2	+X2		= 30
e3	-(1/4)e2		+X1	= 60

Nous sommes au point C, les variables de base sont (e1, X2, X1) les variables hors base (e3, e2). La valeur de la fonction économique est 25500 (car e3 et X2 sont hors base donc valent 0), la solution en ce point correspond à e1=360, X2=30, X1=60.Cette solution n'est toujours pas optimale car il reste un coefficient strictement positif, donc la fonction économique peut s'améliorer localement (les variables ne peuvent qu'être positives).

Choix de la variable entrant dans la base : c'est la variable e3 car son coefficient est le seul positif.

Choix de la variable sortant de la base : les trois variables e1, X2, X1 sont candidates, il nous faut voir quelle est la valeur maximale possible de X2 sans qu'aucune autre variable ne soit négative (ne pas oublier que e2 reste nulle). Examinons les 3 équations :

Si e3 remplace e1, e3 prend la valeur 360/16=22,5

Si e3 remplace X2, e3 prend la valeur 30/(3/4) = 40

Si e3 remplace X1, e3 prend la valeur 60

La valeur maximale que peut prendre e3 est donc le minimum de ces 3 valeurs, c'est à dire 22,5 (sinon on devrait donner des valeurs négatives à e1 ou X1). e3 remplace donc e1. La première contrainte est la contrainte pivot. Nous allons donc :

1. remplacer e3 par 22,5+(15/32)e2-(1/16)e3 dans la fonction économique et les deux dernières contraintes,
2. réécrire la première contrainte de façon à mettre en évidence les variables de base (sous matrice identité) comme dans le système initial.

7.4 Etape 3

En utilisant la relation définie précédemment nous obtenons la formulation équivalente suivante :

MAX -12,5e1 - 18,75e2 +30000

(1/16)e1	-(15/32)e2	+ e3		=22,5
(1/8)e1	- (3/16)e2	+X2		=75
-(1/16)e1	+ (7/32)e2		+X1	=37,5

Nous sommes au point B, les variables de base sont (e3, X2, X1) les variables hors base (e1, e2). La valeur de la fonction économique est 30000, la fonction économique ne peut pas s'améliorer localement car tous les coefficients sont <=0. On a donc atteint le maximum (local donc global).

La solution optimale est donc la suivante :

> Produire 37,5 unités de A(X1), 75 unités de B(X2)et laisser 22h30 inutilisées dans l'atelier (e3) : variables de base.
>
> Utiliser toutes les matières premières (e1=e2=0) : variables hors base.
>
> La marge dégagée est alors de 30000F

Comment interpréter les coefficients de e1 et e2 dans la fonction économique ? La seule façon d'accroître la fonction économique serait de pouvoir leur donner une valeur négative. Par exemple si on donnait à e1 la valeur –1, la fonction économique augmenterait de 12,5. En regardant la première formulation du problème, c'est à dire la définition des variables d'écart, on constate que cela revient à disposer d'une unité supplémentaire de la matière première A. La valeur absolue des coefficients des deux variables d'écart représente le gain que l'on pourrait réaliser en disposant d'une unité de ressource supplémentaire, économiquement cela revient à quantifier le coût d'opportunité associé à une contrainte saturée (à une ressource "rare" pour l'entreprise), contrainte qui empêche d'accroître la production. Bien évidemment cela ne peut pas être valable pour une quantité quelconque, car à partir d'une certaine quantité la ressource n'est plus "rare", et une autre contrainte sera saturée. L'analyse de listing que nous allons voir au paragraphe suivant permet de répondre à ce type de question.

8 Utilisation du solveur Excel pour la programmation linéaire

Pour utiliser Excel en programmation linéaire, il faut formaliser le problème sur une feuille, puis utiliser une macro complémentaire appelée solveur pour résoudre le problème, les solutions sont données sur des feuilles "Rapport" créées par Excel. Nous illustrerons cette utilisation sur l'exemple des paragraphes précédents.

8.1 Formalisation du problème

L'écriture du problème sous Excel se présente sous la forme suivante :

	A	B	C	D	E
1		Produit A	Produit B		
2	Quantité	0	0		
3	Marge Bénéficiaire	300	250		
4	***F.Eco***	***=SOMMEPROD(B2:C2;B3:C3)***			
5					
6		Produit A	Produit B	***Utilisé***	Disponible
7	Matière Première M	12	14	***=SOMMEPROD($B$2:$C$2;B7:C7)***	1500
8	Matière Première P	8	4	***=SOMMEPROD($B$2:$C$2;B8:C8)***	600
9	Temps de fabrication	3	1	***=SOMMEPROD($B$2:$C$2;B9:C9)***	210

Les cellules B2 et C2 contiennent les quantités de produit fabriquées, ici initialisées à 0, variables à déterminer. La cellule B4 contient la formule de la fonction économique, quel 'on peut écrire soit sous la forme B2*B3+C2*C3 ou SOMMEPROD(B2:C23 ; B3:C3).

Les cellules B7:C9 donnent les données technologiques, les cellules E7:E9 donnent les quantités disponibles.

Les cellules D7:D9 contiennent les formules calculant les quantités utilisées : attention aux $ pour la recopie vers le bas.

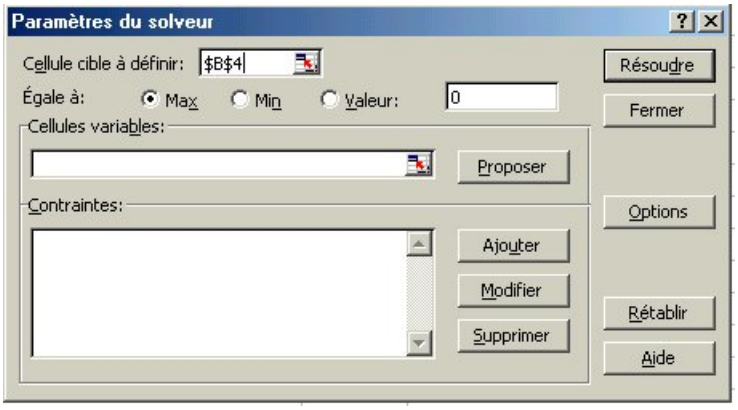
Il est important que le côté droit de chaque contrainte soit une constante, et non pas une fonction des variables de décision, sinon dans certains cas Excel pourrait ne pas accepter que le problème soit linéaire.

La feuille de calcul ainsi écrite ne permet pas seule de résoudre le problème d'optimisation, il nous serait seulement possible de tester certaines solutions (simuler des décisions). Nous vérifierions que ces décisions sont acceptables sans jamais savoir si nous avons atteint l'optimum.

Enfin il n'apparaît pas sur la feuille de calcul le sens des contraintes (<= ou >=), ni le sens de l'optimisation (Maximum ou Minimum). Il est donc nécessaire, pour finaliser la formulation du problème et le résoudre de faire appel à un "add-in" (un programme complémentaire accessible à partir d'Excel, en "français" une macro complémentaire.

8.2 Utilisation du solveur

Après avoir sélectionné la cellule contenant la valeur de la fonction économique, dans le menu ***Outils*** nous choisissons le sous menu *Solveur…*, il apparaît alors la boîte de dialogue suivante :

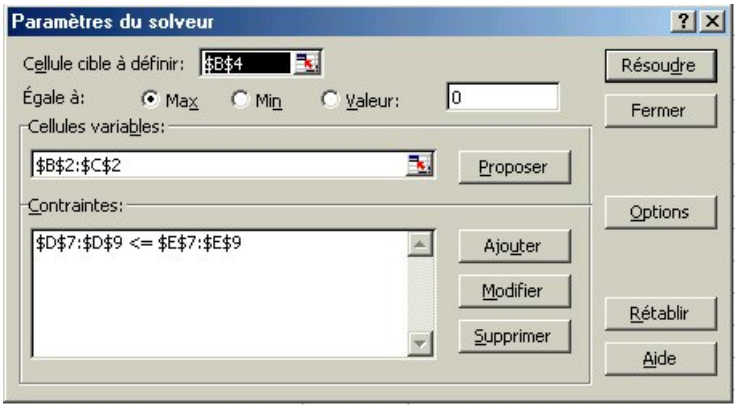


Dans la zone Cellule cible à définir, il est indiqué l'adresse de la cellule contenant la formule de la fonction économique, ici $B$4 ; si vous avez ouvert le solveur à partir d'une autre cellule sélectionnée, c'est l'adresse de cette cellule qui apparaîtra ici, il faudra alors modifier en conséquence cette zone en cliquant sur la cellule de la fonction économique. Ensuite il faut sélectionner le type d'optimisation voulu (Maximisation ou minimisation).

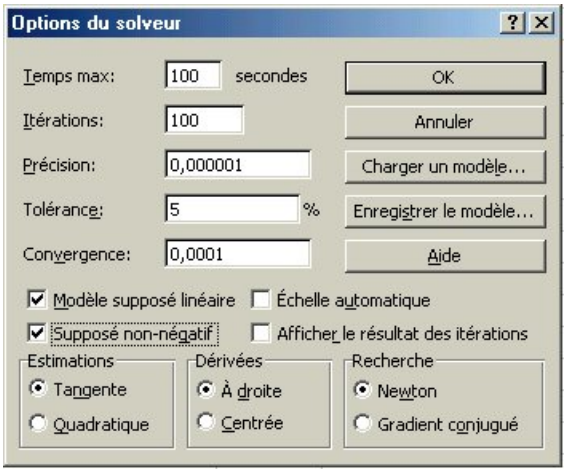
Dans la zone cellules variables, il faut indiquer la zone contenant les variables du problème, ici $B$2:$C$2. Il faut ensuite entrer les contraintes du problème ; pour cela cliquer sur le bouton "Ajouter" de la zone contrainte, une autre boîte de dialogue apparaît :

Dans la zone cellule, il faut indiquer l'adresse de la cellule contenant la formule du côté gauche des contraintes, puis choisir dans la liste déroulante le sens de la contrainte (<=, >= ou =) et enfin, dans la zone contrainte indiquer l'adresse de la cellule contenait la valeur du côté droit de la contrainte. Entre chaque contrainte cliquer le bouton ajouter, vous pouvez entrer les contraintes de même sens sous forme vectorielle, à condition bien sûr que les cellules des mêmes côtés soient adjacentes (par exemple $D$7:$D$9).

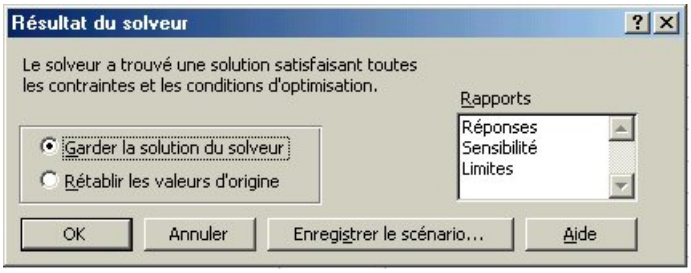
Après la dernière contrainte, valider avec le bouton OK. On revient alors à la première boite de dialogue qui se présente ainsi :



Il nous reste à préciser que le problème est un problème de programmation linéaire, n'utilisant que des variables positives ou nulles. Pour cela cliquer sur le bouton "Options" et dans la zone de dialogue suivante, cocher la case "Modèle supposé linéaire" et "Supposé non négatif" :



Revenu au dialogue initial par le bouton "OK", il faut demander la résolution du problème en cliquant sur le bouton "Résoudre". L'algorithme de résolution s'exécute, en fin de traitement un dernier dialogue apparaît :



Il faut alors sélectionner les rapports de Réponse et Sensibilité, en cliquant sur ces libellés ; mais il est inutile de demander celui des Limites qui en programmation linéaire n'apporte rien.

9 Analyse d'un listing de programmation linéaire

En pratique, on ne résout jamais "à la main" un programme linéaire, on utilise pour ce faire soit des logiciels spécialisés soit un tableur comme Excel.

9.1 Structure d'un listing de programmation linéaire

Les listings de programmation linéaire comportent tous, sous des présentations variables, trois parties :

- la valeur de la fonction économique : valeur optimale de la fonction économique pour le problème posé.
- les résultats concernant les variables naturelles : valeurs des variables naturelles et sensibilité de l'optimum en fonction du coefficient de chacune des variables naturelles dans la fonction économique
- les résultats concernant les contraintes : valeurs des variables d'écart ou de surplus à l'optimum et sensibilité de l'optimum en fonction de chacun des côtés droit des contraintes.

Les phases de l'analyse :

On peut distinguer trois phases dans l'analyse du listing d'un programme linéaire :

1. Déterminer la solution optimale dans la structure actuelle
2. Faire une analyse marginale des contraintes, en vue de déterminer les décisions pouvant améliorer la solution actuelle.
3. Faire une analyse des coefficients de la fonction économique pour déterminer la stabilité de la solution optimale

Les étapes 2 et 3 s'appellent souvent analyse marginale.

Exemple d'un listing Excel

Dans le cas d'un listing produit par Excel, la valeur de la fonction économique et les valeurs des variables naturelles sont données dans la feuille "Rapport des réponses", tandis que les éléments concernant l'analyse marginale se trouvent dans la feuille "Rapport de sensibilité".

9.2 Détermination de la solution optimale

Il s'agit ici de donner les valeurs des variables naturelles, de la fonction économique et l'état des contraintes à l'optimum.

Généralement les listings de Programmation Linéaire donnent pour chaque variable naturelle son statut (Variable de base ou hors base) en plus de sa valeur à l'optimum, certains programmes (SAS par exemple) donnent les mêmes précisions pour les variables d'écart ou de surplus.

Exemple d'un listing Excel

Nous donnons ici le rapport des réponses correspondant au problème initial :

Cellule cible (Max)

Cellule	Nom	Valeur initiale	Valeur finale
$B$4	F.Eco Produit A	0	30000

Cellules variables

Cellule	Nom	Valeur initiale	Valeur finale
$B$2	Quantité Produit A	0	37,5
$C$2	Quantité Produit B	0	75

Contraintes

Cellule	Nom	Valeur	Formule	État	Marge
$D$7	Matière Première M Utilisé	1500	$D$7<=$E$7	Lié	0
$D$8	Matière Première P Utilisé	600	$D$8<=$E$8	Lié	0
$D$9	Temps de Fabrication Utilisé	187,5	$D$9<=$E$9	Non lié	22,5

La cellule cible correspond à la fonction économique : sa valeur à l'optimum est 30000.

Pour atteindre cette valeur les productions sont données dans la partie de la feuille intitulée "Cellules variables" : à l'optimum il faut produire 37,5 unités de A et 75 unités de B.

L'utilisation des ressources est donnée dans la partie "Contraintes". La valeur représente la valeur prise à l'optimum par la partie gauche des contraintes (ici la quantité de ressources utilisée), l'état indique si cette contrainte est saturée (liée) ou non, et la marge représente la valeur de la variable d'écart (ou de surplus) à l'optimum.

Ici toutes les matières premières sont utilisées et il reste 22H30 de disponible dans l'atelier.

9.3 Analyse marginale des contraintes

Il s'agit ici de déterminer des actions modifiant l'environnement (certains paramètres structurels) permettant d'améliorer la fonction économique, ou d'évaluer l'impact de certaines évolutions de cet environnement.

9.3.1 Interprétation du listing

Seuls sont pris en compte les côtés droits des contraintes. L'analyse, théoriquement ne se fait qu'une contrainte à la fois.

Pour les contraintes non saturées, cette analyse est peu intéressante : par exemple, tant que l'entreprise dispose des ressources suffisantes pour la production optimale, la valeur de la

fonction économique ne change pas et il est bien évidemment inutile de se procurer un surplus de ressources.

En revanche une contrainte saturée indique une gène pour l'amélioration de la fonction économique, toute augmentation ou diminution du côté droit de la contrainte va conduire à une modification de l'allocation des ressources et/ou de la production et par conséquent à une modification de la fonction économique. On peut donc associer à chaque contrainte un coût (ou profit) marginal correspondant au resserrement (ou relâchement) de la contrainte, bien évidemment ce coût ne sera valable que sur un intervalle de valeurs pour le côté droit de la contrainte : par exemple si l'on augmente trop une ressource, on se trouvera limité par d'autres ressources ou par le marché, tout apport supplémentaire n'aura alors plus aucun intérêt économique.

Les listings de programmation linéaire donnent à la fois le coût marginal, appelé shadow cost (traduit sous Excel par Ombre Coût) ou shadow price, qui indique le gain associé au relâchement de la contrainte, ainsi que l'intervalle sur lequel cette valeur est valable. Ce shadow price est exprimé en unités de la fonction économique.

Exemple d'un listing Excel

Voici une partie du rapport de sensibilité correspondant à l'exemple:

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$D$7	Matière Première M Utilisé	1500	12,5	1500	600	360
$D$8	Matière Première P Utilisé	600	18,75	600	48	171,4285714
$D$9	Temps de Fabrication Utilisé	187,5	0	210	1E+30	22,5

"Finale Valeur" correspond à la partie gauche des contraintes (ressources utilisées), "Contrainte à droite" correspond à la partie droite des contraintes (ressources disponibles). La valeur de la variable d'écart associée à la contrainte s'obtient comme différence de ces deux valeurs. "Admissible Augmentation" et "Admissible Réduction" définissent l'intervalle sur lequel le shadow price ("Ombre Coût") est valable. Interprétons ces valeurs.

La contrainte de matière première M est saturée, toute augmentation marginale de ressource en cette matière permettra de générer un nouveau profit, la variation marginale du profit par unité de ressource supplémentaire est donnée par le shadow price : 12,5. Cependant, si la quantité de ressource est trop importante son influence économique va diminuer, c'est ce que nous indique l'augmentation admissible : on n'augmentera le profit de 12,5 par unité de ressource supplémente que tant que la quantité supplémentaire restera inférieure ou égale à 600, c'est à dire tant qu'on disposera de moins de 2100 unités de matière première M ; au-delà bien évidemment le profit marginal sera inférieur.

De la même façon, toute diminution d'une unité de ressource dans cette matière première va diminuer le profit de 12,5, et ceci tant que la diminution ne dépassera pas 360 unités ; c'est à dire tant que la quantité de ressource restera supérieure à 1140 unités. Au-delà la perte marginale sera supérieure.

Remarque : *ce shadow price correspond à une restructuration optimale de la production en fonction de la nouvelle quantité, les autres ressources étant inchangées. Le listing ne donne pas cette nouvelle structure de production.*

La contrainte sur la matière première P s'analyse de la même façon, puisque cette contrainte est aussi saturée.

Interprétons maintenant la dernière contrainte : la contrainte d'atelier. Cette contrainte n'est pas saturée, donc augmenter les heures disponibles n'apportera aucun profit supplémentaire, c'est pourquoi le shadow cost est nul et l'augmentation admissible infinie (notée 1E+30 par Excel). De la même manière si on diminue les ressources disponibles, tant que l'on conserve la quantité nécessaire à la production, ceci ne diminuera en rien le profit : la diminution admissible est donc égale à la variable d'écart.

9.3.2 Cas limite : problème dégénéré

Il peut arriver que parmi les variables de base, l'une d'entre elles soit nulle, dans ce cas le shadow cost pour la variable d'écart correspondant (ou le reduced cost pour une variable naturelle) sera lui aussi nul, on dit alors que le problème est dégénéré. Ceci correspond à la valeur limite d'un intervalle de variation d'un coté droit d'une contrainte. Etudions ce cas sur la première contrainte.
Tout d'abord considérons la limite inférieure, le second membre de la contrainte passe à 1500-360=1140, la valeur de la fonction économique est de 30000-12,5*360=25500, le rapport de sensibilité est le suivant :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	60	0	300	200	85,71
$C$2	Quantité Produit B	30	0	250	100	100

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$D$7	Matière Première M Utilisé	***1140***	12,5	***1140***	**960**	**0**
$D$8	Matière Première P Utilisé	***600***	18,75	***600***	**0**	**274,29**
$D$9	Temps de Fabrication Utilisé	***210***	0	***210***	**1E+30**	**0**

Les trois contraintes sont saturées, mais comme il doit y avoir trois variables de base, et que les deux variables naturelles sont dans la base, l'une des variables d'écart nulles est dans la base. C'est celle dont le shadow cost est nul, c'est à dire la troisième contrainte. On constate de plus que les trois contraintes se coupent au même point, ceci apparaît dans le listing par le fait que l'une des deux augmentations limites (admissible augmentation ou admissible réduction) est nulle : dès que l'on bouge un peu l'une des deux premières contraintes (vers le bas pour la première, vers le haut pour la seconde), elle devient inactive (non saturée) et la troisième devient active, alors sont shadow cost va devenir strictement positif.

Ce cas se généralise dans un espace de dimension n, quand n+1 contraintes concourent en un sommet du simplexe : on aura alors une des n+1 contraintes dont le shadow cost sera égal à 0 et pour les n+1 contraintes l'une des limites égale à 0. Cependant la lecture du listing n'est pas très simple et sur beaucoup de logiciel la contrainte correspondant à la variable de base est indiquée comme dégénérée; malheureusement Excel ne l'indique pas.

Il se peut aussi que la variable de base qui est nulle soit une variable naturelle, auquel cas le listing sera un peu différent. C'est le cas pour la valeur maximale de la première contrainte, si le second membre de la contrainte passe de 1500 à 1500+600=2100, on obtient le rapport suivant :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	***0***	***0***	300	200	85,71
$C$2	Quantité Produit B	150	0	250	100	100

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$D$7	Matière Première M Utilisé	***2100***	***12,5***	2100	**0**	**960**
$D$8	Matière Première P Utilisé	***600***	***18,75***	600	**128**	**0**
$D$9	Temps de Fabrication Utilisé	150	0	210	1E+30	60

La production de produit A est nulle, les deux premières contraintes sont toujours saturées, mais l'augmentation admissible de la première et la diminution admissible de la seconde sont nulles, dès que l'on modifiera un peu l'une de ces contraintes dans ce sens le reduced cost du produit A deviendra strictement positif.

Ici c'est la contrainte $A \geq 0$ qui est associée à la troisième contrainte saturée, on constate que l'on est dans le cas de dégénérescence et non pas de solution multiple(voir ci-dessous) d'une part d'après le nombre de variable de base nulle(ici 1) (ou non nulles 2 au lieu de 3) et d'autre part parce que pour la variable naturelle nulle, dont le shadow cost est nul, aucune des deux limites n'est égale à 0, ce qui signifie que ce n'est pas la rentabilité du produit qui est en cause, mais la disponibilité des ressources.

Remarque importante : suivant les arrondis, l'algorithme utilisé par Excel, qui n'est pas exactement le simplexe, conduira à l'une ou l'autre des solutions optimales extrêmes.

9.4 Analyse marginale d'un coefficient de la fonction économique

Il s'agit ici de voir la stabilité de l'optimum (valeur des variables de base) en fonction des variations d'un coefficient de la fonction économique (changement de prix d'un produit par exemple). Cette analyse ne se fait qu'un seul coefficient à la fois. Nous raisonnerons dans le cas d'une maximisation.

9.4.1 Interprétation du listing

Si une variable naturelle n'est pas dans la base, sa valeur est nulle, lui donner une valeur positive ne pourrait que faire baisser la fonction économique, cette baisse est indiquée sur les listings en tant que shadow cost ou reduced cost (traduit en Réduit Coût). Si ce shadow cost est nul ceci signifie qu'il existe au moins un autre sommet solution optimale, donc une infinité de solutions optimales (le segment joignant ces deux sommets).

Si une variable est dans la base, ceci signifie que dans la structure actuelle son coefficient est suffisamment élevé. Puisqu'elle est dans la base son shadow cost est évidemment nul ; mais il existe un intervalle (pour le coefficient de cette variable) pour lequel la solution optimale reste

la même. Si le coefficient augmente trop la solution va changer (augmentant la valeur de cette variable), si le coefficient diminue, cette variable sera moins intéressante économiquement dans la structure actuelle et la solution changera aussi (diminution de la valeur de cette variable).

Cas d'un listing Excel

Pour les besoins de l'analyse nous avons ici ajouté un produit C, qui dans la structure de production actuelle n'est pas rentable, sa contribution est de 291 ("Objectif Coefficient"). Voici la partie du rapport de la sensibilité correspondant à l'analyse marginale des coefficients:

Microsoft Excel 9.0 Rapport de la sensibilité
Feuille: [exempPL.xls]Exemple2

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	37,5	0	300	200	85,71
$C$2	Quantité Produit B	75	0	250	100	64,92
$D$2	Quantité Produit C	0	-52,75	291	52,75	1E+30

Pour les produits A et B, le coût réduit est égal à 0, en effet ces produits sont effectivement fabriqués et s'imposer d'en fabriquer n'est pas une contrainte. Les valeurs "Admissible Augmentation" et "Admissible diminution" nous indique pour chaque produit sur quel intervalle le coefficient de la fonction économique doit rester pour que la production ne soit pas modifier.

Attention : On ne fait varier qu'un coefficient, les autres gardent la même valeur.

Pour le produit A, tant que sa contribution est comprise entre 214,29 (300-85,71) et 500 (300 + 200) (les autres contributions restant respectivement de 250 pour B et 290 pour C), la production optimale restera toujours de 37,5 A et 75 B ; mais la fonction économique sera modifiée en conséquence.

Pour le produit C, l'interprétation est théoriquement la même, tant que sa contribution est inférieure à 343,75 (291+52,75), il est inintéressant à produire. Une autre façon d'aboutir à ce résultat est obtenue avec le coût réduit : si on était obligé de produire ce produit C, on perdrait 52,75 pour chaque unité produite, au moins pour les premières unités, sa contribution minimum est donc égale à sa contribution actuelle (291) + la perte lue ici (52,75) soit 343,75. On ne connaît pas, par le listing, sur quelle quantité s'applique cette perte unitaire ; mais économiquement, il est clair que si le nombre de produits fabriqués augmente, la mauvaise utilisation des ressources conduira à une perte plus importante. De la même manière on ne sait pas quelle quantité on serait conduit à produire si la contribution du produit dépassait 343,75.

9.4.2 Cas limite : problème à solution multiple

Introduisons cette fois ci la marge limite pour le produit C, c'est à dire 343,75; nous obtenons alors le rapport de sensibilité suivant, pour la partie variable, la partie contrainte est restée la même :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	37,50	0	300	200	0
$C$2	Quantité Produit B	75,00	0	250	100	0,00
$D$2	Quantité Produit C	0,00	0	343,75	0,00	1E+30

Le produit C n'est toujours pas produit semble-t-il, mais comme le reduced cost est nul le fait de s'imposer d'en produire ne coûterait rien, il existe donc des solutions optimales contenant des quantités non nulles du produit C. Ici ce n'est pas une variable de base qui vaut 0, les variables de base sont restées les mêmes, mais c'est uniquement le reduced cost (ou pour une variable d'écart le shadow cost) associé à une variable hors base qui est nul.

On voit aussi sur ce listing que dès que l'une des deux premières marges diminue, la solution va changer, de même si la marge du produit C augmente la solution changera ; dans tous les cas la nouvelle solution optimale consistera à commencer la production du produit C. On pourra donc obtenir l'autre solution en modifiant légèrement l'un de ces prix par exemple en mettant 343,751 pour le produit C on obtient alors le listing suivant :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	0	-0,002	300	0,002	1E+30
$C$2	Quantité Produit B	10,000	0	250	0,001	53,571
$D$2	Quantité Produit C	80,000	0	343,751	93,749	0,001

L'autre solution correspond donc à la production de 10 B et 80 C, on vérifie d'ailleurs que :

$$300*37,5+75*250 = 10*250+343,75*80 = 30000$$

En fait toute combinaison convexe entre les deux productions est solution optimale, c'est à dire une production de la forme :

$$\begin{bmatrix} prodA \\ prodB \\ prodC \end{bmatrix} = t\begin{bmatrix} 37,5 \\ 75,0 \\ 0 \end{bmatrix} + (1-t)\begin{bmatrix} 0 \\ 10 \\ 80 \end{bmatrix} = \begin{bmatrix} 37,5*t \\ 65*t+10 \\ 80-80*t \end{bmatrix} \quad \text{pout tout } t \in [0;1]$$

par exemple pour t=0,2 on a une production de 7,5A, 23 B et 64 C qui conduit à une marge totale de 7,5*300 + 23*250 + 64*343,75 = 30000, qui est bien la valeur optimale.

Il se peut aussi que l'on ait une solution multiple qui ne joue que sur les quantités des mêmes produits, et non pas sur l'introduction d'un produit à la limite de la rentabilité, dans ce cas ce sera une variable d'écart (hors base) nulle qui aura un shadow price nul, il faudra éviter de confondre ce cas avec le cas de dégénérescence évoqué plus haut. Pour illustrer ce phénomène, en revenant au cas initial avec deux variables, mettons au minimum admissible le coefficient de B, c'est à dire à 250-100=150. Nous obtenons alors le rapport de sensibilité suivant :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	37,5	0	300	0,00	171,43
$C$2	Quantité Produit B	75	0	150	200,00	0,00

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$E$7	Matière Première M Utilisé	1500	0,00	1500	600	360
$E$8	Matière Première P Utilisé	600	37,50	600	48	171,43
$E$9	Temps de Fabrication Utilisé	187,5	0,00	210	1E+30	22,50

La première contrainte est saturée (valeur finale=contrainte à droite), mais son shadow cost est nul, donc si on dispose de moins de ressources la valeur de la fonction économique ne changera pas; il existe donc une autre production (correspondant à un autre sommet du simplexe) consommant moins de matière première M (et donc plus de temps pour maintenir le nombre de variables de base et hors base) et conduisant à la même valeur de la fonction économique. De façon précise, on peut savoir que cette autre solution consommera exactement 360 unités de moins de matière première M. Pour obtenir cette nouvelle solution, il suffit comme précédemment d'augmenter la marge du produit A ou de diminuer celle du produit B, puisque aucune variation des coefficients dans ce sens n'est acceptée. On obtient alors le résultat suivant, en mettant 149,99 comme valeur de marge pour le produit B :

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Produit A	60	0	300	149,97	0,02
$C$2	Quantité Produit B	30	0	149,99	0,01	49,99

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$E$7	Matière Première M Utilisé	1140	0	1500	1E+30	360
$E$8	Matière Première P Utilisé	600	37,49	600	48	40
$E$9	Temps de Fabrication Utilisé	210	0,02	210	15	22,5

La production de 60A et 30B conduit à la même marge totale :

$$60*300+30*150=37.5*300+75*150=22500$$

Ici encore toute combinaison convexe des deux solutions est aussi optimale.

Les logiciels spécialisés en Programmation linéaires signalent les cas de solutions multiples, malheureusement Excel ne le fait pas.

Ici aussi Excel peut arriver sur l'une quelconque des solutions extrêmes.

9.5 Solution dégénérée ou solutions multiples?

Solution dégénérée et solutions multiples se caractérisent par l'apparition pour une même variable (naturelle ou d'écart ou de surplus) de valeurs nulles à la fois pour la valeur de la variable et pour son shadow cost (ou reduced cost). Comment distinguer alors ces deux cas, si ce n'est pas fait par le logiciel.

La première différence vient de la nature de la variable :

- Un problème admet une solution dégénérée si une variable de base est nulle, tandis qu'un problème admet une solution multiple si le shadow cost associé à une variable hors base est nul.

Si une seule variable présente la particularité d'être nulle et d'avoir son shadow cost (ou reduced cost) nul aussi, il suffit alors de déterminer si cette variable est de base ou hors base. On sait que dans un problème contenant ***n*** variable naturelles et ***p*** contraintes, il y a ***p*** variables de base et donc ***n*** variables hors base ; il suffit alors de compter les variables de base non nulles pour détecter la nature du problème.

Cette détection est plus délicate si plusieurs variables présentent cette particularité, surtout si le problème est à la fois dégénéré et à solutions multiples, dans ce cas on peut faire les remarques suivantes :

- Pour un cas de dégénérescence on a soit une variable naturelle nulle avec un reduced cost nul mais deux valeurs de variations admissibles strictement positives, soit une variable d'écart (ou de surplus) nulle ainsi que son shadow cost, mais dans ce cas l'une des variations admissibles nulle.
- Pour un cas de solution multiple on a soit une variable naturelle nulle avec un reduced cost nul et une des deux valeurs de variation admissible nulle, soit une variable d'écart (ou de surplus) nulle ainsi que son shadow cost, mais dans ce cas les deux valeurs des variations admissibles sont positives strictement.

EXERCICES DE PROGRAMMATION LINEAIRE

10 Coopérative

Une coopérative agricole disposant de 1 000ha. veut définir son plan annuel de production de céréales. Le tableau suivant montre les besoins en irrigation et engrais par type de culture.

	Eau (m3/ha./an)	Engrais (kgs/ha./an)
Blé	1 000	200
Orge	2 000	100
Seigle	250	50

Les profits annuels par ha pour le blé, l'orge et le seigle sont respectivement de 200€, 100€ et 40€. On dispose de 160 tonnes d'engrais et de 1,6 millions de m3 d'eau par an.

Questions

1) Formuler le problème définissant le nombre d'hectares de chaque céréale à cultiver de façon à maximiser le profit.
2) En comparant les ressources utilisées, simplifier le problème autant que possible. En déduire la solution optimale
3) Analyser le listing ci-dessous

Cellule cible (Max)

Cellule	Nom	Valeur initiale	Valeur finale
$B$4	Profit	0	160000

Cellules variables

Cellule	Nom	Valeur initiale	Valeur finale
$B$2	Blé	0	800
$C$2	Orge	0	0
$D$2	Seigle	0	0

Contraintes

Cellule	Nom	Valeur	Formule	État	Marge
$E$8	Surface	800	$E$8<=$F$8	Non lié	200
$E$9	Engrais	160000	$E$9<=$F$9	Lié	0
$E$10	Eau	800000	$E$10<=$F$10	Non lié	800000

Rapport de sensibilité

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Blé	800	0	200	1E+30	0
$C$2	Orge	0	0	100	0	1E+30
$D$2	Seigle	0	-10	40	10	1E+30

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$E$8	Surface	800	0	1000	1E+30	200
$E$9	Engrais	160000	1	160000	40000	160000
$E$10	Eau	800000	0	1600000	1E+30	800000

11 Compagnie Minière

Une compagnie minière possède deux puits différents P1 et P2, pour l'extraction d'uranium. Les puits sont en deux lieux distincts et ne possèdent pas la même capacité de production. Le minerai d'uranium est d'abord concassé, puis analysé et enfin rangé dans l'une des trois qualités U_1, U_2 ou U_3, suivant sa teneur minerai riche, moyen ou pauvre.

La demande du marché pour les trois qualités est supérieure à ce que l'on peut extraire.

La compagnie s'est engagée à fournir à une usine de séparation 12 tonnes de minerai U_1, 8 tonnes de minerai U_2 et 24 tonnes de minerai U_3 par semaine.

L'exploitation de P1 coûte à la compagnie 20 000 € par jour et celle de P2 revient à 16 000 € par jour.

En un jour d'exploitation, le premier puits produit 6 tonnes de U_1, 2 tonnes de U_2 et 4 tonnes de U3 ; les chiffres pour le second puits sont respectivement de 2 tonnes, 2 tonnes et 12 tonnes.

Combien de jours par semaine faut-il exploiter chaque mine pour que les engagements soient tenus le plus économiquement possible? (ci dessous le rapport de sensibilité)

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Nbre jours P1	1	0	20	28	4
$C$2	Nbre jours P2	3	0	16	4	9,333333333

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$D$9	U1 Production	12	1	12	8	4
$D$10	U2 Production	8	7	8	4	2
$D$11	U3 Production	40	0	24	16	1E+30

12 Compagnie du Bois

La Compagnie du Bois veut utiliser au mieux les ressources en bois d'une de ses propriétés forestières.

Dans cette région, il y a une scierie et une fabrique de contreplaqué ; le bois coupé peut ainsi être transformé en bois de charpente ou en contreplaqué.

Pour produire 100 m3 de bois de charpente, il faut 1.000 mètres de planches de sapin et 3.000 mètres de planches de hêtre (ces planches ayant une largeur et une épaisseur fixées). Pour

produire 1.000 mètres de planches de contreplaqué, il faut 2.000 mètres de planches de sapin et 4.000 mètres de planches de hêtre..

La Compagnie du Bois peut couper par période 32.000 m. de planches de sapin et 72.000 m. de planches de hêtre. Les contraintes de vente exigent qu'au moins 400 m^3 de bois de charpente et 12.000 mètres de planches de contreplaqué soient produits pendant la période.

Le profit est de 400 € pour 100 m3 de bois de charpente et de 600 € pour 1.000 m de planches de contreplaqué.

B sera le nombre de centaines de m3 de bois de charpente produits, C correspondant aux milliers de mètres de planches de contreplaqué.

Questions

1) Formuler le problème entant que modèle de programmation linéaire.
2) Résoudre le problème graphiquement.
3) Analyser le listing ci-dessous.

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Quantité Charpente	8	0	400	50	100
$C$2	Quantité Contreplaqué	12	0	600	200	66,66666667

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$E$9	Sapin	32000	0,1	32000	2000	0
$E$10	Hêtre	72000	0,1	72000	0	4000
$E$11	Charpente	8	0	4	4	1E+30
$E$12	Contreplaqué	12	0	12	0	1E+30

13 Le Laboratoire

Un laboratoire fabrique des récepteurs haute performance. Il emploie quatre assembleurs et deux ingénieurs 40 heures par semaine le salaire est de 20 € l'heure pour un assembleur et 30 € l'heure pour un ingénieur. Chacun des six est prêt à faire jusqu'à 10 heures supplémentaires à 50 % par semaine.

Les coûts fixes pour l'entretien du laboratoire s'élèvent à 5.000 € par semaine. Les coûts variables pour l'entretien et les madères premières sont de 5 €/heure pour un assembleur et 10 €/heure pour un ingénieur, le matériel utilisé étant alors plus coûteux.

Le laboratoire vend des récepteurs finis, à 175€ pièce. Le marché peut absorber toute la production. Le laboratoire vend aussi à une compagnie spécialisée des récepteurs non finis, à 130 € pièce le contrat est pour 100 récepteurs minimum, mais la compagnie est prête à en acheter jusqu'à 150.

Pour construire un récepteur non fini, il faut une heure d'assembleur et 30 minutes d'ingénieur. Pour construire directement un récepteur fini, il faut une heure et demie d'assembleur et autant d'ingénieur.

Comment le responsable du laboratoire devrait-il définir sa production et le programme de ses employés pour maximiser son profit?

14 Le Campeur

La société Le Campeur vend des chaises de jardin, des bancs et des tables. Ces objets sont réalisés à l'aide de tubulures métalliques qui doivent être mises en forme (tordues selon la forme désirée, à l'aide d'une machine) puis soudées (par un robot). Durant la période prévue, on dispose d'une capacité de 1.000 minutes pour la mise en forme et de 1.200 pour la soudure.

Une chaise requiert 1,2 minutes de mise en forme et 0,8 de soudure. Pour un banc, il n'y a pas de soudure et il faut 1,7 minutes de mise en forme. La table, quant à elle, nécessite 1,2 minutes de mise en forme et de 2,3 de soudure.

Pour le moment, le fournisseur de tube est en grève, et l'on peut seulement compter sur le stock, qui s'élève actuellement à 2 000 kilos de tubes, achetés 0,4 € le kg. Il en faut 2 pour une chaise, 3 pour un banc et 4,5 pour une table.

La contribution est de 3€ pour une chaise, 3€ pour un banc et 5€ pour une table.

Questions :

1) Formuler mathématiquement le problème à résoudre.
2) A l'aide du listing ci-joint, indiquer la production optimale et la contribution.
3) Un distributeur local propose de livrer du tube supplémentaire à 1€ le kilo (pour une quantité pouvant aller jusqu'à 500 kilos). Est-ce une bonne affaire ?
4) On s'aperçoit qu'un commercial a pris une commande ferme pour 10 bancs. Quelles seront les conséquences si l'on décide d'honorer cette commande ?
5) Le département R&D a conçu une nouvelle façon de réaliser le banc, avec 1,1 minutes de mise en forme, 2 de soudure et 2 kilos de tube. A partir de quelle contribution unitaire un tel produit serait-il intéressant ?
6) Un client est prêt à passer -pour un produit spécifique qui lui est destiné- une commande qui exigerait de la mise en forme, et qu'il payerait 1,5 € la minute. Il est prêt à commander ainsi jusqu'à 8 heures de mise en forme. Que faut-il faire ?
7) Qu'arriverait-il si la contribution pour les chaises diminuait à 2,5 € ?
8) Les bancs sont actuellement vendus 45 € pièce. Quelle augmentation doit on imposer pour qu'ils soient intéressants à produire ?

Listing Excel

Microsoft Excel 9.0 Rapport des réponses
Feuille: [CAMP.XLS]Feuil2

Cellule cible (Max)

Cellule	Nom	Valeur initiale	Valeur finale
$B$4	Marge	0	2766,666667

Cellules variables

Cellule	Nom	Valeur initiale	Valeur finale
$B$2	Chaises	0	700
$C$2	Bancs	0	0
$D$2	Tables	0	133,3333333

Contraintes

Cellule	Nom	Valeur	Formule	État	Marge
$E$7	Mise en Forme	1000	$E$7<=$F$7	Lié	0
$E$8	Soudure	866,6666667	$E$8<=$F$8	Non lié	333,3333333
$E$9	Tubes	2000	$E$9<=$F$9	Lié	0

Microsoft Excel 9.0 Rapport de la sensibilité
Feuille: [CAMP.XLS]Feuil2

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	Chaises	700	0	3	2	0,777777778
$C$2	Bancs	0	-1,383333333	3	1,383333333	1E+30
$D$2	Tables	133,3333333	0	5	1,75	2

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$E$7	Mise en Forme	1000	1,166666667	1000	200	466,6666667
$E$8	Soudure	866,6666667	0	1200	1E+30	333,3333333
$E$9	Tubes	2000	0,8	2000	555,5555556	333,3333333

15 Composition de portefeuille

Un fond de pension veut placer 1 000 000 € dans des actions, des obligations et des bons du trésor. On supposera que le risque du portefeuille est le risque moyen de ses composants, par exemple si l'on place 1000 € dans une action dont le risque est évalué à 10% et 3000 € dans une obligation dont le risque est évalué à 5% le risque moyen est :

(1000*10%+3000*5%)/4000=6,25%

Les caractéristiques des actifs envisagés sont les suivantes :

Actif	Rentabilité moyenne	Risque
Action A	18%	15%
Action B	15%	13%
Obligation A	10%	5%
Obligation B	8%	4%
Bons du trésor	5%	0%

De plus on a le fond de pension veut respecter les contraintes suivantes :

La valeur totale investie en obligations et bons du trésor ne doit pas être inférieure 500 000 €

Le risque du portefeuille doit être inférieur à 10%

La valeur investie en action A et obligation A doit être inférieure d'au moins 100 000 € à celle investie en action B et obligation B

Questions :

1) Formaliser le problème de composition du portefeuille
2) Quelle est la composition optimale du portefeuille ?
3) Quel est son risque ?
4) Quelle est la rentabilité minimum que devraient avoir les obligations A pour qu'il y en ait dans le portefeuille ?
5) La valeur minimum investie en obligations ou bons du trésor passe à 550 000 € quel sera l'impact de cette nouvelle contrainte ?
6) Un des gestionnaires du fond a déjà placé 100000 € en bons du trésor, quel est l'impact de cette action ?
7) On peut se procurer 100000 € supplémentaires à 14% ? Quelle sera l'effet de l'acceptation de ce prêt sur la fonction économique ?

Listing Excel :

Rapport des réponses

Cellule cible (Max)

Cellule	Nom	Valeur initiale	Valeur finale
$B$5	Rendement	0	128500

Cellules variables

Cellule	Nom	Valeur initiale	Valeur finale
$B$3	Montant AA	0	450000
$C$3	Montant AB	0	50000
$D$3	Montant OA	0	0
$E$3	Montant OB	0	500000
$F$3	Montant BT	0	0

Contraintes

Cellule	Nom	Valeur	Formule	État	Marge
$G$10	Dif A B	100000	$G$10>=$H$10	Lié	0
$G$11	Investi	1000000	$G$11<=$H$11	Lié	0
$G$12	Risque	-6000	$G$12<=$H$12	Non lié	6000
$G$13	Obli+trésor	500000	$G$13>=$H$13	Lié	0

Rapport de sensibilité

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$3	Montant AA	450000	0	0,18	1E+30	0,01
$C$3	Montant AB	50000	0	0,15	0,01	0,07
$D$3	Montant OA	0	-0,01	0,1	0,01	1E+30
$E$3	Montant OB	500000	0	0,08	0,07	0,01
$F$3	Montant BT	0	-0,045	0,05	0,045	1E+30

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$G$10	Dif A B	100000	-0,015	100000	900000	100000
$G$11	Investi	1000000	0,165	1000000	150000	100000
$G$12	Risque	-6000	0	0	1E+30	6000
$G$13	Obli+trésor	500000	-0,07	500000	50000	66666,66667

16 Fixation de prix

L'entreprise Toutenkit importe trois nouveaux composants C1, C2, C3 aux prix respectifs unitaires de 3, 5 et 6 $ (transport inclus).

Ces composants peuvent être inclus dans de nombreux produits finis, mais d'après le service Marketing, les produits leaders contenant ces composants et pouvant facilement être assemblés par des amateurs sont les produits PF1, PF2, PF3 et PF4.

D'autre d'après l'expérience des vendeurs de Toutenkit, pour que le montage soit plus intéressant que l'achat du produit tout monté, il faut que le prix d'achat (pour le client) des composants soit inférieur d'au moins 20% au prix du modèle monté.

Pour les 4 produits finis, on a obtenu les renseignements suivants :

Produit	Nombre de C1	Nombre de C2	Nombre de C3	Autres Composants	Prix de vente
PF1	2	1		80 $	125 $
PF2	4	2	2	50 $	125 $
PF3	4		6	90 $	175 $
PF4	1	3	3	70 $	150 $

Où la colonne "Nombre de C1", C2 ou C3 indique le nombre de composants C1, C2 ou C3 dans le produit fini donné, et la colonne "Autres composants" donne le prix d'achat des autres composants nécessaires à la fabrication du modèle. Enfin Prix de vente représente le prix de vente minimum observé sur le marché pour le produit fini donné.

Les ventes hebdomadaires espérées par le service commercial sont de 2000 unités pour C1, 1000 unités pour C2 et 3000 unités pour C3, ces ventes devraient rester stables sur le trimestre.

Enfin le prix de vente d'un produit doit légalement être supérieur à son coût (loi antidumping).

1) Formaliser le problème de fixation de prix de l'entreprise Toutenkit, sachant qu'elle veut maximiser la marge globale dégagée par les trois nouveaux composants.
2) Utiliser Excel pour résoudre le problème. Quels sont les prix que doit fixer l'entreprise pour les trois composants, quelle marge totale l'entreprise réalisera-t-elle ? Quels sont les produits finis dont l'entreprise Toutenkit doit particulièrement surveiller l'évolution ? La loi antidumping est-elle contraignante pour l'entreprise ?
3) Lors du relevé des prix minimums, il y a eu une erreur pour le prix de PF2, la valeur est 120 $ et non pas 125 $, cela a-t-il une influence sur la solution trouvée précédemment ?
4) Quel serait l'effet sur la marge d'une réduction de 10 $ du prix minimum de PF4 ?
5) Si la loi antidumping était abolie, quelle serait la nouvelle marge pour l'entreprise, et les nouveaux prix pratiqués pour les composants ?

Listing Excel

Microsoft Excel 8.0a Rapport des réponses

Cellule cible (Max)

Cellule	Nom	Valeur initiale	Valeur finale
$B$5	fe C1	0	34500

Cellules variables

Cellule	Nom	Valeur initiale	Valeur finale
$B$2	C1	0	3,50
$C$2	C2	0	9,50
$D$2	C3	0	6

Contraintes

Cellule	Nom	Valeur	Formule	État	Marge
$F$7	PF1	96,5	$F$7<=$G$7	Non lié	3,50
$F$8	PF2	95	$F$8<=$G$8	Non lié	5,00
$F$9	PF3	140	$F$9<=$G$9	Lié	0
$F$10	PF4	120	$F$10<=$G$10	Lié	0
$F$11	achatC1	3,50	$F$11>=$G$11	Non lié	1,50
$F$12	achatC2	9,50	$F$12>=$G$12	Non lié	5,50
$F$13	achatC3	6	$F$13>=$G$13	Lié	0

Microsoft Excel 8.0a Rapport de la sensibilité

Cellules variables

Cellule	Nom	Finale Valeur	Réduit Coût	Objectif Coefficient	Admissible Augmentation	Admissible Réduction
$B$2	C1	3,50	0	2000	1E+30	333,33
$C$2	C2	9,50	0	1000	5000,00	1000
$D$2	C3	6	0	3000	500,00	1E+30

Contraintes

Cellule	Nom	Finale Valeur	Ombre Coût	Contrainte à droite	Admissible Augmentation	Admissible Réduction
$F$7	PF1	96,5	0	100	1E+30	3,50
$F$8	PF2	95	0	100	1E+30	5
$F$9	PF3	140	416,67	140	6	6,00
$F$10	PF4	120	333,33	120	7,50	16,5
$F$11	achatC1	3,50	0	2	1,50	1E+30
$F$12	achatC2	9,50	0	4	5,50	1E+30
$F$13	achatC3	6	-500,00	6	1,00	1,00

17 La tannerie Landaise

La tannerie Landaise est une unité de production indépendante située dans la Région Landaise qui traite des peaux de mouton. Elle revend ensuite ses peaux à d'autres entreprises dans toute l'Europe. Actuellement, 3 types de produits finis sont vendus sur le Marché:

- des Cuirs Souples
- du Box
- du Daim

Les peaux passent par 3 ateliers :

- l'atelier de séchage
- l'atelier de tannage
- l'atelier de teinture

Les temps de production sont indiqués dans le tableau suivant :

	Cuir Souple	Box	Daim
Séchage	1h	2h	1h
Teinture	2h	1h	3h
Tannage	1h	1h	4h

L'atelier de Séchage dispose de 50 personnes travaillant 40 heures par semaine ; ce personnel est mensualisé. Le salaire horaire moyen est de 12 €/H. Les coûts variables de production (matières premières, entretien, etc..) sont de 40 €/H.

L'atelier de Teinture dispose de l'équivalent de 37,5 personnes travaillant 40 heures par semaine ; ce personnel est mensualisé. Le salaire horaire moyen est de 12 €/H. Les coûts variables de production (matières premières, entretien, etc..) sont de 90 €/H. D'autre part, la Tannerie Landaise peut sous-traiter à une petite entreprise artisanale l'équivalent de 800H de travail (au maximum) ; dans ce cas, elle paie 25 € par heure sous-traitée.

L'atelier de Tannage n'utilise que des intérimaires qui sont en moyenne payés 28 €/H et l'entreprise peut disposer de 3000 Heures au maximum par semaine ; les coûts variables de production (hors main d'œuvre) sont d'environ 32 €/H.

Enfin la position de l'entreprise sur le marché la conduit à fabriquer moins de peaux en Daim que le total des peaux en Box ou en Cuir souple.

Les prix de ventes unitaires des peaux sont respectivement de 400 € pour le Cuir Souple, 390 € pour le Box, 810 € pour le Daim.

Les coûts fixes hebdomadaires sont de 50 000 € environ.

Questions :

1) Formaliser le problème.
2) Quels sont la production optimale, le chiffre d'affaires correspondant et le profit de l'entreprise.

Les questions suivantes sont indépendantes les unes des autres :

3) Les ouvriers de l'atelier de séchage sont prêts à faire 500 H supplémentaires payées 50% de plus. Que doit faire l'entreprise et quel en sera l'impact sur la fonction économique ?
4) A quel prix devrait-on vendre la peau en Cuir souple pour qu'elle devienne rentable dans la structure de production actuelle ? L'entreprise d'intérim qui fournit les ouvriers de l'atelier de tannage vous propose 1200 H supplémentaires pour un prix global de 40 000 €. Evaluer l'impact de l'acceptation de cette proposition.
5) Un nouveau type de peau utilise 2H de séchage, 2H de teinture et 1H de tannage. A quel prix l'entreprise doit-elle le vendre pour qu'il soit économiquement compétitif avec les produits actuels ?
6) Le prix du Box doit baisser de 10%. Quelle conséquence cette baisse aura-t-elle sur la production et sur le profit de l'entreprise ?

18 L'entreprise ShareGift

L'entreprise ShareGift a reçu une commande d'une association qui veut distribuer à ses membres des portefeuilles, porte-clés ou porte-cartes en tissu enduit à son logo. Le tissu enduit a été fourni par l'association et on dispose de 78 m² (soit 7800 dm²) de tissu.

L'association est prête à acheter 3000 pièces au maximum (toutes catégories confondues). Elle exige aussi la production de 200 parures formées d'un portefeuille et d'un porte-cartes.

Pour fabriquer 1 portefeuille il faut 4 dm² de tissu, 3 minutes de découpe et 2 minutes de couture.

Pour 1 porte-clés, il faut 2 dm² de tissu, 2 minutes de découpe et 1 minute de couture.

Pour 1 porte-cartes, il faut 2 dm² de tissu, 1 minute de découpe et 3 minutes de couture.

Etant donnés les délais de livraison demandés par l'association, on ne pourra disposer que de 100H de découpe et 90 H de couture.

Les coûts variables de découpe sont de 240 F par heure, ceux de couture de 300 F par heure. Ces coût ne prennent pas en compte la main d'œuvre qui est mensualisée.

Les prix de ventes pour chacun des produits sont :

	Prix de Vente
Portefeuille	112 F
Porte-clés	63 F
Porte-cartes	49 F

Questions :

1) Formaliser le problème en prenant comme variables le nombre de portefeuilles, de porte-clés et de porte-cartes fabriqués pour maximiser la marge

Les questions suivantes sont indépendantes les unes des autres

2) L'association demande une réduction de 5 F sur le prix du portefeuille. Quel sera l'impact de cette réduction sur la production et la marge de l'entreprise ?

3) Une panne entraîne une diminution de 10 H des heures disponibles pour la couture. Quel en sera l'impact sur la marge ?
4) 5 m² de tissu ont été endommagés pendant le transport. Quel sera l'impact sur la marge ?
5) Les ouvriers de l'atelier Découpe sont disposés à faire des heures supplémentaires, quel prix maximum êtes vous prêt à les payer et combien d'heures leur demanderez-vous ?
6) Quelles seraient les conséquences si l'association exigeait 300 parures au lieu de 200 ?
7) Le prix du porte-cartes vous semble-t-il bien fixé, par rapport à la structure de production actuelle ? Quel serait d'après vous le prix minimum de vente de cet objet ?
8) Un ouvrier propose une nouvelle façon de fabriquer les porte-clés, qui demande 1,5 dm² de tissu, 1 minute de découpe et 3 minutes de couture. Quel serait le prix minimum de vente pour que ce produit soit intéressant à produire dans la structure actuelle ?

19 Média planning

Une entreprise de jeux pour console veut lancer une campagne publicitaire sur un nouveau jeu, sa cible est constituée des jeunes de 10 à 15 ans, éventuellement de la tranche d'age 15-25 ans. Elle envisage les médias suivants :

		Nombre de contacts (en milliers/spot)		
Média	Prix du spot	10-15 ans	15-25 ans	35-55 ans
TV1	40 000 €	500	180	200
TV2	50 000 €	600	200	200
Radio1	15 000 €	100	50	10
Radio2	12 000 €	70	50	15

Le budget prévu pour le mois à venir est de 2 M€, l'entreprise veut limiter le nombre de spots télévisés diffusés sur la période à 25 au maximum.

Elle souhaite que le nombre de contacts 10-15 ans soit au moins 3 fois supérieur à ceux des contacts 35-55 ans.

Il serait souhaitable aussi que le nombre de contacts 15-25 ans soit au moins de 8000000

Enfin pour des raisons commerciales la différence entre les deux budgets radios ne doit pas excéder 200 000€

20 La Société Electroméga

La société Electroméga fait de l'import de matériel électronique. Elle met les produits (A, B et C) aux normes de sécurité du marché intérieur dans un atelier d'électronique et peint les différents produits. De plus elle a créé un nouveau produit (le produit D)qui est fabriqué à partir des produits (finis et modifiés) A et B(une unité de chaque produit A et B est incorporée dans chacune unité du produit D). Elle peut recevoir par mois jusqu'à 500 produits A, 1200 produits B et 200 produits C.

Ces produits sont respectivement achetés au prix de 400 €, 350 € et 500 € l'unité.

Les consommations dans les différents ateliers sont les suivantes :

	Produit A	Produit B	Produit C	Produit D
Atelier Electronique	1H	1H	2H	1H
Peinture	2H	1H	2H	2H

Les coûts variables de production hors main d'œuvre sont respectivement de 250€ par heure pour l'atelier Electronique et de 200 € par heure pour l'atelier de peinture.

L'atelier d'électronique peut disposer de 2800 H pendant le mois. Les techniciens sont mensualisés et payés en moyenne 150 €/H.

L'atelier de peinture peut disposer de 3000 H par mois et peut éventuellement employer des intérimaires pour l'équivalent de 500 H au maximum. Le coût salarial moyen et de 80€ pour les ouvriers qui sont mensualisés, pour les intérimaires le coût est de 150€ par heure.

Les prix de vente des produits sont respectivement de 1500 € pour le produit A, 1500 € pour le produit B, 2000 € pour le produit C et 4000 € pour le produit D. Les coûts fixes mensuels sont de 500 000 €.

Questions :

1) Formaliser le problème en prenant comme variables d'action les quantités vendues des différents produits et le nombre d'heures d'intérim utilisées.
2) Quelle est la solution optimale en terme de production et en terme de chiffre d'affaires et de profit.

Les questions suivantes sont indépendantes les unes des autres.

3) Les techniciens de l'atelier d'électronique proposent de faire 50H supplémentaires payées 50% plus chères. Quel serait l'impact de l'acceptation sur la fonction économique ?
4) Un autre importateur vous propose un lot de 600 produits B à 400 000 €. Que décidez-vous? Quel serait l'impact sur la fonction économique.
5) On vous annonce que 40 des 200 produits C importés ont été endommagés pendant le voyage et ne sont donc plus disponibles chez l'importateur, quelle conséquence cela aura-t-il sur la solution ?
6) Le prix de vente du produit D peut passer (sans que cela ne gène les ventes) à 4100 €. Quelle sera la conséquence de cette augmentation ?
7) Un ingénieur propose de fabriquer un produit E contenant une unité de B, une unité de C et demandant 4H d'atelier électronique et une demi-heure d'atelier peinture. Ce produit s'il était vendu moins de 5000 € pourrait pénétrer facilement le marché. Quelle décision conseillez-vous à l'entreprise?

PROGRAMMATION DYNAMIQUE

Nous n'étudierons dans ce chapitre que le cas de la programmation dynamique déterministe et où l'ensemble des décisions est fini.

21 Un exemple

Une entreprise doit fabriquer pour les trois semaines à venir 6 unités d'un produit donné. Le coût de production et stockage est des produits, suivant leur semaine de production, est donné dans le tableau suivant :

	Quantités fabriquées						
Semaine	0	1	2	3	4	5	6
1	20	23	29	40	60	75	80
2	20	25	32	42	68	75	80
3	20	26	35	41	66	73	78

Par exemple, si l'on réalise la production demandée avec 1 unité en première semaine et 5 unités en troisième semaine le coût total sera alors de : 23 + 20 +73 = 116

21.1Analyse du problème

Le système S est constitué du département production sur trois semaines. Il peut être considérer comme constitué d'une suite croissante de systèmes emboîtés :

S0=département de production avant la première semaine

S1=département de production la première semaine

S2=département de production les deux premières semaines

S3=S=département de production sur les trois semaines

Avec S0 $\subset$ S1 $\subset$ S2 $\subset$ S3=S, on dit que l'on a décomposé le problème en trois étapes.

Les actions : il s'agit ici de déterminer les quantités à produire chaque semaine. C'est donc une séquence de trois décisions (appelée stratégie) ; à chaque sous système il est possible d'associer une sous séquence de décisions (appelée sous stratégie).

Les paramètres structurels sont : la quantité totale à fabriquer, les coûts de production.

Les variables d'état sont la quantité totale fabriquée, le coût total de production. Remarquons que nous pouvons associer les mêmes variables d'état au différents sous systèmes définis plus haut, c'est à dire à chaque étape.

Les équations de fonctionnement du système consistent à écrire qu'à la fin de la troisième semaine il a été fabriqué 6 unités du produit et que, chaque semaine, on fabrique une quantité positive ou nulle (on ne détruit pas des unités produites).

La conséquence privilégiée est le coût total et *le critère* est le minimum.

21.2Mise en équation du problème

Nous noterons x1 la quantité fabriquée en semaine 1, x2 celle fabriquée en semaine 2 et x3 la quantité fabriquée en semaine 3.

En notant g1(x), g2(x) et g3(x) les fonctions de coût données par le tableau, la formulation du problème est aisée :

Minimiser F3(x1,x2,x3) = g1(x1) + g2(x2) +g3(x3)

Sous les contraintes :

x1 + x2 + x3 =6

x1, x2, x3 >=0

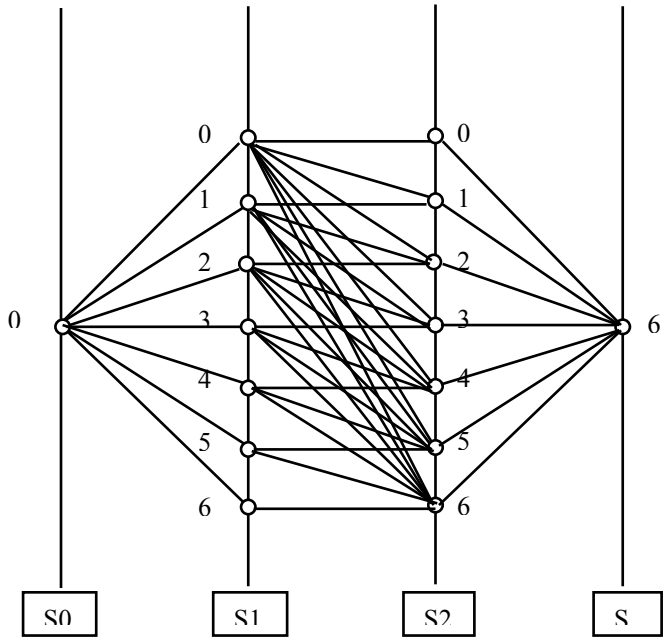
La variable d'état x1 + x2 + x3, se transmet par les sous systèmes S1, S2, la contrainte prenant alors la forme x1<=6 pour S1 et x1 + x2 <=6 pour S3, cette variable joue un rôle particulier pour la programmation dynamique et les différentes valeurs que peut prendre cette variable pour les systèmes S1, S2, S3 s'appellent les états du système pour les différentes étapes. Une décision consiste à passer d'un état à l'étape n à un autre état à l'étape n+1.

La contrainte de positivité des quantités va définir "l'accessibilité" d'un état de l'étape n+1 à partir d'un état de l'étape n : par exemple l'état 4 de l'étape 2 est accessible des états 0,1,2,3,4 de l'étape 1, mais des états 5 ou 6 de l'étape 1.

Enfin la valeur de la fonction économique est pour la stratégie (x1,x2,x3) est égale à la valeur de la fonction économique pour la sous stratégie (x1,x2) plus la valeur de la décision x3. On pourrait faire la même remarque pour la sous stratégie (x1,x2).

21.3 Représentation graphique

On peut donner une représentation graphique du problème sous forme de graphe, en marquant les différentes étapes (sous systèmes), et états du système sur des lignes verticales et en joignant par une ligne un état de l'étape n et un état de l'étape n+1 si celui-ci est accessible. On obtient alors la représentation suivante :



Il s'agit de trouver le chemin de coût minimum qui partant de l'état initial 0 du systeme S0, atteint l'état final 6 du sytème S3. Il serait possible ici d'explorer tous les chemins, mais nous allons montrer sur cet exemple, un algorithme permettant de diminuer de façon significative la combinatoire des chemins.

21.4 Résolution du problème.

La fonction économique peut s'écrire

f(x1,x2,x3)=(g1(x1)+ g2(x2)) + g(x3) avec x1+x2+x3=6

On peur alors écrire :

$$\underset{x1+x2+x3=6}{Max} f(x1,x2,x3) = \underset{x3}{Max}\left(g3(x3) + \underset{x1+x2=6-x3}{Max}\left(g1(x1)+g2(x2)\right)\right)$$

C'est à dire qu'il n'est pas nécessaire de mémoriser tous les chemins qui conduisent de l'état initial à un état donné du système S2, mais seulement ceux qui correspondent au maximum de la fonction économique restreinte à S2. Ceci pourrait s'énoncer de la façon suivante : "toute sous stratégie d'une stratégie optimale est optimale". Attention cela ne signifie pas que pour chaque étape il ne faut conserver que le meilleur état (c.a.d. celui correspondant au coût minimum) mais qu'il suffit de conserver pour *chaque étape et pour tous les états de cette étape la sous stratégie conduisant au coût minimum*. Ceci nous permettra de réduire à chaque étape le nombre de "chemins à explorer".

Appliquons ce principe à la résolution du problème. Nous allons construire des tableaux concernant les différentes étapes, en mettant en ligne les états de l'étape n et en colonne les états de l'étape n+1, chaque case du tableau contenant la valeur de la fonction économique pour l'étape n+1. La colonne la plus à gauche du tableau contenant la valeur optimale de la fonction économique à l'étape n (pour chaque état), la dernière ligne contenant la valeur optimale de la fonction économique pour chacun des états de l'étape n+1.

Première étape : passage du système S0 au système S1

Le seul état possible pour S0 est 0, les états possibles pour S1 sont les productions possibles en première semaine soit (0,1,2,3,4,5,6).

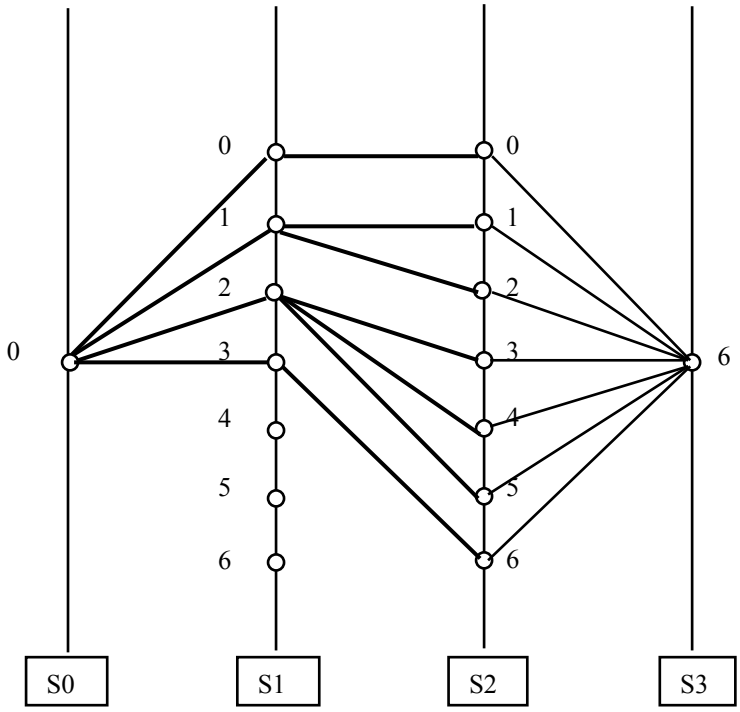
		S1						
Optimum S0	S0	0	1	2	3	4	5	6
0	0	20	23	29	40	60	75	80
Optimum (S1)		**20**	**23**	**29**	**40**	**60**	**75**	**80**

Deuxième étape : passage de S1 à S2

Les états possibles pour S2, correspondent aux productions cumulées des semaines 1 et 2, et sont donc toutes les valeurs (0,1,2,3,4,5,6). Toutefois comme la production de la deuxième semaine ne peut être négative, seuls les états (de S2) de valeur supérieure ou égale sont accessibles à partir d'un état du système S1 ; c'est pour toute la partie sous la diagonale du tableau des valeurs de la fonction économique est vide.

		S2						
Optimum S1	S1	0	1	2	3	4	5	6
20	0	40	45	52	62	88	95	100
23	1		43	48	55	65	91	98
29	2			49	54	61	71	97
40	3				60	65	72	82
60	4					80	85	92
75	5						95	100
80	6							100
Optimum S2		**40**	**43**	**48**	**54**	**61**	**71**	**82**

A ce stade les seuls "chemins" conservés sont ceux qui correspondent à l'optimum de la fonction économique pour chaque état. C'est à dire que le graphe, pour l'étape suivante est réduit à :

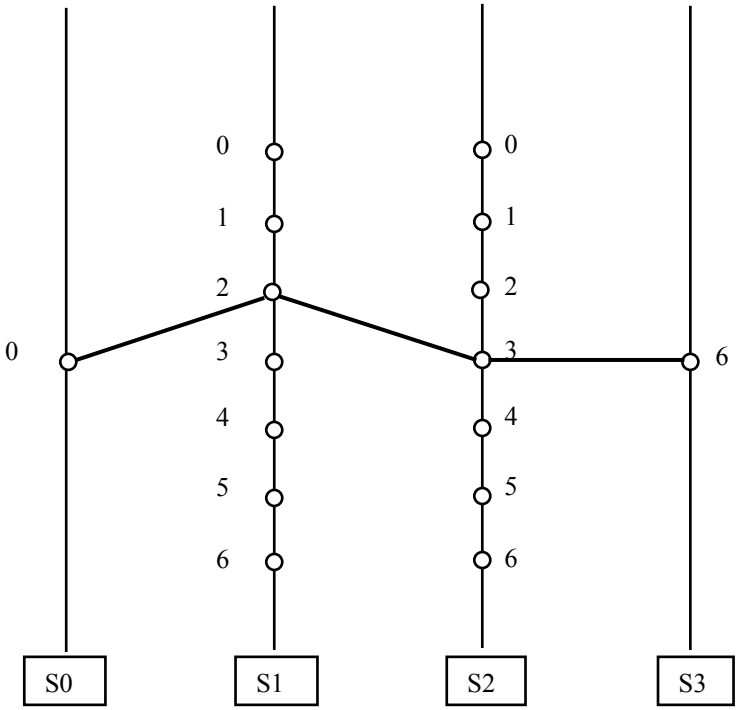


Troisième étape : passage de S2 à S3=S

Les états possibles pour le système S3 se résument au seul état 6, puisque la quantité à fabriquer sur les 3 semaines est fixée. On obtient donc le tableau final suivant :

		S3
Optimum S2	S2	6
40	0	118
43	1	116
48	2	114
54	3	95
61	4	96
71	5	97
82	6	102
Optimum S2		**95**

Le chemin optimal est alors le suivant ("en remontant les tableaux") :



La politique de production correspondante est : fabriquer 2 unités en première semaine, 1 unité en seconde semaine et 3 unités en dernière semaine, pour un coût total de 95.

Remarques :

1. Dans la mesure où l'état final était aussi unique, on aurait pu procéder de façon rétrograde, en partant de la dernière semaine, l'exercice est laissé au lecteur.
2. Une exploration exhaustive de tous le chemins aurait conduit à (7+6+5+4+3+2+1)*2 = 56 additions pour l'évaluation des chemins et 28 comparaisons pour trouver l'optimum. L'algorithme que nous avons utilisé ne demande que 28 additions et (28+6) comparaisons. soit un gain de 22 opérations. La réduction aurait été encore plus importante si le nombre d'étapes avait été plus grand.

22 Formalisation à l'aide de la programmation dynamique

22.1 Caractéristiques d'un problème de programme dynamique discret

Pour qu'un problème puisse être formalisé en termes de programmation dynamique, il faut que l'on puisse définir des étapes c'est à dire une suite croissante de sous systèmes ; nous ne considérerons ici que le cas où cette suite est finie : S0⊂ S1⊂..⊂Sn=S.

A chaque étape i sont associées des décisions qui concerne le passage du sous système Si au sous système Si+1. Nous supposerons ici que ces décisions sont en nombre fini. Une nuplet composé d'une décision pour chaque étape est appelée une stratégie : (d1,d2,..,dn). Un sous-ensemble de décisions consécutives est appelé une sous stratégie (par exemple (d2,d3,d4)).

A chaque étape sont associées des variables d'état privilégiées, dont l'ensemble des valeurs possibles est appelé ensemble des états du système à l'étape i. Nous supposerons aussi que ces valeurs sont en nombre fini. Les états du système S0 s'appellent les états initiaux, ceux du système Sn les états finaux. *Les états du système à l'étape i, représentent les différentes conséquences possibles de toutes les sous stratégies* (d1,d2,..di).

Un état e_{i+1}, de l'étape i+1, est dit accessible à partir d'un état e_i, de l'étape i, s'il existe une décision di permettant de passer de e_i à e_{i+1}. Ce sont les contraintes de fonctionnement du système qui définissent l'accessibilité d'un état par rapport à un autre.

Enfin la fonction économique est définie comme la somme des valeurs des décisions d'une stratégie, cette fonction dépend donc des différents états par lesquels passe la stratégie au cours des n étapes.

22.2 Le principe de Bellman

Nous raisonnerons ici dans le cas d'une maximisation.

Notons di la décision à l'étape i, et ei l'état atteint à cette étape, la fonction économique peut s'écrire :

$$f(d1,d2,...dn,e1,e2,..en) = \sum_{i=1}^{i=n} g_i(di,ei) = \sum_{i=p+1}^{i=n} g_i(di,ei) + \sum_{i=1}^{i=p} g_i(di,ei)$$

ce qui revient simplement à décomposer les n étapes en deux sous-ensembles : les étapes 1 à p et les étapes p+1 à n.

Pour un état *ep+1* fixé, notons :

$$f_p(ep+1) = \max_{d1,d2,dp,e1,e2,ep} (g_1(d1,e1) + g_2(d2,e2).. + g_p(dp,ep))$$

Il est alors clair, d'après l'additivité de la fonction économique que :

$$\max_{d1,d2,dn,e1,e2,en} f(d1,..,dn,e1,..,en) = \max_{dp+1,..,dn,ep+1,..,en} (f_p(ep+1) + g_{p+1}(dp+1,ep+1) + .. + g_n(dn,en))$$

Ce qui revient à dire que la sous stratégie menant de l'état *e1* à l'état *ep+1* est optimale, ce qui s'énonce sous le nom de *principe de Bellman* :

Toute sous stratégie d'une stratégie optimale est elle-même optimale.

On peut alors résoudre le problème par récurrence, pour chaque état terminal de l'étape i, il suffit de déterminer les stratégies optimales conduisant à cet état, les autres stratégies sont sans intérêt pour la suite de la résolution.

Pour démarrer la résolution on partira de l'ensemble des états initiaux ou finaux le plus simple, c'est à dire celui qui a le moins d'éléments ; dans l'exemple traité plus haut ces deux ensembles n'avaient qu'un élément, il était donc indifférent de partir de l'un ou de l'autre.

22.3 Méthode de résolution

Bien qu'il n'y ait pas de méthode générale permettant de résoudre un programme dynamique, avec les restrictions que nous nous sommes imposées (problème déterministe, nombre fini d'étape, de décisions et d'états à chaque étape) il est souvent possible d'utiliser une présentation identique à celle que nous avons utilisée lors de l'exemple.

Pour chaque étape on construira donc un tableau rectangulaire ayant la présentation suivante :

		Etats de l'étape i+1			
Valeurs optimales de l'étape i	Etats de l'étape i	$\mathbf{E_{1,i+1}}$		$\mathbf{E_{k,i+1}}$	
$V_{1,i}$	$E_{1,i}$				
$V_{j,i}$	$E_{j,i}$				
Optima à l'étape i+1					

Dans la cellule se trouvant à l'intersection de la ligne de l'état $E_{j,i}$ (de l'étape i) et de la colonne de l'état $E_{k,i+1}$ (de l'étape i+1) on indiquera la valeur de la fonction économique pour atteindre l'état $E_{k,i+1}$ en passant par l'état $E_{j,i}$ s'il existe une décision di permettant ce passage, cette valeur est : $V_{j,i} + g_i(di)$; sinon on indique la non-accessibilité de l'état.

Dans la dernière ligne on détermine pour chaque état de l'étape i+1, la valeur optimale de la fonction économique pour atteindre cet état.

La dernière étape permet de déterminer la valeur optimale de la fonction économique, pour déterminer la stratégie correspondante, il suffit de "remonter" les tableaux, ce qui donne la suite des états et d'en déduire les décisions correspondantes. Remarque, il est aussi possible de rajouter au tableau une ligne mémorisant, à chaque étape et pour chaque état, la décision optimale.

23 Mise en place sous Excel

Nous allons reprendre l'exercice d'introduction et expliquer les formules utilisées pour la résolution de cet exemple sous Excel (fichier Stocks_Dyn.xls).

	A	B	C
2	*Optimum S1*	S1	0
3	*=INDEX(Couts;1;B3+1)*	0	=SI($B3<=C$2;$A3+INDEX(Couts;2;C$2-$B3+1);"")
4	*=INDEX(Couts;1;B4+1)*	1	=SI($B4<=C$2;$A4+INDEX(Couts;2;C$2-$B4+1);"")
8	*=INDEX(Couts;1;B8+1)*	5	=SI($B8<=C$2;$A8+INDEX(Couts;2;C$2-$B8+1);"")
9	*=INDEX(Couts;1;B9+1)*	6	=SI($B9<=C$2;$A9+INDEX(Couts;2;C$2-$B9+1);"")
10	**Optimum S2**		**=MIN(C3:C9)**

Pour la deuxième étape, nous avons, des formules particulières pour les optima précédents, qui viennent directement du tableau des données :

Le tableau de données des coûts a été nommé « Couts », la colonne B contient les états du système S1 (la production de la première semaine), la ligne 2 contient les états du système S2 (la production des deux premières semaines).

Pour afficher le coût associé à chaque état de S1, chaque production de la première semaine, il suffit d'aller lire dans le tableau de données l'élément de la première ligne correspondant, ceci se fait avec index, l'indice de la ligne est 1, celui de la colonne l'état+1, puisque ces états commencent à 0.

Pour les cases de calcul transitoire, il faut tout d'abord vérifier que l'état de S2 est accessible par l'état de S1, ce qui est fait avec la condition $B3<=C$2 pour la première case (attention aux $ pour la recopie), si cette condition n'est pas vérifiée, rien n'est affichée dans la case, sinon le coût correspondant est affiché : ce coût est égal au minimum de l'état de départ plus le coût de production de la seconde semaine correspondant à C$2-$B3 produits. Ce dernier coût se lit dans le tableau de données initiales, comme précédemment, mais dans la ligne 2.

Enfin la dernière ligne contient le coût minimum de chaque état du système S2, qui nous servira dans l'étape suivante.

	A	B	C
11	***Production Semaine 1***		***=INDEX($B$3:$B$9;EQUIV(C10;C3:C9;0))***
12	***Production Semaine 2***		***=C2-C11***

Il est enfin possible avec Excel de connaître pour chaque état final, le chemin optimal, c'est ce que nous allons faire en rajoutant deux lignes à notre tableau :

Pour trouver la production de la première semaine, il suffit d'aller lire dans la colonne B l'élément qui se trouve sur la ligne du minimum de la colonne courante, c'est ce que fait la fonction EQUIV (avec comme dernier argument 0, et comme premier argument le minimum), la fonction index retourne alors la valeur cherchée.

La production de la deuxième semaine est obtenue par simple différence entre la production des deux semaines et la production de la première semaine.

	A	B	C	D	E	F	G	H	I
1			S2						
2	*Optimum S1*	S1	0	1	2	3	4	5	6
3	*20*	0	40	45	52	62	88	95	100
4	*23*	1		43	48	55	65	91	98
5	*29*	2			49	54	61	71	97
6	*40*	3				60	65	72	82
7	*60*	4					80	85	92
8	*75*	5						95	100
9	*80*	6							100
10	**Optimum S2**		**40**	**43**	**48**	**54**	**61**	**71**	**82**
11	***Production Semaine 1***		***0***	***1***	***1***	***2***	***2***	***2***	***3***
12	***Production Semaine 2***		***0***	***0***	***1***	***1***	***2***	***3***	***3***

On retrouve alors les résultats obtenus en 21.4 :

	A	B	C
14			
15			S3
16	*Optimum S2*	S1	6
17	*=INDEX($C$10:$I$10;B17+1)*	0	=SI($B17<=C$16;$A17+INDEX(Couts;3;C$16-$B17+1);""
23	*=INDEX($C$10:$I$10;B23+1)*	6	=SI($B23<=C$16;$A23+INDEX(Couts;3;C$16-$B23+1);""
24	**Optimum S2**		**=MIN(C17:C23)**
25	***Production Semaine 1+2***		***=INDEX($B$17:$B$23;EQUIV(C24;C17:C23;0))***
26	***Production Semaine 3***		***=C16-C25***
27	***Production Semaine 2***		***=INDEX($C$12:$I$12;C25+1)***
28	***Production Semaine 1***		***=INDEX($C$11:$I$11;C25+1)***

Pour l'étape suivante, signalons simplement les différences, la première colonne du tableau est obtenue en lisant la valeur de l'optimum précédent, les formules internes au tableau sont les mêmes, en changeant cependant la ligne du tableau de données (3 et non 2) :

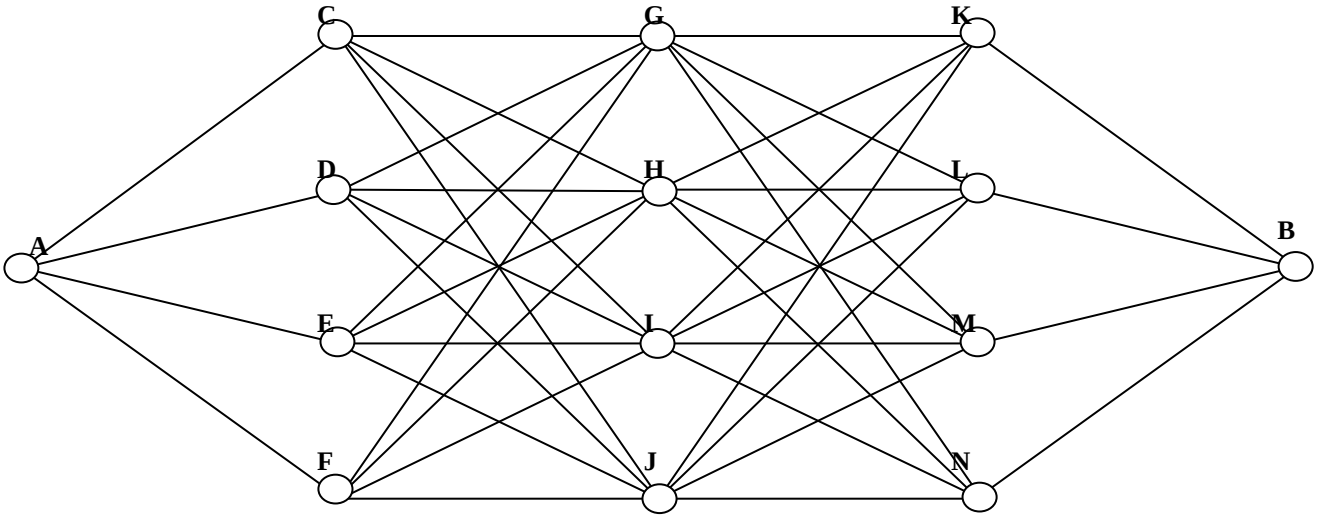
Pour les productions optimales, il faut passer par l'intermédiaire de la production des semaines 1 et 2, que l'on décompose en utilisant les résultats de l'étape précédente.

Remarque : nous avons donner les formules internes au tableau avec les adresses relatives et absolues, bien qu'ici ce soit inutile puisqu'il n'y a qu'un seul état ; mais elles seraient nécessaires si le problème avait plus de trois étapes.

EXERCICES DE PROGRAMMATION DYNAMIQUE

24 Plus court trajet

Il s'agit de déterminer le plus court chemin menant de la ville A à la ville B, les villes intermédiaires et les distances entre ces villes vous sont données ci dessous :



	C	D	E	F
A	47	45	39	38

	G	H	I	J
C	13	48	37	41
D	11	28	48	27
E	27	44	47	20
F	30	16	44	22

	K	L	M	N
G	47	16	31	27
H	50	12	15	42
I	27	35	44	25
J	35	47	39	21

	B
K	39
L	28
M	49
N	14

Questions :

1) Montrer que ce problème peut se formaliser en un programme dynamique, préciser les étapes, les états à chaque étape, les décisions et la fonction économique.
2) Résoudre le problème. Quel(s) est (sont) le(s) chemin(s) optimal(aux)?

25 Aerospa

La société Aerospa doit sous-traiter la construction de 10 ogives de fusée en céramique pour la fin du mois. Elle s'est adressée à trois sous-traitants qui lui ont fait les propositions suivantes :
Sous-traitant 1

Nombre de pièces	1	2	3	4	5
Prix en K€	38	65	100	143	185

Sous-traitant 2

Nombre de pièces	2	4	6	8
Prix en K€	75	150	220	280

Sous-traitant 3

Nombre de pièces	3	6	9
Prix en K€	90	200	315

Questions :

1) Montrer que ce problème peut être traité à l'aide de la programmation dynamique. Préciser les étapes, les états, les décisions et la fonction économique.
2) Résoudre le problème.

26 Choix d'investissement

Une société d'investissement envisage de placer jusqu'à 10M$, qu'elle peut investir dans quatre types de projets collectifs, l'unité d'investissement étant le million de $. Elle peut répartir son investissement comme elle l'entend, par exemple tout placer dans le projet B, ou bien placer 3M$ en A, 2 en B, 4 en D, par exemple.

Le tableau ci-dessous montre le profit qui résultera de chaque investissement : ainsi, un placement de 5M$ en B rapporterait 0,9M$ et un placement de 3M$ en D rapporterait 0,42M$.

Placement	**A**	**B**	**C**	**D**
0	0,00	0,00	0,00	0,00
1	0,28	0,25	0,15	0,20
2	0,45	0,41	0,25	0,33
3	0,65	0,55	0,40	0,42
4	0,78	0,65	0,50	0,48
5	0,90	0,75	0,62	0,53
6	1,02	0,80	0,73	0,56
7	1,13	0,85	0,82	0,58
8	1,23	0,88	0,90	0,60
9	1,32	0,90	0,96	0,60
10	1,38	0,90	1,00	0,60

Questions :

1) Montrer que ce problème peut être traité à l'aide de la programmation dynamique. Préciser les étapes, les états, les décisions et la fonction économique.
2) Résoudre le problème.

27 La société Médiajeux

La société Médiajeux lance une campagne nationale pour un nouveau jeu. Elle veut appuyer cette campagne par une campagne régionale dans 4 régions. Pour cela elle a sélectionné quatre radios locales ayant une forte audience et se propose de passer un certain nombre de spots publicitaires durant le mois à venir.

D'après les campagnes précédentes l'apport de ventes supplémentaires du aux spots peut être évalué, en fonction du nombre de spots diffusés par jour, selon le tableau suivant :

Programmation dynamique

	Nombre de ventes supplémentaires			
Nombre de spots	Région1	Région2	Région3	Région4
0	0	0	0	0
1	1000	700	1400	600
2	2500	2500	9000	1800
3	7500	8000	34000	4000
4	18000	21000	54000	9000
5	32000	42000	59000	19000
6	41000	59000	60000	32000
7	44500	66000	60000	43000
8	45500	69000	60000	49000
9	46000	70000	60000	52000
10	46200	70000	60000	53000
11	46200	70000	60000	53500
12	46200	70000	60000	53600

D'autre part la marge réalisée sur chaque vente est de 20€ et le coût d'un spot publicitaire dépend de la radio locale, le tableau suivant vous donne le coût mensuel d'un spot journalier :

	Région1	Région2	Région3	Région4
Prix mensuel d'un spot	40 000 €	80 000 €	60 000 €	40 000 €

On dispose d'un budget de 400 000 €

Questions :

1) Montrer que ce problème peut se formaliser sous forme de programmation dynamique : on précisera les étapes, l'état du système à chaque étape, les décisions à chaque étape, la fonction économique.
2) Résoudre alors le problème d'allocation du budget aux différentes radios locales.
3) La direction de Médiajeux veut absolument passer au moins un spot dans chaque radio, quelle est alors la meilleure allocation du budget ?

28 Exploitation minière

Le schéma ci-dessous vous donne les estimations profit d'exploitation d'une mine(vue en coupe verticale) :

-4	**-4**	**-4**	**-4**	**8**	**12**	**12**	**0**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**
	-4	**-4**	**-4**	**0**	**12**	**12**	**8**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	**-4**	
		-4	**-4**	**-4**	**8**	**12**	**12**	**0**	**-4**	**8**	**-4**	**4**	**-4**	**-4**	**-4**		
			-4	**-4**	**0**	**12**	**12**	**8**	**-4**	**8**	**4**	**4**	**-4**	**-4**			
				-4	**-4**	**8**	**12**	**12**	**0**	**-4**	**-4**	**-4**	**-4**				
					-4	**0**	**12**	**12**	**8**	**-4**	**-4**	**-4**					
						-4	**8**	**12**	**12**	**0**	**-4**						
							0	**12**	**12**	**8**							

Il s'agit de déterminer la stratégie optimale de creusement de cette mine, sachant que la pente maximale doit rester inférieure à 45°.

Questions :

1) Montrer comment ce problème peut être formalisé en utilisant la programmation dynamique : quelles sont les étapes, l'état du système à chaque étape, les décisions, la fonction économique et le critère.
2) Déterminer alors le (ou les) programme(s) optimal(aux) d'exploitation.

29 Entreprise ABC

Une entreprise ABC doit fabriquer 10 unités d'un produit X dans la journée. Pour ce faire, elle dispose de trois machines M1, M2, M3 de capacité de production journalière respective de 9, 8 et 5 unités.

Les marges dégagées par les différents niveaux de production pour les trois types de machine sont données dans le tableau suivant

	Production									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M1	-360	-120	0	120	360	480	600	840	960	1080
M2	-360	-160	-50	150	260	460	570	770	880	
M3	-300	-140	20	180	340	500				

Questions :

3) Montrer comment ce problème peut être formalisé en utilisant la programmation dynamique : quelles sont les étapes, l'état du système à chaque étape, les décisions, la fonction économique et le critère.
4) Déterminer alors le (ou les) programme(s) optimal(aux) de production.

30 La société Copsi-Cola (univers probabiliste)

La société Copsi-Cola produit des boissons rafraîchissantes dans son usine de la région Midi-Pyrénées, dont la capacité de production est de 1200 T par semaine. La demande est connue une semaine à l'avance, ce qui permet théoriquement de produire exactement la quantité nécessaire, si cette demande est inférieure à 1200 T, ce qui est le cas toute l'année sauf durant les treize semaine de la saison estivale, période durant laquelle, il est possible suivant les conditions météorologiques que la demande excède la capacité de production, ce qui conduit à constituer des stocks. La demande hebdomadaire durant les treize semaines considérées peut être considérée comme prenant six valeurs équiprobables données dans le tableau suivant :

Semaine	Demandes équiprobables					
1	600	700	800	900	1000	1100
2	600	700	800	900	1000	1100
3	800	900	1000	1100	1200	1300
4	900	1000	1100	1200	1300	1400
5	900	1000	1100	1200	1300	1400
6	1000	1100	1200	1300	1400	1500
7	1000	1100	1200	1300	1400	1500
8	800	900	1000	1100	1200	1300
9	700	800	900	1000	1100	1200
10	800	900	1000	1100	1200	1300
11	900	1000	1100	1200	1300	1400
12	800	900	1000	1100	1200	1300
13	800	900	1000	1100	1200	1300

La direction commerciale estime que le pourcentage de rupture ne doit pas excéder 5% de la demande, mais devant les demandes du service de planification, elle admet que l'on peut considérer que le coût de rupture est d'environ 19 fois le coût de stockage, de manière à pourvoir quantifier le coût d'une politique.

Questions :

1) L'an passé, la demande sur les 13 semaines a été la suivante :

Semaine	1	2	3	4	5	6	7	8	9	10	11	12	13
Demande	1000	1100	1300	1300	1100	1200	1400	1300	1200	1100	1300	1200	1300

Quelle aurait du être la politique de production et de stockage pour n'avoir sur la période aucune rupture de stocks?

Quelle aurait du être la politique de production pour minimiser le coût total (stockage °+ rupture) sur la période?

En raisonnant sur l'espérance de coût, quel est d'après vous la meilleure politique de production et de stockage sur la période? (Indication : on essaiera de déterminer pour chaque semaine un niveau idéal de stocks, niveau maximal qu'il ne sera pas toujours possible d'atteindre)

LA SIMULATION

31 Définition

Simulation : méthode de mesure et d'étude consistant à remplacer un phénomène, un système par un modèle plus simple mais ayant un comportement analogue (Larousse).

Le système ou phénomène analysé peut être schématisé sous forme d'un modèle mécanique, électronique ou logico-mathématique. Nous nous intéresserons ici uniquement à la représentation du système sous la forme d'un modèle informatisable.

L'objectif d'un modèle de simulation peut être simplement descriptif : étudier le comportement d'un système sous différentes hypothèses d'évolution de l'environnement, ou aussi normatif (décisionnel): en simulant plusieurs décisions envisagées choisir la meilleure ou la moins mauvaise.

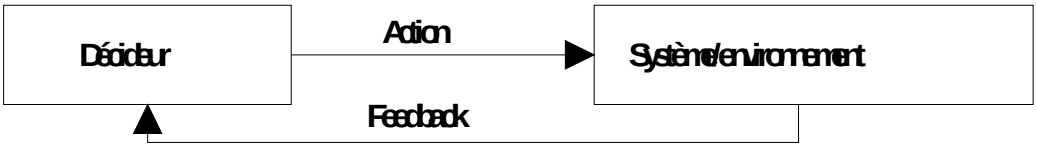
32 Typologie des modèles de simulation

Une première segmentation possible des modèles de simulation peut se faire en fonction du type des connaissances que l'on a sur le système et son environnement. Si cette connaissance est certaine, on parlera de simulation déterministe; s'il est possible (en fonction des expériences passées ou de l'expérience) de probabiliser l'apparition de différents états, on parlera alors de simulation probabiliste.

33 La simulation déterministe

La simulation déterministe est fréquemment utilisée pour la création de scénarii. L'utilisateur teste ainsi les conséquences de diverses hypothèses sur l'évolution du système et de son environnement (cf. les exercices d'introduction à Excel).

La dynamique industrielle, inventée par Forrester, est un autre exemple de modèle de simulation déterministe; elle s'intéresse essentiellement aux systèmes cybernétiques, c'est-à-dire aux systèmes avec boucle de feed-back.



La boucle de feed-back envoie au "décideur" des informations sur le système et son environnement, qui lui permettent de modifier de façon automatique son action à chaque instant. Par exemple un thermostat capte la température ambiante, ce qui lui permet de régler le chauffage en fonction d'un objectif; une usine peut modifier sa production en fonction de la demande constatée sur le marché et du niveau de ses stocks.

34 La simulation probabiliste

Dans ce cas, les événements qui apparaissent lors de l'évolution du système ne sont pas connus avec certitude, mais on est capable de probabiliser cette apparition: par exemple, dans une étude de files d'attente à un guichet, on peut donner la loi de probabilité du temps séparant deux arrivées et éventuellement aussi la loi de probabilité du temps de service.

34.1 Propriétés des modèles de simulation probabiliste

Un modèle de simulation probabiliste permet d'étudier le comportement temporel d'un système dont certains paramètres structurels sont donnés sous forme de loi de probabilité. Les caractéristiques des modèles de simulation probabiliste sont les suivantes :

- Environnement et le système : définis sur une période (jour, mois, année,..) divisée en sous périodes, le nombre de sous périodes peut être fixe (heure, jour,..)ou non (arrivée d'un client, fin de service,..) ; voir plus loin la différence entre simulation événement et simulation temps.
- Les décisions sont en nombre fini, ce nombre est souvent assez faible.
- Les paramètres structurels sont pour certains définis par des lois de probabilité (arrivées de clients à une caisse, temps de service, demande..), d'autres sont déterministes (coûts de production, coût d'un spot)
- Les variables d'état sont des variables aléatoires, c'est à dire que leurs valeurs suivent des lois de probabilités, qu'il n'est généralement pas possible de (ou que l'on ne sait pas) calculer analytiquement. Ces variables d'états sont définies soit au niveau de la sous-période (attente du dernier client arrivé, stock en début de sous période), puis sont éventuellement agrégées au niveau de la période.
- Les équations de fonctionnement sont les équation définissant le passage de la valeur d'une d'état d'une sous période à la sous période suivante.
- Le modèle d'évaluation porte donc sur des variables aléatoires (agrégation sur la période des variables d'état), plus précisément sur des paramètres de ces variables (moyenne, écart type, fractile).Il est donc nécessaire d'approcher la distribution des variables aléatoires de façon empirique en itérant le modèle d'une période.

34.2 Simulation temps et simulation événement

Pour analyser un phénomène aléatoire, on peut raisonner de deux façons différentes : soit on compte le nombre d'événements se produisant pendant un intervalle de temps fixe, soit on détermine le temps séparant deux événements. Dans le premier cas, on parle de simulation temps, dans le second cas de simulation événement.

Reprenons l'exemple de la file d'attente:

- pour une simulation-temps, on se donne la loi de probabilité du nombre d'arrivées pendant un intervalle de temps fixe, par exemple toutes les 10 minutes, dans ce cas la sous période sera l'intervalle de 10mn qui sera considéré comme insécable. Si nous travaillons sur une demi journée de 4H (la période), il y aura donc exactement 24 sous période. Les variables d'état seront donc évaluées toutes les 10mn.
- pour une simulation-événement, on se donne la loi de probabilité du temps séparant deux arrivées. La sous période correspond à ce temps, la fin d'une sous période correspondant à l'arrivée d'un nouveau client. Dans ce cas on ne sait pas à priori combien de sous périodes apparaîtront dans la période, ce nombre va dépendre du nombre d'arrivées de client pendant la demi-journée.

En règle générale une simulation événement permet une analyse plus fine du système, mais sa réalisation informatique (sur tableur du moins) est plus délicate et son coût de traitement plus élevé.

34.3Simulation d'une loi de probabilité

Pour pouvoir simuler le comportement d'un système faisant intervenir des événements probabilisés, il va falloir simuler l'apparition de ces événements; c'est-à-dire générer des événements dont la fréquence observée sur un grand nombre de simulations doit être proche de la loi de probabilité théorique.

Remarquons que la simulation d'une loi de probabilité quelconque peut se ramener à la simulation d'une loi uniforme sur l'intervalle [0;1[.

En effet, soit une loi de probabilité discrète définie par $P(X=x_i)=p_i$ pour i=1,..,n; supposons qu'il existe une méthode m permettant de simuler une loi uniforme sur l'intervalle [0;1[, c'est-à-dire que P(x<=m<x+dx)=dx pour tout x de l'intervalle [0;1[. Définissons la règle d'affectation suivante (m_o étant le résultat obtenu par la méthode m lors d'une expérience):

si $0<=m_o<p_1$ alors $X=x_1$

si $p_1<=m_o<p_1+p_2$ alors $X=x_2$

.....

si $p_1+...+p_k<=m_o<p_1+...+p_k+p_{k+1}$ alors $X=x_{k+1}$

Il est alors clair que l'on simule ainsi la loi de probabilité initiale, puisque la probabilité d'obtenir lors de l'expérience le résultat $X=x_{k+1}$ est égale à la probabilité d'obtenir par la méthode m un résultat dans l'intervalle $[p_1+...+p_k;p_1+...+p_k+p_{k+1}]$ soit p_{k+1}.

Remarque : il est bien évident que l'on pourrait prendre une autre partition de l'intervalle [0;1] qui conduirait à une autre méthode de simulation de la loi de probabilité initiale. Cependant la partition utilisée fait intervenir la fonction de répartition de la loi et est généralisable au cas d'une loi continue, sous réserve de savoir inverser cette fonction de répartition.

Pour une loi de probabilité continue, on peut faire le même raisonnement en considérant la fonction de répartition de la loi (notée F), cette fonction est une fonction croissante continue (pour les cas qui nous intéressent) à valeur dans [0;1[, elle est donc bijective et à tout élément a de l'intervalle [0;1[on peut associer un élément x tel que F(x)=a.

34.3.1 Fonction pseudo aléatoire

On appelle fonction pseudo aléatoire une fonction (évidemment déterministe) qui permet de simuler une loi uniforme sur l'intervalle [0;1[. Cette fonction doit avoir les propriétés suivantes :

Les valeurs prises par cette fonction doivent être uniformément réparties sur l'intervalle [0;1[.

Des résultats consécutifs doivent être indépendants.

Dans la pratique, ces fonctions sont réalisées par des méthodes de congruence, ce qui signifient que les résultats sont périodiques, mais la période est suffisamment longue (plusieurs milliards) pour que cela ne gêne pas la réalisation de simulations. Dans le cas d'Excel, cette fonction se nomme **alea()**.

34.3.2 Simulation d'une loi de probabilité discrète avec Excel.

Pour simuler une loi de probabilité discrète sous Excel on utilise en général la recherche dans une table contenant dans la première ligne(ou colonne) 0 et les probabilités cumulées, et dans

la seconde ligne (ou colonne) les valeurs prises par la variable aléatoire. La valeur recherchée dans la table étant la valeur prise par la fonction alea(). Si cumul est le nom de la table, on utilisera donc la formule :

RECHERCHEH(ALEA();cumul;2) (ou RECHERCHEV(ALEA();cumul;2)).

Dans certains cas particuliers on peut se passer de table de recherche : par exemple pour simuler le jet d'un dé on peut utiliser la formule **ENT(6*ALEA())+1)**. De façon plus générale pour simuler une loi discrète a valeur entière sur l'intervalle [p;q], on utilisera la formule :

ENT((q-p+1)*ALEA())+p

34.3.3 Simulation de certaines loi continue avec Excel

Il est possible avec Excel de simuler toutes les lois continues dont les fonction de répartition inverses sont des fonctions d'Excel. Sinon il faut que l'utilisateur définisse lui-même une fonction permettant de calculer cette inverse (par une macro par exemple).

Nous donnerons ici deux exemples calculables avec Excel, la loi normale et la loi exponentielle.

Pour simuler un tirage aléatoire dans une loi normale $\boldsymbol{N}(\mu,\sigma)$, on utilisera la formule :

LOI.NORMALE.INVERSE(ALEA();μ;σ)

Pour une loi exponentielle de paramètre l, on utilisera le fait le fait que la loi exponentielle de paramètre λ est la loi gamma particulière de paramètres 1 et $1/\lambda$. L' inverse de la fonction de répartition de la loi gamma étant donnée dans Excel, on utilisera la formule

LOI.GAMMA.INVERSE(ALEA();1;λ)

34.4Construction d'un modèle de simulation

Après avoir délimité dans le temps et dans l'espace le système dont on veut étudier le comportement, la construction du modèle comportera deux phases:

- tout d'abord construire un modèle "classique", en séparant bien paramètres et équations, qui permette d'obtenir la réalisation d'une période, divisée en sous période. Ce modèle aura évidemment en entrée des événements aléatoires, donc les sorties correspondant aux critères d'évaluation vont changer à chaque exécution.
- construire une boucle de simulation qui itère le calcul du modèle précédent de façon à obtenir une estimation de la loi ou de certains paramètres des critères

Pour réaliser les itérations sous Excel, on peut procéder de différentes façons ; nous allons en exposer les trois principales sur un exemple.

34.5Exemple: Gestion de stocks

On considère une entreprise distribuant un produit A dont la demande mensuelle suit une loi de probabilité uniforme sur l'intervalle de nombres entiers [400;1000] . Chaque mois l'entreprise envisage de commander 700 unités (quantité appelée **Commande**) qui seront disponibles le mois suivant. Le responsable commercial aimerait estimer les ruptures de stocks sur une année.

34.5.1 Construction d'un modèle annuel

Le système et l'environnement que nous étudions est constitué du magasin, des fournisseurs et des clients sur une année, divisée en mois puisque les commandes sont mensualisées.

La décision que nous avons à prendre est le niveau de commande (actuellement 700).

Les paramètres structurels sont ici simplement la demande qui est probabilisée, on pourrait aussi prendre en compte par exemple un coût unitaire de stockage mensuel moyen, un coût unitaire de rupture.

Les variables d'état sont les éléments qui permettent de suivre mensuellement la satisfaction de la demande, c'est à dire le stock initial, le stock final, le nombre de rupture et le pourcentage de demandes non satisfaites.

Les équations de fonctionnement permettent de calculer au cours du temps l'évolution de ces variables d'état.

Les conséquences retenues par le directeur sont les ruptures, c'est à dire le nombre total de ruptures annuelles et peut-être aussi le pourcentage annuel de demandes non satisfaite.

La mise en équation est la suivante.

Nous allons étudier dans un premier temps le système sur une année soit une période de 12 mois, puisque la demande est mensuelle.

1) Simulation de la demande sur une année.

Chaque mois la demande sera donnée par la formule :

demande(m)=400+ENT(601*ALEA())

2) Calcul des stocks initiaux et finaux du mois(m) :

Stock_initial(m)=Stock_final(m-1)+Commande

Stock_final(m)=Max(Stock_initial(m)-demande(m);0)

On initialisera le stock initial du mois 1 à 0.

3) Calcul de la quantité en rupture chaque mois :

rupture(m)=Max(demande(m)-Stock_initial(m);0)

%rupture(m)=rupture(m)/demande(m)

	A	B	C	D	E	F
1	Commande	700				
2						
3	**Mois**	**Demande**	**Stock Initial**	**Stock Final**	**Rupture**	**%rupture**
4	1	=400+ENT(601*ALEA())	=Commande	=MAX(C4-B4;0)	=MAX(B4-C4;0)	=E4/B4
5	2	=400+ENT(601*ALEA())	=D4+Commande	=MAX(C5-B5;0)	=MAX(B5-C5;0)	=E5/B5
6	3	=400+ENT(601*ALEA())	=D5+Commande	=MAX(C6-B6;0)	=MAX(B6-C6;0)	=E6/B6
7	4	=400+ENT(601*ALEA())	=D6+Commande	=MAX(C7-B7;0)	=MAX(B7-C7;0)	=E7/B7

On peut alors écrire le modèle sous Excel, sur une feuille nommée Modele. Les formules entrées sont les suivantes :

Exemple de simulation sur une année :

	A	B	C	D	E	F
1	Commande	700				
2						
3	**Mois**	**Demande**	**Stock Initial**	**Stock Final**	**Rupture**	**%rupture**
4	1	595	700	105	0	0,00%
5	2	714	805	91	0	0,00%
6	3	563	791	228	0	0,00%
7	4	659	928	269	0	0,00%
8	5	983	969	0	14	1,42%
9	6	902	700	0	202	22,39%
10	7	809	700	0	109	13,47%
11	8	676	700	24	0	0,00%
12	9	893	724	0	169	18,92%
13	10	938	700	0	238	25,37%
14	11	820	700	0	120	14,63%
15	12	639	700	61	0	0,00%
16	Total	9191		Rupture Annuelle	852	
17				% annuel		9,27%

Il nous reste à agréger sur l'année les variables d'état qui vont nous servir de conséquence, par exemple ici le nombre total de rupture sur l'année, ou le pourcentage annuel de rupture :

$$\text{rupture\_annuelle} = \sum_{m=1}^{12} \text{rupture}(m)$$

$$\%\text{rupture\_annuelle} = \frac{\text{rupture\_annuelle}}{\sum_{m=1}^{12} \text{demande}(m)}$$

(La première cellule a pour adresse Modele!E16, la seconde Modele!E17).

Toutefois, comme il a été dit précédemment, à chaque recalcul de la feuille de calcul, les valeurs changent, puisque l'aléa est recalculé. Pour obtenir des résultats utilisables pour la décision, il nous faut donc obtenir des renseignements sur la loi de probabilité des ruptures : par exemple la moyenne des ruptures par an, la fréquence des ruptures supérieures à 5% etc..

34.5.2 Itération du calcul

Il nous faut répéter la simulation annuelle un certain nombre de fois, soit en utilisant des tables pour stocker les résultats, soit en utilisant le mode itératif du tableur soit en programmant une macro.

Utilisation des tables

C'est la méthode la plus simple à mettre en œuvre, les résultats sur modèle de simulation dépendent des valeurs des tirages aléatoires, c'est à dire des résultats de la fonction Alea(), cette fonction n'a pas de paramètres, donc en fait les résultats de notre modèle dépendant d'un paramètre invisible.

Pour obtenir des résultas différents stockés dans une table pour les variables d'état conséquences, il suffira donc de construire une table à un paramètre dont les cellules d'entrée en colonne (ou en ligne)sont associée à une cellule vide, chaque ligne correspondant au résultat d'une itération ; la table doit donc contenir autant de lignes que la taille de l'échantillon que nous voulons constituer.

Le recalcul de la feuille provoquera automatiquement le tirage aléatoire d'autre nombres, donc de nouvelles valeurs des conséquences.

	A	B	C
25	**Itérations**	**=E16**	**=F17**
26	1	=TABLE(;B19)	=TABLE(;B19)
27	=A26+1	=TABLE(;B19)	=TABLE(;B19)
28	=A27+1	=TABLE(;B19)	=TABLE(;B19)
29	=A28+1	=TABLE(;B19)	=TABLE(;B19)
30	=A29+1	=TABLE(;B19)	=TABLE(;B19)

On construit la table stockant les résultats voulus pour un niveau de commande donné, ici le total des ruptures et le pourcentage :

La cellule B19 étant une cellule vide de la feuille.

	A	B	C
22			
23	Moyenne	387,83	0,043851162
24			
25	**Itérations**	**Rupture Annuelle**	**% Annuel**
26	1	811	9,09%
27	2	1859	18,40%
28	3	42	0,50%
29	4	763	8,56%
30	5	480	5,41%
31	6	116	1,36%
32	7	150	1,95%

En utilisant des formats personnalisés simples (pour la première ligne de la table), on obtient alors les résultats suivants :

On peut alors extraire de la table, tous les éléments statistiques qui sont intéressant, sur l'exemple nous nous sommes limités à la moyenne, mais on pourrait (à l'aide delà fonction Fréquence), par exemple, sortir l'histogramme des valeurs.

Il est aussi possible, si l'on veut tester différents niveaux de commande, de construire une table à deux entrées, l'entrée en colonne correspond au numéro de l'itération et l'entrée en

	F	G	H	I
24			**Valeur de la commande**	
25	**=F17**	650	660	670
26	1	=TABLE(B1;B19)	=TABLE(B1;B19)	=TABLE(B1;B19)
27	2	=TABLE(B1;B19)	=TABLE(B1;B19)	=TABLE(B1;B19)
28	3	=TABLE(B1;B19)	=TABLE(B1;B19)	=TABLE(B1;B19)
29	4	=TABLE(B1;B19)	=TABLE(B1;B19)	=TABLE(B1;B19)

ligne à la commande. Cependant dans ce cas, nous sommes limités à un seul critère, ici nous avons choisi le pourcentage de demande non satisfaite :

On pourra alors, faire les statistiques voulues pour chacun des niveaux de commande. Remarque, ici il serait plus intéressant de prendre un indicateur plus synthétique par exemple la somme des coûts de stockage et de rupture.

Notons cependant que, si le nombre de décisions est faible (3 ou 4), il est préférable de construire un modèle permettant de tester, dans un même environnement (c'est à dire avec le même tirage aléatoire), les différentes décisions. Il suffit alors d'une table à une seule entrée pour pouvoir comparer sur plusieurs critères éventuellement les décisions.

Utilisation des itérations

Indiquons par exemple le calcul de la moyenne des ruptures annuelles.

Nous avons besoin de quatre cellules : une cellule drapeau, qui indiquera si les itérations sont commencées, une cellule pour la somme des ruptures obtenues entre l'itération 1 et l'itération

N, une cellule contenant la moyenne des ruptures et enfin une cellule contenant le numéro de l'itération en cours.

Pour calculer la somme des ruptures entre l'itération 1 et N, nous utiliserons la formule :

somme_ruptures(N)=somme_ruptures(N-1)+ ruptures(N)

soit, en ne tenant pas compte des indices,:

somme_ruptures=somme_ruptures+ruptures

la cellule somme_ruptures fait référence à elle-même, il ne faut donc pas oublier de l'initialiser à 0, avant que les itérations ne commencent. La formule contenue dans cette cellule sera alors :

somme_ruptures**=si(drapeau=0;0;somme_ruptures+ruptures)**

D'où la nécessité d'un indicateur de début d'itération, contenu dans la cellule drapeau.

De la même façon, pour obtenir le numéro de l'itération en cours, on écrit la formule :

itération_en_cours**=si(drapeau=0;0;itération_en_cours+1)**

Attention : il nous faut modifier la formule définissant la demande, car l'aléa n'est pas recalculé automatiquement à chaque itération puisqu'Excel ne recalcule que les cellules dépendantes. Dans la demande nous utiliserons la formule :

demande=si(iteration_en_cours>0;400+ent(101*alea());400+ent(101*alea()))

ainsi, comme la cellule iteration_en_cours est modifiée à chaque itération, le test est refait et l'aléa recalculé.

Enfin la moyenne des ruptures sera donnée pour éviter le message d'erreur #DIV/0 (à l'initialisation) par la formule :

moyenne_ruptures**=si(drapeau=0;0;somme_ruptures/itération_en_cours)**

Pour faire fonctionner le modèle, on choisit le mode de calcul manuel et le nombre d'itérations que l'on désire effectuer. On initialise ensuite les valeurs en mettant 0 dans la cellule drapeau, puis en appuyant sur F9. Pour effectuer les itérations on met 1 dans la cellule drapeau, puis on appuiera sur F9.

On obtient alors un tableau semblable à:

drapeau	1
itération en cours	100

Mois	**Demande**	**Stock Initial**	**Stock Final**	**Rupture**	**%rupture**
1	403	450	47	0	0.00%
2	459	497	38	0	0.00%
11	402	588	186	0	0.00%
12	500	636	136	0	0.00%

somme des ruptures	5579
moyenne des ruptures	55.79

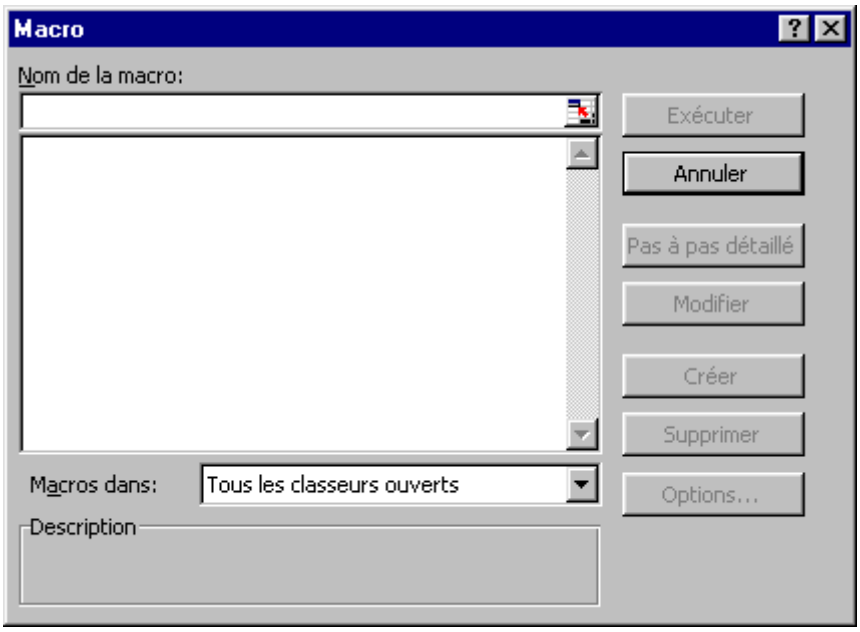
Remarque importante : lors de l'utilisation d'itération dans Excel il faut faire très attention à l'ordre de recalcul de la feuille, de façon à ce que les cellules soient bien mises à jour avec les nouvelle valeurs de chaque itération. Ceci rend délicat l'utilisation de cette méthode si l'on ne maîtrise pas bien l'ordre de recalcul des cellules.

Utilisation d'une macro

Tout d'abord il nous faut créer une feuille macro, pour cela nous passons dans le menu **Macros..**. du bandeau de l'onglet **Développeur**.

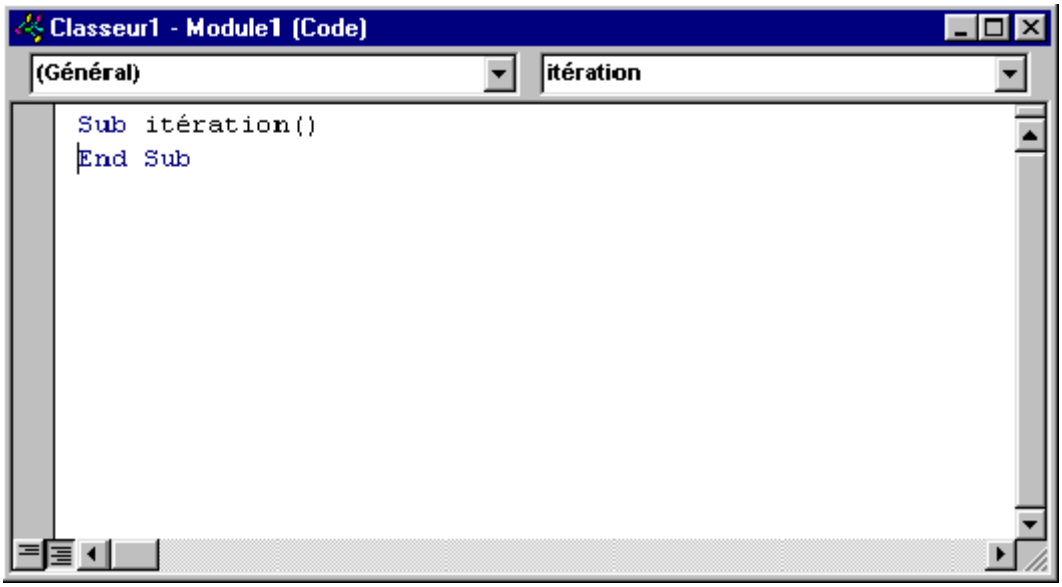
*Remarque : si l'onglet **Développeur** n'apparaît pas, utiliser le bouton Office, Options Excel Standard pour l'afficher.pour l'afficher.*

Nous obtenons alors une boite de dialogue :



Après avoir tapé un nouveau nom de macro le bouton Créer est actif, il suffit de cliquer sur ce bouton pour se retrouver dans l'environnement de Visual Basic (VB) adapté à Excel.

L'utilisateur tape alors le corps de la procédure (Subroutine) là où se trouve le curseur :



Les instructions suivantes mettent dans une cellule nommée **mamoyenne** la moyenne des ruptures de stocks obtenue pour un nombre d'itérations placé dans la cellule nommée **iter**. La somme des ruptures d'une simulation annuelle est stockée dans la cellule nommée **rupture :**

Mois	**Demande**	**Stock Initial**	**Stock Final**	**Rupture**	**%rupture**
1	473	450	0	23	4.86%
…	….	…	…	…	….
11	446	450	4	0	0.00%
12	497	454	0	43	8.65%
				196	← rupture

```
Sub itération()
REM TOTAL EST UNE VARIABLE LOCALE CONTENANT LA SOMME DES RUPTURES
Dim total As Long
total = 0
Application.Calculation = xlCalculationManual
For i = 1 To Range("iter").Value
   Application.Calculate
   total = total + Range("rupture").Value
Next i
Range("mamoyenne").Value = total / Range("iter").Value
Application.Calculation = xlCalculationAutomatic
End Sub
```

Quelques remarques sur ce programme. Les instructions commençant par Rem sont des commentaires non exécutés. Le langage est un langage "objet", ici les objets que nous manipulons sont des zones de cellules.

Range("iter") désigne la zone de cellules ayant pour nom iter. Dans notre exemple cette zone ne contient qu'une seule cellule, nous pouvons alors avoir accès à sa valeur par la propriété Value (propriété en lecture, écriture).

Remarque : Si l'on voulait conserver les résultats de toutes les années simulées pour obtenir différentes statistiques, il suffirait par exemple de définir une zone suffisamment grande nommée résultat :

Itération	Rupture Annuelle
1	112
	← Zone résultat
100	16

et d'utiliser la procédure suivante :

```
Sub iteration2()
Rem conserve dans résultat toutes les ruptures
Application.Calculation = xlCalculationManual
For i = 1 To Range("iter").Value
   Application.Calculate
   Range("résultat").Cells(i, 1).Value = i
   Range("résultat").Cells(i, 2).Value = Range("rupture").Value
Next i
Application.Calculation = xlCalculationAutomatic
End Sub
```

Ici Range("résultat") est une zone contenant deux colonnes et plusieurs lignes pour accéder à une cellule particulière, on utilise la propriété Cells(i,j) qui désigne la cellule se trouvant à la $i^{ème}$ ligne et $j^{ème}$ colonne à partir du coin supérieur gauche de la zone.

Il est aussi possible, après avoir calculé certaines caractéristiques de l'échantillon obtenu (la moyenne par exemple) précédemment, d'écrire une macro permettant de tester différents niveaux de commande. La cellule contenant la moyenne est appelle *mamoyenne*, comme dans le premier cas. En pratique il serait judicieux de garder aussi un indicateur sur le stock moyen, car en augmentant le niveau de commandes on diminue les ruptures mais on gonfle les

stocks !! Sans détailler les instructions, nous donnons ici la procédure permettant d'obtenir ce résultat, il est laissé au lecteur le soin de modifier la procédure pour stocker aussi le niveau moyen de stocks :

```
Sub compare()
    Const commande_min = 550, commande_max = 850, pas = 50
    Rem initialisation de la commande
    Range("Commande").Value = commande_min
    For i = 1 To (commande_max - commande_min) / pas + 1
        Rem on appelle l'iteration
        itération
        Rem On stocke les resultats
        Range("Titre").Cells(1, i) = Range("Commande")
        Range("Rupmoy").Cells(1, i) = Range("mamoyenne")
        Rem on peut se passer de préciser .valeur
        Rem augmenter le niveau de commande
        Range("Commande") = Range("Commande") + pas
    Next i
End Sub
```

On obtient alors les résultats suivants pour 1000 itérations :

	A	B	C	D	E	F	G	H
1	Nbre itérations	1000						
2								
3		Niveau de Stocks						
4		***550***	***600***	***650***	***700***	***750***	***800***	***850***
5	Rupture Moyenne	1 816,03	1 281,09	760,46	402,59	169,78	73,58	32,83
6	Stock Moyen Mensuel	302,04	361,61	450,80	594,62	800,69	1 065,46	1 356,80

Conclusion

Il est assez simple avec Excel de faire de la simulation probabiliste, la plupart du temps l'utilisation des tables est très suffisante, pour les modèles plus important en taille et où les recalculs sont longs, les itérations peuvent être utilisées, si l'on ne veut pas « programmer ».

Les macros offrent bien sur plus de souplesse et, pour qui veut bien investir dans le langage de programmation, permet de construire des modèles plus professionnels.

Signalons enfin qu'il existe aussi des addins permettant de réaliser des simulations sans toujours bien comprendre ce qui est fait, ces addins permettent le tirage au hasard et les itérations sans que l'utilisateur n'interviennent autrement que par un choix de menu.

35 Société Métallurgique et Minière.

La société métallurgique et minière (SMM) a créé, en 1970, une usine sidérurgique dans un port de l'Ouest de la France. L'installation portuaire de cette usine comporte un quai pouvant recevoir en même temps deux bateaux minéraliers de 10.000 tonnes environ. Les équipements du quai ont été conçus pour que chaque minéralier puisse être déchargé dans la journée.

Les besoins actuels de l'usine en minerai sont de 2.500.000 tonnes/an. Cependant, des accords avec des partenaires européens, ont conduit la SMM à prévoir le doublement de la capacité de l'usine d'ici 1993. Des contacts ont déjà été pris avec les fournisseurs de minerai de façon à pouvoir approvisionner l'usine à cette date.

Le contrat qui lie la SMM et les armateurs des minéraliers ne pourra être modifié : la SMM s'est engagée à décharger le bateau dans les 24 heures suivant son arrivée. En cas de retard, la SMM doit payer une indemnité de 7000F par jour d'attente et par bateau.

Les installations portuaires de la SMM peuvent être utilisées 24 heures sur 24, 7 jours sur 7.

Devant le doublement de la capacité de l'usine et donc du nombre de bateaux à décharger, la société SMM craint de voir augmenter dramatiquement les pénalités qu'elle aura à payer aux armateurs, elle a donc demandé à son service des Etudes de proposer des solutions pour augmenter la capacité d'accueil des navires.

Deux solutions ont été proposées :

1) L'agrandissement du quai actuel, qui porterait la capacité journalière de déchargement à 3 bateaux. Le coût de cette solution est de 3.000.000 F. Pour que cet investissement, d'après les normes en vigueur à la SMM, soit considéré comme rentable, il doit permettre d'économiser 500.000F de pénalités par an.

2) Le doublement du quai, ce qui porterait la capacité journalière de déchargement à 4 bateaux. Le coût de cette solution est de 7.500.000 F. Pour que cet investissement, d'après les normes en vigueur à la SMM, soit considéré comme rentable, il doit permettre d'économiser 1.250.000F de pénalités par an.

Les études statistiques réalisées par le passé ont montré que les arrivées journalières des bateaux étaient pratiquement poissonniennes. Le service des études pense que cette adéquation persistera dans le futur.

La société SMM vous demande de l'aider dans sa prise de décision.

Annexe : Probabilités poissonniennes

Moyenne	1,370
0 bateau / jour	0,26
1 bateau / jour	0,35
2 bateaux / jour	0,23
3 bateaux / jour	0,11
4 bateaux / jour	0,04
5 bateaux / jour	0,01

Remarque : le fichier SocMetalMin.xls contient une solution

36 Analyse du travail d'un pompiste.

Vous êtes chargé, par le service Méthodes d'une grande Compagnie Pétrolière, d'analyser le travail du pompiste de la Station Service de cette compagnie située sur la Nationale 20, à Salbris, Loir et Cher.

Dans le cadre d'une expérience, cette station service a embauché un jeune pour servir ses clients. Les clients ne peuvent pas se servir seuls.

Le Gérant de la Station Service trouve que les clients attendent trop longtemps pour être servis, et que cela lui fait perdre des clients. Il souhaite donc embaucher un second pompiste.

Le service Méthodes à envoyé sur place des agents chargés d'analyser le flux de clients ainsi que le temps mis par le pompiste pour les servir.

Après de très nombreux chronométrages, il a pu être établi que le temps séparant l'arrivée de deux clients suit une loi de poisson de moyenne 4 minutes.

Le temps de service est uniformément distribué, mais entre une et sept minutes.

Dans un premier temps, vous pouvez élaborer un tableau comme celui donné ci-dessous (d'autres méthodes sont, bien entendu, possibles).

Arrivées	Temps Service	Chronologie	Début service	Fin service	Attente client	Attente Pompiste
5	6	5	5	11	6	5
2	3	7	11	14	7	0
8	5	15	15	20	5	1
3	3	18	20	23	5	0
3	7	21	23	30	9	0
2	5	23	30	35	12	0
4	5	27	35	40	13	0
4	2	31	40	42	11	0
4	6	35	42	48	13	0
4	7	39	48	55	16	0
2	3	41	55	58	17	0
1	3	42	58	61	19	0

En considérant que le pompiste travaille 540 minutes par jour (sur quatre jours, les trois autres jours la station fonctionne en automatique avec la carte bancaire), calculer le temps total d'attente des clients sur une journée, le temps moyen d'attente par client, puis le temps où le pompiste, lui, attend un client à servir.

Sans élément économique supplémentaire, pouvez-vous donner raison ou tort au Gérant de la Station Service?

Quels sont les éléments économiques dont vous auriez besoin pour aller plus loin?

Comment pourriez-vous introduire un second pompiste dans le modèle, sachant que la station service est équipée de plus de deux pompes pour chaque type de carburant (Sans Plomb 98, Super Plombé 97, et Gazole)?

Remarque : début de solution dans le fichier Pompiste.xls

37 Gestion d'un Cabinet Dentaire.

Un dentiste vous a chargé d'analyser sa procédure de prises de rendez-vous. Il trouve, en effet, que la procédure actuelle conduit à des attentes qui peuvent être insupportables pour certains de ses clients, et il a peur, à terme de perdre une partie de sa clientèle.

Pour le moment, ce Dentiste ne travaille que sur rendez-vous et sa clientèle est suffisamment nombreuse pour qu'il n'y ait pas de "trous" dans son emploi du temps.

Le premier rendez-vous est à 8 heures trente. Ensuite, la Secrétaire médicale programme un rendez-vous de demi-heure en demi-heure jusqu'à midi. Après une pause pour le déjeuner, de nouveaux rendez-vous sont planifiés, sur le même rythme de 13 heures trente à 17 heures.

Vous devez tenir compte de deux sources d'incertitude :

Tout d'abord, les patients n'arrivent pas toujours à l'heure exacte de leur rendez-vous. Ensuite, le temps indispensable pour soigner chaque patient varie en fonction de l'importance du problème dentaire à résoudre.

L'analyse de la clientèle de ce Dentiste vous a fourni les informations suivantes :

En ce qui concerne les arrivées des patients :

Arrivées	Fréquences
15 mn avant	10%
10 mn avant	15%
5 mn avant	20%
A l'heure	20%
5 mn après	20%
10 mn après	10%
15 mn après	5%

Quant à la durée des soins, si 30 minutes est l'occurrence la plus fréquente, la distribution des fréquences est assez large :

Temps soins	Fréquences
15 minutes	5%
20 minutes	10%
25 minutes	15%
30 minutes	30%
35 minutes	15%
40 minutes	10%
45 minutes	10%
60 minutes	5%

Vous devez simuler, pendant une journée, l'arrivée des patients et le travail du Dentiste.

Déterminer le temps moyen d'attente par client, le temps d'attente du dentiste le temps dont il dispose pour le déjeuner.

Quel temps séparant deux prises de rendez-vous préconisez-vous?

38 Gestion des stocks.

Vous êtes chargé, par la Direction d'un magasin textile, à l'enseigne Centmill, d'analyser la politique actuelle de Gestion de stocks et d'Approvisionnement du magasin situé Boulevard Saint Michel, dans le cinquième arrondissement de Paris.

La méthode est toujours la même, et consiste, pour une référence particulière de chemises, à commander, dès que le stock passe en dessous de 200 chemises, la quantité nécessaire pour revenir à un stock de 400 chemises de cette référence[1]. Si par exemple, en fin d'une certaine semaine, le stock final est de 34 chemises, la commande sera de 66 chemises de la référence étudiée.

Il peut alors arriver que le magasin subisse une rupture de stocks. Dans ce cas, la commande est la commande habituelle plus la quantité de ventes manquées.

Si une ou plusieurs nouvelles ruptures de stocks se produisent, en suivant la première, la commande sera, dans ce cas d'un montant égal aux ruptures (les 100 chemises du stock de départ ayant déjà été commandées).

L'analyse des ventes des deux dernières années, pour une catégorie de chemises a donné les résultats suivants :

Quantités	Probabilités
25	5%
30	10%
40	20%
50	25%
60	25%
70	10%
80	5%

Les délais de livraisons, indépendants des quantités achetées, sont donnés dans le tableau ci-dessous :

Semaines	Probabilités
1	10%
2	30%
3	20%
4	30%
5	10%

Sachant que le Directeur du magasin estime que le coût de stockage d'une chemise en stock en début de semaine est de 1 Franc, que le coût d'une rupture de stock est estimé à 25 Francs (coût d'opportunité), et qu'enfin que le coût d'une commande est de 500 Francs, quelle que soit la quantité commandée, pouvez-vous calculer le coût moyen de la politique actuelle calculée sur 52 semaines?

Pouvez-vous proposer une meilleure politique?

Remarque : début de solution dans le fichier Gestocks.xls

[1]Pour ne pas compliquer le problème, nous ne tenons pas compte de la répartition des tailles à l'intérieur d'une référence.

39 Gestion des approvisionnements.

Un grossiste d'appareils électroménagers situé dans la Région Parisienne, commande certains de ses produits en Corée du Sud.

L'un de ces produits (four à micro ondes) présente une demande relativement stable. C'est pourquoi le Directeur de cette Société, en tenant également compte de l'éloignement de son fournisseur a pris l'habitude de commander chaque mois une quantité FIXE d'appareils.

La demande non saisonnière peut être considérée comme suivant une loi normale de moyenne 450 et d'écart type 50 unités. Bien évidemment ce n'est qu'une approximation, puisque les valeurs de la demande doivent être entières et positives : en fait il n'a jamais été constaté de ventes inférieures à 300 unités par mois, ni supérieures à 650.

En fonction des éléments statistiques précédents, le Directeur a décidé de commander 450 appareils par mois.

Pouvez-vous expliquer pourquoi?

Le neveu de notre importateur, après un stage dans l'entreprise en question déclare "la politique d'approvisionnement n'est pas optimale, et donc tu perds de l'argent en commandant 450 appareils par mois".

Sommé de s'expliquer, il répond :"C'est évident, la fonction de profit n'est pas symétrique".

Sachant que la marge nette par produit vendu est de 250 Francs, que le coût de stockage mensuel est de 50 Francs par unité, et que le coût d'opportunité d'une vente manquée est de 500 Francs, pouvez-vous confirmer ou infirmer les déclarations du neveu?

Presque convaincu, l'oncle demande alors à son neveu de lui proposer "la politique optimale".

Pouvez-vous aider le neveu dans la formulation de sa réponse? Sachant que, comme récompense il peut demander à son oncle 10% du gain prouvé, combien peut-il demander à son oncle?

40 Gestion de location de camions

L'agence ADA, de location de véhicules située à Velizy, envisage de diversifier ses produits et de louer des utilitaires (petits camions qui peuvent se conduire avec le permis "voiture").

Le directeur de cette agence vous a demandé d'analyser les chiffres actuellement disponibles (l'agence de Velizy sous traite actuellement ce type de location à l'agence de Versailles) et de lui proposer le nombre "optimum" de camions à mettre dans son parc.

Les statistiques de la demande locale sont résumées dans le tableau suivant :

Nb camions	Probabilités
0	0.20
1	0.20
2	0.30
3	0.15
4	0.15

En ce qui concerne les durées de locations, les chiffres sont basés sur l'ensemble des agences Ile de France et sont les suivants :

Jours de location	Probabilités
1	0.35
2	0.30
3	0.20
4	0.15

Enfin les données économiques sont les suivantes :

Profit net par jour de location par camion : 250 F

Coût d'opportunité d'une location manquée : 300 F

Coût journalier d'inutilisation d'un camion : 50 F

Construire un modèle de simulations permettant au directeur de l'agence de déterminer le nombre de camions à mettre en service.

41 La boucherie Netprix

Une supérette de la chaîne NetPrix vient de rénover son magasin et a modernisé le rayon boucherie. Les deux personnes qui servent à ce rayon se plaignent de leur charge de travail et du fait qu'ils doivent fréquemment faire des heures supplémentaires pour servir les derniers clients de la journée.

Ils ont demandé au responsable du magasin d'être aidés par 2 apprentis. Ces apprentis seraient payés 700€ par mois charges comprises, alors que les professionnels sont payés 2500€, les heures supplémentaires étant payées 25€ l'heure.

Une étude a montré que si les clients attendaient à un rayon, ils prenaient moins de temps pour faire leurs achats et que la perte de chiffre d'affaires occasionnée était d'à peu près 3€ par minute d'attente.

Pour répondre à ses employés le directeur demande une étude sur les temps d'arrivée et de service du rayon boucherie.

Les temps séparant deux arrivées ont été enregistrés à la minute près, c'est à dire que si le temps était inférieur à 1 minute on codait 0, entre 1 et 3 minutes on codait 2 etc..

Les temps de services ont été arrondis à la minute. Les résultats vous sont donnés dans l'annexe.

Au vu de ces résultats, le directeur calcula les moyennes et obtint :

moyenne des temps de service = 16,28 minutes

moyenne des temps séparant deux arrivées = 16,48 minutes

Il convoqua alors les deux bouchers et leur expliqua qu'en fait, c'est plutôt la suppression d'un poste qu'il serait raisonnable d'envisager, puisqu'une seule personne semblait en moyenne suffisante pour satisfaire pratiquement sans attente la clientèle.

Les bouchers ne comprirent pas grand chose aux explications du directeur, mais lui affirmèrent que leur expérience montrait qu'il se constituait des files d'attente importante et que pour s'en convaincre il suffisait de regarder l'état des heures supplémentaires. Il indiquèrent même qu'encas de suppression d'emploi de l'un d'entre eux, on courrait à l'émeute!

Très perplexe le directeur vous demande une étude.

Annexe : Résultat de l'étude des temps entre deux arrivée et des temps de service

Entre deux arrivées	
Temps	Probabilité
0	0,17
2	0,08
4	0,07
6	0,06
8	0,06
10	0,05
12	0,05
14	0,04
16	0,04
18	0,03
20	0,03
22	0,03
24	0,03
26	0,02
28	0,02
30	0,02
32	0,02
34	0,02
36	0,02
38	0,02
40	0,02
42	0,01
44	0,01
46	0,01
48	0,01
50	0,01
52	0,01
54	0,01
56	0,01
58	0,01
60	0,01

Temps de service	
Service	Probabilité
6	0,01
7	0,01
8	0,02
9	0,02
10	0,02
11	0,03
12	0,04
13	0,06
14	0,08
15	0,09
16	0,11
17	0,12
18	0,11
19	0,09
20	0,07
21	0,05
22	0,03
23	0,02
24	0,01
25	0,01

Eléments de Statistique

STATISTIQUES DESCRIPTIVES

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

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.

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. Sous Excel on utilisera une feuille pour ce tableau en indiquant souvent le nom des variables dans la première ligne et éventuellement le numéro de l'individu dans la première colonne :

	A	B	C
1	**Individu**	**Kms**	**Revision**
2	1	25500	Oui
3	2	25700	Oui
4	3	21700	Non
5	4	27300	Oui
6	5	29900	Oui
7	6	21600	Oui
8	7	20200	Oui
9	8	14800	Oui
10	9	19800	Oui
11	10	29800	Oui
12	11	22500	Oui

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).

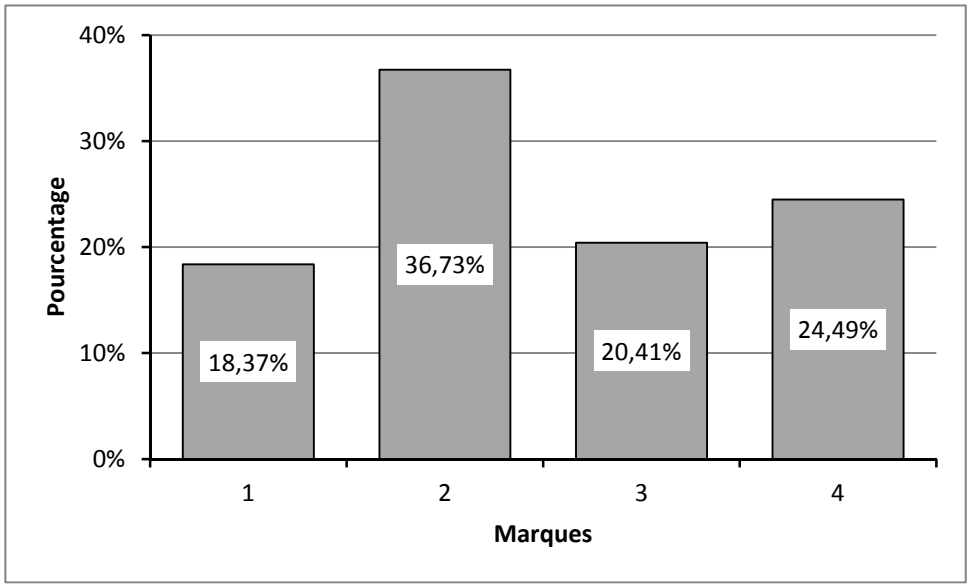
Voici un exemple de résumé fourni pour la variable qualitative Marque du fichier Pfrais.xls :

Formules		Valeurs			
		MARQUES			
	Effectifs		Effectifs	Pourcentage	Pourcentage cumulé
1	=NB.SI(Pfrais!$E$2:$E$50;Feuil1!B3)	Marque 1	9	18,37%	18,37%
2	=NB.SI(Pfrais!$E$2:$E$50;Feuil1!B4)	Marque 2	18	36,73%	55,10%
3	=NB.SI(Pfrais!$E$2:$E$50;Feuil1!B5)	Marque 3	10	20,41%	75,51%
4	=NB.SI(Pfrais!$E$2:$E$50;Feuil1!B6)	Marque 4	12	24,49%	100,00%
	=SOMME(C3:C6)	Total	49	100,00%	

MARQUE

	Effectifs	Pourcentage	Pourcentage cumulé
Marque 1	9	18,37%	18,37%
Marque 2	18	36,73%	55,10%
Marque 3	10	20,41%	75,51%
Marque 4	12	24,49%	100,00%
Total	49	100,00%	

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 le pourcentage 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 $\boldsymbol{N}$ la taille de la population ou sous population et $\boldsymbol{X}$ 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.

Voici un exemple (fichier Forfait.xls) de résumé calculé avec Excel :

Statistiques

Km

N		42	=NBVAL(forfaits!B2:B43)
Moyenne		128,1	=MOYENNE(forfaits!B2:B43)
Médiane		120	=MEDIANE(forfaits!B2:B43)
Ecart-type		54,13	=RACINE(E8) (racine de la variance)
Variance		2930,49	=VAR(forfaits!B2:B43)
Intervalle		233	=E11-E10 (Maximum-Minimum)
Minimum		32	=MIN(forfaits!B2:B43)
Maximum		265	=MAX(forfaits!B2:B43)
Centiles	25	90	=CENTILE(forfaits!$B$2:$B$43;0,25)
	50	120	=CENTILE(forfaits!$B$2:$B$43;0,5)
	75	150	=CENTILE(forfaits!$B$2:$B$43;0,75)

Remarque :

- *en lieu et place de la fonction centile, il est possible d'utiliser la fonction quartile, dont le dernier paramètre est le numéro du quartile.*
- *La fonction VAR de Excel renvoie la variance estimée d'un echantillon, ce qui est le cas ici, et non la variance de la population (voir le chapitre sur l'estimation). Il existe une fonction VARP qui renvoie la variance de la population.*

Les représentations associées aux variables quantitatives 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, il est d'usage d'utiliser les boites à moustaches (Box Plot) et les diagrammes Q-Q (Q-Q Plot).

Réalisation d'histogrammes sous Excel

Il n'existe pas de règles permettant de fixer le nombre de classes utilisées dans un histogramme. Si ce nombre est trop faible, l'"allure" de la loi sous-jacente est gommée, s'il est trop grand, très souvent le graphique sera incohérent. Les logiciels statistiques utilisent très souvent $c = \sqrt{n}$ classes, Sturges suggère que le nombre maximum de classe est $1 + log_2 n$.

Pour obtenir les effectifs des classes, il faut créer un tableau à 2 colonnes, dans la première colonne on indiquera les bornes sipérieures des classes et dans la seconde, on utilisera la fonction matricielle FREQUENCE.

Pour ne pas être gêné par une erreur qui éliminerait les observations correspondant au maximum, nous prendrons comme intervalle l'arrondi supérieur à 3 ou 4 décimales.

Pour entrer une formule matricielle, rappelons que l'utilisateur doit sélectionner la zone dans laquelle cette formule est entrée, puis ensuite valider la formule avec la combinaison de touches Ctrl-Majusule-Entrée. Les paramètres de la fonction FREQUENCE sont :

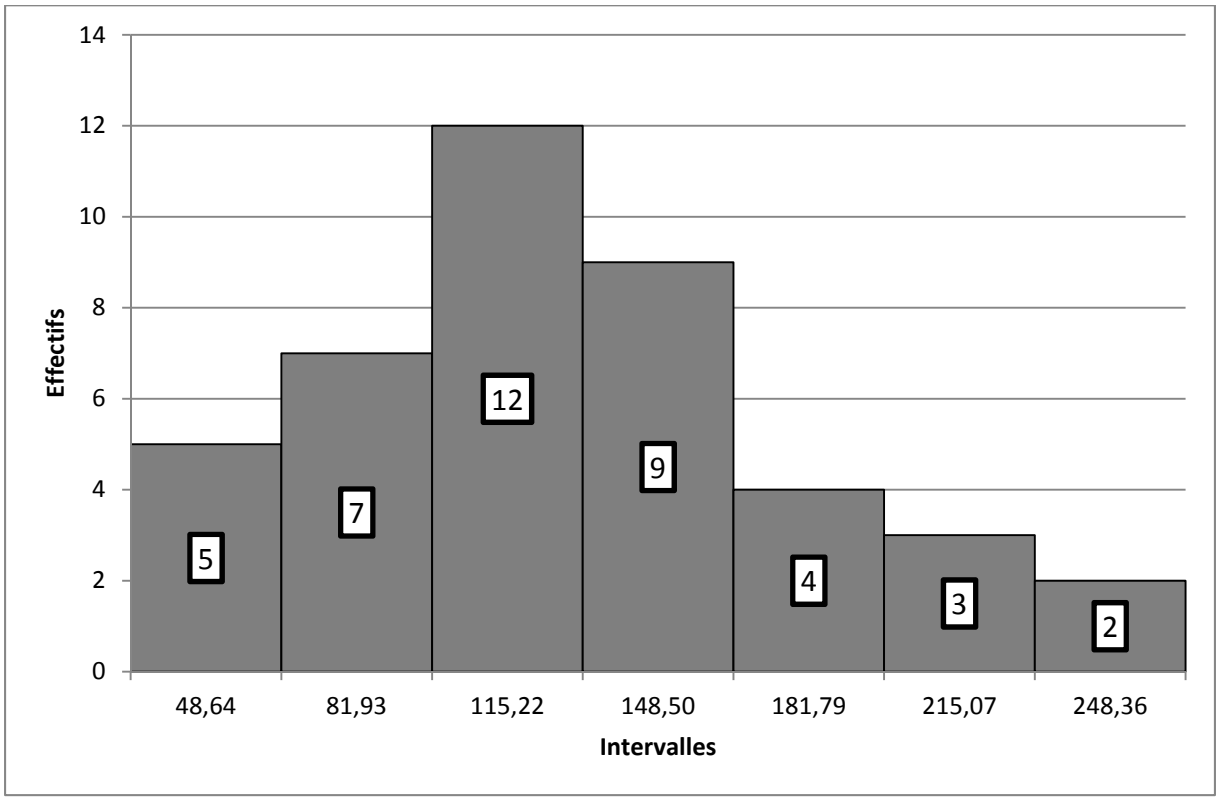
1. La zone de données
2. La zone des bornes supérieures des intervalles

On obtient alors le tableau suivant (pour le fichier Forfaits) :

B4 {=FREQUENCE(forfaits!B:B;Histogramme!A3:A9)}

	A	B	C	D	E	F	G	H
1	**Bornes**	**Effectifs**	**Milieu**		**Calculs intermédiaires**			
2	32,00							
3	65,29	5	48,64		Nombrebre de classes		7	
4	98,57	7	81,93		Largeur intervalle		33,2860	
5	131,86	12	115,22		Minimum		32,00	
6	165,14	9	148,50		Maximum		265,00	
7	198,43	4	181,79					
8	231,72	3	215,07					
9	265,00	2	248,36					

Le graphique associé est obtenu en insérant un histogramme, dont la présentation va être modifiée de façon à satisfaire à l'usage, qui veut que pour une variable quantitative les blocs soient collés pour bien souligner l'aspect continu de la variable. D'où le graphique suivant :



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 dans la plupart des logiciels statistiques limitées à 1,5 fois la distance interquartile. Réaliser de telles boites à moustaches sous Excel demande soit de programmer soit d'utiliser les commandes de base de données pour

extraire les outliers, nous nous limiterons ici à ajouter aux moustaches le minimum et le maximum de la série (qui apparaitront soit extérieurs aux moustaches soit à la limite de celle-ci).

La réalisation de la boîte à moustache se fait en deux étapes :

- Création de la zone des données
- Création du graphique en "détournant" un histogramme empilé.

Création de la zone de données

Les éléments dont nous avons besoin pour créer le graphique sont :

1. Pour la boîte :
    a. Le bas de la boite qui correspond au premier quartile. (ce bas sera rendu transparent)
    b. La hauteur du fond de la boite qui correspond à la différence entre la médiane et le premier quartile.
    c. La hauteur du couvercle de la boîte qui correspond à la différence entre le troisième quartile et la médiane.
2. La longueur des deux moustaches, éventuellement limitées à une fois et demi l'intervalle interquartile.
3. Les outliers éventuellement, ici seul le max et le min

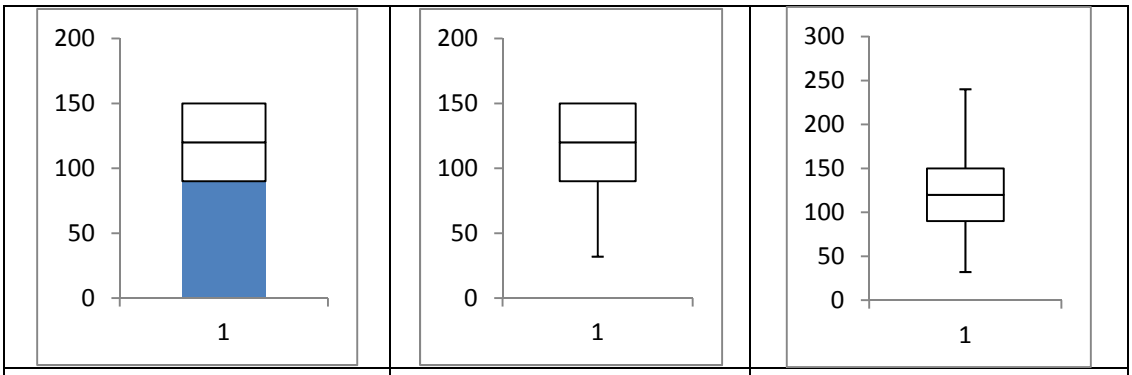
	A	B	C	D	E	F	G
1	**Données Statistiques**			**Pour le graphique**			
2	Min	32		H1	90	1er Quartile	
3	Q1	90		H2	30	Hauteur Boite Bas	B4-B3
4	Médiane	120		H3	30	Hauteur Boite Haut	B5-B4
5	Q3	150		MoustacheBas	58	MIN(1,5*B5-B3;B3-B2)	
6	Max	265		MoustacheHaut	90	MIN(1,5*(B5-B3);B6-B5)	
7							

Création du graphique :

- Etape 1 : Création de la boîte
  Sélectionner les trois premières donnée H1, H2, H3 et insérer un graphique en histogramme empilé (inverser éventuellement les ligne et les colonnes pour obtenir le graphique etape 1 ci-dessous)
  , la partie la plus basse sera ensuite rendue transparente, mais il faut d'abord ajouter la moustache inférieure. Sélectionner un remplissage (blanc par exemple) pour les blocs 2 et 3 et un contour automatique par exemple.
- Etape 2 : Création de la moustache inférieure :
  Sélectionner le bloc inférieur, de l'histogramme et dans le bandeau "Disposition" de

"Outils de graphique", choisir Barres d'erreur, Autres options de barres d'erreur : Dans barres d'erreur verticales, choisir Orientation Moins, style d'arrivée Maj et dans marge d'erreur, choisir Personnalisé et donner comme valeur négative la valeur de la cellule correspondant à la moustache du bas. Choisir enfin dans la mise en forme pour le bloc du bas aucun remplissage. Vous devez alors obtenir le graphique etape 2 ci-dessous.

- Etape 3 : Création de la moustache supérieure
  Sélectionner le bloc supérieur, de l'histogramme et dans le bandeau "Disposition" de "Outils de graphique", choisir Barres d'erreur, Autres options de barres d'erreur : Dans barres d'erreur verticales, choisir Orientation Plus, style d'arrivée Maj et dans marge d'erreur, choisir Personnalisé et donner comme valeur positive la valeur de la cellule correspondant à la moustache du haut. Vous devez alors obtenir le graphique Etape 3 ci-dessous.



Etape 1 | Etape 2 | Etape 3

Il est possible ensuite d'ajouter des outliers comme nouvelles séries associées à un axe secondaire des ordonnées. Pour obtenir par exemple le graphique final suivant, où seuls sont représentés le minimum et le maximum :

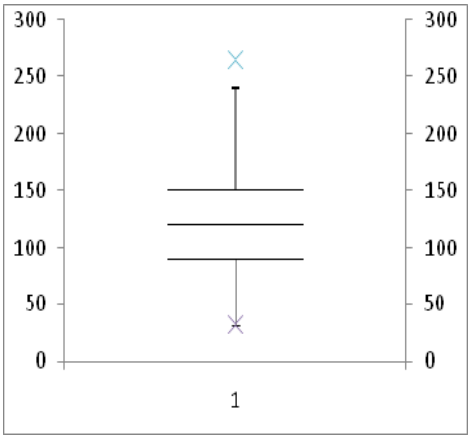


Diagramme Q-Q

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 $\frac{N-0,5}{N}$. Pour ne pas modifier les données par une opération de tri, on utilisera la fonction PETITE.VALEUR(serie,p) qui retourne la pième valeur d'une série d'observations.

Sur l'exemple Forfaits, on obtient le tableau suivant :

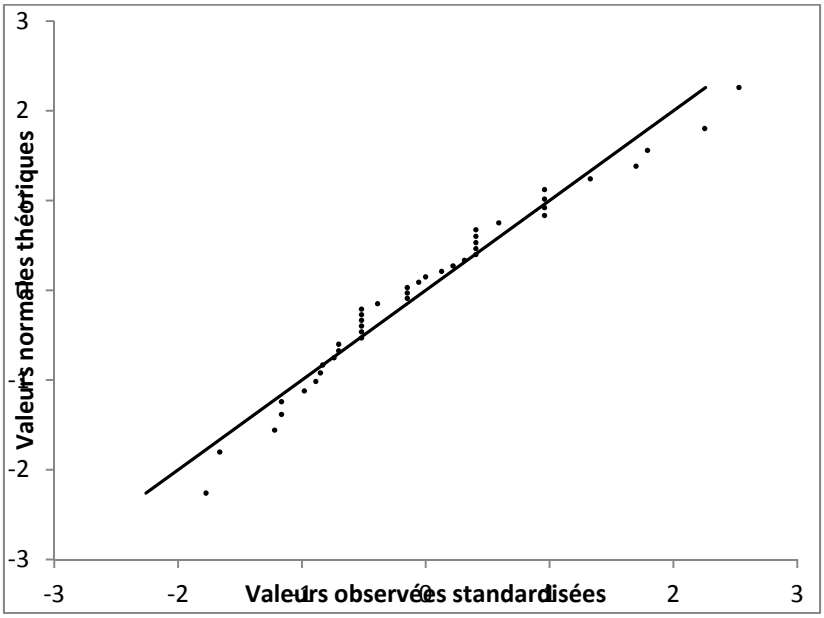
B3 =PETITE.VALEUR(forfaits!$B$2:$B$43;QQ!A3)

	A	B	C	D	E	F	G
1	Observation	Valeur	Standard	Normale		Calculs intermédiaires	
2	1	32	-1,77513994	-2,26018899			
3	2	38	-1,66430365	-1,80274309		Moyenne	128,10
4	3	62	-1,22095849	-1,55878355		Ecart-type	54,13
5	4	65	-1,16554035	-1,38299413		Nbre de données	42

La formule en C3 est : =(B3-$G$3)/$G$4

La formule en D3 est : =LOI.NORMALE.STANDARD.INVERSE((A3-0,5)/$G$5)

Enfin le graphique obtenu :



L'ajustement est correct, bien que l'on retrouve les valeurs extrêmes en queue de distribution.

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) :

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

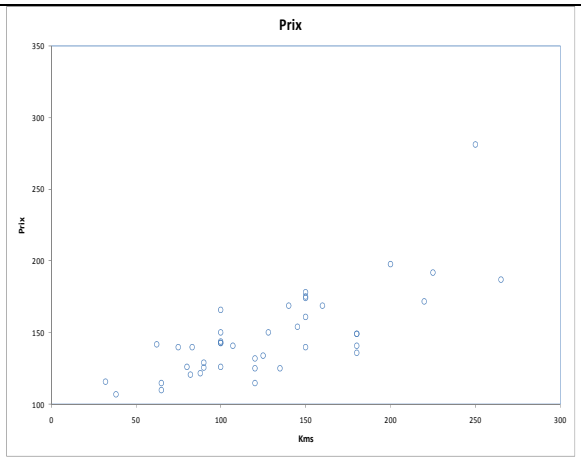
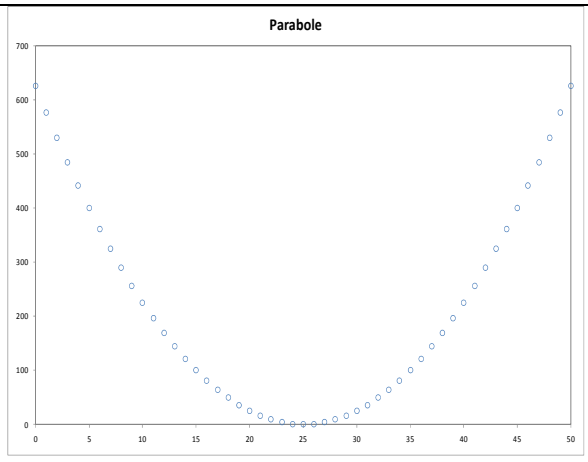
Cette corrélation est toujours comprise entre **-1 et 1**. La liaison entre les variables est d'autant plus forte que la valeur absolue est proche de 1.

Dans Excel, la covariance est données par la fonction COVARIANCE(série1;série2) et le coefficient de corrélation par la fonction COEFFICIENT.CORRELATION(série1;série2).

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 pour les pays de l'Union Européenne, nous avons relevé 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

En utilisant les fonctions de base données, il est aisé d'obtenir un tableau tel que :

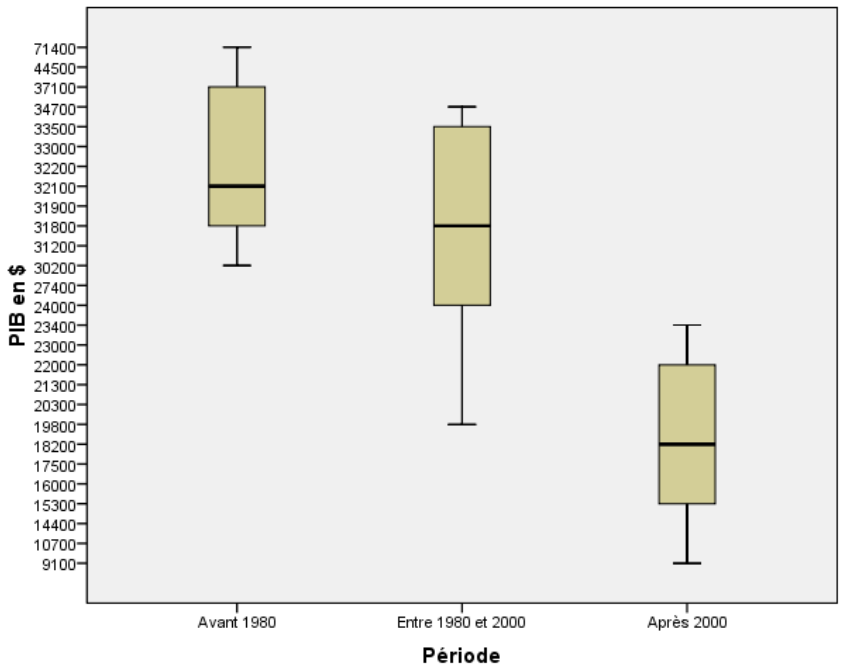
Période		
1	Moyenne	38555,56
	Minimum	30200,00
	Maximum	71400,00
	Ecart-type	12321,90
Période		
2	Moyenne	28600,00
	Minimum	19800,00
	Maximum	34700,00
	Ecart-type	5388,57
Période		
3	Moyenne	17600,00
	Minimum	9100,00
	Maximum	23400,00
	Ecart-type	4485,72

Dont les formules sont les suivantes :

	G	H	I
3	Période		
4	1	Moyenne	=BDMOYENNE($A$1:$D$28;"PIB";G3:G4)
5		Minimum	=BDMIN($A$1:$D$28;"PIB";G3:G4)
6		Maximum	=BDMAX($A$1:$D$28;"PIB";G3:G4)
7		Ecart-type	=BDECARTYPEP($A$1:$D$28;"PIB";G3:G4)
8			
9	Période		
10	2	Moyenne	=BDMOYENNE($A$1:$D$28;"PIB";G9:G10)
11		Minimum	=BDMIN($A$1:$D$28;"PIB";G9:G10)
12		Maximum	=BDMAX($A$1:$D$28;"PIB";G9:G10)
13		Ecart-type	=BDECARTYPEP($A$1:$D$28;"PIB";G9:G10)

La zone A1:D28 contient les données y compris les titres de colonne, la zone de critère est constituée de deux cellules, la première contient le nom du champ "Période" et l'autre la valeur de la période dont on veut les caractéristiques.

On peut constater que les moyennes des PIB sont de plus en plus faibles au cours du temps, on pourrait vérifier graphiquement cela en construisant des boîtes à moustaches. Pour réaliser ces boîtes, il faut extraire les enregistrements correspondant aux trois périodes, car Excel n'a pas de fonction BDmediane ou BDquartile. Après cette extraction, il suffit de procéder comme au paragraphe précédent pour construire les boîtes 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 $\boldsymbol{X}$ selon les modalité de la variable $\boldsymbol{Y}$ sont les mêmes quelque soit la modalité de $\boldsymbol{X}$ prise en compte (et bien sur réciproquement si les répartition de la variable $\boldsymbol{Y}$ selon les modalités de la variable $\boldsymbol{X}$ sont les mêmes quelque soit la modalité de $\boldsymbol{Y}$ 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 $\boldsymbol{f_{i,j}}$ la fréquence dans la population de la présence simultanée des modalités $\boldsymbol{i}$ **et** $\boldsymbol{j}$:

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

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					
			Nord	Est	Centre	Ouest	Sud	Total
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).

SONDAGE-ESTIMATION

1 Un Exemple (Fichier Martin.xls)

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.

1.1 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 :

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

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 $\mathbf{R}$

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

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.

1.2 Utilisation d'Excel.

Dans la feuille Clients, 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.

Cette feuille 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 pour les deux variables qui nous intéressent (mais Monsieur Martin lui ne les aura pas) :

Pour la variable kilométrage :

Moyenne = 25005 (fonction MOYENNE() d'Excel)

Ecart-type = 3978 ici la fonction EcartypeP d'Excel est utilisée et non pas la fonction Ecartype qui ne concerne que les échantillons (voir plus loin)

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 son 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).

La zone contenant les données a été nommée Données. 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 un format personnalisé affichant respectivement Oui ou Non.

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 (cf. exercice), 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.

2.1 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 :

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

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

2.2 Illustration de cette procédure avec Excel.

Nous allons travailler ici sur une nouvelle feuille, que nous nommerons Echantillon. Le modèle existe dans le classeur Martin2.xls, mais nous conseillons au lecteur de refaire lui-même le travail.

La taille de l'échantillon étant fixée dans une cellule nommée *Téchan* (nous préciserons plus loin où doit se trouver cette cellule), nous allons tout d'abord tirer l'échantillon individu. La taille de l'échantillon étant limitée à 500 au maximum.

- Construction de l'échantillon individu

  Sur une zone de 500 lignes, allant de A2 à A501, il suffit de recopier la formule suivante : =ENT(ALEA()*Taille)+1

  En effet ALEA() retourne un nombre (pseudo-) aléatoire compris entre 0 et 1 (1 non compris), cette formule donne donc une valeur entière comprise entre 1 et Taille.

  Pour contrôler le nombre de valeurs obtenues, qui doit être égal à la taille de l'échantillon, nous modifions la formule de la façon suivante :

  =SI(LIGNE()+1<=Téchan ; ENT(ALEA()*Taille)+1 ; '')

  ce qui ne provoquera le tirage aléatoire d'un numéro d'individu que si nous n'avons pas encore atteint le nombre voulu.

- Construction de l'échantillon image

  Sur les zones de 500 lignes allant respectivement de B2 à B501 et de C2 à C501, nous allons indiquer les réponses données aux questions des enquêteurs interrogeant l'individu tiré au hasard, c'est à dire les valeurs correspondant à la première colonne et à la deuxième colonne de la ligne tirée au hasard dans le tableau de données :

  Pour le kilométrage : =SI($A4="";"";INDEX(Données;$A4;1))

  Pour la révision : =SI($A4="";"";INDEX(Données;$A4;2))

- Extrait de la feuille Excel

	A	B	C
1	**Individu n°**	**KM**	**Révision**
2	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A2="";"";INDEX(Données;$A2;1))	=SI($A2="";"";INDEX(Données;$A2;2))
3	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A3="";"";INDEX(Données;$A3;1))	=SI($A3="";"";INDEX(Données;$A3;2))
4	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A4="";"";INDEX(Données;$A4;1))	=SI($A4="";"";INDEX(Données;$A4;2))
5	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A5="";"";INDEX(Données;$A5;1))	=SI($A5="";"";INDEX(Données;$A5;2))
6	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A6="";"";INDEX(Données;$A6;1))	=SI($A6="";"";INDEX(Données;$A6;2))
7	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A7="";"";INDEX(Données;$A7;1))	=SI($A7="";"";INDEX(Données;$A7;2))
8	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A8="";"";INDEX(Données;$A8;1))	=SI($A8="";"";INDEX(Données;$A8;2))
9	=SI(LIGNE()<=techan+1;ENT(Taille*ALEA()+1);"")	=SI($A9="";"";INDEX(Données;$A9;1))	=SI($A9="";"";INDEX(Données;$A9;2))

ou sans l'affichage formule :

	A	B	C
1	**Individu n°**	**KM**	**Révision**
2	14976	22800	1
3	26138	24000	1
4	19783	30500	0

Remarquons, bien sûr, que chaque fois que nous entrons une formule, de façon plus générale chaque fois qu'un recalcul est effectué, les valeurs prises par l'aléa changent, donc l'échantillon individu ainsi que l'échantillon image changent aussi, les valeurs que nous donnerons pour les paramètres recherchés vont donc dépendre de l'échantillon, c'est ce qui sera à l'origine de l'imprécision.

3 Estimation – Estimateur

3.1 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.

3.1.1 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 $\lim_{n\to\infty} Var(\Theta_n)=0$.

Une autre propriété, que nous signalerons simplement, est la consistance : c'est, par rapport aux autres estimateurs possibles d'un même paramètre, le fait d'avoir une dispersion plus faible, c'est à dire une variance inférieure.

3.2 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.

3.2.1 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 $\overline{X}_n$ est défini en fonction de l'échantillon aléatoire $\left(X_1,X_2,...,X_n\right)$ par :

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

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

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

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".

3.2.2 Utilisation d'Excel

Pour calculer l'estimation de la proportion ou de la moyenne du kilométrage, il nous suffira donc d'utiliser la fonction MOYENNE() d'Excel, avec comme argument la zone correspond aux valeurs observées sur l'échantillon respectivement pour le lieu de révision (codé 0 ou 1) et pour le kilométrage.

Mise évidence de la consistance de l'estimateur : pour cela nous allons construire un grand nombre d'échantillons ; il est impossible en effet pour de construire tous les échantillons, par exemple pour une taille de 100 il y a $(42540)^{100}$ échantillons individus différents (même si en fait pour le pourcentage il n'y a que 2^{100} échantillons image et 101 valeurs possibles mais avec des probabilités différentes, comme nous le verrons plus loin). Nous allons donc utiliser une table d'hypothèses à deux entrées, l'entrée en ligne sera liée à la cellule de la taille d'échantillon, l'entrée en colonne sera liée à une cellule vide de la feuille, puisque la fonction ALEA() que nous voulons recalculer ne dépend d'aucun paramètre.

Les entrées en lignes prendront par exemple les valeurs 100,200, 300, 400,500 et nous tirerons 1000 échantillons, donc les entrées en colonne prendront les valeurs de 1 à 1000 pour indiquer le numéro d'ordre de l'échantillon.

Nous obtenons ainsi les valeurs de l'estimation de la proportion (on pourrait faire de même avec la moyenne du kilométrage) pour 1000 échantillons de taille variant entre 100 et 500.

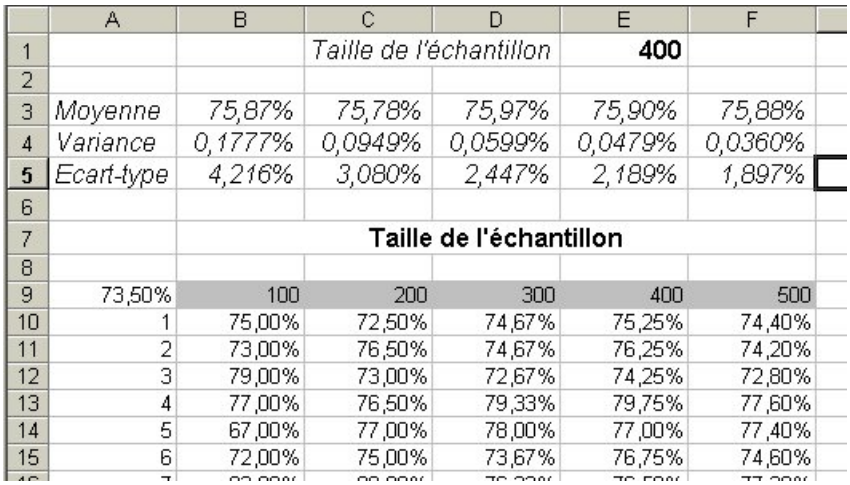
Ensuite, on calculera les caractéristiques de ces moyennes pour chaque taille d'échantillon : moyenne, variance, écart type.

Voici les formules utilisées pour cette construction :

	A	B	C
1			*Taille de l'échantillon*
2			
3	*Moyenne*	*=MOYENNE(B10:B1009)*	*=MOYENNE(C10:C1*
4	*Variance*	*=VAR(B10:B1009)*	*=VAR(C10:C1009)*
5	*Ecart-type*	*=ECARTYPE(B10:B1009)*	*=ECARTYPE(C10:C1*
6			
7			**Taille**
8			
9	=Echantillon!F7	100	200
10	1	=TABLE(E1;I8)	=TABLE(E1;I8)
11	=A10+1	=TABLE(E1;I8)	=TABLE(E1;I8)
12	=A11+1	=TABLE(E1;I8)	=TABLE(E1;I8)
13	=A12+1	=TABLE(E1;I8)	=TABLE(E1;I8)

Rappelons que les cellules d'entrée doivent être sur la même feuille que la table, la cellule E1 correspond à la cellule nommée *Techan* précédemment, la cellule I8 correspond à une cellule vide quelconque de la feuille, enfin la cellule Echantillon !F7 est la cellule contenant la valeur de l'estimation de la proportion dans la feuille de l'échantillon.

On obtient alors les résultats suivants :



	A	B	C	D	E	F
1			*Taille de l'échantillon*		**400**	
2						
3	*Moyenne*	*75,87%*	*75,78%*	*75,97%*	*75,90%*	*75,88%*
4	*Variance*	*0,1777%*	*0,0949%*	*0,0599%*	*0,0479%*	*0,0360%*
5	*Ecart-type*	*4,216%*	*3,080%*	*2,447%*	*2,189%*	*1,897%*
6						
7		**Taille de l'échantillon**				
8						
9	73,50%	100	200	300	400	500
10	1	75,00%	72,50%	74,67%	75,25%	74,40%
11	2	73,00%	76,50%	74,67%	76,25%	74,20%
12	3	79,00%	73,00%	72,67%	74,25%	72,80%
13	4	77,00%	76,50%	79,33%	79,75%	77,60%
14	5	67,00%	77,00%	78,00%	77,00%	77,40%
15	6	72,00%	75,00%	73,67%	76,75%	74,60%

On constate bien que l'estimateur de la moyenne est sans biais, la moyenne des estimations de la proportion est presque égale à la vraie valeur 75,87%.

Mais surtout la variance diminue de façon significative avec la taille de l'échantillon et on observe à peu près le ratio prévu : par rapport à un échantillon de taille 100, la variance des estimations pour un échantillon de taille 200 est à premier la moitié, celle pour un échantillon de taille 300 le tiers, etc.…

3.3 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 :

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

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

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.

3.3.1 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 :

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

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.

3.3.2 Utilisation d'Excel

Nous allons mettre en évidence, le biais de l'estimateur naturel de la variance et visualiser le bon estimateur grâce aux tables d'Excel. Pour que l'écart entre les deux estimateurs soit significatif, nous travaillerons sur des échantillons de petite taille (ici n=10).

Il existe sous Excel deux fonctions associées à la variance :

- La fonction VAR(Zone) ; qui retourne l'estimation de la variance(correspondant à l'estimateur S_n^2), considérant donc que la zone de données est un échantillon. L'estimation de l'écart type est alors la fonction ECARTYPE(Zone).
- La fonction VAR.P(Zone) ; qui retourne la variance des données, c'est à dire ce qui correspond à la variance de la population, de même l'écart type est alors donnée par la fonction ECARTYPEP(Zone).

Cependant même avec cette taille d'échantillon, il est hors de question de tirer tous les échantillons, nous allons tirer un grand nombre d'échantillon (1000 par exemple) et calculer pour chacun des échantillons la valeur des deux fonctions VAR et VARP d'Excel. Nous évaluerons ensuite la moyenne de ces fonctions sur l'échantillon et nous comparerons avec les valeurs calculées sur la population.

Ceci va se faire à l'aide de table à une entrée : une cellule vide, et deux colonnes de résultats. Voici les formules de la feuille de calcul, pour la variable Kilométrage :

	J	K	L
6		Pour un échantillon de taille 10	
7	Moyenne	=MOYENNE(K10:K1009)	=MOYENNE(L10:L1009)
8			
9	Echantillon N°	=VAR(Echantillon!B2:B501)	=VAR.P(Echantillon!B2:B501)
10	1	=TABLE(;I8)	=TABLE(;I8)
11	=J10+1	=TABLE(;I8)	=TABLE(;I8)

La table est dans la zone J9 :L1009, la ligne 7 sert à calculer les moyennes, la zone B2:B501 de la feuille Echantillon contient les valeurs du kilométrage de l'échantillon.

En utilisant des formats personnalisés pour les entêtes de colonne de la table on obtient les résultats suivants :

	J	K	L
6	Pour un échantillon de taille 10		
7	**Moyenne**	**15 934 215**	**14 340 793**
8			
9	**Echantillon N°**	**Fonction VAR**	**Fonction VAR.P**
10	1	16 820 000	15 138 000
11	2	19 140 000	17 226 000
12	3	21 942 222	19 748 000
13	4	13 177 778	11 860 000

La vraie valeur de la variance sur la population est de 15 825 792, la valeur moyenne obtenue avec VAR est très proche de cette valeur (moins de 1% d'erreur), tandis que la valeur obtenue

Avec VAR.P est très en dessous de la vraie valeur, on retrouve comme il était prévu une sous estimation de l'ordre de 10% (9,4%). Si on refait calculer plusieurs fois ces moyennes, on constate que ce n'est pas un résultat exceptionnel, mais que systématiquement la moyenne des variances de 1000 échantillons sous estime la variance de la population ; nos avons donc ainsi mis en évidence le biais calculé plus haut.

4 Estimation par intervalle, précision d'un sondage

Comme nous venons de le voir, les estimations obtenues pour un paramètre à partir d'un échantillon 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 faible 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.

4.1 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.

4.2 Cas du pourcentage

4.2.1 Loi de probabilité de $\overline{X}n$

La loi de X sur la population initiale est, comme nous l'avons vu (2.1), 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) :

$$\text{Pour tout } 0\leq k\leq n \text{ on a } \Pr\left(\overline{X}n=\frac{k}{n}\right)=C_n^k\,p^k(1-p)^{n-k}$$

	A	B
1	Taille de l'échantillon	10
2		
3	**Pourcentage Estimé**	**Probabilité**
4	=(LIGNE()-4)/$B$1	=LOI.BINOMIALE(A4*$B$1;$B$1;Population!$F$9;FAUX)
5	=(LIGNE()-4)/$B$1	=LOI.BINOMIALE(A5*$B$1;$B$1;Population!$F$9;FAUX)
6	=(LIGNE()-4)/$B$1	=LOI.BINOMIALE(A6*$B$1;$B$1;Population!$F$9;FAUX)

Nous pouvons avec Excel, sur une nouvelle feuille, construire cette loi théorique, à l'aide des formules suivantes :

La fonction LOI.BINOMIALE comporte quatre paramètre :

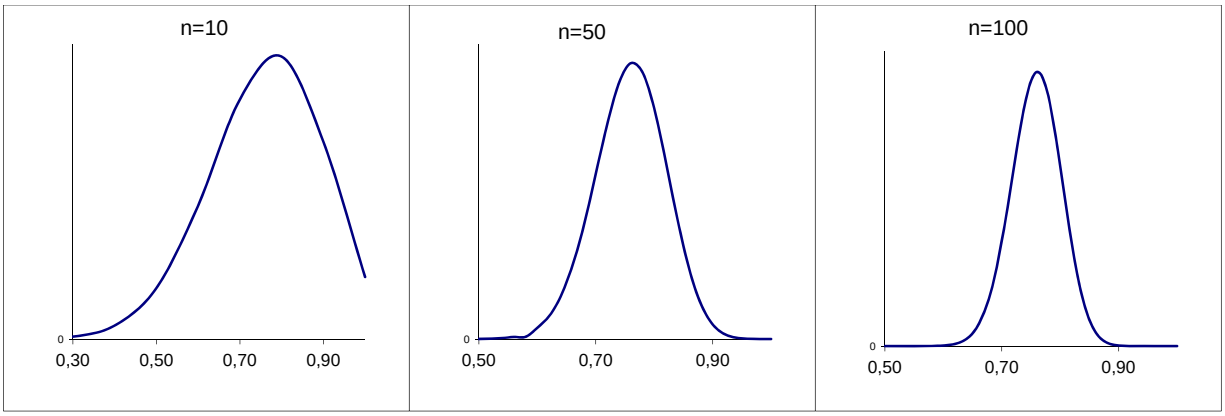
- le premier est le nombre de succès, c'est à dire pour nous le nombre de clients dans l'échantillon faisant leur révision chez un concessionnaire, c'est donc la taille de l'échantillon multiplié par l'estimation du pourcentage.
- Le second est la taille de l'échantillon
- Le troisième est la vraie valeur du paramètre, le pourcentage réel dans la population
- Le dernier est un indicateur logique du cumul de la loi, ici faux car ne voulons pas la loi cumulée.

Par exemple pou un échantillon de taille 10, la loi de probabilité de la proportion estimée sur les échantillons sera la suivante :

	A	B
1	Taille de l'échantillon	10
2		
3	**Pourcentage Estimé**	**Probabilité**
4	0	0,00000067
5	0,1	0,00002104
6	0,2	0,00029774
7	0,3	0,00249642
8	0,4	0,01373609
9	0,5	0,05182648
10	0,6	0,13579312
11	0,7	0,24397594
12	0,8	0,28766404
13	0,9	0,20099273
14	1	0,06319572
15		
16	***Moyenne***	***0,75869770***

Remarquons tout d'abord, que l'on retrouve bien ici la proportion réelle comme espérance de la loi binomiale, et on pourrait conclure par exemple, de l'examen de cette loi, après avoir éliminé les échantillons les plus exceptionnels (dont la probabilité est la plus faible), que 95% des échantillons donneront une proportion comprise entre 60% et 100%, donc une précision de l'ordre de 20% au degré de confiance 0,95.

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

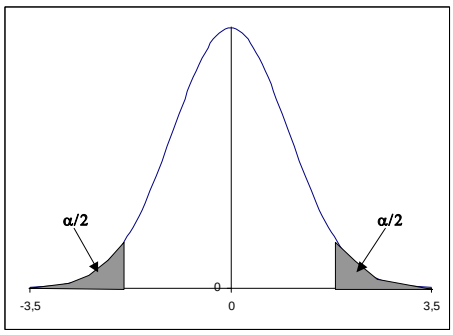
$$\sigma(\overline{X}n)=\sqrt{Var(\overline{X}n)}=\sqrt{\frac{Var(X)}{n}}\ .$$

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

4.2.2 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 :

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



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 $1-\alpha/2$:

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

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 :

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

4.2.3 Utilisation d'Excel

Il est alors facile de mettre en place les formules permettant les calculs de l'intervalle d'estimation, en supposant donnée la taille de l'échantillon n et le degré de confiance $1-\alpha$ voulu. Nous utiliserons la fonction statistique d'Excel :

LOI.NORMALE.STANDARD.INVERSE(probabilité)

qui retourne le fractile d'une probabilité donnée. Il nous reste simplement à exprimer la valeur $1-\alpha/2$ dont nous voulons obtenir le fractile, en fonction du degré de confiance $1-\alpha$, qui est connu.

La formule est simple : $1-\alpha/2 = \frac{1+(1-\alpha)}{2}$. La feuille de calcul se présente alors sous la forme suivante :

	A	B	C	D
1				
2	*Taille Echantillon*	*200*		
3				
4	*Pourcentage Estimé*	*0,72*	**Ecartype Estimé**	=RACINE(B4*(1-B4)/(B2-1))
5				
6	*Degré de Confiance*	*0,95*		
7				
8	**Précision**	=LOI.NORMALE.STANDARD.INVERSE((1+B6)/2)*D4		
9				
10		**Borne Inférieure**	**Borne Supérieure**	
11	**Intervalle**	=B4-B8	=B4+B8	

Les résultats numériques sont alors les suivants :

	A	B	C	D
1				
2	*Taille Echantillon*	*200*		
3				
4	*Pourcentage Estimé*	*72%*	**Ecartype Estimé**	**3,18%**
5				
6	*Degré de Confiance*	*0,95*		
7				
8	**Précision**	**6,24%**		
9				
10		**Borne Inférieure**	**Borne Supérieure**	
11	**Intervalle**	**65,76%**	**78,24%**	

Vérifions que les approximations faites ne conduisent pas à une dégradation des termes du pari. Construisons un grand nombre d'intervalles d'estimations pour un degré de confiance donné (0,95 par exemple) et plusieurs tailles d'échantillons (de 100 à 500) et déterminons le pourcentage de paris gagnés, c'est à dire la fréquence de la présence de la « vraie » valeur du pourcentage dans l'intervalle construit à partir des estimations.

A partir de la table construite plus haut (3.2.2), nous créons un indicateur de réussite qui vaut 1 si l'intervalle d'estimation contient le vrai pourcentage, 0 sinon avec la formule suivante :

=SI(ET(Population!$F$9<B10+$I$4*RACINE(B10*(1-B10)/H$9);Population!$F$9>B10-$I$4*RACINE(B10*(1-B10)/H$9));1;0)

- Population!$F$9 faisant référence à la « vraie valeur » de la proportion
- $I$4 est la référence du fractile de la loi normale centrée réduite
- B10 est le pourcentage estimé
- H$9 est la taille de l'échantillon

On obtient alors les résultats suivants :

	G	H	I	J	K	L
3		**Verification**				
4		Confiance	0,95			
5		Fractile	1,9599611			
6						
7	**Paris réussis**	**94,1%**	**95,7%**	**94,3%**	**94,7%**	**95,6%**
8		**Taille de l'échantillon**				
9		100	200	300	400	500
10		1	1	1	1	1
11		1	1	1	1	1

On obtient bien un résultat proche des 95% de paris réussis (refaire éventuellement une estimation par intervalle !)

4.2.4 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.

1) 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 $1/4$[2]. 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 :

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

[2] 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.

EntierSup(x) désignant le plus petit entier supérieur ou égal à x, ce qui correspond à la fonction d'Excel ARRONDI.SUP(x ;0).

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.

2) 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 :

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

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.

3) Calculs sous Excel et comparaison

Nous allons mettre sur une même feuille, les résultats obtenus dans les deux cas évoqués ci-dessus, les formules sont les suivantes, nous avons créé une cellule contenant le fractile de la loi normale centrée réduite, de façon à obtenir des formules plus lisibles. Nous avons ensuite créé un tableau des tailles correspondant à différentes pré-estimations du pourcentage, il apparaît alors clairement, qu'économiquement il est important de tenir compte d'une estimation antérieure du paramètre recherché.

	A	B
1	*Degré de Confiance*	*0,95*
2	*Précision Voulue*	*0,03*
3		
4	**Fractile**	**=LOI.NORMALE.STANDARD.INVERSE((1+B1)/2)**
5		
6	***Taille***	
7	***Sans Préestimation***	=ARRONDI.SUP($B$4^2/(4*$B$2^2);0)
8	**Estimation**	**Taille**
9	0,1	=ARRONDI.SUP($B$4^2*A9*(1-A9)/($B$2^2);0)+1
10	0,2	=ARRONDI.SUP($B$4^2*A10*(1-A10)/($B$2^2);0)+1

	A	B
1	*Degré de Confiance*	*0,95*
2	*Précision Voulue*	*3,00%*
3		
4	**Fractile**	**1,96**
5		
6	***Taille***	
7	***Sans Préestimation***	***1068***
8	**Estimation**	**Taille**
9	10%	386
10	20%	684
11	30%	898
12	40%	1026
13	50%	1069
14	60%	1026
15	70%	898
16	80%	684
17	90%	386

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.

4.3 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 3.2.1 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 : $Zn=\frac{\overline{Y}n-\mu}{\sigma/\sqrt{n}}$ 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 Zn précédemment définie suit toujours une loi normale.

4.3.1 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.

Sous Excel cette précision se calcule à l'aide de la fonction INTERVALLE.CONFIANCE qui admet trois paramètres :

- Alpha : qui est égal au risque pris, c'est à dire à 1-degré de confiance
- Ecart type : qui est l'écart type connu sur la population.
- Taille : la taille de l'échantillon

Exemple d'application, sur un échantillon de taille 100, tiré du fichier Martin :

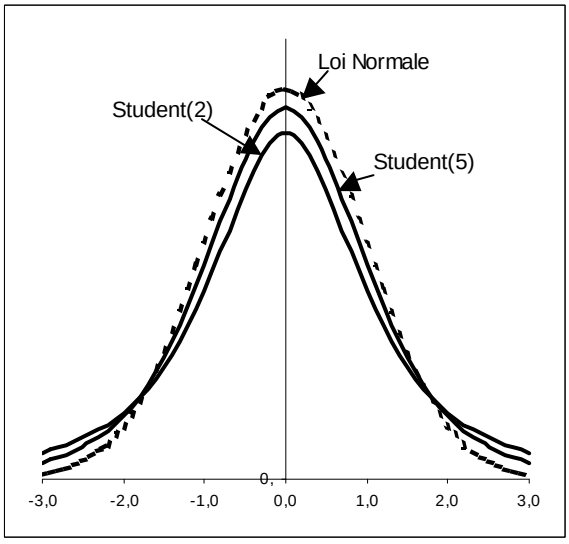
	E	F	G
9	**Degré de confiance**	**0,95**	
10	**Précision**	**=INTERVALLE.CONFIANCE(1-F9;Population!$F$6;techan)**	
11			
12		Borne Inférieure	Borne Supérieure
13	Intervalle	=F6-F10	=F6+F10

La cellule Population !$F$6 est la cellule contenant la valeur de l'écart type du kilométrage parcouru sur toute la population. Les valeurs obtenues sont les suivantes :

	E	F	G	H
5		Moyenne	Variance	Ecart-type
6	Kilométrage	24920,00	14154141,41	3762,20
7				
8				
9	**Degré de confiance**	**0,95**		
10	**Précision**	**779,70**		
11				
12		Borne Inférieure	Borne Supérieure	
13	Intervalle	24140,30	25699,70	

4.3.2 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 $T_n = \frac{\overline{Y}n - \mu}{\sqrt{S\tilde{n}^2 / n}}$ (σ est remplacé par l'estimateur de l'écart type), cette loi est la loi de Student[3] 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 :

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

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.

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

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

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-a, que la moyenne se trouve dans l'intervalle $\left[\bar{y}_n - t_{1-\alpha/2}^{n-1}\hat{\sigma}/\sqrt{n}; \bar{y}_n + t_{1-\alpha/2}^{n-1}\hat{\sigma}/\sqrt{n}\right]$,

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

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

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

Sous Excel nous allons utiliser la fonction donnant le fractile de la loi de Student, il faut noter que Excel ne donne pas le fractile exactement, mais raisonne toujours symétriquement et par complémentarité. De façon précise, la fonction LOI.STUDENT.INVERSE a deux paramètres :

- p : probabilité, qui est un nombre compris entre 0 et 1
- d : nombre de degrés de liberté

Et retourne une valeur t telle que $\Pr(|\text{Student}(d)| \geq t) = p$, pour calculer la précision nous prendrons donc comme valeur : $p = \alpha = 1 - \text{degré de confiance}$ et $d = n - 1$. Nous avons alors les formules suivantes :

	E	F	G
9	**Degré de confiance**	0,95	
10	**Précision**	=LOI.STUDENT.INVERSE(1-F9;techan-1)*H6/RACINE(techan)	
11			
12		Borne Inférieure	Borne Supérieure
13	Intervalle	=F6-F10	=F6+F10

La cellule H6 de la feuille active (Feuille nommée Echantillon) est la cellule contenant *l'estimation de l'écart type* à partir de l'échantillon.

4.4 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 :

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

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.

4.4.1 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. Voici la formule utilisée sous Excel, et les valeurs correspondantes :

	A	B
1	*Degré de confiance*	*0,95*
2	*Précision voulue*	*250*
3	*Ecart type Estimé*	*3950*
4		
5	**Taille nécessaire**	**=ARRONDI.SUP((LOI.NORMALE.STANDARD.INVERSE((1+B1)/2)*B3/B2)^2;0)**

	A	B
1	*Degré de confiance*	*0,95*
2	*Précision voulue*	*250*
3	*Ecart type Estimé*	*3950*
4		
5	**Taille nécessaire**	**959**

L'écart type estimé, était le résultat d'un pré sondage sur 20 individus du fichier Martin, pour la variable kilométrage. Il faudrait donc ajouter environ 940 autres individus pour obtenir une précision sur le kilométrage moyen de l'ordre de 250 km.

Toutefois sur ce nouvel échantillon, l'estimation de l'écart type sera différente, mais plus fiable puisque prise sur un échantillon de taille plus importante, et il faudra donc calculer de nouveau la précision obtenue.

4.4.2 Cas général

Si l'on ne veut pas utiliser l'approximation par une loi normale, il est possible d'utiliser les fonctionnalités d'Excel pour résoudre l'équation implicite définissant la taille de l'échantillon.

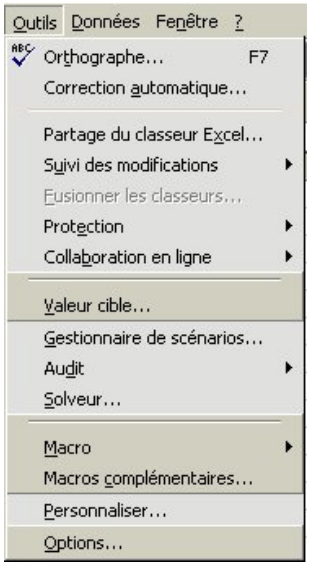
Sur une feuille contenant les résultats du pré sondage, nous allons ajouter trois éléments, le seuil de précision voulu, le seuil de précision obtenue avec la taille d'échantillon, l'écart entre la précision obtenu et la précision voulue. Nous obtenons les éléments suivants :

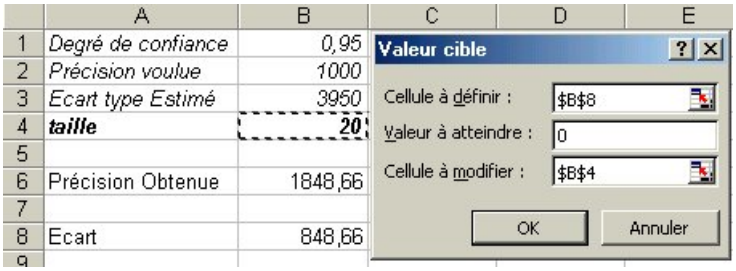
	A	B
1	*Degré de confiance*	*0,95*
2	*Précision voulue*	*1000*
3	*Ecart type Estimé*	*3950*
4	***taille***	***20***
5		
6	Précision Obtenue	=LOI.STUDENT.INVERSE(1-B1;B4-1)*B3/RACINE(B4)
7		
8	Ecart	=B6-B2

	A	B
1	*Degré de confiance*	*0,95*
2	*Précision voulue*	*1000*
3	*Ecart type Estimé*	*3950*
4	***taille***	***20***
5		
6	Précision Obtenue	1848,66
7		
8	Ecart	848,66

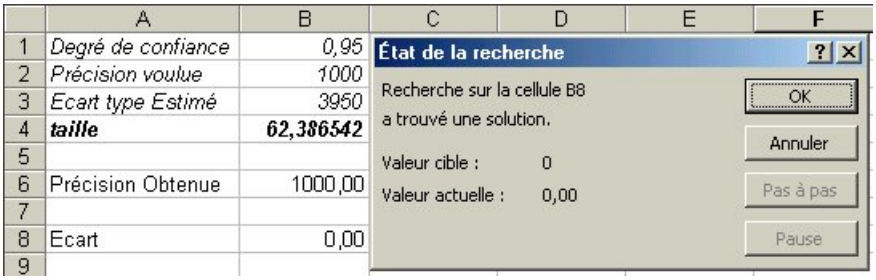
Il nous faut maintenant modifier, la taille de l'échantillon de façon à ce que la précision obtenue soit égale à la précision voulue, c'est à dire que l'écart soit égal à 0. Il est possible de le faire manuellement par tâtonnement, mais il est plus judicieux d'utiliser la commande Valeur Cible d'Excel.

Dans le Menu Outils d'Excel, choisissons cette commande, nous obtenons alors la boite de dialogue :





La cellule à définir correspond à la fonction qui doit atteindre une certaine valeur, donc ici la cellule contenant l'écart entre la précision voulue et la précision obtenue. La valeur à atteindre est ici 0 ; enfin la cellule à modifier, correspond à la taille de l'échantillon. Après avoir validé ces entrées, nous obtenons la boite de dialogue suivante :



Indiquant que la valeur a été atteinte, en validant par OK, la cellule correspondant à la taille contiendra la solution, c'est à dire la taille d'échantillon permettant d'obtenir la précision voulue. Comme cette solution n'est pas obligatoirement entière, il nous faudra, dans une autre cellule, prendre l'entier immédiatement supérieur. Ici, il faudrait donc un échantillon de taille 63 environ, pour atteindre une précision de 1000km, sur le kilométrage moyen annuel des clients.

Si nous calculons, cette taille avec l'approximation normale, nous aurions trouvé 60, un nombre évidemment inférieur, mais peu différent ; c'est pourquoi la plupart du temps on se contentera de l'approximation normale pour la détermination de la taille d'échantillon. La différence entre les deux approches n'étant réellement significative que sur les petits échantillons, auquel cas il est nécessaire de croire à l'hypothèse de normalité, puisque l'on ne dispose pas de données suffisantes pour la tester.

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[4] à n degrés de liberté, X et Z étant de plus indépendantes.

Dans le cas de l'estimation la variable X est l'estimateur de la moyenne $\frac{\overline{Y_n} - \mu}{\sigma / \sqrt{n}}$ 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 $Y_i - \overline{Y_n}$ sont liées par la relation $\sum_{i=1}^{n} Y_i - \overline{Y_n} = 0$; la forme quadratique $(n-1)S_n^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y_n}\right)^2$ 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.

Cette distribution n'est pas donnée directement dans Excel, puis que seule apparaît dans les fonctions d'Excel la fonction de répartition (et pas directement!), si vous voulez tracer cette fonction, il vous faudra donc entrer la formule ci-dessus. On est alors confronté à un nouveau problème, la fonction Gamma; cette fonction n'est pas une fonction d'Excel, seule existe la fonction LNGAMMA(x) qui est le logarithme népérien de la fonction Gamma, il suffira alors de prendre l'exponentielle de cette fonction (voir le fichier Student.xls).

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 α/2 de cette loi, et χ_2 le fractile d'ordre 1- α/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 :

[4] 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

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.

Sous Excel on utilisera la fonction KHIDEUX.INVERSE a deux paramètres :

- p : représente la probabilité d'observer une valeur supérieure au fractile cherché
- d : le nombre de degrés de libertés

Pour un degré de confiance 1-α donné, χ_1 et χ_2 seront définis par :

$$\chi_1 = \text{KHIDEUX.INVERSE}\left(1-\frac{\alpha}{2}, n-1\right) \quad \text{et} \quad \chi_2 = \text{KHIDEUX.INVERSE}\left(\frac{\alpha}{2}, n-1\right)$$

Nous laissons au lecteur le soin d'utiliser ces formules sur l'exemple, nous aurons l'occasion de revenir sur l'utilisation de cette fonction pour le test de contingence.

7 Annexe 4 : Méthode du maximum de vraisemblance

Nous avons jusqu'à présent utiliser des estimateurs "intuitifs" qui se sont avérés efficaces, il existe une méthode mathématique pour trouver systématiquement des estimateurs de paramètres en faisant des hypothèses sur la loi de probabilité suivie par une variable. C'est la méthode du maximum de vraisemblance qui est très utilisée en modélisation statistique et assez facile à mettre en œuvre sur ordinateur. Nous en donnerons le principe ainsi qu'un exemple ici, avec la résolution analytique et avec Excel.

7.1 Formalisme du maximum de vraisemblance

On suppose qu'une variable statistique *X* définie sur une population ***P***, suit une loi donnée dépendant de *p* paramètres $(a_i)_{1\le i\le p}$. La densité de probabilité de *X* (que nous supposons exister) est donc une fonction dépendant à la fois de la valeur *x* prise par X et des paramètres à estimer, nous la noterons $f(x, a_1, \cdots, a_p)$. Par exemple si l'on veut estimer la moyenne et la variance d'une variable X supposée suivre une loi normale, les deux paramètres sont la moyenne μ et l'écart type σ, et la fonction de densité sera donnée par

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Soit maintenant un échantillon aléatoire simple de taille n, tiré dans la population et $(x_j)_{1\le j\le n}$ l'échantillon image, la probabilité d'obtenir cet échantillon est alors $\prod_{j=1}^{n} f(x_j, a_1, \cdots, a_p)$, on appelle vraisemblance de cet échantillon le logarithme de cette probabilité :

$$\mathcal{L}\left((x_j)_{1\le j\le n}, a_1, \cdots, a_p\right) = \sum_{j=1}^{n} \log f(x_j, a_1, \cdots, a_p)$$

le principe du maximum de vraisemblance, consiste à dire que l'échantillon tiré maximise cette vraisemblance, l'estimation des paramètres peut alors être trouvée soit analytiquement (par exemple en annulant les dérivées partielles par rapport aux paramètres) soit

numériquement par un algorithme de maximisation. Si le calcul analytique est possible, l'estimateur associé s'obtient en remplaçant les valeurs x_j par les variables X_j.

L'intérêt de cette méthode est que les estimateurs ainsi trouvé sont de bons estimateurs, asymptotiquement sans biais et convergents. De plus il est possible de trouver la loi limite de ces estimateurs, ce qui permet des estimations par intervalle.

7.2 Estimateurs du maximum de vraisemblance des paramètres d'une loi normale

Soit donc $(x_j)_{1\le j\le n}$ un échantillon de taille *n* d'une variable supposée suivre une loi normale de moyenne μ et d'écart type σ. La fonction de vraisemblance de l'échantillon est définie par :

$$\mathcal{L}\left((x_j)_{1\le j\le n},\mu,\sigma\right)=-nLog\left(\sqrt{2\pi}\right)-nLog(\sigma)-\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{2\sigma^2}$$

Les estimations du maximum de vraisemblance de deux paramètres sont les valeurs *m* et *s* telles que :

$$\mathcal{L}\left((x_j)_{1\le j\le n},m,s\right)=\underset{a,b}{Max}\,\mathcal{L}\left((x_j)_{1\le j\le n},a,b\right)$$

En utilisant les dérivées partielles, nous obtenons les deux équations suivantes :

$$\frac{\partial\mathcal{L}}{\partial a}(m,s)=\sum_{j=1}^{n}\frac{(x_j-m)}{s^2}=0 \text{ et } \frac{\partial\mathcal{L}}{\partial b}(m,s)=-\frac{n}{s}+\sum_{j=1}^{n}\frac{(x_j-m)^2}{s^3}$$

La première équation permet de calculer $\sum_{j=1}^{n}(x_j-m)=0$ soit $m=\frac{1}{n}\sum_{j=1}^{n}x_j$ qui est la moyenne de l'échantillon ce qui correspond à l'estimateur que nous avons utilisé : $m=\overline{x}_n$.

En revanche la seconde équation donne $ns^2=\sum_{j=1}^{n}(x_j-m)^2$, soit $s^2=\frac{1}{n}\sum_{j=1}^{n}(x_j-m)^2$.

L'estimateur du maximum de vraisemblance de la variance est donc la variance de l'échantillon qui comme nous l'avons vu est asymptotiquement sans biais mais non sans biais.

Il est facile de vérifier que les conditions du second ordre sont vérifiées pour le maximum calculé ci-dessus.

7.3 Utilisation d'Excel pour le calcul d'estimation

Le fichier **MaxVrai.xls** contient 10 données qui sont supposées provenir d'une loi normale de moyenne et variance inconnue. Ces données sont dans la plage A5:A14, nous allons construire sur cette feuille de calcul, la fonction de vraisemblance de l'échantillon pour une moyenne et un écart type donnés.

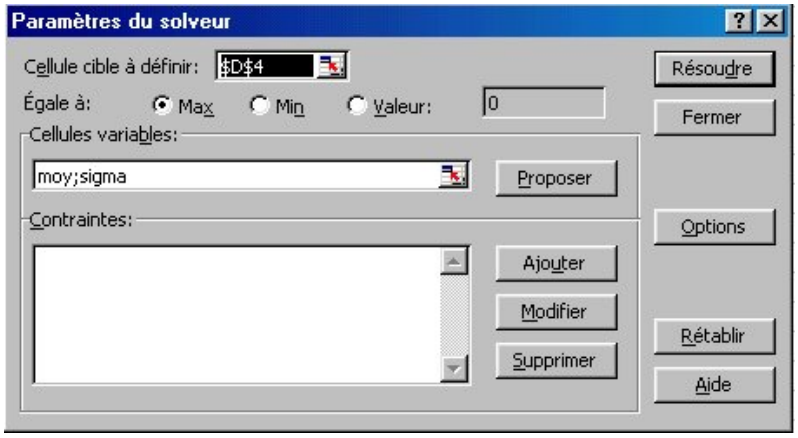
La moyenne est dans la cellule D1, nommée "*moy*", et initialisée à une valeur arbitraire (20). L'écart type est dans la cellule D2, nommée "*sigma*", et initialisé à la valeur 5.

On entre alors les formules permettant de calculer la vraisemblance, c'est à dire la densité de probabilité en chaque valeur et le logarithme népérien de cette probabilité :

	A	B	C
4	**Valeur**	**Proba**	**Vraisemblance**
5	10	=LOI.NORMALE(A5;moyenne;sigma;FAUX)	=LN(B5)
6	30	=LOI.NORMALE(A6;moyenne;sigma;FAUX)	=LN(B6)
7	23	=LOI.NORMALE(A7;moyenne;sigma;FAUX)	=LN(B7)
8	24	=LOI.NORMALE(A8;moyenne;sigma;FAUX)	=LN(B8)
9	12	=LOI.NORMALE(A9;moyenne;sigma;FAUX)	=LN(B9)
10	25	=LOI.NORMALE(A10;moyenne;sigma;FAUX)	=LN(B10)
11	18	=LOI.NORMALE(A11;moyenne;sigma;FAUX)	=LN(B11)
12	15	=LOI.NORMALE(A12;moyenne;sigma;FAUX)	=LN(B12)
13	27	=LOI.NORMALE(A13;moyenne;sigma;FAUX)	=LN(B13)
14	18	=LOI.NORMALE(A14;moyenne;sigma;FAUX)	=LN(B14)

La vraisemblance de l'échantillon est simplement la somme des cellules C5 à C14, valeur de cette vraisemblance est dans la cellule D4.

Une fois ce modèle écrit, il nous faut utiliser le solveur d'Excel pour maximiser la vraisemblance. Après avoir sélectionner la cellule D2, nous choisissons le menu "Outils-Solveur" qui conduit à la boite de dialogue suivante :



Après avoir demandé la résolution nous obtenons les résultats suivants (nous avons affiché les valeurs des fonctions MOYENNE, ECARTYPE et ECARTYPEP sur la feuille :

C	D	E	F	G	H
Moyenne	**20,2**		*20,2*	*Fonction MOYENNE*	
Ecart type	**6,28967399**		*6,62989861*	*Fonction ECARTYPE*	
			6,28967408	*Fonction ECARTYPEP*	

La convergence pour la moyenne s'est bien faite vers la valeur de la fonction moyenne, en revanche pour l'écart type la convergence se fait vers la fonction ECARTYPEP, qui est l'écart type de l'échantillon et non pas vers la fonction ECARTYPE qui est l'estimation habituelle de l'écart type de la population.

Remarque importante : si les valeurs initiales des paramètres sont trop éloignées des valeurs estimées, l'algorithme de recherche de maximum peut échouer, il est donc recommandé avant d'utiliser le solveur de faire une table pour différentes valeurs des paramètres.

EXERCICES ESTIMATION

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

2 La société UVJM (Classeur UVJM.xls)

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 la feuille "Premier sondage" du classeur.

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 donnés dans la feuille "Sondage supplémentaire". 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 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.

3.1 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 1% de la valeur comptable du stock. Conclusion .

3.2 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.

3.3 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

3.4 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

3.5 Annexe 3

Résultats des contrôles des catégories 1 et 3

1. Catégorie d'articles de valeur élevée

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

2. 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 €

4 La société de contrôle et de régulation (Deuxième partie : CasSCR.xls)

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 se présentent sous la forme suivante :

Référence	Quantité C	Prix unitaire	Valeur C	Quantité Stock
0pp10673	369	8	2952	369
0pp12370	389	3	1167	389
0pp15926	402	9	3618	402
0pp29971	434	4	1736	434

Les contenus de chaque colonne sont les suivants :

- Référence : Le numéro de référence du produit.
- Quantité C : quantité comptable, la quantité enregistrée informatiquement
- Prix unitaire : le prix unitaire du produit en euro.
- Valeur comptable : la valeur de la référence enregistrée informatiquement (Quantité C*Prix unitaire).
- Quantité en stock : la quantité physique vérifiée en stock.

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 ? On pourra utiliser soit un histogramme, soit le graphique normal (voir le chapitre Rappel Excel). Justifiez économiquement ce fait.
3. Donner une estimation par intervalle du pourcentage d'erreur dans la seconde catégorie.

5 Maximum de vraisemblance pour la loi exponentielle

La loi exponentielle est une loi à un paramètre λ dont la densité est donnée par la formule.

1. A partir d'un échantillon de taille *n* quelle est l'estimation du maximum de vraisemblance du paramètre λ? Comparer cette estimation à l'estimation de la moyenne.
2. En utilisant les données du fichier MaxVrai.xls, retrouver le résultat précédent (prendre 0,04 comme valeur initiale de lambda, par exemple).

6 Maximum de vraisemblance pour une loi uniforme sur un intervalle

La loi uniforme sur un intervalle $[a,b]$ dépend des deux paramètres a et b, sa densité est donnée par :

$$f(x,a,b) = \frac{1}{b-a} \text{ si } a \le x \le b \text{ et } 0 \text{ sinon}$$

1. A partir d'un échantillon de taille *n* quelle est l'estimation du maximum de vraisemblance des paramètres *a* et *b*?

En utilisant les données du fichier MaxVrai.xls, retrouver le résultat précédent. Que se passe-t-il si l'une des valeurs initiales des paramètres est entre le maximum et le minimum

TESTS D'HYPOTHESE

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.

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 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.

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.

3.1 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.

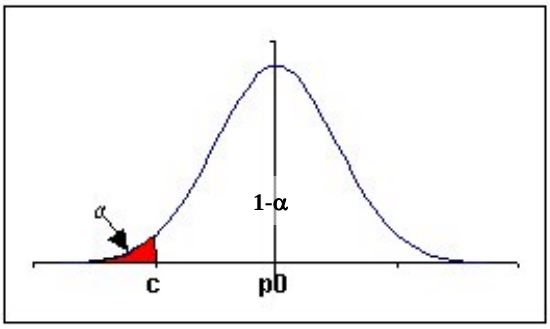
Remarquons que cette loi ne fait pas intervenir des résultats obtenus par sondage, il est donc possible ici de définir la zone de rejet avant même de procéder au sondage. C'est ce que nous allons faire pour les trois cas décrits plus hauts. Nous indiquerons aussi comment calculer avec Excel le niveau de signification du test.

3.2 Tests unilatéraux

Nous traiterons simultanément les deux cas gauche et droite :

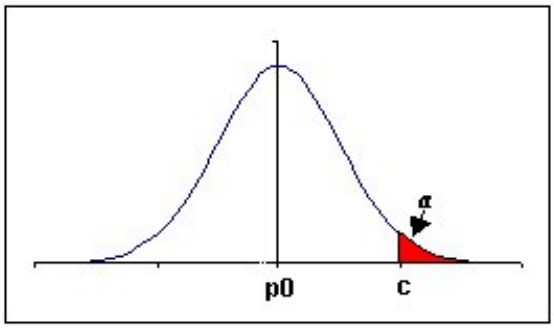
$H_0 : p=p_0$

$H_1 : p<p_0$



$H_0 : p=p_0$

$H_1 : p>p_0$



3.2.1 Cas des petits échantillons, détermination exacte avec Excel

En utilisant la variable binomiale $n\overline{X}_n$ il est facile de déterminer la valeur de nc à l'aide de la fonction CRITERE.LOI.BINOMIALE(*n*, p_0,proba) qui donne la plus valeur pour laquelle la loi cumulée est supérieure à une probabilité donnée. (fichier Standard.xls, feuille proportion), on divisera ensuite par *n* pour obtenir la valeur de *c*.

La probabilité cumulée est ici α

	A	B	C
1	Taille échantillon		40
2	Pourcentage $p\theta$		0,4
3			
4		α	0,1
5	*Valeur de c*		=CRITERE.LOI.BINOMIALE($C$1;$C$2;C4)/$C$1

soit en valeur :

	A	B	C
1	Taille échantillon		40
2	Pourcentage $p\theta$		40%
3			
4		α	10%
5	*Valeur de c*		**30,00%**

La règle de décision est la même que celle qui sera énoncée pour l'approximation normale (cf. ci-dessous).

Ici, la probabilité cumulée est 1-α

	A	B	C
1	Taille échantillon		40
2	Pourcentage $p\theta$		0,4
3			
4		α	0,1
5	*Valeur de c*		=CRITERE.LOI.BINOMIALE($C$1;$C$2;1-C4)/$C$1

soit en valeur :

	A	B	C
1	Taille échantillon		40
2	Pourcentage $p\theta$		40%
3			
4		α	10%
5	*Valeur de c*		**50,00%**

La règle de décision est la même que celle qui sera énoncée pour l'approximation normale (cf. ci-dessous).

3.2.2 Cas des grands échantillons, approximation normale avec Excel

Nous allons ici utiliser, la convergence de la loi de $\overline{X}_n$ vers la loi normale, on peut avec Excel soit utiliser directement la loi normale de paramètre $\left(p_0,\sqrt{p_0(1-p_0)/n}\right)$, soit après centrage et réduction se ramener à la loi normale centrée réduite, nous donnerons les formules de calcul de c en fonction de la loi normale centrée réduite, en revanche nous donnerons les deux formules d'Excel avec la loi normale centrée réduite pour le test gauche, avec la loi non centrée réduite pour le test unilatéral à droite. Nous désignerons, comme d'habitude par z_q le fractile d'ordre q de la loi normale centrée réduite, c'est à dire la valeur définie par : $prob(N(0,1)<z_q)=q$. Comme la variable $\frac{\overline{X}_n-p_0}{\sqrt{p_0(1-p_0)/n}}$ suit une loi normale standard (centrée réduite), il est facile de déterminer dans les deux cas la valeur critique c

Nous avons ici :

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

$c=p_0+z_\alpha*\sqrt{p_0(1-p_0)/n}$ 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.

Formule avec Excel utilisant directement la loi de $\overline{X}_n$, dans ce cas c'est simplement le fractile d'ordre a de la loi de $\overline{X}_n$: la formule utilisée est LOI.NORMALE.INVERSE(α;μ;σ) soit :

	A	B	C
1	Taille échantillon		1000
2	Pourcentage p_0		0,4
3	Sigma de Xn		=RACINE($C$2*(1-$C$2)/$C$1)
4		α	0,05
5	Approx. norr		=LOI.NORMALE.INVERSE(C4;$C$2;$C$3)

Ce qui nous donne les valeurs numériques suivantes pour différentes valeurs du risque de première espèce :

	Alpha	10%	5%	1%
Avec approximation		38,01%	37,45%	36,40%

Nous avons ici :

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

$c=p_0+z_{1-\alpha}*\sqrt{p_0(1-p_0)/n}$ 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.

Formule avec Excel utilisant la loi normale centrée réduite, c'est à dire la formule ci dessus :

=$C$2+RACINE($C$2*(1-$C$2) /$C$1) * LOI.NORMALE.STANDARD.INVERSE(1-C4)

avec la même disposition que pour le test unilatéral gauche.

Ce qui nous donne les valeurs numériques suivantes pour différentes valeurs du risque de première espèce :

	Alpha	10%	5%	1%
Avec approximation		41,99%	42,55%	43,60%

En appliquant la règle de décision, comme sur l'échantillon nous obtenons 45,2%, nous pouvons considérer avec un risque d'erreur inférieur à 1% que le taux de retour est bien supérieur au taux habituel de 40%

3.2.3 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 :

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

Sous Excel on peut utiliser la fonction :

LOI.NORMALE($\hat{p}$;p0;RACINE(p0*(1-p0)/n);VRAI)

Le dernier paramètre indiquant que l'on veut la loi cumulée

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 :

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

ce qui se traduit sous Excel par :

	A	B	C
1	Taille échantillon		1000
2	Nbre de retours		452
3	p chapeau		=C2/Téchan
4	Pourcentage p_0		0,4
5	Sigma Xn		=RACINE(C4*(1-C4)/C1)
6	***ns***		***=1-LOI.NORMALE.STANDARD((C3-C4)/C5)***

La valeur du niveau de signification obtenue *ns*=0,0004 qui est bien inférieur à 1%.

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.

3.2.4 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

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

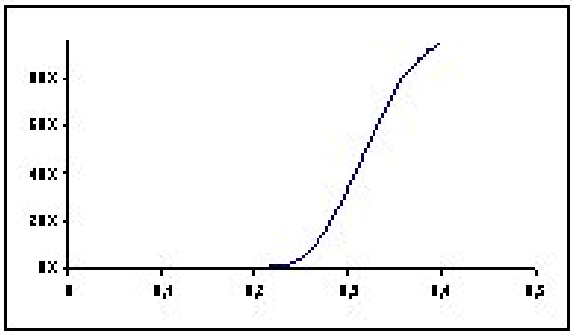
Pour une valeur donnée du risque de première espèce α, la valeur critique c est calculée.

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 :

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

En utilisant cette définition, on obtient alors la courbe suivante (voir le fichier Excel pour le détail des formules) :



Remarque : le test

$$H_0 \ : \ p \geq p_0$$

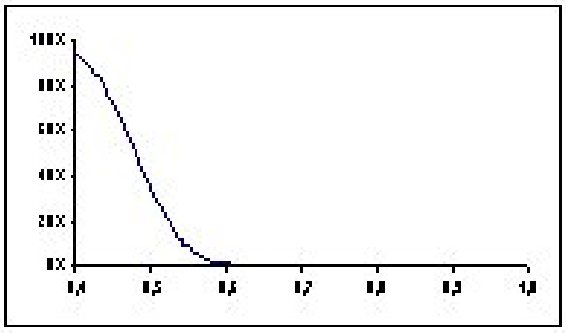
$$\text{contre } H_1 \ : \ p < p_0$$

se ramène à ce test unilatéral

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 :

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

En utilisant cette définition, on obtient alors la courbe suivante (voir le fichier Excel pour le détail des formules):



De même le test

$$H_0 \ : \ p \leq p_0$$

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

se ramène à ce test unilatéral

3.3 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.*

3.3.1 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 :

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

ce qui s'exprime sous Excel sous la forme :

	A	B	C
1	Taille échantillon		500
2	Nbre de retours		175
3	p chapeau		=C2/C1
4	Pourcentage p_0		0,4
5	Sigma Xn		=RACINE(C4*(1-C4)/C1)
6	***ns***		***=2*(1-LOI.NORMALE.STANDARD(ABS(C3-C4)/C5))***

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.

3.3.2 Courbe de puissance du test

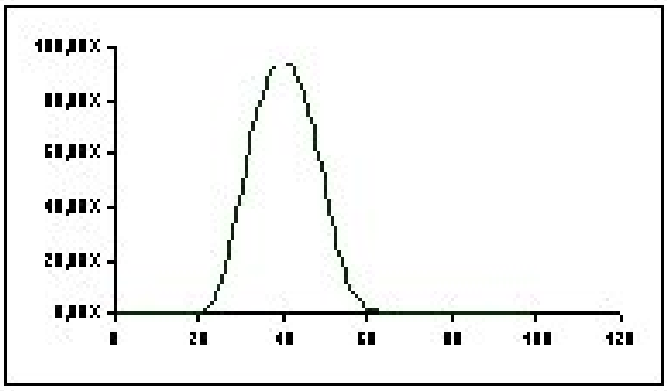
La courbe de puissance du test est symétrique par rapport à p_0, elle n'est pas exactement obtenue comme "recollement" des deux courbes définies précédemment pour les tests unilatéraux. Indiquons rapidement comment on peut avec Excel construire cette courbe. Ici l'ensemble $A_1=[0;p_0[\cup]p_0;1]$, pour 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 :

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

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

En utilisant cette définition, on obtient alors la courbe suivante (voir le fichier Excel pour le détail des formules) avec p_0=40% et n=100 :



4 Comparaison d'une moyenne à un standard

4.1 Un exemple (fichier ptidej.xls)

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 $H_0 \ : \ \mu \le 320$, 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.

4.2 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:

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

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 : 328,27

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

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

4.3 Tests unilatéraux

Nous traiterons simultanément les deux cas gauche et droite :

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

4.3.1 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 :

$prob\left(|T_n| > t_q^n\right) = q$, 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

$c = \mu_0 - t_{2\alpha}^{n-1} * \hat{\sigma}/\sqrt{n}$ 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.

Formule avec Excel, en utilisant la loi de

Nous avons ici :

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

$c = \mu_0 + t_{2\alpha}^{n-1} * \hat{\sigma}/\sqrt{n}$ 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.

Formule avec Excel, en utilisant la loi de

Student : la formule utilisée pour le calcul de la valeur de c est :

μ_0 - LOI.STUDENT.INVERSE(2α;n-1)*s/$\sqrt{n}$

Student inverse, la formule ci dessus devient :

μ_0 - LOI.STUDENT.INVERSE(2α;n-1)*s/$\sqrt{n}$ soit :

	A	B
9	Test μ>$B$7	
10		
11	Risque de 1° espèce	0,1
12	Valeur critique	=$B$7+$B$5/RACINE($B$3)*LOI.STUDENT.INVERSE(2*B11;$B$3-1)

Où B5 est la cellule contenant l'estimation de l'écart type, B3 celle contenant la taille de l'échantillon et B7 celle contenant la valeur μ_0.

Ce qui donne les valeurs numériques suivantes pour différentes valeurs du risque de première espèce :

	A	B	C	D
11	Risque de 1° espèce	0,1	0,05	0,01
12	Valeur critique	323,65	324,69	326,65

En appliquant la règle de décision, comme sur l'échantillon nous obtenons une moyenne de 332, nous pouvons considérer avec un risque d'erreur inférieur à 1% que le seuil de rentabilité est bien atteint, et M. Durlan peut décider de lancer ce produit.

4.3.2 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 $\overline{x}_n$. 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 :

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

C'est à dire la valeur de la fonction de répartition de la loi de Student à (n-1) degrés de liberté, pour la valeur (standardisée) :

$$\frac{\overline{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

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

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

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) :

$$\frac{\overline{x}_n - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

Il nous faut donc, dans les deux cas, utiliser la fonction de répartition de la loi de Student, cette fonction n'existe pas directement sous Excel, mais il existe une fonction qui permet de la calculer, la fonction LOI.STUDENT dont la syntaxe est la suivante :

LOI.STUDENT(Valeur, degrés, uni ou bilatéral)

Pour nous le dernier paramètre sera dans les deux cas égal à 1(unilatéral). Dans ce cas la fonction renvoie pour une valeur positive uniquement, 1- la fonction de répartition, c'est à dire que la fonction sous Excel est définie par :

Si unilatéral (dernier paramètre=1), pour **t>0**

LOI.STUDENT(t,n,1) = $prob(T_n > t)$ où T_n désigne une variable de Student à n degrés de liberté

Si bilatéral (dernier paramètre =2) pour **t>0**

LOI.STUDENT(t,n,2) = $prob(|T_n| > t)$ où T_n désigne une variable de Student à n degrés de liberté

Dans le cas du test unilatéral gauche, il suffira d'utiliser la fonction avec comme premier paramètre l'opposé de la valeur standardisé.

Ici on utilisera directement la formule, ce qui donnera :

	A	B
14	Valeur standard	=(B4-B7)/(B5/RACINE(B3))
15	Niveau Signification	=LOI.STUDENT(B14;B3-1;1)

La valeur du niveau de signification obtenue *ns*=0,0019 qui est bien inférieur à 1%.

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.*

4.4 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 programmées dans Excel.

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.*

Les formules Excel définissant c_1 et c_2 sont les suivantes :

$$c_1 = \mu_0 \text{ -LOI.STUDENT.INVERSE}(\alpha,\text{n-1})* \hat{\sigma}/\sqrt{n}$$

$$c_1 = \mu_0 \text{ -LOI.STUDENT.INVERSE}(\alpha,\text{n-1}) * \hat{\sigma}/\sqrt{n}$$

4.4.1 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.

Etant donné la forme de la fonction de Student sous Excel, ce niveau de signification sera obtenu facilement :

	A	B
19	Nb Observations	250
20	Moyenne Echantillon	312
21	Ecart-type	52
22	Valeur μ0	320
23	**Test Bilatéral**	
24	Valeur standard	=ABS(B20-B22)/(B21/RACINE(B19))
25	Niveau Signification	=LOI.STUDENT(B24;B3-1;2)

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 *.*

5 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 (dans le fichier *Standard.xls*, sur la feuille *Comparaison*, nous avons les résultats par date de distribution, les valeurs estimées) :

	A	B	C	D	E
1	**Semaine**			**Week-End**	
2					
3	Taille échantillon	600		Taille échantillon	400
4	Nbre de retours	264		Nbre de retours	188
5	p1 chapeau	44%		p2 chapeau	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.

5.1 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 que nous verrons plus loin).

5.1.1 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) :

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

5.1.2 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.

5.2 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.

5.3 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.

5.3.1 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.

5.3.2 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 suivi par Z est de moyenne nulle sous l'hypothèse H_0.

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

5.3.3 Utilisation d'Excel

Sous Excel, nous avons la possibilité d'utiliser soit la loi normale, soit la loi normale centrée réduite (nommée standard sous Excel), pour le test bilatéral nous donnerons les formules utilisant la loi normale, et pour le test unilatéral nous utiliserons la loi normale standard.

Sur la feuille de calcul Excel nous calculons tout d'abord l'estimation "la meilleure" sous l'hypothèse H_0, puis l'écart type de la loi normale suivie par Z, ce qui nous permettra de calculer alors la valeur critique pour un risque de première espèce donné ou/et le niveau de signification du test. Les formules sont les suivantes :

	A	B
6	***Valeurs sous H0***	
7	pchapeau	=(B5*B3+E5*E3)/(B3+E3)
8	sigma chap	=RACINE(B7*(1-B7))
9	sigma chap(1/n1+1/n2)	=B8*RACINE(1/B3+1/E3)
10		
11	**Test bilatéral**	
12	alpha	0,05
13	Valeur critique	=LOI.NORMALE.INVERSE(1-B12/2;0;B9)
14	Niveau Signif	=(1-LOI.NORMALE(ABS(E5-B5);0;B9;VRAI))*2

Rappel : le dernier paramètre de la fonction LOI.NORMALE (ici VRAI) indique que l'on utilise la loi cumulée.

Les valeurs obtenues sont alors :

	A	B
11	**Test bilatéral**	
12	alpha	5%
13	Valeur critique	6,30%
14	Niveau Signif	35,04%

On ne pourra donc pas rejeter l'hypothèse H_0, au risque de 5% puisque l'écart observé est de 47%-44%=3% donc inférieur à 6,3%. On voit d'ailleurs par le niveau de signification, que si l'hypothèse H_0 est vraie, plus de 35% des échantillons pourraient conduire à un écart absolu supérieur à celui observé ici.

5.4 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.

5.4.1 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.

5.4.2 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.

5.4.3 Utilisation d'Excel

Comme nous l'avons annoncé, nous utiliserons dans ce paragraphe la loi normale standard, c'est à dire centrée réduite.

Nous ne reprendrons pas ici le calcul intermédiaire de l'estimation du pourcentage commun, les formules spécifiques du test unilatéral sont les suivantes :

	D	E
11	**Test unilatéral (p2>p1)**	
12	alpha	0,05
13	Valeur critique	=B9*LOI.NORMALE.STANDARD.INVERSE(1-E12)
14	Niveau Signif	=(1-LOI.NORMALE.STANDARD((E5-B5)/B9))

Remarque : étant donné les résultats obtenus sur l'échantillon, il est plus "naturel" ici de tester $p_2>p_1$ plutôt que l'inverse..

Les valeurs obtenues sont alors :

	D	E
11	**Test unilatéral (p2>p1)**	
12	alpha	5%
13	Valeur critique	5,28%
14	Niveau Signif	17,52%

On ne pourra donc pas rejeter l'hypothèse H_0, au risque de 5%, puisque l'écart observé (3%) est inférieur à la valeur critique 5,28%. On voit d'ailleurs par le niveau de signification que si l'hypothèse H_0 est vraie, plus de 17,5% des échantillons pourraient conduire à un écart, entre l'estimation de p_2 et celle de p_1,supérieur à 3%.

6 Comparaison de deux moyennes

Reprenons l'exemple de Monsieur Durlan, rassuré sur la rentabilité de son produit, il s'interroge sur le rayon où celui-ci doit être vendu ; en effet en regardant les résultats des magasins tests, il a constaté que certains le vendait avec les produits laitiers et d'autres avec les produits frais (voir la feuille Echantillon du fichier Ptidej.xls). A son avis le choix du rayon produits frais est préférable pour ce type de produit. Dans un premier temps, utilisant les fonctions base de données d'Excel, il obtient les résultats suivants :

	A	B	C	D	E	F	G
1	**Produit frais**		Rayon		**Produits laitiers**		Rayon
2			PF				PL
3							
4	Nb observations	182			Nb observations	150	
5	Moyenne	334,30			Moyenne	320,95	
6	Ecart type	51,99			Ecart type	50,83	

La moyenne des ventes en rayon "produits frais" est effectivement supérieure à celle des ventes en rayon "produits laitiers", cependant la différence est-elle suffisamment importante

pour pouvoir extrapoler ce résultat à l'ensemble de la population, c'est à dire à l'ensemble des supermarchés qui vendront bientôt ce produit. Ce problème est un peu plus compliqué que le problème de pourcentage dans la mesure où interviennent ici les dispersions (écart type) des ventes dans chacun des rayons.

6.1 Formalisation du problème

Nous pouvons ici encore présenter la formalisation de deux façons différentes, soit comme la comparaison de moyennes sur deux populations, soit comme l'étude d'une liaison entre deux variables (une variable quantitative et une variable indicatrice) définies sur une même population (cas particulier de la liaison entre une variable quantitative et une variable qualitative que nous reverrons lors de la régression).

6.1.1 Formalisation sous forme de deux populations

La première population est l'ensemble des rayons "produits frais" des supermarchés que nous noterons P_1, la seconde est l'ensemble des rayons "produits laitiers" notée P_2. Sur chacune de ces populations nous définissons une variable quantitative, notées respectivement X_1 et X_2, qui correspond au volume hebdomadaire des ventes.

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

en désignant par μ_1 et μ_2 les espérances respectives, c'est à dire les moyennes des variables X_1 et X_2 sur chacune des populations (nous noterons $\sigma 1$ et $\sigma 2$ les écarts types), l'hypothèse nulle s'exprime alors sous la forme :

$$H_0 \quad \mu_1 = \mu_2$$

l'hypothèse alternative dans le cas de M Durlan est simplement la préférence pour le rayon "produits frais" (test unilatéral), mais pourrait être simplement un comportement différent (test bilatéral) :

$$H_1 \quad \mu_1 > \mu_2 \qquad \text{ou} \quad \mu_1 \neq \mu_2$$

Nous supposerons de plus que les deux variables suivent une loi normale.

6.1.2 Formalisation à l'aide de deux variables

Dans ce cas la population P unique est l'ensemble des supermarchés où sera distribué le nouveau produit, quelque soit le rayon, la variable X est toujours la variable quantitative du volume des ventes hebdomadaire, et nous allons introduire une nouvelle variable indicatrice Y du rayon : cette variable vaut 1 pour le rayon "produits frais" et 0 pour le rayon "produits laitiers". Le problème de M Durlan se résume à savoir s'il existe une forme de dépendance entre ces variables, une fois la population munie d'une loi de probabilité uniforme ; la loi de X est de plus supposée normale.

Les hypothèses portent uniquement dans la formulation de M Durlan sur les espérances conditionnelles, et non pas sur les deux paramètres. En effet, la moyenne μ_1 représente l'espérance de X sachant Y=1, de même la moyenne μ_2 représente l'espérance de X sachant Y=0.

L'hypothèse H_0 revient alors à écrire :

$$\mu_1 = E(X/Y = 1) = E(X = 0/Y = 0) = \mu_2$$

Ce qui est peut s'interpréter comme une "indépendance" en moyenne.

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

Dans la suite nous utiliserons la formalisation en termes de deux populations, la deuxième formalisation sera généralisée aux variables qualitatives lors de la régression (et en ajoutant une hypothèse supplémentaire sur les variances).

6.2 Statistique associée au test

L'hypothèse nulle peut aussi s'écrire

$$H_0 \quad \mu_1 - \mu_2 = 0$$

Sur un échantillon de taille n_1 tiré de la population P_1, le paramètre μ_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 μ_2 sera $\overline{X}^2_{n_2}$; la statistique utilisée sera donc la variable aléatoire $Z = \overline{X}^1_{n_1} - \overline{X}^2_{n_2}$. En faisant l'hypothèse de normalité des lois, nous connaissons théoriquement la loi de Z :

$$Z \longrightarrow N(\mu, \sigma') \quad \text{avec } \mu = \mu_1 - \mu_2 \quad \text{et} \quad \sigma'^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

sous l'hypothèse H_0, en désignant, nous aurons donc $\mu = 0$.

Malheureusement les écarts type ne sont pas connus et nous allons être conduits à faire une hypothèse sur ceux ci, pour pouvoir mener à bien le test. Nous aurons une connaissance exacte de la loi de la statistique utilisée uniquement dans un cas, le cas d'égalité des variances. Nous indiquons en annexe comment tester éventuellement cette égalité

6.2.1 Egalité des variances(homoscédascité)

Si on ajoute l'hypothèse $\sigma_1 = \sigma_2 = \sigma$, nous pouvons regrouper les deux estimateurs de cette valeur commune, pour obtenir un estimateur de variance inférieure, donc plus précis, en tenant compte des définitions vues au chapitre précédent, nous utiliserons :

$S^2_{n_1+n_2-2} = \dfrac{(n_1-1)S^2_{n_1} + (n_2-1)S^2_{n_2}}{n_1+n_2-2}$, alors $T = \dfrac{Z}{\sqrt{S^2_{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ suit une loi de Student à

$n_1 + n_2 - 2$ degrés de liberté.

6.2.2 Inégalités des variances(hétéroscédascité)

Si nous ne faisons plus l'égalité des variances, une solution simple (voire simpliste) consiste à considérer que les échantillons sont suffisamment grands pour pouvoir remplacer les écarts type réels par leurs estimations et donc utiliser la loi normale. Remarquons que cette solution est d'ailleurs la seule réellement applicable si l'on ne fait pas l'hypothèse de normalité des lois X_1 et X_2 sur les populations.

Toutefois, il est possible sous l'hypothèse de normalité, d'avoir une meilleure approximation en utilisant la statistique $T = \dfrac{Z}{\sqrt{\dfrac{S^2_{n_1}}{n_1} + \dfrac{S^2_{n_2}}{n_2}}}$ qui suit une loi de Student dont l'approximation du nombre de degrés de liberté est donnée par la formule de Satterthwaite :

$$dl = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)}$$ [5]

Cette formule est utilisée par les logiciels statistiques tels que SPSS ou SAS, c'est pourquoi nous l'utiliserons aussi.

6.3 Test unilatéral

Dans ce cas l'hypothèse alternative est $H_1 \quad \mu_1 > \mu_2$, 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.

6.3.1 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 de *Z* sous l'hypothèse H_0. On se ramènera à la loi de Student en divisant par l'estimateur de l'écart type de Z suivant l'hypothèse faite sur l'égalité des variances. En notant $t_{1-\alpha}$ le fractile de la de Student associée, on a alors :

- En cas d'égalité des variances : $c = t_{1-\alpha} * s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ où *s* désigne l'estimation "regroupée" de $\sigma_1 = \sigma_2$ qui est calculé par la formule $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 1}}$, la loi de Student étant à $n_1 + n_2 - 1$ degrés de liberté.
- En cas d'inégalité de variance : $c = t_{1-\alpha}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$, le nombre de degrés de liberté étant donné par la formule de Satterthwaite.

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.

6.3.2 Calcul du niveau de signification

Le niveau de signification est dans ce cas la probabilité, sous l'hypothèse H_0, d'observer un écart entre les deux estimateurs qui soit en valeur absolu au moins égal à l'écart absolu observé sur les échantillons :

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

Ou encore en se ramenant en divisant par l'écart type convenable à la loi de Student :

[5] Satterthwaite, FW "An approximate Distribution of Estimate of Variance Components", Biometrics Bulletin,2, 110-114 (1946)

$ns = 1 - prob\left(T < \frac{\bar{x}_1 - \bar{x}_2}{\sigma'}\right)$ avec le nombre convenable de degrés de liberté.

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

6.3.3 Utilisation d'Excel

Nous présenterons ici les résultats dans les trois cas : égalité de variance, inégalité de variance. Rappelons qu'Excel donne toujours la fonction de répartition symétrique de la loi de Student. Les tailles et estimations des moyennes et écarts types des deux échantillons sont données au début du paragraphe 6, pour leur localisation dans la feuille.

1) Egalité des variances

	A	B
7		
8	**Variances égales**	
9		
10	Variance commune	=((B4-1)*B6*B6+(F4-1)*F6*F6)/(B4+F4-2)
11	Degrés de liberté	=B4+F4-2
12	Ecart type Z	=RACINE(B10*(1/B4+1/F4))
13		
14	Risque de 1° espèce	0,05
15	Valeur critique	=B12*LOI.STUDENT.INVERSE(2*B14;B11)
16		
17	Valeur standard	=($B$5-$F$5)/B12
18	Niveau signification	=LOI.STUDENT(B17;B11;1)

Remarque : Nous avons décomposé les formules de façon à pouvoir facilement les copier pour le cas d'inégalité des variances. L'écart type de Z représente le dénominateur de la loi de Student ; le 2*B14 qui apparaît dans la formule de la cellule C15 est du à la définition de la fonction LOI.STUDENT.INVERSE d'Excel qui est symétrique ; enfin le troisième paramètre 1 de la fonction LOI.STUDENT indique le cumul. La valeur standard représente la différence entre les deux moyennes estimées divisée par l'écart type de Z.

Les valeurs obtenues sont alors :

	A	B
8	**Variances égales**	
9		
10	Variance commune	2649,03
11	Degrés de liberté	330
12	Ecart type Z	5,68
13		
14	Risque de 1° espèce	0,05
15	Valeur critique	***9,36***
16		
17	Valeur standard	2,353
18	Niveau signification	***0,0096***

On pourra donc rejeter l'hypothèse H_0, au risque de 5%, puisque l'écart observé (334,30-320,95=13,35) est supérieur à la valeur critique 9,36. On voit d'ailleurs par le niveau de signification que si l'hypothèse H_0 est vraie, moins de 1% des échantillons pourraient conduire à un écart, entre l'estimation de p_2 et celle de p_1,supérieur à 13,35.

2) Variances inégales

Avec les mêmes conventions et notations que précédemment on a les formules :

	A	B
21	**Variances inégales**	
22		
23	Degrés de liberté	=((B6^2/B4)+(F6^2/F4))^2/((B6^2/B4)^2/(B4-1)+(F6^2/F4)^2/(F4-1))
24	Ecart type Z	=RACINE(B6^2/B4+F6^2/F4)
25		
26	Risque de 1° espèce	0,05
27	Valeur critique	***=B24*LOI.STUDENT.INVERSE(2*B26;B23)***
28		
29	Valeur standard	=($B$5-$F$5)/B24
30	Niveau signification	***=LOI.STUDENT(B29;B23;1)***

Les valeurs obtenues sont alors :

	A	B
21	**Variances inégales**	
22		
23	Degrés de liberté	320,52
24	Ecart type Z	5,66
25		
26	Risque de 1° espèce	0,05
27	Valeur critique	***9,34***
28		
29	Valeur standard	2,358
30	Niveau signification	***0,0095***

On pourra donc rejeter l'hypothèse H_0, au risque de 5%, puisque l'écart observé (334,30-320,95=13,35) est supérieur à la valeur critique 9,34. On voit d'ailleurs par le niveau de signification que si l'hypothèse H_0 est vraie, moins de 1% des échantillons pourraient conduire à un écart, entre l'estimation de p_2 et celle de p_1,supérieur à 13,35.

Remarquons enfin que sur des tailles d'échantillon "raisonnables" comme celles que nous avons ici, il n'y a que peu de différence entre les résultats sous les deux hypothèses d'égalité ou d'inégalité des variances, et il serait tout à fait possible d'utiliser directement la loi normale en remplaçant les écarts types théoriques par leurs estimations (exercice laissé au lecteur).

6.4 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. Nous supposerons ici que les tailles d'échantillons sont suffisamment grandes pour pouvoir utiliser l'approximation normale directement, nous libérant ainsi de l'hypothèse de la normalité des lois sous jacentes.

Le lecteur pourra facilement passer du cas unilatéral au cas bilatéral pour les lois de Student.

6.4.1 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 :

$c = z_{1-\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}$ où s_1 et s_2 désignent les estimations des écarts types de X_1 et X_2.

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.

6.4.2 Calcul du niveau de signification

Le niveau de signification est dans ce cas la probabilité, sous l'hypothèse H_0, d'observer un écart entre les deux estimateurs qui soit en valeur absolu au moins égal à l'écart absolu observé sur les échantillons :

$$ns = prob\left(|Z| >= |\bar{x}_1 - \bar{x}_2|\right) = \left(1 - prob\left(Z < |\bar{x}_1 - \bar{x}_2|\right)\right) * 2$$

Puisque la loi normale suivi par Z est de moyenne nulle sous l'hypothèse H_0.

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

6.4.3 Utilisation d'Excel

Sous Excel, nous avons la possibilité d'utiliser soit la loi normale, soit la loi normale centrée réduite (nommée standard sous Excel), pour le test bilatéral nous donnerons les formules utilisant la loi normale.

Sur la feuille de calcul Excel nous calculons tout d'abord l'estimation de l'écart type de la loi normale suivie par Z, ce qui nous permettra de calculer alors la valeur critique pour un risque de première espèce donné ou/et le niveau de signification du test. Les formules sont les suivantes :

	E	F
21	**Test bilatéral (Loi Normale)**	
22		
23	Ecart type Z	=RACINE(B6^2/B4+F6^2/F4)
24		
25	Risque de 1° espèce	0,05
26	Valeur critique	***=LOI.NORMALE.INVERSE(1-F25/2;0;F23)***
27		
28	Ecart constaté	=ABS($B$5-$F$5)
29	Niveau signification	***=2*(1-LOI.NORMALE(F28;0;F23;VRAI))***

Les valeurs obtenues sont alors :

	E	F
21	**Test bilatéral (Loi Normale)**	
22		
23	Ecart type Z	5,66
24		
25	Risque de 1° espèce	0,05
26	Valeur critique	***11,10***
27		
28	Ecart constaté	13,356
29	Niveau signification	***0,0184***

On pourra rejeter l'hypothèse H_0, au risque de 5% puisque l'écart observé est de 13,35 donc supérieur à 11,10. On voit d'ailleurs par le niveau de signification, que le risque de première espèce assumé est au plus de 1,84%. Cette dernière valeur était attendue, elle correspond

environ au double du niveau de signification du test unilatéral (environ du à l'utilisation de la loi normale et non de la loi de Student).

6.5 La fonction TEST.STUDENT

Il existe sous Excel une fonction nommée TEST.STUDENT, qui permet de déterminer le niveau de signification du test de comparaison des moyennes, si l'on dispose des données dans deux zones matricielles distinctes.

La syntaxe de cette fonction est la suivante :

TEST.STUDENT(matrice1;matrice2;uni/bilatéral;type)

- matrice 1 représente la zone où sont stockées les données du premier échantillon
- matrice 2 représente la zone où sont stockées les données du deuxième échantillon
- uni/bilatéral vaut 1 pour un test unilatéral, 2 pour bilatéral
- type peut prendre 3 valeurs :
  - 1 pour un test dit "pairé", on utilise la variable aléatoire égale à la différence des deux variables, ce qui suppose que cette différence ait un sens et que le nombre d'observations des deux échantillons soit le même.
  - 2 en cas d'égalité des variances
  - 3 en cas d'inégalité des variances.

7 Test du KHI-DEUX

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 (fichier Tchi2.xls) : le fabricant de shampoing DIP, veut déterminer quels sont les critères de choix d'un shampoing suivant les catégories d'ages, de façon plus précise il veut savoir si ces critères différent suivant les tranches d'ages. 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'age et le lieu d'achat habituel du consommateur.

7.1 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'age.

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

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

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.

7.2 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

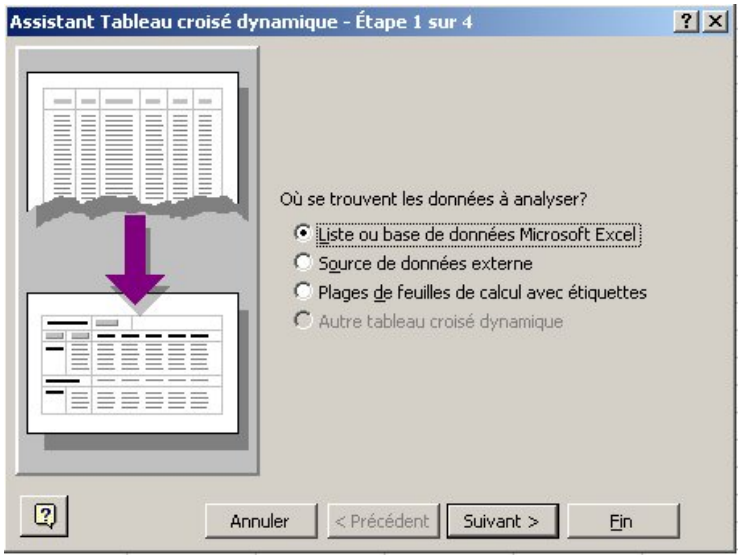
Seules les cellules grisées différent du tableau de contingence observé sur l'échantillon, si ces deux tableaux sont suffisamment différents nous rejetterons l'hypothèse H_0. Il nous faut donc définir une distance entre tableau et connaître la loi de cette distance sous l'hypothèse nulle, pour appliquer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

7.3 Construction des tableaux sous Excel

Si l'on dispose des données brutes comme c'est le cas ici (feuille Enquête), il faut tout d'abord construire le tableau de contingence observé. Pour cela on peut soit utiliser les tables (cf. le chapitre rappel Excel), soit utiliser la commande "Tableau Croisé dynamique" du menu Données, que nous allons illustrer ici.

La cellule active étant une des cellules de données, pour qu'Excel détermine lui-même la zone de données, nous choisissons donc la commande Données, puis Rapport de Tableau Croisé dynamique ; l'assistant va alors nous guider dans le choix des différents éléments.

Tout d'abord nous devons indiquer à partir de quelles données doit être construit le tableau croisé :



Nous confirmons le choix par défaut (Liste ou base de données) en cliquant sur suivant. Si la cellule active est dans la zone de données l'étape suivante est simplement une confirmation de la plage de données (sinon il faudra alors indiquer cette plage) ; nous passons directement à l'étape suivante qui est la création du tableau croisé.

Cette création se fait en précisant la variable en ligne, la variable en colonne et le contenu des cases du tableau, ici le nombre des individus. Il suffit de faire glisser les champs apparaissant à droite de la boîte de dialogue à leur place dans le tableau croisé (figure 1), puis de glisser à l'intérieur du tableau le champ correspondant à une variable qualitative :

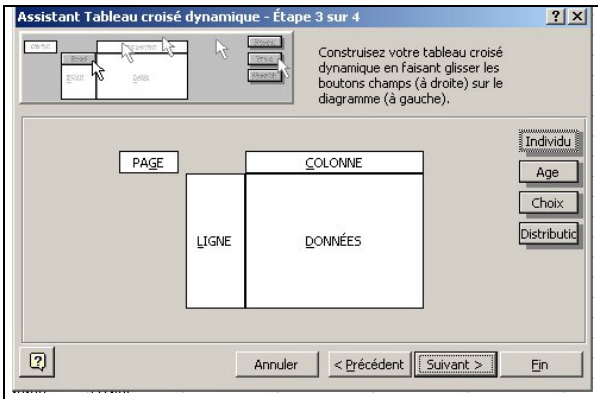


Figure 1

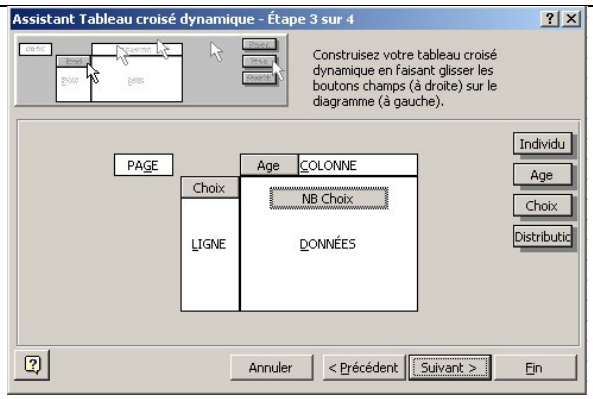


Figure2

Si la variable est quantitative, Excel propose la somme des valeurs de cette variable pour chacun des couples de modalité, en double cliquant sur l'étiquette intérieure au tableau il est possible de modifier cette caractéristique.

En cliquant sur "Suivant", on obtient une dernière boite de dialogue qui permet de choisir où sera créé le rapport, nous choisirons l'option "Nouvelle feuille" et terminerons la création du tableau croisé, ce qui nous donne le résultat suivant sur une feuille qui a été renommée "Choix-Age" :

	A	B	C	D	E
1	NB Choix	Age			
2	Choix	<25	>65	25-45	Total
3	Distribution	63	50	91	204
4	Marque	28	66	8	102
5	Odeur	76	25	52	153
6	Texture	12	33	31	76
7	Total	179	174	182	535

Remarque : contrairement à ce que l'on obtient par les tables d'hypothèse, ce tableau ne contient aucune formule, mais uniquement des valeurs (pour les marges aussi).

Il est alors facile d'obtenir le tableau théorique sous l'hypothèse H_0, par les formules suivantes obtenues par recopie de l'une d'entre elles :

	A	B	E
9	Théorie	<25	Total
10	Distribution	=$E3*B$7/$E$7	=SOMME(B10:D10)
11	Marque	=$E4*B$7/$E$7	=SOMME(B11:D11)
12	Odeur	=$E5*B$7/$E$7	=SOMME(B12:D12)
13	Texture	=$E6*B$7/$E$7	=SOMME(B13:D13)
14	Total	=SOMME(B10:B13)	=SOMME(E10:E13)

ce qui donne les valeurs des effectifs théoriques :

	Zone Nom	B	C	D	E
9	Théorie	<25	>65	25-45	Total
10	Distribution	68,3	66,3	69,4	204
11	Marque	34,1	33,2	34,7	102
12	Odeur	51,2	49,8	52	153
13	Texture	25,4	24,7	25,9	76
14	Total	179	174	182	535

Il est clair dans la mesure où les valeurs ne sont pas entières, ce tableau théorique ne peut évidemment pas être observé. Il nous faut savoir si l'écart entre le tableau observé et le tableau théorique doit être attribué aux aléas de l'échantillonnage ou à une dépendance structurelle entre les variables. Ceci va se faire en définissant une distance entre les tableaux.

7.4 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 signification, en calculant la valeur de la statistique $\hat{d}^2$ sur l'échantillon, le niveau de signification 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 signification est inférieur à α, dans ce cas le risque d'erreur est inférieur ou égal à *ns*.

7.5 Mise en œuvre du test sous Excel

Pour calculer la valeur critique, il suffit d'utiliser la fonction KHIDEUX.INVERSE d'Excel, qui retourne la valeur critique c pour un risque de première espèce donné a. La syntaxe est la suivante :

KHIDEUX.INVERSE(alpha; degrés de liberté)

Il faut alors calculer la statistique sur l'échantillon, voici les formules correspondantes (les colonnes C et D ont été masqué), la valeur de la statistique est dans le coin inférieur droit du tableau, chaque case contient la différence entre l'effectif théorique et l'effectif observé au carré divisée par l'effectif théorique. La statistique est simplement la somme de toutes les cases du tableau :

	A	B	E
16	alpha		0,05
17	Valeur critique		=KHIDEUX.INVERSE(E16;6)
18			
19	**Distance**	<25	
20	Distribution	=(B10-B3)^2/B10	
21	Marque	=(B11-B4)^2/B11	
22	Odeur	=(B12-B5)^2/B12	
23	Texture	=(B13-B6)^2/B13	
24			**=SOMME(B20:D23)**

ce qui conduit aux valeurs numériques :

	A	B	C	D	E
16	alpha				5%
17	Valeur critique				12,59
18					
19	**Distance**	<25	>65	25-45	
20	Distribution	0,4045	4,03	6,72	
21	Marque	1,1	32,5	20,5	
22	Odeur	12,024	12,3	0	
23	Texture	7,0911	2,78	1,02	
24					**100,5**

En appliquant la règle de décision, comme 100,5>12,59 on rejette l'hypothèse H0 avec un risque de première espèce inférieur à 5%.

Pour calculer le niveau de signification, on dispose de deux fonctions, l'une utilise directement les tableaux, l'autre la valeur de la statistique calculée. La fonction **TEST.KHIDEUX** évite le calcul de la statistique, elle retourne directement le niveau de signification avec comme paramètre les deux tableaux : le théorique puis l'observé. La syntaxe est ***TEST.KHIDEUX(théorique, observé)*** sur l'exemple :

TEST.KHIDEUX(B3:D6;B10:D13)

Attention à l'ordre des paramètres!

L'autre méthode consiste à utiliser la fonction ***LOI.KHIDEUX(valeur, DL)*** qui retourne la probabilité pour qu'une loi du CHI2 à DL degrés de liberté soit supérieure à valeur**.** Cette fonction demande bien sur d'avoir calculé la statistique sur l'échantillon, ici la formule est donc **LOI.KHIDEUX(E24;6)**.

Dans les deux cas on trouve comme valeur ns =1,957E-19, on peut donc rejeter l'hypothèse H_0 avec un risque quasi nul (inférieur à 210^{-19}).

8 Annexe : Comparaison de deux variances

Nous allons indiquer ici succinctement la procédure de test d'égalité de deux variances, l'hypothèse alternative étant le fait que les variances sont différentes, les cas unilatéraux étant laissés au lecteur dans la mesure où ils sont très rarement utilisés dans la pratique.

Nous considérons deux variables quantitatives X_1 et X_2 définies sur deux populations P_1 et P_2(comme dans le paragraphe 6- comparaison de deux moyennes), nous supposerons de plus que ces deux variables suivent une loi normale d'écart type respectif σ_1 et σ_2.

L'hypothèse nulle et l'hypothèse alternative sont respectivement :

$$H_0 \quad \sigma_1 = \sigma_2$$
$$H_1 \quad \sigma_1 \neq \sigma_2$$

On utilisera l'hypothèse nulle sous la forme $\sigma_1^2 / \sigma_2^2 = 1$. L'hypothèse alternative peut alors s'écrire sous la forme $\sigma_1^2 / \sigma_2^2 > 1$ ou $\sigma_2^2 / \sigma_1^2 < 1$, soit encore $\max\left(\sigma_1^2 / \sigma_2^2, \sigma_2^2 / \sigma_1^2\right) > 1$.

Sur un échantillon de taille n_1 de la population P_1, l'estimateur de la variance est la statistique que nous avons notée $S_{n_1}^2$ et nous savons que $(n_1 - 1)S_{n_1}^2 / \sigma_1^2$ suit une loi du Chi2 à $(n_1 - 1)$ degrés de liberté, si la loi de X_1 est une loi normale. De même, sur un échantillon de taille n_2 de la population P_2, l'estimateur de la variance est la statistique que nous avons notée $S_{n_2}^2$ et nous savons que $(n_2 - 1)S_{n_2}^2 / \sigma_2^2$ suit une loi du Chi2 à $(n_2 - 1)$ degrés de liberté (voir le chapitre sur l'estimation).

Pour le test nous allons utiliser la statistique $\dfrac{S_{n_1}^2}{S_{n_2}^2}$, dont la loi est connue sous l'hypothèse H0, car alors les deux variances sont égales donc les deux dénominateurs rappelés ci-dessus le sont aussi. Cette loi est la loi de Fisher-Snedecor à $(n_1 - 1, n_2 - 1)$ degrés de liberté, nous noterons $FS_{n,p}$ la loi générique à (n,p) degrés de liberté. D'après la définition même de cette loi, on peut voir que :

$$pour\ f > 1 \qquad prob\left(FS_{n,p} > f\right) = prob\left(FS_{p,n} < 1/f\right)$$

puisque changer le couple (n,p) en (p,n) revient simplement à inverser la fraction définissant la loi.

8.1 Détermination de la valeur critique

8.1.1 Les formules

La valeur critique c de rejet de l'hypothèse H_0 est déterminé par l'équation :

$$prob\left(\frac{S_{n_1}^2}{S_{n_2}^2} < \frac{1}{c}\right) + prob\left(\frac{S_{n_1}^2}{S_{n_2}^2} > c\right) = \alpha$$

en utilisant la remarque faite à la fin du paragraphe précédent, nous obtenons :

$$prob\left(\frac{S_{n_1}^2}{S_{n_2}^2} > c\right) = \alpha/2$$

La règle de décision est alors la suivante : si s_1 et s_2 sont les écarts type estimés sur les échantillons, on rejettera l'hypothèse H_0 avec un risque d'erreur inférieur à α, si :

$$\max\left(\frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2}\right) > c$$

sinon on conservera l'hypothèse H_0, sans connaître le risque d'erreur.

8.1.2 Utilisation d'Excel

Sous Excel nous pouvons utiliser la fonction INVERSE.LOI.F(proba;DL1;DL2)=*f* où *f* est définie par $prob(FS_{DL1,DL2} > f) =$ proba . Pour un risque de première espèce donné α, il suffira donc de donner à proba la valeur $\alpha/2$. Sur l'exemple du paragraphe 6, nous avons les formules et valeurs numériques suivantes :

	E	F
8	**Test Variances**	
9		
10	Risque 1° espèce	0,05
11	Valeur critique	=INVERSE.LOI.F(F10/2;B4-1;F4-1)
12	Valeur observée	=MAX((B6/F6)^2;(F6/B6)^2)

Formules

	E	F
8	**Test Variances**	
9		
10	Risque 1° espèce	5%
11	Valeur critique	1,363
12	Valeur observée	1,046

Valeurs

Comme la valeur critique est inférieure à la valeur observée, nous ne pouvons pas rejeter l'hypothèse H_0 au risque de 5%, nous conserverons donc l'hypothèse d'égalité des variances.

8.2 Calcul du niveau de signification

8.2.1 Les formules

Nous noterons $\hat{f} = \max\left(\frac{s_1^2}{s_2^2}; \frac{s_2^2}{s_1^2}\right)$, la valeur observée sur l'échantillon le niveau de signification est la probabilité d'observer une valeur au moins égale à $\hat{f}$ sous l'hypothèse H_0. Cette probabilité peut s'écrire :

$$ns = 2 * prob\left(FS_{n1-1,n2-1} > \hat{f}\right)$$

La règle de décision consiste à rejeter l'hypothèse H_0, si le niveau de signification *ns* est inférieur au risque de première espèce α.

8.2.2 Utilisation d'Excel

Pour calculer le niveau de signification, on dispose sous Excel de deux fonctions selon que l'on dispose des données brutes ou seulement des résumés.

A partir des données brutes on utilisera la fonction TEST.F(echan1,echan2) où echan1 et echan2 désigne les zones ou sont stockées les données des deux échantillons. Cette fonction retourne directement le niveau de signification du test.

A partir des résumés, ce sera la fonction LOI.F(fobservé,DL1,DL2) qui sera utilisée; cette fonction renvoie la probabilité d'obtenir une valeur supérieure ou égale à fobservé pour une loi de Fisher-Snedecor à (DL1,DL2) degrés de liberté.

Sur l'exemple les formules et les valeurs sont :

	E	F
8	**Test Variances**	
9		
12	Valeur observée	=MAX((B6/F6)^2;(F6/B6)^2)
13	Niveau de signification	=2*LOI.F(F12;B4-1;F4-1)
14	FonctionTest.F	=TEST.F(Echantillon!B2:B183;Echantillon!B184:B333)

Formules

	E	F
8	**Test Variances**	
9		
12	Valeur observée	1,046
13	Niveau de signification	0,779019
14	FonctionTest.F	0,779019

Valeurs

Il y a ici plus de 77% de chances d'observer une telle valeur de $\hat{f}$ sous l'hypothèse H_0, on ne rejette donc pas l'hypothèse nulle au risque de 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%.

1 Taux de phosphate (phos.xls)

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 "phos.xls".

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?

2 AntiSmoke(tabac.xls)

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.xls", la reprise=1 indique que le fumeur a rechuté avant la fin du mois sinon il est indiqué 0.

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?

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%.

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 ?

4 Contrôle de qualité (quali.xls)

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.xls", ainsi que le numéro de la machine ayant fabriqué la pièce.

5. Définissez la population, la variable et le paramètre concernés par l'analyse.
6. Formulez le test du fabricant
7. Le produit vous paraît satisfaisant au point de vue résistance?
8. Peut-on faire une différence suivant la machine ayant servi à la fabrication?

5 La société LOCVIDEO (fichier Videos.xls)

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?

4. Que conseilleriez-vous à LOCVIDEO?

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% [6]. La taille retenue pour chaque échantillon est de 4 900.

Les deux sondages ont donné les résultats suivants :

	Formule F1	Formule F2
Nombre d'acheteurs	801	914

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

7 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 340F. 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

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

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
85140 F	330 F	165 F

8 Exercice 8 : La société Bricoplus

La société Bricoplus a lancé pendant un mois une campagne publicitaire avec bons de réduction dans la presse régionale. Le coût de la campagne a été de 1000KF. 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	308	293	300,88
Ecart-type(Valeur)	207,6	191,2	200

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 avec bon est-il égal au montant moyen des commandes sans bon ? (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 terme de rentabilité, il envisage même une diminution de profit. Sachant que le Chiffre d'affaires mensuel avant la campagne était d'environ 4500000F 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 ?

9 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 ?

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

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) ?

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

LA REGRESSION LINEAIRE

1 Un exemple (fichier Pubradio.xls)

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?

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..

2.1 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.

2.2 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.

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.

3.1 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. Dans la mesure où nous utiliserons des fonctions spécialisées d'Excel pour la régression, nous ne ferons pas de distinction entre régression simple et 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.

3.2 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 appartement 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 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 s.

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.

3.3 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 une fonction d'Excel 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.

3.3.1 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 déterminer 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 : $\hat{e}_i = y_i - \hat{y}_i$

3.3.2 Interprétation géométrique du critère des moindres carrés

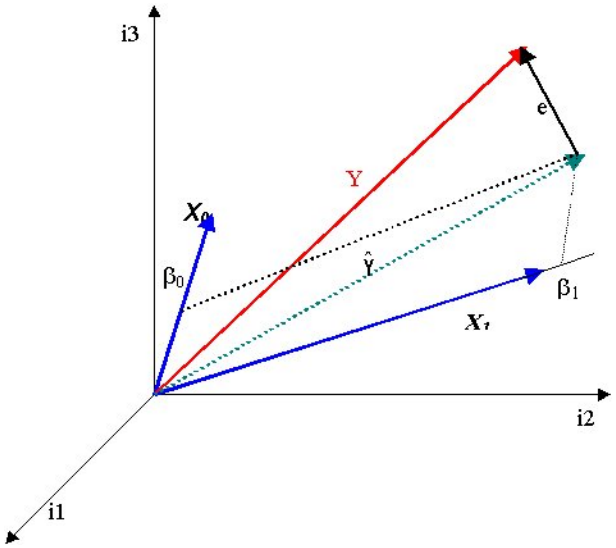
Nous allons interpréter géométriquement la méthode des moindres carrés, ce qui nous permettra d'expliciter certaines propriétés des estimations et estimateurs associés. Pour cela nous allons nous placer dans l'espace des individus, c'est à dire que nous allons considérer un espace vectoriel à n dimensions, 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).

3.3.3 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 nulles.

- 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

3.4 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, nous suivrons d'ailleurs cette présentation sous Excel au paragraphe suivant. Cette présentation est faite sous trois chapitres : indicateurs résumés, validité globale, validité marginale.

3.4.1 Résumés de la régression

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

1) **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 R2 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 régression 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):

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

2) 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.

3) 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.

3.4.2 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, soit par la fonction INVERSE.LOI.F d'Excel. 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			

Nous verrons plus loin comment construire ce tableau sous Excel.

3.4.3 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. Nous indiquerons comment procéder dans Excel. Dans ce cas il faut modifier en conséquence les degrés de liberté des résidus qui ne sont plus *n-p-1* mais *n-p*.

4 Utilisation d'Excel

Nous allons indiquer ici comment construire avec Excel les trois tableaux définis précédemment. La fonction de base permettant de construire ces tableaux est une fonction matricielle nommée DROITEREG, à partir des résultats de cette fonction, nous indiquerons les différentes formules conduisant à générer le listing résultat d'une régression.

Nous utiliserons le fichier Pubradio.xls, renommé pour ce paragraphe Pubradio1.xls, pour illustrer notre propos. Ce fichier comporte une première feuille nommée "Data" contenant les données dans la plage A1:D23. la première ligne de cette plage contient le nom des variables (Ventes, Radio, Journaux, Gratuits), dont les valeurs proprement dites sont dans la plage A2:D23. La colonne A correspond à la variable à expliquer, les autres colonnes aux variables explicatives. Nous nous fixerons un risque de première espèce de 5% pour interpréter les résultats.

Les noms donnés aux plages que nous utiliserons sont les suivants :

Nom	Contenu	Adresse
Xnom	Nom des variables explicatives	$B$1:$D$1
Xdonnees	Valeurs des variables explicatives	$B$2:$D$23
Ydonnees	Valeurs de la variable à expliquer	$A$2:$A$23

Attention : *dans Excel les variables explicatives doivent toujours être dans une zone rectangulaire (une plage) ne contenant pas de colonnes ou lignes vides*. On ne peut pas sélectionner les variables explicatives sur des plages disjointes (même en utilisant l'utilitaire d'analyse).

4.1 La fonction DROITEREG

La fonction DROITEREG d'Excel est une fonction matricielle qui donne tous les éléments permettant de construire un listing standard de régression. La plage contenant les résultats de la fonction est constituée (au maximum) de 5 lignes et p+1 colonnes, p désignant le nombre de variables explicatives. Les arguments de la fonction sont au nombre de 4 :

- La plage contenant les valeurs de la variable à expliquer (une seule colonne ou une seule ligne).
- La plage contenant les valeurs des variables explicatives, comme dit plus haut ces variables doivent être dans des colonnes (ou lignes) adjacentes.
- Un paramètre booléen (Constante) permettant de forcer à 0 la constante (auquel cas la plage de résultats de la fonction n'a plus besoin de comporter que *p* colonnes), si ce paramètre est omis ou vaut VRAI, la constante est incluse dans la régression. Pour nous ce paramètre sera toujours omis, dans la mesure où pour le modèle sans constante, les résultats fondamentaux *SCT=SCE+SCM* et $\sum e_i = 0$ ne sont plus vérifiés, les indicateurs alors utilisés R^2, f_c, t_c ne suivent plus les lois indiquées au paragraphe ci dessus.
- Un paramètre booléen indiquant si l'on veut ou non les statistiques, présentées au paragraphe précédent. Si ce paramètre vaut FAUX ou est omis seuls les coefficients de régression sont donnés en résultat, la plage de résultat ne peut alors contenir qu'une seule ligne. Le paramètre doit être mis à la valeur VRAI explicitement pour pouvoir créer un listing de régression.

Attention : Excel ne fait aucune vérification sur la dimension de la plage de résultats sélectionnée au moment de l'entrée de la formule, si cette plage est trop petite les résultats sont tronqués, par exemple certains coefficients n'apparaîtront pas s'il manque des colonnes, en revanche si la plage est trop grande, cela ne pose aucun problème autre qu'esthétique, dans la mesure où les résultats sont complétés pour remplir la plage par des #NA.

La plage de résultats est structurée de la façon suivante :

- La première ligne contient la valeur des estimations des p coefficients des variables explicatives (en ordre inverse de leurs colonnes dans la fonction) et le coefficient constant. La première valeur correspond au coefficient de la dernière variable explicative b_p, la seconde au coefficient de l'avant dernière variable etc.. Donc on a dans l'ordre les valeurs $(b_p, b_{p-1}, \ldots, b_1, b_0)$.

- La deuxième ligne donne les estimations des écarts typent des estimateurs des coefficients, dans le même ordre que les coefficients. Sur cette ligne nous avons donc $\left(s(B_p), s(B_{p-1}), \ldots, s(B_1), s(B_0)\right)$.

Seules les deux premières lignes ont un nombre d'éléments qui dépend du nombre de variables explicatives, les trois autres lignes comportent toujours exactement deux éléments.

- La troisième ligne contient le coefficient de détermination R^2 et l'erreur type de la régression (estimation de l'écart type des résidus).
- La quatrième ligne contient la valeur de la statistique de Fisher Snedecor globale (f_c) et le nombre de degrés de liberté des résidus (*n-p-1* si il y a une constante, *n-p* sinon).
- Enfin la dernière ligne contient la somme des carrés expliquée (*SCE*) et la somme des carrés résiduelle (*SCR*).

Rappel : pour entrer une formule matricielle, il faut sélectionner la zone de résultat (sur notre feuille $F$1:$I$5), entrer dans la cellule active la formule :

=DROITEREG(Ydonnees;Xdonnees;;VRAI)

puis valider, avec la touche ***Enter***, en maintenant les touches ***Ctrl*** et ⇧***Shift***. La formule est entrée dans l'ensemble de la zone sous la forme :

{=DROITEREG(Ydonnees;Xdonnees;;VRAI)}

Voici les résultats obtenus sur notre exemple :

	F	G	H	I
1	-0,61874	32,62939	23,85	238,4578
2	10,2281	5,368632	4,523787	112,2421
3	0,839445	138,0337	#N/A	#N/A
4	31,37041	18	#N/A	#N/A
5	1793130	342959,5	#N/A	#N/A

Notre modèle estimé s'écrit alors :

Ventes = 238,4578 + 23,85 Radio + 32,6294 Journaux – 0,6187 Gratuits + e

(ecart types) (112,2421) (4,5238) (5,3686) (10,2281) (138,0337)

La deuxième ligne donnant les écart types estimés des coefficients et du terme aléatoire. Avec les renseignements complémentaires :

R^2 = 0,8394 f_c = 31,37 *SCE* = 1 793 130 *SCR* = 342 959,5

Nous avons ainsi presque tous les éléments pour constitutifs du listing, mais les niveaux de signification (par exemple) n'apparaissent pas clairement ici, l'interprétation des résultats n'est donc pas évidente sans calculs supplémentaires. Remarquons que seul manque dans ces résultats, pour construire le listing, le nombre de variables explicatives, que nous stockerons dans une cellule de la feuille de résultats. Nous allons maintenant construire sur une feuille nommée "Listing", construire une sortie standard de régression.

4.2 Listing de régression

Nous allons ici construire pas à pas chacun des éléments d'un listing standard de régression fourni par des package statistiques. Nous avons nommé "Resreg" la plage contenant les résultats de la fonction DROITEREG ci-dessus ($F$1:$I$5). Les éléments dont nous aurons besoin dans cette plage seront obtenus grâce à la fonction INDEX(Resreg;i;j) qui retourne

l'élément à l'intersection de la i^{ème} ligne (relative) et de la j^{ème} colonne (relative) de la plage Resreg.

La cellule B1 de la feuille "Listing" (nommée "Nvar") contient le nombre de variables explicatives (ici 3), voici la première ligne de cette feuille :

4.2.1 Construction du résumé

Ici nous allons donner deux résultats de la plage Resreg, le coefficient de détermination et l'erreur type de régression, et calculer le coefficient de corrélation multiple. Le coefficient de détermination est le premier élément de la troisième ligne de Resreg, l'erreur type le deuxième élément de la même ligne. Nous obtenons alors :

	A	B
3	**Résumé**	
4	R^2	=INDEX(Resreg;3;1)
5	Corrélation multiple	=RACINE(B4)
6	Erreur type	=INDEX(Resreg;3;2)

Formules

	A	B
3	**Résumé**	
4	R^2	0,83944516
5	Corrélation multiple	0,9162124
6	Erreur type	138,033713

Valeurs

Nous constatons que la régression semble a priori intéressante, dans la mesure ou le coefficient de détermination est élevé, le modèle explique "84% des variations" des ventes, l'erreur type serait à comparer avec l'écart type des ventes qui est de 318,9 ; on a donc une diminution très significative de l'incertitude. Toutefois ceci reste très vague et demande à être précisé par des tests.

4.2.2 Construction du tableau d'analyse de la variance

Pour construire ce tableau, nous devons prendre au moins trois éléments de la plage Resreg : la somme des carrés expliquée, la somme des carrés résiduelle et le nombre de degrés de liberté des résidus. La valeur de la statistique de Fisher, peut soit être calculée, soit être importée de cette plage. En revanche tous les autres éléments sont calculés, en particulier le niveau de signification, à l'aide de la fonction LOI.F d'Excel. Le tableau d'analyse de la variance, sous forme de formules, se présente ainsi :

	A	B	C	D	E	F
8	**Analyse de la Variance**					
9	Source	DL	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
10	Régression	=nvar	=INDEX(Resreg;5;1)	=C10/B10	=D10/D11	=LOI.F(E10;B10;B11)
11	Résidus	=INDEX(Resreg;4;2)	=INDEX(Resreg;5;2)	=C11/B11		
12	Totale	=SOMME(B10:B11)	=SOMME(C10:C11)			

et en valeurs:

	A	B	C	D	E	F
8	**Analyse de la Variance**					
9	Source	DL	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
10	Régression	3	1793129,948	597709,9828	31,3704081	2,31065E-07
11	Résidus	18	342959,5063	19053,3059		
12	Totale	21	2136089,455			

Comme ici le niveau de signification de f_c est inférieur à 5%, nous pouvons rejeter l'hypothèse suivant laquelle aucune des variables explicatives n'est significative. Il nous reste à vérifier la validité marginale de notre modèle. Pour cela nous allons construire le tableau des variables du modèle.

4.2.3 Le tableau du modèle

Pour construire ce tableau, nous avons besoin de prendre les coefficients et les écarts types des estimateurs des coefficients dans la plage de résultats. Les autres éléments sont calculés. En particulier le niveau de signification du *T* partiel, doit être calculé par la fonction d'Excel donnant la loi de Student, fonction, qui, rappelons le, a trois arguments :

- Le t_c calculé : rapport entre le coefficient et l'écart type de la variable
- Le nombre de degrés de liberté des résidus : repris de la plage "Resreg"
- Le fait que le test soit bilatéral ou non (ici bilatéral =2)

En tenant compte de l'ordre des éléments de la plage de résultats de la fonction DROITEREG, il est facile de construire le tableau :

	A	B	C	D	E
14	**Modèle**				
15	Variable	Coefficient	Ecart type	t_c calculé	Prob T>|t_c|
16	=INDEX(Xnoms;1)	=INDEX(Resreg;1;nvar)	=INDEX(Resreg;2;nvar)	=B16/C16	=LOI.STUDENT(ABS(D16);INDEX(Resreg;3;2);2)
17	=INDEX(Xnoms;2)	=INDEX(Resreg;1;nvar-1)	=INDEX(Resreg;2;nvar-1)	=B17/C17	=LOI.STUDENT(ABS(D17);INDEX(Resreg;3;2);2)
18	=INDEX(Xnoms;3)	=INDEX(Resreg;1;nvar-2)	=INDEX(Resreg;2;nvar-2)	=B18/C18	=LOI.STUDENT(ABS(D18);INDEX(Resreg;3;2);2)
19	Constante	=INDEX(Resreg;1;nvar+1)	=INDEX(Resreg;2;nvar+1)	=B19/C19	=LOI.STUDENT(ABS(D19);INDEX(Resreg;3;2);2)

Ce qui nous donne les valeurs suivantes :

	A	B	C	D	E
14	**Modèle**				
15	Variable	Coefficient	Ecart type	t_c calculé	Prob T>|t_c|
16	Radio	23,8499964	4,523786884	5,27213085	5,0808E-07
17	Journaux	32,6293884	5,368631858	6,077784678	1,12305E-08
18	Gratuits	-0,618743	10,22809676	-0,060494441	0,951849364
19	Constante	238,457818	112,2421031	2,124495277	0,035411499

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.

La construction de notre feuille listing n'est pas très difficile, mais nous sommes passés par le tableau intermédiaires (plage "Resreg") des résultats de la fonction DROITEREG. Il est possible de se passer de cette plage, pour cela il suffit dans toutes les formules de remplacer Resreg par sa valeur c'est à dire DROITEREG(Ydonnees;Xdonnees;;VRAI), ce qui donne par exemple pour le résumé les formules suivantes (classeur Pubradio2.xls) :

	A	B
1	Nb var. explicatives	3
2		
3	**Résumé**	
4	R^2	=INDEX(DROITEREG(Ydonnees;Xdonnees;;VRAI);3;1)
5	Corrélation multiple	=RACINE(B4)
6	Erreur type	=INDEX(DROITEREG(Ydonnees;Xdonnees;;VRAI);3;2)

L'idéal bien sûr serait de construire une feuille de génération automatique de listing de régression, cet exercice est laissé au lecteur intéressé par la modélisation sous Excel, un exemple en est toutefois donné dans le classeur Listreg.xls. Nous ne détaillerons pas ici les formules dans la mesure où nous donnons un add-in de régression générant ce listing.

4.2.4 Le listing final

Nb var. explicatives 3

Résumé

R^2	0,83945
Corrélation multiple	0,91621
Erreur type	138,03371

Analyse de la Variance

Source	DL	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
Régression	3	1793129,948	597709,9828	31,37040815	2,31065E-07
Résidus	18	342959,5063	19053,3059		
Totale	21	2136089,455			

Modèle

Variable	Coefficient	Ecart type	t_c calculé	Prob T>|t_c|
Radio	23,84999639	4,523786884	5,27213085	5,0808E-07
Journaux	32,62938845	5,368631858	6,077784678	1,12305E-08
Gratuits	-0,61874299	10,22809676	-0,060494441	0,951849364
Constante	238,4578179	112,2421031	2,124495277	0,035411499

4.3 Calcul des estimations $\hat{y}_i$ et, des résidus $\hat{e}_i$

Bien que le modèle trouvé ne soit pas satisfaisant statistiquement, nous allons indiquer comment calculer les estimations des moyennes $\hat{y}_i$ et des résidus $\hat{e}_i$.

4.3.1 Calcul des estimations $\hat{y}_i$

Pour calculer ces estimations il est possible d'utiliser une fonction vectorielle d'Excel, la fonction TENDANCE, cette fonction a la même contrainte que la fonction DROITEREG, les variables explicatives doivent être dans des colonnes adjacentes. La fonction TENDANCE a quatre arguments (un seul obligatoire)

- La plage des valeurs connues de la variable à expliquer (Y connus), ce paramètre est obligatoire.
- La plage des valeurs connues des variables explicatives (X connus), si cette plage est omise, Excel considère que les X sont les valeurs 1,2,….,*n*.
- La plage des X inconnus, si l'on veut prévoir des valeurs de $\hat{Y}$.
- L'existence d'une constante dans la régression, qui sera implicitement refaite, par défaut la valeur de ce paramètre booléen est Vrai, pour indiquer la présence d'une constante.

La formule est entrée matriciellement sur une plage unicolonne contenant autant de lignes que la réunion des plages X connus, X inconnus (classeur Pubradio1.xls) :

E2 = {=TENDANCE(Ydonnees;Xdonnees)}

	A	B	C	D	E	F
1	Ventes	Radio	Journaux	Gratuits	Estimations	Résidus
2	894	0	19	9	852,85	
3	1032	0	19	3	856,56	

Une autre méthode, aussi simple, consiste à utiliser la définition de $\hat{y}_i = b_0 + \sum_{k=1}^{p} b_k x_{ik}$. On entre cette formule dans la première cellule, puis on la recopie sur l'ensemble de la zone (classeur Pubradio2.xls) :

=Listing!$B$19+PRODUITMAT(Data!B2:D2;Listing!$B$16:$B$18)

- Listing!$B$19 est l'adresse de la constante de régression
- Listing!$B$16:$B$18 est l'adresse des autres coefficients de la régression

4.3.2 Calcul des résidus

La formule $\hat{e}_i = y_i - \hat{y}_i$, se traduit de façon simple dans la cellule $F$2 par **=A2-E2** puis est recopiée vers le bas. Il peut être utile de calculer les résidus "standardisés", c'est à dire divisés par leur écart type, dans la mesure où ils sont déjà centrés, la formule sera entrée dans la cellule G2 : **=F2/Listing!$B$6** et recopiée vers le bas, Listing!$B$6 étant l'adresse de l'erreur type de la régression.

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.

5.1 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

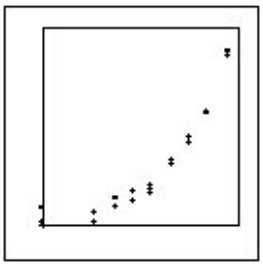


figure 2

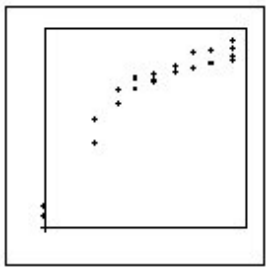


figure 3

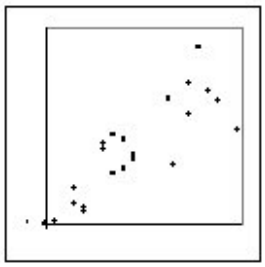


figure 4

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.

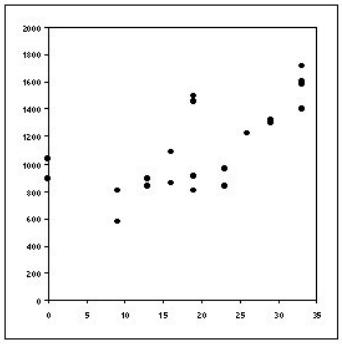
Les figures 2 et 3 indique clairement une allure non linéaire de la moyenne des *y* pour une abscisse *x* donnée, on pourra dans les deux cas essayer une transformation puissance d'exposant supérieur à 1 pour la figure 2(par exemple x^2) et inférieure à 1 pour la figure 3

(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.

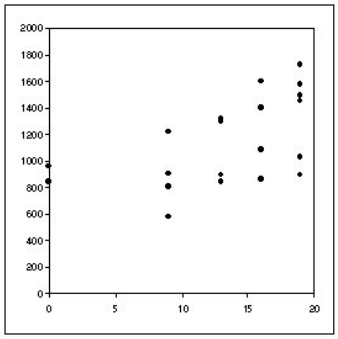
La figure 4 ne met en cause fondamentalement, la linéarité de la moyenne, mais elle montre clairement que la dispersion autour de cette moyenne n'est pas constante, les données ne respectent pas l'hypothèse d' homoscédasticité des résidus, on peut penser ici que la dispersion est proportionnelle à une 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 sous Excel, doivent être validées par un nouveau graphique et aussi par le calcul des corrélations simples éventuellement (fonction COEFFICIENT.CORRELATION(valeursY;valeursX)).

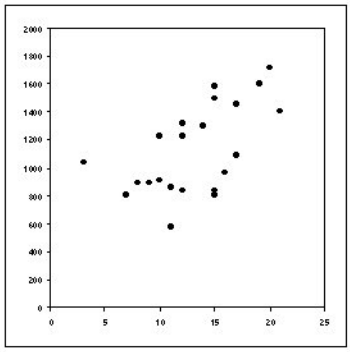
Application à notre exemple, les trois graphiques sont les suivants :



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 (la formule est donnée uniquement dans le cas des valeurs de la variable explicative Radio, elle peut être recopiée pour les autres variables explicatives) :

	Radio/Ventes	Journaux/Ventes	Gratuits/Ventes
Formule	=COEFFICIENT.CORRELATION(Ydonnees;B2:B23)		
Valeur	0,707132	0,539128	0,588683

5.2 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.

5.2.1 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 :

Variable	Coefficient	Ecart type	t_c calculé	Prob T>|t_c|
Radio	23,6460	2,9346	8,0577	0,0000
Journaux	32,5707	5,1400	6,3367	0,0000
Constante	235,1678	95,5770	2,4605	0,0151

Cette fois toutes les variables sont marginalement significatives et le modèle est donc acceptable statistiquement.

5.2.2 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 (Pubradio2.xls et Pub radio3.xls), on obtient les résumés suivants :

Modèle	R2	Erreur Type
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.

5.3 Analyse des résidus

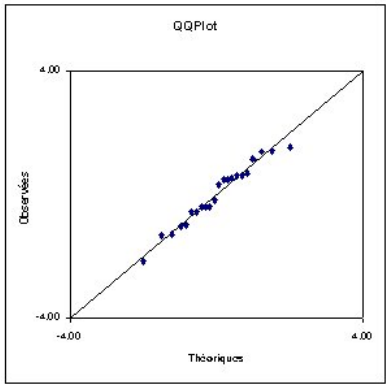
Quand un modèle est satisfaisant statistiquement et économiquement, il nous reste à vérifier que les hypothèses faites sur les résidus, la normalité, l'indépendance et l' homoscédasticité.

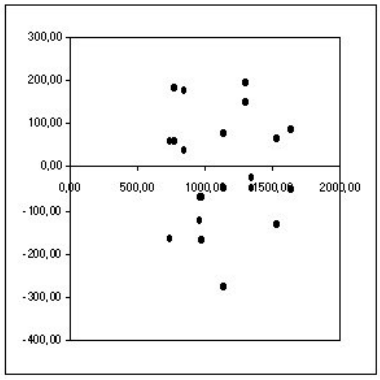
L'indépendance n'est facilement vérifiable que lorsque les variables sont temporelles, dans ce cas le plus simple est de représenter sur un graphique cartésien le résidu en t en fonction du résidu en t-1 (on peut aussi utiliser la statistique de Durbin-Watson).

5.3.1 Normalité et homoscédasticité des résidus

Pour vérifier l'indépendance, on pourra utiliser le graphique normal (voir les rappels d'Excel) 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 $(\hat{y}_i)_{1\le i\le n}$ ce qui résume l'ensemble des graphiques. Sur le modèle retenu pour l'exemple (fichier Pubradio3.xls), 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.

5.3.2 La statistique de Durbin-Watson

La statistique de Durbin-Watson sert à détecter des autocorrélations éventuelles entre les résidus. Cette statistique est définie par :

$$DW = \frac{\sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2} = \frac{\sum_{i=2}^{n}\hat{e}_i^2 + \sum_{i=1}^{n-1}\hat{e}_i^2 - 2\sum_{i=2}^{n-1}\hat{e}_i\hat{e}_{i-1}}{\sum_{i=1}^{n}\hat{e}_i^2} \quad \text{pour n grand} \quad \approx 2 - 2\frac{\sum_{i=2}^{n-1}\hat{e}_i\hat{e}_{i-1}}{\sum_{i=1}^{n}\hat{e}_i^2}$$

Si les résidus ne sont pas corrélés, le second terme sera nul en théorie, donc la statistique sera proche de 2. En revanche si les résidus sont corrélés positivement le second terme sera proche de -2 et la statistique proche de 0, enfin si les résidus sont corrélés négativement le second terme est proche de 2 et la statistique proche de 4. Le problème est de déterminer à partir de quelles valeurs on peut conclure à l'existence d'une autocorrélation, ces valeurs sont données dans table en annexe, et ne sont malheureusement pas accessibles directement par une fonction d'Excel. Sur cette table ne figure que les valeurs correspondant à une autocorrélation positive, le cas d'une autocorrélation négative se traitant par symétrie par rapport à 2. Le test de Durbin-Watson présente une importante particularité, par rapport aux autres tests évoqués dans ce chapitre :

– La valeur critique est double (pour un risque de première espèce donné) : une valeur en dessous de laquelle on conclut à l'autocorrélation positive et une valeur au-dessus de la quelle on conclut à l'absence d'autocorrélation.

Exemple d'utilisation de la table, dont voici un extrait (pour $\alpha = 5\%$) :

	$p=1$		$\mathbf{p=2}$		$p=3$	
n	d_L	d_U	d_L	d_U	d_L	d_U
⋮						
24	1,27	1,45	***1,19***	***1,55***	1,10	1,66

⋮						

Si on a fait une régression (temporelle) à deux variables explicatives, à partir d'un échantillon de 24 données, soit *dw* la valeur de la statistique de Durbin-Watson, calculée sur les résidus. On conclura de la façon suivante :

- Si ***dw<1,19***, on considérera (au risque 5%) qu'il existe une autocorrélation positive entre les résidus et donc que le modèle de régression linéaire ne peut s'appliquer. Il faudra alors utiliser d'autres types de modèles tels que ceux de Box et Jenkins par exemple.
- Si ***1,55<dw<4-1,55=2,45*** on considérera qu'il n'existe pas d'autocorrélation (positive ou négative) entre les résidus, le modèle de régression linéaire est alors applicable.
- Si ***dw>4-1,19=2,81*** on considérera (au risque 5%) qu'il y évidence d'une autocorrélation négative entre les résidus et donc que le modèle de régression linéaire ne peut s'appliquer (voir le premier cas).
- Dans les autres cas on ne peut conclure!

Un extrait de la table est donnée dans le fichier Durbin-Watson.xls

5.4 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.

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.

6.1 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 :

$$\text{pour } 1 \le j \le m \quad X_{1j} = 1 \quad \text{si} \quad X_1 = m \quad , \quad X_{1j} = 0 \quad \text{sinon}$$

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.

6.2 Création des indicatrices sous Excel

La création des indicatrices se fait simplement sous Excel en utilisant la fonction SI. Pour l'utilisation des fonctions standard de régression d'Excel, il est recommandé de ne créer que les *m*-1 indicatrices utiles dans la mesure où, comme nous l'avons signalé plus haut, les variables explicatives doivent être dans une plage constituée de colonnes contiguës. Nous donnons plus loin une macro complémentaire qui permet de se passer de cette contrainte.

Dans notre exemple, la variable Enseigne donne naissance à trois variables indicatrices, nommée Enseigne1, Enseigne2, Enseigne3, dont seules les deux premières seront créées sur la feuille. Les formules sont les suivantes :

	B	C	D	E
1	**Enseigne**	**Région**	**Enseigne1**	**Enseigne2**
2	2	3	=SI(B2=1;1;0)	=SI(B2=2;1;0)
3	3	4	=SI(B3=1;1;0)	=SI(B3=2;1;0)

Ces formules doivent être entrées pour chaque colonne correspondant à une variable indicatrice, si le nombre de modalités est plus important il est possible d'utiliser le nom des variables indicatrices pour entrer une seule formule recopiée sur la droite et vers le bas, c'est ce que nous avons fait pour la région :

	C	F
1	**Région**	**Région1**
2	3	=SI($C2=CNUM(DROITE(F$1;1));1;0)
3	4	=SI($C3=CNUM(DROITE(F$1;1));1;0)

La formule utilise le fait que le dernier caractère du nom (dernier caractère à droite) de la variable indicatrice est égal à la modalité associée à cette variable.

6.3 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 $(X_j)_{1\le j\le m-1}$ 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 :

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne1	-69,76623377	32,35742517	-2,156112033	0,04282314
Enseigne2	119,1428571	37,23317714	3,199911109	0,004304405
Constante	226,8571429	25,29496283	8,968471091	1,25784E-08

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.

6.4 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 :

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne1	-21,4655	45,8613	-0,4681	0,6454
Enseigne2	121,8364	40,8565	2,9821	0,0080
Région1	-66,7396	47,9676	-1,3913	0,1811
Région2	-26,3673	43,6228	-0,6044	0,5531
Région3	10,7324	47,1958	0,2274	0,8227
Constante	235,5585	37,0962	6,3499	0,0000

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".

6.4.1 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 conduit donc 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, avec Excel, 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.

6.4.2 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	$\frac{SCEC}{p} = SME$	$f_g = \frac{SME}{SCRM}$	$prob\left(F_{p,n-p-1} > f_g\right)$
G_m	$SCEC\text{-}SCE_{p-m} = S_m$	m	$\frac{S_m}{m} = SM_m$	$f_m^p = \frac{SM_m}{SCRM}$	$prob\left(F_{m,n-p-1} > f_m^p\right)$
G_{p-m}	$SCEC\text{-}SCE_m = S_{p-m}$	$p-m$	$\frac{S_{p-m}}{p-m} = SM_{p-m}$	$f_{p-m}^p = \frac{SM_{p-m}}{SCRM}$	$prob\left(F_{p-m,n-p-1} > f_{p-m}^p\right)$
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.

6.4.3 Mise en œuvre sous Excel

Pour pouvoir facilement établir le tableau d'analyse de la variance sous Excel, sans avoir recours à des macros, il est nécessaire que les données soient disposées convenablement, c'est à dire que les groupes de variables G_m et G_{p-m} correspondent à des plages de la feuille de calcul (des colonnes contiguës) qui sont adjacentes. C'est le cas pour notre exemple, le groupe de variables des Enseignes (Enseigne1 et Enseigne2) occupe la plage Groupe1=D2:E25, le second groupe (Région1, Région2, Région3) occupe la plage Groupe2=F2:H25, la plage des variables du modèle complet est donc Complet=D2:H25.

La fonction DROITEREG peut alors être utilisée pour calculer les différentes sommes de carrés :

- La somme des carrés expliquée du modèle complet est le premier élément de la cinquième ligne de la fonction DROITEREG appliquée au modèle complet : (*SCEC*=)INDEX(DROITEREG(PlageY;Complet;;VRAI);5;1)
- La somme des carrés résiduelle du modèle complet est le premier élément de la cinquième ligne de la fonction DROITEREG appliquée au modèle complet : (*SCR*=)INDEX(DROITEREG(PlageY;Complet;;VRAI);5;2)
- La somme des carrés expliquée du modèle Groupe1 est le premier élément de la cinquième ligne de la fonction DROITEREG appliquée au modèle ne comprenant

que les variables du Groupe1 :
(*SCEC*=)INDEX(DROITEREG(PlageY;Groupe1;;VRAI);5;1)

- La somme des carrés expliquée du modèle Groupe2 est le premier élément de la cinquième ligne de la fonction DROITEREG appliquée au modèle ne comprenant que les variables du Groupe2 :
(*SCEC*=)INDEX(DROITEREG(PlageY;Groupe2;;VRAI);5;1)

Les autres formules du tableau d'analyse de la variance ne présentent aucune difficulté, les voici :

	A	B	C	D	E	F
1	Anlyse de la varia					
2	Source	Somme des Carrés	DL	Carré Moyen	F	Prob >F
3	Enseigne-Région	=INDEX(DROITEREG(Ventes;Complet;;VRAI);5;1)	5	=B3/C3	=D3/$D$6	=LOI.F(E3;C3;$C$6)
4	*Enseigne*	*=B3-INDEX(DROITEREG(Ventes;Groupe2;;VRAI);5;1)*	*2*	*=B4/C4*	*=D4/$D$6*	*=LOI.F(E4;C4;$C$6)*
5	*Région*	*=B3-INDEX(DROITEREG(Ventes;Groupe1;;VRAI);5;1)*	*3*	*=B5/C5*	*=D5/$D$6*	*=LOI.F(E5;C5;$C$6)*
6	Résidus	=INDEX(DROITEREG(Ventes;Complet;;VRAI);5;2)	18	=B6/C6		
7	Totale	=B6+B3	=C6+C3			

Ce qui donne les valeurs :

Analyse de la variance

Source	Somme des Carrés	DL	Carré Moyen	F	Prob >F
Enseigne-Région	150023,4570	5	30004,6914	6,5363	0,00124561
Enseigne	*53141,3736*	*2*	*26570,6868*	*5,7883*	*0,0114532*
Région	*11427,8899*	*3*	*3809,2966*	*0,8298*	*0,4946877*
Résidus	82627,8764	18	4590,4376		
Totale	232651,3333	23			

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 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.

7.1 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.

7.2 Un exemple

Nous avons déjà vu une illustration de cette méthode au paragraphe 5.2 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%.

Le listing de la première étape est le suivant :

Régression
Ventes en fonction de Région3, Région2, Région1, Enseigne2, Enseigne1

Valeur de R2	0,644842455
Corrélation mult.	0,803020831
Erreur de la régression	67,75276803

Analyse de la variance

Source	D.L.	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
Régression	5	150023,457	30004,69139	6,536346677	0,001245608
Résidus	18	82627,87636	4590,437576		
Total	23	232651,3333	10115,27536		

Modèle Estimé

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne1	-21,46545455	45,86125854	-0,468052017	0,645364908
Enseigne2	121,8363636	40,85645638	2,982059005	0,007991
Région1	-66,73963636	47,9676374	-1,391347166	0,181078888
Région2	-26,36727273	43,62275818	-0,604438459	0,553095538
Région3	10,73236364	47,19583919	0,227400632	0,822675087
Constante	235,5585455	37,09622249	6,349933488	5,54823E-06

Le modèle est valide globalement mais ne l'est pas statistiquement. Quatre variables explicatives ne sont pas significatives marginalement, la variable dont le niveau de

signification est le plus fort est la variable Région 3 qui va donc sortir du modèle. La deuxième étape nous donne les résultats suivants :

Régression
Ventes en fonction de Région2, Région1, Enseigne2, Enseigne1

Valeur de R2	0,643822146
Corrélation mult.	0,803020831
Erreur de la régression	66,04035955

Analyse de la variance

Source	D.L.	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
Régression	4	149786,0806	37446,52016	8,586034072	0,000391824
Résidus	19	82865,2527	4361,329089		
Total	23	232651,3333	10115,27536		

Modèle Estimé

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne1	-23,97482014	43,38880151	-0,552557787	0,587008636
Enseigne2	119,9865108	39,02647901	3,074489778	0,006239747
Région1	-70,14658273	44,41616195	-1,579303111	0,130770816
Région2	-30,42086331	38,80812385	-0,783878742	0,442774438
Constante	241,2239209	26,79060865	9,004047801	2,77329E-08

Le modèle est valide globalement, mais il reste toujours trois variables non significatives marginalement (dont le niveau de signification est supérieur à 0,05). La variable qui va sortir du modèle est la variableEnseigne1. On obtient à l'étape trois :

Régression
Ventes en fonction de Région2, Région1, Enseigne2

Valeur de R2	0,638098553
Corrélation mult.	0,803020831
Erreur de la régression	64,88330069

Analyse de la variance

Source	D.L.	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
Régression	3	148454,4792	49484,82639	11,7545547	0,000116782
Résidus	20	84196,85417	4209,842708		
Total	23	232651,3333	10115,27536		

Modèle Estimé

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne2	126,9791667	36,27086774	3,500858253	0,002250646
Région1	-86,72916667	32,17016733	-2,695950126	0,013900607
Région2	-34,41666667	37,46039112	-0,918748193	0,369175469
Constante	236,2291667	24,77771973	9,533934892	7,01583E-09

Le modèle est toujours globalement acceptable, mais il reste encore une variable explicative non marginalement significative, on obtient alors le modèle final :

Régression
Ventes en fonction de Région1, Enseigne2

Valeur de R2	0,622824534
Corrélation mult.	0,803020831
Erreur de la régression	64,64200868

Analyse de la variance

Source	D.L.	Somme des Carrés	Carré Moyen	f_c calculé	Prob F>f_c
Régression	2	144900,9583	72450,47917	17,33850211	3,57854E-05
Résidus	21	87750,375	4178,589286		
Total	23	232651,3333	10115,27536		

Modèle Estimé

Variable	Coefficient	Ecart type	t_c calculé	prob T>|t_c|
Enseigne2	118,375	34,9106747	3,390796683	0,002756969
Région1	-78,125	30,66239699	-2,547909089	0,018728651
Constante	227,625	22,85440134	9,95978834	2,07897E-09

Ce 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.
- D'un point de vue pratique, la mise en place d'une régression pas à pas est plus lourde avec Excel, car on n'aura pas toujours la chance comme ici de garder des variables explicatives dans des colonnes adjacentes, il sera alors nécessaire de recopier les données sur d'autres feuilles. C'est pour cela qu'une macro complémentaire est proposée avec cet ouvrage.

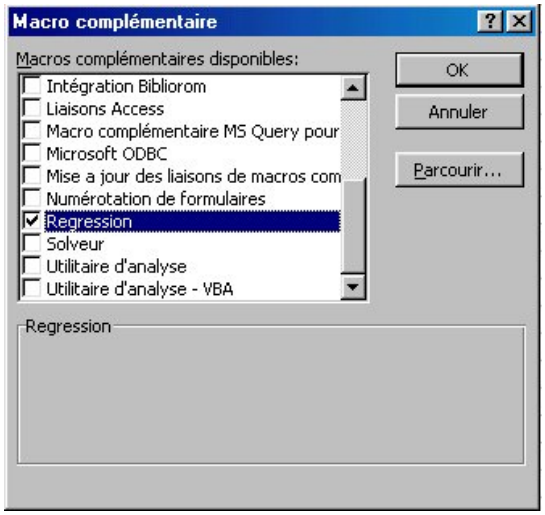
8 La macro complémentaire (add in) ModLinéaire.xla

Cette macro complémentaire, permet de faire des régressions, des régressions pas à pas, et des calculs de F partiel en s'affranchissant de la contrainte portant sur la localisation des variables explicatives dans des colonnes adjacentes. Le tableau de données doit être une base de données Excel (voir Rappels Excel), c'est à dire que les variables sont associées à des colonnes adjacentes et que le nom des variables se trouve dans la première ligne.

8.1 Installation de la macro complémentaire

La macro complémentaire est un fichier qui a pour non "Regression.xla". Copier ce fichier dans un répertoire de votre disque dur, par exemple "Mes macros". Dans le menu Outils d'Excel choisir le sous menu Macros complémentaires… apparaît alors la boite de dialogue suivante :

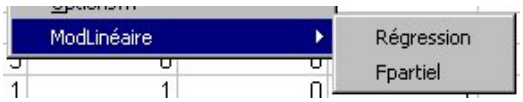
Cliquer alors sur le bouton parcourir pour aller désigner le fichier que vous venez de copier, la macro apparaît alors cochée dans la liste des macro complémentaires disponibles :



Après avoir cliqué sur OK, la macro est installée et le menu Outils mis à jour, un sous menu ModLinéaire est créé..

8.2 Utilisation de la macro complémentaire

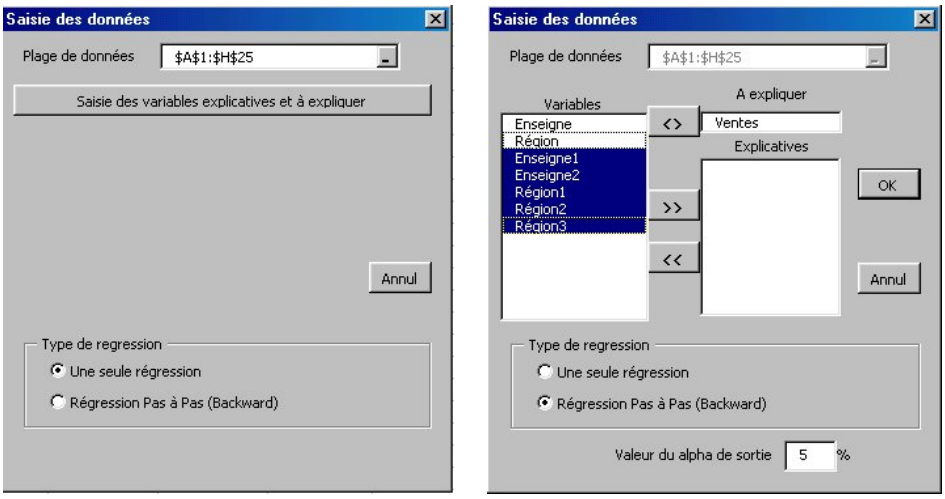
Pour utiliser la macro complémentaire Regression, il est recommandé de choisir une cellule de la plage de données comme cellule active. Dans le menu Outils choisir le sous menu ModLinéaire :



il suffit alors de choisir le sous menu Régression ou Fpartiel qui fait apparaître une boite de dialogue.

8.2.1 Boite de dialogue régression

La boite de dialogue Régression permet de faire soit une régression unique soit une régression pas à pas "backward". Dans un premier temps l'utilisateur doit sélectionner la plage de données, ensuite il choisira les variables explicatives et à expliquer :



La liste de gauche contient les intitules de toutes les variables de la plage de données, correspondant à la première ligne de cette plage. Le bouton <> permet de sélectionner (ou "déselectionner") la variable à expliquer, cette variable est ôtée de la liste en cas de sélection, et rajoutée à la liste si elle avait déjà été sélectionnée comme variable à expliquer.

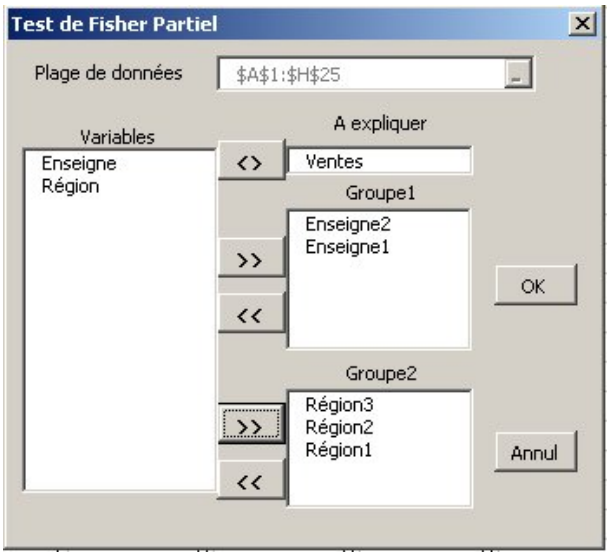
Les deux boutons >> et << servent respectivement à sélectionner ou "déselectionner", une ou plusieurs variables comme variables explicatives, les touches de sélection multiple (majuscule et Ctrl) peuvent être utilisées.

Enfin si la régression pas à pas est choisie, l'utilisateur doit donner la valeur du niveau de signification maximum accepté, seuil de sortie des variables explicatives, cette valeur est par défaut de 5%.

Une fois le dialogue validé, les résultats de la régression ou de la procédure de régression pas à pas sont donnés sur une nouvelle feuille nommée "Rapport de régression n".

8.2.2 Boite de dialogue Fpartiel

Le processus est identique, l'utilisateur fixe d'abord la plage de données, contenant les variables explicatives et à expliquer. La deuxième partie du dialogue consiste à définir la variable à expliquer ainsi que les deux groupes de variables sur lequel doit porter le test partiel :



Les différentes zones se remplissent comme pour le dialogue de régression, le listing de résultat est créé sur une nouvelle feuille de calcul nommée "Fpartieln", et est présenté sous la forme suivante :

Tableau d'analyse de la variance - Test Fisher Partiel
Variable à expliquer :
Ventes

Pemier groupe de Variables :
Enseigne2, Enseigne1

Deuxième groupe de Variables :
Région3, Région2, Région1

Analyse de la variance

Source	D.L.	Somme des Carrés	Carré Moyen	F Calculé	Prob >F
Régression	5	150023,457	30004,69139	6,536346677	0,001245608
Groupe 1	*2*	*138595,5671*	*69297,78355*	*5,788268848*	*0,011453192*
Groupe 2	*3*	*96882,08333*	*32294,02778*	*0,829833008*	*0,494687671*
Résidus	18	82627,87636	4590,437576		

EXERCICES DE REGRESSION LINEAIRE

1 L'entreprisse Elec (Elec.xls)

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. Représenter graphiquement les données, le modèle linéaire vous paraît-il approprié?
2. Etude des régressions à une seule variable explicative : toutes les variables sont-elles individuellement influente sur les variations des ventes? Les régressions vous semblent-elles toutes valides économiquement (en particulier pour la régression Ventes / Prix de l'entreprise)
3. Etudier de la même façon les régressions à deux variables explicatives? Quelle est pour vous la meilleure régression à 2 variables pour expliquer les variations des ventes, pour prévoir les ventes?
4. Que pensez-vous du modèle complet? Comment expliquer que certaines variables individuellement significatives ne le soient plus marginalement? Vérifiez vos assertions à l'aide de régressions linéaires.
5. Appliquer la méthode de régression pas à pas "backward" aux données, puis vérifier à l'aide du tes de Fisher partiel qu'il était possible de passer directement du modèle complet au modèle trouvé par la méthode pas à pas.
6. Sur le modèle trouvé à la question précédente, procédez à l'analyse des résidus. Quelles sont les données mal reconstitué par le modèle (données dont le résidu standardisé est >2) ?

2 Les stylos Runild (Runild.xls)

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

1) Etude des ventes en fonction des deux variables publicité et nombre de représentant.
   a) Représenter graphiquement les ventes en fonction des deux variables, 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 850€, 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. Construire le modèle et analyser les résultats. En prenant un risque a de 0,01 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) Construire le modèle qui vous paraît le plus pertinent avec les trois variables. Analyser les résidus correspondants.

3 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)[7] pour la marque dans la zone de chalandise concernée (entre 0 et 1)

[7] 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.

- 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.
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 ?