BBA1 – 2eme semestre 2015/2016

Statistiques appliquées au marketing

Professeur
M.BELLALAH

OBJECTIFS

L'objectif de ce cours consiste à étudier les différentes formules statistiques et les techniques probabilistes appliquées à la gestion. Les chapitres abordés représentent des outils scientifiques d'aide à la décision. Ainsi le programme ne se limite pas aux formules mathématiques, mais permet d'appréhender la gestion sous un angle scientifique permettant de calculer un chiffre d'affaires moyen en présence d'incertitude sur l'évolution du marché. Par ailleurs, les thèmes abordés dans ces chapitres constituent un élément fondamental dans la compréhension des aléas affectant l'univers économique des entreprises. En effet, ils permettent aux étudiants de comprendre les différentes interactions entre les variables affectant les décisions et les stratégies des entreprises. A l'issu de ce cours, l'étudiant sera en mesure de comprendre la modélisation des variables, et savoir manipuler les opérateurs d'espérance et de variance.
En outre, l'objectif de ce cours consiste à maîtriser les techniques statistiques de base appliquées au Marketing.
Le contenu de cette matière constitue des éléments cruciaux pour le traitement de l'information collectée auprès des ménages. Cet outil statistique permet aux étudiants de comprendre comment sélectionner un échantillon, traiter les informations, et comment étudier les variables caractéristiques de l'échantillon.
Ainsi, les thèmes abordés sont le sondage, l'échantillonnage, l'estimation, les tests de comparaison (proportions, moyennes et variances), les tests d'hypothèse ainsi que la régression linéaire multiple.
La maîtrise de ces techniques statistiques permet un traitement approprié de l'information collectée et de juger de la fiabilité des résultats afin de tirer les conclusions significatives. Plusieurs exercices d'application seront à l'appui pour permettre aux étudiants de parfaire leurs connaissances.

TYPE DE PEDAGOGIE

- Cours magistral
- Exercices pratiques

METHODE D'EVALUATION

Les étudiants seront évalués grâce à un contrôle continu représentant **40 %** de la note finale.
En fin de Semestre, un Examen final de 3h permettra d'obtenir les **60 %** restants.

BIBLIOGRAPHIE

[1]	Probabilités et Statistique	J. FOURASTIE et J.F. LASLIER	**DUNOD**
[2]	Exercices Résolus de Statistiques	J. FOURASTIE et S. SAGUERY	**MASSON**
[3]	Statistiques	Th.H. WONNACOTT et RO.J. WONNACOTT	**ECONOMICA**
[4]	Statistique Descriptive	G. CALOT	**DUNOD**
[5]	Cours de Calcul des Probabilités	G. CALOT	**DUNOD**

DEROULEMENT DES SEANCES

Chapitre 1 : Statistique descriptive

- **Etude des caractères quantitatifs discrets,**
- **Représentation graphique,**
- **Caractéristiques de position et de dispersion**

Chapitre 2 : Statistique descriptive

- **Etude des caractères quantitatifs continus,**
- **Caractéristiques de position et de dispersion, Quantiles,**
- **Histogramme**

Chapitre 3 :

- **Etude des séries statistiques doubles,**
- **Corrélation,**
- **Régression linéaire**

Chapitre 4 : Rappel sur les notions probabiliste (we will discuss about this chapter)

- **Rappel sur les lois : binomiale, Hypergéométrique,**
- **Exercices**
- **Loi normale et les lois dérivant de la loi normale, La Loi de Student**
- **Exercices**

Chapitre 5-6 : Echantillonnage et Estimation

- **Notion sur les sondages**
- **Méthodes d'échantillonnage**
- **Echantillonnage par choix raisonné**
- **Echantillonnage aléatoire simple**
- **Echantillonnage par grappes**
- **Estimation de moyennes, écart type connu dans la population totale.**
- **Estimation de moyennes, écart type inconnu dans la population totale.**
- **Intervalle de confiance d'une estimation**
- **Exercices et corrections**

Chapitre 7 : Estimation des proportions

- **Estimation de proportions : Intervalle de confiance**
- **Estimation de proportions : Taille d'échantillon**
- **Exercices et corrections**

CARACTERES QUANTITATIFS DISCRETS

1- Introduction et exemple

Une variable quantitative est dite discrète si l'étendue des valeurs possibles est dénombrable, c'est-à-dire si les valeurs peuvent être énumérées sous la forme d'une liste de chiffres (a_1 **,** $a_{2,...}$ a_n**) ou plus souvent** d'entiers naturels (0,1,2…).

1.1. Exemple :

Considérons l'exemple suivant : on a interrogé 20 familles d'un immeuble sur leur nombre d'enfants, les résultats sont indiqués dans le tableau suivant :

Nombre d'enfants	0	1	2	3	4	Total
Nombre de familles	3	5	9	2	1	20

Les 20 familles constituent la **population** étudiée. Chaque famille représente un **individu**.

Le nombre d'enfants est le **caractère** étudié. Les chiffres 0, 1, 2, 3 et 4 s'appellent les **modalités**, ce sont les différentes valeurs prises par le caractère.
Puisque les modalités sont numériques, on dit que le caractère est **quantitatif**.
Comme les modalités sont en nombre fini (il y a 5 modalités différentes), on dit qu'on a affaire à un caractère **discret**.

Nous allons étudier un certain nombre de notions et de calculs concernant les caractères quantitatifs discrets :
Les valeurs 3, 5, 9, 2 et 1 s'appellent les **effectifs**, c'est à dire le nombre d'individus dans la population possédant la modalité considérée du caractère étudié.

La donnée des modalités et des effectifs s'appelle une **série statistique simple** (qui, en général, est donnée, comme dans l'exemple sous forme d'un tableau dit **tableau à simple entrée**).

20 représente ce qu'on appelle **l'effectif total**

Le rapport entre l'effectif et l'effectif total s'appelle la **fréquence** : 3/20, 5/20, 9/20, 2/20, 1/20.

On exprime en général les fréquences en pourcentages : 15% , 25%, 45%, 10%, 5%.
La somme des fréquences vaut 1 (ou 100%).

Fréquence	3/20	5/20	9/20	2/20	1/20
Pourcentage	15%	25%	45%	10%	5%

2. Caractéristiques de position ou tendances centrale

2.1 **Le mode**

Le mode est la modalité de plus grand effectif.
Dans l'exemple le mode vaut 2 puisque le plus grand effectif est 9.
Une série statistique peut bien sûr avoir plusieurs modes : on dit qu'elle est **bimodale** avec 2 et **plurimodale** avec plusieurs.

2.2 **La médiane**
La médiane (*symbolisée par med*) est la valeur du caractère partageant la population en deux (sous-populations égales, de même effectif), c'est à dire telle qu'il y ait autant d'individus en dessous qu'au dessus.
La médiane ne s'applique que lorsque les observations peuvent être ordonnées de plus petit à la plus grande. Pour trouver la médiane d'une série de donnée, il est utile de classer ces dernières dans un ordre croissant. On obtient une série ordonnée.

Dans notre exemple, l'effectif total valant 20, la médiane doit partager la population en deux sous-populations de 10 individus.
Il faut donc "couper" la population entre le $10^{ème}$ et le $11^{ème}$ individu, c'est à dire pour une valeur de 2 (le $10^{ème}$ vaut 2 et le $11^{ème}$ aussi); la médiane vaut 2. D'une manière générale si le nombre d'observation est pair la médiane peut être n'importe quelle valeur située entre $(n/2)^e$ observation et la $[(n+2)/2]^e$ observation. On "coupe" ici entre deux individus car il y en a un nombre **pair** au total, mais si l'effectif total avait été **impair**, on aurait coupé **sur** un individu (la médiane aurait alors été la valeur de cet individu). Il s'agit en fait de la valeur $[(n+1)/2]^e$.

2.3. Les moyennes

On peut aussi calculer la moyenne de cette série, pour cela, on multiplie chaque modalité par l'effectif correspondant, on fait la somme, puis on divise par l'effectif total :

$$\bar{x} = \frac{(3\times 0)+(5\times 1)+(9\times 2)+(2\times 3)+(1\times 4)}{20} = 1{,}65$$

C'est le calcul habituel de la moyenne; en fait, ce n'est pas la seule moyenne, on l'appelle la moyenne **arithmétique**. Quand on dira "moyenne" sans préciser, il s'agira toujours de celle-ci.

Il existe d'autres moyennes, nous allons en voir trois : la quadratique, l'harmonique, la géométrique.

La moyenne **quadratique**, c'est la racine carrée de la moyenne des carrés; c'est à dire qu'on calcule la moyenne arithmétique (comme précédemment) en prenant les carrés des modalités (les effectifs, eux, restent les mêmes) et on prend la racine carrée à la fin. Ici, cela donne :

$$x_Q = \sqrt{\frac{(3\times 0^2)+(5\times 1^2)+(9\times 2^2)+(2\times 3^2)+(1\times 4^2)}{20}} = 1{,}803$$

La moyenne **harmonique**, c'est l'inverse de la moyenne des inverses; c'est à dire qu'on calcule la moyenne en prenant les inverses des modalités et qu'on prend l'inverse à la fin. Ici, cela donne :

$$x_H = \frac{1}{\left(\frac{(3\times\frac{1}{0})+(5\times\frac{1}{1})+(9\times\frac{1}{2})+(2\times\frac{1}{3})+(1\times\frac{1}{4})}{20}\right)}$$

Comme l'inverse de 0 n'existe pas, on ne peut pas calculer la moyenne harmonique pour cette série.

La moyenne **géométrique** s'obtient en mettant les modalités à une puissance correspondant à l'effectif, à faire ensuite le produit des termes obtenus, à prendre enfin la racine N^{ème} (N étant l'effectif total) du résultat obtenu. Ici, cela donne :

$$x_G = \sqrt[20]{0^3 \times 1^5 \times 2^9 \times 3^2 \times 4^1} = \left(0^3 \times 1^5 \times 2^9 \times 3^2 \times 4^1\right)^{\frac{1}{20}} = 0 =$$

Quelle que soit la série statistique considérée, ces moyennes sont toujours dans le même ordre : la plus petite est l'harmonique, puis la géométrique, puis l'arithmétique et enfin la quadratique.

Toutes ces notions (mode, médiane, moyennes) s'appellent des caractéristiques (ou des paramètres) de **position** (on dit aussi de tendance centrale). Ils résument globalement la série : le mode c'est la majorité, la médiane c'est la moitié, la moyenne c'est la moyenne.
On peut classer les moyennes arithmétique, géométrique, harmonique et quadratique de la manière suivante :

Harmonique < Géométrique < Arithmétique < Quadratique

3. **Les caractéristiques de dispersion**

Les paramètres de position tels qu'on vient de les définir permettent de se faire une première idée d'une série statistique, mais ils ne suffisent pas.

Considérons par exemple deux élèves; le premier a obtenu les notes suivantes : 3 fois 9, 4 fois 10 et 3 fois 11; le second a obtenu 3 fois 0, 4 fois 10 et 3 fois 20.
Si on calcule mode, médiane ou moyenne pour ces deux séries de notes, on obtiendra 10 à chaque fois; pourtant, ces deux élèves présentent des profils complètement différents : le premier est un élève très régulier avec des notes constamment voisines de la moyenne alors que le second est très irrégulier.

Il faut donc trouver un moyen de mesurer ce phénomène, c'est-à-dire la "dispersion" des valeurs prises par un caractère par rapport à la moyenne (ou par rapport à tout autre paramètre de position).

On va essayer de calculer un écart "global" pour l'ensemble de la série : la première idée qui vient alors consiste à calculer la moyenne arithmétique des écarts par rapport à la moyenne.

Dans l'exemple du début, cela donne :

$$\frac{\left[3 \times (0 - 1{,}65)\right] + \left[5 \times (1 - 1{,}65)\right] + \left[9 \times (2 - 1{,}65)\right] + \left[2 \times (3 - 1{,}65)\right] + \left[1 \times (4 - 1{,}65)\right]}{20} = 0$$

Les écarts négatifs et les écarts positifs se compensent de façon à donner un résultat nul pour toute série.

Il faut donc rendre tous les écarts positifs avant d'en faire la moyenne, il y a deux moyens de le faire :
la valeur absolue et le carré.

Avec la valeur absolue, on obtient ce qu'on appelle l'Ecart (absolu) moyen (noté E) alors qu'avec le carré, on obtient ce qu'on appelle la Variance (notée V). Ici cela donne :

$$\frac{\left[3\times|0-1,65|\right]+\left[5\times|1-1,65|\right]+\left[9\times|2-1,65|\right]+\left[2\times|3-1,65|\right]+\left[1\times|4-1,65|\right]}{20} = 0,82$$

$$\frac{\left[3\times(0-1,65)^2\right]+\left[5\times(1-1,65)^2\right]+\left[9\times(2-1,65)^2\right]+\left[2\times(3-1,65)^2\right]+\left[1\times(4-1,65)^2\right]}{20} = 1,0275$$

En fait, quand on veut calculer pratiquement la variance, on ne procède pas comme indiqué ci-dessus.

Pour calculer la variance, on fait la "moyenne des carrés" moins le "carré de la moyenne" (la moyenne des carrés signifie qu'on calcule la moyenne avec les carrés des modalités, comme on l'a déjà fait pour la moyenne quadratique).

$$V = \frac{(3\times 0^2)+(5\times 1^2)+(9\times 2^2)+(2\times 3^2)+(1\times 4^2)}{20} - (1,65)^2 = 1,0275$$

La racine carrée de la Variance sera appelée **l'écart-type** (noté σ). Ici $\sigma = 1,0137$

CARACTERES QUANTITATIFS DISCRETS

Enoncés des Exercices

Exercice 1

On considère la demande d'un produit A sur le marché international. Cette demande est caractérisée par la modalité qui représente la quantité consommée en fonction du nombre de consommateurs donné par l'effectif. On vous demande de calculer pour cette série statistique
Le Mode, la Médiane, les Moyenne**s**, l'Ecart moyen, la Variance et Ecart-type

Modalités	Effectifs
1	8
2	6
3	5
4	1

Exercice 2

Un gestionnaire cherche à étudier la consommation du coca sur le marché. Après une enquête, il constate que la répartition de la consommation des boites est donnée par le tableau suivant :

La quantité consommée « modalité »	« Effectifs » désigne le nombre de consommateurs
2	6
5	6
8	5
10	3
12	1

Calculer le Mode, la Médiane, les Moyenne**s**, l'Ecart moyen absolu, la Variance et Ecart-type.
+ Q3-Q1, D9-D1

Exercice 3

La demande d'un produit cosmétique est donnée par les modalités. Le nombre de femmes interrogées est représenté par les effectifs. On se fondant sur les données suivantes calculer :
le Mode, la Médiane, les Moyenne**s**, l'Ecart moyen absolu, la Variance et Ecart-type. + Q3-Q1, D9-D1

Modalités	Effectifs
1	8
2	6
5	6
7	6
9	9
10	5

CARACTERES QUANTITATIFS CONTINUS

Définition :
Une variable quantitative continue peut prendre n'importe quelles valeurs à l'intérieur d'un certain **intervalle de variation** qui lui est associé. Les observations obtenues à partir d'une variable continue sont donc espacées. Leur organisation sous la forme d'un tableau statistique consiste à délimiter au préalable l'intervalle de variation.

Considérons l'exemple suivant : on s'intéresse aux salaires mensuels dans une entreprise de 200 personnes, les résultats sont consignés dans le tableau suivant.

	Moins de 400 €	de 400 à 600 €	de 600 à 1000€	de 1000 à 2000 €	de 2000 à 5000€	
Effectifs des classes	40	100	50	8	2	200
Fréquences	20% =40/200	50%	25%	4%	1%	100%

Le caractère étudié, c'est ici le salaire mensuel. La population est composée de 200 individus.

Le caractère a été regroupé en intervalles (qu'on appellera des **classes**) car il aurait été impossible de faire un tableau à simple entrée du type de celui qu'on a vu pour les caractères discrets du fait du trop grand nombre de salaires possibles.

Quand un caractère présente un nombre infini (ou très grand) de modalités, on dit qu'on a affaire à un caractère **continu**.

0, 400, 600,1000, 2000, 5000 sont appelées les **extrémités** de classes.

On remarque que toutes les classes n'ont pas forcément la même "largeur" (on parlera d'**amplitude**) : la première a une amplitude de 400, la deuxième de 200, les autres de 400,1000, 3000.

Le **centre** d'une classe sera le "milieu" de l'intervalle, la demi-somme des extrémités.
Ici, les centres des classes vaudront : 2000, 5000, 8000, 15000, 35000.

Mode :
Si la variable est continue, et si les données sont groupées en classes, on parle plutôt de classe modale. Pour chercher le mode on divise la fréquence par l'amplitude, on obtient ce qu'on appelle la **fréquence relative** de la classe.
Ici, les fréquences relatives seront respectivement de :

0,00005	0,00025	0,0000625	0,000004	0,000000033

Pour déterminer le mode d'un caractère continu, il faut d'abord déterminer la classe modale; le mode sera le centre de celle-ci.

La classe modale sera la classe qui aura la plus grande **fréquence relative** (et non pas la plus grande fréquence ou le plus grand effectif)

Ici, la classe modale est donc la classe 400, 600 et le mode vaut (400 + 600)/2= 500

Médiane :

Si on cumule les fréquences, on obtient ce qu'on appellera **la Fonction de répartition** et qu'on notera F.

La Fonction de répartition représentera, pour une valeur donnée, la proportion d'individus inférieurs à cette valeur (pour le caractère étudié).

Ainsi la fonction de répartition vaut-elle 0% en 0, 20% en 400, 70% en 600, 95% en 1000, 99% en 2000 et 100% en 5000.

On peut exprimer la fonction de répartition en % comme on vient de le faire ou de manière décimale (c'est alors un nombre compris entre 0 et 1).

On remarque aussi que la Fonction de répartition est une fonction croissante qui varie de 0% à 100% (ou de 0 à 1).

On remarque enfin qu'on ne peut calculer les valeurs de la fonction de répartition que pour les extrémités de classes. Quand il faudra estimer la fonction de répartition pour des valeurs intermédiaires, on recourra à l'interpolation linéaire.

La médiane est la valeur qui partage la population en deux, donc la médiane est la valeur pour laquelle la fonction de répartition vaut 1/2 ou 0,5 (en effet si la proportion de la population "avant" vaut 0,5, il en est de même pour la population "après").

Remarque:
Pour ce qui concerne les moyennes (arithmétiques, quadratiques, harmoniques, géométriques), les écarts type ou moyen et la variance, les calculs se mènent exactement comme dans le cas d'un caractère discret en prenant les **centres** des classes comme modalités.

Quantiles :

On appelle **quantile d'ordre α** (α étant compris entre 0 et 1) la valeur du caractère pour lequel la fonction de répartition vaut α.
Les quantiles sont des mesures de position qui ne tentent pas nécessairement de déterminer le centre d'une distribution d'observations, mais de décrire une position particulière.
On a déjà rencontré un quantile : c'est la médiane, quantile d'ordre 0,5.

Les quantiles se déterminent pratiquement comme on a déterminé la médiane, c'est à dire, en général, par interpolation linéaire, mais pas avec la valeur 0,5 comme pour la médiane, mais avec la valeur α.

Les quantiles d'ordre 1/4,2/4,3/4 seront appelés **quartiles** et seront notés Q_1, Q_2, Q_3.
Vous avez reconnu Q_2, ... c'est la médiane.

Les quantiles d'ordre 1/10,2/10,.....,9/10 seront appelés **déciles** et seront notés D_1 à D_9.
On peut définir des centiles, voire des milliles.

La différence $\mathbf{Q_3-Q_1}$ s'appelle **l'écart interquartile**. D_9-D_1 est l'écart interdécile.
Ces deux écarts sont aussi des mesures de dispersion (comme écart-type et écart moyen).

Représentation graphique :

La représentation graphique d'un caractère continu s'appelle un **histogramme**.
En abscisse, on porte les extrémités de classe, chaque classe sera alors représentée par un rectangle dont la base mesure l'amplitude de la classe et la hauteur la **fréquence relative** (pas la fréquence !).
La surface de chaque rectangle représentera donc la fréquence de chaque classe.
L'histogramme est donc constitué d'une juxtaposition de rectangles.

CARACTERES QUANTITATIFS CONTINUS

Enoncés des Exercices

Exercice 1

Une population comportant 20 entreprises est interrogée sur les dépenses mensuelles effectuées pour l'achat d'un input, afin de fabriquer un bien de consommation. Le tableau suivant donne les dépenses en fonction du nombre d'entreprise. A partir de ces données du tableau (1) calculer :
Le mode, médiane (avec deux manières différentes), les moeynnes et les caractéristiques de dispersion (Ecart moyen, Variance et Ecart-type, Ecart interquartile, Ecart interdécile).

Modalités	Effectifs
1 à 3	8
3 à 5	6
5 à 9	5
9 à 15	1

Exercice 2 :

Une multinationale exerce une activité commerciale sur plusieurs marchés. Elle dispose de 20 filiales dont la répartition est reportée dans le tableau suivant. Ce tableau donne également le chiffre d'affaires réalisés sur les cinq marchés. On utilisant les données suivantes calculer le mode, la médiane, les moyennes et toutes les caractéristiques de dispersion ?

Modalités (M.Euro)	Pays	Effectifs
1 à 3	La France	6
3 à 9	L'Italie	6
9 à 15	Allemagne	5
15 à 21	Canada	3
21 à 23	USA	1

Exercice 3 Pour étudier l'importance des investissements effectués en R et D par les laboratoires cosmétiques, un chercheur a interrogé 40 entreprises. Les montants des dépenses effectués sont reportés dans le tableau suivant :

Tableau 2

Modalités (M. Euro)	Effectifs (les entreprises)
1 à 7	8
7 à 9	6
9 à 11	6
11 à 13	6
13 à 15	9
15 à 25	5

a) Quelle est la nature du caractère étudié ?
b) Calculer pour cette série statistique :
Le mode, la médiane, la Moyenn**a arithmétique**, la variance et l'écart-type, l'écart interquartile, l'écart interdécile

CORRECTIONS

	Exercice 1	Exercice 2	Exercice 3
Mode	2	2	14
Médiane	3,667	7,5	11
Moyenne arithmétique	4,35	8,76	10,95
Moyenne quadratique	5,08	10,73	11,98
Moyenne harmonique	3,18	4,54	8,43
Moyenne géométrique	3,68	6,43	9,73
Ecart moyen	2,09	5,442	3,95
Variance	6,928	38,28	23,6
Ecart-type	2,632	6,187	4,858
Ecart interquartile	3,55	10,75	6,22
Ecart interdécile	6,7	17,1	13

REGRESSION LINEAIRE SIMPLE

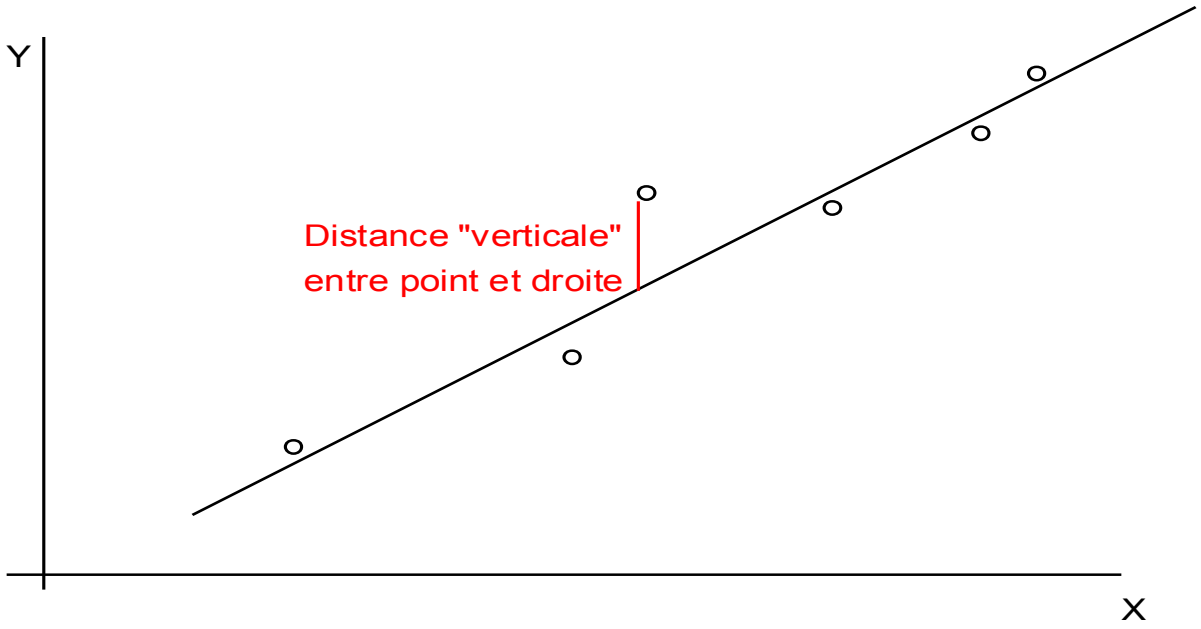
1) POSITION DU PROBLEME :

On s'intéresse ici à deux caractères pour lesquels on dispose de **n** observations simultanées : par exemple, on connaît la consommation et le revenu des ménages Français sur les dix dernières années.

Le problème est le suivant :

- peut-on considérer qu'il existe une relation linéaire (du type **Y = a X + b**) entre les deux caractères ?
- quelles valeurs prennent les coefficients **a** et **b** ?

La première idée qui vient à l'esprit pour commencer à répondre à la question consiste à faire un graphique en portant Y sur l'axe des ordonnées et X sur l'axe des abscisses.



Plus le nuage de points aura une forme allongée et s'orientera autour d'une droite, plus on aura tendance à penser qu'il existe une relation linéaire entre Y et X (rappelons que y = a x +b est l'équation d'une droite).

Une fois cette conviction acquise, il faudra déterminer la "meilleure" droite, celle qui passe "au milieu" du nuage de points.

2) CALCUL DES COEFFICIENTS DE LA DROITE :

Si on la trace à main levée, il y a fort à parier qu'elle sera différente d'un individu à l'autre; il faut donc trouver un moyen de quantifier la distance du nuage de points avec une droite le traversant.

Pour des raisons historiques (qui, a posteriori, ont été justifiées par la théorie, comme on le verra ultérieurement), on a choisi de minimiser la distance "verticale" des points à la droite

On note x_1, x_2 , x_3, ..., x_n les n observations sur le caractère X

On note y_1, y_2 , y_3, ..., y_n les n observations sur le caractère Y

Les points représentés sont donc de coordonnées (x_i, y_i)

La projection verticale d'un point sur la droite d'équation y=ax+b a pour coordonnées (x_i , ax_i+b)

La distance entre un point et sa "projection verticale" (élevée au carré) vaut : (y_i - (ax_i+b))²

La distance "globale" calculée pour l'ensemble du nuage s'obtient en faisant la somme de ces expressions sur les n observations, soit :

$$\sum_{i=1}^{i=n}\left[\left(y_i-(ax_i+b)\right)^2\right]$$

Pour trouver la droite passant "le plus près" possible du nuage de points, il faut trouver les coefficients de la droite (a et b) qui rendent cette somme minimale.

Cette somme est une fonction des deux variables a et b, donc on trouve les valeurs la rendant minimale en annulant les dérivées partielles par rapport à a et à b.

Après calculs, on trouve :

$$a=\frac{\mathrm{cov}(X,Y)}{V(X)} \quad et \quad b=\bar{y}-a\bar{x}$$

Cette méthode pour déterminer les coefficients de la droite s'appelle méthode des moindres carrés ordinaires.

Voyons les notations nécessaires et le détail du calcul sur l'exemple suivant :

Exemple : On a **10** observations simultanées sur les variables X et Y

X	1	2	1	2	3	1	3	2	1	4
Y	2	3	3	4	5	2	5	4	2	7

On calcule $\mathbf{\Sigma x_i}$ ou ΣX = 1+2+1+2+3+1+3+2+1+4 = **20**

On calcule $\mathbf{\Sigma y_i}$ ou ΣY = 2+3+3+4+5+2+5+4+2+7 = **37**

On calcule $\mathbf{\Sigma x_i^2}$ ou ΣX² = 1²+2²+1²+2²+3²+1²+3²+2²+1²+4²= **50**

On calcule Σy_i^2 ou $\Sigma Y^2 = 2^2+3^2+3^2+4^2+5^2+2^2+5^2+4^2+2^2+7^2 =$ **161**

On calcule $\Sigma x_i y_i = (1\times2)+(2\times3)+(1\times3)+(2\times4)+(3\times5)+(1\times2)+(3\times5)+(2\times4)+(1\times2)+(4\times7) =$ **89**

On calcule ensuite les 5 résultats suivants :

$\bar{x} = \frac{\sum x_i}{n}$ *et* $\bar{y} = \frac{\sum y_i}{n}$ **= 2 et 3,7**

$$V(X) = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \frac{50}{10} - (2)^2 = \mathbf{1}$$

$$V(Y) = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{\sum y_i^2}{n} - (\bar{y})^2 = \frac{161}{10} - 3,7^2 = \mathbf{2,41}$$

$$\mathrm{cov}(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - (\bar{x}.\bar{y}) = \frac{89}{10} - (2 \times 3,7) = \mathbf{1,5}$$

Puis $a = \frac{\mathrm{cov}(X,Y)}{V(X)} = \frac{1,5}{1} = 1,5$ *et* $b = \bar{y} - a\bar{x} = 3,7 - (1,5 \times 2) = 0,7$

La droite de régression de Y en X obtenue par la méthode des moindres carrés ordinaires a donc pour équation :

$$Y = 1,5\ X + 0,7$$

3) QUALITE DE LA REGRESSION :

Une fois trouvée l'équation de la droite, on va s'atteler à la seconde question, comment mesurer la qualité de cette relation linéaire qu'on vient de définir, comment tester sa validité ?

Pour répondre à la question, on utilisera l'indicateur suivant : $\boxed{R^2 = \frac{\mathrm{cov}^2(X,Y)}{V(X) \cdot V(Y)}}$

R^2 s'appelle le coefficient de détermination linéaire.
Il est toujours compris entre 0 et 1.
Plus il est proche de 1, meilleure est la régression

Dans l'exemple, $R^2 = 2,25/2,41 =$ **0,9336**

Pour répondre de manière plus précise, on va comparer $(n-2)\frac{R^2}{1-R^2}$ à une valeur de référence issue de la table suivante :

Nombre d'observations	**10**	**15**	**20**	**25**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	+
Valeur de référence	4,7	4,5	4,3	4,2	4,1	4,1	4,0	4,0	4,0	4,0	3,9	3,84

Si $(n-2)\frac{R^2}{1-R^2}$ est plus grand que la valeur de référence, la relation linéaire est valable.

REGRESSION LINEAIRE SIMPLE EXERCICES

Exercice 1 : Un chercheur est intéressé par les deux séries chronologiques suivantes :

	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946
Y enfants morts ayant moins d'un an (en milliers)	60	62	61	55	53	60	63	53	52	48	49	43
X consommation de bière en tonneaux	23	23	25	25	26	26	29	30	30	32	33	31

Donner la droite de régression de Y en X. Calculer R². Tester la pertinence.

Exercice 2 : La réalisation d'un échantillon de 200 observations a permis de calculer les quantités suivantes

$$\sum x_i = 11,34 \quad \sum y_i = 20,72 \quad \sum x_i^2 = 12,16 \quad \sum y_i^2 = 84,93 \quad \sum x_i y_i = 22,13$$

Donner la droite de régression de Y en X. Calculer R². Tester la pertinence.

Exercice 3 : X, Y étant deux séries statistiques, on a trouvé sur 16 observations :

$$\sum x_i = 5,13 \quad \sum y_i = 117,25 \quad \sum (x_i - \bar{x})^2 = 1,27 \quad \sum (y_i - \bar{y})^2 = 4,78 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 1,84$$

Donner la droite de régression de Y en X. Calculer R². Tester la pertinence.

Exercice 4 : X, Y étant deux séries statistiques, on a trouvé sur 16 observations :

$$\sum (y_i - \bar{y})^2 = 4 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 1,8$$
$$\sum x_i = 5 \quad \sum y_i = 117 \quad \sum (x_i - \bar{x})^2 = 1,2$$

Donner la droite de régression de Y en X. Calculer R². Tester la pertinence.

Exercice 5 :

Le tableau 1 donne pour un échantillon de 14 chaînes Hi-Fi le prix X_i de l'amplificateur et le prix Y_i de la chaîne complète. Les prix sont donnés par le tableau suivant :

Tableau 1

X_i	135	147	148	139	171	183	138	145	244	266	166	155	184	143
Y_i	290	420	440	425	450	490	320	420	540	860	540	680	650	510

1) Calculer le prix moyen d'un amplificateur, celui d'une chaîne complète, ainsi que l'écart-type pour chacun de ces caractères.
2) Déterminer l'équation de la droite de régression de Y par rapport à X. Calculer R². Tester la pertinence.
3) Peut-on donner une estimation du prix d'une chaîne Hi-Fi dont l'amplificateur vaut 195.

Vous disposez des informations suivantes : Le tableau 3 donne les valeurs de $(n-2)\frac{R^2}{(1-R^2)}$

Nombre d'observations	10	15	20	25	30	40	50	60	70	90	+	
Valeur de référence	4,7	4,5	4,3	4,2	4,1	4,1	4	4	4	3,9	3,9	

Exercice 6

Le tableau suivant donne des informations communiquées par un responsable de communication d'une entreprise Française. A partir de ce tableau :

Tableau 1

Année	**Budget de communication: x**	**Chiffre d'affaires (y)**
1	7,2	19
2	6	17
3	6,8	18
4	6,6	18
5	6,4	17,5
6	6,2	17,5
7	6,4	18
8	7	18,5
9	6,2	7
10	7,4	20

1) Représenter graphiquement les données de ce tableau, Que peut-on conclure?
2) Calculer $\bar{x}$, $\bar{y}$, Var(x), Var(y) et cov(x,y) ?
3) Déterminer l'équation de la droite de régression de Y par rapport à X. Calculer R². Tester la pertinence ?

LA LOI BINOMIALE

La loi binômiale se définira de la manière suivante :

Considérons une succession d'épreuves identiques et **indépendantes**,

chaque épreuve pouvant se solder soit par un **succès**, soit par un échec.

Si l'on note **p** la **probabilité de succès** à **une** épreuve

et si l'on définit la variable aléatoire X comme étant le **nombre de succès** obtenus **en n épreuves**,

on dira (par définition) que X suit une loi binômiale de paramètres n et p

On note : $X \rightarrow B(n,p)$

(le premier paramètre n représente le nombre d'épreuves, le second p représente la probabilité de succès à une épreuve)

Exemple : Si l'on note X le nombre de "6" obtenu en lançant 12 dés, on sait que X suivra une loi binômiale de paramètres 12 et 1/6.

Nous allons maintenant déterminer la distribution de probabilité d'une loi binômiale, c'est à dire la probabilité qu'elle prenne telle ou telle valeur.

Remarquons tout d'abord que le nombre de succès en n épreuves peut aller de 0 à ... n.

Nous allons donc déterminer P(X = k) pour k prenant toutes les valeurs entre 0 et n

Avoir k succès en n épreuves, c'est aussi avoir n-k échecs.

La probabilité d'avoir un succès est notée p; la probabilité d'avoir un échec sera notée q et vaudra bien sûr 1-p.

La probabilité d'avoir k succès puis n-k échecs lors d'une succession de n épreuves vaudra donc : $\mathbf{p^k q^{n-k}}$.

Mais cette probabilité ne représente qu'un des cas où les n épreuves comportent k succès : c'est celui où tous les succès ont lieu d'abord.

On peut imaginer quantité d'autres cas, par exemple, tous les succès peuvent avoir lieu à la fin, ou au milieu, ou après 3 épreuves, ou en alternance avec les échecs, etc...

De toutes façons, chacun de ces cas aura la même probabilité d'apparition que le premier évoqué, c'est à dire $\mathbf{p^k\, q^{n-k}}$

Pour déterminer la probabilité d'avoir k succès (X=k), il suffit donc de compter tous ces cas.

Or il y a autant de cas que de façons de choisir la place des k succès parmi les n épreuves (la place des échecs est toute désignée une fois choisie celle des succès), il y en a donc : C_n^k.

La probabilité cherchée vaut donc : $P(X = k) = C_n^k p^k q^{n-k}$

Exemple : Si l'on veut appliquer la formule pour déterminer la probabilité qu'il y ait 3 faces "6" lors du lancer de 12 dés, cela donne : $C_{12}^3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^9 = 0{,}1615$.

Nous allons maintenant donner l'Espérance et la Variance d'une loi binômiale (sans calculs) :

$$\mathbf{E(X) = np}$$

$$\mathbf{V(X) = npq}$$

Ce qui donne, toujours dans l'exemple considéré, une espérance de 2 et une variance de 5/3.

LA LOI HYPERGEOMETRIQUE

Considérons une population de N individus partagée en deux : les individus "positifs" en proportion p (donc au nombre de Np) et les individus "négatifs" en proportion 1-p notée q (donc au nombre de Nq).

Supposons qu'on prélève dans cette population un échantillon aléatoire de n personnes, cet échantillon étant prélevé sans remise.

Si on s'intéresse au **nombre de personnes positives dans cet échantillon**, on dira, par définition, qu'il suit une loi **hypergéométrique** de paramètres N,n,p (**on note : X → H (N,n,p)**).

Notez qu'il y **3 paramètres** pour une loi hypergéométrique : le premier, c'est la taille de la population; le deuxième, c'est la taille de l'échantillon; le troisième, c'est la proportion **initiale** d'individus positifs dans la population (initiale car cette proportion est constamment modifiée puisque le tirage est effectué sans remise).

Notez aussi qu'on pourrait parfaitement définir la loi binômiale comme on vient de définir la loi hypergéométrique; la seule différence est qu'alors le tirage devrait être effectué **avec remise**, les individus de l'échantillon représentant des épreuves et les individus positifs, des succès.

Cette remarque sera utile dans la suite, *hypergéométrique : sans remise* et *binômiale : avec remise*.

Cela dit, si n est "petit" par rapport à N (dès que le rapport n/N, appelé **taux de sondage**, est inférieur à 1/10) on pourra assimiler la loi hypergéométrique à la loi binômiale.

La distribution de la loi hypergéométrique est la suivante : $\mathbf{P(X = k) = \frac{C_{Np}^{k} C_{Nq}^{n-k}}{C_{N}^{n}}}$

En effet, le nombre d'échantillons possibles est égal à C_N^n et le nombre de cas favorables s'obtient en multipliant le nombre de façons de choisir les k individus positifs (parmi Np) et les n-k individus négatifs (parmi Nq)

L'**espérance** est égale à : $\mathbf{n \times p}$ (comme la loi binômiale).

La **variance** vaut $\frac{N-n}{N-1} \times npq$ ($\frac{N-n}{N-1}$ s'appelle facteur d'exhaustivité).

LOIS
BINOMIALE et HYPERGEOMETRIQUE
EXERCICES

1) Quelle est la probabilité qu'il y ait 3 personnes nées un Dimanche dans une classe de 19 personnes ?

2) On lance 25 dés. Quelle est la probabilité de faire 7 résultats divisibles par 3 (i.e. 3 ou 6) ?

3) On lance 15 pièces. Quelle est la probabilité de faire "pile" deux fois plus souvent que "face" ?

4) Quels sont les paramètres d'une loi binomiale dont l'espérance vaut 6 et la variance 4 ?

5) Quelle est la variance du nombre de jours de pluie au mois de Mars sachant qu'il y a 64 % de chances qu'il pleuve chaque jour de cette période ?

6) En moyenne, combien y a-t-il de personnes nées un Lundi dans une assemblée de 210 personnes ?

7) Quelle est la probabilité d'avoir 3 étudiants reçus sur 20 quand en moyenne il y en a 4 ?

Exercice 1

Soit une urne contenant 8 boules rouges et 7 boules noires. On tire simultanément 3 boules de l'urne.
On note X le nombre de boules rouges tirées (parmi les 3).Calculer V(X).

Exercice 2

L'oral d'un examen comporte vingt sujets possibles. Le candidat tire trois sujets au hasard ; parmi ces trois sujets il choisit le sujet qu'il désire traiter. Ce candidat a révisé seulement douze sujets. On considère la variable aléatoire X, nombre de sujets révisés parmi les trois sujets tirés.

1- Quelle est la loi de probabilité de X ?
2- Quelle est la probabilité pour que le candidat obtienne au moins un sujet révisé ?

Exercice 3

Douze étudiants se présentent à un examen où le taux de réussite vaut 20 %. Quelle est la probabilité pour que deux soient reçus ?

Exercice 4

Les voyageurs qui louent leur place de chemin de fer l'occupent effectivement dans 90 % des cas. Quelle est la probabilité pour que, sur 20 places louées prises au hasard dans un train, 3 ne soient pas occupées par les voyageurs qui les avaient louées.

Exercice 5

Une variable aléatoire S suit une loi binomiale d'espérance E = 6,3 et d'écart-type $\sigma = 2,1$.
Calculer P(X = 6).

Exercice 6

Quelle est la probabilité d'avoir trois garçons dans un échantillon de 5 personnes prélevé dans une population de 32 personnes dont 14 garçons et 18 filles

- avec remise
- sans remise

Exercice 7

Sachant que la probabilité pour qu'un étudiant soit diplômé est 0,4 , calculer pour un groupe de cinq étudiants, la probabilité pour :

a- qu'aucun ne soit diplômé,
b- qu'un et un seul soit diplômé,
c- que deux soient diplômés,
d- qu'au moins deux soient diplômés,
e- que les cinq soient diplômés.

Exercice 8

Un lapin met au monde une portée de neuf lapereaux, comprenant deux lapereaux noirs, trois blancs et quatre tachetés. Six lapereaux s'échappent. On suppose que chaque lapereau a la même envie et la même possibilité de prendre la clef des champs. Soit X la variable aléatoire qui, à chaque groupe de six lapereaux échappés associe le nombre de lapereaux blancs qui en font partie.

1- Déterminer la loi de probabilité de X .
2- Calculer l'espérance mathématique E(X) et la variance V(X) de cette variable aléatoire.

Exercice 9

Dans une urne, il y a 12 boules : 5 rouges, 3 vertes et 4 bleues. On tire simultanément 3 boules de l'urne. Quelle est la variance du nombre de boules rouges obtenues ?

Exercice 10

On lance 40 noisettes dans 1 kg de chocolat fondu pour faire 4 tablettes de 250 g. Quelle est la probabilité qu'il y ait 9 noisettes dans la première tablette fabriquée ?

CORRECTIONS

1) X → B(19,1/7) P(X=3) = $C_{19}^{3}\left(\frac{1}{7}\right)^{3}\left(\frac{6}{7}\right)^{16}$ = **0,2398**

2) X → B(25,1/3) P(X=7) = $C_{25}^{7}\left(\frac{1}{3}\right)^{7}\left(\frac{2}{3}\right)^{18}$ = **0,1487**

3) X → B(15,1/2) P(X=10) = $C_{15}^{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{5}$ = **0,09164**

4) npq = 4 et np = 6 d'où q = 2/3, p = **1/3** et n = **18**

5) X → B(31;0,64) V(X) = 31×0,64×0,36 = **7,142**

6) X → B(210,1/7) E(X) = **30**

7) 20p = 4, d'où p = 1/5 X → B(20,1/5) P(X=3) = $C_{20}^{3}\left(\frac{1}{5}\right)^{3}\left(\frac{4}{5}\right)^{17}$ = **0,2054**

Exercice 1

X suit une loi hypergéométrique de paramètres 15, 3 et 8/15. V(X) = $\frac{15-3}{15-1}\times 3\times\frac{8}{15}\times\frac{7}{15}$ = **0,64**

Exercice 2

1- X suit une loi hypergéométrique de paramètres 20, 3 et 12/20

2- P(X≥1) = 1- P(X=0) = 1- $\frac{C_{8}^{3}}{C_{20}^{3}}$ = **0,9508**

Exercice 3

Chaque étudiant représente une épreuve indépendante, la probabilité de succès étant de 0,2, on cherche donc la probabilité qu'une loi binomiale de paramètres 12 et 0,2 prenne la valeur 2 :

$$C_{12}^{2}\times 0{,}2^{2}\times 0{,}8^{10} = \textbf{0,2835}$$

Exercice 4

Si l'on note X le nombre de places inoccupées, X suivra une loi binomiale de paramètres 20 et 0,1 et l'on déterminera P(X=3). On peut aussi définir Y le nombre de places occupées et chercher P(Y=17) avec Y suivant une loi binomiale $\mathcal{B}$(20 ; 0,9), ce qui donne le même résultat :

$$C_{20}^{3} \times 0,1^{3} \times 0,9^{17} = \mathbf{0,1901}$$

Exercice 5 :

Si l'écart-type vaut 2,1, la variance vaudra 4,41. Pour une loi binomiale de paramètres n et p, l'espérance vaut np et la variance npq. En faisant le rapport des deux, on obtient q = 4,41/6,3 = 0,7, d'où p = 0,3 et n = 21 . Pour déterminer P(X=6), il ne reste plus qu'à appliquer la formule de la loi binomiale :

$$C_{21}^{6} \times 0,3^{6} \times 0,7^{15} = \mathbf{0,1878}$$

Exercice 6

Avec remise, on a une loi binômiale (5, 14/32). Sans remise, on a une loi Hypergéométrique (32, 5, 14/32)

Avec remise : $P(X = 3) = C_5^3 \left(\frac{7}{16}\right)^3 \left(\frac{9}{16}\right)^2 = \mathbf{0,2650}$ Sans remise : $P(X = 3) = \frac{C_{14}^3 C_{18}^2}{C_{32}^5} = \mathbf{0,2766}$

Exercice 7

Si X est le nombre d'étudiants diplômés, X suit une loi binomiale de paramètres 5 et 0,4.

a) P(X =0) = **0,07776** b) P(X=1) = **0,2592** c) P(X=2) = **0,3456**
d) P(X≥2) = **0,6630** e) P(X=5) = **0,01024**

Exercice 8

1- X suit une loi hypergéométrique de paramètres 9, 6 et 3/9.
2- E(X) = **2** et V(X) = **0,5**

Exercice 9

Le nombre de boules rouges obtenues est une variable aléatoire suivant une loi hypergéométrique de paramètres 12, 3 et 5/12. Sa variance est donc égale à : $\frac{12-3}{12-1} \times 3 \times \frac{5}{12} \times \frac{7}{12} = \mathbf{0,5966}$

Exercice 10

Le nombre de noisettes présentes dans une tablette suit une loi binômiale de paramètres 40 et 1/4, d'où :

$$P(X = 9) = C_{40}^{9} \left(\frac{1}{4}\right)^9 \left(\frac{3}{4}\right)^{31} = \mathbf{0,1397}$$

LA LOI NORMALE ET LOIS DERIVANT DE LA LOI NORMALE

1. Introduction

La Loi Normale est une variable continue, on l'appelle aussi loi de Gauss, loi de Laplace-Gauss.
Une variable suivra une loi normale si : elle dépend d'un grand nombre de causes, indépendantes, dont aucune n'est prépondérante et dont les effets s'additionnent (ces conditions définissant la loi normale sont appelées **conditions de Borel**).
Une Loi normale possède deux paramètres : le premier correspond à son espérance (sa "moyenne") et sera donc noté : m; le second correspond à son écart-type (à la racine carrée de sa Variance) et sera donc noté σ^2.
Une loi normale de paramètres m et $,\sigma^2$ sera notée : **N (m,σ^2).**
On a donc : **E(X) = m** **V(X) = σ^2**

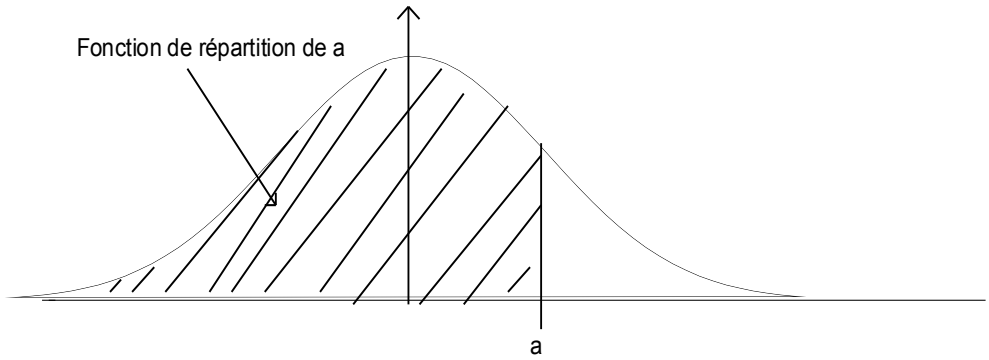
Comme c'est une variable aléatoire continue, les probabilités ponctuelles sont nulles et l'on définit une densité de probabilité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$.

Quand on aura à manipuler une loi normale, on utilisera la propriété suivante :

Si X $\rightarrow$ N(m, σ^2), en posant $Z=\frac{X-m}{\sigma}$, on aura Z $\rightarrow$N(0,1)

Ainsi, par un changement de variable, on peut ramener une loi normale quelconque à une loi normale de paramètres 0 et 1 (appelée **loi normale centrée réduite**).
Cette opération s'appelle : "centrer réduire".
Compte tenu de cette propriété, seule reste à étudier la loi normale centrée réduite N(0,1) :

Si l'on trace la courbe représentative de la densité de probabilité, on obtient une courbe en forme de cloche symétrique par rapport à l'axe des ordonnées :



On sait que la surface sous cette courbe représente la probabilité, donc on peut définir la fonction de répartition (probabilité d'être **avant** une valeur donnée) comme étant la surface sous la courbe de -∞ à la valeur considérée. On note cette fonction de répartition : **F**.

Ainsi : **P(Z $\leq$ a) = F(a)** **P(Z $\geq$ b) = 1-F(b)** **P(a $\leq$ Z $\leq$ b) = F(b) - F(a)**

En effet, la probabilité d'être après b est égale à 1 moins la probabilité d'être avant et la probabilité d'être entre a et b est égale à la probabilité d'être avant b moins la probabilité d'être avant a.

Notez que l'on a utilisé des inégalités "larges" ($\leq$ et $\geq$) mais que l'on aurait pu sans problème utiliser des inégalités strictes (< et >) puisqu'en rajoutant une valeur ponctuelle, on ne change pas la probabilité dans le cas d'une variable continue.

On a donc ramené tout calcul de probabilité sur une loi normale à un calcul de fonction de répartition de N(0,1).

Remarquons aussi que la probabilité d'être avant -z (**F(-z)**) est égale (puisque la courbe est symétrique) à la probabilité d'être après t, c'est à dire à **1 - F(z)**.

Cette remarque permettra de déduire la fonction de répartition d'un nombre négatif de celle du nombre positif opposé.

Pour récapituler, si l'on connaît la fonction de répartition de N(0,1) pour les nombres positifs, on est capable de faire tout calcul de probabilité concernant une loi normale quelconque.

Une Table nous donne pour toute valeur de t positive la valeur de F (t) : les deux premières décimales de t se trouvent sur la première colonne et la troisième décimale sur la première ligne.

Par exemple : F(1,12) = 0,8686 (ligne 1,1 et colonne 0,02)
F(0,37) = 0,6443 (ligne 0,3 et colonne 0,07)
F(-1,12) = 1-0,8686 = 0,1314

2. Combinaison de lois normales :

Toute combinaison linéaire de lois normales indépendantes est une loi normale; les paramètres se déterminent par manipulation d'espérances et de variances.

Par exemple, si $X \rightarrow N(m, \sigma)$, si $Y \rightarrow N(m', \sigma')$, si X et Y sont indépendantes, alors

$$X + Y \longrightarrow N(m + m', \sqrt{\sigma^2 + \sigma'^2})$$

En effet, le premier paramètre d'une loi normale, c'est l'espérance. Or l'espérance de la somme est égale à la somme des espérances (*cf. Variables aléatoires, Manipulation des opérateurs*).
Le second paramètre, c'est l'écart-type, la racine carrée de la variance. Or la variance de la somme se ramène pour des variables indépendantes à la somme des variances (la covariance étant nulle) (*cf. idem qu'au-dessus)*

Remarque : On aurait, par exemple, dans les mêmes conditions :

$$X - Y \longrightarrow N(m - m', \sqrt{\sigma^2 + \sigma'^2})$$

$$X + 2Y \longrightarrow N(m + 2m', \sqrt{\sigma^2 + 4\sigma'^2})$$

$$2X - 3Y \longrightarrow N(2m - 3m', \sqrt{4\sigma^2 + 9\sigma'^2})$$

3. Approximatio : Binômiale par Normale :

On peut approcher une loi binômiale de paramètres n et p par une loi Normale de paramètres np et $\sqrt{npq}$ dès que n est suffisamment grand et p suffisamment proche de 0,5 (on prendra n > 100 et npq > 5)

$$\boxed{\text{Si } n > 100 \text{ et } npq > 5, B(n,p) \approx N(np, \sqrt{npq})}$$

4. LOIS DERIVANT DE LA LOI NORMALE

4.1. La loi du Chi-Deux :

Une loi du **Chi-deux** de paramètre n (on dit "à n degrés de liberté") est définie comme la somme des carrés de n lois normales centrées réduites indépendantes.

Si X_1, X_2, ...,X_n sont des lois normales centrées réduites (N(0,1)) indépendantes,

alors $Y = \sum_{i=1}^{i=n} X_i^2$ suit une loi du Chi-deux à n degrés de liberté (ddl) notée : χ^2**(n)**

L'Espérance d'une loi du Chi-deux à n ddl vaut : n et sa variance vaut : 2n.

Si les lois X_i ne sont pas indépendantes, mais liées entre elles par p relations linéaires **indépendantes** (c'est à dire telles qu'aucune relation ne puisse se déduire des autres), Y suivra toujours une loi du Chi-deux, mais à **n-p** ddl.

4.2 La Loi de Student :

Une loi de Student de paramètre n (on dit "à n degrés de liberté") se définit comme le rapport entre une loi normale centrée réduite et la racine d'une loi du Chi-deux à n ddl "réduite" (c'est-à-dire divisée par n); ces deux lois étant indépendantes.

Ainsi, si $X \rightarrow N(0,1)$, si $Y \rightarrow \chi^2(n)$ et si X et Y sont indépendantes,

alors $Z = \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ suit une loi de Student à n ddl notée : **t(n)**

La densité de probabilité d'une loi de Student est symétrique (comme pour N(0,1)).

L'espérance vaut donc : 0 et la variance : n / n-2

Quand n devient très grand (tend vers l'infini), la loi de Student tend vers N(0,1)

EXERCICES LOI NORMALE

Exercice 1 :
Quelle est la probabilité qu'il y ait entre 312 et 435 hommes dans un échantillon aléatoire de 653 personnes prélevé dans la population française avec remise ? (on rappelle que 54 % des français sont des françaises)

Exercice 2 :
Quelle est la probabilité d'avoir entre 12 et 23 personnes célibataires dans un échantillon aléatoire de 100 individus prélevé avec remise si la proportion de célibataires est de 19 % dans la population totale ?

Exercice 3 :
Sur un échantillon de 78 personnes, quelle est la probabilité d'en avoir entre 6 et 9 nées un Dimanche ?

Exercice 4 : Quelle est la probabilité que le nombre moyen d'enfants sur un échantillon aléatoire prélevé avec remise de 121 personnes soit compris entre 2 et 2,1 sachant que sur la population totale, la moyenne de ce caractère est de 2 avec un écart-type de 0,5 ?

CORRECTIONS

Exercice 1 :
Si on appelle X le nombre d'hommes dans un échantillon aléatoire de 653 personnes, on sait que X suit une loi binômiale de paramètres 653 et 0,46 qu'on peut approcher par une loi normale de paramètres 300,38 et $\sqrt{162,2052}$, d'où :

$$P(312 \leq X \leq 435) = P(\frac{312-300,38}{\sqrt{162,2052}} \leq T \leq \frac{435-300,38}{\sqrt{162,2052}}) = F(10,57) - F(0,91) = \mathbf{0,1814}$$

Exercice 2 :
Dans cet exercice il s'agit d'une approximation de la normale par la loi binômiale (loi binômiale de paramètres 100 et 0,19). La table de la loi normale et le calcul d'une probabilité p(12<X<23) .

$$P(12 \leq X \leq 32) = P(\frac{12-19}{\sqrt{15.39}} \leq T \leq \frac{23-19}{\sqrt{15.39}}) = F(0.2599) - (1 - F(0,4548)) = \quad \mathbf{0,8063.}$$

Exercice 3 :
Dans cet exercice il s'agit d'un calcul de probabilité pour une loi binômiale de paramètres 78 et 1/7.
On utilise la même méthode de l'exercice 2. On trouve 0,1966

Exercice 4 :

Dans cet exercice il s'agit d'un calcul de probabilité pour une loi binômiale de paramètres 2 et 0,5/11= $0.5 / \sqrt{121}$.

On utilise
la même méthode de l'exercice 2. On trouve 0,4861

ÉCHANTILLONNAGE et ESTIMATION

Introduction :
Dans une étude statistique, un dénombrement complet de la population est très souvent pratiquement impossible, soit parce que la population totale est inconnue, soit parce qu'elle comprend beaucoup d'individus pour qu'une telle étude soit complètement réalisable. Toutefois, le but d'une étude statistique est d'obtenir les connaissances sur l'ensemble de la population. Or, si une étude sur l'ensemble de la population est difficilement envisageable, il nous faut malgré tout trouver d'autres moyens pratiques d'y parvenir. Un moyen efficace est de procéder à un **échantillonnage**, qui consiste à choisir parmi les éléments de la population un certain nombre d'unités pour lesquelles nous obtiendrons des observations.
Si l'échantillon étudié est bien choisi, les observations permettront d'acquérir les connaissances voulues sur la population à étudier avec un degré spécifié de précision. Le but de ce chapitre est de présenter les différentes méthodes d'échantillonnage et d'estimation.

1. Méthodes d'échantillonnage :

On distingue deux grandes catégories de méthodes d'échantillonnage :

- l'échantillonnage par choix raisonné
- l'échantillonnage aléatoire.

1.1. l'échantillonnage par choix raisonné

Les méthodes l'échantillonnage par choix raisonné incluent diverses techniques qui consistent à construire l'échantillon sur la base d'informations connues relatives à la population étudiée. Ces méthodes comportent une part d'arbitraire ne permettant pas d'évaluer la précision des estimations, mais elles présentent dans certains cas des avantages de coût et de rapidité par rapport à la méthode d'échantillonnage aléatoire.
L'échantillonnage par choix raisonné est aussi appelé échantillonnage empirique. La méthode principale est celle des quotas. Selon cette méthode l'enquêteur sélectionne les unités, en fonction de quotas qui lui sont données. Dans le cas d'une enquête auprès des ménages ou d'individus, ces quotas portent généralement sur des critères **socio-démographique** tels que **le sexe**, **l'age** ou la catégorie **socio-professionnelle**. Ils sont établis à partir de statistiques officielles et visent à constituer un échantillon possédant la même structure que la population. Dans cette méthode d'échantillonnage le hasard intervient d'une façon très limitée dans la sélection des unités de la population qui feront partie de l'échantillon.

1.2. Échantillonnage aléatoire

L'échantillonnage aléatoire correspond à des méthodes de tirage de l'échantillon où chaque unité de la population à une probabilité positive et connue d'être sélectionnée. Ces méthodes permettent non seulement d'estimer les paramètres, mais encore d'obtenir une mesure de l'erreur susceptible d'avoir été commise.
Il existe trois types d'échantillonnage aléatoire : échantillonnage aléatoire simple, échantillonnage stratifié et l'échantillonnage par grappes.

1.2.1. Echantillonnage aléatoire simple

L'échantillonnage aléatoire simple, ou échantillonnage probabiliste simple est basé sur le principe que tous les éléments de la population ont une probabilité égale (non nulle) de faire partie de l'échantillon.
La population considérée est généralement finie. Soit N le nombre d'unités qui composent la population considérée. Au cours d'un tirage aléatoire, on attribuera à chaque unité de la population la même probabilité d'être choisie soit 1/N. En prélevant au hasard un échantillon de taille n d'une population de N unités, les valeurs obtenues pour les n tirages sont aléatoires. Si l'extraction est réalisée sans remettre les unités tirées de la population, il s'agit d'un échantillon sans remplacement. Si, en revanche, l'extraction est faite avec remise, l'échantillon est avec remplacement.
L'échantillonnage avec remise est utilisé très rarement en pratique, car il y a peu d'intérêt de détenir une même unité deux fois dans l'échantillon.

Pour effectuer un échantillonnage aléatoire simple, il faut d'une part, avoir accès au préalable à une liste complète des éléments de la population et d'autre part, utiliser une méthode de tirage qui garantisse la même probabilité de sélection à tous les éléments de la liste.

1.2.2 Echantillonnage stratifié

L'échantillonnage stratifié consiste à découper la population en strates ou classes homogènes par rapport à l'ensemble de la population puis à réaliser dans chaque strate un échantillonnage aléatoire simple. La méthode d'échantillonnage stratifiée est généralement utilisée lorsque la population étudiée est hétérogène à certains égards. La stratification nécessite donc une connaissance préalable de la structure de cette dernière.
On procède à l'échantillonnage stratifié pour plusieurs raisons. Par exemple, on a parfois besoin d'obtenir des résultats sur un sujet donné pour différentes régions géographiques d'un pays. Dans ce cas, on considère chacune des différentes subdivisions géographiques comme une strate et on procède à un échantillonnage aléatoire à l'intérieur de chaque strate.

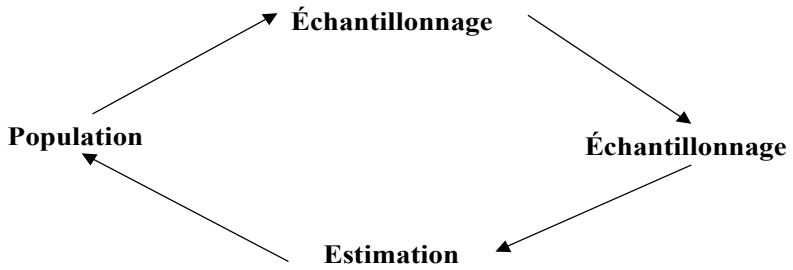
1.2.3 Echantillonnage par grappes

L'échantillonnage par grappes consiste à tirer au hasard des ensembles d'unités de la population, ou grappes, et ensuite à mener l'enquête sur toutes les unités de ces grappes. Les grappes sont souvent constitués par des unités de type géographique comme les quartiers d'une ville. La méthode consiste à diviser une ville en quartiers, puis à sélectionner les quartiers qui feront partie de l'échantillon. On mènera ensuite l'enquête sur toutes les personnes ou ménages, habitant dans les quartiers choisis.
Il y a deux raisons principales de procéder à un échantillonnage par grappes. Dans beaucoup d'enquêtes, il se trouve qu'il n'existe pas une liste complète et fiable des unités de la population pour baser l'échantillonnage, et qu'il est excessivement coûteux de construire une telle liste. Par exemple, dans beaucoup de pays, y compris les pays industrialisés, il est rare que des listes complètes et à jour de la population, des logements ou des exploitations agricoles par exemple soient disponibles.
Le choix de la méthode d'échantillonnage (raisonné, aléatoire, stratifié, par grappe etc,) et donc le choix des unités de la population qui seront observées n'est qu'un des deux aspects du problème des sondages. Un autre aspect est celui du choix de la méthode pour résumer les observations obtenues afin d'obtenir l'estimation la plus proche possible de l'information recherchée. Dans la suite, on examine l'estimation des moyennes et des proportions à partir d'un échantillon aléatoire

simple. La généralisation à d'autres modes d'échantillonnage peut être trouvée dans les ouvrages spécialisés traitant des méthodes d'enquêtes.

2. Estimation

La procédure d'utilisation des informations obtenues à partir de l'échantillon qui permet de déduire des résultats concernant l'ensemble de la population est appelée estimation.
Le graphique suivant montre la relation entre échantillonnage et estimation. L' « échantillonnage » est le passage de la population à l'échantillon, et l'estimation est le passage inverse de l'échantillon à la population.



La valeur inconnue d'une population, à estimer à partir d'échantillon, est appelée un paramètre. Souvent le paramètre à estimer est une moyenne, un total, un pourcentage, un écart type ou une variance. Le paramètre de la population est estimé à partir d'une statistique calculée sur la base d'un échantillon. Un paramètre est donc une caractéristique de l'échantillon. Par exemple, le revenu moyen en France est un paramètre de la population alors que le revenu moyen d'un échantillon représentatif des français est une statistique.

On conclusion, quand un **paramètre** (moyenne, variance, proportion, etc..) est inconnu, on cherche à l'estimer.
On définit un **estimateur**, qui, calculé sur un échantillon, fournit une **estimation**.
Un **estimateur** est une **variable aléatoire** (prenant différentes valeurs en fonction des échantillons prélevés).
Une **estimation** est une valeur numérique, c'est la valeur prise par un estimateur pour un échantillon donné.

On demandera à un estimateur d'être **sans biais**, c'est à dire d'être "en moyenne" égal à la valeur du paramètre qu'il estime; en termes mathématiques, **Espérance (Estimateur) = Paramètre**.
Exemple :

$$E(\overline{X}) = m$$

La **qualité** d'un estimateur se traduit par le fait qu'il ne fluctue pas trop au gré des échantillons prélevés. En d'autres termes, la qualité d'un estimateur est mesurée par sa **Variance**. Le meilleur estimateur sera celui dont la variance est la plus faible (cette variance est limitée par une borne inférieure, appelée **Borne de Cramer-Rao**.

ESTIMATION DE MOYENNES

1.Estimation

On étudie un caractère quantitatif X sur une population de taille N (la valeur de N est éventuellement $\propto$).
Le caractère X y possède une moyenne **m** (inconnue) et un écart-type σ.

Dans cette partie, notre objectif consistera à estimer m. Pour cela, on prélève un échantillon de n individus (**prélevé avec ou sans remise**); on estimera m (la moyenne inconnue du caractère X sur la population totale) par, qui désigne la moyenne du caractère X observée sur l'échantillon.

Pour être plus précis, on notera X_1, X_2,....., X_n les valeurs prises par le caractère respectivement pour le 1er, le 2ème,....., le nème individu d'**un** échantillon.

On notera $\overline{X}=\frac{\sum X_i}{n}$ la moyenne observée sur l'échantillon. $\overline{X}$ **sera l'estimateur de m**.

Si on calcule l'espérance et la variance, on trouve : $E(\overline{X})=m$ *et* $V(\overline{X})=\frac{\sigma^2}{n}$

2.Intervalles de confiance ***:***
Soit m un paramètre à estimer de la population et $\overline{X}$ son estimateur à partir d'un échantillon aléatoire. On cherche à évaluer la précision de $\overline{X}$ comme estimateur de m en construisant un intervalle de confiance autour de l'estimateur, qui souvent s'interprète comme une marge d'erreur.

On déterminera un I.C. (un intervalle de confiance) pour m dans le cas d'un échantillon prélevé avec remise et en supposant, soit que le caractère **X suit une loi normale** sur la population totale (hypothèse dite de normalité), soit que **n est supérieur à 30** (échantillon de taille suffisante).

Dans les deux cas, on peut en déduire que $\overline{X}$ suit une loi normale; dans le premier cas, par combinaison de lois normales et dans le deuxième cas, par application du Théorème Central Limite.

$$\overline{X} \rightarrow N(m,\frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

A partir de la loi de distribution de l'estimateur $\overline{X}$, on détermine un intervalle calculé sur la base de l'échantillon tel que la probabilité soit importante qui englobe la vraie valeur du paramètre recherché.
Soit ($\overline{X}$-e , $\overline{X}$ +e) cet intervalle et (1-α) la probabilité d'appartenance, on peut dire que la marge d'erreur e est liée à α par la probabilité :

$$P(\overline{X}-e \leq m \leq \overline{X}+e)= 1 - \alpha$$

Cela permet de définir un I.C. pour m (avec un niveau de confiance égal à 1 - α) :

$$\left[\overline{X} - u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X} + u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]$$

Cette formule suppose σ connu, or, en général, dans la pratique, quand m est inconnue, σ l'est aussi. Il faut donc trouver un moyen d'**estimer** σ lorsque celui-ci est inconnu.

La première idée qui vient à l'esprit consiste à estimer σ (écart-type du caractère sur **la population totale**) par l'écart-type observé pour le caractère **sur l'échantillon**, qu'on notera : **S'**.

Mais S' ne convient pas; en effet, si l'on calcule E(S'²), on trouve : $\frac{n-1}{n} \times \sigma^2$. Donc S'² n'est pas un estimateur **sans biais** de σ². Pour avoir un estimateur sans biais, il faut "corriger" S', et prendre S = $S' \times \sqrt{\frac{n}{n-1}}$ (démonstration voir cours)

Dans la pratique : $S'^2 = \frac{\sum (X_i - \overline{X})^2}{n} \Rightarrow S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} \Rightarrow S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$

On démontre que $(n-1)\frac{S^2}{\sigma^2}$ suit une loi du chi-deux à n-1 ddl.

En conséquence, quand on remplace σ par S dans $\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$,

on déduit que $\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}}$ suit une loi de Student à n-1 ddl.

Ainsi, la forme d'un I.C. (au niveau 1-α) est-elle : $\left[\overline{X} - t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S}{\sqrt{n}}\right]$

Remarque : Dans le cas d'un échantillon prélevé sans remise, on trouve , dans le cas σ connu et **si n>30**, la formule suivante pour un I.C. pour m :

$$\left[\overline{X} - u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad , \quad \overline{X} + u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right]$$

Dans le cas où σ est inconnu, on ne peut, en toute rigueur, établir aucune formule. Dans la pratique, dès que le **taux de sondage** (n/N) est faible, on appliquera les formules données pour un échantillon prélevé avec remise.

3. Taille d'échantillon :

Pour déterminer la taille d'échantillon nécessaire pour atteindre une précision absolue donnée, on rend le rayon de l'I.C. inférieur à la précision Δ souhaitée. Dans le cas d'un échantillon avec remise on a la formule suivante :

$$\left(\frac{u_{1-\frac{\alpha}{2}} \times \sigma}{\Delta} \right)^2 = n$$

Si l'échantillon est prélevé sans remise en notant n_0 la valeur trouvée ci-dessus , on obtient :

$$n_0 \times \frac{N}{N + n_0 - 1} = n$$

EXERCICES D'ESTIMATION DE MOYENNES 1

1) Le prix moyen d'un bien « A» observée sur un échantillon aléatoire de 101 magasins prélevé avec remise parmi 50000 est de 1,72 euro avec un écart-type de 0,23. Donnez la borne inférieure d'un intervalle de confiance à 98 % pour le prix moyen sur l'ensemble de la population.

2) Quelle taille d'échantillon doit on prélever sans remise pour connaître le prix moyen d'une voiture à 10 € près grâce à un intervalle de confiance à 95 % sachant que l'écart-type est de 100 € sur toute la population interrogée dont le nombre est de 30000?

3) Donnez le rayon de l'intervalle de confiance à 96 % pour le poids moyen des Français obtenu à partir d'un échantillon aléatoire sans remise de 567 personnes si la variance du poids est de 1,5 sur l'ensemble de la population Française.

4) Quelle est la taille de l'échantillon avec remise nécessaire pour connaître une moyenne à 0,5 près sachant que la variance du caractère étudié est de 3 sur la population totale et qu'on désire un niveau de confiance de 90 % ?

5) Donnez la borne inférieure de l'intervalle de confiance à 98 % pour la moyenne d'un caractère sur une population totale de 10000 individus si la moyenne observée sur un échantillon de 123 personnes prélevé sans remise est de 12 et la variance sur la population totale de 4.

Réponses

INDICATIONS :

1) Réponse : 166,56 cm, Formule 8 avec n=101 $\overline{X}$=172 s'=23 et t(100)=2,365
2) Réponse 379.315 donc 380, Formules 9 et 10 avec u=1,96 Δ=10 σ=100 et N=30000
3) Réponse 0.10595, Formule 7 avec u=2,06 n=567 $\sigma = \sqrt{1,5}$ et N=58.000.000
4) Réponse 32.2752 donc 33, Formule 9 avec u=1,6449 Δ=0,5 $\sigma = \sqrt{3}$
5) Réponse 11.58, Formule 7 avec u=2,33 n=123 $\sigma = \sqrt{4}$ et N=10000

EXERCICE D'ESTIMATION DE MOYENNS 2
Enoncés des Exercices

Exercice 1 :

Un important fabricant d'appareils ménagers a besoin d'une estimation précise et systématique des ventes au détail de ses appareils de façon à faciliter son planning de production. En conséquence, il envisage d'utiliser un échantillon de points de vente au détail et d'obtenir ainsi les ventes sur une base mensuelle.

Afin de disposer de bases de réflexion pour réaliser l'enquête définitive, on effectue un pré-test sur un échantillon de 60 détaillants.

Les résultats sont les suivants :

- Vente moyenne par magasin du mois passé : 18,4 appareils
- Ecart-type des ventes dans l'échantillon : 5,35 appareils

1- Quelle serait votre estimation des ventes d'appareils si l'on suppose que l'on dispose de 28. 000 points de vente ? (vous donnerez un intervalle de confiance à 95 %).

2- Le fabricant souhaite obtenir l'estimation des ventes mensuelles moyennes par magasin ± 1 appareil près avec un risque d'erreur de 5 %. Quelle taille d'échantillon faudrait-il prendre pour atteindre cette précision dans le résultat ?

Exercice 2 :

Une enquête devrait permettre de connaître la surface moyenne des exploitations agricoles d'un pays avec une précision de 5 % et un intervalle de confiance à 95 %. L'auteur de l'enquête part précipitamment pour l'étranger; il n'a pas eu le temps de calculer la taille de l'échantillon permettant d'opérer ce sondage, mais il a laissé la note suivante : "l'écart-type représente environ la moitié de la moyenne". Calculer l'effectif de l'échantillon.

Exercice 3 :

La Chambre Syndicale des horticulteurs français souhaite mettre sur pied une enquête périodique pour évaluer l'importance du marché total des arbres, plantes et fleurs pour l'extérieur, achetés par les ménages en France.

On se limitera dans ce texte au marché des rosiers.

1- Pour déterminer la taille d'échantillon nécessaire pour évaluer avec une bonne précision le marché total du rosier, la Chambre Syndicale décide dans un premier temps d'interroger par l'intermédiaire d'une enquête d'un grand institut de sondage un échantillon de 2000 ménages français, que l'on supposera aléatoire simple, sur leurs achats de rosiers au cours des 12 mois précédant l'enquête.

Soit yi le nombre de rosiers achetés au cours des 12 mois précédant l'enquête par le $i^{ème}$ ménage interrogé (i = 1,....,2000).

On obtient les résultats suivants :

$$\sum_{i=1}^{i=2000} y_i = 1260 \qquad \sum_{i=1}^{i=2000} y_i^2 = 19389,5$$

Estimer par un intervalle de confiance bilatéral symétrique à 95 %

m = nombre moyen de rosiers achetés en douze mois par ménage en France

ainsi que le nombre total annuel M de rosiers achetés par les ménages français (le nombre de ménages français est évalué à 20 586 000).

2- Dans l'échantillon de 2000 ménages, 1792 ménages n'ont jamais acheté un rosier. En déduire un intervalle de confiance à 95 % pour p = proportion de ménages ayant acheté au moins une fois un rosier en douze mois.

3- En utilisant les résultats de la préenquête, calculer la taille n d'échantillon nécessaire pour estimer le marché total avec une précision relative de 10 % (on prendra un degré de confiance de 95 %).

CORRECTIONS

Exercice 1 :

1) Il s'agit de déterminer un IC pour une moyenne avec l'écart-type de la population inconnu . Cela donne un IC allant de 17,004 à 19,79 si l'échantillon est prélevé avec remise.

$$18{,}4 \pm 2{,}003 \times \frac{5.35}{\sqrt{59}} = \overline{X} \pm t_{1-\alpha/2}(n-1) \times \frac{S'}{\sqrt{n-1}} = \left[\mathbf{17{,}004 ; 19{,}79}\right]$$

2) Pour répondre à cette question, on est obligé de supposer l'échantillon prélevé avec remise et on estimera σ l'écart-type inconnu de la population totale par s son estimation sans biais.
La formule avec s, u = 1,96 et Δ = 1 donne une taille minimale d'échantillon de **109**.

$$\left(\frac{u_{1-\frac{\alpha}{2}} \times \sigma}{\Delta} \right)^2 = n$$

Exercice 2 :

La précision d'une estimation est donnée par le rayon de l'intervalle de confiance, si la précision doit être de 5%, cela signifie que le rayon de l'IC doit valoir 5% de m au maximum, soit :

$$1{,}96 \times \frac{\sigma}{\sqrt{n}} \leq 0{,}05 \times m \qquad \text{d'où} \qquad n \geq \mathbf{385}\ (384{,}16)$$

Exercice 3 :

1) On estimera m par 1260/2000 = **0,63**
Pour déterminer un IC, on a besoin de S' = **3,049**, et on applique la formule en assimilant la loi de Student à 1999 ddl à une loi normale centrée réduite, ce qui donne un IC allant de **0,4963 à 0,7637** pour m

$$0.63 \pm 1.96 \times \frac{3.049}{\sqrt{1999}} = m = \overline{Y} \pm t_{1-\alpha/2}(n-1) \times \frac{S'}{\sqrt{n-1}} = \left[\mathbf{0.4963; 0.7637}\right]$$

et un IC allant de 0.4963*20 586 000= **10216832 à 0.7637*20 586 000=15721528** pour M.

2)Une simple application de la formule suivante :

$$\left[f \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f(1-f)}{n-1}} \right]$$

avec :
la proportion p sera estimée par f = 208/2000 = **10,4%** , on obtient un IC allant de **9,1 à 11,7 %**.

3) Pour avoir une précision relative de 10%, il faut que le rayon de l'intervalle de confiance soit inférieur ou égal à 0,1 × m, d'où :

$1{,}96\times\frac{\sigma}{\sqrt{n}}\leq 0{,}1\times m \Rightarrow n\geq\left(\frac{1{,}96}{0{,}1}\right)^2\times\left(\frac{\sigma}{m}\right)^2$ = **8997,998** en estimant σ et m

ESTIMATION DE MOYENNES 3
Enoncés des Exercices

Exercice 1 :

Pour connaître la taille moyenne des français, on a prélevé un échantillon aléatoire avec remise de 35 personnes sur lequel on a trouvé 1,75 m en moyenne avec un écart-type de 3 cm.
Donner une estimation par IC à 95 %.

Exercice 2 :
Même question si l'on connaît l'écart-type de la population totale (= 3,2 cm).

Exercice 3 :

On veut connaître le salaire moyen des français salariés à 10 € près. Sachant que l'écart-type du caractère salaire est de 500 € sur la population française, quelle taille d'échantillon doit-on prélever (niveau : 95 %)

a- avec remise ?
b- sans remise ?

Exercice 4:
Considérant que la taille moyenne des français est de 1,70 m. Quelle est la probabilité que, sur une classe de 42 élèves (considérée comme un échantillon aléatoire prélevé sans remise), on observe une taille moyenne supérieure à 1,75 m ?

On connaît bien sûr l'écart-type du caractère taille sur la population française totale, soit 10 cm.

CORRECTIONS

Exercice 1 :

Il s'agit de déterminer un Intervalle de confiance pour une moyenne avec un écart-type sur la population totale inconnu et ce, pour un échantillon prélevé avec remise et avec un niveau de confiance de 95%. La formule s'applique donc avec une moyenne sur l'échantillon de 175 (cm), n = 35, s' = 3 et t = 2,042 (fractile d'ordre 0,975 d'une loi de Student à 34 ddl obtenu par interpolation linéaire entre 30 et 40).
Ce qui donne un IC allant de **173,954 cm à 176,046 cm**

Exercice 2 :

Si l'écart-type de la population totale est connu, on utilisera la formule avec n = 35, σ =3,2 et u = 1,96 fractile d'ordre 0,975 de la loi normale centrée réduite, ce qui donne un IC de **173,94 à 176,06 cm.**

$$\overline{X} \pm u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} = [173{,}94 \ ; 176{,}06]$$

Exercice 3 :

Il s'agit de déterminer la taille d'échantillon nécessaire pour atteindre une précision de 10 F lors de l'estimation de la moyenne inconnue des salaires français par un à 95%.

a) On appliquera donc la formule avec σ = 500, u =1,96 et Δ = 10, ce qui donne n = **9604**

$$\left(\frac{u_{1-\frac{\alpha}{2}} \times \sigma}{\Delta} \right)^2$$

b) **Sans remise**, on corrige la valeur précédente par un facteur N / N+n-1 avec N = 60000000, ce qui donne une taille minimale de **9603.**

$$n_0 \times \frac{N}{N + n_0 - 1}$$

Exercice 4 :

On sait que la taille moyenne sur un échantillon de 42 éléments suit une loi normale de paramètres 170 et $\frac{10}{\sqrt{42}}$, d'où : $P(\overline{X} \geq 175) = P\left(T \geq \frac{175-170}{10/\sqrt{42}} \right) = 1 - \Pi(3{,}24) = 1 - 0{,}999394 = \mathbf{0{,}000606}$

ESTIMATION DE MOYENNES : Formules

(6) $\left[\overline{X} \pm u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]$: Intervalle de confiance au niveau 1-α pour une moyenne inconnue si l'échantillon est prélevé avec remise et σ connu

(7) $\left[\overline{X} \pm u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right]$: idem si l'échantillon est prélevé sans remise

(8) $\left[\overline{X} \pm t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S}{\sqrt{n}}\right]$ ou $\left[\overline{X} \pm t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S'}{\sqrt{n-1}}\right]$: idem que (6) si σ inconnu

(9) $\left(\frac{u_{1-\frac{\alpha}{2}} \times \sigma}{\Delta}\right)^2$: Taille minimale d'échantillon avec remise nécessaire pour atteindre une précision absolue Δ

(10) $n_0 \times \frac{N}{N+n_0-1}$: idem si l'échantillon est prélevé sans remise en notant n_0 la valeur trouvée en (9)

$\overline{X}$ = Moyenne observée sur l'échantillon	**σ** = Ecart-type sur la population totale
n = Taille de l'échantillon	**N** = Taille de la population totale
S' = Ecart-type observé sur l'échantillon	**S** = Estimation de σ = $s' \times \sqrt{\frac{n}{n-1}}$
Δ = Précision absolue (Rayon de l'intervalle de confiance) à atteindre pour l'estimation de la moyenne inconnue sur la population totale	