Haute École Louvain en Hainaut

Traitement de données

Partie 2 : statistiques à deux caractères

Année scolaire 2019 - 20

ÉCONOMIQUE

TABLE DES MATIÈRES

1 AJUSTEMENT

Les séries statistiques à un caractère étudient les différentes valeurs prises par un caractère commun à tous les individus d'une population.

Or, pour les individus d'une population, les observations peuvent porter sur plusieurs caractères.

L'étude de séries statistiques à deux caractères conduit à chercher s'il existe un lien entre deux caractères.
Par exemples, existe-t-il un lien entre la masse et la taille d'un individu, entre la masse et la pointure ?

Il existe plusieurs types de lien entre deux caractères.

- Dépendance fonctionnelle *exemple : $E = m.C^2$*
  Graphiquement, tous les points de coordonnées (m , E) sont sur une droite.

- Indépendance *exemple : la taille T et le nombre K de km D-T d'un employé de la société S*
  Graphiquement, tous les points de coordonnées (T , D) sont éparpillés au hasard.

- Dépendance statistique *exemple : le revenu R des ménages et leurs dépenses de consommation D*
  Graphiquement, les points de coordonnées (R , D) ne sont ni situés sur une même courbe ni éparpillés au hasard.
  Les variations de valeurs entre les deux caractères sont similaires. Il existe donc une **corrélation** possible entre ces deux caractères.

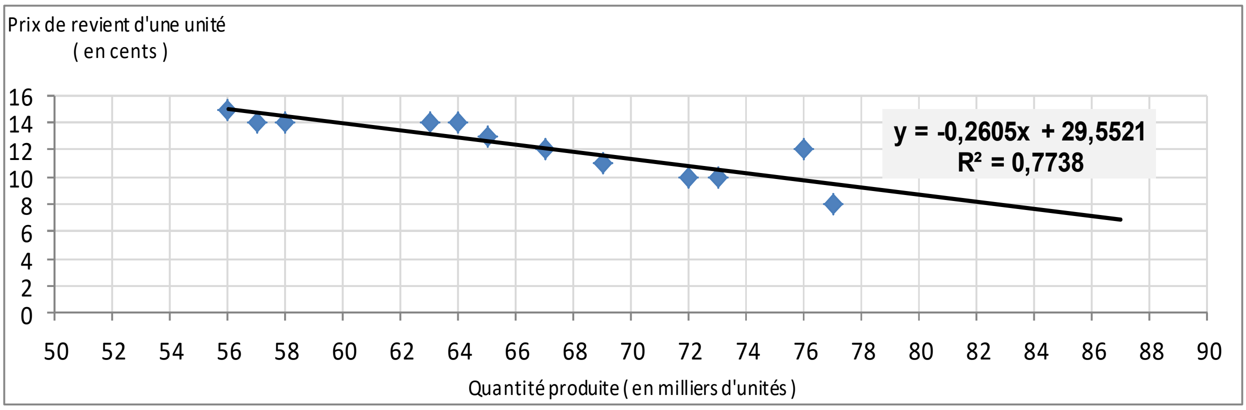
Une dépendance statistique est représentée par une courbe appelée **COURBE D'AJUSTEMENT.**

Il existe différents types de courbes d'ajustement parmi lesquelles

- la courbe linéaire $y = ax + b$
- la courbe parabolique $y = ax^2 + bx + c$
- la courbe puissance $y = A.x^a$
- la courbe exponentielle $y = A.B^{ax}$ (B valant généralement e ou 10)
- la courbe logarithmique $y = \log_a (x)$

Exemple :
Pour déterminer la fonction d'offre d'un certain bien, on dispose de 12 observations indiquant Y, le prix de revient (en cents), en fonction de la quantité produite X (en milliers d'unités)

x	69	76	72	56	57	77	58	65	67	63	73	64
y	11	12	10	15	14	8	14	13	12	14	10	14



On peut tracer, à travers le nuage de points, une droite d (courbe d'ajustement) d'équation y = ax + b reflétant aussi bien que possible le lien qui apparaît entre les deux caractères.
Cette droite doit passer "le plus près possible" de chacun des points sans nécessairement passer par chacun des points.

1.1 Méthode empirique

Notre jugement personnel peut nous permettre d'obtenir rapidement une courbe d'ajustement d'un ensemble de données. Selon la courbe à tracer, il suffit de choisir autant de points qu'il y a de paramètres à définir.

Exemple
Construire graphiquement une droite d'ajustement qui ajuste les données suivantes et trouver l'équation de cette droite.

xi	2	3	5	7	9	10
yi	2	3	7	10	14	17

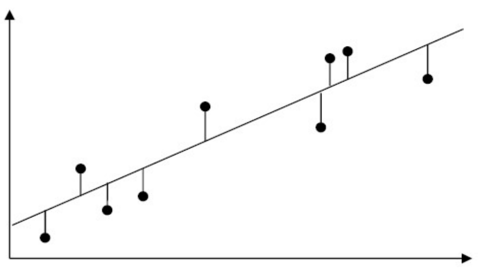
L'inconvénient d'une telle méthode est que des observateurs différents obtiendront des courbes différentes selon le choix opéré.

1.2 Méthode des moindres carrés

Ce qui manque dans la méthode ci-dessus est un critère permettant de définir la "meilleure courbe d'ajustement".
Les statisticiens ont donc proposé la **MÉTHODE DES MOINDRES CARRÉS** qui consiste à minimiser les écarts entre les valeurs des observations réelles et celles mesurées sur la courbe.

Principe :
Lorsque la courbe est tracée, à toute valeur x_i d'un point Pi correspondent deux ordonnées :
la valeur réellement observée, y_i et l'ordonnée mesurée sur la courbe, c_i

A chaque point Pi correspond donc un nombre $d_i = c_i - y_i$



Le principe de la méthode des moindres carrés est de rendre minimale la somme des di²

1.2.1 Cas d'une droite d'ajustement d'équation y = ax + b

Les paramètres a et b de la droite d'ajustement d'un nuage de N points correspondant à N observations obtenue par la méthode des moindres carrés correspondent à la solution du système de 2 équations à 2 inconnues

$$\begin{cases} a = \dfrac{\sum_{i=1}^{N} x_i.y_i - N.\overline{x}.\overline{y}}{\sum_{i=1}^{N} x_i^2 - N.\overline{x}^2} \\ b = \overline{y} - a.\overline{x} \end{cases}$$

La valeur de b indique que la droite d'ajustement passe par le point moyen PM de coordonnées $(\overline{X}, \overline{Y})$.

Autre forme de la valeur de a : utilisation des variables centrées.
Si on effectue le changement de variables suivant $X_i = x_i - \overline{X}$ et $Y_i = y_i - \overline{Y}$

alors $a = \dfrac{\sum_i X_i.Y_i}{\sum_i X_i^2}$

[Exercice 1]

1.2.2 Cas d'une courbe exponentielle en base e d'équation $y = A.e^{ax}$

Les paramètres A et a de la courbe exponentielle sont obtenus par linéarisation de l'équation.

Principe de calcul : utiliser la fonction ln ainsi qu'un changement de variable pour revenir à l'équation d'une droite.

Si $z_i = \ln(y_i)$ et $b = \ln(A)$ alors l'équation $y = A.e^{ax}$ est équivalente à l'équation $z = ax + b$

REM : Si la base de la courbe exponentielle est 10, la fonction ln est à remplacer par la fonction Log_{10}

[Exercice 2]

1.2.3 Cas d'une courbe associée à une fonction « puissance » d'équation $y = A.x^a$

Le principe est similaire à celui du 2.2.2

Si $z_i = \ln(y_i)$, $v_i = \ln(x_i)$ et $b = \ln(A)$ alors l'équation $y = A.x^a$ est équivalente à l'équation $z = av + b$

[Exercice 3]

2 CORRÉLATION

2.1 Cas général – corrélation de valeurs

La méthode des moindres carrés permet de déterminer une courbe exprimant un lien entre deux caractères.
Or, l'objectif est de pouvoir estimer la valeur d'un caractère par la valeur prise par l'autre.
Pour cela, il est donc nécessaire de mesurer la qualité de l'ajustement pour pouvoir juger de la qualité prédictive de la relation mise en évidence entre les deux caractères.

Dans ce but, on calcule le **COEFFICIENT DE CORRÉLATION**.

D'une manière générale, quel que soit le type d'ajustement, le **COEFFICIENT DE CORRÉLATION** est déterminé par

$$r = \pm\sqrt{\frac{\sum_i (y_{i\,estimé} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}}$$

$y_{iestimé}$ représentant l'ordonnée sur la courbe d'ajustement

Caractéristiques de r

- C'est un nombre sans unité, compris entre –1 et 1.
- Le signe de r indique le sens de la corrélation.
- La valeur absolue de r indique la « justesse » de la corrélation.

- **Plus | r | s'approche de 1, meilleur est l'ajustement.**
- **Il est souvent admis que si 0,7 < | r | < 1 alors l'ajustement est considéré comme valide**

Mais il faut être prudent car la valeur de r mesure le degré d'ajustement propre au type de courbe d'ajustement considérée.
Ainsi, si on suppose que la courbe est une droite et si on calcule une valeur de r proche de 0, on peut tout au plus en déduire qu'il n'y a pas de corrélation linéaire. Ce qui ne signifie pas qu'il n'y ait pas de corrélation d'un autre type (Puissance, Exponentielle, Logarithmique…)

De plus, le fait que 2 caractères soient en corrélation ne signifie pas nécessairement qu'il y ait une relation de causalité entre les deux, les variations de ces 2 caractères pouvant être la conséquence des variations d'un troisième caractère.

Exemple
Supposons que des relevés montrent une corrélation négative entre le prix du blé et le nombre de rongeurs dans un champ. Ces deux variables dépendent non pas l'une de l'autre mais bien du niveau de production de blé. Une production abondante fait baisser les prix mais fait augmenter le nombre de rongeurs dans le champ.

Pour estimer la corrélation, on définit parfois le **COEFFICIENT DE DÉTERMINATION** qui vaut r².

2.2 Cas d'un ajustement linéaire (Coefficient de Pearson)

Comme dans le cas de l'Ecart-type dont le calcul passe par celui de la variance, le calcul du coefficient de corrélation linéaire passe par celui de la **COVARIANCE.**

La **COVARIANCE** correspond à la moyenne arithmétique des produits des écarts par rapport à la moyenne arithmétique calculés pour chaque caractère.

$$\mathrm{COV}(x,y) = \frac{\sum_i (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{N} \quad \text{ou} \quad \mathrm{COV}(x,y) = \frac{\sum_i x_i \cdot y_i}{N} - \overline{X}.\overline{Y} \quad \text{et} \quad r = \frac{COV(x,y)}{\sigma_x . \sigma_y}$$

[Exercices 4]

3 RÉGRESSION

La **RÉGRESSION** (Estimation) consiste à effectuer une estimation de la valeur d'un caractère par la valeur prise par l'autre à partir d'une courbe d'ajustement des moindres carrés qui ajuste les données de l'échantillon de base.

Si on veut estimer y en fonction de x, il faut utiliser l'équation de la courbe $y = f(x)$
Si on veut estimer x en fonction de y, il faut utiliser l'équation de la courbe $x = g(y)$

Dans le cas d'un ajustement linéaire,
- si on veut estimer y en fonction de x, il faut utiliser l'équation de la droite d $y = ax + b$
- si on veut estimer x en fonction de y, il faut utiliser l'équation de la droite d' $x = a'y + b'$

Dans ce cas, le coefficient de détermination $r^2 = a.a'$

Si la corrélation est parfaite, $r^2 = 1$ donc $a.a' = 1$ donc $a'=1/a$ donc les deux droite sont identiques.

[Exercice 5]

[Exercices 6 - 7 – 8]

4 SÉRIE AVEC DONNÉES GROUPÉES

Exemple :
Dans l' agence immobilière HELMONS, une étude porte sur la répartition des appartements répertoriés selon le nombre de chambres (yi) en fonction du nombre d'enfants (xi) du ménage.
L'étude a été réalisée sur un échantillon de 100 ménages.

Le **TABLEAU DE CONTINGENCE** *donne*

Tableau de contingence (données)			
yj	1	2	3
xi			
1	5	12	3
2	1	13	17
3	0	14	20
4	0	0	15

Marche à suivre :

1) Etudier séparément la distribution des xi et celle des yj

Tableaux de distribution												
Distribution des xi							Distribution des yj					
xi	ni	ni.xi	ni.xi²	Lig i	xi.Lig i		yj	nj	nj.yj	nj.yj²	Lig j	yj.Lig j
1	20	20	20	38	38		1	6	6	6	7	7
2	31	62	124	78	156		2	39	78	156	80	160
3	34	102	306	88	264		3	55	165	495	157	471
4	15	60	240	45	180							
	100	244	690		638			100	249	657		638

Lig i = $\sum$ **nij . yj** Lig 1 = 5*1 + 12*2 + 3*3

2) Calculer les paramètres de chacune des distributions

Paramètres			
Moy xi	2,4400	*Moy yj*	2,4900
Var xi	0,9464	*Var yj*	0,3699
σ xi	0,9728	*σ yj*	0,6082

3) Calculer la corrélation

$$\mathrm{COV} = \frac{\sum_i \sum_j \mathrm{nij}.(\mathrm{xi}-\bar{\mathrm{X}}).(\mathrm{yi}-\bar{\mathrm{Y}})}{\mathrm{N}} \quad \text{ou} \quad \mathrm{COV} = \frac{\sum_i \sum_j \mathrm{nij.xi.yj}}{\mathrm{N}} - \bar{\mathrm{X}}.\bar{\mathrm{Y}} \quad \text{et} \quad \mathrm{r} = \frac{\mathrm{COV}}{\sigma\mathrm{x}\,.\sigma\mathrm{y}}$$

Corrélation	
COV	0,3044
r =	0,5145
r² =	0,2647

4) Si la corrélation est pertinente, déterminer l'équation de la droite d'ajustement Y = aX + b

$$\mathrm{a} = \frac{\sum_i \sum_j \mathrm{nij.xi.yj} - \mathrm{N}.\bar{\mathrm{X}}.\bar{\mathrm{Y}}}{\sum_i \mathrm{ni.xi}^2 - \mathrm{N}.\bar{\mathrm{X}}^2} \quad \textit{et} \quad \mathrm{b} = \bar{\mathrm{Y}} - \mathrm{a}.\bar{\mathrm{X}}$$

Dans le cas de l'exemple, la corrélation n'est pas suffisante donc la droite d'ajustement est inutile.

[Exercices 10 – 11]

5 SÉRIE CHRONOLOGIQUE

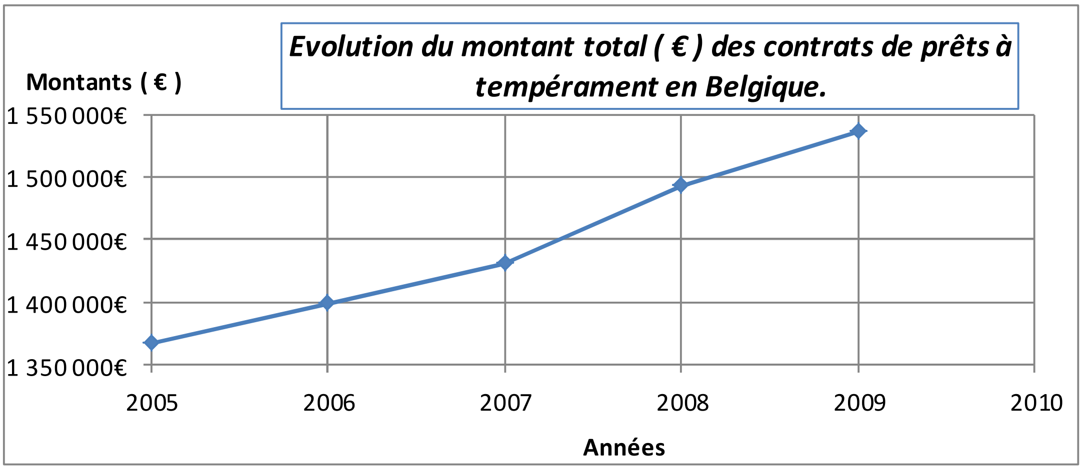
5.1 Définition

Une **SÉRIE CHRONOLOGIQUE** ou série temporelle est une suite d'observations d'un caractère ordonnées dans le temps, habituellement à intervalles égaux.
La représentation graphique de la série sous forme d'une ligne polygonale permet d'en obtenir une bonne image.

Exemple : *Evolution du montant total (€) des ontrats de prêts à tempérament en Belgique.*

Année	2005	2006	2007	2008	2009
Montant	1 367 647€	1 398 962€	1 431 078€	1 493 628€	1 536 174€

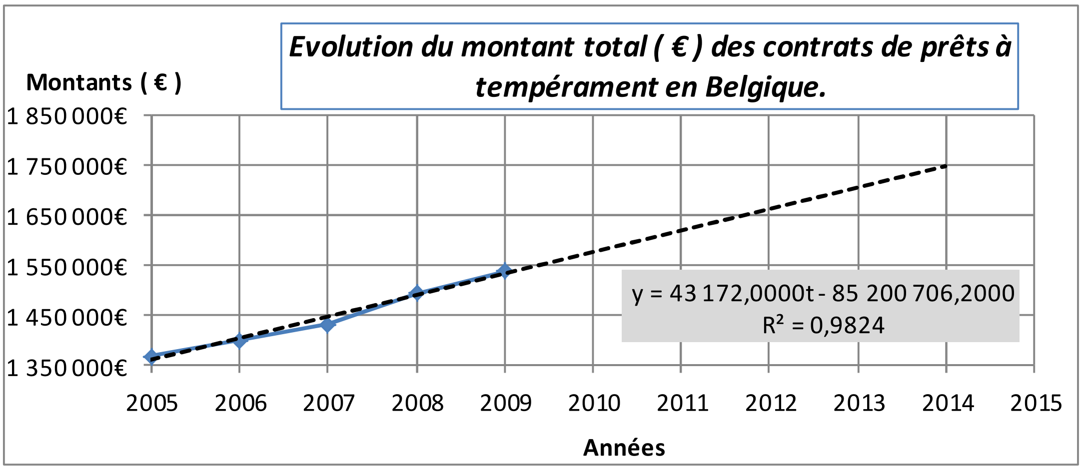
Source : Banque Nationale – Centrale des crédits aux particuliers – Statistiques 2009



5.2 Tendance d'une série chronologique (Impact de longue durée)

La méthode des moindres carrés reste valable en prenant comme variable xi la variable ti qui détermine le temps.

Droite d'ajustement y = at + b		Corrélation		Estimation		
a =	43172	r =	0,9911	en	2013	2014
b =	-85200706,2	r² =	0,9824	y	1 704 529,80€	1 747 701,80€



[Exercices 14 – 15 – 16]

6 EXERCICES

Exercice 1

Enoncé :

Dans un magasin de meubles en kit, le service vente dispose de données relatives au volume de rangement (x en m³) et au prix de vente (y en €). Le relevé de ces données donne le tableau suivant.

Volume (m³)	0,15	0,20	0,25	0,30	0,40	0,50
Prix (€)	110,75	140,00	153,25	168,75	205,00	266,75

Marche à suivre :

- Tracer le nuage de points de y en fonction de x et compléter le graphique par une droite de tendance. *
- Déterminer les paramètres de la droite d'ajustement de y en fonction de x en utilisant les fonctions. *
- * Idem pour x en fonction de y.

Exercice 2

Enoncé :

Dans un laboratoire de résistance des matériaux, on mesure, toutes les 30s, la température d'une plaque de métal qui se refroidit à partir de 100 °C. Le relevé de ces données donne le tableau suivant.

x	Instant (en min)	0	0,5	1	1,5	2	2,5	3	3,5	4
y	Température (en °C)	105	95,1	85,6	76,6	69,3	62,8	56,3	51,1	48,2

Marche à suivre :

- Tracer le nuage de points de y en fonction de x et compléter le graphique par une courbe de tendance de type exponentielle en base e.
- Déterminer les paramètres de cette courbe en utilisant le principe de linéarisation.

Exercice 3

Enoncé :

Dans un laboratoire de physique, on mesure les variations de l'intensité d'un courant en fonction de l'augmentation de la tension aux bornes d'une lampe. Le relevé de ces données donne le tableau suivant.

U → tension aux bornes de la lampe (en volts)

I → intensité du courant (en milliampères)

U	0,1	0,2	0,5	0,8	1,5	2	2,5	3	4	5
I	36	51	81	100	132	155	175	200	222	250

Marche à suivre :

- Tracer le nuage de points de I en fonction de U et compléter le graphique par une courbe de tendance de type puissance.
- Déterminer les paramètres de cette courbe en utilisant le principe de linéarisation.

Exercice 4

Marche à suivre :

En partant du relevé de données suivant

Tableau des relevés				
x	20	30	36	40
y	10	20	30	44

- Tracer le nuage de points de y en fonction de x et compléter le graphique par 3 courbes de tendance de types droite, puissance et fonction exponentielle en base e.
- Déterminer les paramètres de ces trois courbes d'ajustement.
- Déterminer le meilleur ajustement.

Exercice 5

- Dans l'exercice 2,
    - estimer le prix de vente d'un meuble dont le volume de rangement est de 0,80 m³,
    - estimer le volume de rangement d'un meuble dont le prix de vente serait de 350€,
    - Vérifier graphiquement ces estimations

- Dans l'exercice 3, estimer la température de la plaque après 6 minutes.

- Dans l'exercice 4, estimer l'intensité du courant lorsque la tension aux bornes de la lampe est de 7,5 volts.

- Dans l'exercice 5, donner la meilleure estimation de y pour une valeur de x de 50.

Exercice 6

Enoncé :

La société HELTECh a mis au point un nouveau produit. Selon une étude préalable, le responsable du service commercial essaie d'étudier le lien entre le prix de vente proposé (x - en €) et le nombre d'acheteurs potentiels à ce prix (y) en interrogeant un échantillon de 500 clients de la société.
Le relevé des données donne le tableau suivant.

Tableau des relevés										
Prix de vente proposé (€)	xi	20,00	19,50	19,00	18,50	18,00	17,50	17,00	16,50	16,00
Nb.acheteurs potentiels	yi	60,00	80,00	130,00	200,00	240,00	350,00	390,00	420,00	440,00

Marche à suivre :

- Tracer le nuage de points de Y en fonction de X et compléter le graphique par une droite de tendance.
- Déterminer les paramètres de la droite d'ajustement de Y en fonction de X.
- Idem pour X en fonction de Y.
- Déterminer les coefficients de corrélation et de détermination.
- Estimer le nombre d'acheteurs potentiels si le prix proposé était de 15€.
- Estimer le prix qu'il faudrait proposer pour avoir un quota de 500 acheteurs potentiels.
- Vérifier graphiquement toutes les réponses.

Exercice 7

Enoncé :
Dans un bureau d'études, afin de mettre au point de nouveaux pneumatiques pour un type de moto M, on étudie le lien possible entre la vitesse à l'instant de début de freinage (X) et la distance parcourue entre cet instant et l'arrêt complet de la moto (Y).Le relevé des essais donne le tableau suivant.

Vitesse (en km/h)	20	30	40	50	60	70
Distance parcourue (m)	59	96	152	206	308	401

Marche à suivre :
- Tracer le nuage de points de Y en fonction de X
- Insérer une droite d'ajustement de Y en fonction de X et afficher ses paramètres.
- Cet ajustement est-il pertinent ?
- Insérer une courbe d'ajustement de type « fonction du second degré » et afficher ses paramètres.
- Ce nouvel ajustement est-il plus pertinent ?
- Estimer la distance de freinage pour une vitesse est de 100 km/h par les deux courbes d'ajustement.
- Vérifier en prolongeant ces courbes.

Exercice 8

Enoncé :
Dans un service de maintenance d'une usine, on étudie le lien entre le coût annuel d'entretien et de réparation (Y) d'un type de machine Mch et l'âge de ces machines (X)
Le relevé des données donne le tableau suivant

Age de la machine (en années)	1	2	3	4	5
Coût d'entretien et de réparation (en milliers d'€)	13,3	14,2	16,1	18,9	23,6

Marche à suivre :
- Tracer le nuage de points de Y en fonction de X et compléter le graphique par 4 courbes de tendance de types droite, parabole,puissance et fonction exponentielle en base e.
- Déterminer, pour chacune des courbes, la corrélation.
- Pour chacun des ajustements, estimer le coût d'entretien d'une machine de 8 ans.

Exercice 9

Enoncé :
Dans un échantillon de 50 couples mariés dont les maris sont issus de familles « non nombreuses », on étudie le lien entre le nombre de frères/sœurs du mari (xi) et le nombre de frères/sœurs de la femme (yi).

Le tableau de contingence donne

xi \ yi	0	1	2	3	4
0	5	4	0	0	0
1	2	20	2	3	2
2	0	0	10	1	1

Marche à suivre :
- Estimer si la corrélation linéaire entre les 2 variables est significative.
- Si oui, estimer le nombre de frères/sœurs de la femme lorsque le nombre de frères/sœurs du mari est de 3.

Exercice 10

Enoncé :

Dans le cadre d'une étude de diététique, un sondage porte sur le lien entre la masse corporelle (xi en kg) et la taille (yj en cm) de 65 garçons de 12 ans ne pratiquant aucun sport hebdomadaire.

Le tableau de contingence donne

	yj	150		155		160		165		170
xi										
40										
			12		4		3		0	
45										
			1		15		12		5	
50										
			1		2		5		5	
55										

Marche à suivre :

- Estimer si la corrélation linéaire entre les 2 variables est significative.
- Si oui, estimer la taille d'un garçon lorsque sa masse corporelle est comprise entre 55 kg et 60 kg..

Exercice 11

Enoncé :

Le tableau suivant donne les chiffres relatifs à l'évolution de l'espérance de vie en Wallonie.

Evolution de l'espérance de vie à la naissance en Wallonie *					
	2004	2006	2008	2010	2012
Hommes	74	74,3	74,4	75,3	75,7
Femmes	80,9	81,2	81,3	81,5	81,8
*	Tableau tiré du rapport *LA FÉDÉRATION WALLONIE-BRUXELLES EN CHIFFRES* édité par la FW-B en 2015				

Marche à suivre :

- Tracer les deux nuages de points représentant les deux séries de données.
- Insérer, dans chacun de ces nuages, une droite d'ajustement et afficher ses paramètres.
- Estimer, si cela a un sens, l'espérance de vie des hommes wallons en 2020 (idem femmes).

Exercice 12

Enoncé :

Le tableau suivant donne les chiffres relatifs à l'évolution du nombre de postes de travail dans le secteur TIC, en Belgique. (Les données sont exprimées en milliers – Chiffres ONSS)

	2010	2011	2012	2013	2014
Nbr.spécialistes des TIC (en milliers)	134,4	159,4	197,8	195,3	199,7
Source : Enquête sur les forces de travail (EFT) relatives à l'emploi, Eurostat http://statbel.fgov.be/fr/binaries/Barometre_de_la_societe_de_l_information_2017_tcm326-278973.pdf					

Marche à suivre :

- Calculer la valeur du coefficient de corrélation pour chacun des ajustements de type linéaire, exponentiel en base e et puissance.
- Vérifier graphiquement
- Estimer au mieux le nombre de personnes employées, en Belgique, comme spécialistes des TIC à l'horizon 2020.

Exercice 13

Enoncé :

Le tableau suivant donne les chiffres relatifs à l'évolution de la récolte des déchets municipaux en Belgique.
Les déchets municipaux sont les déchets récoltés par les services communaux de collecte, les parcs à conteneurs, les balayeurs,… à l'exclusion des matériaux de construction.
Les données sont exprimées en milliers de tonnes.

	2008	2009	2010	2011	2012	2013	2014
Collecte	5 134	5 037	4 973	5 035	4 970	4 892	4 886
Données exprimées en milliers de tonnes							
Source : Direction générale Statistique - Statistics Belgium sur base de données administratives http://statbel.fgov.be/fr/statistiques/chiffres/environnement/dechets/municipaux/							

Marche à suivre :

- Tracer le nuage de points représentant la série de données.
- Graphiquement, déterminer le type de courbe donnant le meilleur ajustement.
- Estimer le tonnage de déchets municipaux collectés pour cette année et les 4 suivantes.

Exercice 14

Au 01/09/2010, un troupeau de 200 cerfs a été introduit dans la région R.
Depuis, le responsable chargé de la surveillance du troupeau effectue, annuellement, à la même date, le relevé du nombre de têtes.

	2010	2011	2012	2013	2014	2015	2016
t (années)	0	1	2	3	4	5	6
Nbr.têtes	200	267	386	602	832	902	915

D'après ses expériences antérieures, il suppose que l'ajustement de ces premières données suit une fonction polynomiale de degré 4.

Estimer le nombre de têtes pour les 4 prochaines années.

BIBLIOGRAPHIE

1 Ouvrages imprimés

- DROESBEKE JJ, *Eléments de statistique*, Bruxelles, Editions de l'Université de Bruxelles, 2001.

- OUELLET Gilles, *Statistique et probabilités*, Sainte-Foy (Québec) , Les éditions Le griffon d'argile, 1998.

2 Notes de cours

- JOMAUX F, *Statistique descriptive*, Fucam, Septembre 2005

- DENORME C, *Statistique 2è année Chimie clinique*, Haute école Charleroi Europe, Septembre 2009

3 Sites Web

http://www.iweps.be/
http://walstat.iweps.be/carto/cartographie.php
http://www.cons-dev.org/elearning/ando/06/61/61.html