S4 2019-2020

Nathan TONNELLE

Statistiques

MINFO0401

Table des matières

[Statistique descriptive 3](#_Toc33653418)

[Vocabulaire 3](#_Toc33653419)

[Population 3](#_Toc33653420)

[Echantillon 3](#_Toc33653421)

[Caractère et modalité 3](#_Toc33653422)

[Représentation graphique 4](#_Toc33653423)

[Variables qualitatives 4](#_Toc33653424)

[Variable quantitative 4](#_Toc33653425)

[La fonction cumulative 5](#_Toc33653426)

[Variables continues 5](#_Toc33653427)

[Définition (représentation graphique différentielle) 5](#_Toc33653428)

[Exemple 5](#_Toc33653429)

[Courbe cumulative 6](#_Toc33653430)

[Description numérique d'une variable 6](#_Toc33653431)

[Paramètres de position 6](#_Toc33653432)

[Paramètre de dispersion 7](#_Toc33653433)

[Coefficient de variation 8](#_Toc33653434)

[Les moments 8](#_Toc33653435)

[Caractéristiques de forme 9](#_Toc33653436)

[Distributions à deux dimensions 10](#_Toc33653437)

[Notations 10](#_Toc33653438)

[Distribution des fréquences du couple (X,Y) 10](#_Toc33653439)

[Notation 10](#_Toc33653440)

[Remarque 10](#_Toc33653441)

[Distribution marginale 11](#_Toc33653442)

[Moyennes et variances marginales 11](#_Toc33653443)

[Distributions conditionnelles 11](#_Toc33653444)

[Définition 11](#_Toc33653445)

[Remarque 11](#_Toc33653446)

[Indépendance des variables X et Y 12](#_Toc33653447)

[Conséquence 12](#_Toc33653448)

[Moyennes et variances conditionnelles 12](#_Toc33653449)

[Résultat 12](#_Toc33653450)

[Notion de corrélation 13](#_Toc33653451)

[Coefficient de corrélation linéaire 13](#_Toc33653452)

[Rapports de corrélation 13](#_Toc33653453)

[Définition 13](#_Toc33653454)

[Propriété 14](#_Toc33653455)

[Courbes de régression 14](#_Toc33653456)

[Définition 14](#_Toc33653457)

[Commentaire 14](#_Toc33653458)

[Liaison entre deux variables 14](#_Toc33653459)

# Statistique descriptive

## Vocabulaire

### Population

Une population est l'ensemble des individus ou objets sur lesquels portent une étude statistique, on le note P

#### Exemple

1. Logement d'une ville
2. Personnel d'une entreprise
3. Animaux d'un parc naturel

### Echantillon

Un échantillon est une partie de la population e étudier sur laquelle porte l'étude statistique.

Une étude statistique portant sur un échantillon est appelée sondage.

On appelle étude associative ou recensement si elle porte sur l'ensemble de la population.

### Caractère et modalité

Une étude statistique porte sur 1 ou plusieurs caractères communs à tous les individus de la population à étudier. Un caractère est aussi appelé variable.

#### Exemple

1. Surface de logement
2. Age, ancienneté, revenu
3. Régime alimentaire, espèces

#### Modalité

Les modalités d'un caractère sont les différentes valeurs que peut prendre se caractère sur les individu de la population étudiée

##### Exemple

1. R+
2. [16,10)
3. Herbivore, carnivore, omnivore

On distingue 2 types de caractères :

* Caractère qualitatif :

Les modalités sont des attributs qualitatif (ex: régime alimentaire)

* Caractère quantitatif

Les modalités sont des quantités numériques (ex : age)

Les caractères quantitatifs sont de 2 types

* Variable discrète :

Les modalités de la variable appartiennent à un ensemble discret tel N, Z, N²

* Variable continue :

Les modalités de la variable prennent des valeurs dans un ensemble continu tel que R

#### Remarque :

Pour étudier une variable continue on constitue des classes de valeurs possibles, ces classes sont des intervalles d'amplitude égale ou inégale est constitué alors de nouvelle modalité ou des caractères.

#### Attention :

Le découpage en classes peut influer sur les résultats et les interprétations que l'on peut faire.

S'il est trop important, il risque de faire apparaitre des irrégularités artificielles car les effectifs des classes seront trop faibles.

S'il est trop grossier, il conduira à une perte d'information.

#### Effectif et fréquence

L'effectif d'une modalité ou d'une classe de modalité est le nombre d'individu de la population correspondant à cette modalité ou à cette classe de modalités. On ne note ni pour la ième modalité.

La fréquence de la ième modalité (ou classe) est donnée par le rapport de son effectif sur l'effectif total de la population noté n.

On a alors fi=ni/n

#### Remarque

Si la variable X possède K modalités x1,..,xk alors

i=n i

## Représentation graphique

Il existe différentes façons de représenter graphiquement des variables

### Variables qualitatives

#### Diagramme circulaire

C'est un disque dans lequel chaque modalité est représenté par un secteur angulaire proportionnel à sa fréquence.

Pour 1<=i<=f **α** i=360xfi

#### Diagramme en tuyaux d'orgue

C'est un diagramme formé de rectangles tous de même largeur et dont les hauteurs sont proportionnelles aux fréquences des modalités

### Variable quantitative

Il y a 2 sortes de représentation graphique des variables quantitatives :

* Diagramme différentiel
* Diagramme intégrale

#### Variables discrètes

Le diagramme différentiel utilisé dans le cas d'une variable discrète est le diagramme à bâton.

Où chaque bâton est de longueur ou de hauteur proportionnel à la fréquence de la modalité correspondante

##### Exemple

Nombre d'enfants dans un échantillon de 50 familles.

Nombre d'enfant ni fi

0 9 0.18

1 7 0.14

2 12 0.24

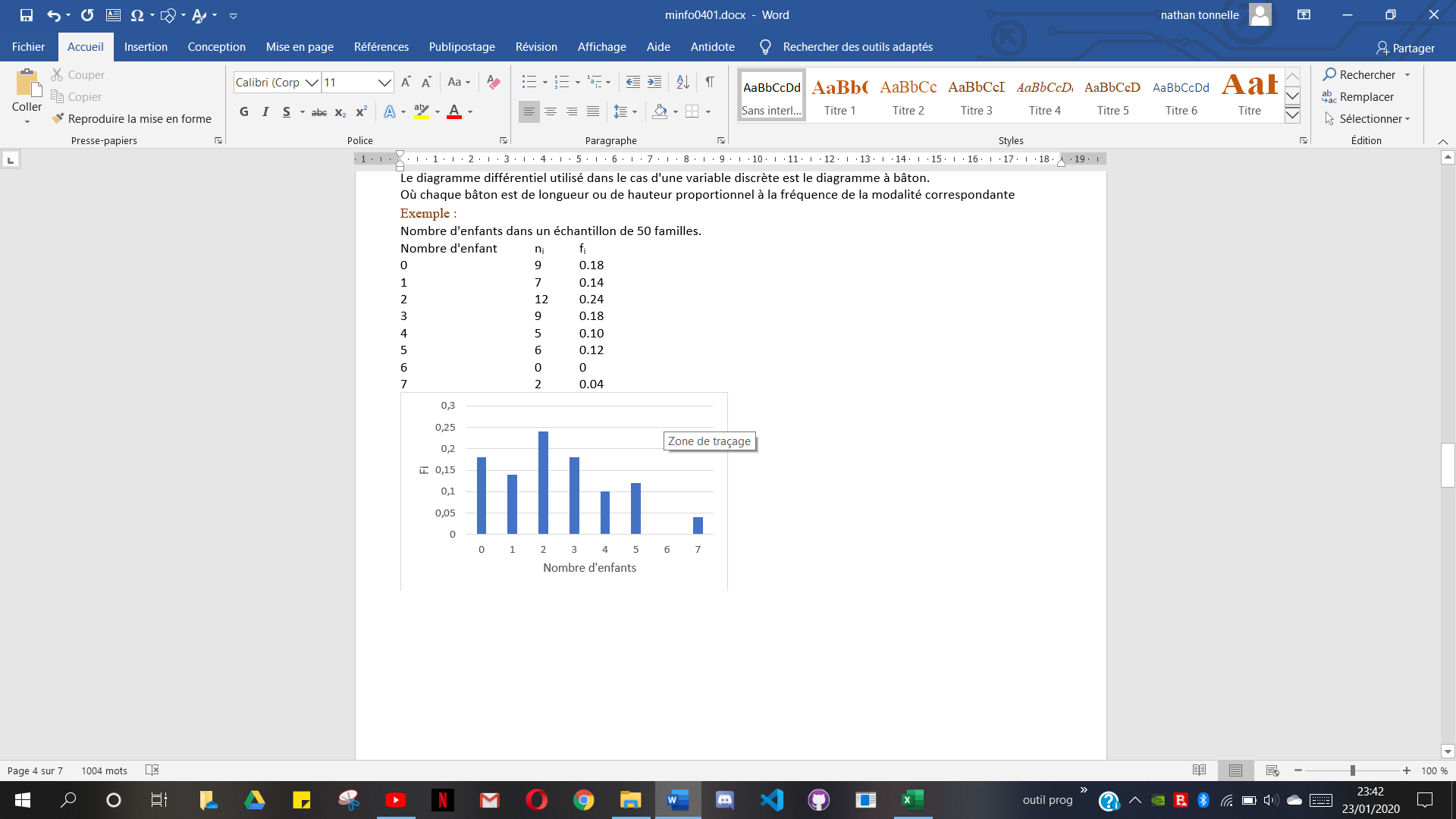
3 9 0.18

4 5 0.10

5 6 0.12

6 0 0

7 2 0.04



### La fonction cumulative

La fonction cumulative d'une variable X prise en un point x , noté F(x) est définie comme la proportion de la population pour laquelle la variable X prend des valeurs >= à x

Si les modalités de X sont x1,x2,…xk alors pour xi<=x<xi+1, F(x)=

Exemple :

Si x<0 F(x)=0

0<=x<1 F(x)=0.18

1<=x<2 F(x)=0.32

2<=x<3 F(x)=0.56

3<=x<4 F(x)=0.74

4<=x<5 F(x)=0.84

5<=x<7 F(x)=0.86

x>=7 F(x)=1

## Variables continues

Si ei et ei+1 sont les extrémités de la classe n°i, noté [ei,ei+1], on notera ci soit milieu et ai son amplitude, 1<=i<=k

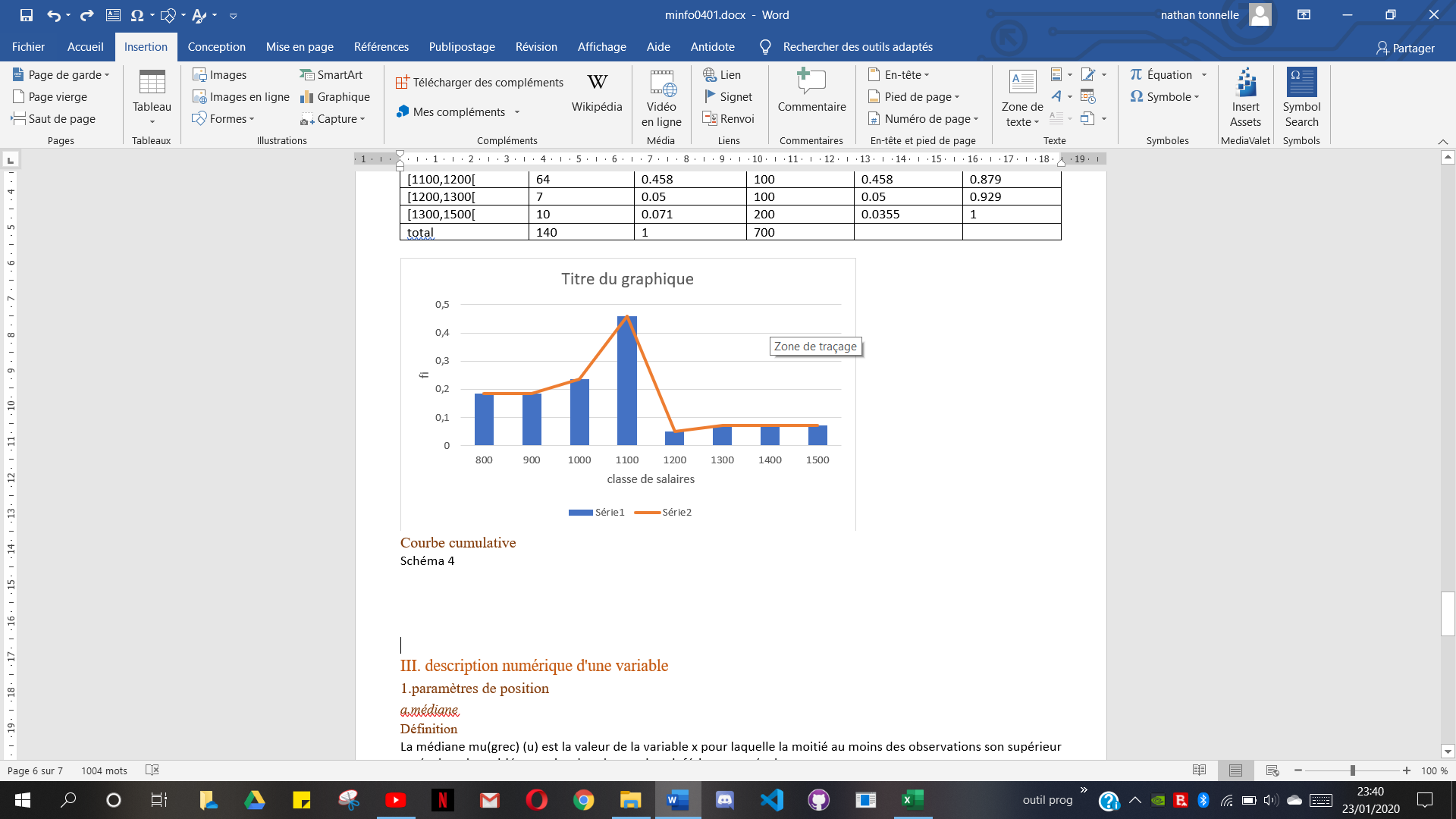
### Définition (représentation graphique différentielle)

Un histogramme est une représentation graphique où chaque classe est représenté par un rectangle de base proportionnelle à son amplitude et de surface proportionnelle à sa fréquence. Ainsi la hauteur de la classe n°i est hi=fi/ai. Un polygone statistique est un polygone reliant le milieu des bases supérieurs des rectangles de l'histogramme.

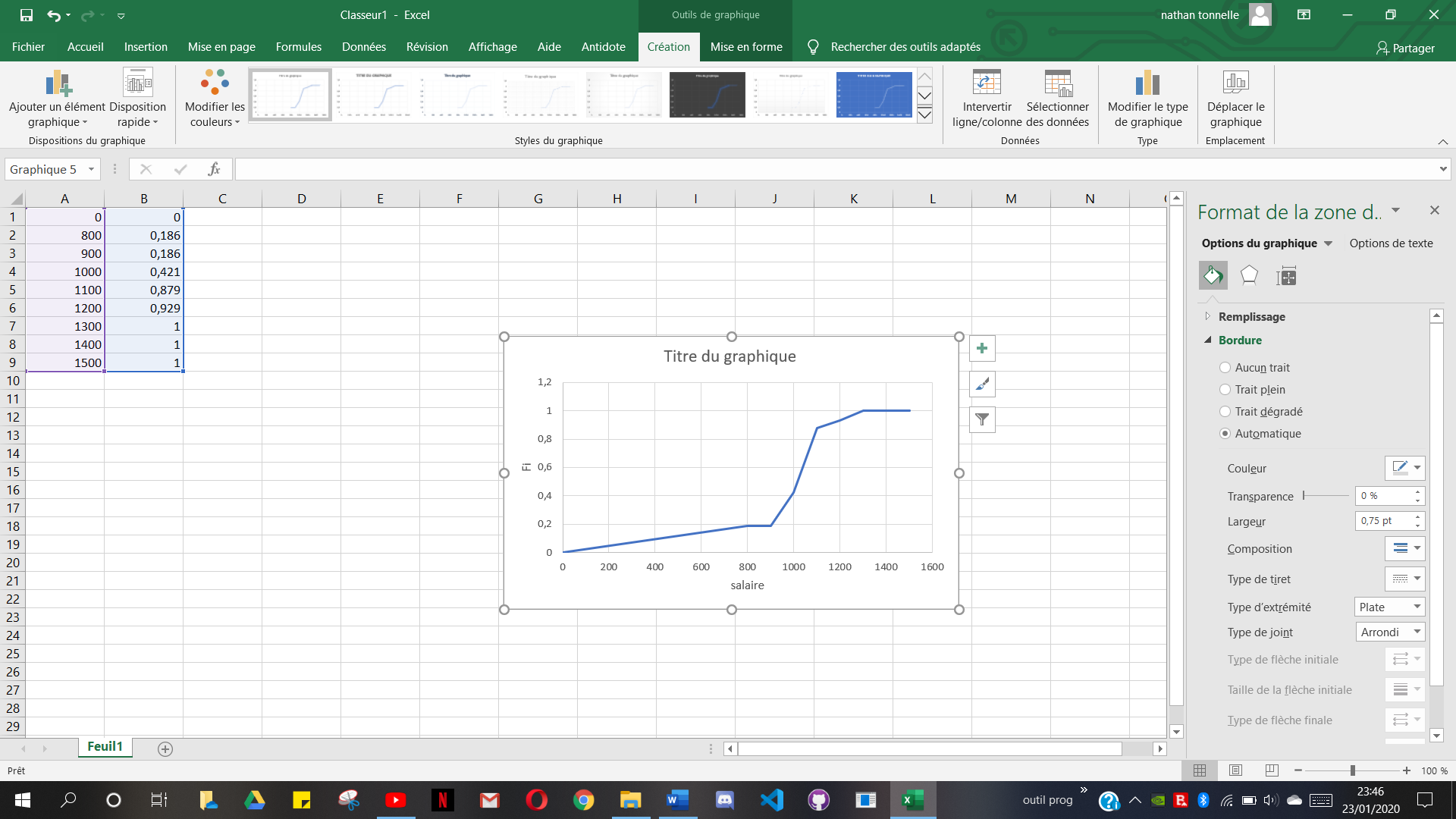
### Exemple

Salaire mensuel net des ouvriers d'un établissement industriel.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classes de salaire | ni | fi | ai | hi=fi/ai (echelle x100) | Fi |
| [800,1000[ | 26 | 0.186 | 200 | 0.096 | 0.186 |
| [1000,1100[ | 33 | 0.235 | 100 | 0.235 | 0.421 |
| [1100,1200[ | 64 | 0.458 | 100 | 0.458 | 0.879 |
| [1200,1300[ | 7 | 0.05 | 100 | 0.05 | 0.929 |
| [1300,1500[ | 10 | 0.071 | 200 | 0.0355 | 1 |
| total | 140 | 1 | 700 |  |  |



### Courbe cumulative



## Description numérique d'une variable

### Paramètres de position

#### Médiane

##### Définition

La médiane mu(grec) (u) est la valeur de la variable x pour laquelle la moitié au moins des observations son supérieur ou égale et la moitié au moins des observations inférieures ou égales

##### Remarque

Dans le cas continue on détermine d'abords la classe médiane avant de calculer le point le médian par la méthode d'interpellation linéaire.

##### Exemple

1. (nombre d'enfants par famille) u=2 car (F2=0.32 et F3=0.56)
2. (salaires des ouvriers)

Classe médiane = [1100,1200[

Calcul de la médiane par la méthode d'interpellation linéaire

#### Le mode

##### Définition

Le mode est la valeur de la variable x ayant la plus grande fréquence

##### Remarque

Certaines séries statistiques peuvent avoir plusieurs modes. Dans le cas continu on parle de classe modale, on veillera cependant a tenir compte de l'amplitude des classe. La classe modale correspond à la classe ayant la plus grande hauteur hi.

##### Exemple

1. (salaire des ouvriers)

Classe modale =[1100,1200[

Equation (M M2)

Equation(M3 M4)

#### la moyenne

Soit x une variable prenant les valeurs x1,…,xk avec les effectifs n1,…,nk respectivement (avec les fréquences f1,…,fk). Alors la moyenne de la variable x est donnée par

Où

Dans le cas continu, la moyenne d'une variable x est définie par

Où les Ci sont les milieux des classes [ei,ei+1[

##### Propriétés

1. Linéarité

Si on considère la transformation Y=aX+b

Alors la moyenne de Y est = a +b

1. Si on définit la fonction

Alors L(c) prend son minimum pour c=

1. Si P=P1UP2

Où la moyenne de X sur P1 est , et l'effectif est n1; la moyenne de X sur P2 est 2 et l'effectif de P2 est n2, alors la moyenne de X sur P est :

### Paramètre de dispersion

#### Etendu

L'étendu est la différence entre la plus grande et la plus petite valeur de la série statistique.

#### L'écart moyen absolu

On considère une série de n observations dont les modalités sont x1,…,xk.

On note sa moyenne empirique.

Alors l'écart-type moyen absolu est défini par

#### Variance (écart quadratique moyen)

Si x1,…,xn sont les modalités d'une variance X observée n fois, alors la variance de X est définie par

##### Propriété

1. Si , alors

##### Preuve

1. Si P1 et P2 sont 2 sous populations d'une population P tel que les moyennes, les variences et les effectifs de X sont :

Pour P1 :

Pour P2 :

Alors la variance de X sur l'ensemble de P est

Où est la moyenne de X sur P

#### Ecart type

L'écart type est défini par :

##### Intérêt :

X est en Km => en Km²

L'écart type exprime la dispersion dans la même unité de mesure que la variable X

### Coefficient de variation

Le coefficient de variation est défini par

#### Intérêt :

Ce coefficient est indépendant de l'unité de mesure. Il permet alors de comparer les dispersions de sens statistiques exprimés dans des unités de mesure différents.

### Les moments

Définition

On appelle moment d'ordre t (t appartient à N) par rapport à une constante a d'une variable statistique x

Les moments non centrés correspondent à

Les moments centrés correspondent à a=

#### Remarque

### Caractéristiques de forme

#### Coefficient d'asymétrie

Le coefficient d'asymétrie de Fisher est défini par

Si A=>B alors =>

#### Remarque :

Si la distribution est symétrique, alors

Si alors la distribution n'est as symétrique



Distribution non symétrique, oblique à gauche et étroite à droite

#### Coefficient d'aplatissement (Kertosis)

Il compare l'aplatissement de courbe statistique à la courbe de la loi N(0,1).



Si la courbe est identique à celle de la loi N(0,1) alors

Si , la courbe ne présente pas le même aplatissement que la courbe de la loi N(0,1).

#### Box-plot (boite à moustache)

##### Définition :

Premier quartile q1 est la valeur de la variable x pour laquelle au moins ¼ des observations lui sont inférieur ou égales et au moins ¾ des observations supérieures ou égales.

Le troisième quartile q3 est la valeur de la variable x pour laquelle au moins ¾ des observations lui sont inférieur ou égales et au moins ¼ des observations supérieures ou égales.

##### Remarque :

Dans le cas d'une variable continue, on détermine d'abord les classes contenant le premier quartile q1 et le 3eme quartile q3 avant de procéder par la méthode de l'interpellation linéaire au calcul de q1 et q3

Le box-plot :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 25% | | 25% | 25% | 25% | |  |
|  |  |  |  |  |  |  |  |
| Min | | Q1 | | Q3 | | Max | |

# Distributions à deux dimensions

On considère un couple de variables (X,Y) observé sur une population P de taille n. on suppose que la variable X possède les modalités x1,…,xk et que les modalités de la variable y sont y1,…,yl. Ces modalités peuvent être des valeurs discrètes ou des classes de modalités (cas où les variables sont continues). L'observation du vecteur (X,Y) donne lieu à la table de contingence qui se présente sous la forme :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X\Y |  |  | … |  | … |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Où nij est l'effectif connu portant un nombre de fois où la modalité xi de x et la modalité de yj de Y sont observé, simultanément sur les individus de la population P.

## Notations

,

La table de contingence donne la distribution des effectifs du couple (X,Y).

## Distribution des fréquences du couple (X,Y)

La fréquence de la modalité (xi,yj) de (X,Y) est définie par

, 1<= i<=k et 1<=j<=l

## Notation

Pour 1<=i<=k et 1<=j<=l , on note

## Remarque

# Distribution marginale

La distribution marginale des effectifs de X est de définir par (n1,…,nk). la distribution marginale des fréquences de X est (f1.,…,fk.).

Les distributions marginales des effectifs et des fréquences de la variable Y sont respectivement (n.1,…,n.l) et (f.1,…,f.l).

## Moyennes et variances marginales

La moyenne et la variance marginales de X sont définies par

La moyenne de la variance marginale de Y sont :

## Distributions conditionnelles

### Définition

La fréquence conditionnelle de X=xi sachant Y=yj est définie, pour , par

La fréquence conditionnelle de Y=yj sachant X=xi est définie pour tous par

La distribution conditionnelle de X sachant Y=yj est

La distribution conditionnelle de Y sachant X=xi est

### Remarque

Par conséquent

Conséquence

### Indépendance des variables X et Y

Les variables X et y sont indépendantes lorsqu'on a

### Conséquence

Si X et Y sont indépendants alors

Cela peut s'exprimer aussi sous la forme

## Moyennes et variances conditionnelles

La moyenne conditionnelle de X sachant Y=yj est définie par

La moyenne conditionnelle de Y sachant X=xi est

### Résultat

Les moyens marginales et les moyennes conditionnelles sont liées par les relations suivantes :

Les variances conditionnelles de X sachant Y=yj et sachant X=xi sont définies respectivement par

Et

Les variances marginales de X et Y peuvent être décomposées dans les termes suivants :

= moyennes des variances conditionnelles et variances des moyennes conditionnelles

## Notion de corrélation

### Coefficient de corrélation linéaire

#### Définition

Le coefficient de corrélation linéaire entre deux variables X et Y est défini par

Où

#### Propriétés

#### Remarque

Si X et Y ne sont pas corrélées, cela n'implique pas forcément que X et Y sont indépendants.

Il y a identité entre indépendance et non corrélation uniquement dans le cas de variable gaussiennes.

#### Propriétés (suite)

* Si X et Y présentaient une liaison fonctionnelle, i.e., il existe une fonction f tant que

### Rapports de corrélation

On considère deux variables X et Y prennent les valeurs x1,…,xk et y1,…,yl respectivement.

On note Xbarre et Ybarre les moyennes empiriques de X et Y respectivement.

On note var(X) et var(Y) les variances marginales de X et Y.

Si est la moyenne de X sachant , la moyenne de Y sachant X=xi ,

On note

Et

### Définition

Le rapport de corrélation de X en Y est défini par

Le rapport de corrélation de Y en X est défini par

### Propriété

1. Si , cela veut dire que pourtant

Le conditionnement de X par Y ne donne aucune liaison significative

1. Si , cela veut dire que

X est lié de façon complète à Y

1. Plus est proche de 1 plus la liaison de X à Y est importante

## Courbes de régression

Les courbes de régression ont pour objet de donner une représentation graphique sur le plan de la distribution conjointe du vecteur (X,Y). Il y a deux courbes de régression. La courbe de régression de X en Y et la courbe régression de Y en X.

### Définition

Si la variable X prend ses valeurs dans l'ensemble {x1,…,xk } et Y dans l'ensemble {y1,…, yl}. On note la moyenne conditionnelle de X sachant Y= et la moyenne conditionnelle de Y sachant X=xi . Alors la courbe de régression de X en Y notée est la courbe passant par les points de coordonnées . La courbe passait par les points de coordonnées.

### Commentaire

Lorsque X=xi, la valeur qui synthétise le plus la variable Y est la moyenne conditionnelle de Y sachant X = xi



Pour ajuster une droite à une courbe on utilise le critère des moindres carrées donné dans l'exemple par

Il faudra alors minimiser CMC(C,D) par rapport à a et b pour trouver la meilleure droite d'ajustement.

### Liaison entre deux variables

#### Liaison nulle ente variables X et Y

Cela signifie qu'il n'y a pas d'influence d'une variable sur l'autre

Les variables X et Y sont indépendantes, signifie que la variation de l'une des variables n'entraine pas d'effets sur l'autre variable.

Autrement dit,

Ainsi on a

Les courbes de régression ont pour équations alors :



#### Liaison fonctionnelle

La liaison fonctionnelle est donnée par :

#### Remarque

Lorsque X et Y présentent une liaison fonctionnelle, les courbes de régression sont confondues.

#### Liaison relative

Dans ce cas le nuage de points du vecteur (X,Y) est résumé par les courbes qui se coupent au point centre de gravité

Ces courbes donnent des informations sur la nature de la liaison entre les variables X et Y.

* On dira que la corrélation est positive si les deux variables varient dans le même sens.
* On dira que la corrélation est négative si les varient dans les sens opposés.
* On dira que la corrélation entre X et Y est linéaire si les courbes de régression sont des droites parallèles aux axes.

La corrélation est d'autant plus grande que l'angle formé par les courbes est petit

#### Remarque

A la différence avec l'indépendance la corrélation n'est pas propriété réciproque.

On peut avoir



X est corrélé avec Y et Y est non corrélé avec X



Y dépend de X et X ne dépend pas de Y.