

STATISTIQUE DESCRIPTIVE

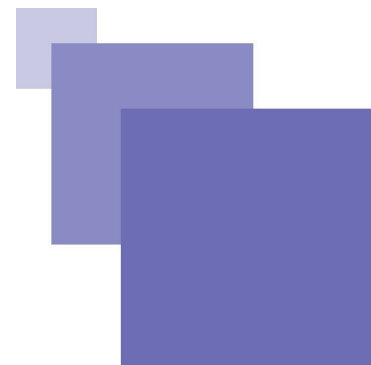
1.0

PROBL104

ATIAMPO KODJO ARMAND © UVCi 2017

Novembre 2017

Table des matières



Objectifs.....	5
Introduction.....	7
I - Représentation des données statistiques.....	9
A. Série statistique.....	9
B. Exercice : Série statistique discrète.....	16
C. Exercice.....	16
D. Exercice.....	16
E. Exercice.....	17
F. Exercice.....	17
G. Exercice : Série statistique continue.....	17
H. Exercice : Effectif corrigés dans une série statistique.....	18
I. Exercice : Fonctions de repartition.....	19
J. Exercice.....	20
II - Paramètres statistiques.....	23
A. Moyenne et variance.....	23
B. Exercice.....	26
C. Exercice : Exercice de niveau plus difficile.....	27
III - Corrélation.....	29
A. Ajustement linéaire.....	30
IV - Exercice.....	33
V - Exercice.....	35
Conclusion.....	37
Bibliographie.....	39

Objectifs

À la fin de cette leçon, vous serez capable :

- **Organiser les données sous formes de classes représentatives**
- **Estimer les paramètres tels que la moyenne et la variance.**
- **Estimer les corrélations existantes entre les données observées**

Introduction



Le but de la statistique est de traiter des données (en général en grand nombre) en vue d'en tirer une information utile. Ces données proviennent d'un sondage, de mesures sur un central téléphonique, d'une enquête, d'un recensement, etc..., et sont donc des réalisations d'un phénomène aléatoire.

En statistique, la donnée de base est la donnée d'une réalisation x de la variable aléatoire X de loi inconnue P supposée appartenir à un certain ensemble de probabilités.

A partir d'une observation $X(\omega)$, on essaie d'avoir des informations sur la loi.

La famille des lois possibles est ce que l'on appelle le **modèle statistique**. Le modèle statistique décrit l'information dont on dispose a priori sur le phénomène aléatoire considéré et l'ensemble des paramètres inconnus. Il décrit tout ce qu'on ignore sur le phénomène aléatoire.

Dans cet leçon, notre sujet est d'introduire les méthodes de traitement des données collectées

Il existe de nombreux logiciels de statistique qui permettent de mieux se familiariser avec les notions de statistique (Excel, R, IBM SPSS, Scilab, Matlab, ...). Dans ce cours, nous utilisons le logiciel Scilab qui est un très bon logiciel assez complet. Les exemples des différentes sections sont accompagnés de leur code Scilab.

Le logiciel Scilab est téléchargeable à l'adresse suivante : <https://www.scilab.org/fr/download/latest> (Télécharger la version 32 bits pour windows)

Une vidéo¹ pour vos premiers pas sous Scilab est disponible

Un polycopié introductif (cf.) à l'utilisation de Scilab est disponible (chapitre 1)

Un polycopié introductif (cf.) de l'utilisation de Scilab dans la statistique est également disponible

1 - https://youtu.be/_GgtRCtds2s

Représentation des données statistiques



Objectifs

À la fin de cette section, vous serez capable :

- Organiser les données sous formes de classes représentatives

A. Série statistique

Il existe deux types de caractères statistiques qui permettent de qualifier les données utilisées:

- les caractères de type qualitatif (qui ne peuvent pas être mesurés) qui constitue une nomenclature, par exemple le sexe (Masculin/Féminin), les couleurs . . .
- les caractères de type quantitatif (qui peuvent être mesurés) et que l'on peut représenter par un nombre réel, par exemple taille, le poids, . . .

Dans nous allons nous intéresser aux données de type quantitatif que nous allons représenter par des variables statistiques.



Définition

Soit P un ensemble, appelé « population », alors une variable statistique est une application :

$$X: P \rightarrow \mathbb{R}$$

$$i \rightarrow X(i)$$

$X(P)$ est appelé univers image de X et on dira que :

- X est discrète si $X(P)$ est un ensemble discret (fini ou dénombrable)
- X est continue si $X(P)$ est un intervalle de \mathbb{R}

$N = \text{Card}(P)$ sera appelé l'effectif total



Attention

Dans la pratique on s'intéresse surtout à l'univers image $X(P)^{-1}$ et on identifie souvent une variable statistique avec la liste des valeurs $X(i)$ prises par la variable. Dans ce cas on parle en général de série statistique.



Exemple : Exemple illustratif

Exemples de séries statistiques : on considère la population P composée des étudiants inscrits au semestre 1 de la licence 1 MULTIMÉDIA à l'UVCI (on admettra qu'on a un effectif total $N = 100$) et nous définissons deux variables statistiques :

X = "nombre de participation au forum de mathématiques au semestre 1"

Y = "La moyenne sur 20 obtenue au premier semestre à l'examen de mathématique"

Nous avons bien des données de type quantitatif qui peuvent être décrits par une série statistique représentés par ces deux séries de nombres

$X = [0 \ 5 \ 1 \ 0 \ 1 \ 2 \ 1 \ 5 \ 0 \ 3 \ 0 \ 5 \ 1 \ 3 \ 5 \ 2 \ 0 \ 1 \ 2 \ 2 \ 1 \ 5 \ 4 \ 5 \ 2 \ 5 \ 0 \ 3 \ 4 \ 1 \ 5 \ 4 \ 4 \ 3 \ 4 \ 1 \ 1 \ 1 \ 3 \ 4 \ 1 \ 2 \ 5 \ 3 \ 5 \ 1 \ 4 \ 4 \ 1 \ 2 \ 2 \ 3 \ 4 \ 0 \ 4 \ 5 \ 2 \ 2 \ 3 \ 4 \ 4 \ 4 \ 2 \ 2 \ 5 \ 3 \ 5 \ 1 \ 3 \ 3 \ 5 \ 0 \ 3 \ 3 \ 0 \ 5 \ 5 \ 2 \ 4 \ 0 \ 3 \ 1 \ 2 \ 5 \ 1 \ 1 \ 2 \ 2 \ 4 \ 5 \ 0 \ 1 \ 4 \ 2 \ 5 \ 5 \ 3 \ 5 \ 2 \ 3]$

et

$Y = [0,84 \ 17,15 \ 9,79 \ 16,40 \ 5,20 \ 9,07 \ 13,96 \ 15,97 \ 7,47 \ 1,18 \ 0,16 \ 10,86 \ 17,82 \ 12,74 \ 8,87 \ 15,94 \ 8,16 \ 16,53 \ 4,46 \ 3,42 \ 13,60 \ 16,52 \ 17,39 \ 5,83 \ 3,25 \ 17,63 \ 13,53 \ 4,27 \ 7,64 \ 2,14 \ 13,50 \ 17,57 \ 7,27 \ 8,20 \ 4,73 \ 13,22 \ 15,82 \ 10,19 \ 9,01 \ 9,60 \ 6,37 \ 13,40 \ 15,61 \ 12,67 \ 15,09 \ 6,70 \ 17,52 \ 3,99 \ 2,22 \ 5,22 \ 7,34 \ 9,04 \ 7,79 \ 13,18 \ 12,67 \ 13,78 \ 8,59 \ 8,68 \ 13,87 \ 9,78 \ 0,73 \ 14,41 \ 2,36 \ 2,45 \ 12,63 \ 9,38 \ 2,26 \ 1,24 \ 12,00 \ 2,27 \ 4,41 \ 17,23 \ 7,17 \ 10,62 \ 7,52 \ 17,00 \ 6,40 \ 10,67 \ 1,10 \ 15,03 \ 8,62 \ 13,04 \ 4,60 \ 9,42 \ 13,29 \ 4,16 \ 7,91 \ 3,92 \ 16,17 \ 8,65 \ 0,28 \ 13,76 \ 10,21 \ 2,77 \ 0,67 \ 4,09 \ 14,20 \ 5,92 \ 4,20 \ 6,44]$

-La série statistique X est discrète puisque l'ensemble des valeurs prises qui est le nombre de participations effectives au forum est l'ensemble discret $\{0, 1, 2, 3, 4, 5\}$.

-La série Y est continue puis l'ensemble des valeurs en l'occurrence les moyennes sont dans l'intervalle $[0, 20]$

Une telle représentation des données n'est pas commode pour leur traitement. Les données ainsi recueillies sont appelées **données brutes**. Nous donnons une première forme de représentation de ces données brutes



Définition : Modalités

Soit X une variable statistique alors on appelle modalités de X :

- l'ensemble des valeurs différentes $\{x_1; x_2; \dots; x_n\}$ prises par X si X est discrète
- un ensemble d'intervalles $\{[x_1; x_2]; [x_2; x_3]; \dots; [x_{n-1}; x_n]\}$ formant une partition de $X(P)$ si X est continue

un tableau statistique est un tableau regroupant les caractéristiques d'une série statistiques par modalités :

Série discrète

modalité	x_1	x_2	x_3	x_4	...
effectif n_i					
fréquence f_i					

Série continue

modalité	$]x_1; x_2]$	$]x_2; x_3[$	$]x_3; x_4]$	$]x_4; x_5]$...
effectif n_i					
fréquence f_i					

Si N est l'effectif total on a

$$f_i = \frac{n_i}{N} \quad \forall i = 1, 2, \dots, n$$

On représentera graphiquement les données d'une série discrète (resp. continue) en dessinant des colonnes de hauteurs (resp. surface) proportionnelle à l'effectif de chaque modalité, c'est ce qu'on appelle **un diagramme en bâtons** (resp. **un histogramme**)



Exemple : Tableau statistique d'une série statistique discrète

Il faut se rendre à l'évidence compte que la série statistique X telle que formulée précédemment n'est pas très pratique à traiter. La première étape est de construire un tableau statistique à partir des données brutes.

Pour faire le tableau statistique de la série X , dont les modalités sont $\{0; 1; 2; 3; 4; 5\}$. La série X a un effectif total $N=100$, il faudra à partir des données brutes construire le tableau :

modalité	0	1	2	3	4	5
effectif n_i	11	18	18	16	16	21
fréquence f_i	0.11	0.18	0.18	0.16	0.16	0.21

il est facile de voir que :

$$n_1 + n_2 + n_3 + n_4 + n_5 = 100$$

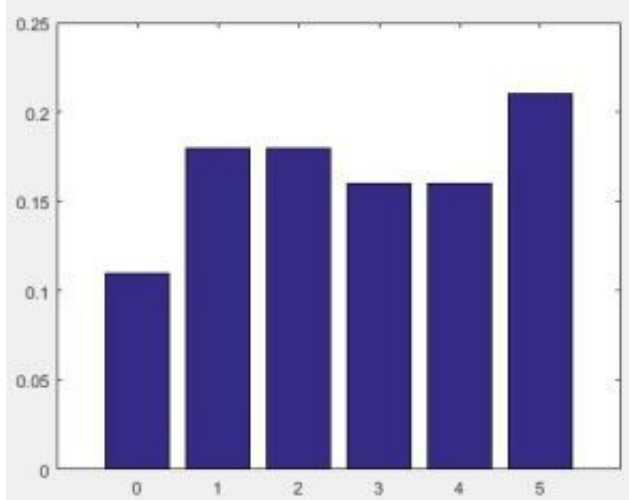
et

$$f_1 + f_2 + f_3 + f_4 + f_5 = 1$$



Exemple : Représentation sous forme d'un diagramme en bâtons

La représentation de la série X à partir du tableau précédent peut être vue graphiquement à l'aide d'un diagramme en bâtons dans laquelle la hauteur de chaque cellule est la fréquence de la modalité.





Exemple : Série statistique continue

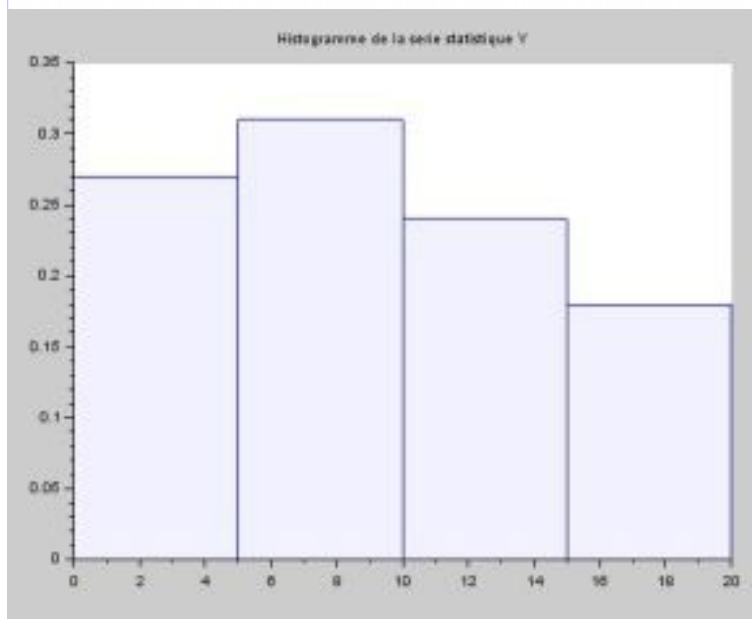
Reprenons notre série statistique Y. Cette série est continue, Pour mieux l'étudier, nous allons regrouper les données par intervalle de taille 5. le choix de largeur optimal des intervalles est hors programme. Nous avons alors le tableau suivant :

modalité]0,5]]5,10[]10, 15]]15, 20]
effectif n_i	27	31	24	18
fréquence f_i	0.27	0.31	0.24	0.18



Exemple : Représentation sous forme d'histogramme

A partir du tableau précédent des effectifs, nous allons représenter la série statistique y sous forme d'histogramme qui est équivalent à la représentation en diagramme en battons pour les séries statistiques discrètes. dans un histogramme, la hauteur de chaque cellule est la fréquence de la classe ou intervalle $]x_{i-1}; x_i]$. nous avons donc :



Un autre représentation permettant d'identifier la nature de la répartition des données statistiques est donnée par la fonction de répartition, qui est étroitement associée à la notion de cumul des fréquences.



Définition

Soit X une série statistique alors la fonction de répartition de X est donnée par :

$$F : \mathbb{R} \rightarrow [0; 1]$$

$$x \rightarrow \sum_{\{i|x_i \leq x\}} f_i$$



Rappel : Propriétés des fonctions de répartition

Soit F la fonction de répartition d'une série statistique X alors

- F est croissante
- $\lim_{t \rightarrow -\infty} F(t) = 0$
- $\lim_{t \rightarrow +\infty} F(t) = 1$
- si X est une variable discrète alors F est constante par morceau.



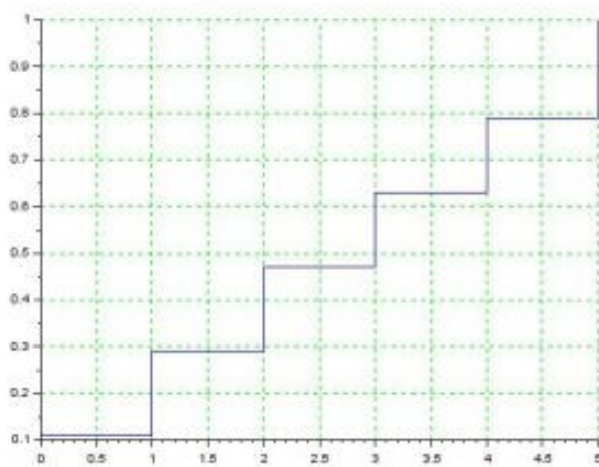
Exemple : Fonction de répartition de la série statistique discrète X

Pour construire la fonction de répartition de la variable X il faut donc faire le cumul des fréquences. Pour cela repartons du tableau statistique et ajoutons

une ligne pour calculer le cumul des fréquences :

modalité	0	1	2	3	4	5
effectif n_i	11	18	18	16	16	21
fréquence f_i	0.11	0.18	0.18	0.16	0.16	0.21
cumul	0.11	0.29	0.47	0.63	0.79	1

Une représentation graphique de la fonction de répartition est donnée sur la figure suivante. Ici elle est limitée aux modalités. Dans la pratique elle doit être prolongée à droite de 5 où elle vaut 1 et à gauche de 0 où elle vaut 0





Exemple : Fonction de répartition de la série Y

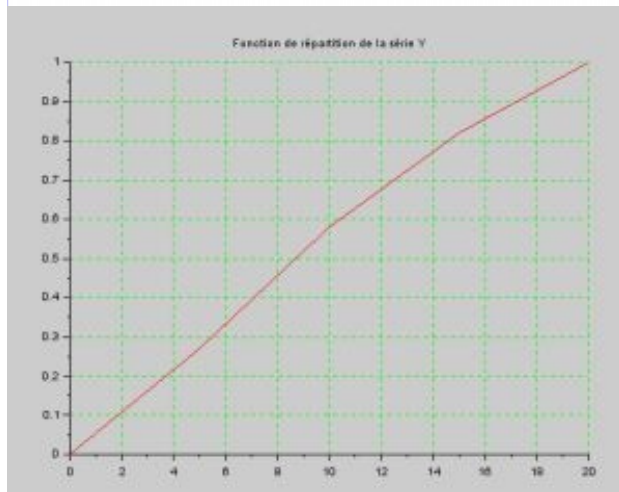
Pour calculer la fonction de répartition de la série statistique Y, on procède comme pour la série statistique X. On calcule les fréquences cumulées.

modalité]0,5]]5,10]]10, 15]]15, 20]
fréquence f_i	0.27	0.31	0.24	0.18
cumul	0.27	0.58	0.72	1

La différence majeure est que l'on calcule les fréquences cumulées aux bornes des intervalles sauf pour le premier intervalle, dans lequel la fréquence cumulée vaut 0. Ainsi dans le cas de notre série statistique précédente, Y nous avons le tableau suivant :

t	$]-\infty, 0[$	5	10	15	20	$[20, +\infty[$
F(t)	0	0.27	0.58	0.72	1	1

Il faut également noter que dans le cas d'une série statistique Y la fonction de répartition est une fonction affine par morceaux (de la forme $y=ax+b$ sur chaque segment)





Attention

Les fichiers contenant le code Scilab pour la série discrète X (cf.) et la série continue Y sont disponibles (cf.)

B. Exercice : Série statistique discrète

Exercice

Dans un test objectif comportant 10 questions, un professeur a relevé le nombre de bonnes réponses de chacun des ses 80 étudiants. Il a obtenu les données brutes suivantes :

2 3 5 5 4 6 6 5 4 3 7 7 7 6 2 7 7 9 8 10 5 6 6 8 6 6 3 7 3 5 9 7 6 4 7 5 9 9 6 9 6 3
9 8 8 7 5 6 10 6 9 7 7 7 4 7 10 8 7 10 3 5 8 5 8 7 4 8 10 7 4 6 6 8 7 7 7 8 8 9

Quelle est la population totale ?

Exercice

Déterminer la variable statistique X de l'exercice précédent. la réponse doit être donnée sous la forme d'une assertion entre double-côtes. Par exemple $X = \text{"Le numéro de la face supérieure du dé"}$

Exercice

La variable statistique X est continue ?

☐ Vrai

☐ Faux

C. Exercice

Donner le tableau des effectifs de la série statistique. On notera les résultats sous forme d'un ensemble de couples . par exemple $\{(1,10), (4,13), (10,23)\}$. le premier terme du couple correspond à la modalité et le second à l'effectif

D. Exercice

Donner le tableau des fréquences de la série statistique. On notera les résultats sous forme d'un ensemble de couples . par exemple $\{(1,10), (4,13), (10,23)\}$. le

premier terme du couple correspond à la modalité et le second à la fréquence. la précision est de quatre chiffres après la virgule)

E. Exercice

Donner le tableau des effectifs cumulés de la série statistique. On notera les résultats sous forme d'un ensemble de couples . par exemple $\{(1,10), (4,13), (10,23)\}$. le premier terme du couple correspond à la modalité et le second à la fréquence cumulée

F. Exercice

Combien d'étudiants ont répondu moins de sept fois correctement ?

G. Exercice : Série statistique continue

On considère les données brutes de la série statistique Y qui représente la taille (en cm) des étudiants de la filière Multimédia de l'UVCI (on relève la taille sur une population de 100 étudiants). les données brutes sont représentées par la matrice Y suivante :

$Y=[188, 169, 181, 180, 159, 180, 164, 184, 177, 159, 187, 174, 170, 173, 175, 150, 171, 166, 173, 182, 167, 173, 174, 175, 166, 173, 180, 188, 165, 173, 178, 182, 175, 186, 162, 193, 173, 184, 184, 184, 182, 187, 193, 161, 193, 169, 163, 179, 179, 184, 182, 182, 172, 175, 193, 170, 176, 166, 177, 163, 191, 171, 189, 183, 178, 197, 166, 187, 180, 172, 181, 167, 177, 177, 186, 174, 168, 182, 183, 182, 170, 186, 193, 167, 184, 159, 169, 192, 172, 167, 178, 176, 177, 175, 172, 175, 182, 176, 179, 188]$

Exercice

La série statistique Y est-elle continue ?

☐ vrai

☐ faux

Exercice

Nous voulons regrouper les données en classes de largeur 10 Lequel des tableaux suivants représente les données de la série statistique Y



modalité] 150,160]] 160,170]] 170,180]] 180,190]]190,200]
Effectif	12	24	30	12	24
fréquence	0.12	0.24	0.3	0.12	0.24



modalité	[150, 160[[160, 170[[170,180 [[180, 195]	[190,200[
Effectif	4	21	39	28	8
fréquence	0.04	0.21	0.39	0.28	0.08



modalité] 150,160]]160, 170]] 170,180]]180, 190]]190, 200]
Effectif	4	21	39	28	8.
fréquence	0.04	0.21	0.39	0.28	0.08



modalité]150, 165]]165, 170]] 170,175]]175, 180]] 180,190]] 190,200]
Effectif	10	15	21	18	28	8
fréquence	0.1	0.15	0.21	0.18	0.28	0.08



Aucun des tableaux présentés

H. Exercice : Effectif corrigés dans une série statistique

Lorsque l'on est en face d'une série statistique continue dans laquelle la largeur des intervalles varie, il est de coutume de corriger la hauteur de ces classes afin de faciliter le dessin de l'histogramme et de mieux affiner la répartition des données. Pour ce faire on utilise le résultats suivant

Soit X une variable statistique continue alors ayant des modalités $]x_1; x_2]$ $]x_2; x_3]$. . . $]x_{n-1}; x_n]$ d'amplitude différentes alors pour chaque modalité on appelle « effectif corrigé » ou « hauteur corrigée » la hauteur qu'il faut donner dans l'histogramme à la colonne pour que sa surface soit proportionnelle à son effectif réel.

$$h_i = \frac{f_i}{|x_i - x_{i-1}|} = \frac{n_i}{N|x_i - x_{i-1}|}$$

Dans un histogramme normalisé on calcule les hauteurs corrigées par la formule : de telle sorte que la surface totale des colonnes soit égale à 1.

Exercice

En reprenant les hypothèses de l'exercice précédent ? lequel de ces tableaux représentent la série statistique Y

$Y_i=$]150, 165]]165, 170]]170,175]]175, 180]]180,190]]190,200]
f_i	0.1	0.15	0.21	0.18	0.28	0.08
$ x_i - x_{i-1} $	14	5	5	5	10	10
h_i	0.006667	0.03	0.042	0.036	0.028	0.008

$Y_i=$]150, 165]]165, 170]]170,175]]175, 80]]180,190]]190,200]
f_i	0.1	0.21	0.08	0.15	0.28	0.18
$ x_i - x_{i-1} $	15	5	5	5	10	10
h_i	0.006667	0.03	0.042	0.036	0.028	0.008

$Y_i=$]150, 165]]165, 170]]170,175]]175, 180]]180,190]]190,200]
f_i	0.1	0.15	0.21	0.18	0.28	0.08
$ x_i - x_{i-1} $	15	5	5	5	10	10
h_i	0.006667	0.03	0.042	0.036	0.028	0.008

$Y_i=$]150, 165]]165, 170]]170,175]]175, 180]]180,190]]190,200]
f_i	0.1	0.28	0.15	0.18	0.21	0.08
$ x_i - x_{i-1} $	15	5	5	5	10	10
h_i	0.03	0.036	0.036	0.042	0.028	0.006667

I. Exercice : Fonctions de repartition

Exercice

Yao qui est gardien de but de l'équipe de football de son équipe note le nombre de buts encaissés par son équipe au cours de chaque match. Pour la dernière saison, il a compilé toutes ses données sous la forme du tableau suivant :

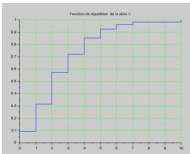
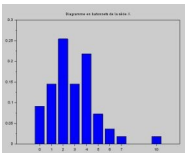
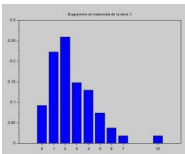
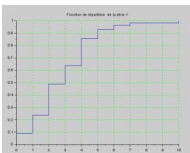
$X_i =$	0	1	2	3	4	5	6	7	10
n_i	5	12	14	8	7	4	2	1	1
f_i									
cumul									

Lesquelles des assertions suivantes sont correctes ?

- ☐ La population totale est 54
- ☐ La population totale est 55
- ☐ Les différentes fréquences sont [0.0925926 0.2222222 0.2592593 0.1481481 0.1296296 0.0740741 0.037037 0.0185185 0.0185185]
- ☐ les différentes fréquences cumulées sont [0.0925926 0.3148148 0.5740741 0.7222222 0.8518519 0.9259259 0.962963 0.9814815 1]

Exercice

Lesquels des graphiques sont représentatifs de la série statistique X ?

- ☐ 
- ☐ 
- ☐ 
- ☐ 

J. Exercice

Exercice

On considère la série statistique Y qui mesure les temps de parcours (en

minutes) des 40 participants d'une course de vitesse. Les données ont été relevées et sont résumées dans le tableau suivant :

Classes (minutes) $]b_{i-1}, b_i]$	[43, 45]	[45, 47]	[47, 49]	[49, 51]	[51, 53]	[53, 55]	[55, 57]
Effectif n_i	2	3	7	11	8	6	3

On s'intéresse à la fonction de répartition dans la classe $]49, 51]$. laquelle des fonctions suivantes représentent la restriction de la fonction de répartition à cette classe ?

☐ $y = (11/240)x - 467/240$

☐ $y = (11/80)x - 515/80$

☐ $y = (25/512)x + 10/180$

Paramètres statistiques



Objectifs

A la fin de cette section, vous serez capables de :

- Estimer les paramètres tels que la moyenne et la variance

L'analyse des données statistiques doit permettre d'isoler certaines valeurs remarque définissant la manière dont son réparties les données statistiques. Le plus connu est la moyenne empirique ainsi qu'un indicateur qui lui est étroitement associé la variance.

A. Moyenne et variance

La moyenne permet de cerner facilement la valeur centrale de la série statistique.



Attention

Il faut distinguer le calcul de la moyenne dans le cas d'une série statistique discrète de celui d'une série statistique continue



Définition : Moyenne d'une série statistique discrète

Soit X une série statistique alors la moyenne de X , notée \bar{X} , est la valeur :

$$\bar{X} = \frac{1}{N} \sum_{k \in P} X(k) = \frac{1}{N} \sum_{i=1}^n n_i x_i = \sum_{i=1}^n f_i x_i$$

n désigne le cardinal de l'ensemble des modalités



Exemple

Nous reprenons notre série statistique discrète X . Nous pouvons calculer directement la moyenne à partir du tableau construit dans la section précédente

$$\bar{X} = 0.11 \times 0 + 0.18 \times 1 + 0.18 \times 2 + 0.16 \times 3 + 0.16 \times 4 + 0.21 \times 5 = 2.71$$



Définition : Moyenne d'une série statistique continue

Soit X une variable continue regroupée suivant les modalités $]x_1; x_2]$ $]x_2; x_3]$. . . $]x_{n-1};$

$x_n]$. Si on pose le milieu de chaque modalité $xc_i = \frac{x_i + x_{i+1}}{2}$ alors la

moyenne
approchée est définie par

$$\tilde{X} = \sum_{i=1}^n f_i \times xc_i$$



Exemple

Nous allons estimer la moyenne de la série statistique Y. Nous avons regroupé les données de Y en quatre classes de largeur 5.

modalité]0 ;5]]5 ;10]]10 ;15]]15 ;20]
effectif n_i	27	31	24	18
fréquence f_i	0.27	0.31	0.24	0.18

la moyenne approchée est d'après la formule ci-dessus est

$$\tilde{Y} = 0.27 \times 2.5 + 0.31 \times 7.5 + 0.24 \times 12.5 + 0.18 \times 17.5 = 9.15$$



Définition : Variance et écart-type

Soit X une série statistique de moyenne $\mu = \bar{X}$ alors on appelle variance de X la valeur

$$Var(X) = \overline{(X - \mu)^2}$$

et l'écart-type de X noté σ_X est défini par : $\sigma_X = \sqrt{Var(X)}$



Complément

Pour une variable continue si on a pas accès aux données brutes on pourra aussi calculer une variance et un écart-type approché (comme pour la moyenne approchée).



Complément

Dans la pratique on calcule la variance à l'aide de la formule de Koenig suivante

$$Var(X) = \overline{X^2} - \bar{X}^2$$



Exemple : Variance d'une série statistique

Nous allons à nouveau considérer notre série statistique discrète X de la section précédente. Nous allons estimer sa variance et son écart-type. Nous disposons de deux méthodes :

Méthode 1 utilisation du théorème de Koenig

Nous allons rappeler dans le tableau suivant, les principaux paramètres de la série X

modalité	0	1	2	3	4	5
effectif n_i	11	18	18	16	16	21
fréquence f_i	0.11	0.18	0.18	0.16	0.16	0.21

Calculons la moyenne de la série X^2 nous avons :

$$\text{moy}(X^2) = 0 \times 0.11 + 1 \times 0.18 + 2^2 \times 0.18 + 3^2 \times 0.16 + 4^2 \times 0.16 + 5^2 \times 0.21 = 10.15$$

puis à l'aide du théorème de Koenig la variance est alors :

$Var(X) = 10.15 - 2.71^2 = 2.8059$. l'écart-type qui est la racine carrée de la variance est alors de 1.67

Méthode 2. Directement à partir des données brutes

Un estimateur sans biais de la variance est donné par la formule suivante :

$$Var(X) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

Tout d'abord on estime la moyenne de X. Cela peut être fait par les méthodes vues précédemment et on obtient que $\text{moy}(X) = 2.71$.

On construit la nouvelle série statistique $Y = X - \text{moy}(X)$.

La variance est alors la moyenne de la série Y^2 . Un calcul nous donne $\text{var}(X) = 2.8059$. l'écart-type qui est la racine carrée de la variance est alors de 1.67



Remarque

Cette deuxième méthode sera surtout utile dans l'estimation de la variance d'une série statistique continue

Nous allons donner une définition qui généralise la notion de variance au cas deux séries statistiques distinctes



Définition : Covariance

Soient X et Y deux séries statistiques d'effectifs N alors la covariance de X et Y est donnée par :

$$COV(X, Y) = \overline{(X - \bar{X})(Y - \bar{Y})}$$

Dans la pratique elle est calculée de la façon suivante :

$$COV(X,Y) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

B. Exercice

Dans un groupe de 10 personnes, on mesure les tailles t_i , en cm :

t_i	150	153	157	158	158	162	165	165	168	74
-------	-----	-----	-----	-----	-----	-----	-----	-----	-----	----

Exercice

On définit une nouvelle série $x_i = t_i - 160$. On construit la nouvelle série $x_i = t_i - \text{moy}(x_i)$. laquelle des assertions suivantes est vérifiée ?

☐

t _i	150	153	157	158	158	162	165	165	168	174
----------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

☐

x _i	149	152	156	157	157	161	164	164	167	173
----------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

☐

t _i	-10	-7	-3	-2	-2	2	5	5	8	14
----------------	-----	----	----	----	----	---	---	---	---	----

☐

x _i	150	153	157	158	158	162	165	165	168	74
----------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	----

Exercice

Moyenne de la série statistique x_i

<input type="radio"/>	0
<input type="radio"/>	161
<input type="radio"/>	160

Exercice

Variance de la série statistique x_i

- ☐ $\text{Var}(X_i) = 52.222$
- ☐ $\text{Var}(X_i) = 0$
- ☐ $\text{Var}(X_i) = 2457$

Exercice

Exprimez les paramètres de la série initiale t_i

- ☐ moy(t_i)=161 var(t_i)=52.222
- ☐ moy(t_i)=160 var(t_i)= 47
- ☐ moy(t_i)=0 var(t_i)=52.222

C. Exercice : Exercice de niveau plus difficile

On considère la série statistique

$X = [2, 4, 2, 1, 0, 0, 3, 1, 3, 0, 2, 3, 0, 3, 1, 0, 2, 0, 2, 1, 0, 3, 1, 2, 2, 3, 1, 3, 2, 2, 0, 2, 0, 3, 2, 1, 2, 2, 3, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 0, 3, 1, 0, 2, 0, 2, 1, 2, 2, 2, 1, 2, 0, 2, 1, 0, 0, 4, 1, 0, 3, 0, 2, 0, 0, 3, 1, 3, 0, 0, 1, 2, 1, 2, 6, 2, 2, 1, 1, 2, 3, 1, 0, 1, 1, 1, 3, 1, 3, 2]$ correspondant aux nombre d'absences injustifiées d'une population de $N=100$ étudiants.

Exercise

Que est le mode de cette série statistique

- | | |
|---|---|
| ○ | 2 |
| ○ | 4 |
| ○ | 3 |
| ○ | 5 |

Exercise

Quelle est la valeur de la médiane cette série statistique

Paramètres statistiques

<input type="radio"/>	3.5
<input type="radio"/>	6
<input type="radio"/>	2.5
<input type="radio"/>	7

Exercice

Donner la valeur de l'inter-quartile de cette série statistique

Corrélation



Objectifs

A la fin de cette section, vous serez capables de :

- **Estimer les corrélations existantes entre les données observées**

En statistique il est souvent important de rechercher s'il existe un lien entre deux variables X et Y , lien qui, dans l'idéal, pourrait s'exprimer par une équation de la forme $Y = aX + b$.

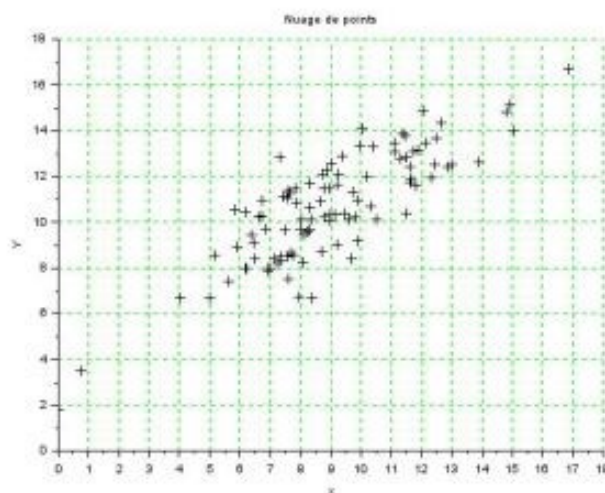
Considérons par exemple les 2 séries statistiques suivantes :

X = "La moyenne sur 20 en mathématiques au semestre 1"

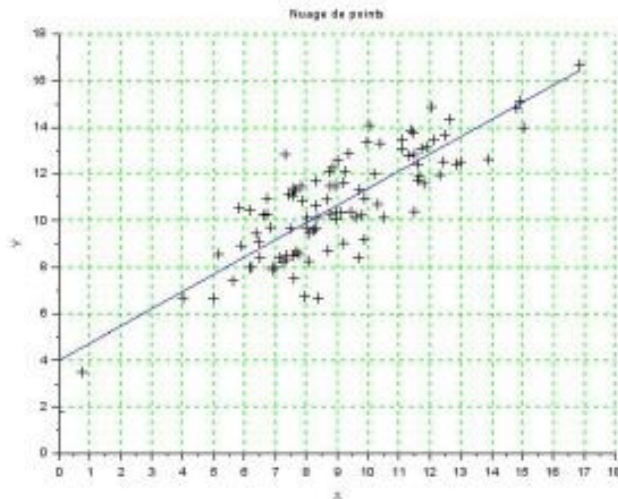
Y = "La moyenne générale sur 20 obtenue au premier semestre "

Les données brutes sont dans le fichier suivant (cf.)

On voudrait quantifier le lien entre ces deux variables statistiques. Si on place les points de coordonnées (X_i, Y_i) sur un graphe on obtient le "nuage" suivant :



ces points semblent à première vue alignés sur une droite, cela signifie qu'on peut quantifier le lien entre Y et X par une équation de la forme $Y = aX + b$. **Le but est donc de trouver les coefficients a et b** de telle sorte que la droite $y = ax + b$ passe au plus près du maximum de points comme sur la figure suivante 'droite en bleue). C'est ce qu'on appelle faire **une régression linéaire**.



A. Ajustement linéaire

Nous allons apprendre à déterminer les coefficients a et b . Pour cela nous allons utiliser la méthode des moindres carrés. Cette méthode consiste à minimiser l'erreur globale qu'on comment en écrivant que $Y = aX + b$.



Fondamental

Soient X et Y deux séries statistiques alors la méthode des moindres carrés consiste à chercher la "meilleure" droite, d'équation $y = ax + b$, passant par le nuage de points (X, Y) comme étant la droite qui minimise la somme des carrés des écarts entre les points (X_i, Y_i) et $(X_i, aX_i + b)$ c'est à dire on cherche a et b qui minimisent la fonction

$$\varphi(a, b) = \sum_{i=1}^N (Y_i - a \times X_i - b)^2$$



Fondamental : Droite d'ajustement linéaire

Soient X et Y deux séries statistiques, alors la droite d'ajustement linéaire de Y en X a pour équation $y = ax + b$ avec :

$$a = \frac{COV(X, Y)}{Var(X)} \quad \text{et} \quad b = \bar{Y} - a\bar{X}$$



Attention

on remarquera que l'ajustement de Y par rapport à X n'est pas le même que l'ajustement de X par rapport à Y

Afin de quantifier la qualité de l'approximation $Y = aX + b$ on utilise un paramètre appelé le coefficient de corrélation linéaire.



Définition : Coefficient de corrélation linéaire

Soient X et Y deux séries statistiques, on suppose que la droite d'ajustement linéaire de Y en X a pour équation $y = ax + b$ et on définit son coefficient de corrélation par :

$$\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

alors le nuage de points (X, Y) est d'autant plus proche de la droite d'ajustement linéaire de Y en X que $|\rho|$ est proche de 1. En particulier tous les points du nuage sont sur la droite $y = ax + b$ si et seulement si $\rho = 1$

La corrélation sera dite forte (resp. faible) si

$$\rho \geq \frac{\sqrt{3}}{2} \approx 0.8666 \text{ (resp } \rho < \frac{\sqrt{3}}{2} \text{)}$$



Exemple

Nous allons reprendre l'exemple introductif de cette section avec nos deux séries statistiques :

X = "La moyenne sur 20 en mathématiques au semestre 1"

Y = "La moyenne générale sur 20 obtenue au premier semestre "

Le fichier suivant contient les données brutes des deux séries ainsi que le code Scilab (cf.) que vous pourrez utiliser pour générer les paramètres et tracer les courbes présentées dans le cours. Afin de déterminer les paramètres de la droite d'ajustement, on effectue les étapes suivantes :

1) Estimation des paramètres de chaque série :

$$\bar{X} = 8.996117 \quad Var(X) = 7.255691 \quad \sigma_X = 2.693639$$

$$\bar{Y} = 10.636117 \quad Var(Y) = 5.818773 \quad \sigma_Y = 2.412213$$

$$COV(X,Y) = 5.358226$$

2) Paramètres de la droite de régression linéaire

$$a = \frac{COV(X,Y)}{Var(X)} = \frac{5.358226}{7.255691} = 0.738486$$

$$b = \bar{Y} - a\bar{X} = 10.636117 - 0.738486 \times 8.996117 = 3.992611$$

$$\rho = \frac{COV(X,Y)}{\sigma_X \times \sigma_Y} = \frac{5.358226}{2.693639 \times 2.412213} = 0.824643$$

La corrélation est faible.

la courbe en bleue donne la corrélation de X par rapport à Y et comme vous pouvez le constater elles sont différentes.

Exercice

IV

On reprend les hypothèses de l'exemple et on considère nos deux séries statistiques X et Y . Cette fois ci on va étudier la corrélation entre la série X et la série Y . Les données des deux séries ont données dans le fichier suivant. Vous pouvez utiliser le logiciel Scilab pour calculer les paramètres et tracer les courbes graphiques

Exercice

Parù les assertions suivantes, lesquelles sont vérifiées ?

- ☐ $M(Y) = 8.996117$ $Var(X) = 7.255691$ $Ecart\text{-}type(X) = 2.693639$
- ☐ $M(Y) = 10.636117$ $Var(Y) = 5.818773$ $Ecart\text{-}type(Y) = 2.412213$
- ☐ $M(X) = 10.636117$ $Var(Y) = 5.818773$ $Ecart\text{-}type(Y) = 2.412213$
- ☐ $M(X) = 8.996117$ $Var(X) = 7.255691$ $Ecart\text{-}type(X) = 2.693639$
- ☐ $M(X) = 10.996117$ $Var(X) = 12.255$ $Ecart\text{-}type(X) = 3.563639$

Exercice

Lesquelles des affirmations suivantes sont vérifiées ?

- ☐ $cov(X, Y) = cov(Y, X)$
- ☐ $COV(X, Y) = 5.358226$
- ☐ $COV(X, Y) = 93345643$
- ☐ $COV(X, Y) = Var(X)Var(Y)$
- ☐ $COV(X, Y) = E(XY)$

Exercice

Estimer les paramètres de la droite d'ajustement linéaire. Les données seront exprimées sous la forme d'une liste de 3 nombres réels. Par exemple a

Exercice

$=0.234564$ $b=3.214321$ $corr=62.432534$ (précision de 6 chiffres après la virgule)

Exercice

La corrélation entre Y et X est forte :

☐ Vrai

☐ Faux

Exercice

V

On soupçonne que l'acidité d'un sol (ph) est liée à la présence d'aluminium échangeable (qae) suivant la loi

$$qae = k \times A^{ph} \Leftrightarrow \ln(qae) = \ln(k) + ph \times \ln(A)$$

Pour vérifier cette hypothèse on a mesuré le ph et la quantité qae d'aluminium échangeable (en p.p.m.) en divers points du sol :

Le tableau suivant résume les données brutes recueillies :

ph	4.2	4.4	4.8	5.1	5.4	5.6	6.2
qae	400	260	120	60	30	15	4

Exercice

Soit les séries statistiques $Y = \ln(qae)$ et $X = ph$. Estimez les paramètres de l'ajustement linéaire de Y par rapport à X . Les résultats seront donnés sous forme d'une liste de 3 nombres réels. Par exemple $a = 0.234564$ $b = 3.214321$ $corr = 62.432534$ (précision de 6 chiffres après la virgule)

Exercice

Calculez les valeurs de paramètres k et A . Les résultats seront donnés sous forme d'une liste de 2 nombres réels. Par exemple $A = 0.234564$ $k = 3.214321$ (précision de 6 chiffres après la virgule)

Exercice

Qualifier la qualité de l'ajustement

☐ La corrélation est faible

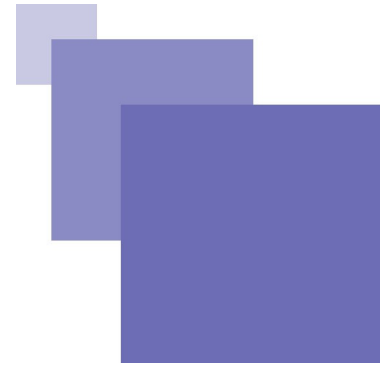
☐ La corrélation est forte

Exercice

Estimer la quantité d'aluminium échangeable pour un $pH = 5$

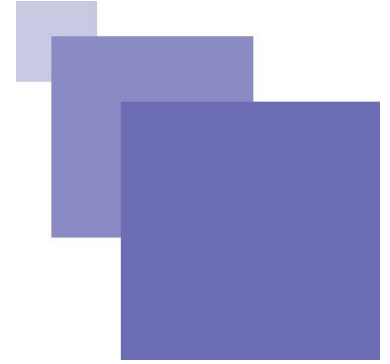


Conclusion



Ce chapitre met fin à ce cours introductif sur les probabilités et la statistique. Les notions abordées vous seront utiles dans différents domaines de l'informatique comme l'étude de l'évolution d'un processus informatique, la modélisation et le dimensionnement des réseaux de télécommunication et même dans le domaine du traitement de l'image. La maîtrise de ces outils est donc primordiale.

Bibliographie



[1] Pierre Andreoletti, Support du cours de Probabilités et Statistiques, IUT d'Orléans, Département Informatique, 2008

[2] Ph. Roux, Probabilités discrètes et statistique descriptive, DUT Informatique, semestre 2, 2010