Statistique descriptive et prévision

Année 2010/2011

L. Chaumont

Contents

1. Étude d'une variable **5**

1.1. Définitions 5

1.2. Représentations graphiques usuelles 8

1.3. Paramètres caractéristiques de tendance centrale 12

1.4. Paramètres caractéristiques de dispertion 13

2. Étude conjointe de deux variables **18**

2.1. Définitions et notations 18

2.2. Covariance 20

2.3. Régression linéaire 21

3. Prévision **24**

3.1. Introduction 24

3.2. Méthode des moyennes mobiles 25

3.3. Méthode de lissage exponentiel 26

3.4. Méthodes de régression linéaire 28

4. Intervalles de confiance **32**

4.1. Quelques notions de probabilités 32

4.2. Intervalles de confiance 36

Chapitre 1

Étude d'une variable.

1.1. Définitions.

1.1.1 Le vocabulaire usuel : L'ensemble des objets sur lesquels porte une étude statistique s'appelle une *population*. Il peut s'agir de personnes, de plantes, de produits alimentaires,... Chaque élément d'une population est un *individu*. On désigne généralement par Ω la population étudiée et par ω ses individus. Lorsque Ω est un ensemble fini, on énumère ses individus en les désignant par $\omega_1, \omega_2, \ldots, \omega_n$. L'indice n qui correspond au nombre d'individus de Ω est appelé la *taille* de la population, on note ceci par $\text{card}(\Omega) = n$.

Lorsque l'on étudie une population Ω, celle-ci n'est généralement pas considérée dans sa totalité (en particulier si Ω n'est pas un ensemble fini). On en étudie simplement une partie finie que l'on appelle un *échantillon*. De même on parle de la taille d'un échantillon pour désigner le nombre d'individus qu'il contient. L'échantillon étudié est souvent aussi noté Ω.

La première étape d'une étude satistique est de recueillir des données. A chaque individu d'une population ou d'un échantillon correspond une *donnée quantitative* ou *qualitative* qui est la valeur du *caractère* étudié. L'ensemble de ces données, lorsqu'elles n'ont pas encore été traitées, s'appelle une *série statistique brute*.

1.1.2 Variables : Dans une population (ou un échantillon d'une population) Ω, supposons que nous voulions étudier un certain *caractère* : le poids, la taille, la couleur,... des individus. Ce caractère prend généralement plusieurs valeurs que l'on appelle des *modalités*.

A ce caractère, on associe une variable qui est l'objet de base du formalisme de la statistique descriptive. Cette variable que nous noterons ici X, est une application qui à chaque individu associe une modalité du caractère étudié. L'espace des modalités sera noté E, c'est l'ensemble des valeurs que peut prendre X. Ceci se représente dans le formalisme mathématique usuel de la manière suivante :

$$\begin{array}{rcl} X : \Omega & \to & E \\ \omega & \mapsto & X(\omega) \end{array}$$

Lorsque l'ensemble des modalités est fini, nous noterons r son cardinal, c'est à dire $\text{card}(E) = r$.

Exemple : Ω = ensemble des produits fabriqués par une usine,
$X(\omega)$ = poids du produit $\omega \in \Omega$.

Il existe deux types de variable : les variables *qualitatives* et les variables *quantitatives*. Les variables quatitatives se distinguent des autres par le fait que l'on peut effectuer sur leurs modalités les opérations algébriques (+,-,×,÷). Il s'agit par exemple pour des personnes du poids de l'âge ou du revenu mensuel. Parmi les variables qualitatives on peut distinguer :

_ Les variables *qualitatitves nominales* comme la couleur ou le numéro de téléphone pour lesquelles on ne peut faire d'opérations algébriques et pour lesquelles il n'existe pas d'ordre naturel.

_ Les variables *qualitatives ordinales* comme l'année de naissance ou le rang de classement pour lesquelles on ne peut pas faire d'opérations mais dont on peut ordonner les modalités.

Rappelons qu'un ensemble E est dit dénombrable s'il peut être mis en bijection avec l'ensemble des entiers naturels $\mathbb{N}$, c'est à dire si l'on peut énumérer ses éléments : $E = \{e_1, e_2, \dots, e_k, \dots\}$. Si E est un ensemble fini ou dénombrable, on dira que la variable est *discrète* et si E est l'ensemble $\mathbb{R}$ des nombres réels ou un intervalle de $\mathbb{R}$, on dira que X est une variable *continue*.

1.1.3 Effectifs et fréquences : Nous supposons dans cette section que la variable considérée est finie (c'est à dire que E est fini) et qu'elle est soit quantitative, soit qualitative ordinale. La manière la plus simple d'analyser une série statistique brute est de la représenter dans un tableau indiquant le nombre d'individus présentant chacun des différents caractères.

Supposons que la population étudiée Ω soit de taille n et que l'ensemble E des modalités de la variable étudiée X est $E = \{x_1, \dots, x_r\}$. Pour $i \in \{1, \dots, r\}$, on note x_i la i-ième modalité et l'on suppose que ces modalités sont rangées par ordre croissant : $x_1 < x_2 < \cdots < x_r$.

Définitions :

1. *L'effectif correspondant à la modalité x_i est le nombre d'individus de la population Ω ayant comme modalité x_i. On le notera n_i.*

2. *La féquence correspondant à la modalité x_i sera notée f_i. C'est le rapport $f_i = \dfrac{n_i}{n}$.*

3. *L'effectif cumulé de la modalité x_i est le nombre d'individus, noté N_i, ayant une modalité inférieure ou égale à x_i. Ainsi on a $N_i = n_1 + n_2 + \cdots + n_i$.*

4. *La fréquence cumulée F_i se définit de même : $F_i = f_1 + f_2 + \cdots + f_i$.*

En particulier, nous avons les relations suivantes :

$$N_i = N_{i-1} + n_i\,, \quad F_i = F_{i-1} + f_i\,, \quad F_i = \frac{N_i}{n}\,.$$

On peut désormais présenter une série statistique brute par le tableau suivant.

X	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_r	f_r	N_r	F_r
	n	1		

Remarquons que l'on a toujours $N_r = n$ et $F_r = 1$.

Exemple : On a relevé le nombre d'enfants parmi 20 familles d'une certaine ville. On a obtenu la série statistique brute suivante :

$$4\,;3\,;2\,;1\,;3\,;0\,;2\,;3\,;2\,;4\,;5\,;2\,;0\,;2\,;3\,;4\,;5\,;2\,;3\,;2$$

Ces données s'organisent en le tableau suivant :

X	n_i	f_i	N_i	F_i
0	2	$0,1$	2	$0,1$
1	1	$0,05$	3	$0,15$
2	7	$0,35$	10	$0,5$
3	5	$0,25$	15	$0,75$
4	3	$0,15$	18	$0,9$
5	2	$0,1$	20	1
	20	1		

1.1.4 Regroupement par classes : Il arrive que l'on ait à regrouper les modalités d'une variable en un certain nombre d'intervalles que l'on appelle des *classes*. C'est le cas lorsque la variable est continue ou bien lorsque celle-ci est discrète et que seules certaines de ses modalités ont un intérêt pour l'étude.

Dans ce cas, on découpe l'ensemble E des modalités en des intervalles $[a_0, a_1[,\ [a_1, a_2[,\ \ldots,\ [a_{r-2}, a_{r-1}[,\ [a_{r-1}, a_r]$ qui forment une partition de E :

$$E = [a_0, a_1[\cup[a_1, a_2[\cup\cdots\cup[a_{r-2}, a_{r-1}[\cup[a_{r-1}, a_r]\,.$$

(Remarquons que pour définir une partition de E, il faut que le dernier intervalle soit fermé en a_r.)

L'intevalle $[a_{i-1}, a_i[$ est appelé la *classe* i.

Dans ce cas, l'effectif n_i est le nombre d'individus dont la modalité appartient à l'intervalle $[a_{i-1}, a_i[$, (ou $[a_{r-1}, a_r]$, s'il s'agit de n_r). On définit alors f_i, N_i et F_i de la même manière qu'à la section précédente.

Dans le cas d'un regroupement par classes, on définit encore le *centre* et l'*amplitude* de la classe i respectivement par

$$c_i = \frac{a_{i-1} + a_i}{2} \quad \text{et} \quad l_i = a_i - a_{i-1}\,.$$

Enfin, de manière à tenir compte de l'amplitude de la classe i lorsque l'on évalue son effectif, on définit la *densité de fréquence* (ou *fréquence relative*) par :

$$h_i = \frac{f_i}{l_i}\,.$$

C'est le cas par exemple dans la définition de la fonction de répartition empirique (voir section 1.2.2) ci-dessous.

1.2. Représentations graphiques usuelles.

Certains paramètres caractéristiques de dispertion ou de tendance centrale peuvent aussi se représenter graphiquement. On se reportera pour ceci à la section 2.3

1.2.1 Le cas discret : Considérons une variable finie, quantitative ou qualitative ordinale : $X : \Omega \to E$, avec $\Omega = \{\omega_1, \ldots, \omega_n\}$ et $E = \{x_1, \ldots, x_r\}$. On ordonne les modalités de X dans l'ordre croissant :

$$x_1 < x_2 < \cdots < x_r\,.$$

Le diagramme en bâtons.

En abscisse figurent les modalités rangées dans l'ordre croissant et en ordonnée figurent les effecifs ou bien les fréquences. La hauteur du bâton issu de x_i est égale à n_i (ou f_i). Pour chaque cas on précise si le diagramme est relatif aux fréquences ou bien au effectifs.

Le diagramme en bâton (relatif aux fréquences) correspondant à l'exemple de la section 1.1.3 est le suivant :

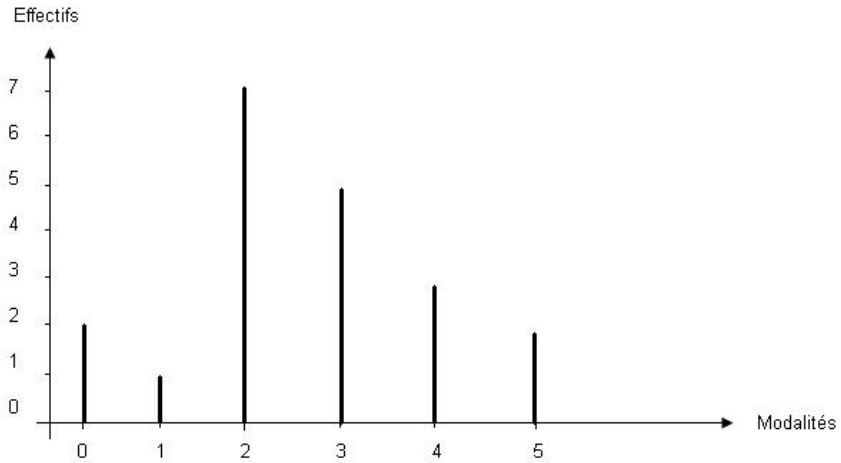


Diagramme en bâtons

Remarquons que l'origine des abscisses ne correspond pas nécessairement à l'origine des ordonnées.

Le diagramme cumulatif.

En abscisse figurent les modalités rangées dans l'ordre croissant et en ordonnée figurent les effecifs cumulés ou bien les fréquences cumulées. De même que pour le diagramme en bâtons, dans chaque cas on précise si le diagramme cumulatif est relatif aux effectifs cumulés ou au fréquences cumulées.

DÉFINITION : *Le diagramme cumulatif relatif aux fréquences cumulées est le graphe de la fonction F_X définie par :*

$$F_X(x) = \begin{cases} 0 & si\ x < x_1 \\ F_i & si\ x \in [x_i, x_{i+1}[,\ i = 1, \ldots, r-1 \\ 1 & si\ x \geq x_r. \end{cases}$$

F_X est appelée la fonction de répartition empirique de la variable X. On définit de la même manière le diagramme cumulatif relatif aux effectifs cumulés.

Le diagramme cumulatif (relatif au fréquences cumulées) correspondant à l'exemple de la section 1.1.3 est le suivant :

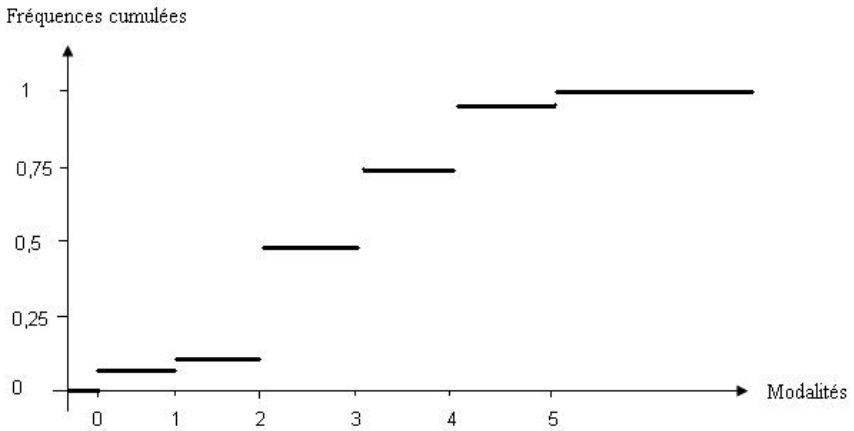


Diagramme cumulatif

1.2.2 Le cas continu : Nous nous plaçons ici dans le cas d'une variable continue (ou considérée comme telle) pour laquelle on a regroupé les modalités en les classes suivantes :

$$[a_0, a_1[, \quad [a_1, a_2[, \quad \ldots, [a_{r-1}, a_r[$$

La courbe cumulative.

C'est l'équivalent du diagramme cumulatif (relatif aux fréquences cumulées) dans le cas discret. Dans ce cas la fonction de répartition empirique peut être représentée au moyen d'interpolations linéaires.

DÉFINITION : *La courbe cumulative de la variable X est le graphe de la fonction F_X définie par :*

$$F_X(x) = \begin{cases} 0 & si\ x < a_0 \\ F_{i-1} + h_i(x - a_{i-1}) & si\ x \in [a_{i-1}, a_i[,\ i = 1, \ldots, r \\ 1 & si\ x \geq a_r, \end{cases}$$

avec $F_0 = 0$. La fonction F_X est appelée la fonction de répartition empirique de la variable X.

On rappelle que $h_i = \frac{f_i}{a_i - a_{i-1}}$ est la densité de fréquence associée à la classe $[a_{i-1}, a_i[$ définie en section 1.1.4.

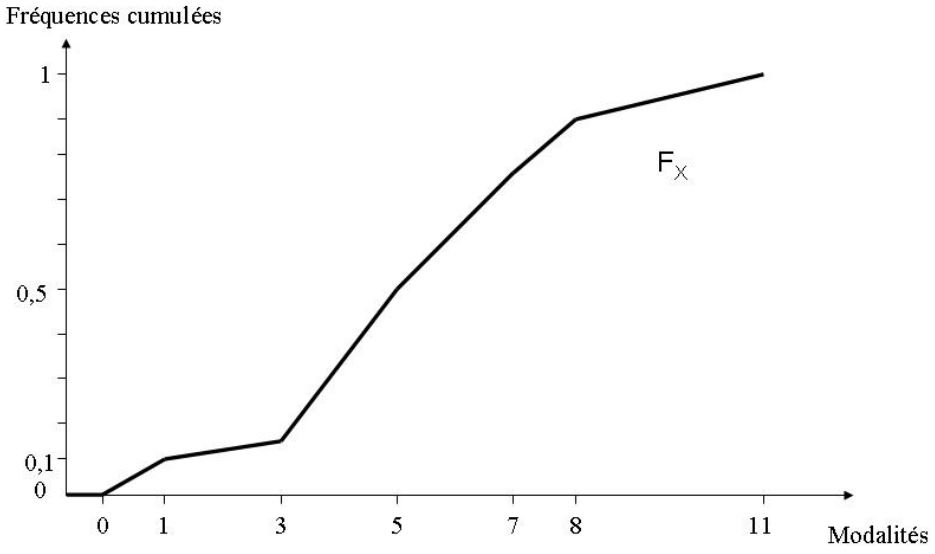
Remarque : Lorsque l'on considère une variable ayant pour modalités $x_1, \ldots, x_r$,

on choisit généralement les classes $[a_{i-1}, a_i[$ de telle sorte que $a_i = x_i$, a_0 étant une valeur arbitraire (origine des modalités).

Exemple : On donne le tableau de fréquences cumulées suivant :

Modalités	Fréquences cumulées
1	$0,1$
3	$0,15$
5	$0,5$
7	$0,75$
8	$0,9$
11	1

On a $a_1 = 1, a_2 = 3, a_3 = 5, a_4 = 7, a_5 = 8$ et $a_6 = 11$ et l'on suppose que la valeur arbitraire de a_0 choisie pour origine des modalités est $a_0 = 0$ (remarquons que ce choix dépend de la nature de la caractéristique étudiée).



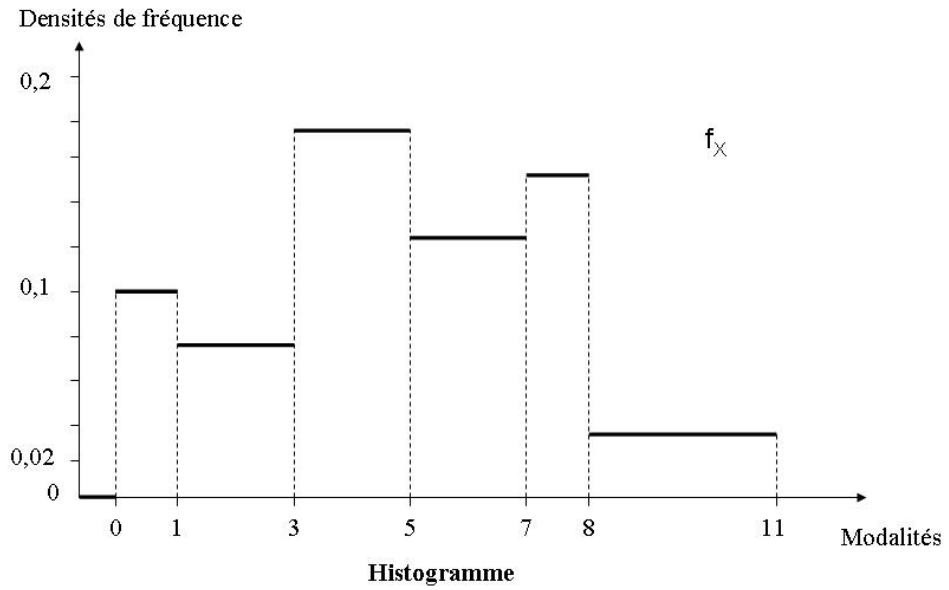
L'histogramme.

Il est clair que la fonction de répartition empirique est une fonction continue et dérivable sauf éventuellement en les points d'abscisses $a_0, a_1, \ldots, a_r$. On prolonge sa fonction dérivée de telle sorte que celle-ci est continue à droite. La fonction obtenue

est appelée la densité empirique de X et est notée f_X :

$$f_X(x) = \begin{cases} 0 & \text{si } x < a_0 \\ h_i & \text{si } x \in [a_{i-1}, a_i[,\ i = 1, \ldots, r \\ 0 & \text{si } x \geq a_r \end{cases}$$

Le graphe de la fonction empirique est appelé l'*histogramme* de X. Celui qui correspond à l'exemple ci-dessus est le suivant :



Histogramme

REMARQUE : Dans les exemples qui précèdent, les variables prennent toujours des valeurs positives. Il faut cependant noter que ceci n'est pas nécessairement les cas. Par exemple la température d'un local de stockage en degré Celcius est une variable qualitative ordinale qui peut prendre des valeurs négatives.

1.3. Paramètres caractéristiques de tendance centrale.

Ce sont des valeurs autour desquelles les données d'une série statistique brute se répartissent de manière 'équilibrée'.

La médiane. Celle-ci concerne les variables quantitatives ou bien qualitatives ordinales. C'est une valeur notée $q_{1/2}$ qui partage l'échantillon (ou la population)

en 2 groupes d'effectifs égaux : celui des individus dont la valeur de modalité est inférieure à $q_{1/2}$ et celui des individus dont la valeur de modalité est supérieure à $q_{1/2}$.

Exemple : La médiane de la série suivante $1, 5, 6, 7, 34, 50, 176$ est égale à 7.

Nous donnerons une définition plus précise de la médiane à la section suivante.

La moyenne. Celle-ci concerne uniquement les variables quantitatives. On considère une variable $X : \Omega \to E$, avec $\Omega = \{\omega_1, \ldots, \omega_n\}$ et $E = \{x_1, \ldots, x_r\}$.

DÉFINITION : *La moyenne de la variable X se note $\overline{X}$. Celle-ci est définie par :*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X(\omega_i) \, .$$

En regroupant les individus suivant leur valeur de modalité, on obtient la définition équivalente suivante :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{r} n_i x_i = \sum_{i=1}^{r} f_i x_i \, .$$

Les propriétés suivantes se vérifient facilement:

Propriétés : 1. Si la variable X est constante, c'est à dire si il existe une valeur réelle x telle que $X(\omega) = x$ pour tout $\omega \in \Omega$, alors $\overline{X}$.

2. Considérons une variable $Y : \Omega \to F$ définie sur la même population que X (dont l'ensemble de modalités F n'est pas nécessairement égal à E). Soient a et b deux nombres réels alors $aX + bY$ définit une troisième variable sur Ω dont la moyenne est :

$$\overline{aX + bY} = a\overline{X} + b\overline{Y} \, .$$

1.4. Paramètres caractéristiques de dispertion.

Ceux-ci indiquent la manière dont les données sont réparties les unes relativement aux autres.

L'étendue. C'est la différence entre la valeur minimale et la valeur maximale que peuvent prendre les modalités :

$$\max\{x_1, \ldots, x_r\} - \min\{x_1, \ldots, x_r\} \, .$$

Les quantiles. Pour un réel $\alpha \in]0, 1[$, le quantile d'ordre α est une valeur de modalité notée q_α qui sépare la population en deux parties : celle des individus dont la valeur

de modalité est inférieure à q_α et celle des individus dont la modalité est supérieure à q_α. Cette valeur n'est pas déterminée de manière unique en général. Voici une définition qui permet de remédier à ce problème.

DÉFINITION. *Le quantile d'ordre* $\alpha \in]0,1[$ *est la plus petite valeur* q_α *telle que* $F_X(q_\alpha) = \alpha$. *Plus formellement :*

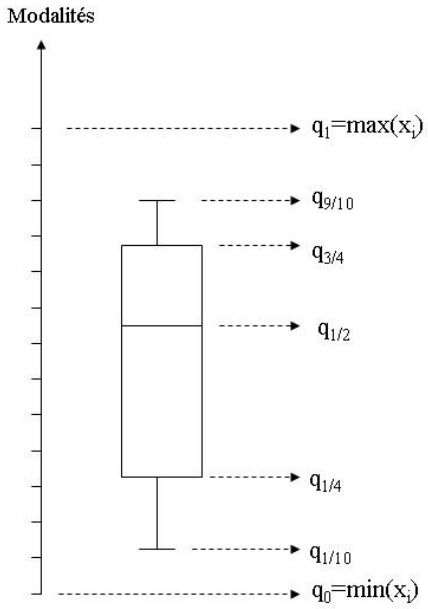
$$q_\alpha = \inf\{x : F_X(x) \geq \alpha\}\,.$$

Notons que selon cette définition, la valeur $q_{1/2}$ n'est autre que la médiane. Voici d'autres quantiles particuliers très souvent utilisés :

$q_{1/4}$ et $q_{3/4}$ sont les *quartiles*.
$q_{i/10}$ pour $i = 1, 2, \ldots, 9$ sont les *déciles*.
$q_{i/100}$, pour $i = 1, 2, \ldots, 99$ sont les *centiles*.

Remarque : La moitié de la population a une valeur de modalité comprise entre $q_{1/4}$ et $q_{3/4}$ et 80% de la population a une valeur de modalité comprise entre $q_{1/10}$ et $q_{9/10}$. Plus généralement, on voit que pour une valeur $\alpha < 1/2$ fixée, plus $q_{1-\alpha} - q_\alpha$ est petit et plus les valeurs de la série statistique sont regroupées autour de la médiane.

Cette remarque nous conduit à représenter les principaux quantiles dans un diagramme que l'on appelle la "boîte à moustaches". Celle-ci met en valeur la dispertion de la série statistique. On indique les modalités sur un axe vertical à côté duquel une "boîte" représente les quantiles : $q_0 = \min\{x_1, \ldots, x_r\}$, $q_{1/10}$, $q_{1/4}$, $q_{1/2}$, $q_{3/4}$ et $q_1 = \max\{x_1, \ldots, x_r\}$ de la manière suivante :



Boîte à moustaches

Détermination des quantiles. Pour déterminer q_α, on repère tout d'abord l'intervalle $[a_i, a_{i+1}[$ auquel appartient q_α à l'aide du tableau des fréquences cumulées ou de la courbe cumulative.

Puisque F_X est une fonction affine sur l'intervalle $[a_i, a_{i+1}[$, on a alors la règle d'interpolation linéaire suivante :

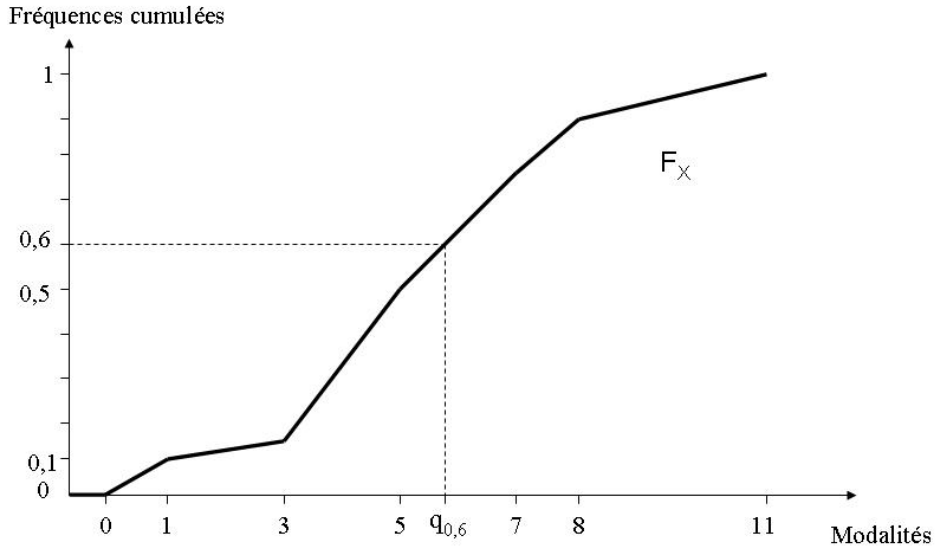
$$\frac{F_X(a_{i+1}) - F_X(a_i)}{a_{i+1} - a_i} = \frac{F_X(q_\alpha) - F_X(a_i)}{q_\alpha - a_i}\,.$$

Ainsi, avec $F_X(q_\alpha) = \alpha$ par définition de q_α, on en déduit la formule suivante :

$$q_\alpha = a_i + (a_{i+1} - a_i)\frac{\alpha - F_X(a_i)}{F_X(a_{i+1}) - F_X(a_i)}\,.$$

Exemple. On reprend l'exemple de la section 1.2.2. On détermine $q_{3/5} = q_{0,6}$ par le

tableau des fréquences cumulées ou bien sur le graphe de F_X, comme ci-dessous.



Ici, on voit que $q_{3/5} \in [5, 7[$. On applique alors la formule ci-dessus avec $a_i = 5$, $a_{i+1} = 7$, $F_X(a_i) = 0,5$, $F_X(a_{i+1}) = 0,75$ et $F_X(q_{0,6}) = 0,6$. Soit

$$q_{0,6} = 5 + 2 \times \frac{0,6 - 0,5}{0,75 - 0,5} = 5,8\,.$$

La variance. On considère une variable quantitative $X : \Omega \to E$, avec $\Omega = \{\omega_1, \ldots, \omega_n\}$ et $E = \{x_1, \ldots, x_r\}$.

DÉFINITION : *La variance de la variable X se note $Var(X)$. Celle-ci est définie par :*

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (X(\omega_i) - \overline{X})^2\,.$$

En regroupant les individus suivant leur valeur de modalité, on obtient la définition équivalente suivante :

$$Var(X) = \frac{1}{n} \sum_{i=1}^{r} n_i (x_i - \overline{X})^2 = \sum_{i=1}^{r} f_i (x_i - \overline{X})^2\,.$$

Il existe une troisième manière d'exprimer la variance qui est en fait la forme la

plus utilisée :

$$Var(X) = \left(\frac{1}{n} \sum_{i=1}^{r} n_i x_i^2 \right) - \overline{X}^2 .$$

Voici quelques propriétés élémentaires de la variance :

La variance d'une variable est toujours positive ou nulle. Celle-ci est nulle si et seulement si la variable est constante. Si a et b sont deux paramètres réels, alors la variance ce la variable $aX + b$ est

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X) .$$

L'écart type. Celui-ci se note $\sigma(X)$ et est défini par

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} .$$

L'écart type représente l'écart moyen (en moyenne quadratique) entre les valeurs de la variable X et sa moyenne. On déduit des propriétés de la variance qu'il est toujours positif ou nul et que pour deux valeurs réelles quelconques a et b, on a

$$\sigma(aX + b) = |a| \sigma(X) .$$

Exemple : La moyenne, la variance et l'écart type de la série statistique brute suivante : 3, 5, 8, 3, 16, 5, 5, 1, 0, 10, 10 sont respectivement

$$\overline{X} \simeq 5,82 , \quad \mathrm{Var}(X) \simeq 19,85 , \quad \sigma(X) \simeq 4,46 .$$

Chapitre 2

Étude conjointe de deux variables.

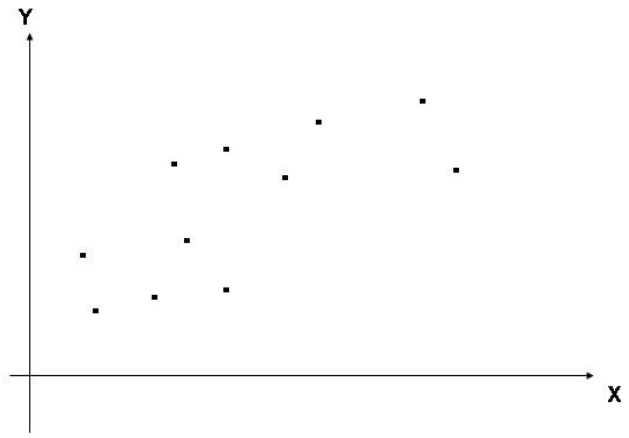
2.1. Définitions et notations.

Lorsque l'on doit observer plusieurs caractères sur la même population, l'étude séparée de chacun d'eux telle qu'elle est présentée au chapitre précédent ne fournit aucune indication sur leurs relations éventuelles. Le problème consiste donc à étudier simultanément les caractères sans perdre l'information conjointe due au fait qu'ils ont été observés sur les mêmes individus.

On se limitera dans ce chapitre au cas où deux caractères sont observés donnant lieu à deux variables X et Y définies sur une même population $\Omega = \{\omega_1, \ldots, \omega_n\}$.

Exemple : X = poids d'un colis et Y = volume de ce colis.

Dans le cas où ces deux variables sont qualitatives ordinales ou bien quantitatives, on peut représenter les données $\{X(\omega_1), X(\omega_2), \ldots, X(\omega_n)\}$ comme des points du plan des coordonnées $(X(\omega_i), Y(\omega_i))$, $i = 1, 2 \ldots, n)$. Le graphe ainsi obtenu est appelé un *nuage de points* :



Suppposons que les deux variables X et Y soient à valeurs dans E et F (parties finies de $\mathbb{R}$), avec $\text{Card}(E) = r$ et $\text{Card}(F) = p$. Les données recueuillies et les effectifs associés se représentent dans un *tableau de contingences* que nous allons définir maintenant.

L'effectif conjoint de la modalité x_i et de la modalité x_j est défini par :

$$n_{ij} = \text{Card}\{\omega \in \Omega : X(\omega) = x_i,\, Y(\omega) = y_j\}\,.$$

L'effectif des individus présentant la modalité x_i est noté $n_{i\cdot}$. Celui-ci est égal à

$$n_{i\cdot} = \sum_{j=1}^{p} n_{ij}$$

et l'effectif des individus présentant la modalité y_j est noté $n_{\cdot j}$. Celui-ci est égal à

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

Tableau de contingences :

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_p	
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\dots$	n_{rj}	$\dots$	n_{rp}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	n

Remarquons les relations suivantes :

$$n = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{p} n_{\cdot j} = \sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij} = \sum_{j=1}^{p}\sum_{i=1}^{r} n_{ij}\,.$$

De même, on établit le *tableau des fréquences conjointes* en définissant

$$f_{ij} = \frac{n_{ij}}{n}, \quad f_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{et} \quad f_{\cdot j} = \frac{n_{\cdot j}}{n}\,.$$

Tableau des fréquences conjointes :

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_p	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1p}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{ip}	$f_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	f_{r1}	$\dots$	f_{rj}	$\dots$	f_{rp}	$f_{r\cdot}$
	$f_{\cdot 1}$	$\dots$	$f_{\cdot j}$	$\dots$	$f_{\cdot p}$	1

On vérifie alors les relations suivantes :

$$\sum_{i=1}^{r} f_{i\cdot} = \sum_{j=1}^{p} f_{\cdot j} = \sum_{i=1}^{r}\sum_{j=1}^{p} f_{ij} = \sum_{j=1}^{p}\sum_{i=1}^{r} f_{ij} = 1\,.$$

2.2. Covariance.

La covariance et le coefficient de corrélation linéaire permettent de mesurer la dépendance entre les variables X et Y.

DÉFINITION : *La covariance entre les variables X et Y se note $Cov(X,Y)$. Celle-ci est définie par :*

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X(\omega_i) - \overline{X})(Y(\omega_i) - \overline{Y})\,.$$

En regroupant les individus suivant leur valeur de modalité, on obtient la définition équivalente suivante :

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y}) = \sum_{i=1}^{r} f_{ij}(x_i - \overline{X})(y_j - \overline{Y})\,.$$

En développant les expressions sous le signe somme, on obtient après simplification :

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij}x_i y_j\right) - \overline{X}\,\overline{Y}\,.$$

Propriétés :
1. Pour toute variable X, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
2. Pour toutes variables X et Y, $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
3. Pour toutes variables X, Y et Z et pour tous réels α et β,

$$\mathrm{Cov}(\alpha X + \beta Y, Z) = \alpha\mathrm{Cov}(X,Z) + \beta\mathrm{Cov}(Y,Z)\,.$$

4. Pour toutes variables X et Y,

$$|\mathrm{Cov}(X,Y)| \leq \sigma(X)\sigma(Y)\,.$$

DÉFINITION : *Lorsque la variance des variables X et Y n'est pas nulle (c'est à dire lorsque celles-ci ne sont pas constantes) on définit le coefficient de corrélation (linéaire) entre les variables X et Y par :*

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}\,.$$

D'après la propriété 4. ci-dessus, on a toujours :

$$-1 \leq r(X,Y) \leq 1\,.$$

On voit facilement que si il existe une relation de linéarité entre X et Y, c'est à dire si il existe deux réels a et b tels que $Y = aX + b$, alors $|r(X,Y)| = 1$. Plus généralement, plus le nuage de points (X,Y) est "aplati" autour d'une droite et plus $r(X,Y)$ est proche de 1 en valeur absolue. Par conséquent, si le coefficent de corrélation linéaire est proche de 0 alors il n'y a pas de relation de linéarité entre X et Y. Toutefois, même si un coefficient de corrélation linéaire est proche de 0, il peut y avoir une forte dépendance (non linéaire) entre X et Y. C'est le cas par exemple pour les variables :

X	-3	-2	-1	0	1	2	3
Y	3	2	1	0	1	2	3

Lorsque $r(X,Y)$ est proche de 1, on dira que les variables X et Y sont *fortement corrélées*. Remarquons que cela ne signifie pas pour autant qu'il y a une relation directe de cause à effet entre X et Y.

2.3. Régression linéaire.

Lorsque l'hypothèse d'une relation de linéarité entre deux caractères sur une même population est plausible, il convient déterminer l'équation de la *droite de regression linéaire* entre les deux variables correspondantes. Cette droite est souvent aussi appelée la *droite des moindres carrés*.

DÉFINITION : *Supposons que la v.a. X ne soit pas constante (de telle sorte que $Var(X) \neq 0$). On appelle droite de régression linéraire de Y sur X, la droite d'équation $y = \hat{a}x + \hat{b}$, avec*

$$\hat{a} = \frac{Cov(X,Y)}{Var(X)} \quad et \quad \hat{b} = \overline{Y} - \hat{a}\overline{X}\,.$$

Au sens mathématique, il s'agit de la droite d'équation $y = f(x)$ qui minimise l'erreur quadratique :

$$\frac{1}{n}\sum_{i=1}^{n}(Y(\omega_i) - f(X(\omega_i)))^2\,,$$

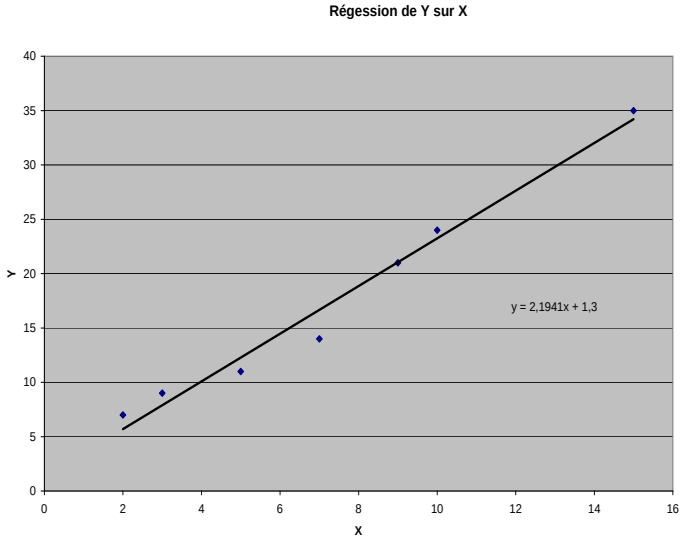
où $(X(\omega_i), Y(\omega_i))$, $i = 1, \ldots, n$ est la suite des points du plan correspondant caractères observés. Plus les variables X et Y sont fortement corrélées et plus le nuage

de points (X, Y) se resserre autour de la doite de régression. En fait la droite de régression linéaire de Y sur X est celle qui, minimise les distances verticales des points à la droite dans le nuage de points (X, Y), (alors que a droite de régression linéaire de X sur Y est celle qui, minimise les distances horizontales). Il y a donc dissymétrie entre les deux régressions.

EXEMPLE : Pour les données suivantes :

X	5	2	7	9	15	3	10
Y	11	7	14	21	35	9	24

Le coefficient de corrélation linéaire est $r(X, Y) = 0,9891$ et les coefficients de la droite de régression linéaire de Y sur X valent $\hat{a} = 2,19$ et $\hat{b} = 1,3$.



Il peut exister un lien assez "explicite" entre les variables X et Y sans que celui ci ne soit linaire. Autrement dit, le nuage de points (X, Y) peut se trouver "proche" du graphe d'une fonction explicite f, qu'elle soit polynomiale, logarithmique, exponentielle,... Dans la pratique, on ne connaît pas la fonction f. Au vu du nuage de point, on peut faire une hypothèse concernant la nature de la fonction.

Une fois qu'une hypothèse a été formulée, de manière à ajuster au mieux les paramètres de la fonction que l'on veut mettre en évidence, on a fait une régression linéaire de $T = g(Y)$ sur $Z = k(X)$, où g et k sont des fonctions bien choisies qui dépendent de la fonction inverse de f lorsqu'elle existe. Nous ne considérerons que les fonctions suivantes :

Exponentielle	$f(x) = ae^{bx}$	$g(y) = \ln y$	$k(x) = x$
Puissance	$f(x) = ax^b$	$g(y) = \ln y$	$k(x) = \ln x$
Inverse	$f(x) = a + \frac{b}{x}$	$g(y) = y$	$k(x) = \frac{1}{x}$
Logistique	$F(x) = \frac{1}{1+e^{-(ax+b)}}$	$g(y) = \ln\left(\frac{y}{y-1}\right)$	$k(x) = x$.

Lorsque la représentation graphique d'un nuage de points (P_n, P_{n+1}) laisse montrer que la suite numérique (P_n) est définie par une équation de récurrence du type $P_{n+1} = F(P_n)$. Pour déterminer la fonction F, on a recours à la méthode de régression linéaire. On considère alors comme variable X la suite $X = (P_0, P_1, \ldots, P_{n-1})$ et comme variable Y la suite $Y = (P_1, P_2, \ldots, P_n)$.

Chapitre 3

Prévision.

3.1. Introduction.

Une prévision est l'interprétation dans le futur d'une série d'observations effectuées à des dates fixes. Ces observations correspondent aux enregistrements des quantités de consommation ou de commandes de certains produits et sont généralement exprimées en effectifs ou en unités de mesures quelconques. On les appelle séries *temporelles* ou *chronolgiques*.

En gestion de production les prévisions sont utiles pour :

- La gestion des stocks afin de savoir quand et de combien approvisionner ?
- Le calcul des besoins externes, afin d'établir des règles de production.
- L'évaluation des charges des différents postes de travail au sein de l'entreprise.

Les données de consommations sont généralement enregistrées sur des intervalles de temps réguliers, semaines, mois, années,... que l'on appelle des *périodes*. Une fois enregistrée, une série chronologique est représentée sous forme de graphique. On détermine alors à quel type de tendance celle-ci obéit afin de déterminer la méthode de prévision à appliquer.

Les méthodes les plus courantes sont :

1. Méthodes des *moyennes mobiles*, cf. section 3.2. Un certain nombre n de périodes étant fixé, les prévisions correspondent à la moyenne des n périodes antérieures.

2. Méthode de *lissage exponentiel*, cf. section 3.3. C'est une méthode qui prend en compte la prévision de la période antérieure. La prévision pour la période n est une moyenne pondérée de la prévision et de la valeur réelle de consommation à la période $n-1$.

3. Méthode de la *droite de régression linéaire*, cf. section 4.2. Dans ce cas la courbe des prévisions est la droite de régression linéaire des consommations réelles (variable Y) sur le temps (variable X), cf. Chapitre 2.

Cette méthode peut être rendue plus précise en utilisant les coefficients de saisonnalité afin de modifier la pente de la droite de régression sur chaque saison.

3.2. Méthode des moyennes mobiles.

Son principe est simple : un nombre n impair de périodes étant choisi, on remplace chaque valeur par la moyenne de cette valeur, des $(n-1)/2$ valeurs de consommation qui la précèdent et des $(n-1)/2$ valeurs de consommation qui lui succèdent :

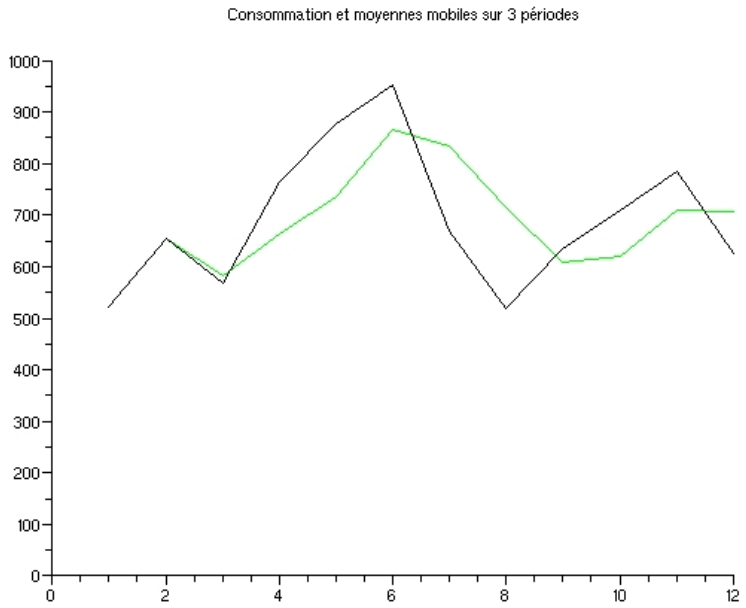
$$P_k = \frac{1}{n} \sum_{i=k-(n-1)/2}^{k+(n-1)/2} D_i \,.$$

Son avantage est qu'elle atténue les fluctuations brutales des consommations tout en préservant la tendance générale de la courbe. Toutefois lorsque n est trop grand, cette méthode a l'inconvénient de cacher les éventuels changements de tendance survenus au cours du temps.

Exemple : Une scierie a enregistré sa consommation en m^3 de bois sur les 12 mois de l'année et a obtenu la série chronologique suivante.

Mois	Consommation (en m^3)	Moyennes mobiles (sur 3 mois)
1	521	
2	654	581
3	567	662
4	765	737
5	878	866
6	954	834
7	670	715
8	520	608
9	634	621
10	710	710
11	786	707
12	625	

Il est préférable de représenter ces deux séries de données sur un même graphique.



3.3. Méthode de lissage exponentiel.

Notons D_n la valeur réelle de la consommation enregistrée pour la période n et P_n la valeur de la prévision pour cette période.

La valeur de P_n est égale à la valeur P_{n-1} à laquelle on ajoute "une partie" de l'écart $D_{n-1} - P_{n-1}$. Cet partie dépend d'un coefficient multiplicateur α compris entre 0 et 1, appelé *coefficient de lissage* que l'on choisit à l'avance.

$$P_n = P_{n-1} + \alpha(D_{n-1} - P_{n-1})\,.$$

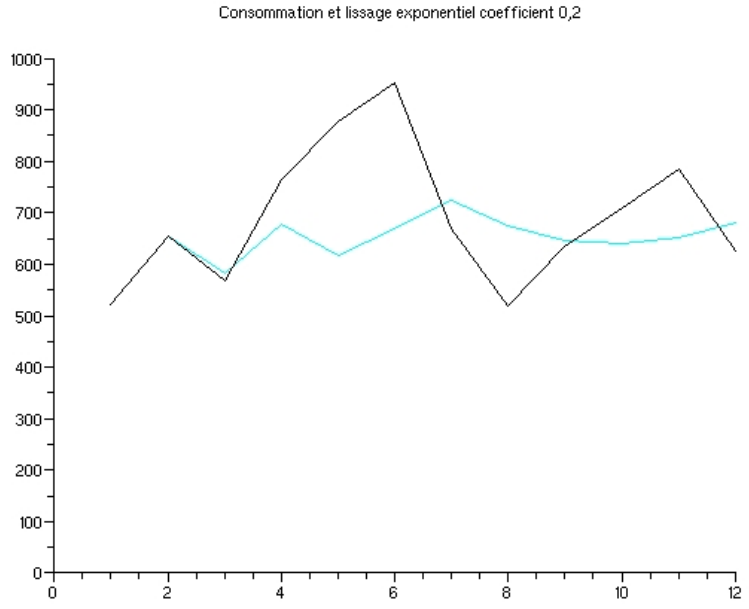
On peut aussi définir cette prévision en disant que P_n est la moyenne pondérée de D_{n-1} et de P_{n-1} en affectant à la première valeur le poids α et à la seconde, le poids $1 - \alpha$. On a alors l'expression équivalente :

$$P_n = \alpha D_{n-1} + (1 - \alpha)P_{n-1}\,.$$

On convient que la première valeur d'estimation P_1 est toujours une moyenne mobile. (Toute autre valeur pourrait être choisie pourvue qu'elle estime la valeur réelle avec une faible erreur). Sur l'exemple de la section 3.2, la période 1 correpondra au troisième mois avec $P_1 = 581$. Choisissons par exemple $\alpha = 0,2$. On a ensuite $P_2 = \alpha D_1 + (1-\alpha)P_1 = 0,2 \times 567 + 0,8 \times 581 = 578,2$, puis $P_3 = \alpha D_2 + (1-\alpha)P_2 = 0,2 \times 765 + 0,8 \times 578,2 = 615,56$. Les valeurs arrondies à l'unité sont alors :

Mois	Consommation (en m^3)	Lissage exponentiel (coefficient 0,2)
1	521	
2	654	581
3	567	596
4	765	590
5	878	625
6	954	676
7	670	732
8	520	720
9	634	680
10	710	671
11	786	679
12	625	700

Les données effectives et les prévisions par lissage exponentiel sont représentées dans le graphique ci-dessous. Ces courbes montrent que le lissage exponentiel peut parfois atténuer très fortement les fluctuations observées dans la réalité.



Plus le coefficient de lissage est grand et plus l'on tient compte des estimations récentes. Plus précisément, l'influence d'un résultat antérieur sur le calcul de la prévision décroît exponentiellement au cours du temps. La méthode tient son nom de cette propriété. Le choix du coefficient de lissage est arbitraire. Dans la pratique,

on retient celui qui minimise l'erreur de prévision.

3.4. Méthodes de régression linéaire.

3.4.1 La droite de régression

Cette méthode est à utiliser plutôt dans les cas où l'évolution des consommations est dite *à tendance*, c'est à dire lorsque l'allure de la courbe des consommations effectives au cours du temps est croissante ou bien décroissante.

La courbe des prévisions est la droite de régression des consommations sur le temps. On se reportera pour ceci au chapitre précédent. Notons P_n la prévision pour la période n, on cherche les coefficients a et b tel que la droite d'équation

$$T_n = an + b$$

soit la plus proche du nuage de points constitué par les valeurs effectives au sens des moindre carrés. Notons D_n la valeurs de consommation réelle sur la période n. Si les données ont été relevées sur N périodes alors le calcul fait au chapitre 2 donne :

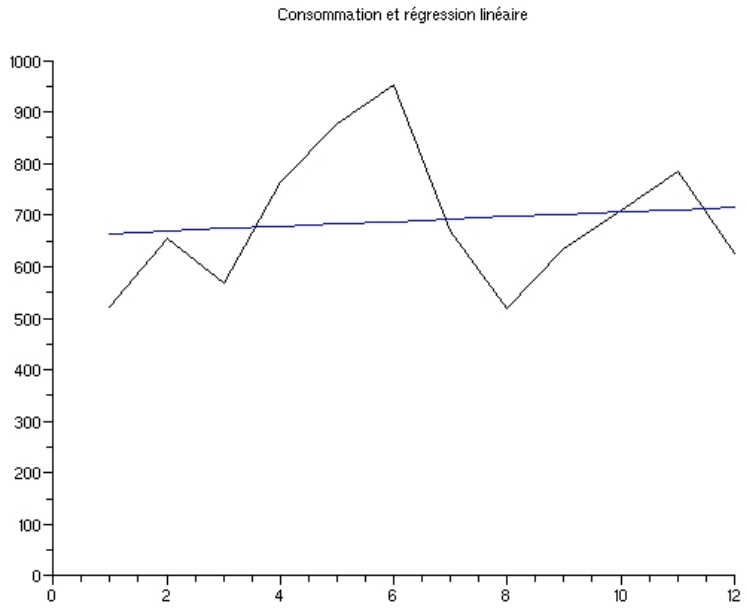
$$a = \frac{N\sum_{n=1}^{N} nD_n - \sum_{n=1}^{N} n \sum_{n=1}^{N} D_n}{N\sum_{n=1}^{N} n^2 - \left(\sum_{n=1}^{N} n\right)^2}\,, \quad b = \frac{\sum_{n=1}^{N} D_n}{N} - a\frac{\sum_{n=1}^{N} n}{N}\,.$$

Pour simplifier cette formule, on peut utiliser les relations suivante :

$$\sum_{n=1}^{N} n = \frac{N(N+1)}{2} \quad \text{et} \quad \sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}\,.$$

Sur l'exemple de la section 3.2 on obtient

$$T_n = 4,614n + 660,33\,.$$



La prévision pour la période n se fait soit en lisant la valeur de T_n sur le graphe de la droite de régression, soit par un calcul direct en utilisant la formule donnée plus haut pour l'expression de T_n.

Nous avons présenté ici le cas d'une croissance linéaire. Si l'on note plutôt une croissance exponentielle ou logarithmique, on effectuera alors la régression qui correspond comme au chapitre précédent.

3.4.1 Prise en compte des coefficients cycliques

La méthode de régression linéaire peut être rendue plus précise si en plus de la tendance croissante ou décroissante observée, on remarque des variations cycliques. On peut alors prendre en compte ces cycles en définissant pour chacun d'eux un coefficient appelé *coefficient cyclique* par lequel on multiplie les valeurs prévues par la régression linéaire sur le cycle considéré. Ces cycles sont aussi parfois appelés des *saisons* et l'on parle alors de *coefficient de saisonnalité*.

Nous noterons C_i le coefficient cyclique qui se rapporte au i-ième cycle. Celui-ci est calculé de la manière suivante :

$$C_i = \frac{\text{Consommation moyenne sur la durée du cycle } i}{\text{Consommation moyenne de toute la série chronologique}}.$$

Plus précisément si sur N données on observe k cylces, avec $N = kn$, où n est la

longueur des cycles, alors

$$C_i = k \sum_{j=(i-1)n+1}^{in} D_j / \sum_{j=1}^{N} D_j \,, \quad i = 1, \ldots, k \,.$$

Dans l'exemple de la section 3.2, on peut considérer qu'il existe des cycles trimestriels. Le tableau suivant représente le calcul des coefficients cycliques arrondies au centième près :

Moyenne annuelle	$690,33$			
Moyenne trimestrielle	581	$865,67$	608	707
Coefficient de saisonnalité	$0,84$	$1,25$	$0,88$	1

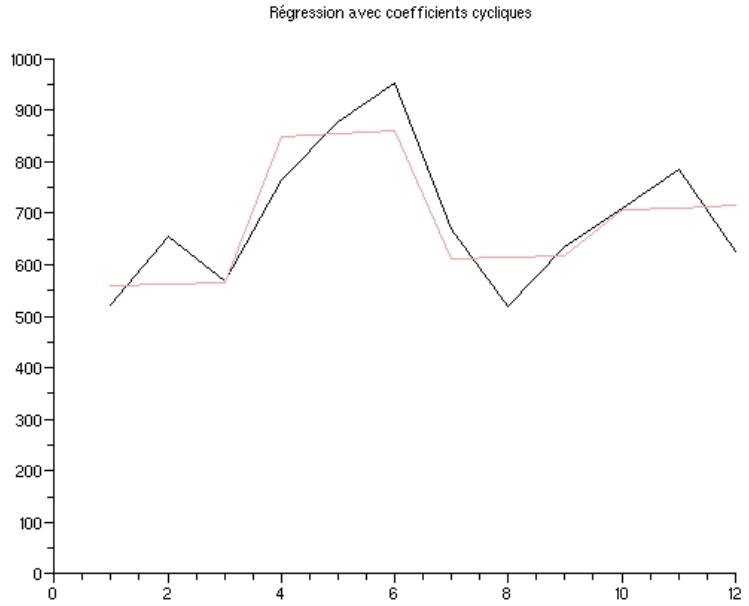
Les prévisions données par la régression linéaire des consommations corrigées par les coefficients de saisonnalité définissent la prévisions

$$P_n = T_n C_n \,, \quad n = 1, 2, \ldots, N \,,$$

où les prévisions T_n sont définies à la section précédante. On en déduit le tableau de prévisions suivant :

Mois	Consommation	$P_n = T_n C_n$
1	521	559
2	654	562
3	567	566
4	765	848
5	878	854
6	954	860
7	670	610
8	520	614
9	634	618
10	710	706
11	786	711
12	625	716

La courbe des prévisions épouse alors les contours des cycles apparents sur la courbe des valeurs réelles.


Régression avec coefficients cycliques
1000
900
800
700
600
500
400
300
200
100
0
0
2
4
6
8
10
12

Chapitre 4

Intervalles de confiance

4.1. Quelques notions de probabilités.

Une probabilité $\mathbb{P}$ sur un espace Ω est une fonction d'ensemble

$$\begin{aligned} \mathbb{P} : \Omega &\to [0,1] \\ A &\mapsto \mathbb{P}(A) \end{aligned}$$

telle que pour une certaine famille $\mathcal{F}$ de sous ensembles de Ω (appelée tribu) on ait

- pour tout $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$,
- pour tous $A, B \in \mathcal{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- pour tout $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
- $\mathbb{P}(\emptyset) = 0$ et $\mathbb{P}(\Omega) = 1$.

Pour une suite d'événements $A_1, \ldots, A_n$ de $\mathcal{F}$, la probabilité de l'intersection $A_1 \cap A_2 \cap \cdots \cap A_n$ sera notée

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1, A_2, \ldots, A_n)\,.$$

Une variable aléatoire X définie sur Ω est une application

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

telle que pour tous $a, b \in \mathbb{R}$, l'ensemble $\{\omega : a < X(\omega) < b\}$ appartient à la tribu $\mathcal{F}$. Dans la suite, pour tout sous ensemble A de $\mathbb{R}$, l'ensemble $\{\omega \in \Omega : X(\omega) \in A\}$ sera noté simplement $\{X \in A\}$. De même on notera $\{\omega \in \Omega : X(\omega) = x\} = \{X = x\}$, ou encore $\{\omega : a < X(\omega) < b\} = \{a < X < b\}$.

4.1.1 Les lois discrètes : On dit qu'une variable aléatoire X est discrète si il existe un sous ensemble au plus dénombrable N de $\mathbb{R}$ tel que l'ensemble $\{X \in N\}$ soit de probabilité égale à 1. Ceci revient à dire que l'ensemble des valeurs que X peut prendre est au plus dénombrable. Généralement il s'agira d'un sous ensemble de $\mathbb{Z}$. Dans ce cas la loi de la variable aléatoire X est donnée par l'ensemble des valeurs

$$\mathbb{P}(X = x)\,, \quad x \in N\,.$$

Des v.a. $X_1, \ldots, X_n$ à valeurs dans l'ensemble N sont dites indépendantes si pour tous les éléments $x_1, \ldots, x_n$ de N on a

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) \ldots \mathbb{P}(X_n = x_n)\,.$$

(Remarquons que selon la notation adoptée ci-dessus, le membre de gauche de l'égalité ci-dessus n'est autre que : $\mathbb{P}(\{X_1 = x_1\} \cap \{X_2 = y\} \dots \{X_n = x_n\})$).

L'*espérance* (aussi appelée la *moyenne*) d'une v.a. discrète est la quantité :

$$\mathbb{E}(X) = \sum_{x \in N} x \mathbb{P}(X = x)\,,$$

lorsque cette somme est définie (en particulier lorsque $\sum_{x \in N} |x| \mathbb{P}(X = x)$). C'est la moyenne des valeurs que peut prendre X pondérées par les probabilités correspondantes. L'espérance est un opérateur linéaire sur l'ensemble des variables aléatoires. C'est à dire que pour toutes v.a. $X_1, \dots, X_n$ et tous réels $a_1, \dots, a_n$ l'espérance de la v.a. $a_1X_1 + a_2X_2 + \cdots + a_nX_n$ s'exprime en fonction des espérances des v.a. $X_1, \dots, X_n$ de la manière suivante :

$$\mathbb{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \cdots + a_n\mathbb{E}(X_n)\,. \quad (4.1.1)$$

Lorsque la somme $\mathbb{E}(X^2) = \sum_{x \in N} x^2 \mathbb{P}(X = x)$ est finie, on définit aussi la *variance* d'une v.a. par :

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2\,.$$

Lorsque les v.a. $X_1, \dots, X_n$, la variance de la somme $X_1 + \cdots + X_n$ est égale à la somme des variances des v.a. $X_1, \dots, X_n$:

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\,.$$

Exemples :

1. La loi de Bernoulli.
C'est la loi d'une v.a. à valeurs dans l'ensemble $N = \{0, 1\}$. Celle-ci est donnée par

$$\mathbb{P}(X = 0) = 1 - p \text{ et } \mathbb{P}(X = 1) = p$$

pour une certaine valeur $p \in]0, 1[$ appelée le paramètre de X. Cette loi est aussi appelée la loi du pile ou face. Il est facile de vérifier que l'espérance et la variance de la loi de Bernoulli de paramètre p sont données respectivement par

$$\mathbb{E}(X) = p \quad \text{et} \quad \mathrm{Var}(X) = p(1 - p)\,.$$

2. La loi binomiale.
Une v.a. X suit la loi binomiale de paramètres n (entier) et $p \in]0, 1[$ si

$$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}\,, \quad \text{pour tous } k = 0, 1, \dots, n,$$

où par définition $C_n^k = \frac{n!}{k!(n-k)!}$. Cette loi est aussi appelée la loi du tirage avec remise. En effet, supposons qu'un urne contienne K boules dont M sont rouges et

$N - M$ sont bleues. On tire au hasard avec remise $n < M$ boules de cette urne et on note X le nombre de boules rouges obtenues. On peut alors vérifier que

$$\mathbb{P}(X = k) = C_n^k \left(\frac{M}{K}\right)^k \left(1 - \frac{M}{K}\right)^{n-k}, \quad \text{pour tous } k = 0, 1, \ldots, n.$$

En posant $p = M/K$, on voit que X suit une loi binomiale de paramètres n et p.

3. La loi géométrique.
Une v.a. X suit une loi géométrique de paramètre $a \in]0, 1[$ si elle est à valeurs dans $\mathbb{N}$ et si

$$\mathbb{P}(X = k) = a(1 - a)^k, \quad \text{pour tout } k \in \mathbb{N}.$$

4. La loi de Poisson.
Une v.a. X suit une loi de Poisson de paramètre $\lambda > 0$ si elle est à valeurs dans $\mathbb{N}$ et si

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{pour tout } k \in \mathbb{N}.$$

Propriété : La somme de n v.a. de Bernoulli indépendantes de paramètre p suit une loi binomiale de paramètres n et p. Ainsi on déduit de la propriété de linéarité de l'espérance et du fait que la variance de la somme de n v.a. indépendantes est égale à la somme des variances que si X suit une loi binomiale de paramètres n et p alors

$$\mathbb{E}(X) = np \quad \text{et} \quad \mathrm{Var}(X) = np(1 - p)\,.$$

4.1.2 Les lois continues : Une v.a. X est dite continue si il existe une fonction positive ou nulle f appelée *densité* de X et telle que pour tous $a, b \in \mathbb{R}$, $a \leq b$ on a

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\, dx\,.$$

L'égalité ci-dessus pour tous $a, b \in \mathbb{R}$ définit la loi de X.

Lorsque l'intégrale $\int_{-\infty}^{\infty} |x| f(x)\, dx$ est finie, l'espérance d'une v.a. continue de densité f est définie de la manière suivante :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x)\, dx\,.$$

Lorsque l'intégrale $\int_{-\infty}^{\infty} x^2 f(x)\, dx$ est finie, la variance d'une v.a. continue de densité f est définie de la manière suivante :

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2\,.$$

Notons que la variance mesure la différence moyenne entre une variable aléatoire et son espérance.

Exemples :

Pour tout sous ensemble A de $\mathbb{R}$ nous noterons par $\mathbb{1}_A$ la fonction

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon.} \end{cases}$$

1. La loi uniforme sur $[0,1]$.
La densité d'une v.a. X de loi uniforme est donnée par

$$\mathbb{1}_{[0,1]}(x)\,.$$

On peut facilement vérifier que l'espérance et la variance de X sont respectivement données par

$$\mathbb{E}(X) = \frac{1}{2} \quad \text{et} \quad \mathrm{Var}(X) = \frac{1}{12}\,.$$

2. La loi gaussienne.
Une v.a. de loi gaussienne de paramètres $m \in \mathbb{R}$ et $\sigma > 0$ a pour densité la fonction :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}\,, \quad x \in \mathbb{R}\,.$$

L'espérance et la variance de X sont respectivement données par

$$\mathbb{E}(X) = m \quad \text{et} \quad \mathrm{Var}(X) = \sigma^2\,.$$

3. La loi exponentielle.
Une v.a. de loi exponentielle de paramètre $\lambda > 0$ a pour densité :

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty[}\,.$$

On a alors

$$\mathbb{E}(X) = \frac{1}{\lambda} \quad \text{et} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}\,.$$

4.1.3 Inégalités utiles :

Pour toute variable aléatoire positive X (discrète ou continue), on a l'*inégalité de Markov*:

$$\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a\,, \quad \text{pour tout } a > 0\,.$$

Pour tout variable aléatoire X telle que $\mathbb{E}(X^2) < \infty$, on a l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \mathrm{Var}(X)/a^2\,, \quad \text{pour tout } a > 0\,.$$

4.2. Intervalles de confiance.

4.2.1 Le cas binomial :

Considérons une machine industrielle qui produit des pièces dont une certaine proportion $p \in]0,1[$ sont défectueuses. Si l'on teste 10^6 pièces et que l'on compte 180 pièces défectueuses, on dira que p est voisin de $0,00018$ mais que sait-on de la marge d'erreur qui correspond à cette estimation ? Plus précisément, pour un certain *seuil de confiance* $\alpha \in]0,1[$, on cherche un *intervalle de confiance* I tel que la probabilité pour que la proportion exacte p de pièces défectueuses produites par la machine appartienne à l'intervalle I avec une probabilité supérieure ou égale à $1-p$. La valeur $1-\alpha$ sera alors appelée le *niveau de confiance*.

Les résultats des 10^6 tests constituent un *échantillon statistique* que l'on modélise par une suite de variables aléatoires indépendantes et de même loi de Bernoulli de paramètre p, $\{X_n,\, 0 \leq n \leq 10^6\}$. La v.a. X_n vaut 1 si la n-ième pièce testée est conforme et 0 sinon. Formellement, on cherche un intervalle I dont les bornes dépendent de la suite $\{X_n,\, 0 \leq n \leq 10^6\}$ et tel que

$$\mathbb{P}(p \in I) \geq 1-\alpha\,.$$

Pour cela, définissons la moyenne empirique $\overline{X}_n = (X_1 + \cdots + X_n)/n$ et appliquons l'inégalité de Bienaymé-Thebychev : pour tout $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{p-p^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}\,,$$

où la dernière égalité vient de ce que $p - p^2 \leq 1/4$, pour tout $p \in [0,1]$. En posant $\alpha = 1/4n\varepsilon^2$, on peut encore écrire cette inégalité de la manière suivante :

$$\mathbb{P}\left(\overline{X}_n - \frac{1}{2\sqrt{n\alpha}} \leq p \leq \overline{X}_n - \frac{1}{2\sqrt{n\alpha}}\right) \geq 1-\alpha\,.$$

Ainsi on peut énoncer :

Un intervalle de confiance au niveau $1-\alpha$ (i.e. au seuil α) pour le paramètre p est

$$I = [\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}]\,.$$

Il apparaît clairement dans cette expression que plus le seuil de confiance diminue et plus l'intervalle de confiance I devient grand. Dans l'exemple ci-dessus, si l'on prend $\alpha = 0,3$, i.e. $1-\alpha = 0,7$, on a avec $n = 10^6$ et $\overline{X}_n = 0,00018$,

$$I \simeq [0,0009\,;0,00027]\,.$$

Par conséquent, on voit que ce test n'est pas très sophistiqué. On lui préfère le cas gaussien.

4.2.2 Le cas gaussien :

Dans ce cas, on cherche à tester la la moyenne $m \in \mathbb{R}$ de la loi d'un échantillon gaussien $X_1, X_2, \ldots, X_n$. Plus précisément, on suppose que $X_1, X_2, \ldots, X_n$ sont n variables aléatoires indépendantes et de même loi gaussienne $\mathcal{N}(m, \sigma^2)$. On suppose la variance σ^2 connue. Dans le cas où celle-ci n'est pas connue, on en calcule une valeur approchée à l'aide de la variance empirique :

$$\sigma^2 \simeq \sigma^2_{\text{emp}} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) - \overline{X}_n^2\,.$$

Le test gaussien repose sur le fait que la variable aléatoire

$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m)$$

suit une loi gaussienne centrée et réduite, i.e. $\mathcal{N}(0,1)$.

De même que dans le cas binomial décrit plus haut, on cherche un intervalle de confiance au seuil $\alpha \in]0,1[$ (au niveau $1-\alpha$) du paramètre m. Puisque $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m)$ suit une loi $\mathcal{N}(0,1)$, nous avons, pour tout réel $a \geq 0$,

$$\mathbb{P}\left(-a \leq \frac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \leq a\right) = \int_{-a}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx\,.$$

Soit en utilisant la propriété de symétrie de la loi gaussienne :

$$\mathbb{P}\left(\overline{X}_n - \frac{\sigma a}{\sqrt{n}} \leq m \leq \overline{X}_n + \frac{\sigma a}{\sqrt{n}}\right) = 2\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx - 1\,.$$

Dans la pratique, pour déterminer la valeur $a \geq 0$ de l'intervalle de confiance

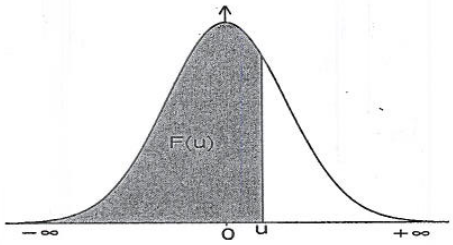
$$I = [\overline{X}_n - \frac{\sigma a}{\sqrt{n}}\,;\overline{X}_n + \frac{\sigma a}{\sqrt{n}}]$$

au niveau $1-\alpha = 2\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx - 1$, on détermine dans la table des quantiles de la loi gaussienne ci-dessous la plus petite valeur de a telle que

$$\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \geq (2-\alpha)/2\,.$$

Par exemple pour $\alpha = 0,2$, on obitent $a = 1,29$.

Remarque : L'échantillon gaussien est utilisé pour modéliser de très nombreuses situations. Il arrive très souvent que des valeurs négatives pour les variables X_n n'aient aucun sens dans la réalité bien que celles-ci soient modélisées par des gaussiennes. Ceci s'explique par le fait que la probabilité pour qu'une variable aléatoire de loi $\mathcal{N}(m, \sigma^2)$ soit négative est d'autant plus petite que la moyenne m est grande et la variance σ^2 est petite.



TAB. 1. Fonction de répartion F de la loi normale standard $X \sim \mathcal{N}(0,1)$. La table ci-dessous donne la valeur $F(u) = P(X \leq u)$ en fonction de u. Par exemple si $u = 1.96 = 1.9 + 0.06$ alors $F(u) = 0.975$

$u = u_1 + u_2$ (u_1 \ u_2)	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0.1	0.5398	0.5437	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0.2	0.5792	0.5831	0.587	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.614
0.3	0.6179	0.6217	0.6255	0.6293	0.633	0.6368	0.6405	0.6443	0.648	0.6517
0.4	0.6554	0.659	0.6627	0.6664	0.67	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.719	0.7224
0.6	0.7257	0.729	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0.8	0.7881	0.791	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0.9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1	0.8413	0.8437	0.8461	0.8484	0.8508	0.8531	0.8554	0.8576	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8789	0.8809	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1.3	0.9031	0.9049	0.9065	0.9082	0.9098	0.9114	0.913	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9221	0.9236	0.925	0.9264	0.9278	0.9292	0.9305	0.9318
1.5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.944
1.6	0.9452	0.9463	0.9473	0.9484	0.9494	0.9505	0.9515	0.9525	0.9535	0.9544
1.7	0.9554	0.9563	0.9572	0.9581	0.959	0.9599	0.9607	0.9616	0.9624	0.9632
1.8	0.964	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9712	0.9719	0.9725	0.9731	0.9738	0.9744	0.975	0.9755	0.9761	0.9767
2	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2.1	0.9821	0.9825	0.9829	0.9834	0.9838	0.9842	0.9846	0.9849	0.9853	0.9857
2.2	0.986	0.9864	0.9867	0.9871	0.9874	0.9877	0.988	0.9883	0.9886	0.9889
2.3	0.9892	0.9895	0.9898	0.99	0.9903	0.9906	0.9908	0.9911	0.9913	0.9915
2.4	0.9918	0.992	0.9922	0.9924	0.9926	0.9928	0.993	0.9932	0.9934	0.9936
2.5	0.9937	0.9939	0.9941	0.9942	0.9944	0.9946	0.9947	0.9949	0.995	0.9952
2.6	0.9953	0.9954	0.9956	0.9957	0.9958	0.9959	0.996	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9975	0.9976	0.9977	0.9978	0.9978	0.9979	0.998	0.998
2.9	0.9981	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986