Probabilités et Statistiques

Statistiques descriptives

Julian Tugaut

Télécom Saint-Étienne

Sommaire

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
   - Introduction
   - Série à valeurs isolées
   - Série à valeurs classées
4. Séries statistiques simples : caractéristiques
   - Moyenne arithmétique
   - Quantiles
   - Variance, Écart-type
   - Autres caractéristiques
5. Séries statistiques doubles
   - Définitions
   - Tableau de contingence
   - Fréquence conditionnelle
6. Indépendance de deux caractères
   - Nuage de points
   - Covariance
   - Description empirique de la dépendance

Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
4. Séries statistiques simples : caractéristiques
5. Séries statistiques doubles
6. Indépendance de deux caractères

Vocabulaire - 1

Ici, on introduit le vocabulaire propre aux statistiques et plus précisément aux statistiques descriptives. Ces dernières ont pour objet de décrire des caractéristiques dans une population donnée. On parle d'enquête statistique.

Vocabulaire - 1

Ici, on introduit le vocabulaire propre aux statistiques et plus précisément aux statistiques descriptives. Ces dernières ont pour objet de décrire des caractéristiques dans une population donnée. On parle d'enquête statistique.

Une enquête statistique consiste à effectuer des observations sur des éléments d'un ensemble (la population) que l'on appelle individus de la population. Par exemple, on peut considérer l'ensemble des pièces fabriquées par une machine et l'on souhaite savoir si elles sont défectueuses. Il peut aussi s'agir de la population des actifs en France et l'on veut décrire la répartition des différentes catégories socio-professionnelles.

Vocabulaire - 2

On peut distinguer deux types d'enquêtes statistiques.

Vocabulaire - 2

On peut distinguer deux types d'enquêtes statistiques.

- On peut faire des observations sur tous les individus de la population. On dit alors que l'on fait un recensement. C'est ce qui est fait en démographie. Les premières études statistiques ont été les recensements de la population par les états. Ce type d'enquêtes est le plus naturel et il ne nécessite aucune inférence statistique puisque l'échantillon est la population elle-même. Néanmoins, on ne peut pas toujours effectuer de recensement.

Vocabulaire - 2

On peut distinguer deux types d'enquêtes statistiques.

- On peut faire des observations sur tous les individus de la population. On dit alors que l'on fait un recensement. C'est ce qui est fait en démographie. Les premières études statistiques ont été les recensements de la population par les états. Ce type d'enquêtes est le plus naturel et il ne nécessite aucune inférence statistique puisque l'échantillon est la population elle-même. Néanmoins, on ne peut pas toujours effectuer de recensement.
- On peut faire des observations sur des individus tirés de la population qui constituent un échantillon. On dit alors que l'on fait un sondage. Par exemple, certains contrôles qualité détruisent le produit. Dans ce cas, on ne les pratique pas sur tous les produits. Également, il peut être coûteux en temps de faire un recensement. Ainsi, un sondage pré-électoral ne peut pas être fait sur tous les individus de la population.

Vocabulaire - 2

On peut distinguer deux types d'enquêtes statistiques.

- On peut faire des observations sur tous les individus de la population. On dit alors que l'on fait un recensement. C'est ce qui est fait en démographie. Les premières études statistiques ont été les recensements de la population par les états. Ce type d'enquêtes est le plus naturel et il ne nécessite aucune inférence statistique puisque l'échantillon est la population elle-même. Néanmoins, on ne peut pas toujours effectuer de recensement.
- On peut faire des observations sur des individus tirés de la population qui constituent un échantillon. On dit alors que l'on fait un sondage. Par exemple, certains contrôles qualité détruisent le produit. Dans ce cas, on ne les pratique pas sur tous les produits. Également, il peut être coûteux en temps de faire un recensement. Ainsi, un sondage pré-électoral ne peut pas être fait sur tous les individus de la population.

En général, on observe sur des individus des variables. En statistiques, on parle plutôt de caractères.

Vocabulaire - 3

Il y a plusieurs types de caractères. Le premier type auquel on peut penser est celui des caractères dits qualitatifs. Ici, les possibilités (on parle plutôt de modalités) que peut avoir le caractère qualitatif ne sont pas des nombres. C'est ainsi le cas dans une enquête d'opinion. On peut donner l'exemple de l'avis des étudiants sur un cours magistral à Télécom Saint-Étienne ; les modalités pouvant être "Très Bien", "Bien", "Moyen", "Mauvais", "Très Mauvais" ou "Absent en CM". On peut objecter à cet exemple qu'un ordre naturel apparaît dans les modalités. Ainsi, bien que le caractère soit qualitatif, il peut être vu comme s'il était à valeurs numériques. Ici, "Très Bien" serait 20, "Bien" serait 16, "Moyen" serait 12, "Mauvais" serait 8 et "Très Mauvais" serait 4 tandis que la modalité "Absent" pourrait être associée à 0.

Vocabulaire - 4

Dans ce cas, on dira que le caractère est qualitatif ordinal. Un autre exemple est celui de la qualité gustative dans la population des vins : on peut établir un ordre naturel (très bonne, bonne, moyenne, médiocre, mauvaise, très mauvaise).

Vocabulaire - 4

Dans ce cas, on dira que le caractère est qualitatif ordinal. Un autre exemple est celui de la qualité gustative dans la population des vins : on peut établir un ordre naturel (très bonne, bonne, moyenne, médiocre, mauvaise, très mauvaise).

Au contraire, si l'on ne peut pas établir un ordre naturel, on parlera de caractère qualitatif nominal. Ainsi, la catégorie socio-professionnelle de la population des actifs est un caractère qualitatif nominal. De même, le genre d'un enfant parmi la population des nouveau-nés en France en 2022 est un caractère qualitatif nominal.

Vocabulaire - 5

Enfin, si les modalités sont des nombres, on dit que le caractère est quantitatif. Ainsi, la taille des nouveau-nés en France en 2022 est un caractère quantitatif. Il en est de même du poids. Si on prend l'exemple des actifs en France, l'âge est un caractère quantitatif également. Globalement, un caractère quantitatif peut être vu comme la réalisation d'une variable aléatoire.

Vocabulaire - 5

Enfin, si les modalités sont des nombres, on dit que le caractère est quantitatif. Ainsi, la taille des nouveau-nés en France en 2022 est un caractère quantitatif. Il en est de même du poids. Si on prend l'exemple des actifs en France, l'âge est un caractère quantitatif également. Globalement, un caractère quantitatif peut être vu comme la réalisation d'une variable aléatoire.

Comme pour les variables aléatoires, on distingue les caractères quantitatifs en fonction de l'ensemble des modalités possibles. Si l'ensemble est fini ou infini dénombrable (par exemple, la durée de vie en années d'une ampoule), on dit que le caractère est discret. Si l'ensemble des modalités que peut prendre un caractère quantitatif est infini non dénombrable, on dit que le caractère est continu.

Vocabulaire - 6

En réalité, en statistiques, la simplicité de la lecture étant le maître-mot, la différence est plus ténue. Par exemple, la taille en cm est un caractère quantitatif discret sur la population des actifs. Néanmoins, le nombre de valeurs étant très élevé, il n'est pas envisageable de représenter ce caractère quantitatif comme s'il s'agissait du nombre d'enfants (que l'on peut, sans trop de restrictions, majorer par 10) au sein des ménages français. Il est alors pertinent de le considérer comme étant continu.

Vocabulaire - 6

En réalité, en statistiques, la simplicité de la lecture étant le maître-mot, la différence est plus ténue. Par exemple, la taille en cm est un caractère quantitatif discret sur la population des actifs. Néanmoins, le nombre de valeurs étant très élevé, il n'est pas envisageable de représenter ce caractère quantitatif comme s'il s'agissait du nombre d'enfants (que l'on peut, sans trop de restrictions, majorer par 10) au sein des ménages français. Il est alors pertinent de le considérer comme étant continu.

Il est crucial de comprendre que c'est l'usage qui déterminera souvent si le caractère sera à considérer comme discret ou comme continu.

Vocabulaire - 7

Enfin, les données peuvent être vues de deux manières. On peut les garder telles quelles ou on peut les grouper. Plutôt que d'essayer de théoriser, donnons un exemple. On s'intéresse ainsi aux notes obtenues sur 20 par 15 étudiants. Voici les données brutes :

$$10, 20, 12, 10, 20, 15, 16, 20, 10, 3, 4, 15, 20, 12, 12.$$

Vocabulaire - 7

Enfin, les données peuvent être vues de deux manières. On peut les garder telles quelles ou on peut les grouper. Plutôt que d'essayer de théoriser, donnons un exemple. On s'intéresse ainsi aux notes obtenues sur 20 par 15 étudiants. Voici les données brutes :

$$10, 20, 12, 10, 20, 15, 16, 20, 10, 3, 4, 15, 20, 12, 12.$$

Lire des données est faisable car le nombre 15 n'est pas trop élevé. Néanmoins, dessiner des tendances n'est pas immédiat.

Vocabulaire - 7

Enfin, les données peuvent être vues de deux manières. On peut les garder telles quelles ou on peut les grouper. Plutôt que d'essayer de théoriser, donnons un exemple. On s'intéresse ainsi aux notes obtenues sur 20 par 15 étudiants. Voici les données brutes :

$$10, 20, 12, 10, 20, 15, 16, 20, 10, 3, 4, 15, 20, 12, 12.$$

Lire des données est faisable car le nombre 15 n'est pas trop élevé. Néanmoins, dessiner des tendances n'est pas immédiat.

On les groupe donc comme suit pour obtenir les données dites groupées :

$$3(1), 4(1), 10(3), 12(3), 15(2), 16(1), 20(4).$$

Le nombre entre parenthèses correspond au nombre d'occurrences de la modalité qui la précède. On peut ensuite présenter de manière plus adéquate pour vraiment étudier cette série.

Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
4. Séries statistiques simples : caractéristiques
5. Séries statistiques doubles
6. Indépendance de deux caractères

Caractère qualitatif - 1

Nous verrons exclusivement un moyen de représenter graphiquement les séries statistiques ayant des modalités non numériques.

Caractère qualitatif - 1

Nous verrons exclusivement un moyen de représenter graphiquement les séries statistiques ayant des modalités non numériques.

La représentation utilisée est celle du diagramme circulaire (ou camembert). Ce diagramme se prête bien au cas où il y a un nombre fini et peu élevé de modalités r. Alors, on trace un disque partitionné en différents secteurs. Chaque secteur en représente une. Et, son angle est proportionnel à la fréquence f_k de la modalité a_k. Ici, la fréquence est $f_k := \frac{n_k}{n}$ où n_k est le nombre d'individus ayant la modalité a_k tandis que n est le nombre total d'individus de la population. Pour simplifier, on utilise plus facilement les degrés que les radians.

Caractère qualitatif - 2

Exemple

Pour un bilan de bloc à Télécom Saint-Étienne, les délégués réalisent un sondage sur l'appréciation d'un cours magistral. Les modalités sont : Très Bien (TB) ; Bien (B) ; Moyen ; Mauvais (M) ; Très Mauvais (TM) ; Absent au cours (A).

Caractère qualitatif - 2

Exemple

Pour un bilan de bloc à Télécom Saint-Étienne, les délégués réalisent un sondage sur l'appréciation d'un cours magistral. Les modalités sont : Très Bien (TB) ; Bien (B) ; Moyen ; Mauvais (M) ; Très Mauvais (TM) ; Absent au cours (A).

Sur les cent dix étudiants, soixante ont répondu. Et, voici la répartition des avis :

Caractère qualitatif - 2

Exemple

Pour un bilan de bloc à Télécom Saint-Étienne, les délégués réalisent un sondage sur l'appréciation d'un cours magistral. Les modalités sont : Très Bien (TB) ; Bien (B) ; Moyen ; Mauvais (M) ; Très Mauvais (TM) ; Absent au cours (A).

Sur les cent dix étudiants, soixante ont répondu. Et, voici la répartition des avis :

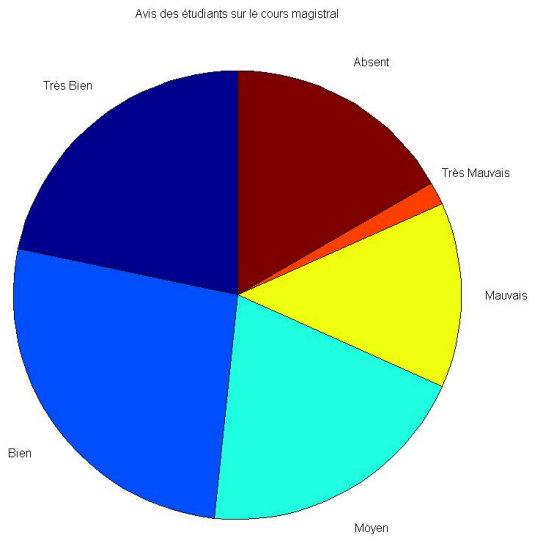
Avis	TB	B	Moyen	M	TM	A
Nombre d'étudiants	13	16	12	8	1	10

Caractère qualitatif - 3

On obtient ainsi le diagramme circulaire suivant :

Caractère qualitatif - 3

On obtient ainsi le diagramme circulaire suivant :

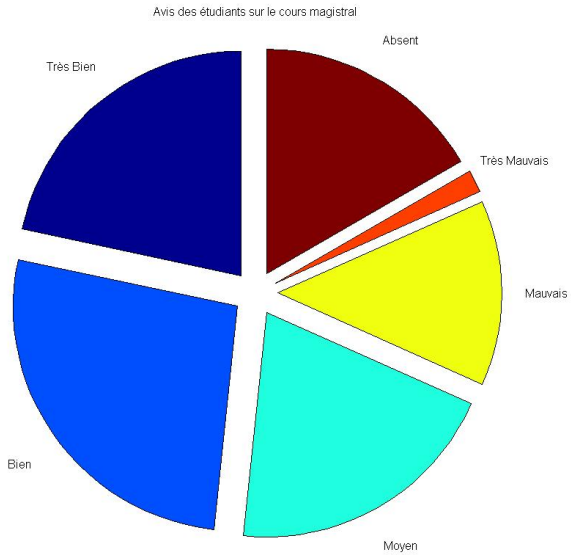


Caractère qualitatif - 4

On peut aussi “exploser” le diagramme pour le rendre plus lisible :

Caractère qualitatif - 4

On peut aussi “exploser” le diagramme pour le rendre plus lisible :

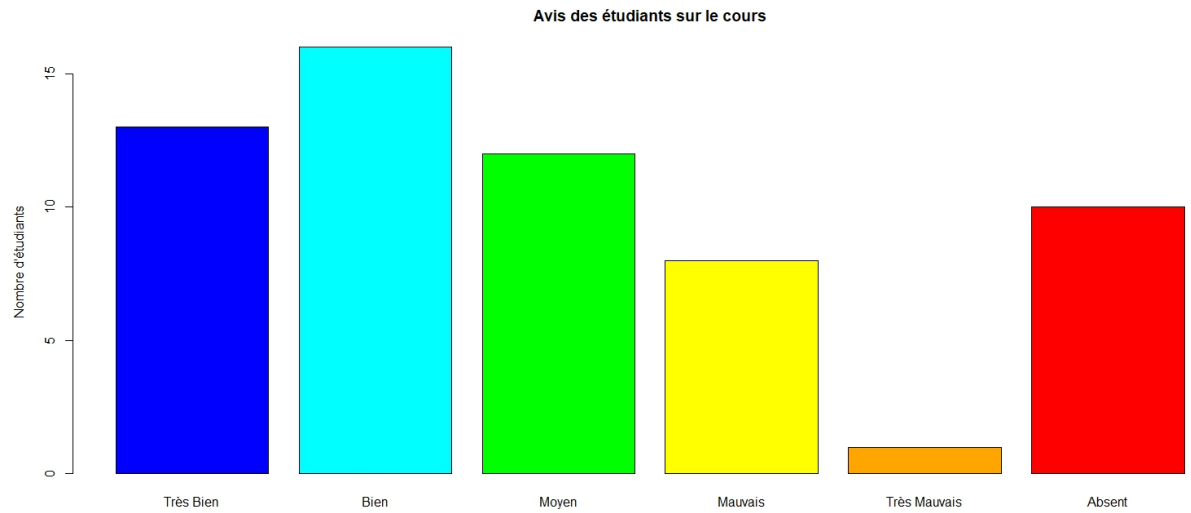


Caractère qualitatif - 5

Il convient de noter que l'on peut aussi utiliser un diagramme en barres, lequel est, la plupart du temps, plus facile à lire.

Caractère qualitatif - 5

Il convient de noter que l'on peut aussi utiliser un diagramme en barres, lequel est, la plupart du temps, plus facile à lire.



Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
   - Introduction
   - Série à valeurs isolées
   - Série à valeurs classées
4. Séries statistiques simples : caractéristiques
5. Séries statistiques doubles
6. Indépendance de deux caractères

Introduction

On a observé sur n individus d'une population I un caractère quantitatif x dont on note les valeurs observées : $x[1], \cdots, x[n]$. La valeur observée de l'individu i est $x[i]$.

Introduction

On a observé sur n individus d'une population I un caractère quantitatif x dont on note les valeurs observées : $x[1], \cdots, x[n]$. La valeur observée de l'individu i est $x[i]$.

Définition

La suite de valeurs $(x[1], \cdots, x[n])$ est appelée une série statistique simple.

Série à valeurs isolées - 1

Les séries à valeurs isolées sont utilisées pour les caractères quantitatifs discrets. Plus exactement, on les utilise pour les caractères quantitatifs dont le nombre de modalités possibles n'est pas trop élevé.

Série à valeurs isolées - 1

Les séries à valeurs isolées sont utilisées pour les caractères quantitatifs discrets. Plus exactement, on les utilise pour les caractères quantitatifs dont le nombre de modalités possibles n'est pas trop élevé.

On regroupe les valeurs égales de la série, on note l'effectif de chaque valeur isolée et on les range par ordre croissant. On suppose qu'il y a r valeurs différentes : $y_1 < \cdots < y_k < \cdots < y_r$.

Série à valeurs isolées - 2

Valeurs isolées	y_1	y_2	$\cdots$	y_k	$\cdots$	y_r
Effectifs	n_1	n_2	$\cdots$	n_k	$\cdots$	n_r
Fréquences	$f_1 = \frac{n_1}{n}$	$f_2 = \frac{n_2}{n}$	$\cdots$	$f_k = \frac{n_k}{n}$	$\cdots$	$f_r = \frac{n_r}{n}$

Série à valeurs isolées - 3

Proposition

D'abord, $r \leq n$. Ensuite :

$$\sum_{k=1}^{r} n_k = n \quad \text{et} \quad \sum_{k=1}^{r} f_k = 1\,.$$

Série à valeurs isolées - 3

Proposition

D'abord, $r \leq n$. Ensuite :

$$\sum_{k=1}^{r} n_k = n \quad \text{et} \quad \sum_{k=1}^{r} f_k = 1\,.$$

Ici, f_k est la proportion d'éléments de la série égaux à y_k.

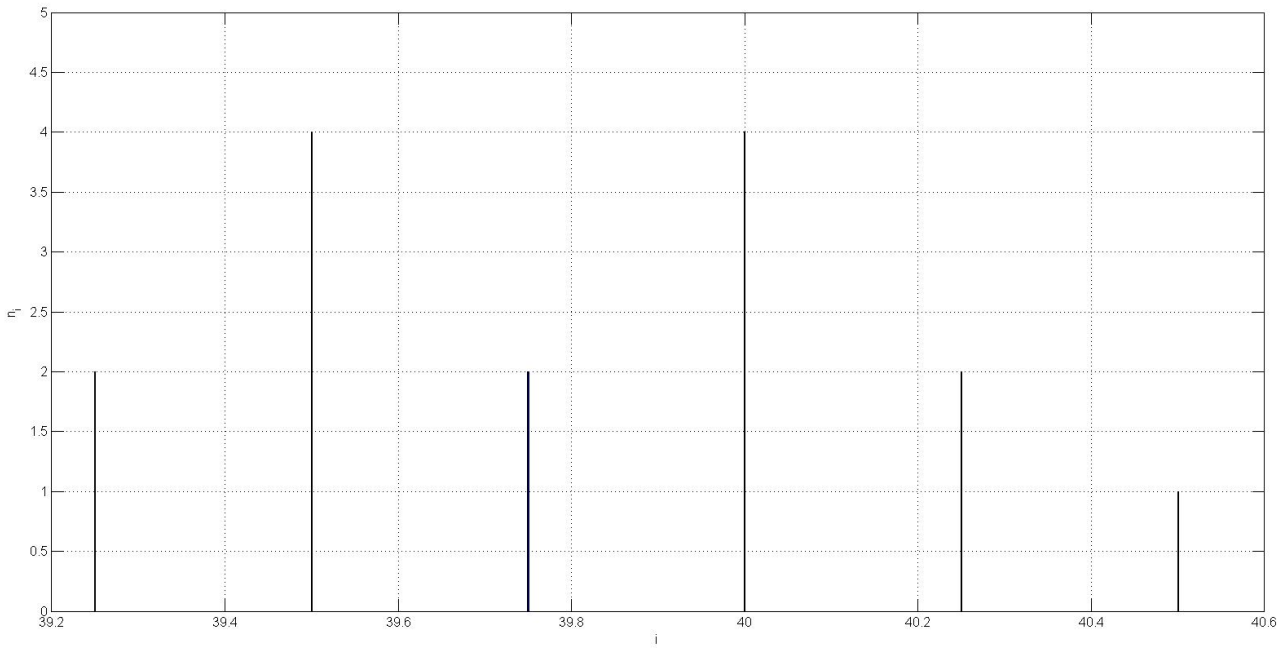
Série à valeurs isolées - 4

Un exemple avec $n = 15$ et $r = 6$:

y_i	n_i (effectifs)	f_i (fréquence)	N_i	F_i
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Diagramme en bâtons des effectifs

Diagramme en bâtons des effectifs



Modes

Définition

Les points où l'on a un maximum relatif (local) de l'effectif (ou de la fréquence) sont appelés des modes.

Modes

Définition

Les points où l'on a un maximum relatif (local) de l'effectif (ou de la fréquence) sont appelés des modes.

Remarque

La plurimodalité, c'est-à-dire le fait d'avoir plusieurs modes, signifie que la population que l'on regarde n'est pas homogène. Ainsi, il est judicieux de la scinder pour récupérer des sous-populations homogènes.

Modes

Définition

Les points où l'on a un maximum relatif (local) de l'effectif (ou de la fréquence) sont appelés des modes.

Remarque

La plurimodalité, c'est-à-dire le fait d'avoir plusieurs modes, signifie que la population que l'on regarde n'est pas homogène. Ainsi, il est judicieux de la scinder pour récupérer des sous-populations homogènes.

Ici, les modes sont en 39.50 et 40.

Effectifs cumulés

On calcule les effectifs cumulés et les fréquences cumulées (l'équivalent de la fonction de répartition) :

$$N_k := n_1 + \cdots + n_k = \sum_{l=1}^{k} n_l \,,$$

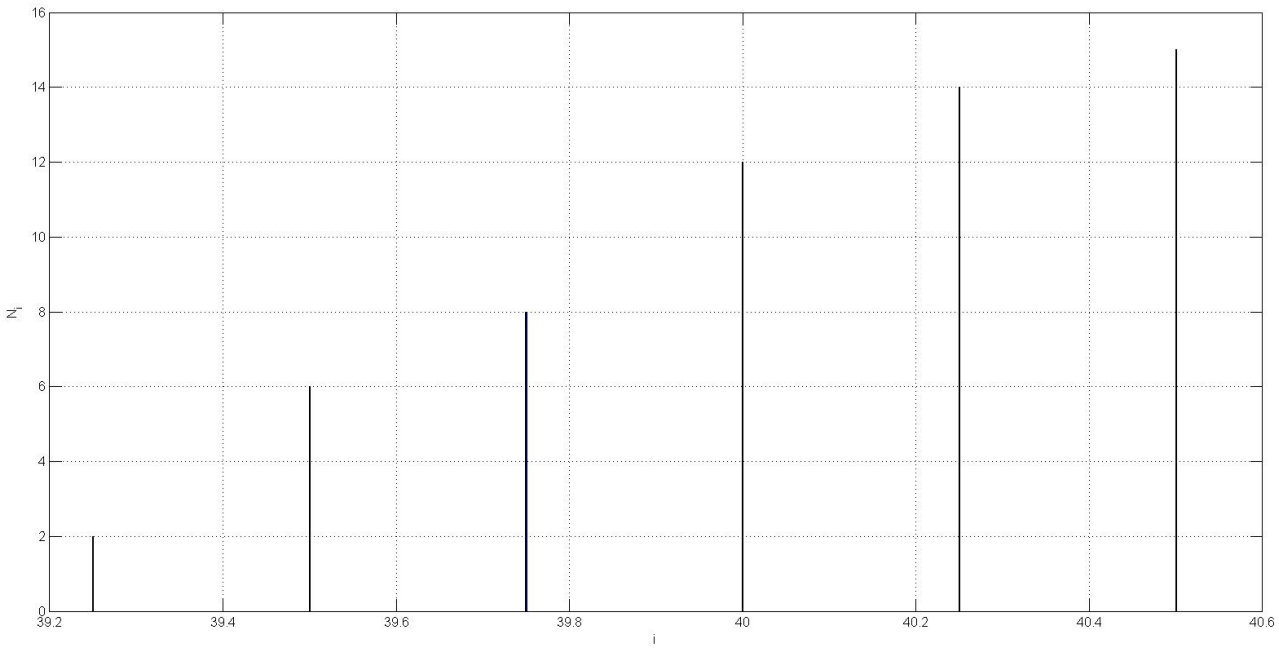
et

$$F_k := f_1 + \cdots + f_k = \sum_{l=1}^{k} f_l = \frac{N_k}{n} \,.$$

N_k est le nombre d'éléments de la série qui sont inférieurs ou égaux à y_k. Et, F_k est la proportion de tels éléments. Ainsi, douze éléments de la série sont inférieurs ou égaux à 40.

Diagramme en bâtons des effectifs cumulés

Diagramme en bâtons des effectifs cumulés



Série à valeurs classées - 1

Les séries à valeurs classées sont utilisées pour les caractères quantitatifs continus ou pour les caractères quantitatifs discrets qui prennent un grand nombre de valeurs possibles.

Série à valeurs classées - 1

Les séries à valeurs classées sont utilisées pour les caractères quantitatifs continus ou pour les caractères quantitatifs discrets qui prennent un grand nombre de valeurs possibles.

On regroupe les éléments de la série dans des intervalles semi-ouverts $[z_k; z_{k+1}[$ appelés classes ; où $z_{k-1} < z_k$ pour tout $k \in [\![1; s]\!]$. Et, on note l'effectif (et la fréquence de chaque classe). On considère s classes.

Série à valeurs classées - 2

Classes	$[z_0; z_1[$	$[z_1; z_2[$	$\cdots$	$[z_{k-1}; z_k[$	$\cdots$	$[z_{s-1}; z_s[$
Effectifs	n_1	n_2	$\cdots$	n_k	$\cdots$	n_s
Fréquences	$f_1 = \frac{n_1}{n}$	$f_2 = \frac{n_2}{n}$	$\cdots$	$f_k = \frac{n_k}{n}$	$\cdots$	$f_s = \frac{n_s}{n}$

La quantité f_k représente la proportion d'éléments de la série qui sont dans la classe $[z_{k-1}; z_k[$.

Série à valeurs classées - 3

Proposition

D'abord, $s \leq n$. Ensuite :

$$\sum_{k=1}^{s} n_k = n \quad \text{et} \quad \sum_{k=1}^{s} f_k = 1\,.$$

Série à valeurs classées - 4

On donne un exemple avec $n = 80$:

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

Série à valeurs classées - 5

Série à valeurs classées - 5

Remarque

La répartition des éléments de la série à l'intérieur d'une même classe ne doit pas être trop éloignée d'une répartition uniforme. Cette hypothèse de distribution uniforme au sein des classes est nécessaire pour effectuer des calculs sur la moyenne et la variance. De même, le calcul des quantiles nécessite que la courbe des fréquences cumulées soit affine par morceaux et donc qu'au sein de chaque classe la répartition soit uniforme. De fait, il est important que les classes ne soient pas trop larges non plus.

Histogramme - 1

On regarde maintenant l'histogramme des effectifs (ou des fréquences). En abscisse, on met les bornes des classes et en ordonnée les effectifs par unité de longueur de classe (effectifs corrigés). Ainsi, il est essentiel de bien comprendre que l'on ne représente pas simplement l'histogramme tel qu'il a pu être vu au lycée.

Histogramme - 1

On regarde maintenant l'histogramme des effectifs (ou des fréquences). En abscisse, on met les bornes des classes et en ordonnée les effectifs par unité de longueur de classe (effectifs corrigés). Ainsi, il est essentiel de bien comprendre que l'on ne représente pas simplement l'histogramme tel qu'il a pu être vu au lycée.

L'objectif principal de cette renormalisation est d'approximer la densité de probabilité de la variable aléatoire continue sous-jacente.

Histogramme - 1

On regarde maintenant l'histogramme des effectifs (ou des fréquences). En abscisse, on met les bornes des classes et en ordonnée les effectifs par unité de longueur de classe (effectifs corrigés). Ainsi, il est essentiel de bien comprendre que l'on ne représente pas simplement l'histogramme tel qu'il a pu être vu au lycée.

L'objectif principal de cette renormalisation est d'approximer la densité de probabilité de la variable aléatoire continue sous-jacente. La première chose à faire est de choisir une unité adéquate. Par exemple, ici, les valeurs de z_k vont de 100 à 170. Ainsi, on pose $u = 10$.

Histogramme - 2

Proposition

L'aire du rectangle correspondant à $[z_{k-1}; z_k[$ dans histogramme des effectifs corrigés est proportionnelle à l'effectif de la classe. En effet, il vaut $A_k = (z_k - z_{k-1}) \times \frac{n_k \times u}{z_k - z_{k-1}} = un_k$. On en déduit par ailleurs que l'aire totale est égale à
$un_1 + \cdots + un_k + \cdots + un_s = un$.

Histogramme - 2

Proposition

L'aire du rectangle correspondant à $[z_{k-1}; z_k[$ dans histogramme des effectifs corrigés est proportionnelle à l'effectif de la classe. En effet, il vaut $A_k = (z_k - z_{k-1}) \times \frac{n_k \times u}{z_k - z_{k-1}} = un_k$. On en déduit par ailleurs que l'aire totale est égale à
$un_1 + \cdots + un_k + \cdots + un_s = un$.

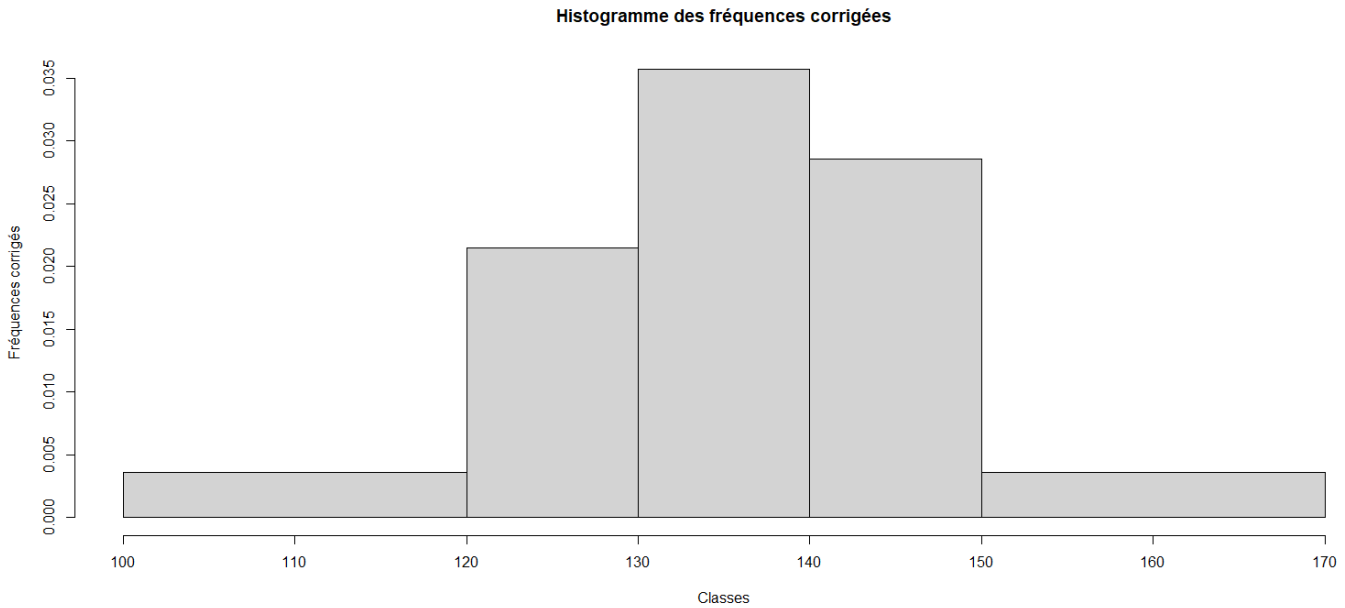
On peut noter que pour l'histogramme des fréquences corrigées, en prenant $u = 1$, l'aire totale vaut 1 ; comme c'est le cas pour toute densité de probabilité.

Histogramme - 3

$[z_{k-1}; z_k[$	n_k	f_k	$z_k - z_{k-1}$	$\frac{z_k - z_{k-1}}{u}$	$\frac{f_k \times u}{z_k - z_{k-1}}$
$[100; 120[$	10	$\frac{10}{80}$	20	2	$\frac{5}{80}$
$[120; 130[$	15	$\frac{15}{80}$	10	1	$\frac{15}{80}$
$[130; 140[$	25	$\frac{25}{80}$	10	1	$\frac{25}{80}$
$[140; 150[$	20	$\frac{20}{80}$	10	1	$\frac{20}{80}$
$[150; 170[$	10	$\frac{10}{80}$	20	2	$\frac{5}{80}$

Histogramme - 4

Histogramme - 4



Classes modales

Définition

Les classes pour lesquelles on a un maximum relatif ou local de l'effectif corrigé sont appelées "classes modales". On prend pour modes les milieux des classes modales.

Classes modales

Définition

Les classes pour lesquelles on a un maximum relatif ou local de l'effectif corrigé sont appelées "classes modales". On prend pour modes les milieux des classes modales.

Dans l'exemple plus haut, la classe modale est $[130; 140[$ et le mode est donc 135.

Effectifs cumulés

Les effectifs cumulés et les fréquences cumulées sont calculés de la même manière que pour les séries statistiques simples à valeurs isolées. Ainsi, cinquante éléments sont strictement inférieurs à 140. N_k (respectivement F_k) représente le nombre (respectivement la proportion) d'éléments de la série strictement inférieurs à z_k.

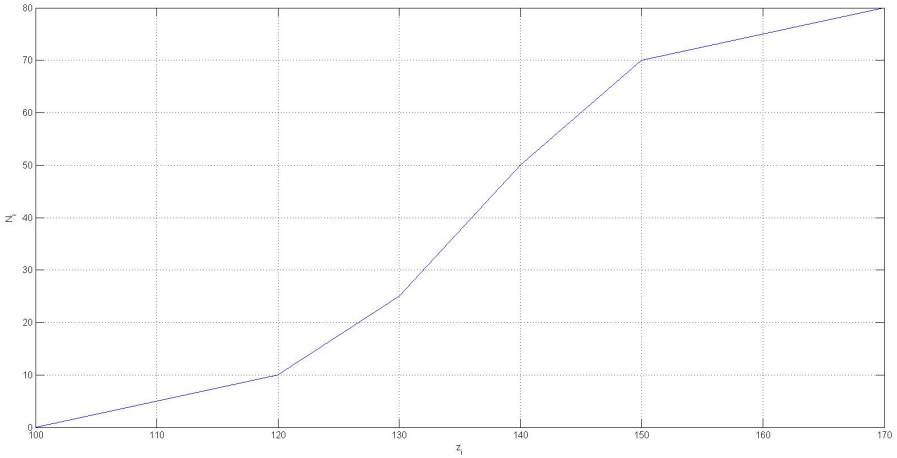
Effectifs cumulés

Les effectifs cumulés et les fréquences cumulées sont calculés de la même manière que pour les séries statistiques simples à valeurs isolées. Ainsi, cinquante éléments sont strictement inférieurs à 140. N_k (respectivement F_k) représente le nombre (respectivement la proportion) d'éléments de la série strictement inférieurs à z_k.

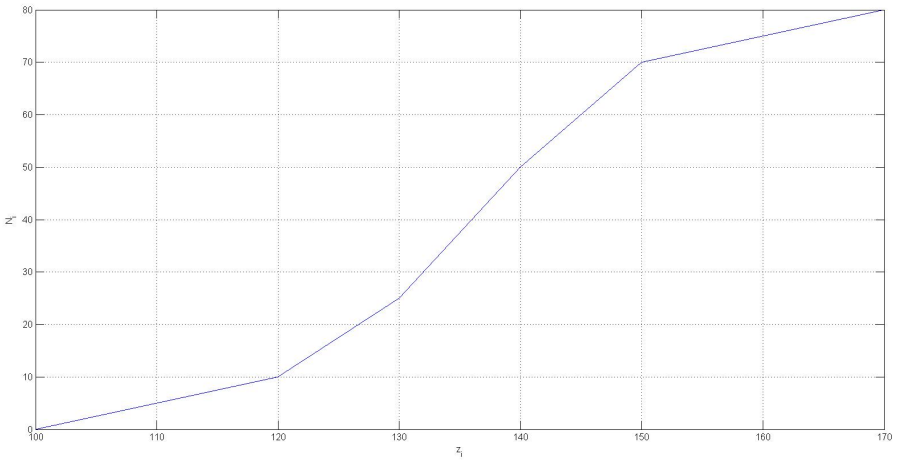
On obtient alors la courbe des effectifs cumulés (ou celle des fréquences cumulées) en joignant par des segments de droite les points (z_k, N_k) (ou (z_k, F_k)). En effet, comme on suppose que la répartition au sein d'une classe est uniforme, la fonction de répartition sous-jacente est affine par morceaux.

Courbe des effectifs cumulés

Courbe des effectifs cumulés

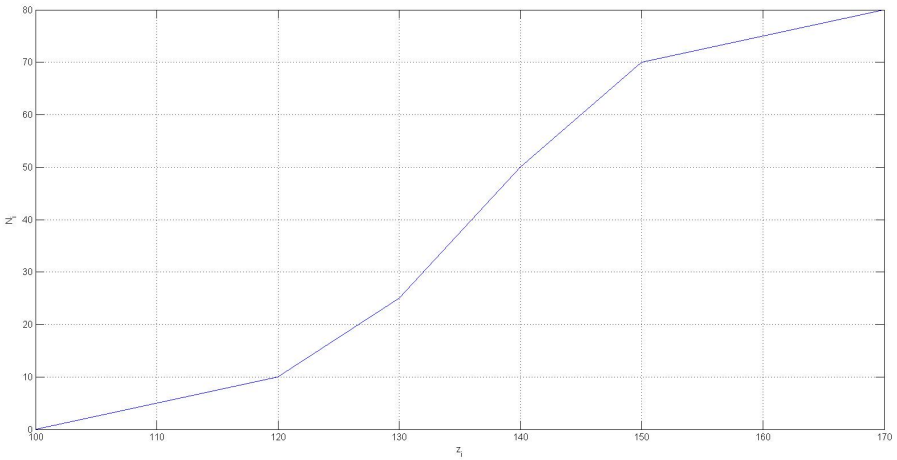


Courbe des effectifs cumulés



Par convention, on pose $N_0 = F_0 = 0$.

Courbe des effectifs cumulés



Par convention, on pose $N_0 = F_0 = 0$.

Remarque

La courbe des fréquences cumulées peut s'assimiler à la fonction de répartition.

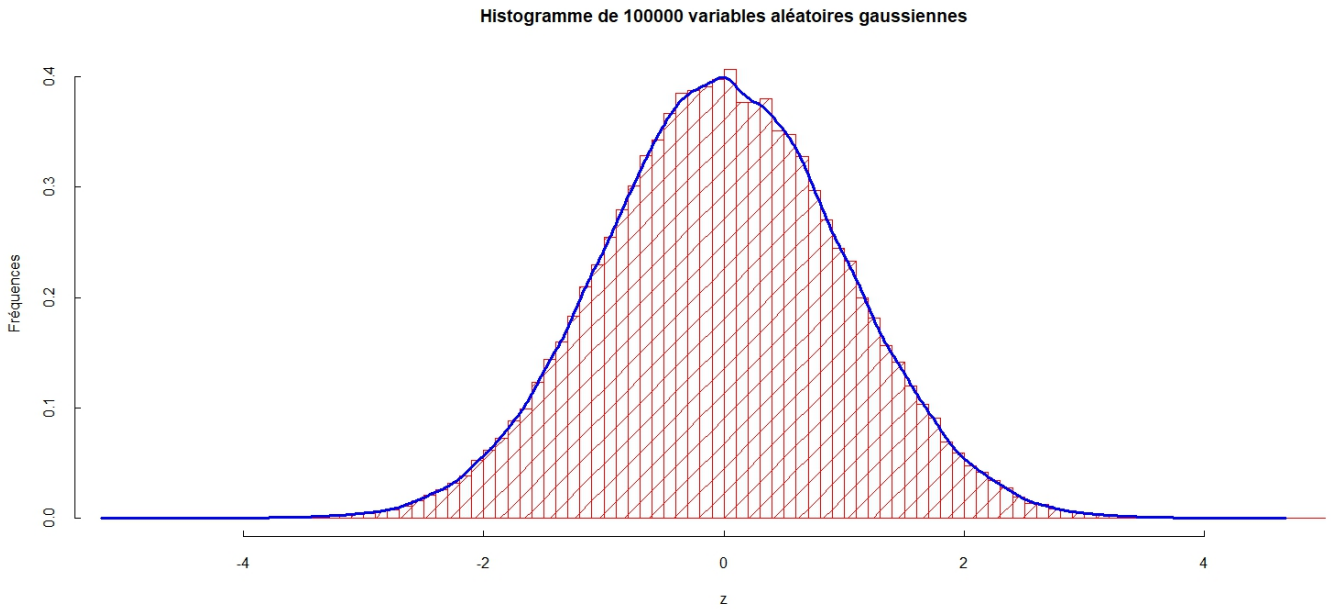
Histogramme avec longueur de classe constante - 1

Remarque

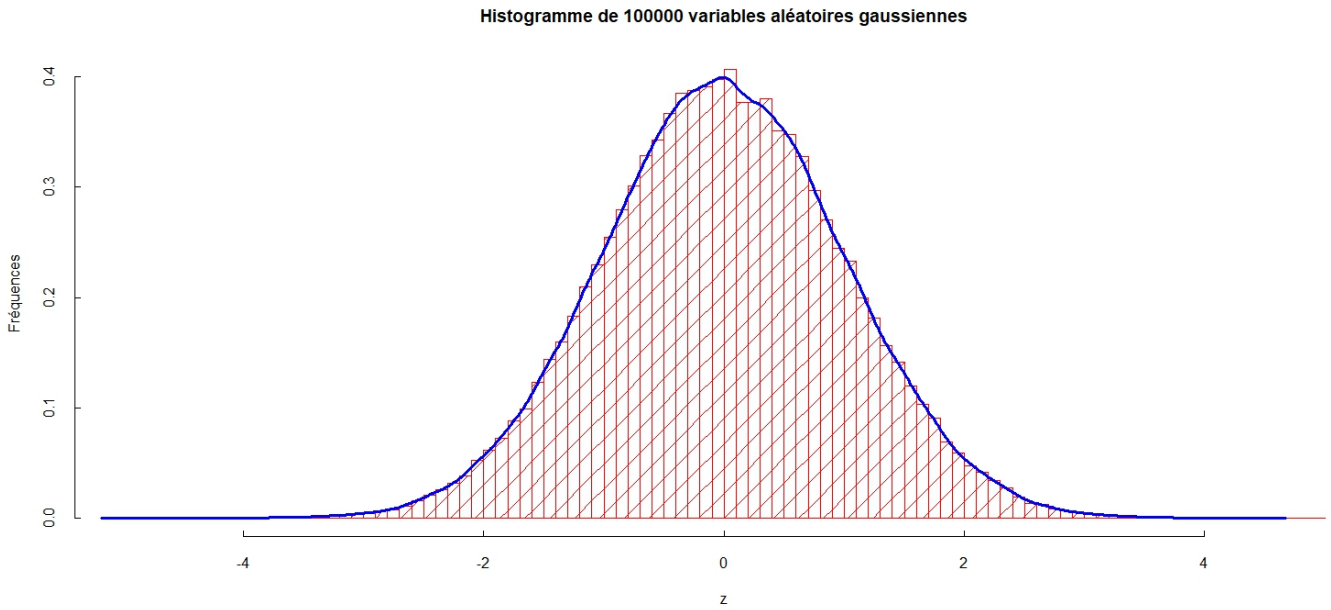
Il est généralement plus simple de considérer que la longueur des classes est constante. Alors, il n'y a pas besoin de corriger les effectifs, ou les fréquences. Voici par exemple un histogramme de 100 000 variables aléatoires suivant la loi normale centrée réduite où chaque classe est de longueur 0.1 :

Histogramme avec longueur de classe constante - 2

Histogramme avec longueur de classe constante - 2



Histogramme avec longueur de classe constante - 2



Par ailleurs, la densité de probabilité empirique a été superposée, en bleu.

Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
4. Séries statistiques simples : caractéristiques
   - Moyenne arithmétique
   - Quantiles
   - Variance, Écart-type
   - Autres caractéristiques
5. Séries statistiques doubles
6. Indépendance de deux caractères

Moyenne arithmétique

On se donne une série statistique simple $(x[1], \cdots, x[i], \cdots, x[n])$.

Moyenne arithmétique

On se donne une série statistique simple $(x[1], \cdots, x[i], \cdots, x[n])$.

Définition

On appelle moyenne arithmétique de cette série le nombre égal à la somme des éléments de la série divisé par l'effectif n. On note $\overline{x}$ cette moyenne arithmétique :

$$\overline{x} = \frac{\sum_{i=1}^{n} x[i]}{n}.$$

Moyenne arithmétique

On se donne une série statistique simple $(x[1], \cdots, x[i], \cdots, x[n])$.

Définition

On appelle moyenne arithmétique de cette série le nombre égal à la somme des éléments de la série divisé par l'effectif n. On note $\overline{x}$ cette moyenne arithmétique :

$$\overline{x} = \frac{\sum_{i=1}^{n} x[i]}{n}.$$

Cette quantité correspond à l'espérance dans le cas d'une variable aléatoire réelle discrète.

Moyenne arithmétique

On se donne une série statistique simple $(x[1], \cdots, x[i], \cdots, x[n])$.

Définition

On appelle moyenne arithmétique de cette série le nombre égal à la somme des éléments de la série divisé par l'effectif n. On note $\overline{x}$ cette moyenne arithmétique :

$$\overline{x} = \frac{\sum_{i=1}^{n} x[i]}{n}.$$

Cette quantité correspond à l'espérance dans le cas d'une variable aléatoire réelle discrète.

Remarque

La moyenne est sensible aux valeurs extrêmes (qui peuvent être aberrantes) de la série.

Calcul dans le cas d'une série à valeurs isolées - 1

On suppose que l'on dispose de r modalités $y_1, \cdots, y_k, \cdots, y_r$ et l'effectif associé à la modalité y_k est n_k. La fréquence est ainsi $f_k = \frac{n_k}{n}$.

Calcul dans le cas d'une série à valeurs isolées - 1

On suppose que l'on dispose de r modalités $y_1, \cdots, y_k, \cdots, y_r$ et l'effectif associé à la modalité y_k est n_k. La fréquence est ainsi $f_k = \frac{n_k}{n}$.

Proposition

Dans ce cas, la moyenne arithmétique est :

$$\overline{x} = \frac{\sum_{k=1}^{r} n_k y_k}{n} = \sum_{k=1}^{r} f_k y_k \, .$$

Calcul dans le cas d'une série à valeurs isolées - 2

Exemple

y_k	n_k (effectifs)	f_k (fréquence)	N_k	F_k
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Calcul dans le cas d'une série à valeurs isolées - 2

Exemple

y_k	n_k (effectifs)	f_k (fréquence)	N_k	F_k
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Alors, la moyenne arithmétique est égale à

$$\begin{aligned}\overline{x} &= \frac{\sum_{k=1}^{r} n_k y_k}{n} \\ &= \frac{2 \times 39.25 + 4 \times 39.50 + \cdots + 2 \times 40.25 + 1 \times 40.50}{15} \\ &= 39.80\,.\end{aligned}$$

Calcul dans le cas d'une série à valeurs classées - 1

On suppose que l'on dispose de s classes $[z_0; z_1[, \cdots, [z_{k-1}; z_k[, \cdots, [z_{s-1}; z_s[$ et l'effectif associé à la classe numéro k est n_k. La fréquence est ainsi $f_k = \frac{n_k}{n}$.

Calcul dans le cas d'une série à valeurs classées - 1

On suppose que l'on dispose de s classes $[z_0; z_1[, \cdots, [z_{k-1}; z_k[, \cdots, [z_{s-1}; z_s[$ et l'effectif associé à la classe numéro k est n_k. La fréquence est ainsi $f_k = \frac{n_k}{n}$.

Ici, il est vraiment crucial que les répartitions au sein des différentes classes soient uniformes.

Calcul dans le cas d'une série à valeurs classées - 1

On suppose que l'on dispose de s classes $[z_0; z_1[, \cdots, [z_{k-1}; z_k[, \cdots, [z_{s-1}; z_s[$ et l'effectif associé à la classe numéro k est n_k. La fréquence est ainsi $f_k = \frac{n_k}{n}$.

Ici, il est vraiment crucial que les répartitions au sein des différentes classes soient uniformes.

Proposition

Si la répartition au sein de chaque classe est uniforme, alors la moyenne $\overline{x} = \frac{\sum_{i=1}^{n} x[i]}{n}$ peut être approximée par

$$\frac{1}{n}\sum_{k=1}^{r} n_k \frac{z_{k-1}+z_k}{2} = \sum_{k=1}^{r} f_k \frac{z_{k-1}+z_k}{2}.$$

Calcul dans le cas d'une série à valeurs classées - 2

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

Calcul dans le cas d'une série à valeurs classées - 2

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

Alors, on peut calculer comme suit

$$\overline{x} = \frac{10 \times 110 + \cdots + 20 \times 145 + 10 \times 160}{80} = 135.625\,.$$

Définition

Définition

Soit $0 < p < 1$. Le quantile (ou fractile) d'ordre p est le plus petit nombre $q_p(x)$ tel que la proportion des éléments de la série x strictement inférieurs à $q_p(x)$ soit supérieure ou égale à p.

Quelques quantiles particuliers

Notation

Pour $p = \frac{1}{2}$, la quantité $q_{\frac{1}{2}}(x)$ porte le nom de médiane.

Quelques quantiles particuliers

Notation

Pour $p = \frac{1}{2}$, la quantité $q_{\frac{1}{2}}(x)$ porte le nom de médiane.

Pour $p = \frac{1}{4}$, on parle de premier quartile. Pour $p = \frac{3}{4}$, on parle de troisième quartile.

Quelques quantiles particuliers

Notation

Pour $p = \frac{1}{2}$, la quantité $q_{\frac{1}{2}}(x)$ porte le nom de médiane.

Pour $p = \frac{1}{4}$, on parle de premier quartile. Pour $p = \frac{3}{4}$, on parle de troisième quartile.

Pour $p = \frac{k}{10}$ (avec $k \in [\![1; 9]\!]$), on parle de déciles.

Quelques quantiles particuliers

Notation

Pour $p = \frac{1}{2}$, la quantité $q_{\frac{1}{2}}(x)$ porte le nom de médiane.

Pour $p = \frac{1}{4}$, on parle de premier quartile. Pour $p = \frac{3}{4}$, on parle de troisième quartile.

Pour $p = \frac{k}{10}$ (avec $k \in [\![1; 9]\!]$), on parle de déciles.

Pour $p = \frac{k}{100}$ (avec $k \in [\![1; 99]\!]$), on parle de centiles.

Box-plot - 1

Une représentation graphique fort utile pour visualiser d'un coup d'œil les quantiles est le box-plot aussi appelée le diagramme à moustaches. Ce diagramme représente un rectangle dont les extrémités sont le premier quartile et le troisième quartile. Ce rectangle est coupé d'une barre qui représente la médiane.

Box-plot - 1

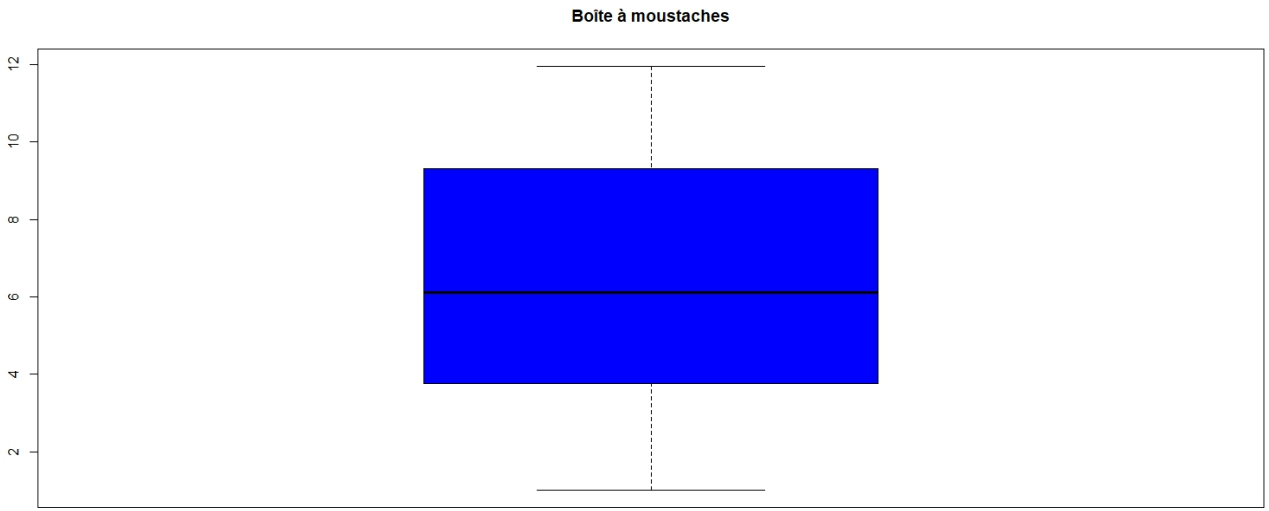
Une représentation graphique fort utile pour visualiser d'un coup d'œil les quantiles est le box-plot aussi appelée le diagramme à moustaches. Ce diagramme représente un rectangle dont les extrémités sont le premier quartile et le troisième quartile. Ce rectangle est coupé d'une barre qui représente la médiane. Enfin, la boîte a deux "moustaches" qui représente le premier décile et le neuvième décile.

Box-plot - 2

Voici un exemple pour une série statistique constituée de 1 000 simulations d'une loi uniforme sur $[1; 12]$:

Box-plot - 2

Voici un exemple pour une série statistique constituée de 1 000 simulations d'une loi uniforme sur $[1; 12]$:

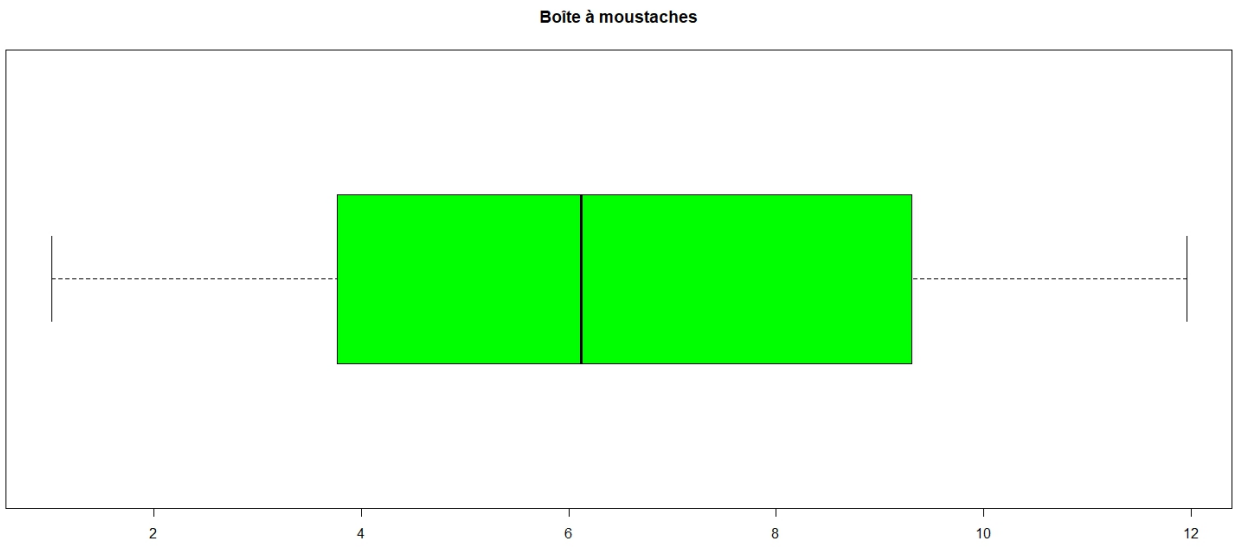


Box-plot - 3

On peut parfois la représenter horizontalement :

Box-plot - 3

On peut parfois la représenter horizontalement :

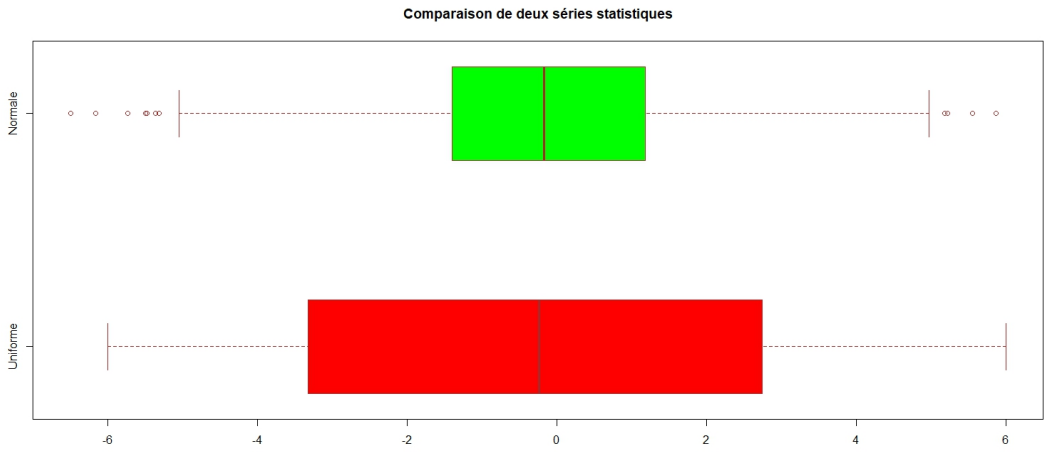


Box-plot - 4

Cette représentation peut servir à comparer deux séries statistiques comme dans le cas suivant avec 1 000 simulations de loi uniforme sur $[-6; 6]$ et 1 000 simulations de loi normale centrée et d'écart-type 2 :

Box-plot - 4

Cette représentation peut servir à comparer deux séries statistiques comme dans le cas suivant avec 1 000 simulations de loi uniforme sur $[-6; 6]$ et 1 000 simulations de loi normale centrée et d'écart-type 2 :

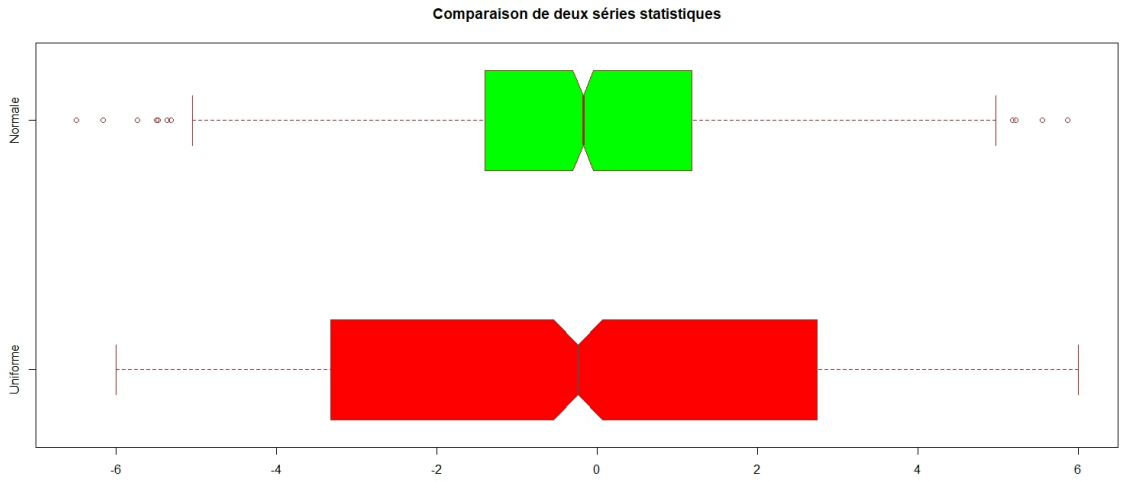


Box-plot - 5

Il convient de noter que certains logiciels (comme R), fournissent la possibilité de couper un peu le rectangle pour que celui-ci ressemble plus à des moustaches :

Box-plot - 5

Il convient de noter que certains logiciels (comme R), fournissent la possibilité de couper un peu le rectangle pour que celui-ci ressemble plus à des moustaches :



Calcul dans le cas d'une série à valeurs isolées - 1

y_i	n_i (effectifs)	f_i (fréquence)	N_i	F_i
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Calcul dans le cas d'une série à valeurs isolées - 2

On calcule la médiane (quantile d'ordre $p = \frac{1}{2}$). Ici, $n = 15$ donc $n_p = \frac{1}{2} \times n = 7.5 \notin \mathbb{N}$. On regarde la plus petite valeur telle que les effectifs cumulés sont supérieurs ou égaux à 7.5 donc à 8. Ainsi, on a $q_{\frac{1}{2}}(x) = 39.75$.

Calcul dans le cas d'une série à valeurs isolées - 2

On calcule la médiane (quantile d'ordre $p = \frac{1}{2}$). Ici, $n = 15$ donc $n_p = \frac{1}{2} \times n = 7.5 \notin \mathbb{N}$. On regarde la plus petite valeur telle que les effectifs cumulés sont supérieurs ou égaux à 7.5 donc à 8. Ainsi, on a $q_{\frac{1}{2}}(x) = 39.75$.

Exercice

Donner le premier quartile et le troisième quartile de la série statistique simple.

Calcul dans le cas d'une série à valeurs isolées - 2

On calcule la médiane (quantile d'ordre $p = \frac{1}{2}$). Ici, $n = 15$ donc $n_p = \frac{1}{2} \times n = 7.5 \notin \mathbb{N}$. On regarde la plus petite valeur telle que les effectifs cumulés sont supérieurs ou égaux à 7.5 donc à 8. Ainsi, on a $q_{\frac{1}{2}}(x) = 39.75$.

Exercice

Donner le premier quartile et le troisième quartile de la série statistique simple.

On peut aussi déterminer les quantiles en utilisant le diagramme en bâtons des fréquences cumulées et en traçant les droites d'équations $(y = \frac{1}{4})$ et $(y = \frac{3}{4})$.

Calcul dans le cas d'une série à valeurs classées - 1

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

Calcul dans le cas d'une série à valeurs classées - 1

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

On regarde le premier quartile (quantile d'ordre $\frac{1}{4}$) : $q_{\frac{1}{4}}(x)$. Ici, $n = 80$ donc $n_p = p \times n = 20 \in \mathbb{N}$. L'intervalle de premier quartile est donc $[120; 130[$. En d'autres termes, $q_{\frac{1}{4}}(x) \in [120; 130[$. Pour calculer le premier quartile, on procède alors à une interpolation linéaire. *On fait implicitement l'hypothèse que les éléments de la série dans la classe* $[120; 130[$ *sont répartis uniformément*.

Calcul dans le cas d'une série à valeurs classées - 2

On suppose ainsi que la courbe des effectifs cumulés a pour équation :

$$y = ax + b\,.$$

Calcul dans le cas d'une série à valeurs classées - 2

On suppose ainsi que la courbe des effectifs cumulés a pour équation :

$$y = ax + b\,.$$

Il reste à déterminer a et b. On sait que l'on a

$$10 = 120a + b \quad \text{et} \quad 25 = 130a + b\,.$$

Calcul dans le cas d'une série à valeurs classées - 2

On suppose ainsi que la courbe des effectifs cumulés a pour équation :

$$y = ax + b\,.$$

Il reste à déterminer a et b. On sait que l'on a

$$10 = 120a + b \quad \text{et} \quad 25 = 130a + b\,.$$

On en déduit immédiatement $a = \frac{15}{10}$ et $b = 10 - 120 \times a = -170$. Par conséquent, on obtient

$$20 = \frac{15}{10} q_{\frac{1}{4}}(x) - 170\,.$$

Calcul dans le cas d'une série à valeurs classées - 3

Ceci nous amène directement à $q_{\frac{1}{4}}(x) = \frac{20+170}{\frac{15}{10}} = \frac{380}{3} \approx 126.67$.

Calcul dans le cas d'une série à valeurs classées - 3

Ceci nous amène directement à $q_{\frac{1}{4}}(x) = \frac{20+170}{\frac{15}{10}} = \frac{380}{3} \approx 126.67$.

Exercice

Calculer le septième décile de la série x définie dans le Tableau.

Variance

Définition

La variance de la série statistique simple x est la moyenne des carrés des écarts des éléments de la série à leur moyenne arithmétique $\overline{x}$:

$$s_x^2 := \frac{1}{n} \sum_{i=1}^{n} (x[i] - \overline{x})^2$$

Variance

Définition

La variance de la série statistique simple x est la moyenne des carrés des écarts des éléments de la série à leur moyenne arithmétique $\overline{x}$:

$$s_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x[i]-\overline{x})^2$$

Notation

La variance de la série x est notée s_x^2. Le s provient de l'anglais "standard deviation".

Écart-type

Définition

On appelle écart-type de la série la racine carrée de la variance :

$$s_x := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x[i]-\overline{x})^2}\,.$$

Exemples

Exemple

On considère la série $x := (9, 11)$. Sa moyenne est 10 et sa variance est

$$s_x^2 = \frac{1}{2}\left[(9-10)^2 + (11-10)^2\right] = 1\,.$$

Exemples

Exemple

On considère la série $x := (9, 11)$. Sa moyenne est 10 et sa variance est

$$s_x^2 = \frac{1}{2}\left[(9-10)^2 + (11-10)^2\right] = 1\,.$$

Exemple

On considère la série $x := (5, 15)$. Sa moyenne est 10 et sa variance est

$$s_x^2 = \frac{1}{2}\left[(5-10)^2 + (15-10)^2\right] = 25\,.$$

Formule de calcul

Formule de calcul

Soit une série statistique simple $x := (x[1], \cdots, x[n])$. On note x^2 la série statistique simple : $x^2 := (x[1]^2, \cdots, x[n]^2)$. On a alors

$$s_x^2 = \overline{x^2} - \overline{x}^2 .$$

Formule de calcul

Formule de calcul

Soit une série statistique simple $x := (x[1], \cdots, x[n])$. On note x^2 la série statistique simple : $x^2 := (x[1]^2, \cdots, x[n]^2)$. On a alors

$$s_x^2 = \overline{x^2} - \overline{x}^2 .$$

Cette formule est à rapprocher de $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Calcul dans le cas d'une série à valeurs isolées

Proposition

Si l'on a une série statistique simple à valeurs isolées, la variance est égale à

$$s_x^2 = \frac{1}{n} \sum_{k=1}^{r} n_k \left(y_k - \overline{x}\right)^2 .$$

Calcul dans le cas d'une série à valeurs isolées

Proposition

Si l'on a une série statistique simple à valeurs isolées, la variance est égale à

$$s_x^2 = \frac{1}{n} \sum_{k=1}^{r} n_k \left(y_k - \overline{x}\right)^2 .$$

Ainsi, l'écart-type est :

$$s_x = \sqrt{\frac{1}{n} \sum_{k=1}^{r} n_k \left(y_k - \overline{x}\right)^2} .$$

Version corrigée

Définition

Parfois, on introduit aussi la variance dite corrigée :

$$\widetilde{s_x^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x[i]-\overline{x})^2 = \frac{n}{n-1}s_x^2\,.$$

Version corrigée

Définition

Parfois, on introduit aussi la variance dite corrigée :

$$\widetilde{s_x^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x[i]-\overline{x})^2 = \frac{n}{n-1}s_x^2\,.$$

L'intérêt de cette définition alternative prend tout son sens quand on fait de l'estimation ponctuelle ou par intervalles de confiance.

Calcul dans le cas des valeurs isolées : exemple

y_i	n_i (effectifs)	f_i (fréquence)	N_i	F_i
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Calcul dans le cas des valeurs isolées : exemple

y_i	n_i (effectifs)	f_i (fréquence)	N_i	F_i
39.25	2	$\frac{2}{15}$	2	$\frac{2}{15}$
39.50	4	$\frac{4}{15}$	6	$\frac{6}{15}$
39.75	2	$\frac{2}{15}$	8	$\frac{8}{15}$
40.00	4	$\frac{4}{15}$	12	$\frac{12}{15}$
40.25	2	$\frac{2}{15}$	14	$\frac{14}{15}$
40.50	1	$\frac{1}{15}$	15	$\frac{15}{15}$

Ici, $\overline{x} = 39.80$. La variance est $s_x^2 = 0.135$. Et l'écart-type est $s_x \approx 0.3674$. Quant à la variance corrigée, elle vaut $\widetilde{s_x^2} = \frac{80}{79} \times 0.135 \approx 0.1367$ et l'écart-type associé est alors $\widetilde{s_x} \approx 0.3697$.

Calcul dans le cas d'une série à valeurs classées - 1

Proposition

Pour une série à valeurs classées, la formule

Calcul dans le cas d'une série à valeurs classées - 1

Proposition

Pour une série à valeurs classées, la formule

$$s_x^2 = \sum_{k=1}^{s} f_k \left(\frac{z_{k-1} + z_k}{2} \right)^2 - \left(\sum_{k=1}^{n} f_k \frac{z_{k-1} + z_k}{2} \right)^2$$

est une bonne approximation de la réalité.

Calcul dans le cas d'une série à valeurs classées - 1

Proposition

Pour une série à valeurs classées, la formule

$$s_x^2 = \sum_{k=1}^{s} f_k \left(\frac{z_{k-1} + z_k}{2} \right)^2 - \left(\sum_{k=1}^{n} f_k \frac{z_{k-1} + z_k}{2} \right)^2 + \sum_{k=1}^{s} f_k \frac{(z_k - z_{k-1})^2}{12},$$

est une bonne approximation de la réalité.

Calcul dans le cas d'une série à valeurs classées - 2

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

Calcul dans le cas d'une série à valeurs classées - 2

$[z_{k-1}; z_k[$	n_k	f_k	N_k	F_k
$[100; 120[$	10	$\frac{10}{80}$	10	$\frac{10}{80}$
$[120; 130[$	15	$\frac{15}{80}$	25	$\frac{25}{80}$
$[130; 140[$	25	$\frac{25}{80}$	50	$\frac{50}{80}$
$[140; 150[$	20	$\frac{20}{80}$	70	$\frac{70}{80}$
$[150; 170[$	10	$\frac{10}{80}$	80	$\frac{80}{80} = 1$

On a déjà vu que l'on avait $\overline{x} = 135.625$. Alors, la variance est $s_x^2 = 214.1927$. Ainsi, l'écart-type vaut $s_x \approx 14.6353$. Ici, la variance intra-classe vaut 14.58 soit plus de 6.8% de la variance totale. Quant à la variance corrigée, elle vaut $\widetilde{s_x^2} \approx 216.9040$ et l'écart-type associé est $\widetilde{s_x} \approx 14.7277$.

Coefficient de dispersion

Le coefficient de dispersion mesure la dispersion de la série en tenant compte de sa moyenne arithmétique. Il s'agit de

$$\frac{s_X}{\overline{X}}.$$

Ce coefficient ne dépend pas de l'unité de mesure choisie.

Coefficient de dispersion

Le coefficient de dispersion mesure la dispersion de la série en tenant compte de sa moyenne arithmétique. Il s'agit de

$$\frac{s_X}{\overline{X}}.$$

Ce coefficient ne dépend pas de l'unité de mesure choisie.

L'intérêt est qu'en renormalisant par la moyenne, l'ordre de grandeur n'influera pas. En effet, un écart-type de 1 pour une série ayant une moyenne de 10 correspond à un écart de 10% tandis que pour une moyenne de 1000, ça ne correspond qu'à 0.1%.

Néanmoins, quand la moyenne est proche de zéro, ce coefficient tend à être très grand et donc d'autant plus sensible aux variations de la moyenne. De plus, il ne sert pas pour la construction des intervalles de confiance.

L'étendue

Définition

On appelle étendue ("range" en anglais) de la série statistique simple x la différence entre le plus grand élément de la série et le plus petit.

L'étendue

Définition

On appelle étendue ("range" en anglais) de la série statistique simple x la différence entre le plus grand élément de la série et le plus petit.

Exemple avec une série à valeurs isolées

Dans le cas de la série définie précédemment, l'étendue est $40.50 - 39.25 = 1.25$.

L'étendue

Définition

On appelle étendue ("range" en anglais) de la série statistique simple x la différence entre le plus grand élément de la série et le plus petit.

Exemple avec une série à valeurs isolées

Dans le cas de la série définie précédemment, l'étendue est $40.50 - 39.25 = 1.25$.

Exemple avec une série à valeurs classées

Dans le cas de la série définie précédemment, l'étendue est $170 - 110 = 60$.

L'étendue

Définition

On appelle étendue ("range" en anglais) de la série statistique simple x la différence entre le plus grand élément de la série et le plus petit.

Exemple avec une série à valeurs isolées

Dans le cas de la série définie précédemment, l'étendue est $40.50 - 39.25 = 1.25$.

Exemple avec une série à valeurs classées

Dans le cas de la série définie précédemment, l'étendue est $170 - 110 = 60$.

Remarque

L'étendue est sensible aux valeurs aberrantes.

L'interquartile

Intervalle interquartile

On appelle intervalle interquartile de la série statistique simple $x := (x[1], \cdots, x[i], \cdots, x[n])$ l'intervalle semi-ouvert $[q_{\frac{1}{4}}(x); q_{\frac{3}{4}}(x)[$ où l'on rappelle que $q_{\frac{1}{4}}(x)$ (respectivement $q_{\frac{3}{4}}(x)$) est le premier quartile (respectivement le troisième quartile) de la série.

L'interquartile

Intervalle interquartile

On appelle intervalle interquartile de la série statistique simple $x := (x[1], \cdots, x[i], \cdots, x[n])$ l'intervalle semi-ouvert $[q_{\frac{1}{4}}(x); q_{\frac{3}{4}}(x)[$ où l'on rappelle que $q_{\frac{1}{4}}(x)$ (respectivement $q_{\frac{3}{4}}(x)$) est le premier quartile (respectivement le troisième quartile) de la série.

Définition : Interquartile

On appelle interquartile de la série x la longueur de l'intervalle interquartile.

L'interquartile

Intervalle interquartile

On appelle intervalle interquartile de la série statistique simple $x := (x[1], \cdots, x[i], \cdots, x[n])$ l'intervalle semi-ouvert $[q_{\frac{1}{4}}(x); q_{\frac{3}{4}}(x)[$ où l'on rappelle que $q_{\frac{1}{4}}(x)$ (respectivement $q_{\frac{3}{4}}(x)$) est le premier quartile (respectivement le troisième quartile) de la série.

Définition : Interquartile

On appelle interquartile de la série x la longueur de l'intervalle interquartile.

L'intervalle interquartile contient 50% des éléments de la série.

Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
4. Séries statistiques simples : caractéristiques
5. Séries statistiques doubles
   - Définitions
   - Tableau de contingence
   - Fréquence conditionnelle
6. Indépendance de deux caractères

Définitions

On mesure deux caractères quantitatifs x et y sur un ensemble de n individus.

Définitions

On mesure deux caractères quantitatifs x et y sur un ensemble de n individus.

Soit $x[i]$ la mesure de x pour l'individu i et $y[i]$ celle de y.

Définition

La suite $((x[1], y[1]), \cdots, (x[i], y[i]), \cdots, (x[n], y[n]))$ est appelée une série statistique double.

Définitions

On mesure deux caractères quantitatifs x et y sur un ensemble de n individus.

Soit $x[i]$ la mesure de x pour l'individu i et $y[i]$ celle de y.

Définition

La suite $((x[1], y[1]), \cdots, (x[i], y[i]), \cdots, (x[n], y[n]))$ est appelée une série statistique double.

Marginales

Les deux séries statistiques simples x et y avec $x := (x[1], \cdots, x[n])$ et $y := (y[1], \cdots, y[n])$ sont appelées les séries statistiques marginales de la série statistique double. On dit que x est la première marginale et que y est la deuxième.

Exemple

Exemple

On considère l'ensemble des habitants de la France. Alors, si $x[i]$ est la taille de l'individu i et si $y[i]$ est le poids de ce même individu, la série x est la première marginale de (x, y) tandis que y en est la deuxième marginale.

Tableau de contingence - 1

On présente la série statistique double dans un tableau de contingence.

Tableau de contingence - 1

On présente la série statistique double dans un tableau de contingence.

Pour ce faire, on commence par grouper les données.

Tableau de contingence - 2

On donne chaque série marginale sous la forme d'une série à valeurs isolées ou une série à valeurs classées. On suppose que $a_1, \cdots, a_k, \cdots, a_r$ sont les r modalités de la série x. Et, $b_1, \cdots, b_l, \cdots, b_s$ sont les s modalités de la série y. Ici, les modalités a_k et b_l représentent une valeur isolée ou une classe. On dit qu'un individu a la modalité a_k pour le caractère x lorsque :

- la valeur de x pour cet individu est égale à la valeur isolée représentée par a_k,
- la valeur de x appartient à la classe représentée par a_k.

Tableau de contingence - 3

On fait de même avec y et b_l pour tout $l \in [\![1; s]\!]$.
À l'intersection de la ligne a_k et de la colonne b_l, on porte le nombre total $n_{k,l}$ d'individus qui ont la modalité a_k pour x et la modalité b_l pour y. On l'appelle effectif de (a_k, b_l).

Tableau de contingence - 4

x \ y	b_1	$\cdots$	b_l	$\cdots$	b_s
a_1	$n_{1,1}$	$\cdots$	$n_{1,l}$	$\cdots$	$n_{1,s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	$n_{k,1}$	$\cdots$	$n_{k,l}$	$\cdots$	$n_{k,s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	$n_{r,1}$	$\cdots$	$n_{r,l}$	$\cdots$	$n_{r,s}$

Tableau de contingence - 5

La somme des effectifs sur une ligne est importante :

$$n_{k,\bullet} := \sum_{l=1}^{s} n_{k,l} \, .$$

Tableau de contingence - 5

La somme des effectifs sur une ligne est importante :

$$n_{k,\bullet} := \sum_{l=1}^{s} n_{k,l} \, .$$

Alors, $n_{k,\bullet}$ représente le nombre d'individus qui ont la modalité a_k de x. On l'appelle effectif marginal de la modalité a_k.

Tableau de contingence - 6

De même, on somme sur une colonne :

$$n_{\bullet,l} := \sum_{k=1}^{r} n_{k,l} \, .$$

Tableau de contingence - 6

De même, on somme sur une colonne :

$$n_{\bullet,l} := \sum_{k=1}^{r} n_{k,l} \, .$$

Alors, $n_{\bullet,l}$ représente le nombre d'individus qui ont la modalité b_l de y. On l'appelle effectif marginal de la modalité b_l. On a les égalités suivantes :

$$\sum_{k=1}^{r} n_{k,\bullet} = \sum_{l=1}^{s} n_{\bullet,l} = \sum_{k=1}^{r} \sum_{l=1}^{s} n_{k,l} = \sum_{l=1}^{s} \sum_{k=1}^{r} n_{k,l} = n \, .$$

Tableau de contingence - 7

On peut définir la fréquence du couple (a_k, b_l) comme on le fit pour les séries statistiques simples :

$$f_{k,l} := \frac{n_{k,l}}{n}.$$

C'est la proportion d'individus de la population qui ont la modalité a_k de x et la modalité b_l de y. On introduit également

$$f_{k,\bullet} := \frac{n_{k,\bullet}}{n} \quad \text{et} \quad f_{\bullet,l} := \frac{n_{\bullet,l}}{n},$$

pour tout $k \in [\![1; r]\!]$ et pour tout $l \in [\![1; s]\!]$. $f_{k,\bullet}$ est la fréquence marginale de la modalité a_k : c'est la proportion d'éléments de la population qui ont la modalité a_k de x. De même, $f_{\bullet,l}$ est la fréquence marginale de la modalité b_l : c'est la proportion d'éléments de la population qui ont la modalité b_l de y.

Tableau de contingence - 8

Or, $f_{k,\bullet} = \sum_{l=1}^{s} n_{k,l}$ donc

$$f_{k,\bullet} := \sum_{l=1}^{s} f_{k,l} \quad \text{et de même} \quad f_{\bullet,l} := \sum_{k=1}^{r} f_{k,l} \, .$$

Tableau de contingence - 8

Or, $f_{k,\bullet} = \sum_{l=1}^{s} n_{k,l}$ donc

$$f_{k,\bullet} := \sum_{l=1}^{s} f_{k,l} \quad \text{et de même} \quad f_{\bullet,l} := \sum_{k=1}^{r} f_{k,l}\,.$$

Puis, l'on a

$$\sum_{k=1}^{r} f_{k,\bullet} = \sum_{l=1}^{s} f_{\bullet,l} = \sum_{k=1}^{r}\sum_{l=1}^{s} f_{k,l} = \sum_{l=1}^{s}\sum_{k=1}^{r} f_{k,l} = 1\,.$$

Tableau de contingence - 9

On obtient ainsi le tableau de contingence suivant :

x \ y	b_1	$\cdots$	b_l	$\cdots$	b_s	
a_1	$n_{1,1}$	$\cdots$	$n_{1,l}$	$\cdots$	$n_{1,s}$	$n_{1,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	$n_{k,1}$	$\cdots$	$n_{k,l}$	$\cdots$	$n_{k,s}$	$n_{k,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	$n_{r,1}$	$\cdots$	$n_{r,l}$	$\cdots$	$n_{r,s}$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$\cdots$	$n_{\bullet,l}$	$\cdots$	$n_{\bullet,s}$	n

Fréquence conditionnelle - 1

Définition

On pose : $f(b_l \mid a_k) := \frac{f_{k,l}}{f_{k,\bullet}} = \frac{n_{k,l}}{n_{k,\bullet}}$. Cette quantité est appelée fréquence conditionnelle de la modalité b_l par rapport à la modalité a_k.

Fréquence conditionnelle - 1

Définition

On pose : $f(b_l \mid a_k) := \frac{f_{k,l}}{f_{k,\bullet}} = \frac{n_{k,l}}{n_{k,\bullet}}$. Cette quantité est appelée fréquence conditionnelle de la modalité b_l par rapport à la modalité a_k.

C'est aussi la proportion d'individus qui ont la modalité b_l parmi les $n_{k,\bullet}$ individus qui ont la modalité a_k. C'est à rapprocher de $\mathbb{P}(Y = b_l \mid X = a_k)$.

Fréquence conditionnelle - 2

Remarque

On a ainsi une suite de s valeurs positives telles que leur somme vaut

$$\sum_{l=1}^{s} \frac{f_{k,l}}{f_{k,\bullet}} = \frac{\sum_{l=1}^{s} f_{k,l}}{f_{k,\bullet}} = \frac{f_{k,\bullet}}{f_{k,\bullet}} = 1\,.$$

Fréquence conditionnelle - 2

Remarque

On a ainsi une suite de s valeurs positives telles que leur somme vaut

$$\sum_{l=1}^{s} \frac{f_{k,l}}{f_{k,\bullet}} = \frac{\sum_{l=1}^{s} f_{k,l}}{f_{k,\bullet}} = \frac{f_{k,\bullet}}{f_{k,\bullet}} = 1\,.$$

Il y a donc bien une loi de probabilité sous-jacente.

Fréquence conditionnelle - 3

Définition

L'ensemble des couples $(b_l, f(b_l \mid a_k))$ s'appelle la distribution conditionnelle du caractère y par rapport à la modalité a_k de x.

Fréquence conditionnelle - 3

Définition

L'ensemble des couples $(b_l, f(b_l \mid a_k))$ s'appelle la distribution conditionnelle du caractère y par rapport à la modalité a_k de x.

Remarque

Pour simplifier l'écriture, on fera l'abus de notation :

$$f(b_1 \mid a_k)\,\delta_{b_1} + \cdots + f(b_l \mid a_k)\,\delta_{b_l} + \cdots + f(b_s \mid a_k)\,\delta_{b_s}\,.$$

Fréquence conditionnelle - 4

De la même manière, on définit ensuite $f(a_k \mid b_l)$ comme étant le rapport $\frac{f_{k,l}}{f_{\bullet,l}} = \frac{n_{k,l}}{n_{\bullet,l}}$, appelée fréquence conditionnelle de la modalité a_k par rapport à la modalité b_l : Il s'agit de la proportion d'individus qui ont la modalité a_k parmi les $n_{\bullet,l}$ individus qui ont la modalité b_l.

Fréquence conditionnelle - 4

De la même manière, on définit ensuite $f(a_k \mid b_l)$ comme étant le rapport $\frac{f_{k,l}}{f_{\bullet,l}} = \frac{n_{k,l}}{n_{\bullet,l}}$, appelée fréquence conditionnelle de la modalité a_k par rapport à la modalité b_l : Il s'agit de la proportion d'individus qui ont la modalité a_k parmi les $n_{\bullet,l}$ individus qui ont la modalité b_l.

Définition

L'ensemble des couples $(a_k, f(a_k \mid b_l))$ s'appelle la distribution conditionnelle du caractère x par rapport à la modalité b_l de y.

Plan

1. Vocabulaire
2. Caractère qualitatif
3. Séries statistiques simples : présentation
4. Séries statistiques simples : caractéristiques
5. Séries statistiques doubles
6. Indépendance de deux caractères
   - Nuage de points
   - Covariance
   - Description empirique de la dépendance

Définition

Définition

On dit que le caractère y est indépendant du caractère x dans l'ensemble I lorsque la distribution conditionnelle du caractère y par rapport à la modalité a_k de x est la même pour tout $k \in [\![1; r]\!]$.

Définition

Définition

On dit que le caractère y est indépendant du caractère x dans l'ensemble I lorsque la distribution conditionnelle du caractère y par rapport à la modalité a_k de x est la même pour tout $k \in [\![1; r]\!]$.

En d'autres termes, pour tout $k \in [\![1; r]\!]$, cette distribution conditionnelle est égale à celle de y par rapport à la modalité a_1 de x :

$$\sum_{l=1}^{s} f\left(b_l \mid a_k\right) \delta_{b_l} = \sum_{l=1}^{s} f\left(b_l \mid a_1\right) \delta_{b_l} .$$

Définition

Définition

On dit que le caractère y est indépendant du caractère x dans l'ensemble I lorsque la distribution conditionnelle du caractère y par rapport à la modalité a_k de x est la même pour tout $k \in [\![1; r]\!]$.

En d'autres termes, pour tout $k \in [\![1; r]\!]$, cette distribution conditionnelle est égale à celle de y par rapport à la modalité a_1 de x :

$$\sum_{l=1}^{s} f\left(b_l \mid a_k\right) \delta_{b_l} = \sum_{l=1}^{s} f\left(b_l \mid a_1\right) \delta_{b_l} .$$

Par conséquent, l'indépendance du caractère y par rapport au caractère x peut se traduire par l'égalité $f\left(b_l \mid a_k\right) = f\left(b_l \mid a_1\right)$ pour tout $k \in [\![1; r]\!]$ et pour tout $l \in [\![1; s]\!]$.

Exemple

x \ y	b_1	b_2	b_3	b_4	
a_1	2	4	5	1	12
a_2	4	8	10	2	24
	6	12	15	3	36

Exemple

x \ y	b_1	b_2	b_3	b_4	
a_1	2	4	5	1	12
a_2	4	8	10	2	24
	6	12	15	3	36

L'égalité des distributions conditionnelles de y par rapport à x est immédiate. On en déduit que y est indépendante de x dans cette population.

Résultats classiques - 1

Théorème

La relation d'indépendance est symétrique. Si y est indépendant de x alors x est indépendant de y. On dit d'ailleurs que x et y sont indépendants.

Résultats classiques - 1

Théorème

La relation d'indépendance est symétrique. Si y est indépendant de x alors x est indépendant de y. On dit d'ailleurs que x et y sont indépendants.

En effet :

Théorème

Les caractères x et y sont indépendants si et seulement si $f_{k,l} = f_{k,\bullet}\, f_{\bullet,l}$ pour tout $k \in [\![1; r]\!]$ et pour tout $l \in [\![1; s]\!]$.

Résultats classiques - 2

Preuve

On a déjà vu que l'indépendance de y par rapport à x signifie que l'on a

$$f\left(b_l \mid a_k\right) = f\left(b_l \mid a_1\right),$$

pour tout $k \in [\![1; r]\!]$ et pour tout $l \in [\![1; s]\!]$. On en déduit l'égalité

$$\frac{f_{k,l}}{f_{k,\bullet}} = \frac{f_{1,l}}{f_{1,\bullet}}.$$

Résultats classiques - 3

Puis, l'on a :

$$\frac{f_{k,l}}{f_{k,\bullet}} = \frac{f_{k',l}}{f_{k',\bullet}},$$

pour tout $1 \leq k, k' \leq r$ et tout $1 \leq l \leq s$. Ainsi :

$$f_{k,l} f_{k',\bullet} = f_{k,\bullet} f_{k',l}.$$

On fait la somme pour k' allant de 1 à r et l'on trouve :

$$f_{k,l} \sum_{k'=1}^{r} f_{k',\bullet} = f_{k,\bullet} \sum_{k'=1}^{r} f_{k',l}.$$

Résultats classiques - 44

Or, $\sum_{k'=1}^{r} f_{k',\bullet} = 1$ et $\sum_{k'=1}^{r} f_{k',l} = f_{\bullet,l}$; ce qui achève la preuve.

Remarque cruciale

Il est important de comprendre que cette notion d'indépendance est extrêmement restrictive. En effet, si l'on a un échantillon de taille n finie, même si les deux variables aléatoires sous-jacentes sont bien indépendantes, il est quasiment impossible que les égalités soient vérifiées.

Remarque cruciale

Il est important de comprendre que cette notion d'indépendance est extrêmement restrictive. En effet, si l'on a un échantillon de taille n finie, même si les deux variables aléatoires sous-jacentes sont bien indépendantes, il est quasiment impossible que les égalités soient vérifiées.

De fait, dans la pratique, il faut effectuer un test d'indépendance pour vérifier si l'on peut conclure ou non sur l'indépendance des variables aléatoires sous-jacentes.

Nuage de points - 1

Soit une série statistique double $((x[1], y[1]), \cdots, (x[n], y[n]))$ sur une population de n individus.

Nuage de points - 1

Soit une série statistique double $((x[1], y[1]), \cdots, (x[n], y[n]))$ sur une population de n individus.

On porte dans un système d'axes orthogonaux les points dont les coordonnées sont $(x[i], y[i])$.

Nuage de points - 1

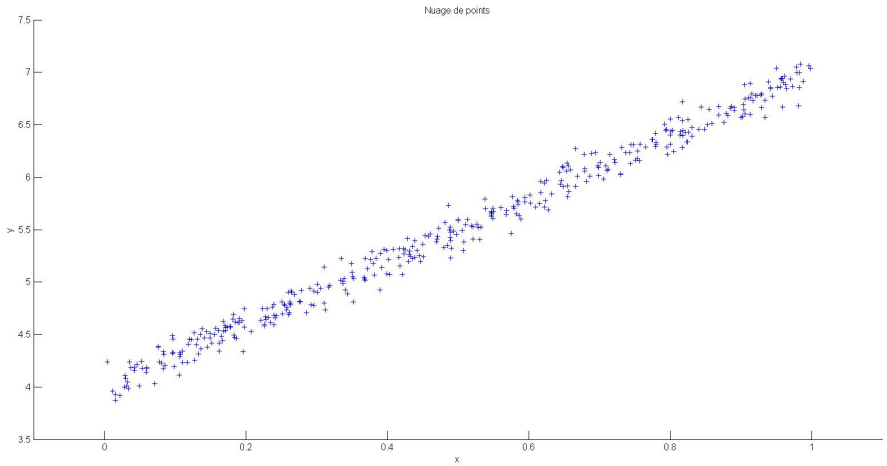
Soit une série statistique double $((x[1], y[1]), \cdots, (x[n], y[n]))$ sur une population de n individus.

On porte dans un système d'axes orthogonaux les points dont les coordonnées sont $(x[i], y[i])$.

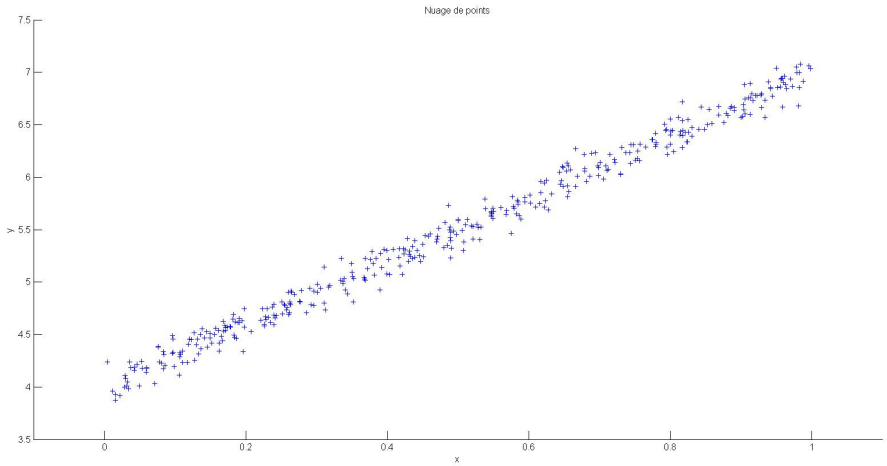
L'un des intérêts de ce nuage de points est de vérifier, à l'œil nu s'il est pertinent ou non d'effectuer une régression linéaire.

Nuage de points - 2

Nuage de points - 2



Nuage de points - 2



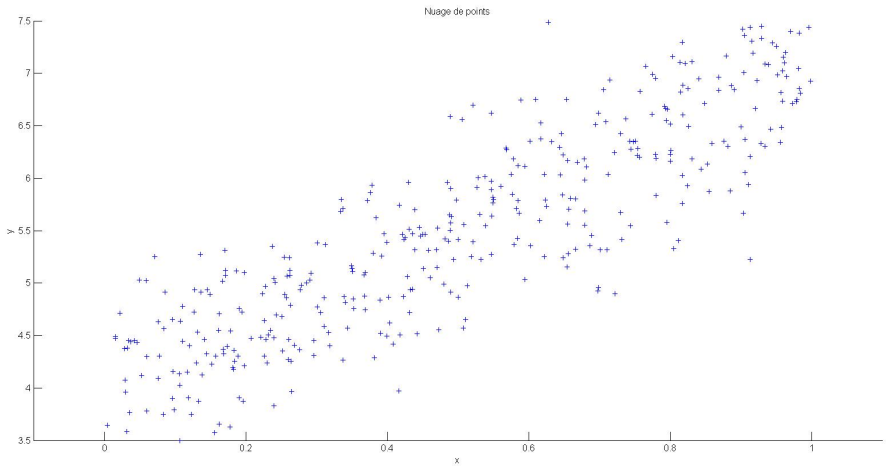
Dans ce nuage, il est assez clair qu'une direction se précise et que ce nuage peut être approché par une droite. Ainsi, la régression linéaire sera pertinente.

Nuage de points - 3

Voici un deuxième nuage de points :

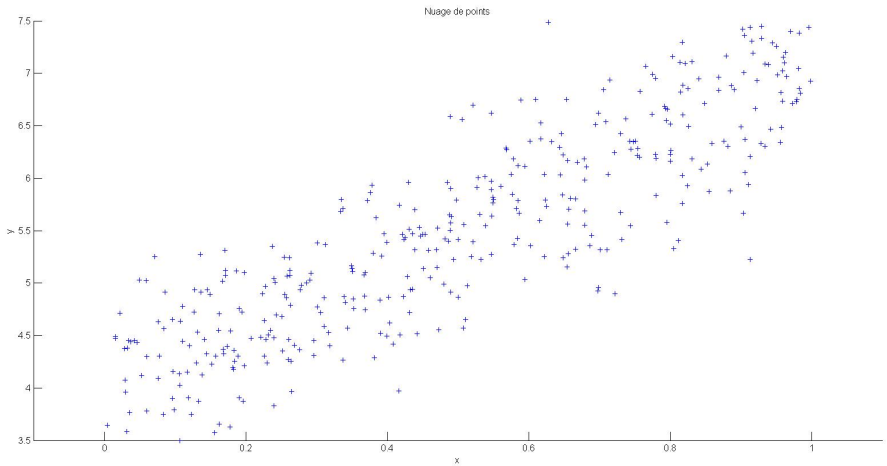
Nuage de points - 3

Voici un deuxième nuage de points :



Nuage de points - 3

Voici un deuxième nuage de points :



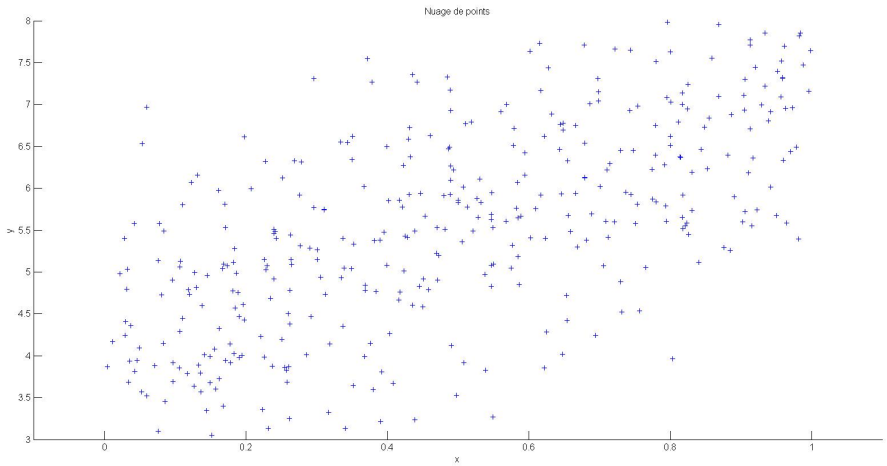
Dans ce nuage, la direction est moins nette mais l'on peut distinguer une certaine droite qui, par ailleurs, aurait la même équation que la précédente.

Nuage de points - 4

Voici un troisième nuage de points :

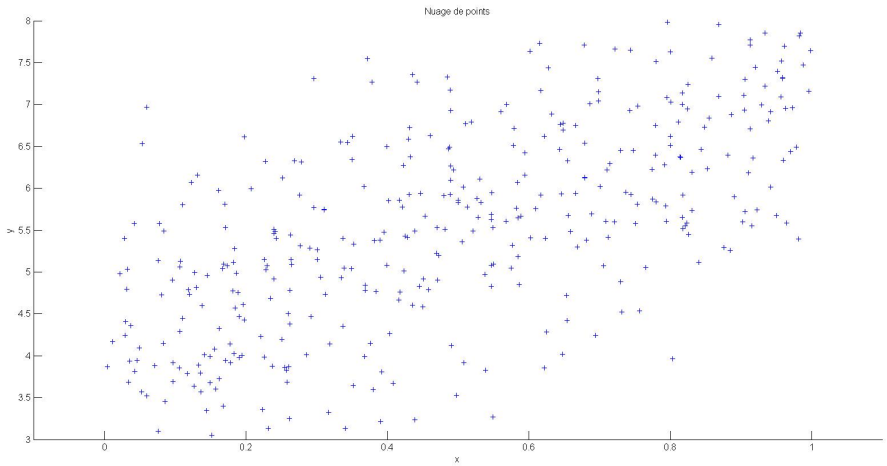
Nuage de points - 4

Voici un troisième nuage de points :



Nuage de points - 4

Voici un troisième nuage de points :



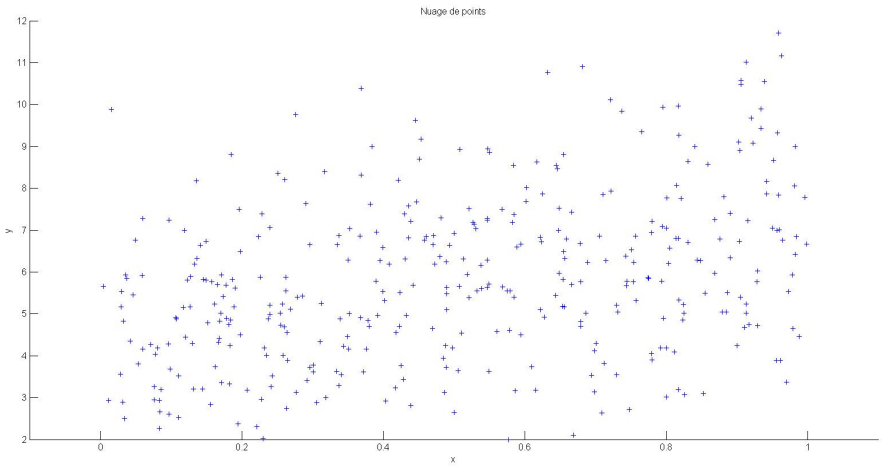
L'existence ou non d'une direction privilégiée est encore moins perceptible.

Nuage de points - 5

Voici un quatrième nuage de points :

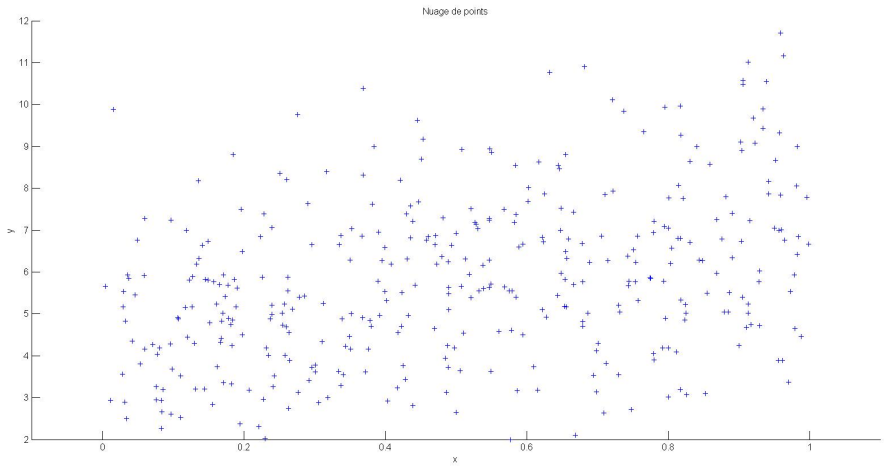
Nuage de points - 5

Voici un quatrième nuage de points :



Nuage de points - 5

Voici un quatrième nuage de points :



Cette fois, il est difficilement envisageable qu'une régression linéaire apporte quoi que ce soit de pertinent.

Nuage de points - 6

Remarque

Pour réaliser ces nuages, on a simulé 400 variables aléatoires suivant la loi uniforme sur $[0; 1]$. Ceci nous donne la série x. Puis, pour y, on a posé :

$$y[i] := ax[i] + b + \rho\epsilon[i]\,,$$

Nuage de points - 6

Remarque

Pour réaliser ces nuages, on a simulé 400 variables aléatoires suivant la loi uniforme sur $[0; 1]$. Ceci nous donne la série x. Puis, pour y, on a posé :

$$y[i] := ax[i] + b + \rho\epsilon[i]\,,$$

où ϵ_i suit une loi normale centrée réduite. Quant à ρ, la valeur du bruit, elle a été prise égale à 0.1 dans le premier nuage, à 0.5 dans le deuxième, à 1 dans le troisième et à 2 dans le quatrième.

Covariance - 1

La covariance est une quantité essentielle pour mesurer la corrélation linéaire entre deux caractères quantitatifs mesurés sur une même population.

Covariance - 1

La covariance est une quantité essentielle pour mesurer la corrélation linéaire entre deux caractères quantitatifs mesurés sur une même population.

Définition

La covariance entre x et y est

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x[i]-\overline{x})\,(y[i]-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n}x[i]y[i] - \overline{xy}\,.$$

Covariance - 1

La covariance est une quantité essentielle pour mesurer la corrélation linéaire entre deux caractères quantitatifs mesurés sur une même population.

Définition

La covariance entre x et y est

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x[i]-\overline{x})\,(y[i]-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n}x[i]y[i] - \overline{xy}\,.$$

En ce qui concerne le calcul pratique, si x et y sont deux séries à valeurs isolées, le calcul donne simplement

$$s_{xy} = \sum_{k=1}^{r}\sum_{l=1}^{s} f_{k,l}\,a_k\,b_l - \overline{xy}\,.$$

Covariance - 2

Si x est à valeurs classées alors que y est à valeurs isolées, la covariance sera approchée par

$$s_{xy} = \sum_{k=1}^{r} \sum_{l=1}^{s} f_{k,l} \frac{z_{k-1} + z_k}{2} b_l - \overline{xy}\,.$$

Covariance - 2

Si x est à valeurs classées alors que y est à valeurs isolées, la covariance sera approchée par

$$s_{xy} = \sum_{k=1}^{r} \sum_{l=1}^{s} f_{k,l} \frac{z_{k-1} + z_k}{2} b_l - \overline{xy}\,.$$

Enfin, dans le cas où x et y sont à valeurs classées, supposer l'indépendance des perturbations de x avec celles de y conduit au résultat suivant :

$$s_{xy} = \sum_{k=1}^{r} \sum_{l=1}^{s} f_{k,l} \frac{z_{k-1} + z_k}{2} \frac{\widetilde{z_{l-1}} + \widetilde{z}_l}{2} - \overline{xy}\,.$$

Courbe des moyennes conditionnelles

On suppose que le caractère x a une action sur le caractère y. Pour les individus ayant la modalité a_k du caractère x, on calcule la moyenne des valeurs du caractère y, notée $\overline{y_k}$, appelée moyenne conditionnelle de y pour x ayant la modalité a_k.

Exemple

Courbe des moyennes conditionnelles

On suppose que le caractère x a une action sur le caractère y. Pour les individus ayant la modalité a_k du caractère x, on calcule la moyenne des valeurs du caractère y, notée $\overline{y_k}$, appelée moyenne conditionnelle de y pour x ayant la modalité a_k.

Exemple

On prend un Tableau (poids/taille). On calcule la moyenne conditionnelle du poids pour les individus de taille dans $[150; 160[$:

taille \ poids	$[45; 55[$	$[55; 65[$	$[65; 75[$	$[75; 85[$
$[150; 160[$	2	7	5	2

Courbe des moyennes conditionnelles

On suppose que le caractère x a une action sur le caractère y. Pour les individus ayant la modalité a_k du caractère x, on calcule la moyenne des valeurs du caractère y, notée $\overline{y_k}$, appelée moyenne conditionnelle de y pour x ayant la modalité a_k.

Exemple

On prend un Tableau (poids/taille). On calcule la moyenne conditionnelle du poids pour les individus de taille dans $[150; 160[$:

taille \ poids	$[45; 55[$	$[55; 65[$	$[65; 75[$	$[75; 85[$
$[150; 160[$	2	7	5	2

On a alors $\overline{y_1} = \frac{2\times50+7\times60+5\times70+2\times80}{16} = 64.375$. Pour obtenir la courbe des moyennes conditionnelles, on joint par des segments de droite les points de coordonnées $(\xi_k, \overline{y_k})$ où ξ_k est la valeur isolée en x ou le milieu de la classe de x. Dans le cas présent, on joint les points $(155, 64.375)$, $(165, 74.89)$, $(175, 81.64)$, $(185, 88.69)$ et $(195, 91.48)$.