Mathématiques

-

Statistiques descriptives

Ajustements linéaires et non-linéaires

Statistiques inférentielles

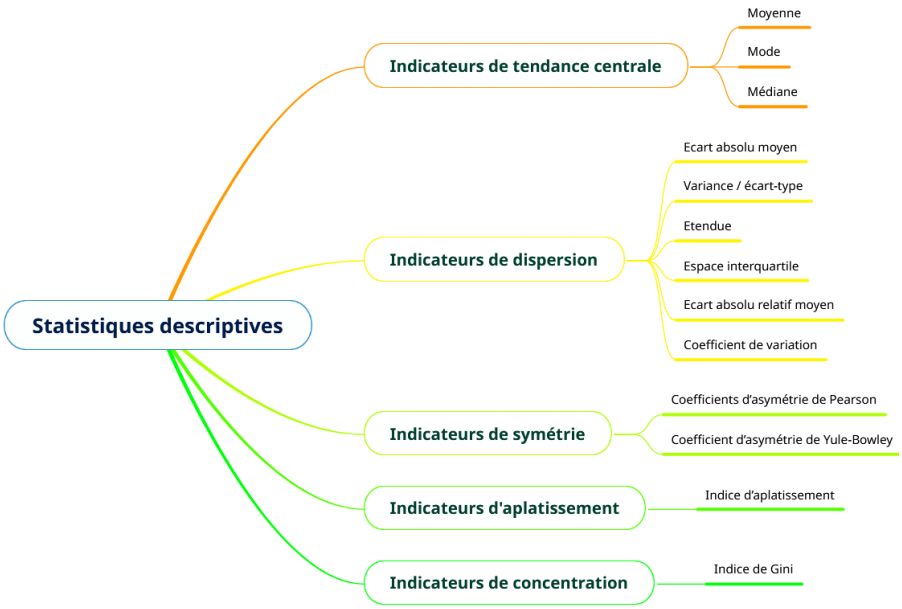
Frédéric Menan

fmenan@cesi.fr

02/2025

Table des matières

1 Statistiques descriptives et classification



La statistique descriptive sert à décrire une population à l'aide d'indicateurs numériques ou de techniques graphiques.

En statistique descriptive, le caractère statistique étudié est connu pour tous les individus de la population. On n'a donc pas à « estimer » la moyenne ou la variance par exemple, ils peuvent être calculés exactement.

1.1 Définitions et Classification

1.1.1 Population et individu

L'individu est l'unité statistique à laquelle on s'intéresse.

La population P est l'ensemble fini des individus statistiques que l'on souhaite décrire.

La taille de la population est le nombre N d'individus distincts de la population.

1.1.2 Variable (ou caractère statistique)

Une variable statistique est un moyen de décrire chacun des individus de la population.

1.1.3 Modalité

Une modalité d'une variable est une des façons possibles d'effectuer la description d'un individu au moyen de cette variable.

L'ensemble des modalités d'une variable est l'ensemble des différentes façons possibles de décrire les individus de la population avec la variable.

Soit X une variable statistique. L'ensemble de ses modalités sera noté M_X. Un élément de cet ensemble sera noté x.

On peut maintenant donner la définition mathématique d'une variable statistique : une variable statistique X est une fonction définie sur P (ensemble de départ) à valeurs dans M_X (ensemble d'arrivée) représentée par

$$\begin{cases} X: P \to M_X \\ i \to X(i) \end{cases}$$

où X(i) est l'élément de M_X qui désigne la modalité de la variable X présentée par l'individu i de P.

1.1.4 Classification des variables statistiques

1.1.4.1 Classification 1 : variables continues, variables discrètes

Une variable X est discrète si M_X est un ensemble dénombrable, càd si X possède un nombre dénombrable de modalités.

Une variable X est une variable continue si M_X est un ensemble non-dénombrable. Pour une variable continue, si on choisit deux modalités, alors toutes les valeurs comprises entre ces deux modalités sont aussi des modalités de la variable.

1.1.4.2 Classification 2 : variables qualitatives, variables quantitatives

Une variable est dite qualitative si ses modalités ne sont pas quantifiables.

Une variable est dite quantitative si ses modalités peuvent être considérées comme des quantités exprimées dans une échelle de valeurs.

1.1.4.3 Limites des classifications usuelles

1/ Il n'existe en pratique aucune variable continue.

Pour une variable continue, la précision avec laquelle les modalités de cette variable sont relevées rend la variable discrète. Par exemple, si on s'intéresse à l'épaisseur des livres d'une bibliothèque, l'instrument avec lequel on évalue cette épaisseur aura un degré de précision qui limitera les modalités possibles de cette variable. L'ensemble des modalités (épaisseurs) que l'on peut observer (mesurer) est alors discret.

2/ Ce n'est pas parce que les modalités sont mesurables que la variable présente les propriétés d'une variable quantitative.

La variable qui décrit les dimensions des livres de la bibliothèque a des modalités x = (h ; L ; e) qui ne présentent pas les propriétés usuelles des variables quantitatives : il n'y a pas d'ordre naturel sur ces modalités. Ainsi peut-on dire qu'un livre présentant la modalité (21 ; 15 ; 2) est plus grand qu'un livre présentant la modalité (14 ; 22,5 ; 2) ?

Les classifications 1 et 2 montrent donc des limites. La classification 3 est plus pertinente.

1.1.4.4 Classification 3 : classification par opérations statistiques possibles

Variable nominale : X est une variable nominale s'il n'est pas possible de définir un ordre sur l'ensemble M_X de ses modalités.

Variable ordinale : X est une variable ordinale si

1. il est possible de définir un ordre sur l'ensemble M_X des ses modalités ;
2. les écarts et les relations de proportionnalité entre modalités de X n'ont pas de sens

Variable numérique : X est une variable numérique si

1. il est possible de définir un ordre sur l'ensemble M_X des ses modalités ;
2. les écarts et les relations de proportionnalité entre modalités de X ont un sens

1.1.4.5 Traitements statistiques possibles avec les variables de la classification 3

Variable	Ordre	Ecarts Proportionnalité	Traitements statistiques possibles	Traitements non possibles
Nominale	NON	NON	Dénombrement (effectifs, fréquences, mode)	Cumul, quantiles Opérations usuelles
Ordinale	OUI	NON	Dénombrement (effectifs, fréquences, mode), Cumul, quantiles	Opérations usuelles
Numérique	OUI	OUI	Tous	Aucun

1.2 Organisation des données

1.3 Données brutes

Les données brutes sont l'ensemble des modalités observées pour chaque variable et chaque individu de la population.

Soit une population P de N individus décrite par trois variables X, Y et Z.

On observera pour chaque individu i de P, trois modalités, notées X(i), Y (i) et Z(i).

Les données brutes seront la liste de ces modalités écrites pour tous les individus : X(1), Y(1), Z(1),X(2), Y (2), Z(2),X(3), . . . , Z(N − 1),X(N), Y (N), Z(N).

1.4 Tableau élémentaire

On appelle tableau élémentaire le tableau qui à chaque individu de la population reporté en première colonne associe la modalité que présente cet individu pour chacune des variables.

Individu	Modalité de X	Modalité de Y	Modalité de Z
1	X(1)	Y(1)	Z(1)
2	X(2)	Y(2)	Z(2)

...	...	...	...
N	X(N)	Y(N)	Z(N)

1.5 Comptage

1.5.1 Comptage pour un petit nombre de modalités

1.5.1.1 Tri à plat et effectifs

Lors d'un « tri à plat », on dénombre, pour chaque modalité x_k de la variable X, les individus de la population qui présentent cette modalité.

On peut alors établir les effectifs des modalités de la variable et créer le tableau statistique.

L'effectif n_k de la modalité x_k est le cardinal de l'ensemble des individus de la population présentant la modalité x_k de la variable X :

$$n_k=\#\{i\in P \mid X(i)=x_k\}$$

1.5.1.2 Tableau statistique

Le tableau statistique est alors :

Modalité de X	Effectif de la modalité
x_1	n_1
x_2	n_2
…	…
x_k	n_k

1.5.2 Comptage pour un grand nombre de modalités

1.5.2.1 Regroupement par classes

On construit des classes de modalités notées C_1, C_2,...,C_j telles que

$$\forall j, \forall k, \text{si } j \neq k \text{ alors } C_j \cap C_k = \emptyset$$

$$C_1 \cup C_2 \cup \ldots\ldots \cup C_j = M_X$$

Remarque : pour tout x de Mx, il existe une et une seule classe contenant x.

Pour une variable numérique, les classes sont des intervalles de nombres réels :

$$C_k = [e_{k-1}, e_k[\text{ ou } C_k =]e_{k-1}, e_k] \text{ ou } C_k = [e_{k-1}, e_k] \text{ ou } C_k =]e_{k-1}, e_k[$$

Remarque : si l'on définit $C_k = [ek_{-1}, e_k[$, alors pour tout individu i, $C(i)=C_k \Leftrightarrow e_{k-1} \leq X(i) < e_k$

1.5.2.2 Tri à plat et effectifs

Dans le cas d'un regroupement par classes, le tri à plat consistera à dénombrer, pour chaque classe de modalités C_j de la variable X, les individus de la population présentant une modalité de la variable appartenant à C_j.

On définira ensuite l'effectif de la classe C_j. C'est le nombre nj défini par :

$$n_j=\#\{i\in P \mid X(i)\in C_j\}$$

1.5.2.3 Tableau statistique après regroupement par classes

Classe de modalités de X	Effectif de la classe	Centre de classe
C1	n1	x1
C2	n2	x2
...	...	...
Cj	nj	xj

1.5.2.4 Centre de classe

Lorsque les classes de modalités sont des intervalles d'extrémités e_0, e_1, . . . , e_J, on peut faire figurer dans le tableau statistique les centres de ces classes, que l'on note x_j, définis par

$$x_j = \frac{e_j + e_{j-1}}{2}$$

1.6 Fréquences & densités

1.6.1 Fréquence

On appelle fréquence de la modalité x_k de X le nombre f_k défini par : $f_k = \frac{n_k}{N}$

1.6.2 Distribution statistique

Les K couples (x_k, f_k) avec k = 1, . . . ,K, sont appelés la distribution statistique de la variable X dans la population P.

La distribution statistique d'une variable est le point de départ de toute analyse statistique.

1.6.3 Notion de densités

Source du problème : le regroupement par classes : l'effectif ou la fréquence d'une classe dépend de son amplitude. Sens à donner à l'effectif ou la fréquence d'une classe ?

On utilisera les densités d'effectif ou les densités de fréquence.

1.6.3.1 Définitions

L'amplitude a_k d'une classe de modalité $C_k = [e_{k-1} ; e_k [$ est définie par : $a_k = e_k - e_{k-1}$

On appelle densité d'effectif de la classe $C_k = [e_{k-1} ; e_k [$, le nombre d_k défini par

$$d_k = n_k / a_k$$

On appelle densité de fréquence de la classe $C_k = [e_{k-1} ; e_k [$ le nombre δ_k défini par

$$\delta_k = f_k / a_k$$

1.7 Effectifs et fréquences cumulés

1.7.1 Effectif cumulé croissant

On appelle effectif cumulé croissant de la modalité x_k de X, le nombre N_k défini par :

$$N_k = \sum_{j=1}^{k} n_j = n_1 + n_2 + \cdots + n_k$$

Si X est une variable ordinale ou numérique, N_k est le nombre d'individus présentant une modalité de X inférieure ou égale à x_k.

1.7.2 Fréquence cumulée croissante

On appelle fréquence cumulée croissante de la modalité x_k de X, le nombre noté F_k et défini par

$$F_k = \sum_{j=1}^{k} f_j = f_1 + f_2 + \cdots + f_k$$

1.7.3 Fonction de répartition

Si X est une variable numérique à modalités réelles, alors la fonction F_X telle que

$$\begin{cases} F_X : \mathrm{IR} \rightarrow [0 ; 1] \\ x \rightarrow F_X(x) = F_k \Leftrightarrow x_k \leq x < x_{k+1} \end{cases}$$

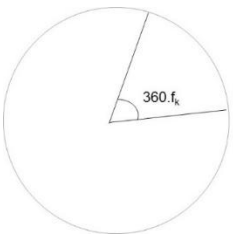
est appelée fonction de répartition de X.

1.8 Représentation graphique des données

1.8.1 Diagrammes en secteurs

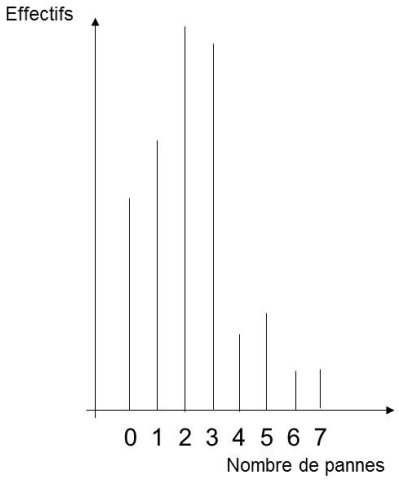
Pour la modalité x_k de fréquence f_k, la longueur de l'arc, en degrés, est

$$l_k = 360.f_k$$



1.8.2 Diagrammes en bâtons

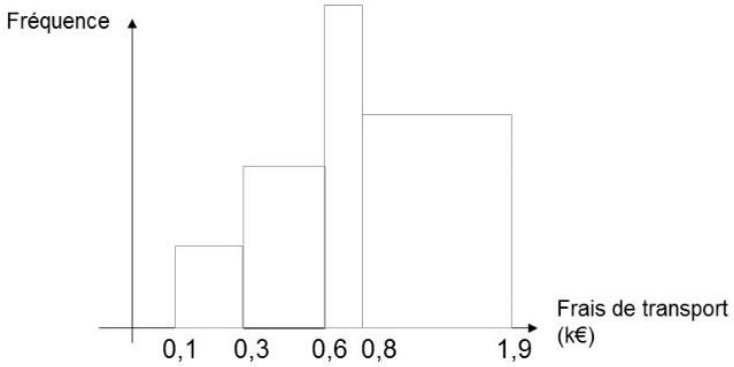
1/ Si pas de regroupement par classes



Nombre de pannes (modalités)	0	1	2	3	4	5	6	7
Nombre de machines (effectifs)	11	14	20	19	4	5	2	2

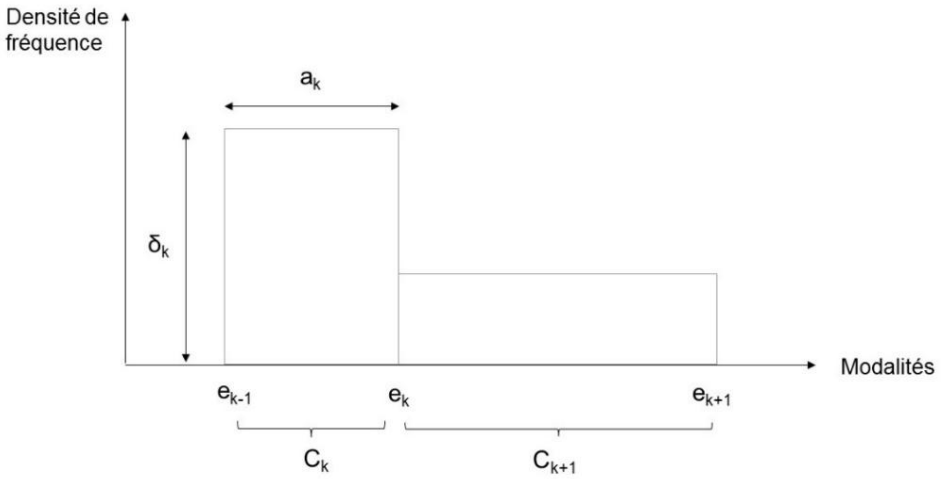
2/ Si regroupement par classes

Frais de transport (k€)	Fréquence
[0,1;0,3[	0,10
[0,3;0,6[	0,20
[0,6;0,8[	0,40
[0,8;1,9]	0,30



1.8.3 Histogramme des fréquences

Cet histogramme est utilisé pour les variables numériques dont les modalités ont été regroupées par classes. Il est plus pertinent que le diagramme précédent dans le cas de classes d'amplitude inégales.



L'histogramme des fréquences ne s'utilise que pour une variable numérique. Il présente un intérêt particulier par rapport au diagramme en bâtons quand les classes sont d'amplitudes inégales.

Surface de la classe à gauche : f_k ; Surface de la classe à droite : f_{k+1}

L'histogramme permet de lire la valeur du mode (voir suite du cours). La classe modale est la classe correspondant au rectangle de plus grande hauteur. Le mode est le centre de cette classe.

1.8.4 Fréquences cumulées croissantes : fonction de répartition

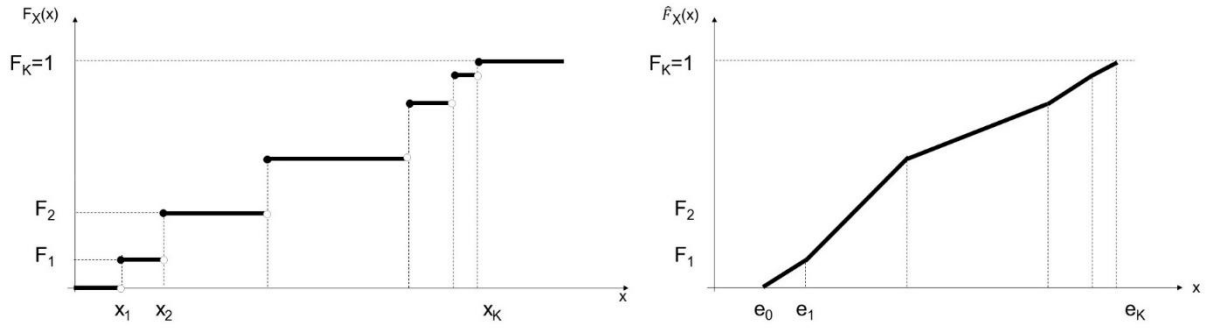


Figure 1. Fréquences cumulées croissantes. Gauche : fonction de répartition. Droite : Fréquences cumulées croissantes lors d'un regroupement par classes : polygone des fréquences cumulées

1.9 Description de données statistiques

1.9.1 Indicateurs de position (indicateurs de tendance centrale)

1.9.1.1 Moyenne arithmétique

La moyenne arithmétique de la variable X est le nombre noté $\bar{X}$ et défini par

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X(i)$$

Remarque : un regroupement par effectif donne alors :

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{K} n_k . x_k = \sum_{k=1}^{K} f_k . x_k$$

Remarque : dans le cas d'un regroupement par classes, on considèrera que tous les individus d'une classe ont leur modalité égale au centre de classe

1.9.1.2 Mode

Le mode Mo(X) de la variable X est la modalité la plus fréquemment observée dans la population P.

Le mode se calcule à partir des fréquences. La modalité x_k sera le mode de X si $f_k \geq f_j$, $\forall$ j=1,…..,K

Remarque : dans le cas d'un regroupement par classes, on ne connait pas la fréquence de chaque modalité donc on ne peut pas estimer le mode. On ne pourra que définir la classe modale comme étant la classe associée à la plus forte densité. Le mode de X sera par convention le centre de cette classe.

1.9.1.3 Médiane

Soit X une variable statistique et soit i un individu de la population.

Le rang de i, noté R(i), est sa position dans un classement des individus de la population par ordre croissant des modalités de X, une fois les ex æquo départagés.

L'individu médian est un individu désigné par i_{Me} tel que $R(i_{Me}) = \lceil N/2 \rceil$ où pour tout nombre réel y, $\lceil y \rceil$ désigne le plus petit nombre entier supérieur ou égal à y.

Ainsi on a

$$R(i_{Me}) = (N+1) / 2 \text{ si N est impair}$$

$$R(i_{Me}) = N / 2 \text{ si N est pair}$$

La médiane de X, que l'on note Me(X), est la modalité de X présentée par l'individu médian :

$$Me(X) = X(iMe).$$

Remarque : On déduit immédiatement de la définition que

- (N−1) / 2 individus précèdent i_{Me} et (N−1) / 2 suivent i_{Me}, si N est impair ;
- N / 2 − 1 individus précèdent i_{Me} et N / 2 individus suivent i_{Me} si N est pair.

1.9.1.4 Analyse et comparaison des différents indicateurs de tendance centrale

Robustesse des indicateurs

Soient les données brutes :

i	1	2	3	4	5	6	7	8	9	10
X(i)	17	1	10	11	18	10	9	18	19	10

Mode de X : Mo(X) = 10 ; Médiane : Me(X) = 10 ; Moyenne = 12,3

Si pour l'individu 2, on s'est trompé et qu'au lieu de 1, X(2) = 8

Le mode et la médiane ne sont pas modifiés. La moyenne passe à 13,0 !

Multimodalité d'une distribution

Soient les données brutes :

i	1	2	3	4	5	6	7	8	9	10	11	12
X(i)	17	3	10	11	18	10	9	18	19	10	9	18

Il y a 2 modes : 10 et 18, donc 2 groupes : les étudiants autour de 10 et ceux autour de 18

Moyenne et médiane ne peuvent capter cette information (Moyenne : 12,7)

Synthèse

Le mode et la médiane sont dits robustes à des erreurs d'observation. Notamment, la médiane n'est généralement pas modifiée par les erreurs d'observation faites sur des individus à très faible ou très forte modalité. La moyenne est modifiée par ces erreurs d'observation.

Seul le mode peut nous renseigner sur la bi- ou multimodalité de la distribution de X

La moyenne et la médiane peuvent être différentes et sont donc intéressantes à utiliser en même temps.

1.9.2 Indicateurs de dispersion

1.9.2.1 Indicateurs de dispersion absolue

Etendue

Définition

On appelle étendue de la distribution de la variable X le nombre noté ETD(X) défini par

$ETD(X)=X^M-X_m$ avec $X^M=\max\{X(i), i \in P\}$ et $X_m=\min\{X(i), i \in P\}$

C'est la distance maximale observée entre deux modalités.

Analyse de cet indicateur

Les modalités extrêmes sont généralement les plus difficiles à mesurer avec précision. L'ETD est donc peu précise. On prendra souvent cette étendue sur une population débarrassée des extrêmes. Cela mène à la définition d'étendue interquartile, plus fiable.

Etendue interquartile

Notion de quartile

Soit P une population et X une variable (numérique ou ordinale). Soient les populations P1 et P2 constituées par la médiane de X.

– Le premier quartile de X est la modalité de X, notée $Q_1(X)$, définie comme la médiane de X pour la sous-population P1.

– Le deuxième quartile de X est la modalité de X, notée $Q_2(X)$, définie par $Q_2(X) = Me(X)$.

– Le troisième quartile de X est la modalité de X, notée $Q_3(X)$, définie comme la médiane de X pour la sous-population P2.

Etendue interquartile

On appelle étendue interquartile de X la valeur EIQ(X) définie par

$$EIQ(X) = Q_3(X) - Q_1(X)$$

1.9.2.2 Les indicateurs de dispersion absolue autour d'une tendance centrale

Soit une population P décrite par une variable X.

Soit TC(X) une mesure de tendance centrale (mode, médiane, moyenne).

Soient une distance d et une variable D telles que pour tout i ϵ P : D(i) = d(X(i) ; TC(X)).

Les valeurs de D nous fournissent la dispersion autour de TC(X).

Construction d'un indicateur de dispersion autour d'une tendance centrale

- Choix de d (d(a, b) = |a − b|, ou d'(a, b) =d(a, b)² = (a − b)²)
- Choix de TC(X) ((mode, médiane, moyenne)
- Choix de la grandeur qui va résumer la dispersion D (mode, médiane, moyenne)

Sur ce principe on construit les indicateurs suivants :

Pour une variable statistique X, l'écart absolu moyen (EAM) est la quantité définie par :

$$EAM(X) = \frac{1}{N}\sum_{i=1}^{N}|X(i) - \bar{X}|$$

Pour une variable statistique X, la variance est la quantité V(X) définie par :

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

Pour une variable statistique X, l'écart-type est la quantité σ(X) définie par :

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

Théorème de König-Huygens

On peut écrire

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

Démonstration

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

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

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

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

1.9.2.3 Les indicateurs de dispersion relative autour d'une tendance centrale

Ecart absolu relatif moyen

On appelle écart absolu relatif moyen la quantité définie par

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{X(i)-\bar{X}}{\bar{X}}\right|$$

Coefficient de variation

On appelle coefficient de variation de X, et on note CV(X), la quantité définie par

$$COV(X)=\frac{\sigma(X)}{|\bar{X}|}$$

1.9.3 Indicateurs de symétrie

Notion d'asymétrie à gauche / asymétrie à droite

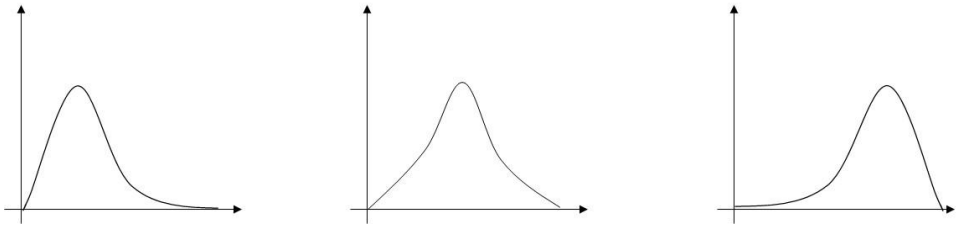


Figure 2. Gauche : distribution asymétrique à gauche. Milieu : distribution symétrique. Droite : distribution asymétrique à droite

Lorsque l'on considère des variables quelconques, les positions relatives du mode, de la médiane et de la moyenne sont

- Mo(X) < Me(X) < $\bar{X}$ si la distribution de X est asymétrique à gauche
- Mo(X) > Me(X) > $\bar{X}$ si la distribution de X est asymétrique à droite

1.9.3.1 Coefficients d'asymétrie de Pearson

On appelle coefficients d'asymétrie de Pearson les deux coefficients définis par

$$P_1(X)=\frac{\bar{X}-Mo(X)}{\sigma(X)}$$

$$P_2(X)=\frac{\bar{X}-Me(X)}{\sigma(X)}$$

Observer P2(X) < 0, càd $\bar{X}$ < Me(X), indique une asymétrie à droite, et P2(X) > 0 une asymétrie à gauche.

1.9.3.2 Coefficient d'asymétrie γ1

Le coefficient d'asymétrie γ1 est défini par

$$\gamma_1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X(i)-\bar{X}}{\sigma(X)}\right)^3$$

– si γ1(X) est nul, la distribution de X est symétrique ;

– si γ1(X) est négatif (positif), la distribution de X est asymétrique à gauche (à droite).

1.9.3.3 Coefficient d'asymétrie de Yule-Bowley

Soient $Q_1(X)$, $Q_2(X)$, $Q_3(X)$ les quartiles de X. Le coefficient d'asymétrie de Yule-Bowley B(X) est défini par :

$$B(X) = \frac{[Q_3(X)-Q_2(X)]-[Q_2(X)-Q_1(X)]}{Q_3(X)-Q_1(X)}$$

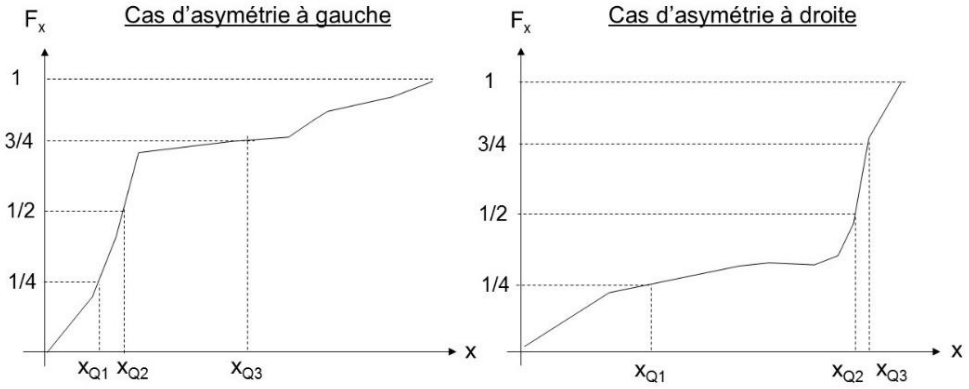
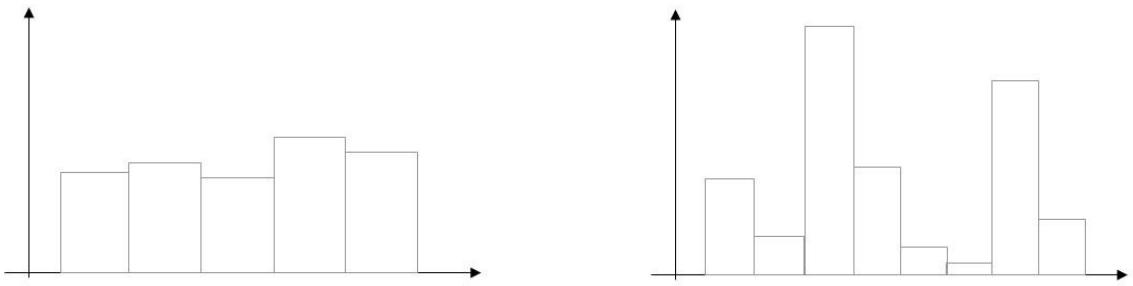


Figure 3. Polygone des fréquences cumulées en cas d'asymétrie

1.9.4 Indicateurs d'aplatissement

Notion d'aplatissement



L'indice d'aplatissement κ(X) de la variable X est défini par

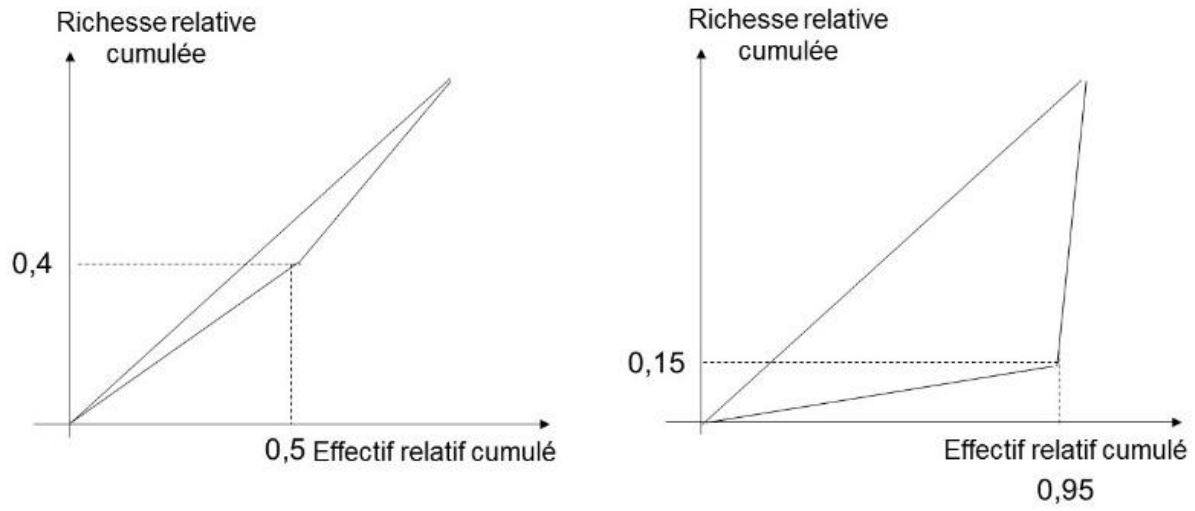
$$\kappa(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\bar{X})^4}{\sigma(X)^4}$$

L'indice d'aplatissement κ(X) est d'autant plus grand que les modalités de la variable X sont dispersées autour des valeurs $\bar{X}$ − σ(X) et $\bar{X}$ + σ(X).

L'indice κ(X) atteint sa valeur minimale 1 si et seulement si pour tout individu i = 1, . . . ,N, on a : X(i) = $\bar{X}$ + σ(X) ou X(i) = $\bar{X}$ − σ(X).

1.9.5 Indicateurs de concentration

1.9.5.1 Courbe de Lorenz



Gauche : 50% détient 40% de la richesse, 50% détient 60% de la richesse

Droite : 95% détient 15% de la richesse, 5% détient 85% de la richesse

1.9.5.2 Médiale

La médiale est la médiane de la nouvelle série calculée (série associée aux richesses).

1.9.5.3 Indice de Gini

L'indice de Gini I_G est le rapport entre l'aire de concentration et.

$$I_G = \frac{Aire\ de\ concentration}{Aire\ du\ triangle\ sous\ la\ ligne\ d'équirépartition}$$

Aire du triangle sous la ligne d'équirépartition = ½

Aire de concentration = ½ - Aire sous la courbe de Lorenz

I_G = 0 : on parle d'équirépartition. I_G = 1 : maximum d'inégalité

1.9.6 Représentation à l'aide d'une boite à moustache

Une boite à moustache est une figure permettant la représentation graphique des indicateurs de tendance centrale et de la dispersion d'une variable statistique.

Elle se présente sous forme de rectangle dont la longueur représente le premier quartile et le troisième quartile. Des segments sont ensuite ajoutés pour représenter les valeurs extrêmes. Dans le rectangle, on ajoute les indicateurs de tendance centrale comme la médiane.

https://www.bibmath.net/dico/index.php?action=affiche&quoi=./m/moustache.html&special=imprimable

2 Ajustements linéaires et non linéaires

Soient deux variables x et y. On trace la variable y en fonction de la variable x.

Soient les exemples ci-dessous.

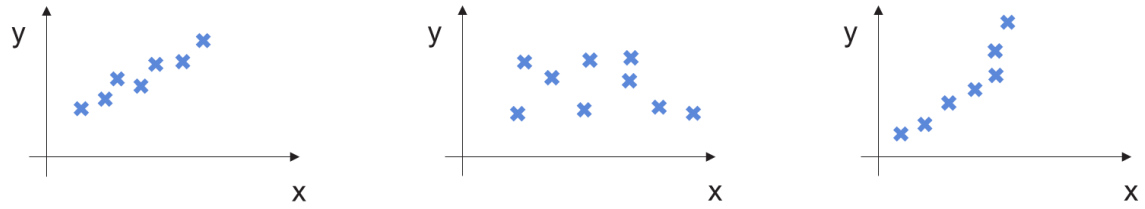


Figure 4. Exemples de nuages de points

Sur le nuage de points à gauche, on peut raisonnablement envisager une corrélation linéaire, soit un modèle de type $y = ax + b$ avec a et b réels. Sur le nuage de points au milieu, les deux variables ne semblent pas corrélées, c'est-à-dire que y ne semble pas varier lorsque x varie. Sur le nuage de points de droite, une corrélation semble exister mais non linéaire.

On peut ainsi se poser les questions suivantes :

- Existe-t-il une corrélation entre x et y ? si oui peut-elle être supposée linéaire ?
- Comment établir une relation mathématique entre x et y ?
- Comment définir et quantifier la « qualité » du modèle mathématique proposé ?

2.1 Régression linéaire

2.1.1 Coefficient de corrélation linéaire

2.1.1.1 Rappels mathématiques

- Propriétés de l'espérance

Soient X et Y deux variables aléatoires et soient a et b deux réels. Alors

$$E(a.x + b) = a.E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

- Définition et Propriétés de la variance

$$V(X) = E[(X - E(X))^2]$$

$$V(aX + b) = a^2.V(X)$$

- Ecart-type

$$\sigma(X) = \sqrt{V(X)}$$

- Covariance

$$cov(X, Y) = E\big[\big(X - E(X)\big).(Y - E(Y))\big]$$

2.1.1.2 Définition : coefficient de corrélation linéaire

On définit r tel que

Coefficient de corrélation linéaire

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

$cov(X,Y)$ est la covariance de X et Y. $\sigma(X)$ écart-type de X.

Supposons que Y=aX+b avec a>0.

Par les propriétés de la variance et de la covariance on a

$\sigma(Y) = a.\sigma(X)$

$cov(X,Y) = E\big[\big(X - E(X)\big).(Y - E(Y))\big]$

$cov(X,Y) = E[X.Y - X.E(Y) - E(X).Y + E(X).E(Y)]$

$cov(X,Y) = E[aX^2 + bX - X.a.E(X) - X.b - E(X).a.X - E(X).b + E(X).a.E(X) + E(X).b]$

$cov(X,Y) = E[aX^2 - 2aXE(X) + aE(X)^2] = aE[(X - E(X))^2] = a.V(X)$

Donc

$$r = \frac{a.V(X)}{\sigma(X).a.\sigma(X)} = 1$$

De même, soient X et Y deux variables telles que Y=aX+b avec a<0, alors

$$r = -1$$

Lorsqu'une relation linéaire existe entre X et Y, alors r=1 ou -1.

Coefficient de corrélation linéaire r

$$0 \leq |r| \leq 1$$

Plus r est proche de 1, plus la relation linéaire entre X et Y se rapproche d'une relation linéaire.

Par ailleurs, si r=0 alors $cov(X,Y) = 0$

Remarque

Une valeur r=0 ne veut pas dire que les variables sont indépendantes ! Cela veut seulement dire qu'il n'existe pas de relation linéaire entre elles.

Exemple. Soit la variable aléatoire X de loi symétrique par-rapport à 0. Soit $Y=X^2$.

La variable XY est aussi symétrique par-rapport à 0 donc E(XY)=0.

De plus, on a E(X).E(Y)=0 car E(X)=0 et E(Y)=0, donc cov(X,Y)=0 et r=0 or X et Y ne sont pas indépendantes puisque $Y=X^2$.

2.1.2 Méthodes de régression linéaire

Quand r nous indique que l'on peut supposer une relation linéaire entre X et Y, on peut déterminer a et b tels que $Y = aX + b$.

2.1.2.1 Méthode des points extrêmes

Dans cette méthode, on considère une droite passant par les points extrêmes du nuage de points, comme ci-dessous.

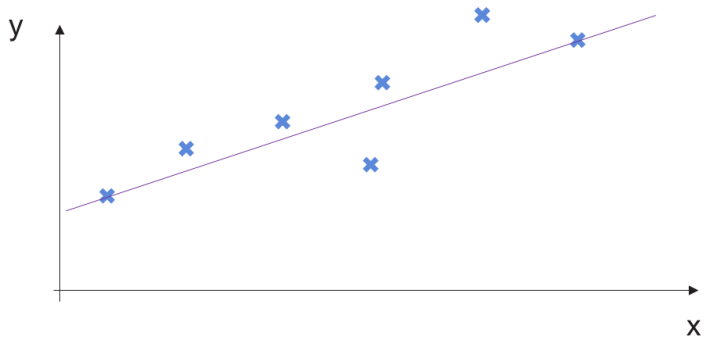


Figure 5. Méthode des points extrêmes

Les coefficients a et b du modèle $Y = aX + b$ se déterminent ensuite classiquement comme lorsque l'on connait deux points d'une droite.

2.1.2.2 Méthode de Mayer (points moyens)

Dans la méthode de Mayer, on regroupe le nuage de points en deux groupes et l'on calcule le point moyen pour chaque groupe. On trace une droite passant par ces deux points appelée la droite de Mayer.

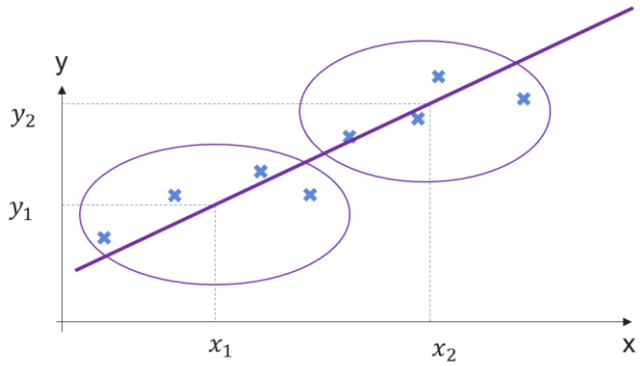


Figure 6. Méthode de Mayer (points moyens)

2.1.2.3 Méthode des moindres carrés

Introduction

La traque d'une planète, Gauss, et la méthode des moindres carrés (*Extrait du livre « Une révolution de la théorie des nombres. Gauss » 2018)* [1]

Mais où est passée Cérès ?

Figure 7. Carl Friedrich Gauss 1777-1855 Mathématicien, astronome et physicien allemand

« A tout juste vingt-cinq ans, Gauss réussit à déterminer l'orbite de Cérès à l'aide seulement de quelques données et de ses connaissances en mathématiques. Ce haut fait lui valut une brillante réputation dans toute l'Europe.

Le vide entre Mars et Jupiter attirait particulièrement l'attention. Les astronomes s'étaient toujours étonnés qu'il y eut une si grande distance entre ces deux planètes. Mais ils avaient beau scruter cette région, ils n'y avaient jamais rien découvert.

Or le 1er janvier 1801, Piazzi annonça avoir aperçu un objet sidéral qui se déplaçait sur la voûte céleste et semblait tourner autour du soleil, dans la région entre Mars et Jupiter. Piazzi décida d'appeler sa planète Cérès. Piazzi avait passé des nuits entières collé à son objectif. Malheureusement, dans la nuit du 11 février, il fut pris d'une grippe si violente qu'il dut rester alité et ne put travailler pendant plusieurs nuits de suite. Quand enfin il se fut remis et qu'il retourna à son télescope, Cérès n'était plus là. Les observations ne s'étalaient que sur 42 jours, période trop courte pour pouvoir déterminer la trajectoire de Cérès.

Le baron Von Zach se rappela qu'à Brunswick vivait un mathématicien de 24 ans qui selon les dires étudiait non seulement les mathématiques mais les réinventait. Le baron rassembla toutes les informations dont il disposait et les fit parvenir au jeune Gauss.

Gauss ne pose aucune hypothèse de départ, il se contenta d'exploiter les chiffres qu'on lui avait fournis.

Dans l'équipe d'astronomes, le moral était au plus bas. Gauss n'était qu'un mathématicien, il n'avait jamais touché à un télescope de sa vie. Von Zach se résolut enfin à suivre les indications que lui avait fait parvenir le jeune mathématicien. Et, la nuit du 7 décembre 1801, les télescopes de Lilienthal détectèrent un petit point légèrement brillant qui émergeait de l'obscurité, tout près de l'emplacement que Gauss avait déterminé par ses calculs.

Une explosion de joie accueillit cette nouvelle. A partir de rien, Gauss avait accompli ce qu'il est bien convenu d'appeler des merveilles.

Pour parvenir à ce résultat, Gauss eut recours à la **méthode dite des moindres carrés**, l'une des contributions les plus importantes de sa carrière au domaine des mathématiques. »

La méthode des moindres carrés

La méthode des moindres carrés développée par Gauss s'inscrit dans le domaine de l'optimisation en mathématiques. Elle a pour but de donner la fonction qui s'ajuste le mieux aux données expérimentales.

Etant donnée une relation entre deux variables qui fait que l'une soit dépendante de l'autre, cette relation est définie par une fonction f de sorte que $f(t_i) = y_i$. Le but est de trouver cette fonction.

Dans le cas de Cérès, on suppose que la variable indépendante est la situation dans le temps, tandis que la variable dépendante est la position dans l'espace.

Soient les paires de données (t_1, y_1), (t_2, y_2), (t_3, y_3) chacune indiquant une situation dans le temps et une position dans l'espace. Déterminer la trajectoire de la planète revient à trouver une fonction $\hat{y}_i = f(t_i)$. Il faut minimiser la différence entre la position réelle du corps céleste, y_i, et sa position estimée $\hat{y}_i$. L'écart entre ces deux valeurs se nomme erreur $e_i = y_i - f(t_i)$.

Pour éviter que les erreurs par excès annulent les erreurs par défaut, on élève les erreurs au carré

$$S_{CR} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - f(t_i))^2$$

S_{CR} : somme des carrés des résidus.

La méthode des moindres carrés tente de définir une droite qui minimise la somme des segments rouges sur la figure ci-dessous.

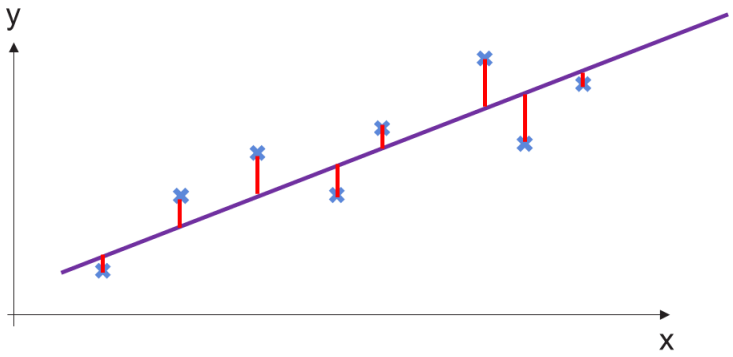


Figure 8. Illustration visuelle du principe de la méthode des moindres carrés

La régression linéaire

On dispose de n observations (x_i, y_i) qui associent les valeurs de ces deux variables. Comme il existe des raisons de supposer que l'ajustement est linéaire, la relation qui s'établit entre les variables est une droite :

$$f(x_i) = b_1 + b_2.x_i$$

On réduit les erreurs au minimum, avec $e_i = y_i - (b_1 + b_2.x_i)$. On a alors

$$S_{CR} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} \left(y_i - (b_1 + b_2.x_i)\right)^2$$

Questions

L'objectif est de déterminer b_1 et b_2 qui minimisent S_{CR}. On rappelle que si a est un extremum local d'une fonction g continue en a, alors en a, g'(a)=0

1/ Montrer que l'on a le système suivant

$$\begin{cases} -2.\sum_{i=1}^{n}\left(y_i - (b_1 + b_2.x_i)\right) = 0 \\ -2.\sum_{i=1}^{n}(y_i.x_i - (b_1 + b_2.x_i).x_i) = 0 \end{cases}$$

On définit

$\alpha = \frac{1}{n}\sum_{i=1}^{n} y_i.x_i$; $\beta = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2$; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

2/ Montrer que le système d'équations de la question 1 devient

$$\begin{cases} \bar{y} = b_1 + b_2.\bar{x} \\ \alpha = b_1.\bar{x} + b_2.\beta \end{cases}$$

Si on multiplie par $\bar{x}$ les deux membres de la première équation :

$$\begin{cases} \bar{y}.\bar{x} = (b_1 + b_2.\bar{x}).\bar{x} \\ \alpha = b_1.\bar{x} + b_2.\beta \end{cases}$$

puis si l'on soustrait les expressions on obtient :

$$\alpha - \bar{y}.\bar{x} = b_2.(\beta - \bar{x}^2)$$

3/ En déduire les expressions de b_1 et b_2 en fonction de α, β, $\bar{x}$ et $\bar{y}$

Correction

1/ Dériver partiellement Scr par-rapport aux variables b1 et b2

$$S_{CR} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}\left(y_i - (b_1 + b_2.x_i)\right)^2$$

Rappel : dérivée de u^n = n.u'.u^{n-1}

$$\frac{\partial S_{CR}}{\partial b_1} = \sum_{i=1}^{n} -2.\left(y_i - (b_1 + b_2.x_i)\right) = -2.\sum_{i=1}^{n}\left(y_i - (b_1 + b_2.x_i)\right)$$

$$\frac{\partial S_{CR}}{\partial b_2} = \sum_{i=1}^{n} 2.\left(y_i - (b_1 + b_2.x_i)\right)(-x_i) = -2\sum_{i=1}^{n}(x_i y_i - (b_1 + b_2.x_i)x_i)$$

Si b1 et b2 minimisent Scr alors

$$\frac{\partial S_{CR}}{\partial b_1} = 0$$

$$\frac{\partial S_{CR}}{\partial b_2} = 0$$

Donc

$$\begin{cases} -2.\sum_{i=1}^{n}\left(y_i - (b_1 + b_2.x_i)\right) = 0 \\ -2.\sum_{i=1}^{n}(y_i.x_i - (b_1 + b_2.x_i).x_i) = 0 \end{cases}$$

2/ Réagencement des termes et division par n

$$\sum_{i=1}^{n}(y_i) = \sum_{i=1}^{n}\left((b_1 + b_2.x_i)\right) = \sum_{i=1}^{n}\left((b_1)\right) + \sum_{i=1}^{n}\left((b_2.x_i)\right)$$

On divise par n

$$\bar{y} = b_1 + b_2.\bar{x}$$

Idem pour l'autre équation

$$\begin{cases} \bar{y} = b_1 + b_2.\bar{x} \\ \alpha = b_1.\bar{x} + b_2.\beta \end{cases}$$

3/ En déduire les expressions de b_1 et b_2 en fonction de a_{11}, a_{20}, $\bar{x}$ et $\bar{y}$

$$\alpha - \bar{y}.\bar{x} = b_2.(\beta - \bar{x}^2)$$

$$b_2 = \frac{\sum_{i=1}^{n} y_i.x_i - n.\bar{y}.\bar{x}}{\sum_{i=1}^{n} {x_i}^2 - n.\bar{x}^2}$$

$$b_1 = \bar{y} - b_2.\bar{x}$$

La méthodes des moindres carrés a comme paramètres b_1 et b_2, définis par $y = b_1 + b_2.x$ et tels que

$$b_2 = \frac{\sum_{i=1}^{n} y_i.x_i - n.\bar{y}.\bar{x}}{\sum_{i=1}^{n} {x_i}^2 - n.\bar{x}^2}$$

$$b_1 = \bar{y} - b_2.\bar{x}$$

Avec $\bar{x}$ moyenne des x, $\bar{y}$ moyenne des y.

La méthode des moindres carrés minimise la grandeur

$$S_{CR} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

Avec $f(t_i)$ valeur prédite de y_i (donnée par le modèle).

2.1.2.4 Synthèse

La méthode des points extrêmes est la moins précise car elle ne prend en compte que deux points et n'exploite pas les autres points. De plus on sait que l'incertitude sur les mesures est souvent plus grande pour les valeurs extrêmes, soit précisément les points utilisés pour ajuster le modèle. Elle est cependant simple et rapide à utiliser.

La méthode de Mayer prend en compte l'ensemble des points, et reste simple et rapide à mettre en œuvre.

La méthode des moindres carrés est plus lourde de mise en œuvre mais elle est la plus précise des trois méthodes.

EXCEL utiliser la méthode des moindres carrés lorsque l'on trace la courbe de tendance d'une série de données.

2.1.3 "Qualité" d'un ajustement linéaire : coefficient de détermination

Par une des méthodes décrites précédemment on a déterminé un ajustment linéaire entre x et y tel que $y = f(x)$.

Comment estimer la qualité de cet ajustement ? On définira le coefficient de détermination R tel que

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$\sum_{i=1}^{n}(y_i - f(x_i))^2$ est l'écart entre la variable y_i et la valeur prédite par le modèle, $f(x_i)$.

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ est l'écart entre la variable y_i et la moyenne des y.

https://www.canal-u.tv/chaines/canal-unisciel/regression-lineaire-mise-en-oeuvre-de-la-regression-lineaire-simple

https://fr.wikipedia.org/wiki/Coefficient_de_d%C3%A9termination

2.2 Régression non linéaire

Reprenons la Figure 4, à droite. Il semble exister une relation entre X et Y mais elle ne semble pas linéaire.

Dans le cas d'une relation non linéaire entre X et Y, une première méthode sera de se rapporter à un ajustement linéaire en effectuant un changement de variable.

2.2.1 Relation de type $Y = A.X^n$

Soient deux variables X et Y tels que

$$Y = A.X^n$$

Avec A et n réels.

L'égalité peut s'écrire

$$\log(Y) = \log(A.X^n) = \log(A) + n.logX$$

On voit alors une relation linéaire entre logY et logX. On ajustera un modèle linéaire entre logY et logX afin de calculer les réels A et n.

2.2.2 Relation de type $Y = A.n^X$

Soient deux variables X et Y tels que

$$Y = A.n^X$$

Avec A et n réels.

L'égalité peut s'écrire

$$\log(Y) = \log(A) + X.logn$$

Cette fois-ci, on ajustera un modèle linéaire entre logY et X afin de déterminer les constantes A et n.

2.3 Sur l'effet cigogne (où l'on apprend que corrélation et causalité sont deux notions bien distinctes)

2.3.1 Présentation

https://fr.wikipedia.org/wiki/Corr%C3%A9lation_n%27implique_pas_causalit%C3%A9

2.3.2 Exemple : le snack de la plage

Sur un an, on a noté la quantité de crèmes glacées vendues par le snack d'une plage et on a mesuré la quantité de mégots de cigarettes ramassés dans le sable chaque semaine. Les résultats sont donnés ci-dessous.

1/ Tracer le nuage de points : mégots en fonction des glaces vendues

2/ Peut-on envisager une corrélation linéaire ?

4/ Peut-on conclure que la quantité de mégots augmente parce que la quantité de glaces augmente, ou devrait-on dire que lorsque la quantité de mégots augmente alors la quantité de glaces vendues augmente, mais qu'une troisième variable influence les deux avec la même tendance ?

Glaces	Mégots
1	10
12	25
45	90
33	71
60	105
2	6
9	25
98	180
14	36
56	120

3	12
19	45
16	30
18	40
20	50
22	44
7	20
4	14
40	83
36	71
25	50
80	161
77	155
65	141
53	116
70	149

3 Echantillonnage et estimation

3.1 Introduction

On considèrera une population composée d'un ensemble d'individus.

On s'intéressera à un caractère particulier des individus de cette population. On suppose ce caractère quantifiable par un nombre réel. On suppose qu'on a mesuré la valeur du caractère de n individus et qu'on a obtenu les nombres x_1, …, x_n.

Un échantillon de taille n est une partie de n éléments choisis aléatoirement dans une population P.

Si l'on étudie un caractère statistique de cette population, l'échantillon nous donne une suite X_1, …, X_n de variables aléatoires indépendantes avec X_i valeur du caractère statistique de l'individu i.

Remarque : pour que les variables X_i soient indépendantes, on considérera des tirages « avec remise » (on prélève X_1 on note son caractère et on le réinjecte dans P). P n'est alors pas modifiée par le tirage et la probabilité de tirer X_i n'est pas modifiée par le tirage. Si l'échantillon est suffisamment petit par-rapport à P, on considérera que le résultat d'un tirage sans remise fournit quand même des variables indépendantes.

3.2 Pré-requis sur les variables aléatoires

3.3 Définitions et Propriétés générales

Soit X une variable aléatoire prenant les valeurs $x_1, \ldots, x_n$ avec les probabilités respectives $p_1, \ldots, p_n$.

On appelle espérance mathématique de X le nombre réel, noté E(X), défini par :

$$E(x) = x_1p_1 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i$$

Soit X une variable aléatoire. On appelle variance de X le nombre réel V(X), défini par

$$V(X) = E\left[\left(X - E(X)\right)^2\right]$$

La variance est donc la moyenne des carrés des écarts à la moyenne.

On appelle écart type de X le nombre réel σ(X), défini par

$$\sigma(X) = \sqrt{V(X)}$$

Soient X et Y deux variables aléatoires sur un espace de probabilité (Ω,P), et soient a et b deux réels. Alors

$$E(aX + b) = a.E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

$$V(aX + b) = a^2.V(X)$$

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

3.4 Variance de la somme de deux variables aléatoires

On propose de calculer

$$(X+Y)-E(X+Y)=\big(X-E(X)\big)+\big(Y-E(Y)\big)$$

Par définition

$$V(X+Y)=E[(X+Y)-E(X+Y)]^2$$

Or

$$[(X+Y)-E(X+Y)]^2=[X-E(X)]^2+[Y-E(Y)]^2+2\big(X-E(X)\big).\big(Y-E(Y)\big)$$

Donc

$$V(X+Y)=V(X)+V(Y)+2E\big[\big(X-E(X)\big).\big(Y-E(Y)\big)\big]$$

Soit la covariance de X et Y définie par

$$cov(X,Y)=E\big[\big(X-E(X)\big).\big(Y-E(Y)\big)\big]$$

Proposition admise

Si X et Y sont indépendantes, alors

$$cov(X,Y)=0$$

et donc

$$V(X+Y)=V(X)+V(Y)$$

Conséquence

Soient n variables aléatoires X_1, ……., Xn indépendantes, suivant la même loi d'espérance μ et d'écart-type σ. Soit M la moyenne arithmétique de ces variables aléatoires

$$M=\frac{X_1+X_2+\cdots..+X_n}{n}$$

L'espérance et la variance de M sont :

E(M) = μ

V(M)=σ^2/n

3.5 Loi binomiale

Une épreuve de Bernoulli est une épreuve à deux éventualités (choix d'un joueur entre deux options, succès ou échec, pile ou face) dont les probabilités respectives sont notées p et 1 − p (p ϵ [0, 1]). Une telle épreuve est appelée épreuve de Bernoulli de paramètre p.

Soit A la variable aléatoire telle que

- A=1 si succès de probabilité p
- A=0 si non succès de probabilité 1-p

Alors, par définition de l'espérance,

$$E(A) = 1 * p + 0 * (1 - p) = p$$

De plus, par la formule de König-Huyghens,

$$V(A) = E(A^2) - E^2(A) = p - p^2 = p.(1 - p)$$

Une expérience aléatoire consistant à répéter n fois (n entier > 0), de manière indépendante, une même épreuve de Bernoulli de paramètre p est un schéma de Bernoulli de paramètres n et p. A un schéma de Bernoulli de paramètres n et p, on peut associer la variable aléatoire X égale au nombre de succès en n tentatives. Cette variable aléatoire prend donc les valeurs 0, 1, 2, . . . , n − 1, n. La loi binomiale est la loi de probabilité associée à cette variable aléatoire et si k est un entier élément de {0, 1, . . . , n}, on a

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

L'espérance de la loi binomiale de paramètre n et p est

$$E(X) = E(A_1 + A_2 + \cdots A_n) = n.E(A) = n.p$$

La variance de la loi binomiale de paramètre n et p est

$$V(X) = V(A_1 + A_2 + \cdots A_n) = n.V(A) = n.p.(1 - p)$$

3.6 Théorèmes limites

3.6.1 Inégalité de Tchebychev

Soit X une variable aléatoire d'espérance E(X) et de variance V(X). Alors pour tout ε > 0,

$$P\{|X - E(X)| \geq \varepsilon\} \leq \frac{V(X)}{\varepsilon^2}$$

Démonstration

Soit X une variable aléatoire d'espérance E(X) et de variance V(X), et soit a un réel positif. On construit la variable aléatoire Y telle que :

$$\begin{cases} Y(\omega) = a^2 \text{ si } |X(\omega) - E(X)| \geq a \\ \qquad\quad Y(\omega) = 0 \text{ sinon} \end{cases}$$

On a alors : $Y \leq [X - E(X))]^2$

Donc : $E(Y) \leq E([X - E(X))]^2)$

Donc par définition de la variance : $E(Y) \leq V(X)$

Or par définition de l'espérance : $E(Y) = a^2 . P(|X-E(X)| \geq a)$

On a ainsi obtenu l'inégalité de Tchebychev : $P(|X-E(X)| \geq a) \leq V(X) / a^2$

3.6.2 Loi des grands nombres

Soit (Xn) une suite de variables aléatoires indépendantes de même espérance et de même variance (càd pour tout entier n, d'espérance μ et variance σ²). Soit $\overline{Xn}$ la moyenne arithmétique de l'échantillon de taille n :

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Alors :

$$\lim_{n\to\infty} p(\mu - \varepsilon < \overline{X_n} < \mu + \varepsilon) = 1$$

Démonstration

Les variables Xi ont même espérance μ, même variance σ² et sont indépendantes, donc

$$E(\overline{X_n}) = \mu \text{ et } V(\overline{X_n}) = \sigma^2 / n$$

On écrit l'inégalité de Tchebychev pour la variable $\overline{X_n}$:

$$P\{|\overline{X_n} - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{n.\varepsilon^2}$$

Par conséquent :

$$\lim_{n\to+\infty} P\{|\overline{X_n} - \mu| > \varepsilon\} = 0$$

$$\lim_{n\to+\infty} P\{|\overline{X_n} - \mu| < \varepsilon\} = 1$$

Remarque : on dira que $\overline{X_n}$ converge en probabilité vers μ

3.7 Théorème central limite

Soit (Xn) une suite de variables aléatoires indépendantes et de même loi, d'espérance μ et variance σ².

Alors la variable aléatoire

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n.\mu}{\sigma\sqrt{n}}$$

converge en loi vers la loi normale centrée réduite N(0,1).

Remarque : on peut aussi écrire la variable aléatoire Z_n de la sorte :

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n.\mu}{\sigma\sqrt{n}} = \frac{\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu}{\frac{\sigma\sqrt{n}}{n}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Démonstration du théorème central limite

Elle fait appel à la notion de fonction caractéristique (dont le cadre sort de ce cours)

- https://fr.wikipedia.org/wiki/Fonction_caract%C3%A9ristique_%28probabilit%C3%A9s%29

Une fois cette notion introduite, on peut s'intéresser à la démonstration du théorème :

- florian.bouguet.free.fr/doc/developpements/TCL.pdf
- https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_central_limite#D.C3.A9monstration_du_th.C3.A9or.C3.A8me_central_limite
- math.univ-lille1.fr/~suquet/Polys/TLC.pdf

Il sera nécessaire dans cette démonstration d'exploiter le théorème de convergence de Lévy dont l'énoncé peut se trouver ci-dessous :

- https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_convergence_de_L%C3%A9vy

3.8 Estimation

3.8.1 Notion d'estimateur

Soit un échantillon aléatoire $X_1,...,X_n$ dont la loi de probabilité a pour paramètre θ.

θ n'est pas connu, on veut l'estimer à l'aide d'un échantillon aléatoire de la population.

Un estimateur du paramètre θ est une valeur expérimentale (moyenne, variance...) utilisée comme estimation de θ.

3.8.2 Convergence d'un estimateur

Soit un estimateur θ_n du paramètre θ. Cet estimateur est issu d'un échantillon aléatoire (X1,.....,Xn).

Pour 1 individu, on calculera θ_1. Pour 2 individus, on calculera θ_2.... On peut donc définir une suite θ_n de variables aléatoires.

On dira que l'estimateur θ_n est convergent s'il converge en probabilité vers θ.

3.8.3 Efficacité et biais d'un estimateur

L'efficacité d'un estimateur est quantifiée par la grandeur $E((\theta_n - \theta)^2)$. Plus cette erreur est petite, ou converge vite vers 0, plus l'efficacité de l'estimateur est grande.

On montre que $E((\theta_n - \theta)^2) = V(\theta_n) + (E(\theta_n) - \theta)^2$

La grandeur $E(\theta_n) - \theta$ est le biais de l'estimateur.

Si $E(\theta_n) \rightarrow \theta$ quand $n \rightarrow \infty$, on parlera d'estimateur sans biais.

On dit que l'estimateur est sans biais quand, si on moyenne les estimations sur tous les échantillons de taille n on retrouve le paramètre θ.

3.8.4 Estimateurs usuels

Soit $X_1,...,X_n$ un échantillon aléatoire. Soient μ et σ l'espérance et l'écart-type de la variable X.

3.8.4.1 Moyenne empirique

La moyenne empirique d'un échantillon aléatoire $X_1,...,X_n$ est la variable aléatoire $\bar{X}$ définie par :

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$E(\bar{X}) = \mu$ et $V(\bar{X}) = \sigma^2 / n$ donc la moyenne $\bar{X}$ est un estimateur sans biais et convergent de l'espérance μ de loi de X.

3.8.4.2 Variance

Un estimateur sans biais et convergent de la variance de X est la variance de l'échantillon aléatoire $X_1,...,X_n$ définie par

$$S^2{}_{n-1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Remarque : la variance $S^2{}_n = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$, est un estimateur biaisé de l'échantillon. En effet son espérance $E(S^2{}_n)$ vaut $\frac{n-1}{n}.V(X)$ et non V(X). (démo : https://fr.wikipedia.org/wiki/Variance). Par conséquent on la corrige en la multipliant par $\frac{n}{n-1}$, afin d'obtenir une espérance égale à V(X) :

$$E\left(\frac{n}{n-1}.\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right) = \frac{n}{n-1}.\frac{n-1}{n}.V(X) = V(X)$$

3.8.5 Estimation par intervalle de confiance

On n'essaie pas de trouver une valeur estimée la meilleure possible de θ (moyenne, écart-type...) mais on définit un intervalle dans lequel la vraie valeur se trouve avec une probabilité définie à l'avance.

On cherchera ainsi les valeurs a et b telles que

$$p(x \in [a, b]) = 1 - \alpha$$

α est appelé le « risque ».

3.8.6 Estimation d'une moyenne μ par intervalle de confiance

On peut obtenir un intervalle de confiance [a, b] dans lequel une moyenne μ se trouve avec un risque donné.

3 questions sont à se poser :

1. Le caractère statistique suit-il une loi normale ou une loi quelconque ?
2. L'échantillon est-il de grande taille ? (càd : $n < 30$ ou $n \geq 30$?)
3. L'écart-type du caractère statistique est-il connu ou doit-il être estimé sur l'échantillon ?

3.8.6.1 Dans les cas {caractère de loi normale + écart-type connu + taille quelconque} ou {caractère de loi quelconque + écart-type connu + n ≥ 30}

Soit $\bar{X}_n$ l'estimateur : "moyenne d'un échantillon de taille n."

On sait que $E(\bar{X}_n) = \mu$ et $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$.

D'après le théorème central limite, la variable $Y_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers la loi normale centrée réduite N(0, 1).

On recherchera, dans la loi normale centrée réduite, le réel t_α tel que : $p(|Yn| < t_\alpha) = 1 - \alpha$

Or :

$$p(|Yn| < t_\alpha) = p\left(-t_\alpha < \frac{\overline{Xn} - \mu}{\sigma/\sqrt{n}} < t_\alpha\right) = p(\overline{Xn} - t_\alpha.\frac{\sigma}{\sqrt{n}} < \mu < \overline{Xn} + t_\alpha.\frac{\sigma}{\sqrt{n}})$$

Après une évaluation ponctuelle $\overline{Xn}$ de la statistique θ, on saura que θ se trouve, avec un risque d'erreur α, dans l'intervalle :

$$[\overline{Xn} - t_\alpha.\frac{\sigma}{\sqrt{n}}, \overline{Xn} + t_\alpha.\frac{\sigma}{\sqrt{n}}]$$

Remarque : si l'écart-type est inconnu on l'estime à l'aide de la variance de l'échantillon

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

3.8.6.2 Dans le cas {caractère de loi normale + écart-type inconnu}

On utilise la variance de l'échantillon définie par :

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

La variable

$$Y_n = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

suit une loi de Student à n-1 degrés de libertés (https://fr.wikipedia.org/wiki/Loi_de_Student).

On estime t_α avec cette loi. L'intervalle de confiance est ensuite donné par le même raisonnement et donc par la même formule que dans le cas précédent.

3.8.7 Estimation d'une proportion par intervalle de confiance

Dans une population P, un caractère C ne peut prendre que deux valeurs (1,0) et la proportion de la population vérifiant C = 1 est p. Celle de l'évènement contraire est q = 1−p. On veut donner un intervalle de confiance pour p à partir de son calcul p_n sur un échantillon de n individus. On considère la variable aléatoire P_n définie sur l'ensemble des échantillons de taille n.

Son espérance est p et son écart-type √(p.q/n).

D'après le théorème central limite,

$$Y_n = \frac{P_n - p}{\sqrt{\frac{pq}{n}}}$$

converge vers la loi normale centrée réduite N(0, 1)

On estime t_α tel que $p(|Yn| < t_\alpha) = 1- \alpha/2$ et avec une estimation ponctuelle de p (p_n) sur un échantillon, on obtient l'intervalle de confiance :

$$[p_n - t_\alpha\sqrt{\frac{pq}{n}}, p_n + t_\alpha\sqrt{\frac{pq}{n}}]$$

Cependant, on ne connait pas p et q. On estime donc l'intervalle de confiance par approximation ponctuelle ou par majoration.

- par approximation ponctuelle : on remplace p et q par p_n et q_n connus sur un échantillon

$$[p_n - t_\alpha\sqrt{\frac{p_n.q_n}{n}}, p_n + t_\alpha\sqrt{\frac{p_n.q_n}{n}}]$$

- par majoration : si 0 < p < 1, alors p.(1−p) < 1/4. On peut donc majorer l'intervalle de confiance par :

$$[p_n - \frac{t_\alpha}{2\sqrt{n}}, p_n + \frac{t_\alpha}{2\sqrt{n}}]$$

4 Logiciel R

4.1 Présentation de R

https://www.r-project.org/

https://cran.r-project.org/

The “Comprehensive R Archive Network” (CRAN) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.

4.2 Installation

Etape 1/ Installer R

R-4.3.0 for Windows : https://cran.r-project.org/

Etape 2/ Installer R commander (https://www.brunoy-osteopathe.fr/installer-et-configurer-r-commander/)

Dans la fenêtre de commande : install.packages("Rcmdr")

Dans la fenêtre de commande : library(Rcmdr)

(on va vous signaler qu’il manque des paquets, cliquez sur **Oui**, puis sur **OK** pour les installer.)

Quittez R Commander, puis quittez R. Une boîte de dialogue va vous proposer de sauver une image de la session. Répondez oui. Relancez R, tapez flèche haut sur le clavier, jusqu’à ce que la commande library(Rcmdr) s’affiche de nouveau, puis appuyez sur Entrée.

https://www.r-project.org/

https://cran.r-project.org/

https://fr.wikipedia.org/wiki/R_(langage)

https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/

https://www.brunoy-osteopathe.fr/installer-et-configurer-r-commander/

R Commander : Petit guide pratique 1. Statistiques de base Jean-Philippe Gaudron : https://cran.r-project.org/doc/contrib/Gaudron-GuideRcmdr.pdf

The R Foundation sur les réseaux sociaux

https://www.linkedin.com/company/the-r-foundation-for-statistical-computing?original_referer=https%3A%2F%2Fwww.r-project.org%2F

The R Foundation for Statistical Computing

Institute for Statistics and Mathematics
Wirtschaftsuniversität Wien Welthandelsplatz 1 1020 Vienna, Austria
Tel: (+43 1) 31336 4754

5 Exercices

5.1.1 Exercice

On demande sa couleur préférée à 102 personnes. 12 personnes répondent le rouge.

Quelle est l'intervalle de confiance à 95% au sein de la population du caractère « couleur préféré = rouge » ?

Correction

$$[p_n - t\sqrt{\frac{p_n \cdot q_n}{n}}, p_n + t\sqrt{\frac{p_n \cdot q_n}{n}}]$$

I=[0,06 ;0,18]

5.1.2 Clash of Clans™

Clash of Clans™ est un jeu de stratégie en ligne dont le but est de créer et développer un camp en combattant les autres joueurs. Clash of Clans™ est développé par SUPERCELL.

Exercice : les attaques des autres joueurs : la taille des armées

Le nombre de soldats que contiennent les armées des joueurs suit une loi normale de moyenne 150 et d'écart-type 30. Quelle est la probabilité pour que votre camp soit attaqué par une armée de plus de 200 soldats ?

Exercice : expérience des joueurs d'un clan

Les joueurs peuvent former des clans pour s'allier et affronter d'autres clans. Chaque joueur a une expérience d'espérance μ = 200, d'écart-type 50, suivant une loi normale. On prend un joueur au hasard parmi l'ensemble des joueurs du jeu. Quelle est la probabilité pour que son expérience soit inférieure à 210 ?

Exercice : pièges

Des pièges peuvent être installés pour repousser les ennemis. Ils se déclenchent automatiquement au passage de l'ennemi mais doivent être réactionnés manuellement.

Vous partez en vacances dans une zone sans WIFI et sans réseau mobile. Impossible de vous connecter à votre profil Clash of Clans pour réactionner vos pièges !

Juste avant votre départ, vous actionnez vos deux pièges.

Avec l'expérience, vous avez déduit que chaque jour, un piège a la probabilité p=0.8 de ne pas être actionné par un ennemi.

Xn est le nombre de pièges actionnés au début de la n-ième journée de vacances (n = 0, 1, ….).

a. Combien d'états possède cette chaîne de Markov ?

b. Déterminer les probabilités de transition correspondantes et réaliser le graphe des transitions.

c. Cette chaîne de Markov est-elle absorbante ? Si oui, quel est l'état absorbant et quelle est la distribution limite de cette chaîne ?

d. Ecrire la matrice de transition P

e. Les deux éléments sont en état de fonctionnement à n = 0. Quelle est la distribution du nombre d'éléments en panne après n jours ?

f. Combien de jours seront nécessaires pour passer de 0 à 2 machines en panne ?

CORRECTION

Exercice : les attaques des autres joueurs : la taille des armées (3 pts)

Le nombre de soldats que contiennent les armées des joueurs suit une loi normale de moyenne 150 et d'écart-type 30. Quelle est la probabilité pour que votre camp soit attaqué par une armée de plus de 200 soldats ?

Soit X une variable aléatoire qui suit la loi normale de moyenne 150 et d'écart-type 30

On cherche

$$P(X > 200)$$

Or

$$P(X > 200) = P(\frac{X - 150}{30} > \frac{200 - 150}{30})$$

Soit $Y = \frac{X-150}{30}$, alors

$$P(X > 200) = P(Y > 1.67) = 1 - P(Y < 1.67)$$

X suit la loi normale de moyenne 150 et d'écart-type 30 donc Y suit la **loi normale centrée réduite N(0,1)** donc par lecture sur la table de la loi normale centrée réduite

$$P(Y < 1.67) = 0.9525$$

donc

$$P(X > 200) = 1 - 0.9525 = 0.0475$$

Exercice : expérience des joueurs d'un clan (3 pts)

Les joueurs peuvent former des clans pour s'allier et affronter d'autres clans. Chaque joueur a une expérience d'espérance μ = 200, d'écart-type 50, suivant une loi normale. On prend un joueur au hasard parmi l'ensemble des joueurs du jeu. Quelle est la probabilité pour que son expérience soit inférieure à 210 ?

Soit X une variable aléatoire qui suit la loi normale de moyenne 200 et d'écart-type 50
On cherche

$$P(X < 210)$$

Or

$$P(X < 210) = P(\frac{X - 200}{50} < \frac{210 - 200}{50})$$

Soit $Y = \frac{X-200}{50}$, alors

$$P(X < 210) = P(Y < 0.2)$$

X suit la loi normale de moyenne 200 et d'écart-type 50 donc Y suit la loi normale centrée réduite donc par lecture sur la table de la loi normale centrée réduite

$$P(Y < 0.2) = 0.5793$$

Exercice : pièges (4 pts)

a. Combien d'états possède cette chaîne de Markov ?

3 états : 0, 1 et 2 pièges actionnés

b. Déterminer les probabilités de transition correspondantes et réaliser le graphe des transitions.

$$p_{00} = p(A \cap B) = p(A).p(B) = p.p = 0.8^2$$

$$p_{01} = p.(1-p) + p.(1-p) = 2p(1-p) = 2 * 0.8 * (1-0.8) = 2 * 0.8 * 0.2$$

$$p_{02} = p(piège\ A\ soit\ actionné) * p(piège\ B\ soit\ actionné) = 0.2^2$$

$$p_{10} = 0$$

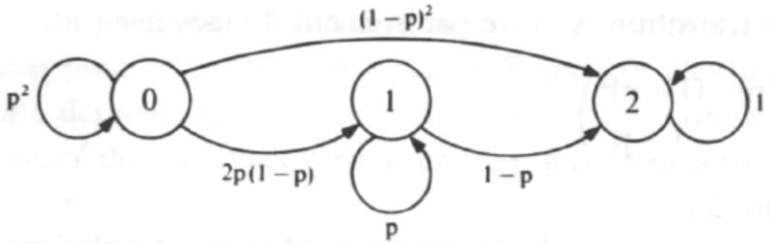
$$p_{20} = 0$$

$$p_{21} = 0$$

$$p_{11} = 0.8$$

$$p_{12} = 0.2$$

$$p_{22} = 1$$



p=0.8

c. Cette chaîne de Markov est-elle absorbante ? Si oui, quel est l'état absorbant et quelle est la distribution limite de cette chaîne ?

Oui, l'état 2 est absorbant. La distribution limite est donc $\pi = (0,0,1)$

d. Ecrire la matrice de transition P

$$P = \begin{pmatrix} p^2 & 2p(1-p) & (1-p)^2 \\ 0 & p & 1-p \\ 0 & 0 & 1 \end{pmatrix}$$

e. Les deux éléments sont en état de fonctionnement à n = 0. Quelle est la distribution du nombre d'éléments en panne après n jours ?

$$\pi(n) = \pi(0).p^n = (p^{2n}, 2p^n(1-p)^n, (1-p^n)^2)$$

f. Combien de jours seront nécessaires pour passer de 0 à 2 machines en panne ?

$$n_0 = 1 + p^2 n_0 + 2p(1-p)n_1$$

$$n_1 = 1 + p.n_1$$

$$n_0 = \frac{1+2p}{1-p^2}$$

$$n_1 = \frac{1}{1-p}$$

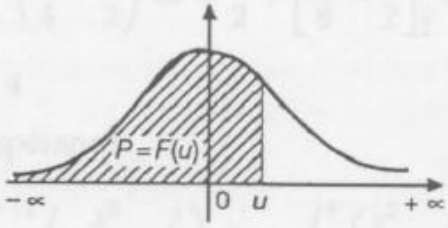
Si p=0.9, n0=14.7 jours et n1=10 jours

Si p=0.8, n0=7.22 jours

6 Annexes

Fonction de répartition de la loi normale centrée réduite

Probabilité $F(u)$ d'une valeur inférieure à u

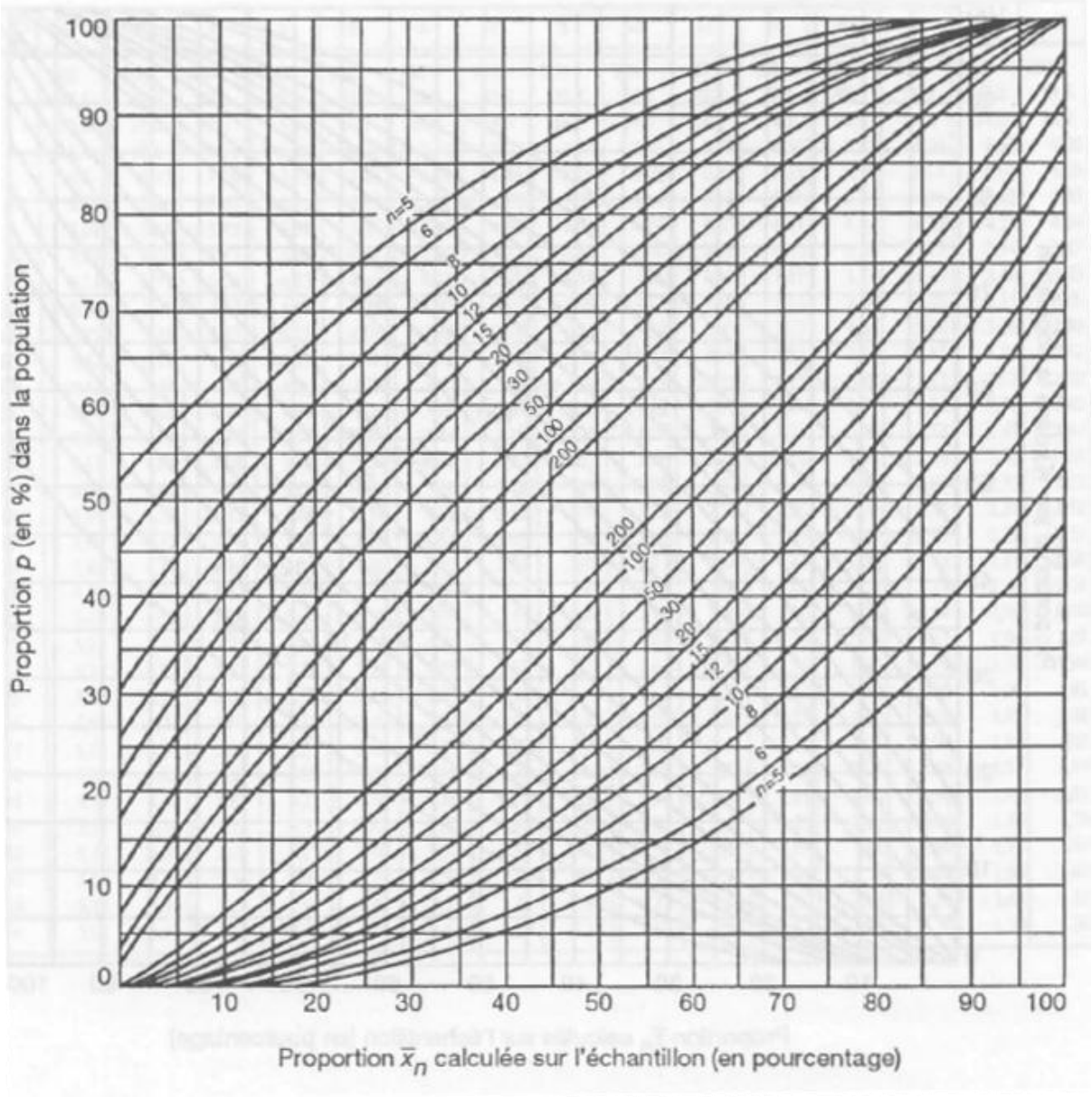


u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Tables pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$F(u)$	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Intervalles de confiance pour une proportion p
Intervalle bilatéral de niveau de confiance 0,95
Intervalles unilatéraux de niveau de confiance 0,975



Fractiles d'ordre P de la loi de Student T_ν

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

7 Références bibliographiques

[2], [3]

[1] *Gauss: une révolution de la théorie des nombres*. Barcelone: RBA Coleccionables, 2018.

[2] REDER, « Probabilités et statistiques. Cours et Exercices Bordeaux I ». 2002.

[3] J. Fourastié et J.-F. Laslier, *Probabilités et statistique*, 3e éd. Paris: Dunod, 1987.