

Notes de cours de Statistiques univariées

Jean-Etienne Poirrier*

26 septembre 2006

Table des matières

1	Introduction et définitions	3
1.1	Population, échantillon, unité et variable	3
1.2	Les trois types de variables	3
2	La statistique descriptive	4
2.1	L'approche graphique de réduction des données	4
2.1.1	Variable discrète	4
2.1.2	Variable continue	7
2.2	L'approche numérique de réduction des données	8
3	Paramètres numériques de réduction des données	8
3.1	Les paramètres de position	8
3.1.1	Le mode	9
3.1.2	La moyenne	9
3.1.3	La médiane	10
3.1.4	Le quantile	10
3.2	Les paramètres de variabilité	11
3.2.1	L'étendue	11
3.2.2	L'intervalle inter-quartiles	11
3.2.3	L'écart moyen	12
3.2.4	La variance	12
3.2.5	L'écart-type	12
3.3	Les paramètres de forme	14
3.3.1	Coefficiente de symétrie	14
3.3.2	Coefficient d'aplatissement	14
3.4	Les paramètres d'association	14
3.4.1	Covariance	15
3.4.2	Coefficient de corrélation	15

*Ces notes de cours de statistiques sont largement inspirées de mes notes prises au cours du professeur A. Albert pour les deuxièmes candidatures en biologie à l'Université de Liège (Belgique). Je ne suis donc pas statisticien et, si vous aviez des remarques, des suggestions ou si vous trouviez des erreurs, vous pouvez m'en faire part à l'adresse suivante : jepoirrier@gmail.com. La dernière version de ces notes se trouvent ici : <http://www.poirrier.be/jean-etienne/notes/>. Ces notes sont sous la GNU Free Documentation Licence dont le texte se trouve ici : <http://www.gnu.org/copyleft/fdl.html>

<i>TABLE DES MATIÈRES</i>	2
3.4.3 Droite de régression	16
3.4.4 Coefficient de détermination	16
4 Échantillonnage, probabilité et variables aléatoires	17
A Démonstration de la formule de la variance dans le cas où $n = 2$	18
B Démonstration de la formule de travail de la variance	18
C Démonstration de la formule de travail de la covariance	19

1 Introduction et définitions

R.A. Fisher a défini la statistique comme la **discipline qui étudie les méthodes de réduction de données, la variabilité et les populations**.

- Les méthodes de réduction des données font partie de la *statistique descriptive* (ou exploratoire). Elles consistent à essayer de résumer un échantillon de données via des graphiques ou des caractéristiques numériques. Elle est présentée en détail à la section 2.
- L'étude de la variabilité cherche à l'expliquer. Elle fait partie de la théorie de l'échantillonnage.
- L'étude des populations fait partie de la *statistique inférentielle* qui prend un échantillon et en tire des conclusions pour toute la population. Elle part donc de l'expérience à l'hypothèse (faite au départ).

1.1 Population, échantillon, unité et variable

Avant d'aller plus loin, il est important de définir clairement quelques termes ...

La **population** est un ensemble de sujets (= objets = éléments) qui ont au-moins une propriété en commun.

L'**échantillon** de la population est un sous-ensemble de la population. Cet échantillon doit être *représentatif* de la population.

L'**unité statistique** est l'élément de la population sur lequel on travaille. Par exemple, si on s'intéresse aux étudiants d'une école, l'unité sera l'étudiant.

Finalement, la **variable** est une grandeur caractéristique à laquelle on s'intéresse. Si on s'intéresse à une seule variable, on parlera de *statistique univariée*. Si on s'intéresse à deux ou plusieurs variables, on parlera de *statistique multivariée*.

1.2 Les trois types de variables

Il existe trois types de variables : les variables quantitatives, qualitatives et binaires.

Les **variables quantitatives** expriment une quantité : $x = 0, 1, 2, 3, \dots, n$. Elles sont donc mesurables, numériques. On les classe en variables quantitatives discrètes et variables quantitatives continues.

Une variable quantitative discrète peut être représenté par un nombre *fini* de valeurs. Ce sera, par exemple, le nombre d'enfants par famille, le nombre d'hospitalisations par patient, le nombre de pétales dans une fleur, etc. Ces valeurs peuvent être traitées mathématiquement (par exemple, par des opérations de base comme l'addition, la soustraction, etc.).

Une variable quantitative continue peut prendre *toutes* les valeurs possibles dans un intervalle donné $[a, b]$ ¹. Par exemple, le poids, la taille, l'âge, la concentration en ozone ou en calcium, ... sont des variables quantitatives continues. En effet, si je dis que je pèse 77 kg, c'est une approximation : je pèse, en réalité entre 76.5 et 77.5 kg ou entre 76.6 et 77.4 kg ou ...

Les **variables qualitatives** expriment une qualité ; ce sont des données catégorisées (on parlera aussi de variables nominales) : $x = m_1, m_2, m_3, \dots, m_q$. Les valeurs prises par la variable sont des *modalités*, se traduisant par des noms. Par exemple, la couleur de peau est une variable qualitative : on est blanc, jaune, noir, rouge, etc. Le groupe sanguin est un autre exemple de variable qualitative : on est A, B O ou AB mais rien d'autre.

Il arrive que l'on associe un chiffre à une modalité, généralement pour en faciliter l'encodage. Mais il faut bien faire attention qu'on ne peut pas les traiter mathématiquement !

¹Cette notation $[a, b]$ signifie un ensemble allant de "a" à "b" en incluant ces deux valeurs signifie un ensemble allant de "a" à "b" en incluant ces deux valeurs. La notation $]a, b[$ signifierait l'ensemble des valeurs comprises entre a et b avec seul b compris dans l'ensemble. Toutes les variations sont permises.

Parmi les variables qualitatives, il y a les variables ordinales dans lesquelles il y a un ordre dans les modalités : $m_1 < m_2 < \dots < m_p$. Par exemple, le grade à un examen est une variable qualitative ordinale : $AJ < S < D < GD < PGD^2$. On pourrait traiter ces variables mathématiquement car, en-dessous, il y a (ou il peut y avoir) une variable quantitative continue.

Pour être complet, signalons qu'il existe des variables quantitatives continues qu'on catégorise pour en faire des variables qualitatives. C'est moins bien. Exemple : vous avez entre 0 - 20 ans, 20 - 40 ans, 40 - 60 ans, 60+ ans (alors que l'âge est une variable quantitative continue).

Finalement, les **variables binaires** peuvent être de deux types. Soit c'est une variable qualitative qui ne prend que deux modalités (exemples : le sexe M/F, le statut de fumeur O/N, l'anomalie génétique : O/N). Soit c'est une variable quantitative discrète ne prenant que deux valeurs. On peut toujours alors la ramener à 0 ou 1 ($N = 2; x = 0/1$).

2 La statistique descriptive

L'objectif de la statistique descriptive est de résumer un échantillon de données. Au départ, on a l'échantillon et une variable X supposée quantitative. On désigne par n l'**effectif** de l'échantillon (en anglais : "*sample size*"). L'effectif est le nombre d'objets, de sujets, de personnes, ... dans l'échantillon. On représente l'*échantillon des données* dans un tableau brut des données de la manière suivante ³ :

$$\{x_1, x_2, x_3, \dots, x_n\}$$

Deux remarques à propos des données :

1. les *données manquantes* (*missing values*) doivent quand même être encodées. On choisit pour cela un signe ou une valeur particulière. Il faut donc indiquer au programme de ne pas prendre en considération ce signe ou cette valeur. Par exemple, dans le logiciel SAS, les données manquantes sont signalées par un point (".") mais d'autres logiciels utilisent éventuellement d'autres spécifications.
2. les *données censurées* (*censored values*) sont des valeurs qu'on n'a pas pu obtenir ou observer mais dont on a une borne (inférieure ou supérieure). Par exemple, en parlant du poids d'une personne, on peut ne pas avoir son poids mais être sûr qu'elle fait plus de 40 kg. Il ne faut pas laisser tomber ces données et trouver un moyen pour les encoder.

Pour résumer l'échantillon, la statistique descriptive dispose de deux moyens : l'approche graphique et l'approche numérique.

2.1 L'approche graphique de réduction des données

2.1.1 Variable discrète

Soit l'échantillon de données suivant : $\{x_1, x_2, x_3, \dots, x_n\}$. Si on le trie, on obtient un *échantillon ordonné* : $\{x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}\}$. Les chiffres en indice entre parenthèses indiquent le *rang de l'observation*, c.à.d. la position de la valeur dans l'échantillon s'il est trié par ordre croissant.

Il y a, ainsi, trois types de tableaux :

1. le *tableau brut* qui ne contient que les données telles que récoltées ;
2. le *tableau ordonné* qui contient les données triées par ordre croissant ;
3. le *tableau recensé* : $(x_1, x_2, x_3, \dots, x_n)$.

²AJ = ajourné ; S = satisfaction ; D = distinction ; GD = grande distinction ; PGD = plus grande distinction

³le chiffre en indice indique le numéro de la valeur dans l'échantillon

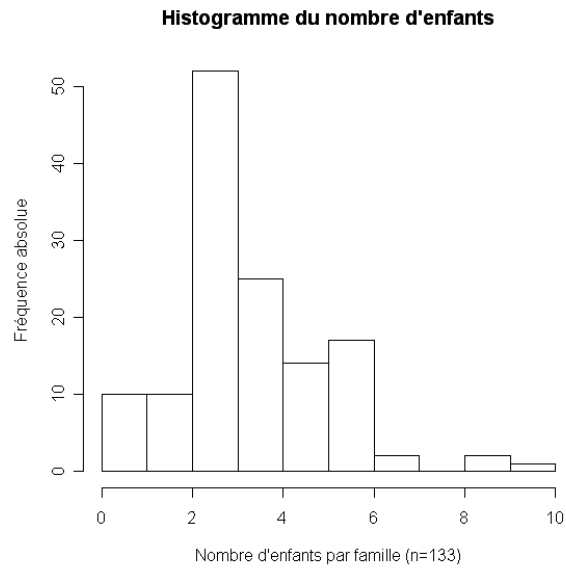


FIG. 1 – Histogramme des densités du nombre d'enfants par famille

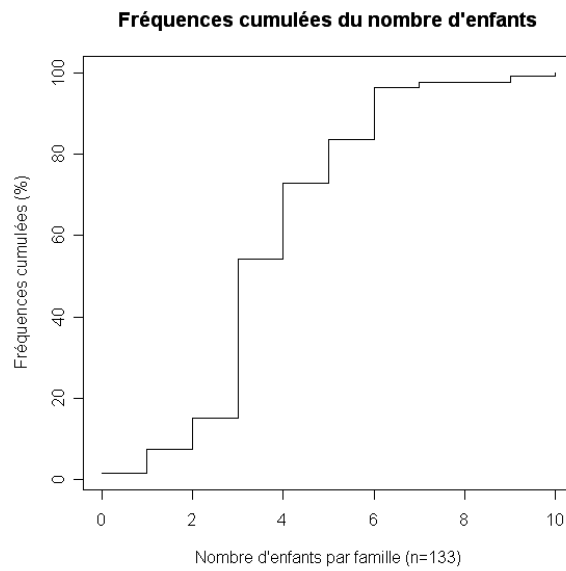


FIG. 2 – Diagramme des fréquences cumulées

10	22	24	42	37	77	89	85	28	63
9	10	7	51	2	1	52	7	48	54
32	29	2	15	46	48	39	6	72	14
36	69	40	61	12	21	54	53	58	32
27	33	1	25	22	6	81	11	56	5
63	53	88	48	52	87	71	51	52	33
46	33	85	22	5	87	28	2	85	61
16	42	69	7	10	53	33	3	85	8
51	60	58	9	14	74	24	87	7	81
30	76	7	6	27	18	17	53	70	49

TAB. 3 – Age à l’admission à l’hôpital (variable X) pour un échantillon de 100 patients

Classes d’âges (en années)	Centres C_i	Répétitions r_i	Fréquences f_i (en %)	Fréq. cumulées c_i (en %)
0-10	5	22	22	22
10-20	15	8	8	30
20-30	25	13	13	43
30-40	35	10	10	53
40-50	45	8	8	61
50-60	55	16	16	77
60-70	65	7	7	84
70-80	75	5	5	89
80-90	85	11	11	100
Total		$n = 100$	100	

TAB. 4 – Tableau de classes de l’âge à l’admission à l’hôpital chez 100 patients

2.1.2 Variable continue

Comme toujours, il va être plus facile de montrer l’approche graphique de réduction des données de variable continue par un exemple. Dans le tableau brut des données suivant

- l’unité statistique est la suivante : les patients entrant à l’hôpital,
- la variable (continue) est : l’âge en années.

Trier ce tableau sera lourd et peu intéressant (surtout si nous avons énormément de données). C’est pourquoi nous allons créer un **tableau de classes**. Dans ce tableau (voir tableau 4), on définit 10 classes (dans la colonne 1) : de 0 à 10 ans (inclus : 0-10), de 10 (exclus) à 20 ans (10-20), ... Théoriquement, on définira k classes pour son échantillon, où l’heuristique nous dit que $k = \sqrt{n}$ (avec $n =$ le nombre de données dans son échantillon).

Dans la deuxième colonne, on définira le *centre de la classe* (C_i). Comme le nom l’indique, le centre de la classe est la valeur numérique du milieu de la classe. Par exemple, le centre de la classe 0-10 est 5.

Dans la troisième colonne, on définira la *répétition* : $\sum r_i = n$. Ce nombre représente le nombre de valeurs continues se retrouvant dans chaque classe. Cette manière de procéder va plus vite que le classement “classique”. Sinon, on représentera encore les fréquences (f_i) et les fréquences cumulées (C_i).

Ce tableau des classes est donc caractérisés par les classes, C_i , r_i , f_i , c_i .

On peut dériver un premier graphique de ces tableaux : l’histogramme (ou diagramme d’aires) où on représente les fréquences (f_i , en ordonnées) en fonction des classes : $f_i - vs - classes$ (voir

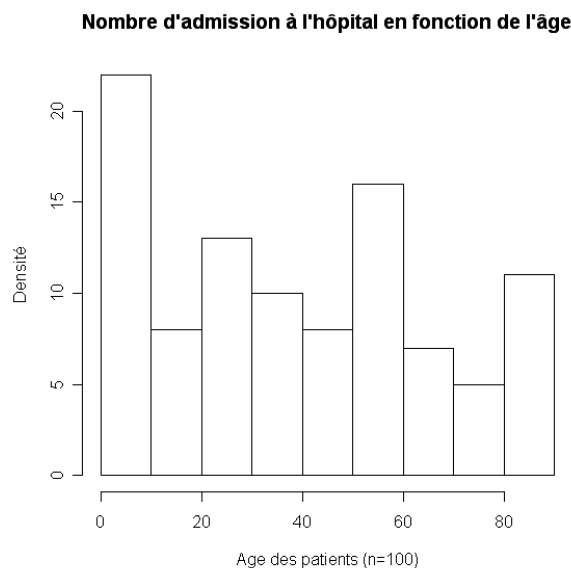


FIG. 3 – Histogramme des densités des âges d'admission à l'hôpital

graphique 3). Trois remarques :

1. les classes d'âges doivent être équidistantes ;
2. si on veut regrouper deux classes, on doit additionner les fréquences et la base de l'aire doit être agrandie (exemple, si on veut regrouper les 2 dernières classes, l'aire doit faire 16%) ;
3. il ne faut pas oublier d'indiquer le n de l'effectif.

Nous pouvons également dériver un second graphique de ces tableaux : le diagramme cumulé approché (voir graphique 4). Ici, cela donne beaucoup plus d'informations ; cela permet, par exemple, de répondre à la question "quelle est la proportion des gens qui ont tel âge ou plus / moins ?". Comme nous le verrons plus tard (section 3.1.3), nous avons une valeur particulière importante : la *médiane*, valeur où 50% des valeurs sont en-dessous et 50% des valeurs sont au-dessus.

2.2 L'approche numérique de réduction des données

Soit l'échantillon d'effectif n suivant :

$$\{x_1, x_2, x_3, \dots, x_n\}$$

Deux familles de paramètres vont pouvoir réduire les données numériquement : les paramètres de position et les paramètres de variabilité. D'autres familles de paramètres pourront être définies. Elles sont toutes décrites dans la section suivante.

3 Paramètres numériques de réduction des données

3.1 Les paramètres de position

Les paramètres de position sont le mode, la moyenne arithmétique, la médiane et le quartile. Ils permettent de savoir autour de quelles valeurs tournent les données de l'échantillon, de trouver

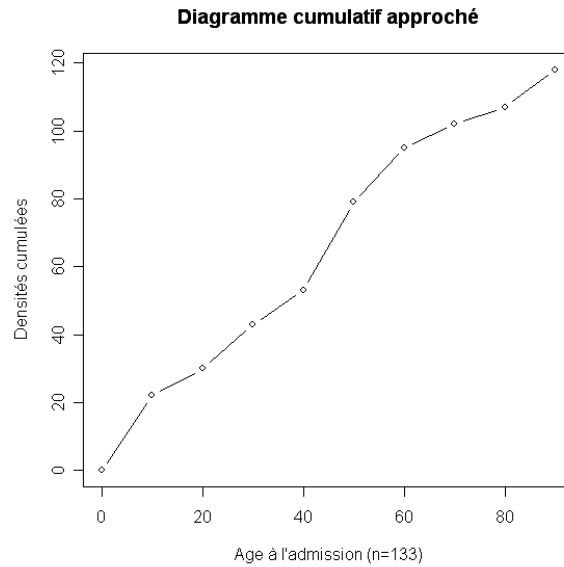


FIG. 4 – Diagramme cumulatif approché

une valeur centrale de l'échantillon.

3.1.1 Le mode

Le **mode** est la valeur la plus fréquente dans l'échantillon. Par exemple, pour le nombre d'enfants par famille, le mode est 3 (en d'autres termes, pour la variable discrète du nombre d'enfants par famille, la valeur la plus fréquente est 3). Par contre, pour l'âge d'admission à l'hôpital, la *classe modale* est 0-10 ans (cette classe de la variable continue contient le plus d'individus).

3.1.2 La moyenne

La **moyenne arithmétique** (en anglais : *mean, average*) est définie par l'équation suivante :

$$\bar{x} = m = \frac{\sum x_i}{n} \quad (1)$$

Dans l'exemple du nombre d'enfants par famille, $n = 133$ et $\sum x = 498$. Donc, $\bar{x} = 3.74$. Ce résultat est bizarre pour une variable discrète. On dira ici que la moyenne se situe entre 3 et 4, qu'elle est plus proche de 4 que de 3.

Dans l'exemple de l'âge d'admission à l'hôpital, $n = 100$ et $\sum x = 3920$. Donc, $\bar{x} = 39.2$ ans.

La moyenne arithmétique a les propriétés suivantes :

1. simplicité (d'emploi et de concept)
2. généralité (utilisé partout)
3. sensible aux valeurs aberrantes (erreurs de données)
4. si $X = 0/1$ avec $0 = \text{non fumeur}$ et $1 = \text{fumeur}$, $n = 100$: $\bar{x} = \frac{\text{nbrede}1}{n}$. Cela donne une proportion p ! Une proportion est donc une moyenne arithmétique de variables binaires. Une proportion peut donc être traitée comme une moyenne arithmétique.

3.1.3 La médiane

La **médiane** M est la valeur qui laisse 50% des observations en-dessous et 50% des observations au-dessus. On l'appelle également parfois “*percentile 50*” : c'est la valeur centrale par excellence.

Pour la calculer, il faut d'abord trier l'échantillon. Ensuite,

- si l'effectif n de l'échantillon est impair,

$$M = x_{(\frac{n+1}{2})}$$

Par exemple, si l'échantillon est (28, 14, 11, 21, 13), il devient, une fois trié : (11, 13, 14, 21, 28).

$$M = x_{(\frac{5+1}{2})} = x_3 = 14.$$

- si l'effectif n de l'échantillon est pair,

$$M = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Par exemple, si l'échantillon est (28, 14, 11, 13, 12, 23), il devient, une fois trié : (11, 12, 13, 14, 23, 28)

$$\text{et } M = \frac{x_{(\frac{6}{2})} + x_{(\frac{6}{2}+1)}}{2} = \frac{x_3 + x_4}{2} = \frac{13+14}{2} = 13.5.$$

La médiane a, comme propriété, d'être peu sensible aux valeurs extrêmes.

3.1.4 Le quantile

Le **quantile** α ⁶ est la valeur P_α qui laisse α % des observations en-dessous et $(1 - \alpha)$ % des observations au-dessus d'elle. Les deux “quartiles”⁷ les plus importants sont P_{25} (qui laisse 25 % des observations en-dessous) et P_{75} .

Ces deux quartiles peuvent également être définis de manière graphique. Si on reporte sur un graphique la fréquence des observations en fonction de ces observations, on obtient le graphique 5. P_{25} est la valeur en abscisse pour laquelle la droite d'équation $x = P_{25}$ découpe une aire représentant 25 % de l'aire totale sous les points.

Afin d'avoir un aperçu des données, on peut comparer la moyenne et la médiane. Trois cas sont possibles ...

1. Si $\bar{x} \simeq M$, nous nous trouvons dans le cas idéal : c'est un indicateur de symétrie. Afin d'obtenir cette courbe, nous pouvons éventuellement *normaliser* les données, c'est-à-dire leur appliquer une transformation comme $\ln x$ ou \sqrt{x} .
2. Si $\bar{x} \gg \gg M$, c'est
 - soit un indicateur d'erreur(s) dans les données,
 - soit signe d'une distribution dissymétrique à droite (c'est le cas, notamment, de la durée de vie (MTBF du néon, temps d'hospitalisation, etc.), d'études sur les enzymes, les hormones, ...).
3. Si $\bar{x} \ll \ll M$, c'est
 - soit, de nouveau, un indicateur d'erreur(s) dans les données,
 - soit signe d'une distribution dissymétrique à gauche (c'est le cas, notamment, de l'âge lors d'une intervention chirurgicale, si on intervient après un certain âge). On peut également normaliser ces données avec une transformation comme x^2 ou x^3 .

⁶A la place de quartile, on parle aussi de percentile

⁷les 4 quartiles découpent l'échantillon en 4 morceaux

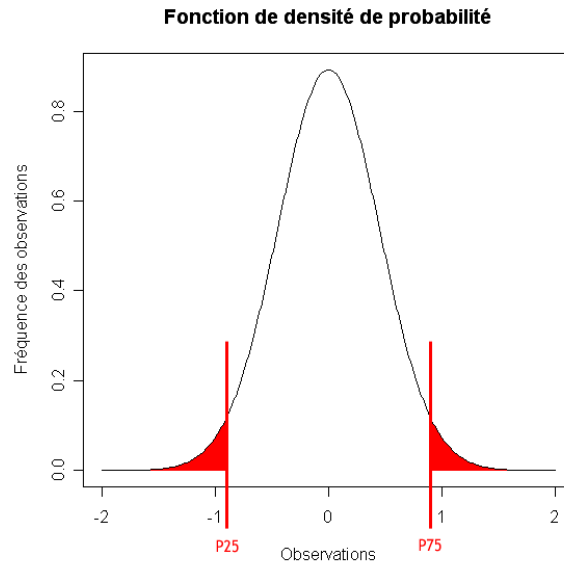


FIG. 5 – Représentation schématique des quartiles sur une fonction de densité de probabilité. Les barres verticales rouges représentent $\pm 2 \sigma$ (P₂₅ à gauche et P₇₅ à droite).

3.2 Les paramètres de variabilité

Ces paramètres permettent d'étudier la dispersion des observations. Leur objectif est de trouver un indicateur de cette dispersion.

Il faut noter qu'un indicateur de dispersion est toujours ≥ 0 . S'il n'y a pas de variabilité dans les observations, l'indice de dispersion = 0.

3.2.1 L'étendue

L'étendue (ou amplitude, en anglais : *range*) est l'écart entre la plus grande valeur et la plus petite valeur. Elle est définie par :

$$E = x_{(n)} - x_{(1)} \quad (2)$$

Comme énoncé précédemment, $E \geq 0$. Si $E = 0$, c'est que $x_{(n)} = x_{(1)}$.

Ce paramètre est simple mais très sensible aux valeurs extrêmes ou aberrantes !

3.2.2 L'intervalle inter-quartiles

L'intervalle inter-quartiles est défini par la relation suivante :

$$H = P_{75} - P_{25} \quad (3)$$

Dans ce cas, la relation $H \geq 0$ est toujours vérifiée puisque $P_{75} \geq P_{25}$.

On dit que ce paramètre est "robuste" car il est peu sensible aux valeurs extrêmes (c'est du au fait que les quartiles jouent avec les rangs et non les valeurs des observations).

3.2.3 L'écart moyen

L'écart moyen est défini comme

$$EM = \sum_{i=1}^n \frac{|x_i - \bar{x}|}{n}$$

On emploie une valeur absolue car, par définition de la moyenne, $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Même avec cette "précaution", $EM = 0$ si tous les $x_i = \bar{x}$. Ce paramètre est peu utilisé.

3.2.4 La variance

La **variance** est définie par la relation

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (4)$$

Cette variance a quelques propriétés intéressantes :

- $s^2 \geq 0$ (car on utilise un carré au numérateur); $s^2 = 0$ si $x_i = \bar{x}$.
- le numérateur est parfois appelé "somme des carrés" (*sum of squares*)
- le dénominateur est parfois appelé "degré de liberté" (*degree of freedom, d_f*)⁸
- les unités de s^2 sont celles des unités de X au carré (pratiquement, c'est inutilisable)
- s^2 est très sensible aux valeurs extrêmes (car elles affectent la moyenne)

Et aussi quelques cas particuliers intéressants :

- si $n = 1$, on ne peut calculer la variance; il faut donc au-moins 2 données pour pouvoir calculer une variance
- si $n = 2$, $s^2 = \frac{\Delta^2}{2} = \frac{(x_1 - x_2)^2}{2}$ (la formule de la variance dans le cas où $n = 2$ est démontré à la section A)

Afin de faciliter les calculs, il existe une "formule de travail" qui n'introduit pas d'erreurs d'arrondis et où il suffit de calculer $\sum x_i$ et $\sum x_i^2$ (la formule de travail de la variance est démontrée à la section B). Cette formule est toujours ≥ 0 :

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n - 1} \quad (5)$$

Si la variable est binaire ($X = 0/1$), les observations $\{x_1, \dots, x_n\}$ ne sont que des 0 et des 1. On a vu que $\bar{x} = p$ (une proportion). Dans ce cas, la variance devient (≥ 0) :

$$s^2 = p(1 - p)$$

3.2.5 L'écart-type

Suite à la difficulté d'interpréter la variance (notamment du au problème des unités, cf. plus haut), on a introduit l'écart-type (en anglais, *standard deviation*) dont la formule est :

$$s = +\sqrt{s^2} \quad (6)$$

En tenant compte de la formule de travail de la variance (équation 5), on peut écrire une représentation complète et pratique de l'écart-type, tant les équations 6 que 7 sont ≥ 0 :

⁸dans le dénominateur de la variance, on retire 1 à n car, si on connaît $x_{(n-1)}$, on connaît immédiatement x_n puisque $\sum x_i = \sum \bar{x}$

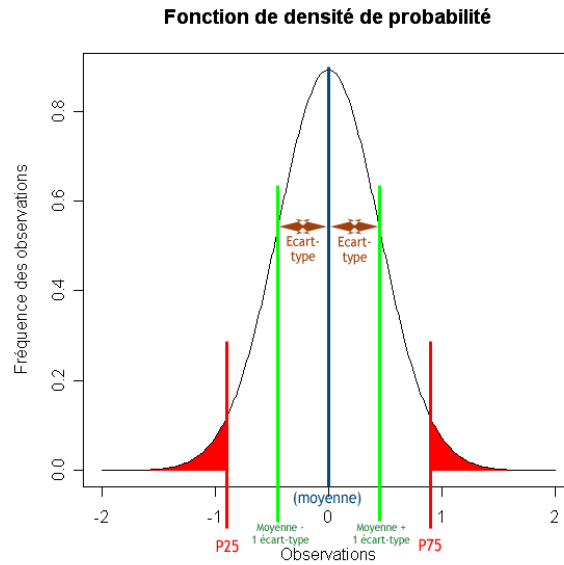


FIG. 6 – Distribution symétrique, normale, “gaussienne”

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1} \quad (7)$$

Remarquons que, si on a un échantillon $[x_1, x_2, x_3, \dots, x_n]$,

– et si on pose $y_i = x_i + a$ (avec $a = \text{constante}$), $\rightarrow \bar{y} = \bar{x} + a$ mais $s_y = s_x$: bien que la moyenne ait changé, la dispersion n’a pas changé !

– et si on pose $y_i = \lambda x_i$ (avec $\lambda = \text{constante} \geq 0$), $\rightarrow \bar{y} = \lambda \bar{x}$ et $s_y = \lambda s_x$: la dispersion a été multipliée également !

Reprenons l’exemple du nombre d’enfants par famille : $n = 133$, $\sum x = 498 \rightarrow \bar{x} = 3.74$, $\sum x^2 = 2238$, $s^2 = 2.8281$ (enfants² $\rightarrow s = 1.68$ (enfants)). Dans les familles, il y a donc en moyenne (3.74 ± 1.68) enfants par famille.

Si on reprend l’exemple de l’âge à l’admission à l’hôpital, $n = 100$, $\sum x = 3920 \rightarrow \bar{x} = 39.2$, $\sum x^2 = 224452$, $s^2 = 715.0303$ (ans² $\rightarrow s = 26.74$ (ans)). A l’entrée, il y a donc en moyenne des patients âgés de (39.2 ± 26.74) ans. Entre environ 10 et 60 ans, cette population est très variable !

Note importante : la moyenne (\bar{x}) et l’écart-type (s) sont les paramètres les plus importants ! La figure 6 montre une distribution symétrique, normale (au sens “gaussienne”) (cela ne fonctionne pas avec une distribution non symétrique !).

- entre $\bar{x} \pm 1s$: 68% des observations
- entre $\bar{x} \pm 2s$: 95% des observations
- entre $\bar{x} \pm 3s$: 99.9% des observations

Le **coefficient de variation** (en pourcents) quantifie ce que représente l’écart-type par rapport à la moyenne. Sa formule est :

$$CV = \frac{100s}{\bar{x}} \quad (8)$$

Cela permet, par exemple, de vérifier la reproductibilité de techniques. Ici, dans le cas d'un dosage, on a une erreur commise $CV = 5\%$; si $\bar{x} = 80\frac{g}{l} \rightarrow s = 4\frac{g}{l}$. On a donc $4\frac{g}{l}$ de variabilité sur $80\frac{g}{l}$.

3.3 Les paramètres de forme

Si nous avons l'échantillon $x_1, x_2, x_3, \dots, x_n$, nous pouvons obtenir très facilement \bar{x} et s . Maintenant, nous allons faire subir à cet échantillon une **transformation en valeurs centrées réduites** :

$$z_i = \frac{x_i - \bar{x}}{s}$$

Cette transformation sera donc appliquée à chacun des éléments de l'échantillon (pour $i = 1, 2, 3, \dots, n$). Le nombre obtenu en z_i est un nombre pur, sans unité. Grâce à cette transformation, nous aurons un nouvel échantillon :

$$z_1, z_2, z_3, \dots, z_n$$

Ce qu'il y a d'intéressant, avec ce nouvel échantillon, est que $\bar{z} = 0$ et $s_z^2 = 1$.

3.3.1 Coefficient de symétrie

Le coefficient de symétrie (*skewness* en anglais) est défini par la formule suivante :

$$g_1 = \frac{\sum_{i=1}^n z_i^3}{n} \quad (9)$$

- Si $g_1 = 0$, c'est un indicateur de symétrie ;
- Si $g_1 \gg 0$, c'est un indicateur de dissymétrie à droite ;
- Si $g_1 \ll 0$, c'est un indicateur de dissymétrie à gauche.

3.3.2 Coefficient d'aplatissement

Le coefficient d'aplatissement (*kurtosis* en anglais) indique si le sommet de la courbe est "pointu" ou "plat" et est défini par la formule suivante ⁹ :

$$g_4 = \frac{\sum_{i=1}^n z_i^4}{n} - 3 \quad (10)$$

- Si $g_4 = 0$, la courbe est "standard" ;
- Si $g_4 \gg 0$, la courbe est plus (trop) pointue ;
- Si $g_4 \ll 0$, la courbe est plus (trop) plate.

3.4 Les paramètres d'association

Ces paramètres permettent d'établir des relations entre des variables. Il leur faut donc au-moins 2 variables : X et Y (par exemple, le poids et la taille). On représentera alors le "nouvel" échantillon ainsi (échantillon "bivarié") :

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \quad (11)$$

⁹Le coefficient d'aplatissement est toujours ≥ 0

On peut prendre une approche graphique et indiquer autant d'axes qu'il y a de variables (graphique X -vs- Y , stéréogramme, etc.). Le plus simple est encore l'approche numérique, avec le calcul de paramètres comme la covariance, le coefficient de corrélation, les paramètres d'une droite de régression et le coefficient de détermination ...

3.4.1 Covariance

La covariance est un nombre réel (positif, négatif ou nul) donné par la formule suivante, pour les deux variables x et y :

$$\mathcal{S}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1} \quad (12)$$

Au numérateur, nous avons donc une somme de produits croisés et, au dénominateur, nous avons les degrés de liberté. Les unités de \mathcal{S} sont en fait un produit des unités de X par celles de Y ; leur compréhension est difficile, voire inutile. En fonction du signe de (\mathcal{S}), nous pouvons déterminer 3 options :

- Si $\mathcal{S}_{xy} > 0$, la relation entre X et Y est croissante
- Si $\mathcal{S}_{xy} < 0$, la relation entre X et Y est décroissante
- Si $\mathcal{S}_{xy} \approx 0$, il n'y a pas de relation ni d'association entre X et Y (graphiquement, on aura une sorte de "patate"). Dans ce cas, une variable n'a pas d'influence sur l'autre, et vice-versa (par exemple : une comparaison entre la taille d'individus et les deux derniers chiffres de leur carte d'identité)

Comme pour la variance, il existe une formule "de travail" qui facilite les calculs (sa démonstration se trouve à la section C) :

$$\mathcal{S}_{xy} = \frac{\sum_{i=1}^n xy - \frac{\sum x \cdot \sum y}{n}}{n - 1} \quad (13)$$

Notons, finalement, que $\mathcal{S}_{xy} = \mathcal{S}_{yx}$ et que, si $X = Y$, $\mathcal{S}_{xy} = \frac{\sum xy - \frac{\sum(x) \cdot \sum y}{n}}{n - 1}$, c'est-à-dire : la variance ! La covariance d'une variable par elle-même est donc la variance de cette variable (covariance \supset variance).

3.4.2 Coefficient de corrélation

Le coefficient de corrélation entre deux variables est un nombre réel (positif, négatif ou nul) pur (sans unité). Il représente la corrélation divisée par le produit des écart-types :

$$r = r_{xy} = \frac{\mathcal{S}_{xy}}{s_x \cdot s_y} \quad (14)$$

Le signe de r peut nous renseigner déjà sur le sens de la relation entre les deux variables :

- Si $r > 0$, la relation est linéaire croissante
- Si $r < 0$, la relation est linéaire décroissante
- Si $r \approx 0$, il n'y a pas de relation

On peut également montrer que $-1 \leq r \leq +1$ ainsi que, si $r = +1$, la relation est linéaire croissante parfaite mais, si $r = -1$, la relation est linéaire décroissante parfaite. Il va de soi que la corrélation d'une variable avec elle-même est parfaite : $r_{xx} = 1$. Et, de nouveau, le coefficient de corrélation possède une formule de travail :

$$r = \frac{\sum_{i=1}^n xy - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \cdot \left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} \quad (15)$$

Quelques derniers conseils ... Il faut faire attention aux valeurs aberrantes qui falsifient la corrélation. Il faut donc toujours regarder les données avant de calculer r ! Enfin, ce coefficient de corrélation est valable pour une relation linéaire entre les deux variables : d'autres types de relations peuvent exister ...

3.4.3 Droite de régression

Lorsqu'on examine 2 variables (X et Y), 2 situations sont possibles :

1. X et Y sont observés simultanément, X et Y sont des variables aléatoires (par exemple, le poids et la taille). Dans ce cas, on utilisera un coefficient de corrélation r et 2 droites de régression (qui ne seront pas les mêmes).
2. X est fixé par l'utilisateur et Y est observé, X est une variable mathématique et Y est une variable aléatoire. Dans ce cas, on utilisera un coefficient de détermination r^2 et une seule droite de régression (Y sur X), appelée droite des moindres carrés.

Dans le cas d'une droite des moindres carrés, la moyenne de Y ¹⁰ sera de type $a + b \cdot x$. a est appelé l'ordonnée à l'origine (équation 17) et b , la pente de régression (*slope*, en anglais, (équation 16)) :

$$b = r \cdot \frac{s_x}{s_y} = \frac{\sum_{i=1}^n xy - \frac{\sum x \cdot \sum y}{n}}{\sum_{i=1}^n y^2 - \frac{(\sum y)^2}{n}} \quad (16)$$

$$a = \bar{a} - b \cdot \bar{x} \quad (17)$$

Par contre, nous aurons une droite de régression de X sur Y uniquement dans le premier cas, celui où X et Y ne sont pas fixés mathématiquement. On retrouve une équation de droite de même type : $\bar{X} \equiv$ moyenne de $X = a' + b' \cdot y$, avec :

$$b' = r \cdot \frac{s_x}{s_y} = \frac{\sum_{i=1}^n xy - \frac{\sum x \cdot \sum y}{n}}{\sum_{i=1}^n y^2 - \frac{(\sum y)^2}{n}} \quad (18)$$

$$a' = \bar{a} - b' \cdot \bar{x} \quad (19)$$

Les deux droites de régression (X sur Y et Y sur X) se coupent au point moyen (\bar{x}, \bar{y}) .

Grâce à cette droite de régression, nous pouvons prédire (si $x = x_0$, que vaut y ?) et extrapoler des valeurs.

3.4.4 Coefficient de détermination

Ce coefficient de détermination donne la proportion (le pourcentage de la variabilité d'une variable qui serait expliquée par l'autre variable. Il se calcule comme ceci :

$$r^2 = b \cdot b' \quad (20)$$

¹⁰droite telle qu'en moyenne, elle passe par les points donnés

Par exemple, si on trouve un coefficient de corrélation $r = 0.8$ entre le poids et la taille d'un groupe, le coefficient de détermination sera $r^2 = 0.64$: 64% de la variabilité de la taille est expliquée par le poids, 35% reste inexpliquée (par le poids).

La suite au prochain numéro ...

4 Enchantillonnage, probabilité et variables aléatoires

A Démonstration de la formule de la variance dans le cas où $n = 2$

Soit l'échantillon $\{a, b\}$. L'effectif est donc $n = 2$ (variables quantitatives) et la moyenne est, par définition, $c = \frac{a+b}{2}$.

La théorie dit que la formule de la variance est $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. Appliquons-la à notre cas particuliers ...

$$s^2 = \frac{(a-c)^2 + (b-c)^2}{1} \quad (21)$$

$$= a^2 + c^2 - 2ac + b^2 + c^2 - 2bc \quad (22)$$

$$= a^2 + b^2 + 2c^2 - 2c(a+b) \quad (23)$$

$$= a^2 + b^2 + 2\left(\frac{a+b}{2}\right)^2 - 2\frac{a+b}{2}(a+b) \quad (24)$$

$$= a^2 + b^2 - (a+b)^2 + \frac{(a+b)^2}{2} \quad (25)$$

$$= a^2 + b^2 - \frac{a^2}{2} - \frac{b^2}{2} - \frac{2ab}{2} \quad (26)$$

$$= \frac{a^2}{2} + \frac{b^2}{2} - ab \quad (27)$$

$$= \frac{a^2 + b^2 - 2ab}{2} \quad (28)$$

$$= \frac{(a-b)^2}{2} \quad (29)$$

Cqfd

B Démonstration de la formule de travail de la variance

Soit l'échantillon $\{x_1, x_2, x_3, \dots, x_n\}$. L'effectif est donc n (variables quantitatives) et la moyenne est \bar{x} .

La théorie dit que la formule de la variance est $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. Essayons de trouver quelque chose de plus fonctionnel ...

$$s^2 = \frac{1}{n-1} \cdot [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] \quad (30)$$

$$= \frac{1}{n-1} \cdot [(x_1^2 + \bar{x}^2 - 2x_1\bar{x}) + (x_2^2 + \bar{x}^2 - 2x_2\bar{x}) + \dots + (x_n^2 + \bar{x}^2 - 2x_n\bar{x})] \quad (31)$$

$$= \frac{1}{n-1} \cdot [(x_1^2 + x_2^2 + \dots + x_n^2) + n\bar{x}^2 - 2\bar{x}(x_1 + x_2 + \dots + x_n)] \quad (32)$$

$$= \frac{1}{n-1} \cdot \left[\sum(x_i^2) + n\left(\frac{\sum x_i}{n}\right)^2 - 2\left(\frac{\sum x_i}{n}\right)(\sum x_i) \right] \quad (33)$$

$$= \frac{1}{n-1} \cdot \left[\sum(x_i^2) + \frac{(\sum x_i)^2}{n} - \frac{2(\sum x_i)^2}{n} \right] \quad (34)$$

$$= \frac{1}{n-1} \cdot \left[\sum (x_i^2) - \frac{(\sum x_i)^2}{n} \right] \quad (35)$$

$$= \frac{\sum (x_i^2) - \frac{(\sum x_i)^2}{n}}{n-1} \quad (36)$$

Cqfd! Il suffit donc de connaître n et de calculer $\sum x_i$ et $\sum x_i^2$ pour avoir la variance.

C Démonstration de la formule de travail de la covariance

A écrire ...