Cours de Statistiques

Focus sur le Chapitre 1: Rappels

Licence 3 – Parcours Gestion et Finance

Stéphane ROBIN
Université Paris 1 Panthéon-Sorbonne
Département de Gestion – EM Sorbonne

Contenu du Chapitre 1

1. Vocabulaire fondamental
2. Types de variables statistiques
3. Paramètres de position et de forme
4. Mesures d'association (covariance et corrélation)
5. Introduction à l'inférence statistique et rappels sur les probabilités

1. Vocabulaire statistique fondamental

Outre les concepts fondamentaux de population et d'échantillon déjà rencontrés dans l'introduction, la statistique utilise la terminologie suivante:

VARIABLE STATISTIQUE (VS) ou CARACTERE
Application qui à chaque individu de la population associe une valeur.

MODALITES
Ce sont les différentes valeurs que peut prendre une variable statistique.

Vocabulaire statistique fondamental

TAILLE ou EFFECTIF TOTAL
Nombre d'individus de la population ou de l'échantillon

EFFECTIF D'UNE MODALITE
Nombre d'individus qui présentent cette modalité particulière d'une VS.

FREQUENCE D'UNE MODALITE (souvent notée f_i)
Effectif de cette modalité divisé par l'effectif total. C'est la **proportion** d'individus qui présentent cette modalité.

Vocabulaire statistique fondamental

DEFINITION OPERATIONNELLE
Les modalités d'une variable n'ont de sens que si la variable a une **définition opérationnelle**, c'est-à-dire une signification claire et acceptée universellement.

Dans le cadre d'un problème ou d'une étude statistique, il importe donc de toujours bien préciser la nature de chacune des variables rencontrées.

2. Types de VS (ou caractères)

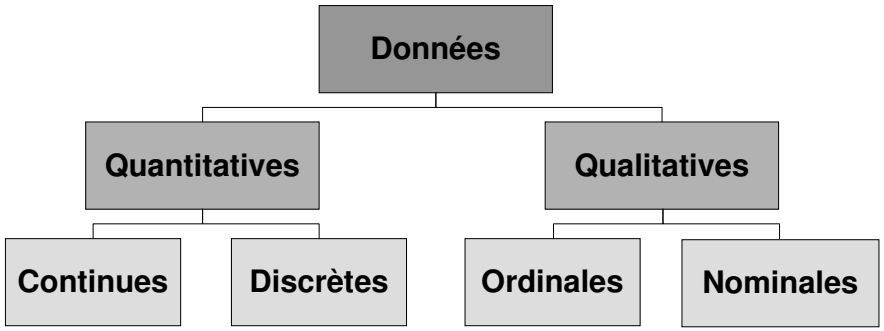
Le type d'une VS est défini à partir de ses **modalités**

Une VS est **quantitative** si ses modalités représentent des quantités (souvent exprimées en unités de mesure, comme le gramme, le mètre ou l'euro).

Elle est **qualitative** si ses modalités peuvent seulement être classées dans des catégories. Exemples:

"oui" et "non"
"souvent", "parfois" et "jamais"

Types de Variables



Echelle des variables qualitatives

Pour les **variables qualitatives**, il existe deux niveaux (ou **échelles**) de mesure :

1. L'échelle **ordinale**
2. L'échelle **nominale**

Attention à leur utilisation !

Echelle ordinale

Une **échelle ordinale** classe les données dans des catégories distinctes pour lesquelles un ordre existe.

Variables qualitatives		*Catégories ordonnées*
Classe d'âge	⟷	Junior / Moyen / Senior
Satisfaction produit	⟷	Insatisfait / Neutre / Satisfait
Notes élèves du primaire	⟷	A / B / C / D / E

Echelle nominale

Une **échelle nominale** classe les données dans des catégories distinctes entre lesquelles aucun ordre n'est sous-entendu.

Variables qualitatives		*Catégories*
Possède un PC	⟷	Oui / Non
Type d'actions détenues	⟷	Croissance / Autres
Fournisseur d'accès Internet	⟷	Orange / SFR / Free

3. Paramètres de position et de forme

Ce sont des indicateurs statistiques qui nous renseignent sur la **distribution** des données (certains correspondent aux **moments** de la distribution):

La **tendance centrale** ou **position** indique autour de quelle valeur les données sont groupées.

La **dispersion** mesure la variation des données autour de la valeur centrale

L'**allure** de la distribution informe sur l'étendue de ses valeurs, de la plus petite à la plus grande.

3.1. Les indicateurs de position

Moyenne arithmétique

Médiane

Mode

Quantiles

Moyenne géométrique (utile en finance)

La moyenne arithmétique

La **moyenne arithmétique** est la mesure la plus commune de tendance centrale (au point que le langage courant la désigne simplement comme "la moyenne").

Moyenne arithmétique = somme des valeurs divisée par nombre de valeurs

La moyenne arithmétique est donc sensible aux minimum et maximum de la distribution

Elle peut donc être affectée par les valeurs extrêmes ("points aberrants"), auxquelles on doit prêter attention quand on la calcule.

Formule de la moyenne arithmétique

Notée *m* (ou μ) dans la population et "X barre" dans un échantillon, la moyenne arithmétique se calcule de la même manière dans les deux cas. Dans une population de taille *N*:

$$m = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + ... + X_N}{N}$$

Et dans un échantillon de taille *n*:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + ... + X_n}{n}$$

Dans les deux cas, X_i désigne la $i^{ème}$ valeur de la VS X

Moyenne et valeurs extrêmes: exemple

Avec l'échantillon ci-dessous, on trouve **"X barre" = 3.94**:

i	1	2	3	4	5
X	3.0	4.5	5.2	3.0	4.0

Avec cet autre échantillon, tiré de la même population, on trouve **"X barre" = 5.14**:

i	1	2	3	4	5
X	3.0	4.5	5.2	3.0	10.0

Si l'on supprime la dernière observation (i=5), car on considère sa valeur comme aberrante, on trouve une moyenne proche de celle obtenue de le premier échantillon: **"X barre" = 3.925**.

La médiane

La **médiane** est une autre mesure de tendance centrale, qui **n'est pas affectée par les valeurs extrêmes**

La médiane est la valeur centrale d'une **série ordonnée** (50% des valeurs sont au-dessus et 50% au-dessous)

La médiane d'une série ordonnée de 1 à n se situe au niveau de la "(n+1)/2"^{ème} valeur

Calcul de la médiane

Si le nombre de valeurs est **impair**, la médiane est la valeur au milieu de la série.

Si le nombre de valeurs est **pair**, la médiane est la moyenne des deux valeurs au milieu de la série.

Attention, **(n+1)/2** n'est pas la valeur de la médiane, il s'agit de la **position de la médiane** dans la **série ordonnée**.

Médiane et valeurs extrêmes: exemple

Soit un premier échantillon:

i	1	2	3	4	5
X	2.0	2.5	4.0	4.5	5.0

Position de la médiane: (5+1)/2=3; Médiane = 4

Soit un second échantillon tiré de la même population :

i	1	2	3	4	5
X	2.0	2.5	4.0	6.0	7.0

Position de la médiane: (5+1)/2=3; Médiane = 4

Le mode (1)

Le **mode** d'une série de données est la valeur la plus fréquente dans cette série.

Le mode n'est pas affecté par les valeurs extrêmes

Le mode peut être utilisé pour les variables quantitatives et qualitatives

Il est possible qu'une série ne comporte pas de mode

Le mode (2)

Il est possible qu'une série comporte **plusieurs modes.** On parle alors de série **plurimodale** (par opposition à une série unimodale, qui ne comporte qu'un seul mode).

Pour des données groupées dans des classes, on parle de **classe modale** (classe la plus fréquente)

Dans ce cas, la détermination du mode comporte une part d'arbitraire, liée au choix des classes.

Le mode: exemple de calcul

- Soit la série suivante, qui comporte 13 observations :

X	1	9	3	7	5	10	5	12	9	13	12	14	9

- Identifions ses modalités et l'effectif de chacune d'elles :

Modalité	1	3	5	7	9	10	12	13	14
Effectif	1	1	2	1	3	1	2	1	1

- Le mode est la modalité (valeur) la plus fréquente, soit ici 9

Le mode: autres exemples

- Exemple de série ne comportant pas de mode :

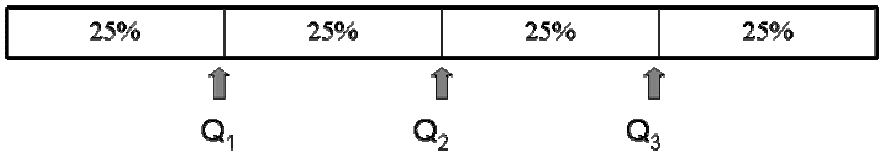
Modalité	2	3	5	7	10	12
Effectif	1	1	1	1	1	1

- Exemple de série multimodale:

Modalité	1	3	5	7	9	10	12	13	14
Effectif	1	1	3	1	3	1	3	1	1

Autres indicateurs de position: les quartiles

Les quartiles séparent une série **ordonnée** en 4 segments de taille égale (c-à-d. contenant le même nombre de valeurs). Il y a 3 quartiles notés Q_1, Q_2 et Q_3, du plus faible au plus élevé :



La taille de chacun des 4 segments représente ¼ (soit 25%) de la taille de la série.

Les trois quartiles

- Le premier quartile, $\mathbf{Q_1}$, est la valeur qui divise l'échantillon selon les proportions ¼ et ¾ : 25% des observations sont inférieures à Q_1, et 75% lui sont supérieures.

- Le second quartile, $\mathbf{Q_2}$, est la valeur qui divise l'échantillon en deux parties égales : 50% des observations sont inférieures à Q_2, et 50% lui sont supérieures. $\mathbf{Q_2}$ **est donc la médiane**.

- Le dernier quartile, $\mathbf{Q_3}$, est la valeur qui divise l'échantillon selon les proportion ¾ et ¼ : 75% des observations sont inférieures à Q_3, et 25% lui sont supérieures.

Remarque importante sur les quartiles

- Les quartiles $\mathbf{Q_1}$ et $\mathbf{Q_3}$ sont des certes des indicateurs de position, mais pas des indicateurs de position centrale.
- Seul le second quartile, $\mathbf{Q_2}$, est un indicateur de centralité, et ceci parce qu'il correspond à la médiane.
- On peut considérer que le second quartile, Q_2, est simplement un autre nom de la médiane.

Situation des quartiles

On peut trouver où se situent les quartiles dans la série ordonnée à l'aide des formules ci-dessous.

Position du premier quartile Q_1 : (n+1)/4 ème valeur

Position du second quartile (médiane) Q_2 : (n+1)/2 ème valeur

Position du troisième quartile Q_3 : 3(n+1)/4 ème valeur

où n est le nombre d'observations dans la série.

Principes d'application des formules

1. Si le résultat est un nombre entier, le quartile est égal à l'observation située à ce rang.
2. Si le résultat est une fraction dont le dénominateur est 2 (1.5, 2.5, 3.5, etc.), le quartile est égal à la moyenne des deux observations dont les rangs encadrent cette fraction.
3. Sinon, le résultat est arrondi à l'entier le plus proche et on applique le premier principe.

Exemple: situation du 1er quartile

Trouver Q_1 dans la série **ordonnée** suivante:

i	1	2	3	4	5	6	7	8	9
X	11	12	13	16	16	17	18	21	22

On remarque que **n = 9**.

Q_1 se situe donc au **(9+1)/4 = 2.5ème rang** de la série ordonnée, on retient donc la moyenne des 2nde et 3ème observations.

Donc **Q_1 = 12.5**

Généralisation : les quantiles

Les **quartiles** Q_1, Q_2, Q_3 sont les **quantiles d'ordre 25%, 50% et 75%**. Dans la série ordonnée, 25% des observations précèdent Q1 et 75% des observations suivent Q1.

Les **déciles** D_1, D_2, …, D_9 sont les **quantiles d'ordre 10%, 20%, … 90%.** Dans la série ordonnée, 20% des observations précèdent D_2 et 80% des observations suivent D_2.

Les **centiles** C_1, C_2, …, C_{99} sont les **quantiles d'ordre 1%, 2%, …, 99%**. Dans la série ordonnée, 1% des observations précèdent C_1 et 99% des observations suivent C_1.

La moyenne géométrique

La **moyenne géométrique** est un indicateur de position moins courant que la moyenne arithmétique. Elle sert surtout à mesurer le taux de variation moyen d'une variable dans le temps. Sur un échantillon de taille *n*, sa formule est:

$$\overline{X}_G = (X_1 \times X_2 \times ... \times X_n)^{1/n}$$

ou encore, sur une série temporelle avec $t = 1, \ldots \tau$:

$$\overline{X}_G = (X_1 \times X_2 \times ... \times X_\tau)^{1/\tau}$$

3.2. Les indicateurs de dispersion

La **dispersion** indique l'étendue ou la variation des observations (autour d'un indicateur de centralité) dans une série de données. Les indicateurs passés en revue ici sont:

- L'écart absolu ou étendue
- L'écart interquartiles
- La variance
- L'écart-type
- Le coefficient de variation
- Le score Z

L'écart absolu ou étendue

C'est la mesure de dispersion la plus simple. Elle est égale à la différence entre la plus grande et la plus petite valeur :

$$\text{Ecart absolu} = X_{\text{maximum}} - X_{\text{minimum}}$$

Exemple:

Modalité de X	1	2	4	6	8	9	11	12	13
Effectif	1	1	2	1	3	1	2	1	1

Ecart absolu (étendue) = 13 - 1 = 12

Inconvénients de l'écart absolu

1. L'écart absolu ignore la distribution des données:

X	7	8	9	10	11	12
Eff.	1	1	1	1	1	1

Ecart absolu = 12 - 7 = 5

X	7	8	9	10	11	12
Eff.	1	0	0	1	1	3

Ecart absolu = 12 - 7 = 5

2. L'écart absolu est sensible aux valeurs extrêmes:

Variable X1	**1**	1	1	1	2	2	2	2	3	3	3	4	**5**
Variable X2	**1**	1	1	1	2	2	2	2	3	3	3	4	**120**

Ecart absolu X_1 = 5 - 1 = 4

Ecart absolu X_2 = 120 - 1 = 119

L'écart interquartiles

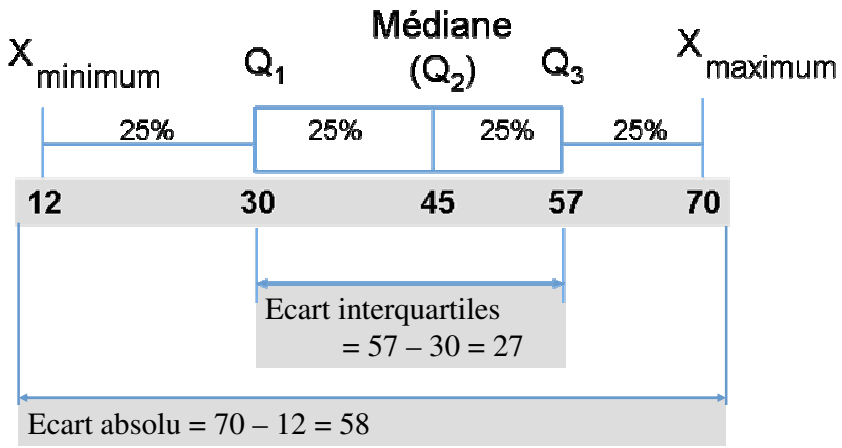
Les problèmes posés par les valeurs extrêmes peuvent être éliminés en utilisant l'écart interquartiles.

L'écart interquartiles élimine les valeurs hautes et basses et se calcule comme l'écart absolu sur les observations restantes.

Ecart interquartiles = Troisième quartile – premier quartile

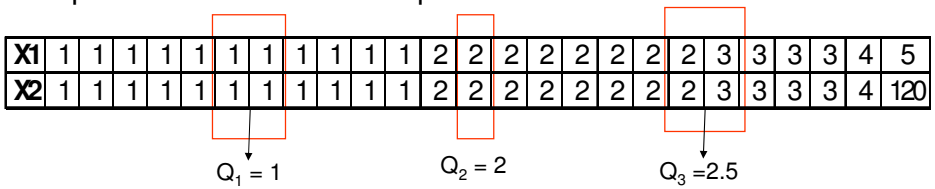
$= Q_3 - Q_1$

L'écart interquartile: exemple



Utilisation de l'écart interquartiles

De par sa construction, l'écart interquartile est très utile pour comparer deux séries statistiques:



Ecart Interquartiles $X_1 = Q_3 - Q_1 = 2,5 - 1 = 1,5$

Ecart Interquartiles $X_2 = Q_3 - Q_1 = 2,5 - 1 = 1,5$

Variance de la population

La **variance de la population**, notée σ, est la moyenne des carrés des écarts à la moyenne (c-à-d. des écarts entre chaque observation et la moyenne):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - m)^2$$

Avec m = moyenne de la population

N = taille de la population

X_i = i^{ème} valeur de la variable X

Variance empirique d'un échantillon

La **variance empirique**, notée S², se calcule sur un échantillon. C'est la moyenne (approximative) des carrés des écarts entre les observations et leur moyenne. Sa formule se distingue de celle de la variance de la population par son dénominateur:

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

Avec $\overline{X}$ = moyenne arithmétique

n = taille de l'échantillon

X_i = i^{ème} valeur de la variable X

La variance empirique: formule alternative

En récrivant la variance empirique comme:

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

on démontre facilement que:

$$S^2 = \frac{n}{n-1}\left[\overline{X^2} - \overline{X}^2\right]$$

Ecart-type de la population

L'**écart-type de la population**, noté σ, est la racine carrée de la variance de la population. Il est exprimé dans la même unité que les données de la population:

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

Avec m = moyenne de la population

N = taille de la population

X_i = i^{ème} valeur de la variable X

L'écart-type empirique

L'**écart-type empirique** est la mesure de dispersion la plus couramment utilisée. Il est défini comme la racine carrée de la variance empirique:

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

L'écart-type empirique mesure la **'dispersion moyenne' autour de la moyenne**. Il est exprimé dans la **même unité** que les observations ayant servi à son calcul.

Ecart-type empirique: exemple

Echantillon de temps de trajet quotidien (en minutes):

Jour (i)	1	2	3	4	5	6	7	8	9	10
Temps de trajet (Xi)	39	29	43	52	39	44	40	31	44	35

Sur cet échantillon, "X barre" = 39.6, S^2 = 45.82 et S = 6.77.
Interprétation: *"Un trajet dure en moyenne 39,6 minutes avec un écart-type de 6,77 minutes".*

Remarque: ici la somme des écarts à la moyenne est nulle car les écarts négatifs et positifs se compensent. D'où l'importance d'utiliser la somme des **carrés** des écarts à la moyenne dans les formules de la variance et de l'écart-type!

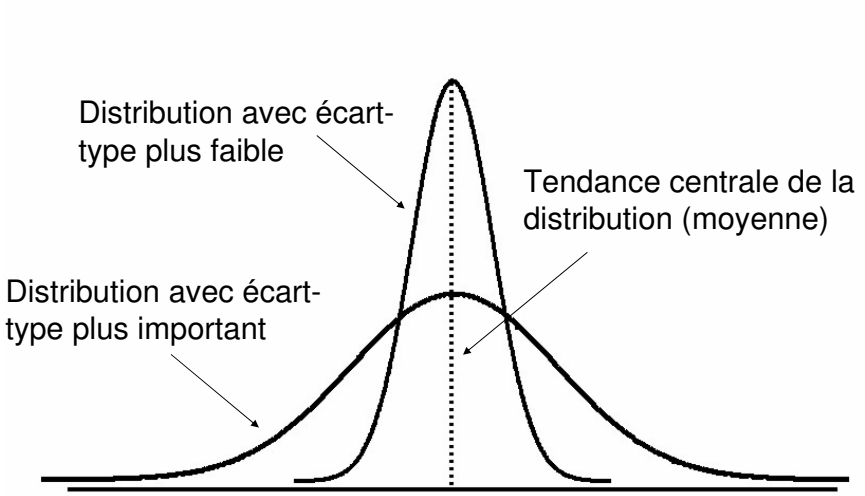
Remarques

- Plus une série statistique est dispersée, plus les indicateurs de dispersion précédents (écart absolu, écart interquartile, variance et écart-type) sont élevés.
- Inversement, ces indicateurs de dispersion sont faibles quand une série est peu dispersée.
- Si toutes les observations étaient identiques ces indicateurs seraient nuls: il n'y aurait pas de dispersion.
- Par construction, ces indicateurs ne peuvent pas être négatifs!

Ecart-type et courbe "en cloche"

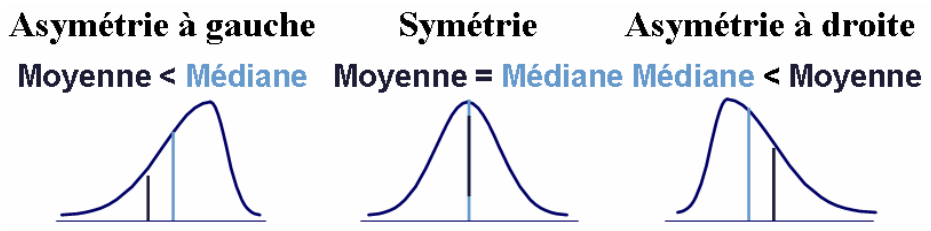
- Graphiquement, la distribution d'une série statistique peut souvent être représentée par une courbe "en cloche" (ou courbe de Gauss) **à partir du moment où** le nombre d'observations devient suffisamment grand (on parle alors de "distribution normale« , on y reviendra plus loin).
- Dans ce contexte, une plus grande dispersion de la série (c'est-à-dire un écart-type empirique plus élevé) se traduit par une courbe en cloche plus "plate" et "large".

Ecart-type et courbes "en cloche"



3.3. Allure d'une distribution

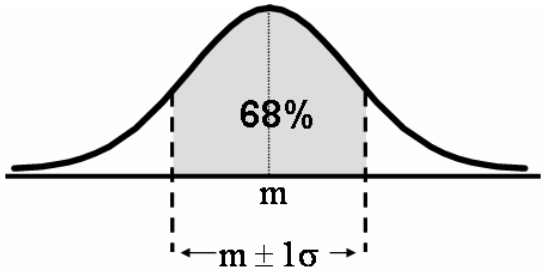
- L'**allure** d'une série statistique décrit la distribution des données dans cette série, en termes de **symétrie** / **asymétrie**.
- L'allure de la distribution dépend de la position relative de la médiane par rapport à la moyenne:



La règle empirique

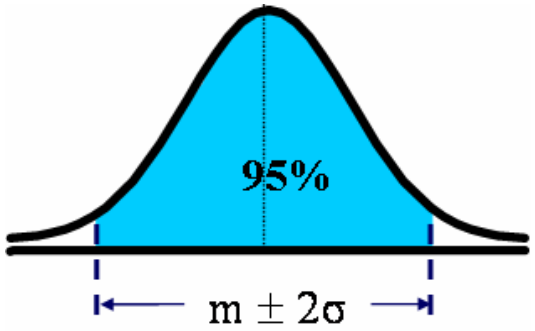
La **règle empirique** permet de se faire une idée de la dispersion des données pour les distributions normales (ou gaussiennes, "en forme de cloche").

« Environ 68% des données d'une distribution normale se situent à un écart-type de la moyenne, soit à $m \pm 1.\sigma$ »



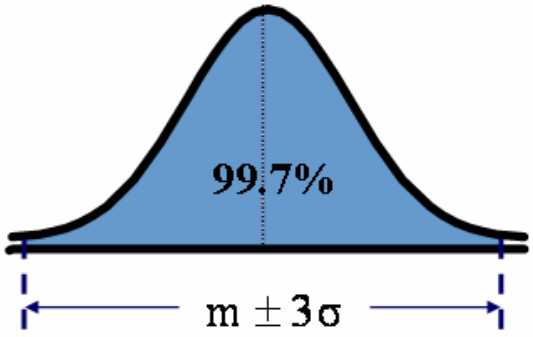
Règle empirique: formulation 2

« Environ 95% des données d'une distribution normale se situent à deux écart-types de la moyenne, soit à $m \pm 2.\sigma$ »



Règle empirique: formulation 3

« Environ 99.7% des données d'une distribution normale se situent à trois écart-types de la moyenne, soit $m \pm 3.\sigma$ »



Exemple d'utilisation de la règle empirique

Soit une variable statistique présentant une distribution normale ("en forme de cloche"), de moyenne 500 et d'écart-type 90. Alors:

- 68% des données se situent entre 410 et 590 (500 ± 90).
- 95% des données se situent entre 320 et 680 (500 ± 180).
- 99.7% des données se situent entre 230 et 770 (500 ± 270).

3.4. Mesures d'association: covariance et corrélation

Les indicateurs présentés plus haut ne concernent que la distribution d'un seul caractère.

Il est utile de disposer d'indicateurs permettant de préciser la nature des relations entre deux caractères.

Les deux indicateurs les plus courant pour ce faire sont la covariance et le coefficient de corrélation

La covariance

La **covariance** renseigne sur l'existence d'une relation linéaire (plus ou moins forte) entre deux variables quantitatives. Sa formule est:

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

La covariance n'est qu'un indicateur de la **force de la relation**: on ne peut en déduire **aucun lien de causalité** entre les variables étudiées!

Interprétation de la covariance

On peut considérer trois cas:

$\text{cov}(X,Y) > 0$: X et Y ont tendance à évoluer linéairement dans la même direction

$\text{cov}(X,Y) < 0$: X et Y ont tendance à évoluer linéairement dans des directions opposées

$\text{cov}(X,Y) = 0$: X et Y n'évoluent pas selon une relation linéaire (pas de liaison linéaire entre X et Y)

Indépendance et covariance

Si deux variables X et Y sont indépendantes alors leur covariance est nulle

Attention au sens de l'implication:
on a "indépendance $\Rightarrow$ $\text{cov}(X,Y) = 0$" mais pas l'inverse!

Il peut en effet exister une relation non linéaire entre X et Y, qui n'est pas mesurée par la covariance.

La covariance peut alors être nulle, mais X et Y ne sont pas indépendantes pour autant!

Somme de variables statistiques

La somme de deux variables statistiques X et Y est une variable statistique S.

La **moyenne** de cette variable S est simplement égale à la **somme des moyennes** de X et de Y

Le calcul de la **variance de la somme** de X et Y, en revanche, fait intervenir la **covariance** en plus des variances de X et Y:

$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$

$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y)$

Cas particulier: si X et Y sont indépendantes, alors $\text{cov}(X,Y) = 0$ et on observe que $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) = \text{Var}(X-Y)$.

Application numérique

X	Y	$S = X + Y$
46	46	92
40	49	89
48	44	92
45	46	91
43	39	82
46	44	90
48	39	87
44	46	90
45	43	88
46	49	95
52	52	104

On calcule:

$$\overline{X} = 45.73$$
$$\overline{Y} = 45.18$$
$$\overline{S} = \overline{X} + \overline{Y} = 90.91$$
$$Var\,X = 9.42$$
$$Var\,Y = 16.16$$
$$Cov(X,Y) = 2.15$$
$$Var\,S = Var\,X + var\,Y + 2cov(X,Y)$$
$$= 29.89$$

Le coefficient de corrélation

Le **coefficient de corrélation linéaire**, noté *r* ou ρ, est une mesure normalisée de la force relative de la relation linéaire existant entre deux variables quantitatives *X* et *Y*.

Sa formule s'appuie sur celle de la covariance:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{Cov(X,Y)}{S_X S_Y}$$

Tout comme la covariance, le coefficient de corrélation n'est qu'un indicateur de la force de la relation linéaire: il ne dit rien sur un éventuel lien de causalité entre les variables étudiées!

Propriétés et interprétation

Le coefficient de corrélation n'a pas d'unité de mesure

Par construction, il prend ses valeurs entre –1 et 1

Plus il est proche de –1, plus X et Y seront liées par une relation linéaire fortement *négative*

Plus il est proche de 1, X et Y seront liées par une relation linéaire fortement *positive*

Si le coefficient de corrélation est proche de 0, il n'y a aucune relation linéaire entre X et Y.

Dans l'application numérique précédente, on trouve $r = 0.17$

4. De la description à l'inférence statistique

Les indicateurs et techniques abordés ci-dessus permettent avant tout de **décrire un échantillon** de données. Il ne permettent de décrire une population **qu'à partir du moment** où toutes les données de la population sont observées (!).

Dès que l'on veut tirer des conclusions sur une population à partir de données d'échantillons, on entre dans le domaine de la **statistique inférentielle**

Celle-ci fournit des outils permettant de déterminer dans quelle mesure les valeurs des **statistiques d'échantillons** reflètent celles des **paramètres de la population**

Paramètres de la population et statistiques d'échantillon

Indicateur	Population	Echantillon
Moyenne	m	$\overline{X}$
Variance	σ^2	S^2
Ecart-type	σ	S

Inconnus Faciles à calculer

4.1. Variables aléatoires et probabilités

En statistique inférentielle, le hasard joue un rôle central car les échantillons sont "tirés au sort" dans la population.

Les valeurs des variables retenues dans l'échantillon seront donc considérées comme résultant du hasard, d'un **tirage aléatoire** parmi un grand nombre de valeurs possibles dans la population.

On ne parlera donc plus de variable statistique, mais de **variable aléatoire (VA)**. La distribution d'une VA dans la population est caractérisée par une **loi de probabilité**, et les valeurs que prend la VA dans l'échantillon sont appelées ses **réalisations**.

L'outil probabiliste: intuition

On a présenté plus haut le concept de **fréquence** d'une VS. La **probabilité** peut être comprise comme une autre manière d'aborder ce concept. Par exemple, on peut dire:

- Si, parmi 100 individus (corbeaux), 90 présentent le trait A (plumage noir) et 10 présentent le trait B (plumage gris) alors la **fréquence** du trait B (plumage gris) est 0.10 soit 10%

- Un individu (corbeau) tiré au sort parmi ces 100 a 10 chances sur 100 (1 chance sur 10) de présenter le trait B (plumage gris). La **probabilité** qu'il présente ce trait est 0.10 (10%).

Fréquences et probabilités

Dans l'exemple précédent:

100 individus	Vocabulaire "descriptif"	Vocabulaire "probabiliste"
Couleur du plumage	Variable statistique X	Variable aléatoire X
"plumage gris", "plumage noir"	Modalités de la VS	Valeurs de la VA (X = 1 si gris, 0 si noir)
0.10	Fréquence de la modalité "plumage gris"	Probabilité d'observer X =1 (plumage gris), notée P(X=1)
0.90	Fréquence de la modalité "plumage noir"	Probabilité d'observer X =0 (plumage noir), notée P(X=0)

Approche "fréquentiste" des probabilités

Cette approche des probabilités est qualifiée de "fréquentiste".

Au lieu de parler des fréquences de la modalité d'une VS, on parlera de la probabilité d'observer un évènement (valeur(s) d'une VA). Par exemple: "probabilité d'observer un corbeau gris".

Tout comme une fréquence, une probabilité prend ses valeurs entre 0 et 1 (ou 0 et 100 si on l'exprime en %)

Tout comme une VS, une VA est caractérisée par une distribution, appelée **distribution de probabilité** ou **loi de probabilité.**

Définition formelle

Dans l'approche fréquentiste, la probabilité est formellement définie comme la limite de la fréquence quand la taille de la population (ou d'un hypothétique échantillon) devient infinie:

$$\Pr = \lim f_n \text{ quand } n \to \infty$$

Exemple: si on jette un dé à 6 faces non truqué un nombre infini de fois, la fréquence doit se stabiliser autour d'une valeur limite (ici, 1/6), qui est justement la *probabilité* d'obtenir une chiffre donné entre 1 et 6.

Vocabulaire des probabilités (1)

Soit X une VA et S l'ensemble de ses valeurs possibles. Comme X est le résultat d'un processus aléatoire (ou d'une "expérience"), on appellera S "ensemble des valeurs possible".

Si X peut prendre k valeurs e_1, …, e_k, on appelle **évènement** tout sous-ensemble de S: $\{e_1\}$ est un évènement, $\{e_1, e_2, e_3\}$ aussi.

La **probabilité d'un évènement** est la somme des probabilités de tous les résultats (ou "points" ou "évènements élémentaires") inclus dans cet évènement. Par exemple, si $E = \{e_1, e_2, e_3\}$, alors $\Pr(E) = \Pr(e_1)+\Pr(e_2)+\Pr(e_3)$

Vocabulaire des probabilités (2)

Soit G et H deux évènements. $\mathbf{G \cup H}$ ("G ou H") est l'ensemble des points appartenant à G, à H ou aux deux. $\mathbf{G \cap H}$ ("G et H") est l'ensemble des points appartenant à la fois à G et à H. On a:

$$\Pr(G \cup H) = \Pr(G) + \Pr(H) - \Pr(G \cap H)$$

Si G et H sont **mutuellement incompatibles**, alors

$$\Pr(G \cup H) = \Pr(G) + \Pr(H)$$

Soit E un évènement. On note E^C **l'évènement complémentaire** de E (ensemble des points ne figurant pas dans E). On a:

$$\Pr(E) = 1 - \Pr(E^C)$$

Probabilités conditionnelles

Soit G et H deux évènements. On note Pr(G | H) la probabilité de G *conditionnellement* à la réalisation de H, et Pr(H | G) la probabilité de H *conditionnellement* à la réalisation de G.

Ces **probabilités conditionnelles** s'écrivent:
$\Pr(G \mid H) = \Pr(G \cap H)/\Pr(H)$
$\Pr(H \mid G) = \Pr(H \cap G)/\Pr(G)$

N.B.: Si $\Pr(G \mid H) = \Pr(G)$ alors G et H sont **statistiquement indépendants**. On en déduit que si G et H sont indépendants, alors $\Pr(G \cap H) = \Pr(G).\Pr(H)$, ce qui découle directement de la formule des probabilités conditionnelles

Approche bayésienne

L'approche bayesienne s'appuie sur la **formule de Bayes**, selon laquelle, pour deux évènements G et H, on a:

$$\Pr(G \mid H) = \Pr(G) \times \Pr(H \mid G) / \Pr(H)$$

Origine de la formule:

$\Pr(G \mid H) = \Pr(G \cap H)/\Pr(H) \Leftrightarrow \Pr(G \cap H) = \Pr(G \mid H) \times \Pr(H)$
$\Pr(H \mid G) = \Pr(H \cap G)/\Pr(G) \Leftrightarrow \Pr(H \cap G) = \Pr(H \mid G) \times \Pr(G)$

Comme $\Pr(G \cap H) = \Pr(H \cap G)$, on a :

$\Pr(G \mid H) \times \Pr(H) = \Pr(H \mid G) \times \Pr(G)$
$\Leftrightarrow \Pr(G \mid H) = \Pr(H \mid G) \times \Pr(G) / \Pr(H) = \Pr(G) \times \Pr(H \mid G) / \Pr(H)$

Cette approche a été traitée en cours au moyen d'un exemple.

Autres approches

- Probabilités symétriques: appliquées aux jeux de hasard, comme les dés, où tous les résultats sont équiprobables

- Approche axiomatique: construction mathématique rigoureuse

- Probabilités subjectives: individuelles, "révélées" par exemple par les préférences des individus face à des choix associés à des probabilités objectives.

- Cotes (ou rapports de cotes): $d = p / (1 - p)$ où p est la probabilité d'un évènement dont la réalisation peut se traduire par un gain (ou une perte) comme dans le cas d'un pari.

4.2. Lois de probabilités

Les lois de probabilités sont des outils mathématiques élaborés par les statisticiens pour représenter des phénomènes aléatoires.

Elles sont caractérisées à l'aide de paramètres (espérance, équivalent probabiliste de la moyenne, et variance) et de fonctions mathématiques (fonction de répartition et/ou densité)

Les VA peuvent être discrètes (valeurs entières, de deux à une infinité) ou continues (valeurs réelles).

Cela conduit à distinguer deux grandes familles de lois de probabilités: les lois discrètes et les lois continues

Espérance et variance

L'espérance indique la tendance centrale d'une distribution de probabilité, et la variance indique sa dispersion.

Ces indicateurs ont les propriétés déjà rencontrées pour la moyenne et la variance. En particulier, pour deux VA X et Y, et pour deux constantes a et b, on a:

$$E(aX+bY) = aE(X)+bE(Y)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X, Y)$$
$$\mathrm{Var}(X-Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)-2\mathrm{Cov}(X, Y)$$

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

La fonction de répartition

La fonction de répartition décrit la manière dont se répartissent les valeurs d'une VA.

Soit X une VA. On note $X(\Omega)$ l'ensemble de ses valeurs possibles. Pour toute valeur x appartenant à $X(\Omega)$, la fonction de répartition donne la probabilité d'observer une valeur de X inférieure à x

Si x n'appartient pas à $X(\Omega)$, la fonction de répartition sera égale soit à 0, soit à 1 (voir plus loin).

L'écriture formelle de la fonction de répartition est différente pour les lois (VA) discrètes et continues.

Cas discret et cas continu

Soit X une VA de loi discrète (ou VA discrète). On note $X(\Omega)$ l'ensemble des valeurs de X.
La fonction de répartition de X s'écrit, $\forall\ x \in X(\Omega)$:

$$\Pr(X \leq x) = \sum_{k \leq x} \Pr(X = k)$$

Soit X une VA de loi continue (ou VA continue) telle que $X(\Omega) = \mathbb{R}$.
La fonction de répartition de X s'écrit, $\forall\ x \in \mathbb{R}$:

$$\Pr(X \leq x) = F(x)$$

A chaque loi continue correspond une forme spécifique de la fonction F (voir plus loin)

Propriétés de $F(x)$

Soit X une VA continue telle que $X(\Omega) = \mathbb{R}$, et soit F(x) sa fonction de répartition. On peut noter les propriétés suivantes:

1. $\forall\ x \in \mathbb{R}$, $F(x) \in [0, 1]$
2. F est une fonction croissante sur $\mathbb{R}$
3. Pour tous réels a et b, $\Pr(a < X \leq b) = F(b) - F(a)$
4. $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to +\infty} F(x) = 1$

Densité de probabilité

Soit X une VA dont l'ensemble des valeurs est $X(\Omega)$. Si X est discrète, alors la probabilité que X prenne la valeur k, $k \in X(\Omega)$, peut s'écrire ***Pr***$(X = k) = p$.

Si X est continue ($X(\Omega) = \mathbb{R}$), on peut la doter d'une fonction f appelée **densité de probabilité**, définie comme la dérivée de la fonction de répartition. Autrement dit, $\forall\ x \in \mathbb{R}$, f est donnée par:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

L'une ou l'autre de ces fonctions suffit à complètement caractériser une distribution de probabilité.