



Modélisation – Partie I : Statistique

Semaine Formation	Semaine Civile	Du lundi	Au samedi	Séances	Contenu	Logiciel principal
1	36	02/09/2019	07/09/2019	2 séances de TP (dont 1 en autonomie)	<ul style="list-style-type: none">• Rappels de stat• Premières manipulations sous Excel• Choix d'un jeu de donné• Préparation des données• Début de l'analyse univariée	
2	37	09/09/2019	14/09/2019	1 séance TD-TP	<ul style="list-style-type: none">• Manipulations Excel• Analyse univariée	
3	38	16/09/2019	21/09/2019	1 séance de TP en autonomie	<ul style="list-style-type: none">• Analyse univariée• Réflexion sur l'analyse bivariée	
4	39	23/09/2019	28/09/2019	2 séances de TP (dont 1 en autonomie)	<ul style="list-style-type: none">• Analyse bivariée	
5	40	30/09/2019	05/10/2019	1 séance de TP en autonomie	<ul style="list-style-type: none">• Finalisation• Rendu Travail Excel (noté sur 7 pts)• Importation des données sous R• Oraux éventuels	
6	41	07/10/2019	12/10/2019			
7	42	14/10/2019	19/10/2019	1 séance TD-TP + 1TP en autonomie	<ul style="list-style-type: none">• Finalisation de l'importation• Analyse univariée sous R	
8	43	21/10/2019	26/10/2019	1 séance TD-TP	<ul style="list-style-type: none">• Analyse bivariée sous R	
	44	28/10/2019	02/11/2019	Vacances Toussaint		
9	45	04/11/2019	09/11/2019	2 séances de TP (dont 1 en autonomie)	<ul style="list-style-type: none">• Finalisation - Approfondissement	
10	46	11/11/2019	16/11/2019	1 séance TD-TP	<ul style="list-style-type: none">• Rendu Travail R (noté sur 7 pts)• Oraux éventuels• QCM d'évaluation (noté sur 6 points)	

RAPPELS DE STATISTIQUE DESCRIPTIVE

La **statistique descriptive** est un ensemble de méthodes et de techniques permettant de présenter, de décrire, de résumer un ensemble de données observées sur une **population**.

Les **individus** (ou unités statistiques) sont les éléments de la population étudiée. Pour chaque individu, on dispose d'une ou de plusieurs observations.

Une **variable statistique** (ou caractère statistique) est ce qui est observé ou mesuré sur les individus de la population.

A l'origine : science qui a pour but de faire connaître l'étendue, la population, les ressources agricoles et industrielles d'un État (Littré).

RAPPELS DE STATISTIQUE DESCRIPTIVE

La **statistique descriptive** est un ensemble de méthodes et de techniques permettant de présenter, de décrire, de résumer un ensemble de données observées sur une **population**.

Les **individus** (ou unités statistiques) sont les éléments de la population étudiée. Pour chaque individu, on dispose d'une ou de plusieurs observations.

Une **variable statistique** (ou caractère statistique) est ce qui est observé ou mesuré sur les individus de la population.

Le tableau individu variable (fichier [CAF-cours1.xlsx](#))

num_alloc	COMDO	RMI	dtdeMRI	nbenf	fam	agealloc
5872	64372	0		2	Isolé	29
6149	64053	0		2	Isolé	51
6108	64238	0		2	Couple	39
5939	64343	0		2	Isolé	36
5558	64519	0		2	Isolé	50
1195	64399	0		2	Isolé	35
4629	64243	0		2	Isolé	35
694	64511	0		1	Isolé	42
1279	64347	1	30/01/1998	1	Couple	45
2763	64370	1	10/11/2000	0	Couple	31
5657	64482	0		0	Couple	42
5235	64405	0		0	Isolé	22
5190	64192	0		2	Isolé	35
1570	64516	0		3	Isolé	37
4910	64526	0		3	Isolé	43
4311	64511	0		2	Isolé	50
159	64388	0		2	Isolé	30
4089	64374	0		0	Couple	38
1578	64405	0		2	Isolé	33
3253	64238	1	05/05/1998	1	Couple	51
4951	64415	0		1	Isolé	30

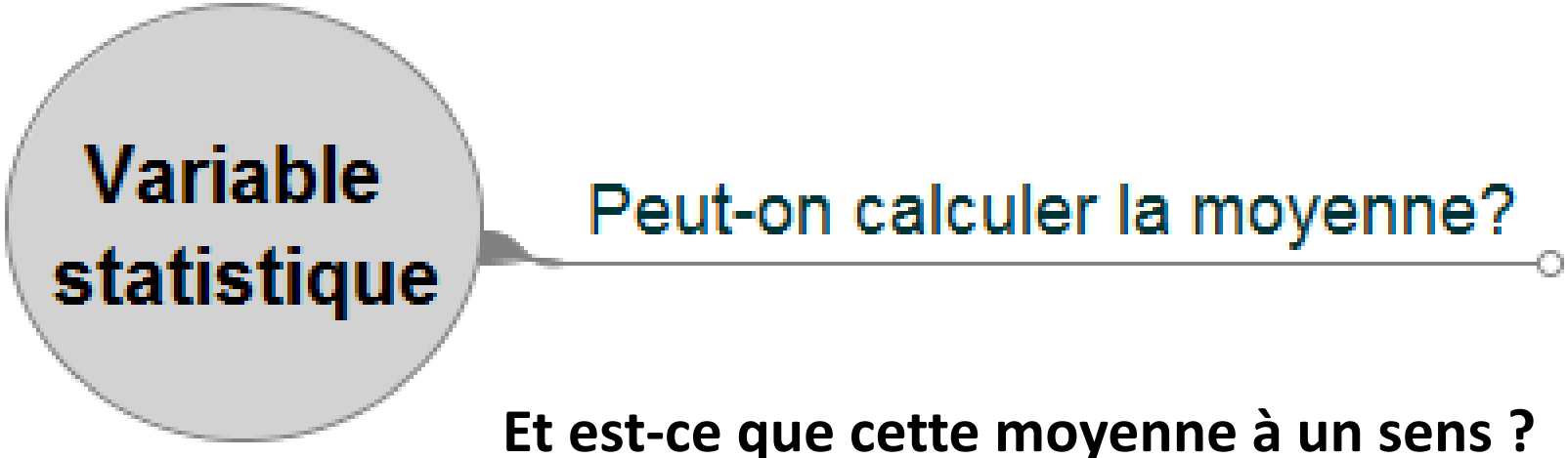
- Tableau individus × variables
- **Individus** : allocataires des CAF Béarn-Soul et Landes. (1 allocataire = 1 famille)
- **Population** : ensemble des allocataires de six communautés de communes des CAF Béarn-Soule et Landes en 2006 (**6328 individus**).

Les différents types de variables statistiques

Qualitatives nominales et ordinales

Quantitatives discrètes et continues

DIFFERENTS TYPES DE VARIABLES STATISTIQUES



**Variable
statistique**

Peut-on calculer la moyenne?

Et est-ce que cette moyenne à un sens ?

DIFFERENTS TYPES DE VARIABLES STATISTIQUES

**Variable
statistique**

Peut-on calculer la moyenne?

Oui : variables quantitatives

Non : variables qualitatives

Variables quantitatives

Discrètes

Continues

VARIABLES QUANTITATIVE

Grand nombre de valeurs ?

Oui : **VARIABLES QUANTITATIVES
CONTINUES**

Non : **VARIABLES QUANTITATIVES
DISCRETES**

VARIABLES QUANTITATIVES

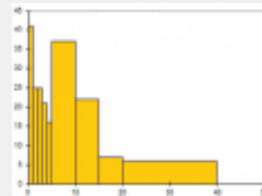
Oui : **VARIABLES QUANTITATIVES
CONTINUES**

Exemples de variables continues :

- taille, poids d'un individu dans une population,
- mesure d'une pièce dans une production,
- salaire, revenu,...

Calcul d'indicateurs :

- moyenne,
- médiane,
- écart-type,...



GRAPHIQUE : histogramme

VARIABLES QUANTITATIVES

Non : **VARIABLES QUANTITATIVES
DISCRETES**

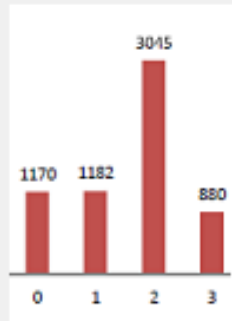
Exemples de variables discrètes :

- nombre d'enfants,
- âge dans une population de lycéens,
- nombre d'année d'études supérieures,...

Calcul d'indicateurs :

- moyenne,
- médiane,
- écart-type,...

Tableau de fréquences : effectifs, pourcentages,...



GRAPHIQUE : diagramme en barres

Variables qualitatives

Ordinales

Nominales

VARIABLES QUALITATIVES

Modalités ordonnées ?

Oui : **VARIABLES QUALITATIVES
ORDINALES**

Non : **VARIABLES QUALITATIVES
NOMINALES**

VARIABLES QUALITATIVES

Oui : **VARIABLES QUALITATIVES
ORDINALES**

Exemples de variables qualitatives ordinales :

- grade,
- jugement de valeur (Satisfait, Très satisfait, ...)
- mise en classe d'une variable quantitative (Petit, Moyen, Grand)
- date, mois, jour de la semaine,...

Tableau de fréquences : effectifs, pourcentages,...



GRAPHIQUE : diagramme en barre

VARIABLES QUALITATIVES

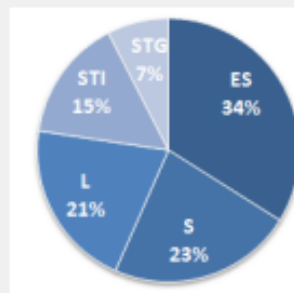
Exemples de variables nominales:

- CSP, statut marital,
- Région d'origine,
- Diplôme,
- Sexe,...

Tableau de fréquences : effectifs, pourcentages,...



GRAPHIQUE : diagramme en barres
(modalités rangées par effectifs décroissants)



GRAPHIQUE : diagramme en secteurs
(modalités rangées par effectifs décroissants)

Non : **VARIABLES QUALITATIVES
NOMINALES**

VARIABLES QUALITATIVES

Exemples de variables nominales:

- CSP, statut marital,
- Région d'origine,
- Diplôme,
- Sexe,...

Variables à deux modalités : homme/femme, Oui/Non,...

Elles sont souvent codées 0/1 et il est alors possible de calculer la **moyenne**.

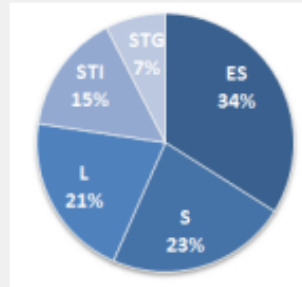
Mais celle-ci ne s'interprète pas comme telle mais comme une **proportion**.

Raison pour laquelle, ces variables sont considérées, malgré tout comme qualitatives.

Tableau de fréquences : effectifs, pourcentages,...



GRAPHIQUE : diagramme en barres
(modalités rangées par effectifs décroissants)

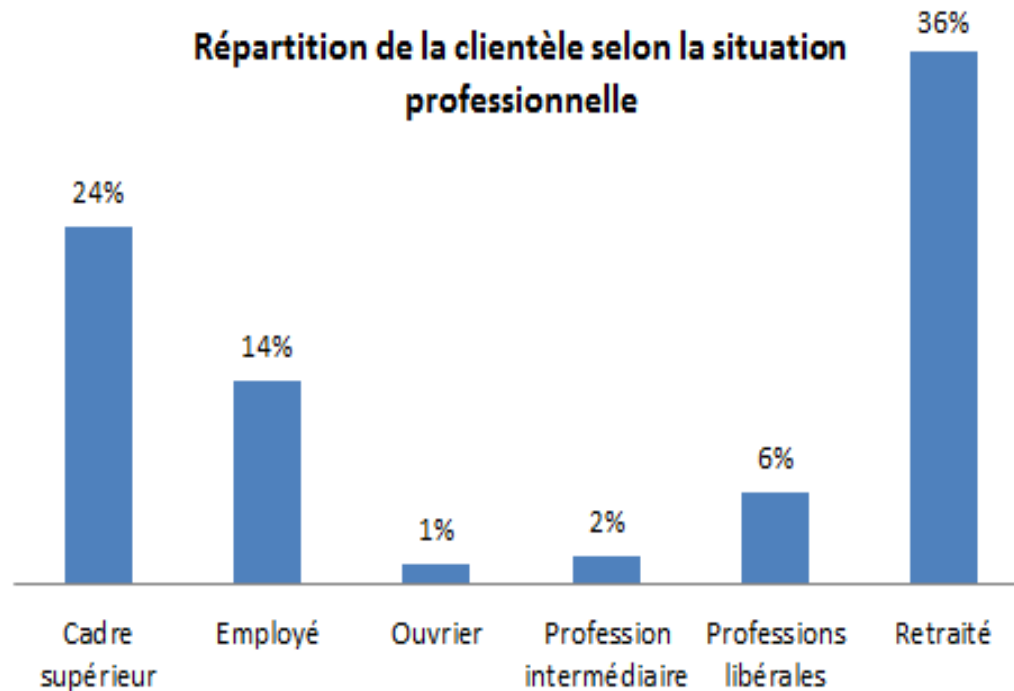


GRAPHIQUE : diagramme en secteurs
(modalités rangées par effectifs décroissants)

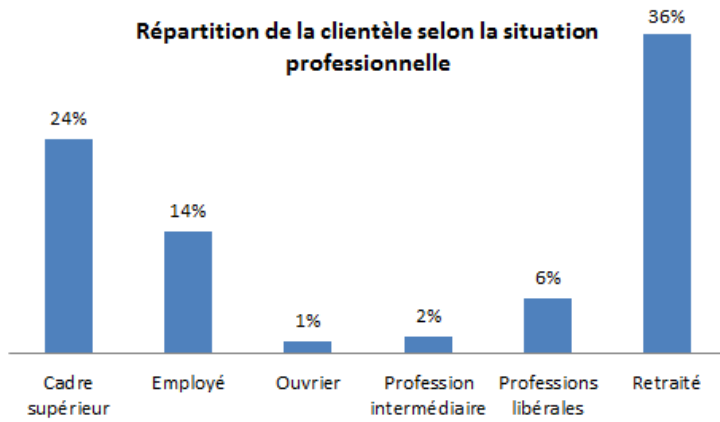
Les graphiques

Quelques écueils à éviter

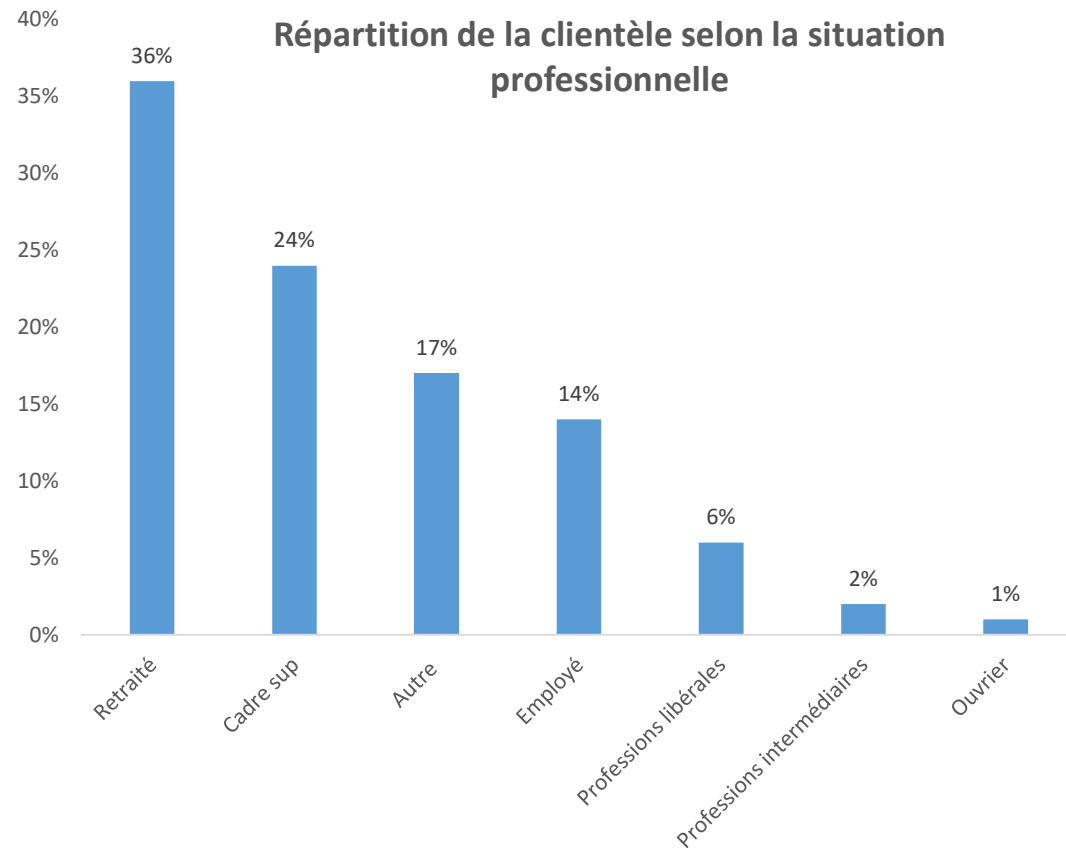
Graphique correct ?



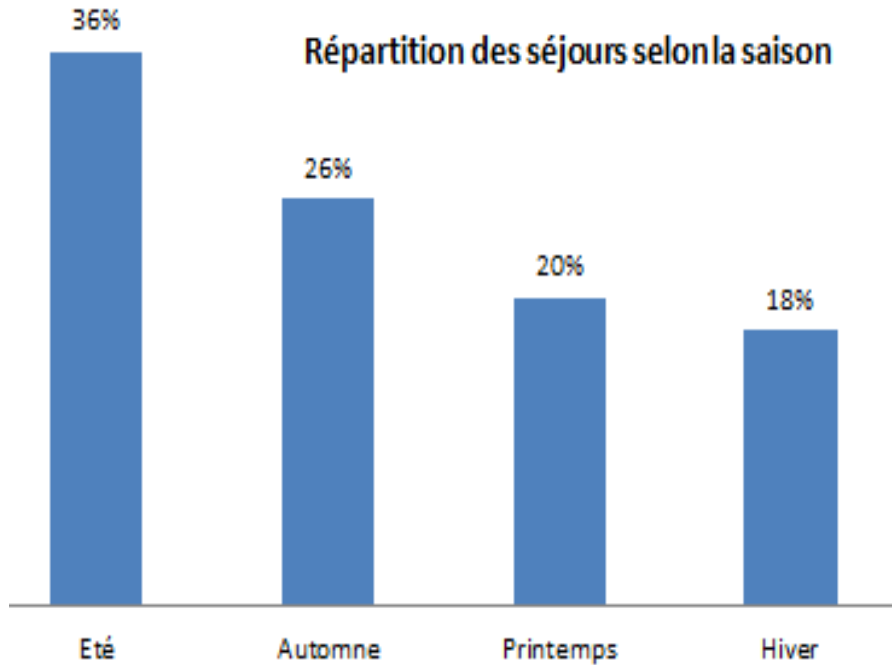
Répartition de la clientèle selon la situation professionnelle

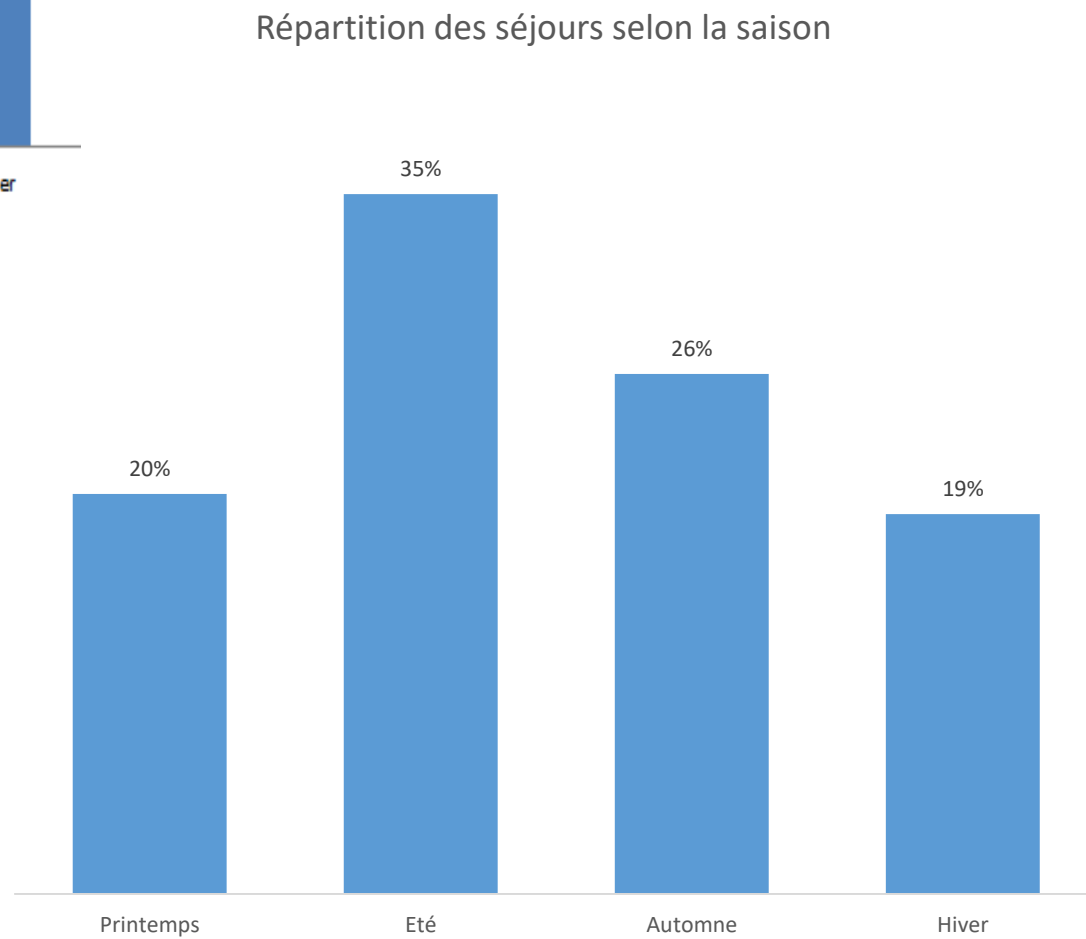
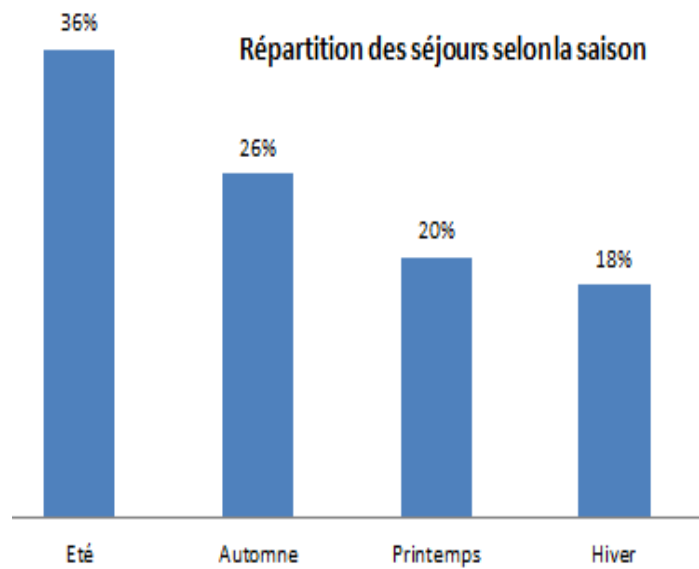


Répartition de la clientèle selon la situation professionnelle



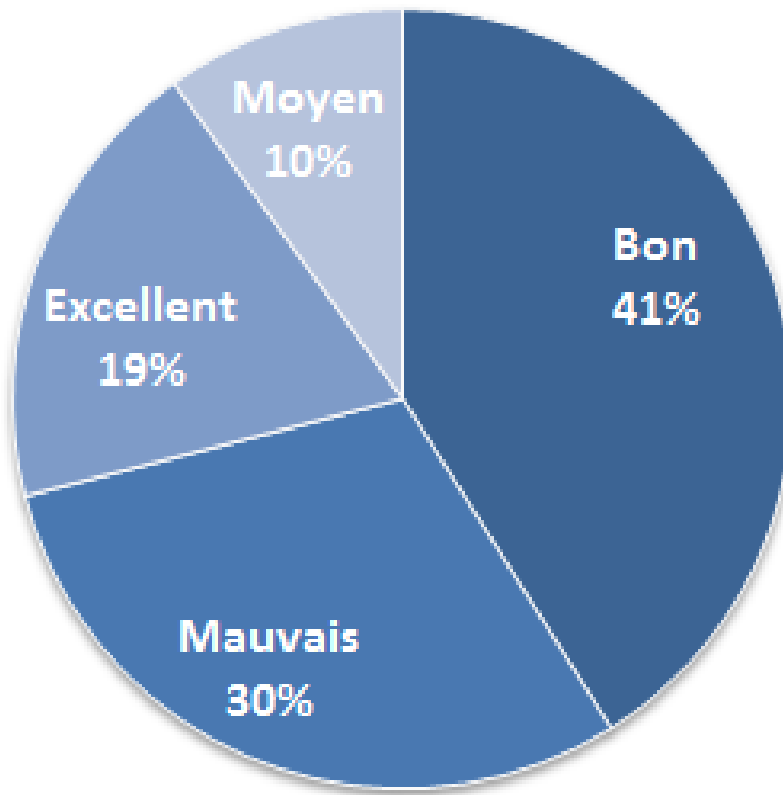
Graphique correct ?



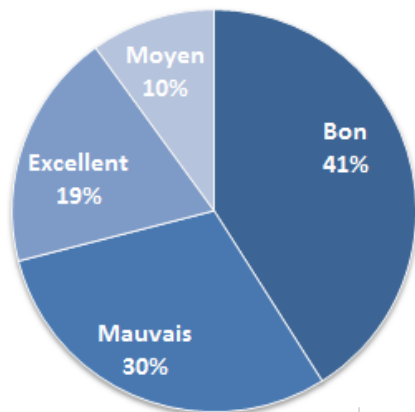


Graphique correct ?

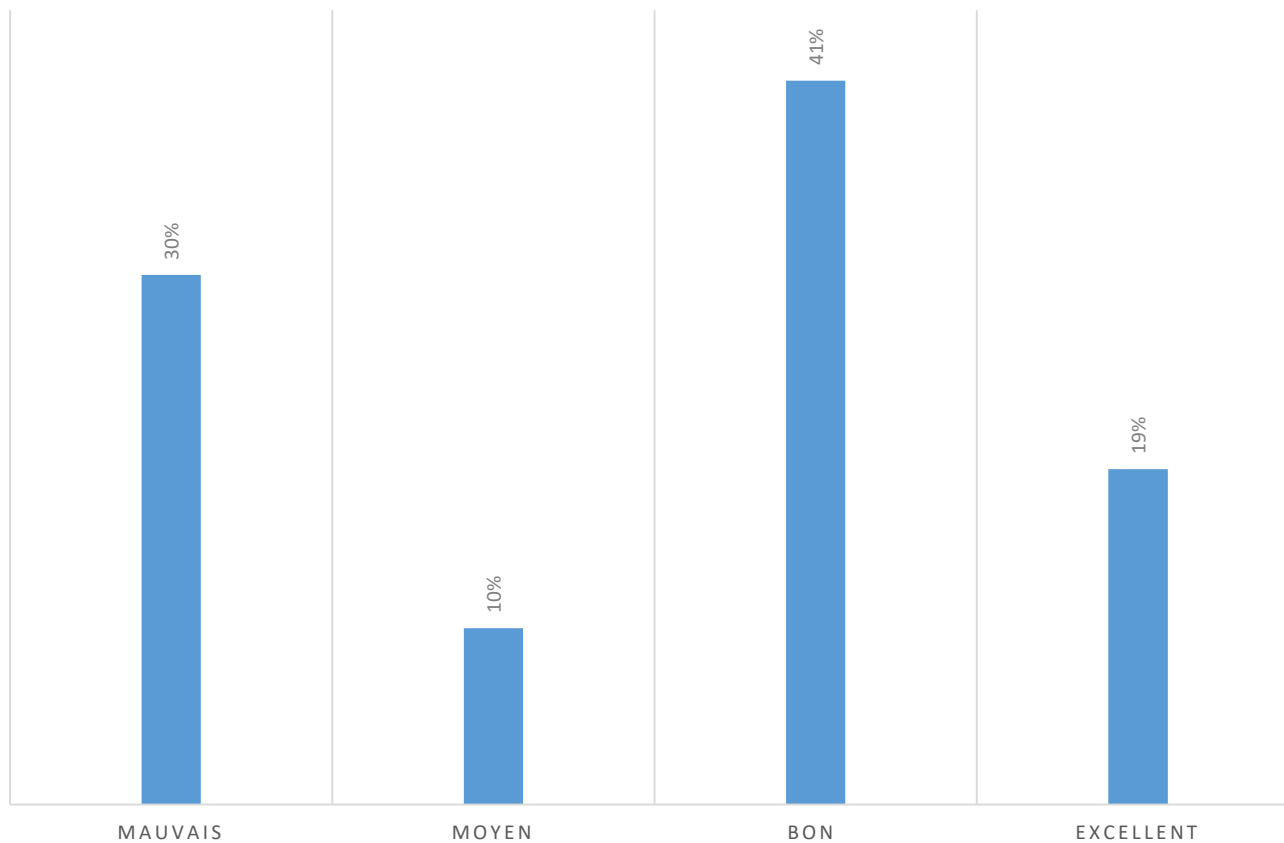
Qualité de l'encadrement



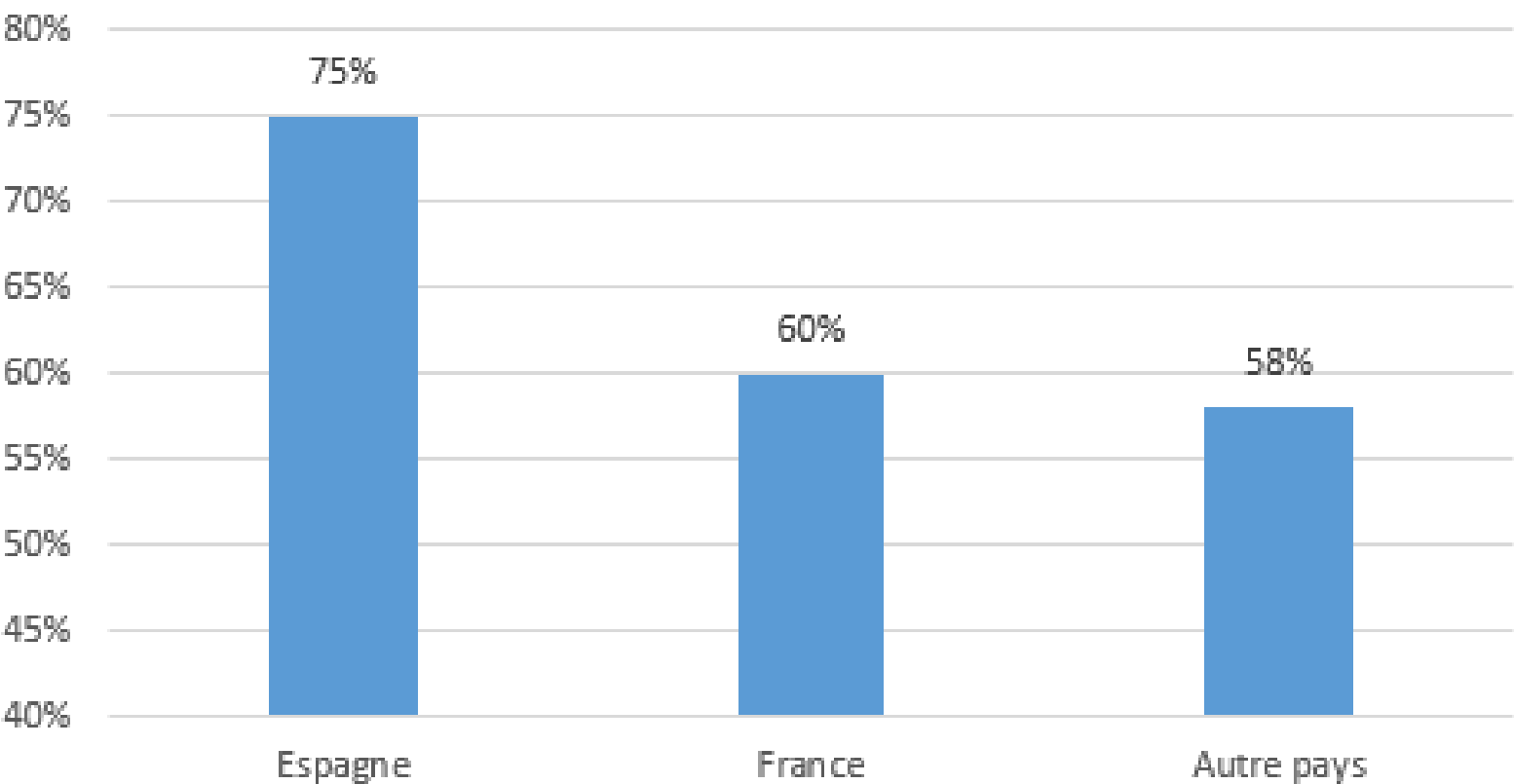
Qualité de l'encadrement



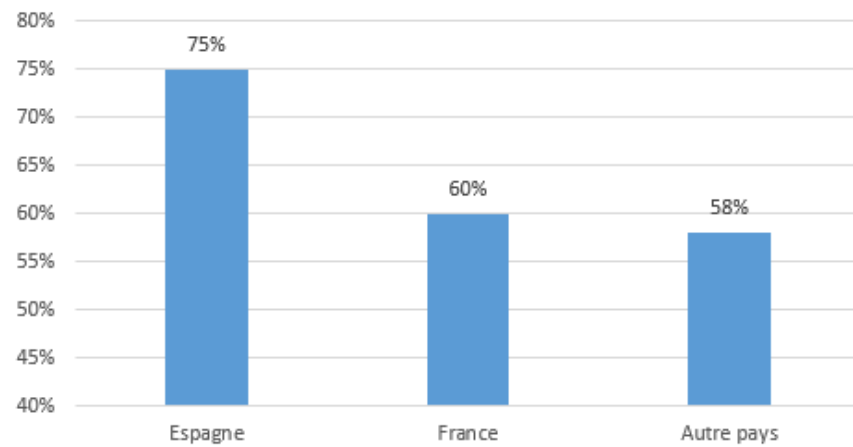
RÉPARTITION DES CLIENTS SELON LEUR APPRÉCIATION DE LA QUALITÉ DE L'ENCADREMENT



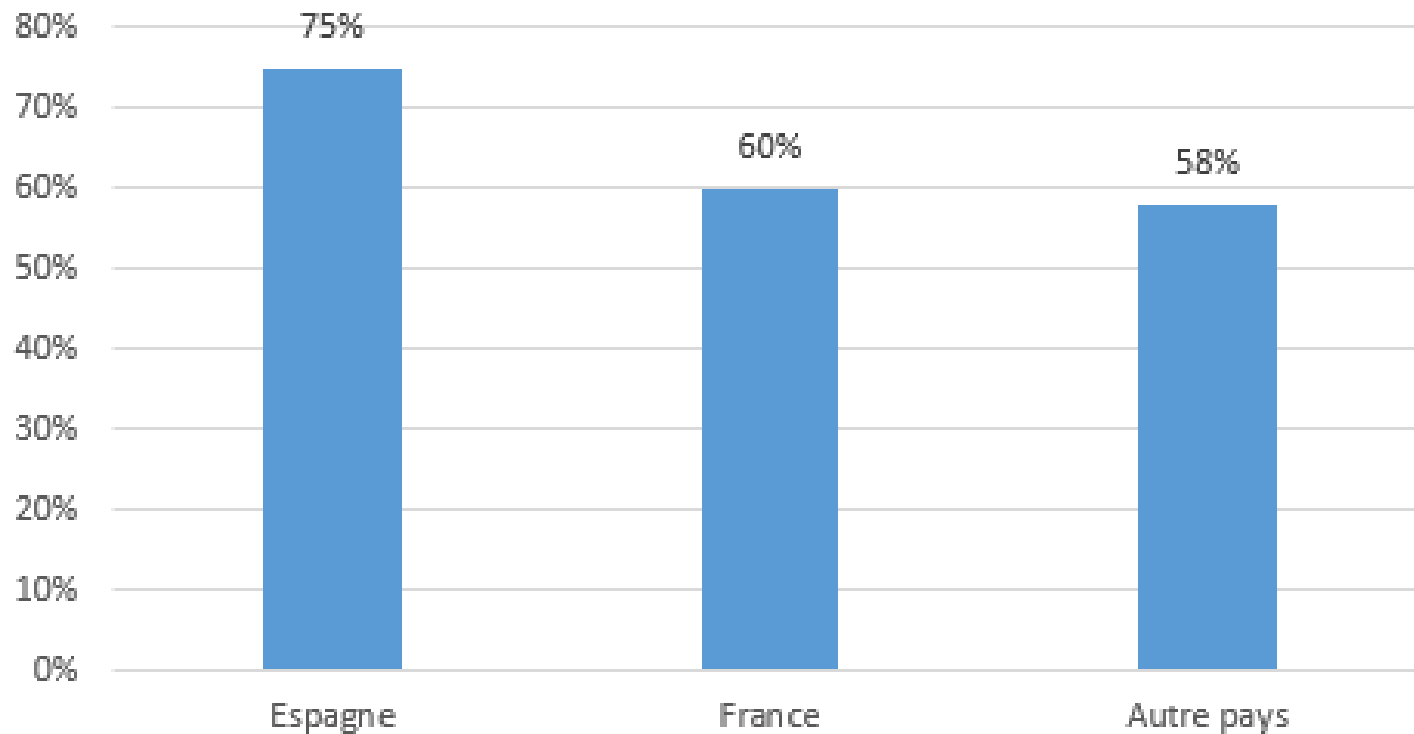
Proportion de groupes avec un GPS



Proportion de groupes avec un GPS

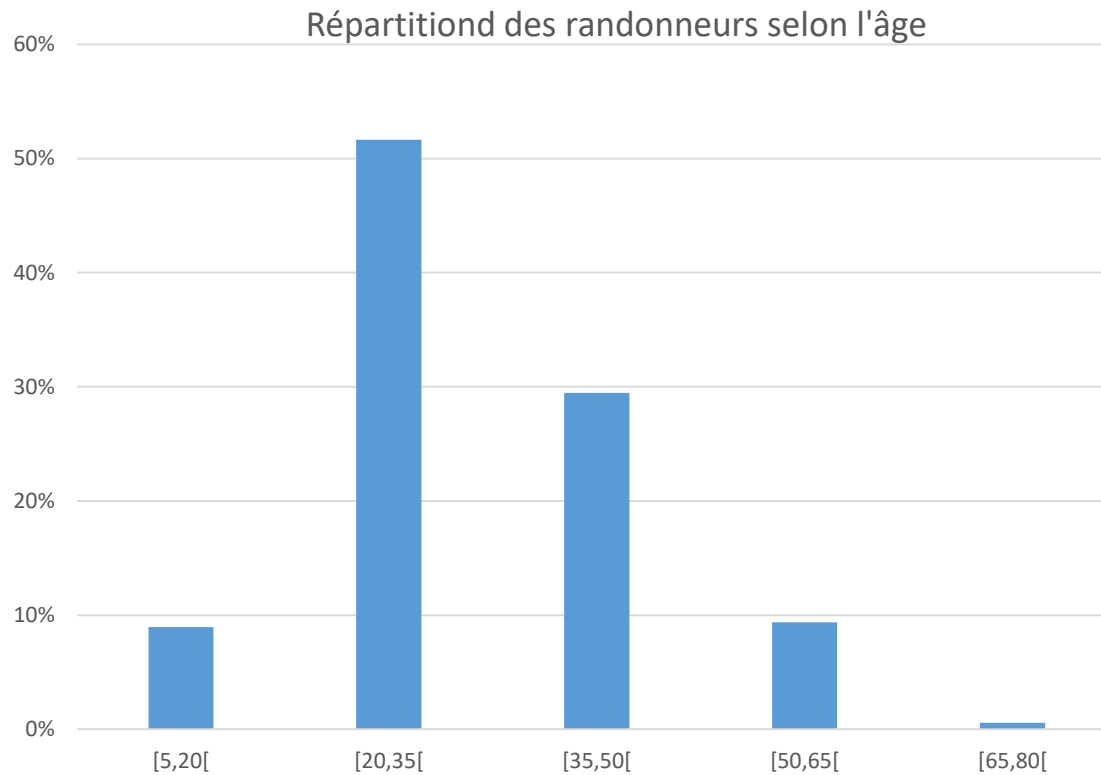


Proportion de groupes avec un GPS



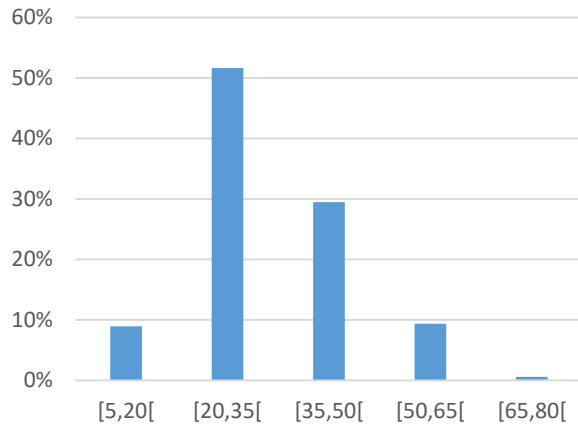
Variable quantitative
continue

Variable quantitative continue

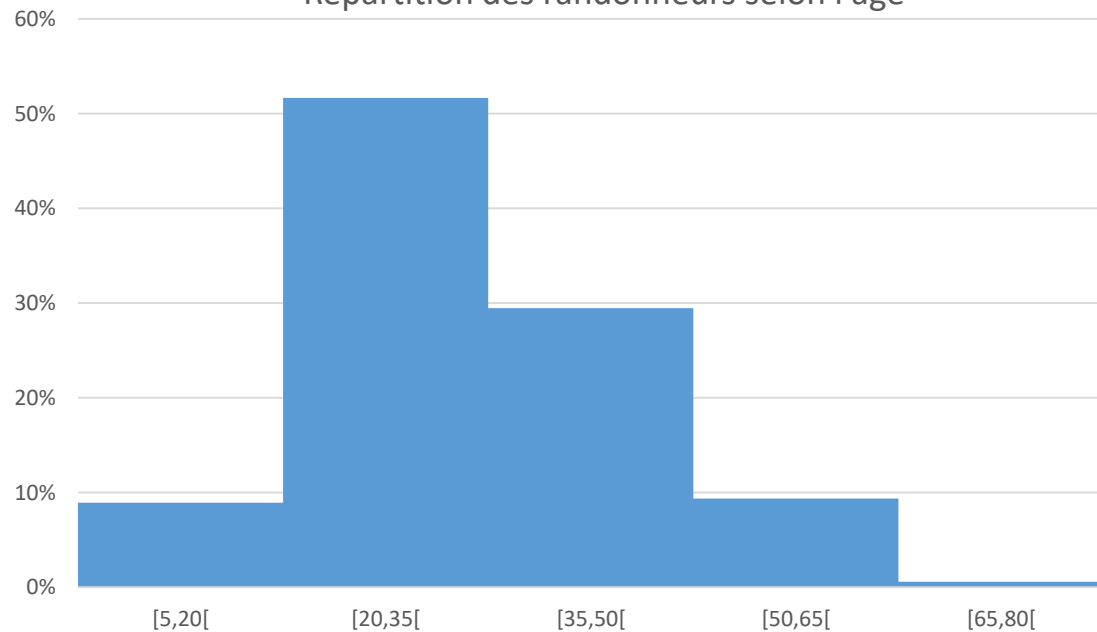


Variable quantitative continue

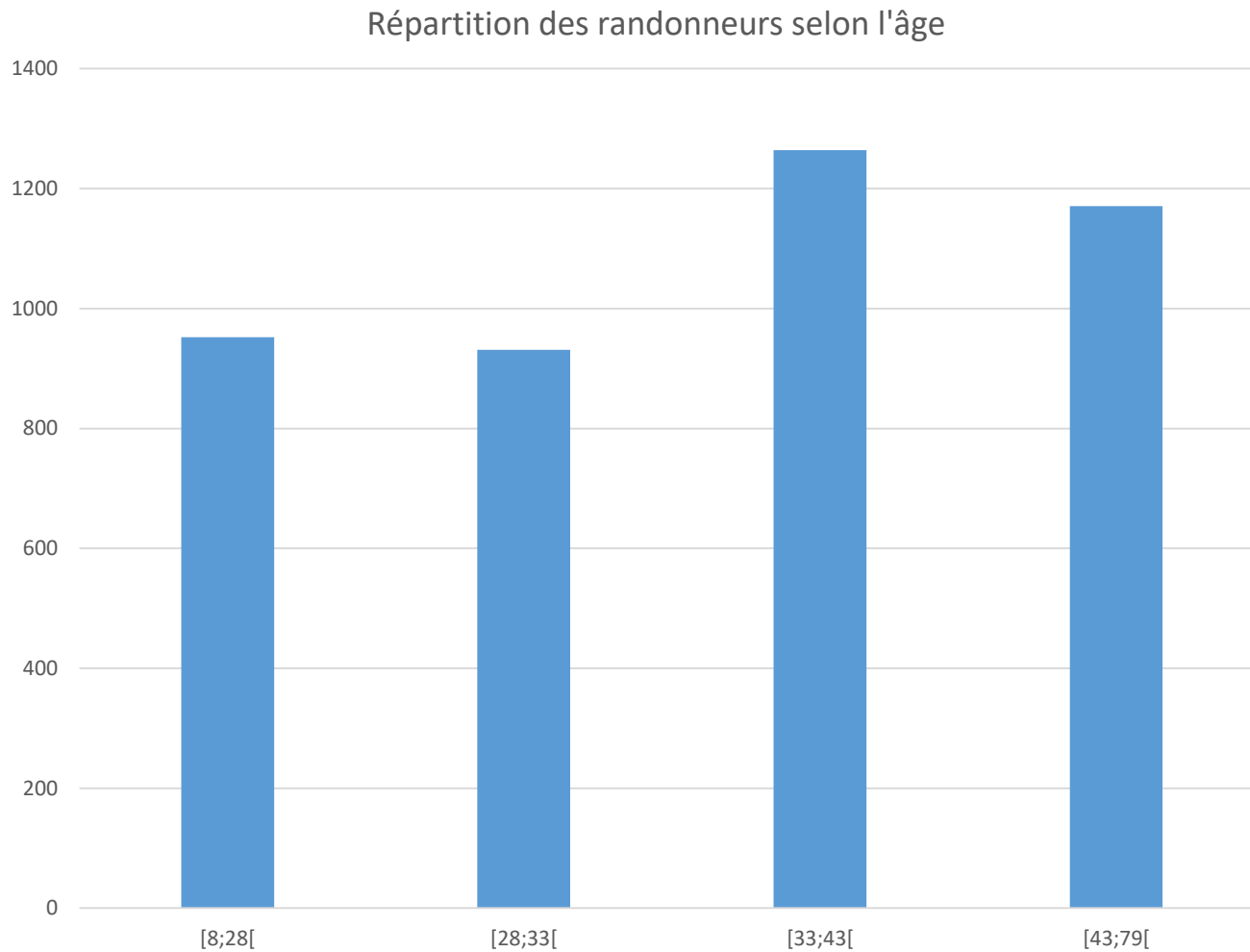
Répartition des randonneurs
selon l'âge



Répartition des randonneurs selon l'âge

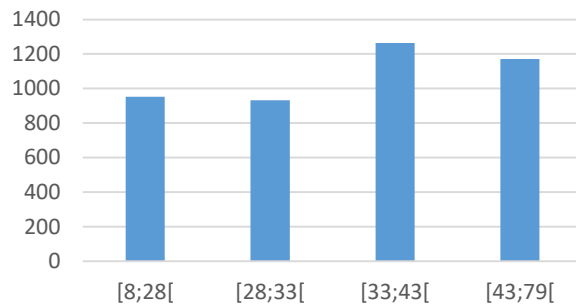


Autres classes

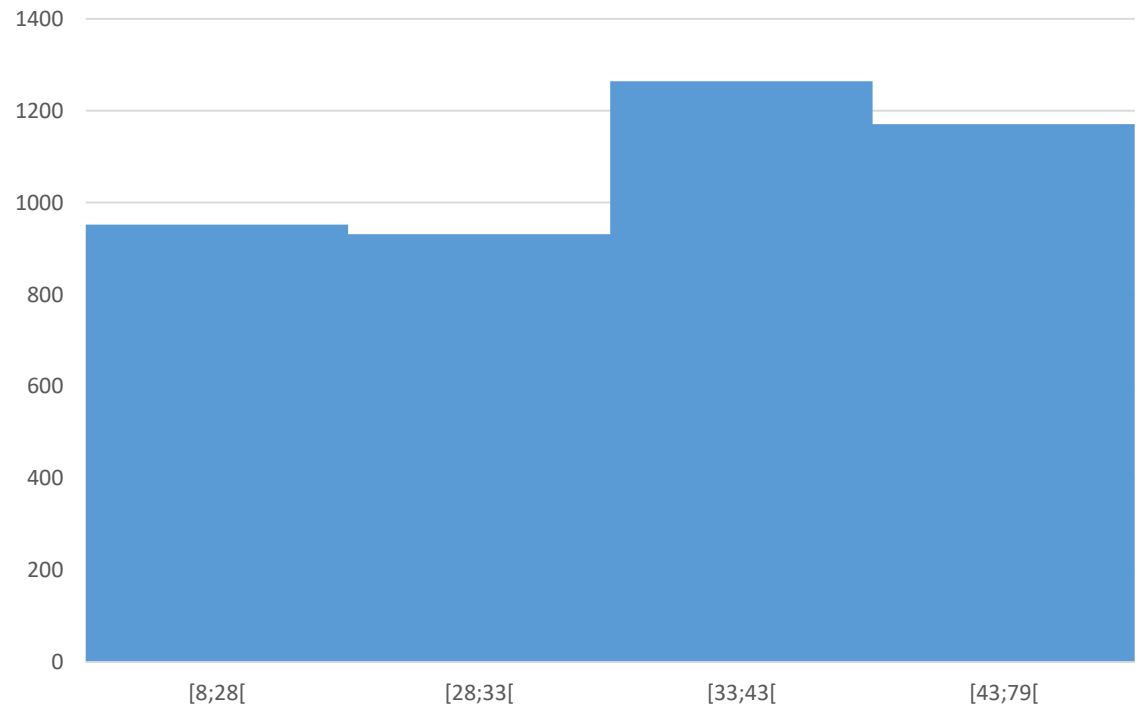


Autres classes

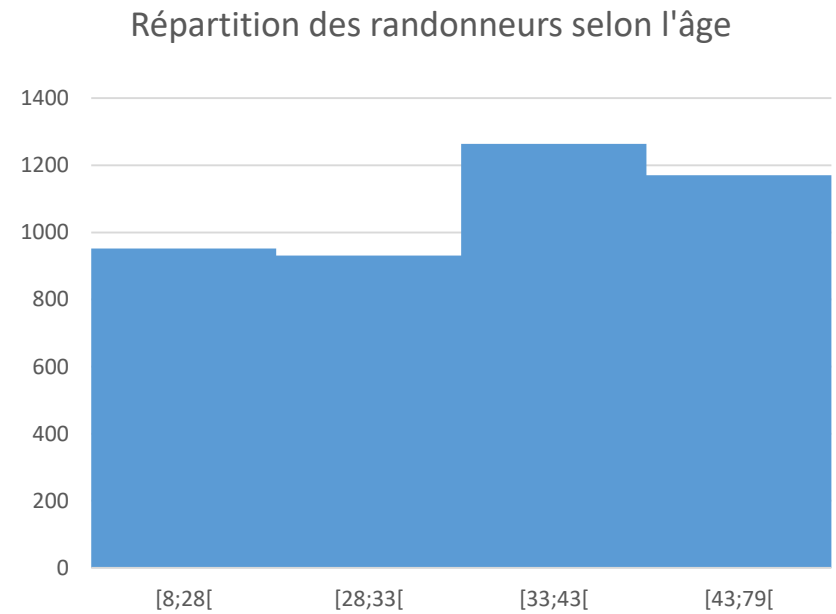
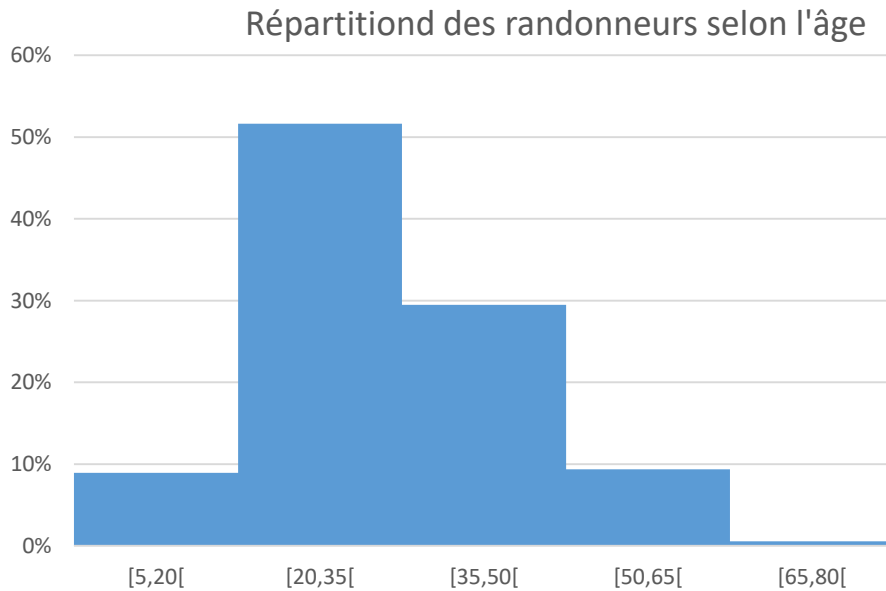
Répartition des randonneurs
selon l'âge



Répartition des randonneurs selon l'âge

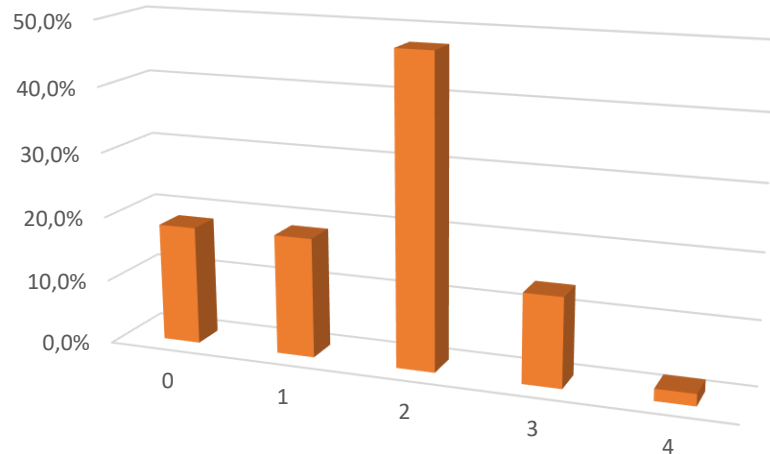


Autres classes

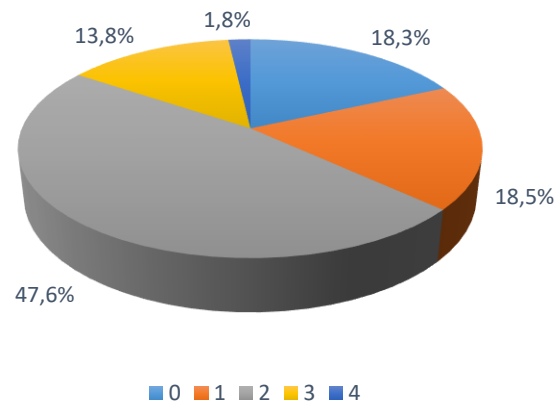


Derniers conseils

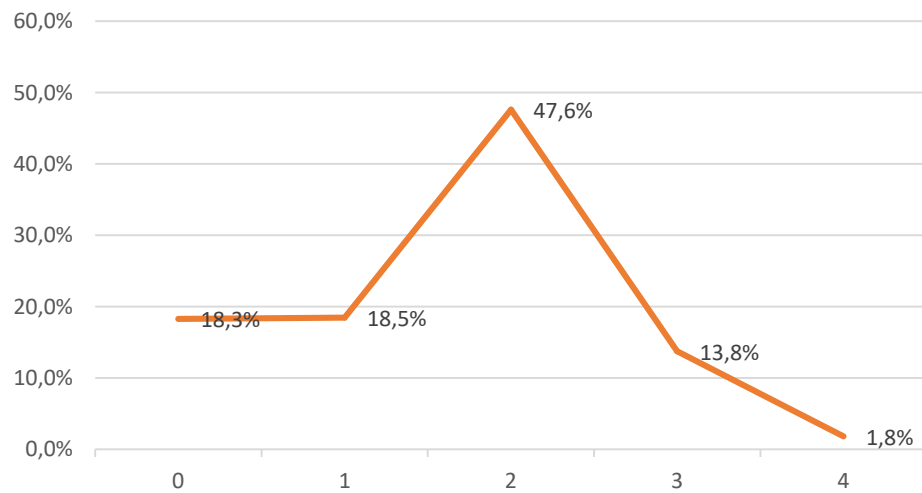
Répartition des allocataires de la CAF selon le nombre d'enfants



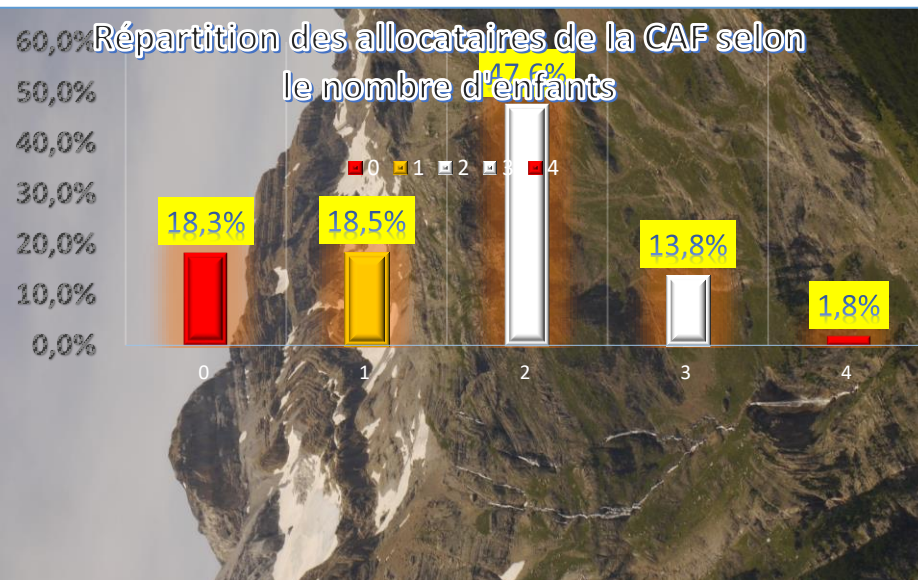
Répartition des allocataires de la CAF selon le nombre d'enfants



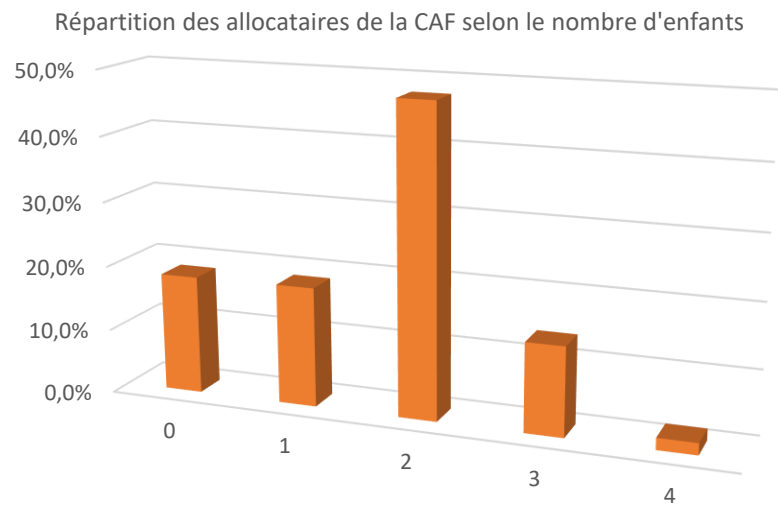
Répartition des allocataires de la CAF selon le nombre d'enfants



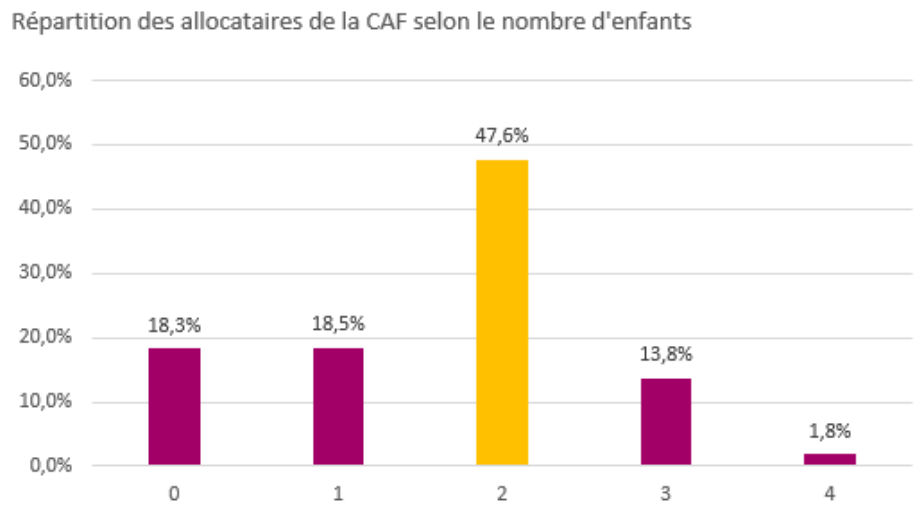
Répartition des allocataires de la CAF selon le nombre d'enfants



Exemple 1 : Nombre d'enfants



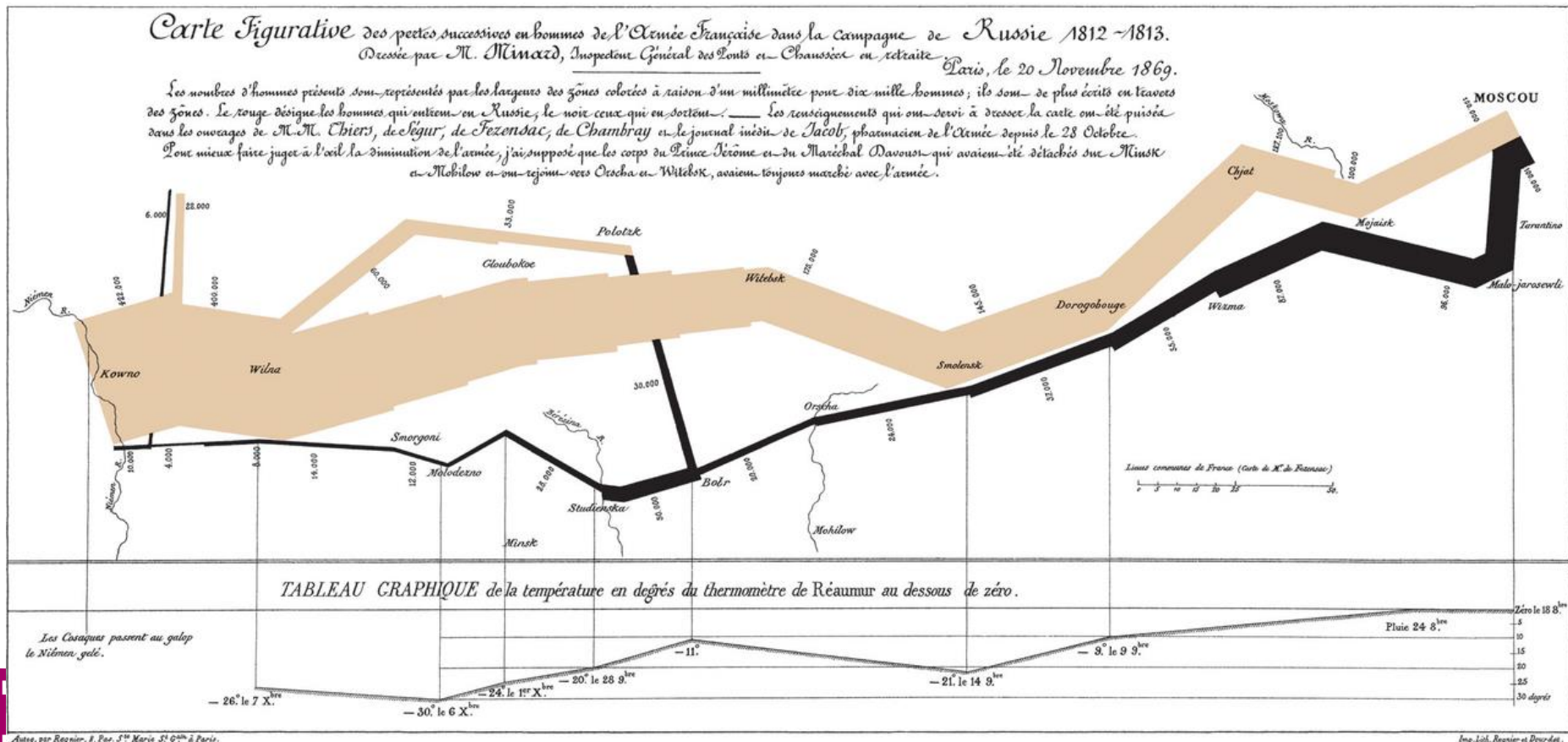
Près de la moitié des allocataires sont parents de deux enfants



Source : CAF Pau-Béarn – Décembre 2006

Bilan : qualité d'un bon graphique

- Restitution **fidèle** des données
- « Donne au lecteur **le plus grand nombre** d'informations pour **un temps d'attention le plus limité** tout en ayant recours à **la plus faible quantité d'encre** et au **plus petit espace** de représentation. » (Edward Tufte-1983)



Indicateurs

Variables quantitatives

NOTATIONS

- Variable statistique (ou caractère statistique) : ce qui est observé ou mesuré sur les individus de la population statistique.
- Notations : x , lettre minuscule
- x_i : valeur ou modalité observée chez le $i^{\text{ème}}$ individu
- $(x_i) = (x_1, \dots, x_n)$: **série statistique**

poids

88

80

75

91

83

...



Variable x

x_1

x_2

x_3

x_4

x_5

...

Indicateurs de tendance centrale (ou de position)

Ces indicateurs correspondent à une valeur « centrale » autour de laquelle les valeurs de la série sont censées se concentrer.

LA MOYENNE (ARITHMÉTIQUE)

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

- Si tous les individus avaient la même valeur, la moyenne correspond à la valeur qui devrait être observée chez chacun.
- manque de « robustesse »

LA MEDIANE

LA MEDIANE

- La médiane est une valeur telle que la moitié des individus ont une valeur inférieure à la médiane.
- La médiane sépare donc la population en deux sous-populations de même effectif (\approx même).
- *quantiles, décile d'ordre $\alpha\%$, quartile.*
- Indicateur robuste

Distribution des salaires mensuels nets en 2013 et évolution entre 2012 et 2013 en euros constants

Déciles	Ensemble		Hommes		Femmes	
	2013	Évolution (%)	2013	Évolution (%)	2013	Évolution (%)
D1	1 200	-0,6	1 254	-0,9	1 154	-0,3
D2	1 342	-0,5	1 415	-0,6	1 268	-0,3
D3	1 471	-0,4	1 559	-0,5	1 374	-0,2
D4	1 609	-0,3	1 709	-0,3	1 485	-0,1
D5 ou Médiane	1 772	-0,1	1 882	-0,2	1 619	0,1
D6	1 974	0,0	2 100	-0,1	1 794	0,3
D7	2 244	0,0	2 405	0,0	2 029	0,3
D8	2 682	0,0	2 921	-0,1	2 368	0,3
D9	3 544	-0,2	3 892	-0,1	3 036	0,0
C95	4 526	-0,2	5 030	-0,4	3 756	0,2
C99	8 061	-0,5	9 253	-0,6	6 053	0,1
Moyenne	2 202	-0,3	2 389	-0,4	1 934	0,0

Champ : France, salariés en EQTP du privé et des entreprises publiques, y compris les bénéficiaires de contrats aidés. Sont exclus les apprentis, les stagiaires, les salariés agricoles et les salariés des particuliers employeurs.

Source : Insee, DADS, fichier semi-définitif.

LE MODE

- Valeur la plus fréquente de la série statistique
- Variable continue : classe modale
- Lorsque qu'une variable se répartie de façon symétrique, la médiane, la moyenne et le mode sont proches.

Indicateurs de dispersion

Complètent l'information donnée par une caractéristique centrale (moyenne,...)

« Mesurent » la tendance qu'ont les valeurs de la série à se concentrer ou se disperser autour de cet indicateur.

INTERVALLES

- **Intervalle Min-max** : il permet de calculer l'étendue
Max-Min
- **Intervalle Interquartile** : $[Q1 ; Q3]$ représente 50%
de la population

VARIANCE ET ÉCART-TYPE

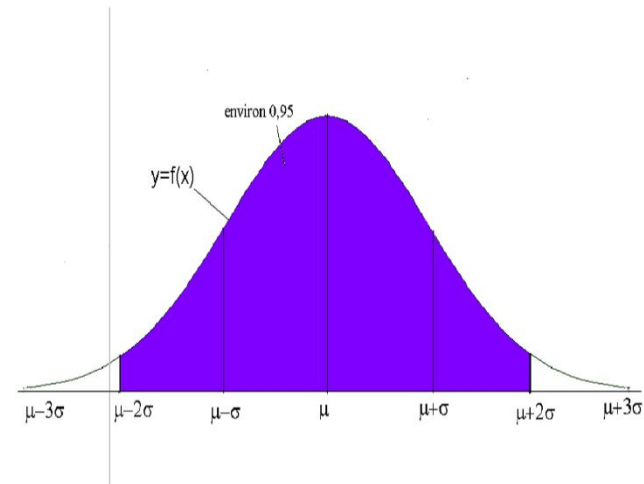
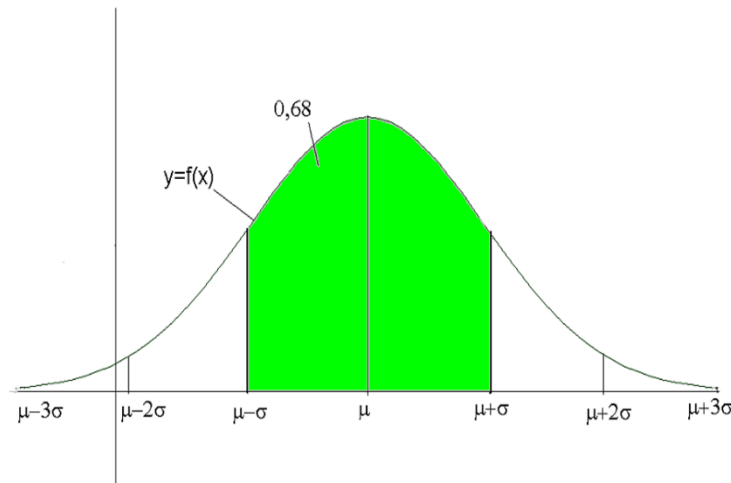
- Dispersion autour de la moyenne.
- Variance : moyenne des carrés des écarts de ces valeurs à la moyenne.

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

- Ecart-type : racine carrée de $Var(x)$

PAS DE RÈGLE D'INTERPRÉTATION

- Pour **comparer** la dispersion de plusieurs séries
- Cas particulier : réalisation d'une **loi normale**,



Calculs à partir des
distributions de fréquences
et des effectifs

Calcul à partir des fréquences

Poids x_i	Effectifs n_i	Fréquences $f_i = n_i/n$
...
97	22	0,01
98	48	0,03
99	21	0,01
100	66	0,04
101	23	0,01
102	47	0,03
103	27	0,02
104	43	0,03
...

p valeurs de la variable pour n individus

$$n = \sum_{i=1}^p n_i$$

Calcul à partir des fréquences

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \Rightarrow \quad \bar{x} = \frac{1}{n} \sum_{j=1}^p n_j x_j = \sum_{j=1}^p f_j x_j$$

Poids x_i	Effectifs n_i	Fréquences $f_i = n_i/n$
...
97	22	0,01
98	48	0,03
99	21	0,01
100	66	0,04
101	23	0,01
102	47	0,03
103	27	0,02
104	43	0,03
...

p valeurs de la variable pour n individus

$$n = \sum_{i=1}^p n_i$$

Calcul à partir des fréquences

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \Rightarrow \quad \bar{x} = \frac{1}{n} \sum_{j=1}^p n_j x_j = \sum_{j=1}^p f_j x_j$$

$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \Rightarrow \quad \text{Var}(x) = \frac{1}{n} \sum_{i=1}^n n_i (x_i - \bar{x})^2 = \sum_{i=1}^n f_i (x_i - \bar{x})^2$$

Poids x_i	Effectifs n_i	Fréquences $f_i = n_i/n$
...
97	22	0,01
98	48	0,03
99	21	0,01
100	66	0,04
101	23	0,01
102	47	0,03
103	27	0,02
104	43	0,03
...

p valeurs de la variable pour n individus

$$n = \sum_{i=1}^p n_i$$

Travail sous R

Fonctions R les plus courantes pour une analyse univariée :

- Graphiques : barplot, hist, pie
- Indicateurs : summary, mean, sd, quantile
- Tableaux de fréquence : table

Dans les fichiers cimas ou rugby :

1. Choisir une variable de chaque type :

- Faire un diagramme
- Calculer des indicateurs le cas échéant
- Faire un tableau de fréquence le cas échéant

2. Prendre une variable quantitative continue et la mettre en classe.

Comment faire pour obtenir des classes d'effectifs proches ?