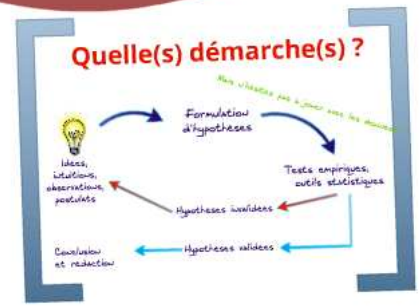
DataJournalisme Lab

Comprendre l'utilisation des statistiques dans le datajournalisme

Alexandre Bertin
Responsable Etudes et Diagnostic

2 février 2012



Un peu d'histoire

Premières traces lors des recensements du bétail en Egypte ancienne (20 siècles avant J.C.)

18ème siècle, premières tables de mortalité (Deparcieux, 1746)

19ème siècle, application industrielle des statistiques en sciences de gestion aux Etats-Unis d'Amérique

20ème siècle, premières lois de probabilité établies par les théoriciens eugénistes (Pearson, Fisher)

Années 1950, apparition de l'informatique et des traitements multidimensionnels des grandes bases de données

Quelques définitions

La statistique désigne l'ensemble des nombres qui résument quantitativement des informations brutes aussi appelées données

Les statistiques renvoient à la méthode utilisée pour calculer ces résumés numériques et en tirer des constats généraux

On parle de recensement lorsque l'on fait une étude exhaustive d'une population.
Si on n'étudie qu'une sous-population, qu'un groupe issu de la population mère, on parle d'échantillon

Les éléments de cette population et des échantillons sont des individus (qu'ils soient ou non des êtres humains : des boulons, par exemple, sont considérés comme les individus d'une population)

A chaque individu d'une population sont associés des caractères, appelés variables : par exemple, un être humain est caractérisé par son sexe, son poids, sa taille mais aussi sa commune de résidence et ses goûts artistiques. Un lot de boulons peut être décrit suivant le poids, le diamètre ou la matière.

Chaque variable peut présenter deux ou plusieurs modalités : sexe ->homme/femme, âge, PCS, etc.
Les modalités d'une variable doivent être "mutuellement exclusives" et "collectivement exhaustives"

Ce qui rend la variable utile scientifiquement est la mesure, qui est la procédure qui nous permet de trouver les valeurs d'une variable pour des cas différents

Les niveaux de mesure

Variables nominales : variables permettant d'identifier une personne appartenant à une classe. Se mesure de manière à ce que ses valeurs diffèrent les unes des autres. Ses valeurs sont des catégories sans ordre : Homme/Femme, les types de transports utilisés, les modes de logement etc.

Variables ordinales : variables dans lesquelles les catégories possibles peuvent être classées dans un ordre spécifique ou dans un ordre naturel quelconque : la PCS, le niveau d'éducation

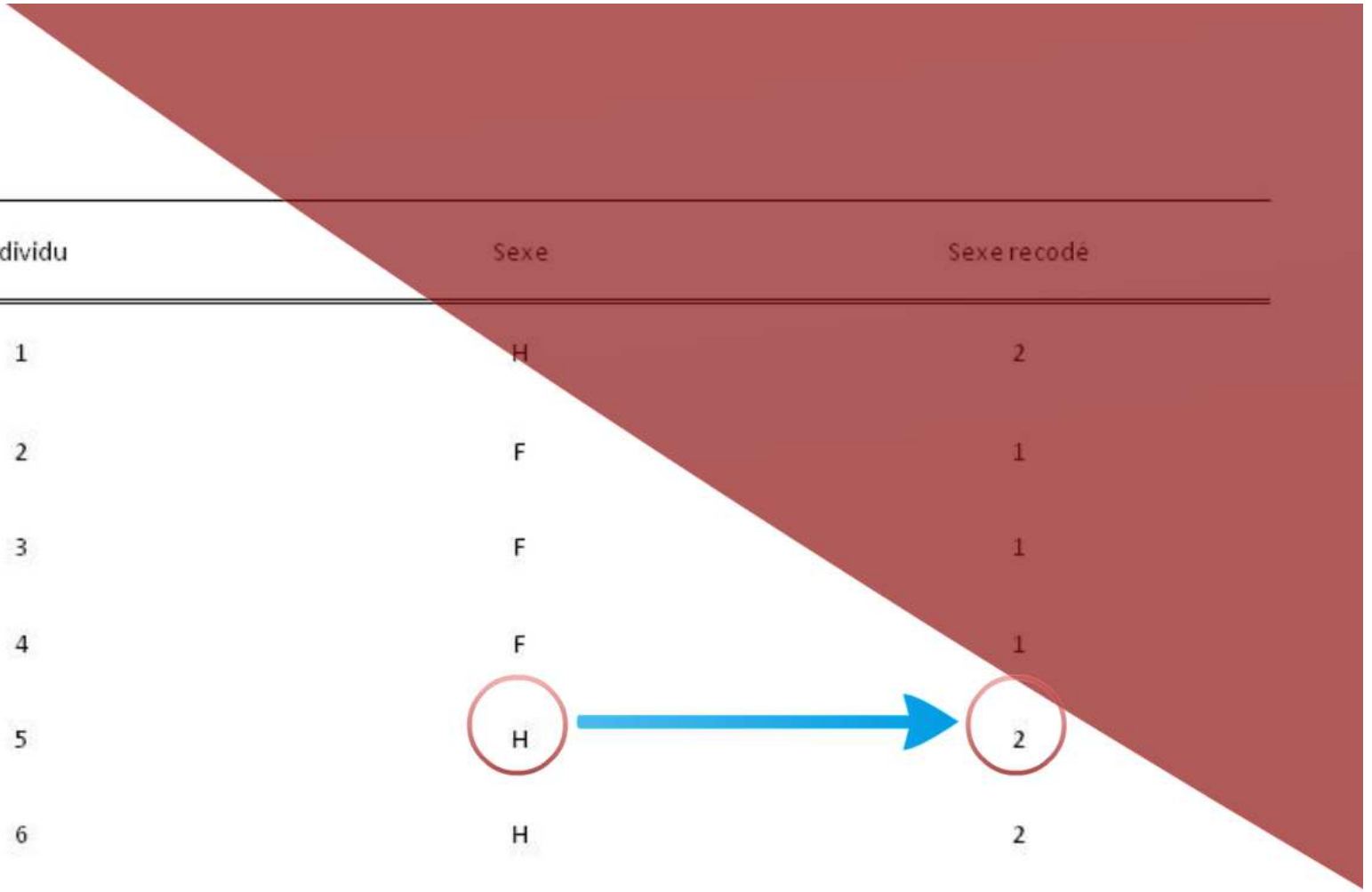
Variables d'intervalles : variables dont les valeurs peuvent être ordonnées mais également mesurées à l'aune d'une unité de mesure fixe ou standard : le poids, la température

Variables continues : On dit qu'une variable est continue si elle peut supposer un nombre infini de valeurs réelles. La taille, la température ou l'âge sont des variables continues.

Variables discrètes : variables qui ne peuvent prendre que certaines valeurs dans son étendue. La taille d'une famille par exemple, le nombre de salariés d'une entreprise, le nombre d'amis sur Facebook

Variables de classe : lorsque l'information est trop importante on peut faire le choix de perdre en qualité mais gagner en praticité en recodant les valeurs des variables en classes : l'âge par exemple peut être classifier, la superficie, la taille

Individu	Sexe	Sexe recodé
1	H	2
2	F	1
3	F	1
4	F	1
5	H	2
6	H	2
7	H	2
8	F	1
9	F	1
10	H	2



Individu	Diplôme	Diplôme recodé
1	Brevet	2
2	CAP	3
3	Bac	4
4	Bac	4
5	Bac+3	5
6	Bac+5	6
7	Bac	2
8	Sans diplôme	1
9	Bac	4
10	Bac	4

Individu	Taille (en cm)
1	182
2	159
3	175
4	167
5	189
6	178
7	179
8	162
9	160
10	198

Individu (ménages)	Taille du ménage (en nombre de personnes)
1	1
2	7
3	4
4	4
5	3
6	2
7	4
8	5
9	3
10	1

Individu	Taille (en cm)
1	182
2	159
3	175
4	167
5	189
6	178
7	179
8	162
9	160
10	198

Taille (en classe)	Nombre d'individus
[150;160[	1
[160;170[	3
[170;180[	3
[180;190[	2
[190;200[	1
Total	10

Quelle(s) démarche(s) ?

Idées, intuitions, observations, postulats

Formulation d'hypothèses

Mais n'hésitez pas à jouer avec les données !

Tests empiriques, outils statistiques

Hypothèses invalidées

Hypothèses validées

Conclusion et rédaction

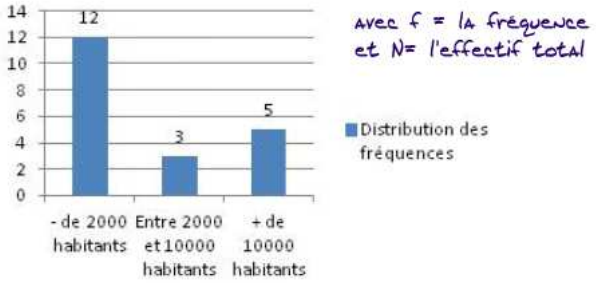
La distribution des fréquences (ou des effectifs) et des pourcentages

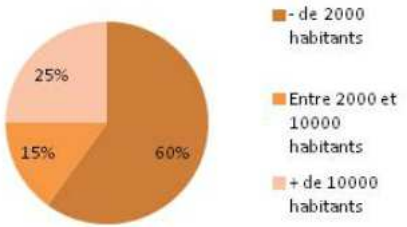
Nom	Taille	Département
Aast	- de 2000 habitants	Pyrénées-Atlantiques
Abjat-sur-Bandiat	- de 2000 habitants	Dordogne
Agen	+ de 10000 habitants	Lot-et-Garonne
Agnos	- de 2000 habitants	Pyrénées-Atlantiques
Ahetze	- de 2000 habitants	Pyrénées-Atlantiques
Aiguillon	Entre 2000 et 10000 habitants	Lot-et-Garonne
Ainhoa	- de 2000 habitants	Pyrénées-Atlantiques
Aire-sur-l'Adour	Entre 2000 et 10000 habitants	Landes
Allemans-du-Dropt	- de 2000 habitants	Lot-et-Garonne
Alos-Sibas-Abense	- de 2000 habitants	Pyrénées-Atlantiques
Ambarès-et-Lagrave	+ de 10000 habitants	Gironde
Ambès	Entre 2000 et 10000 habitants	Gironde
Andernos-les-Bains	+ de 10000 habitants	Gironde
Anglet	+ de 10000 habitants	Pyrénées-Atlantiques
Annesse-et-Beaulieu	- de 2000 habitants	Dordogne
Aramits	- de 2000 habitants	Pyrénées-Atlantiques
Arbanats	- de 2000 habitants	Gironde
Arbonne	- de 2000 habitants	Pyrénées-Atlantiques
Arbus	- de 2000 habitants	Pyrénées-Atlantiques
Arcachon	+ de 10000 habitants	Gironde

Taille de la commune	Distribution des fréquences	Distribution des pourcentages
- de 2000 habitants	12	60%
Entre 2000 et 10000 habitants	3	15%
+ de 10000 habitants	5	25%
Total	*20*	100%

Pourcentage = f/N*100

avec f = la fréquence et N= l'effectif total





Tableaux croisés, effectifs marginaux

Tableau de contingence (Taille / Département):

	Dordogne	Gironde	Landes	Garonne	Pyrénées Atlantiques	*Total*
- de 2000 habitants	2	1	0	1	8	*12*
Entre 2000 et 10000 habitants	0	1	1	1	0	*3*
+ de 10000 habitants	0	3	0	1	1	*5*
Total	*2*	*5*	*1*	*3*	*9*	*20*

Proportions / Colonne (Taille / Département):

	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées Atlantiques	*Total*
- de 2000 habitants	100%	20%	0%	33%	89%	*60%*
Entre 2000 et 10000 habitants	0%	20%	100%	33%	0%	*15%*
+ de 10000 habitants	0%	60%	0%	33%	11%	*25%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

Proportions / Ligne (Taille / Département):

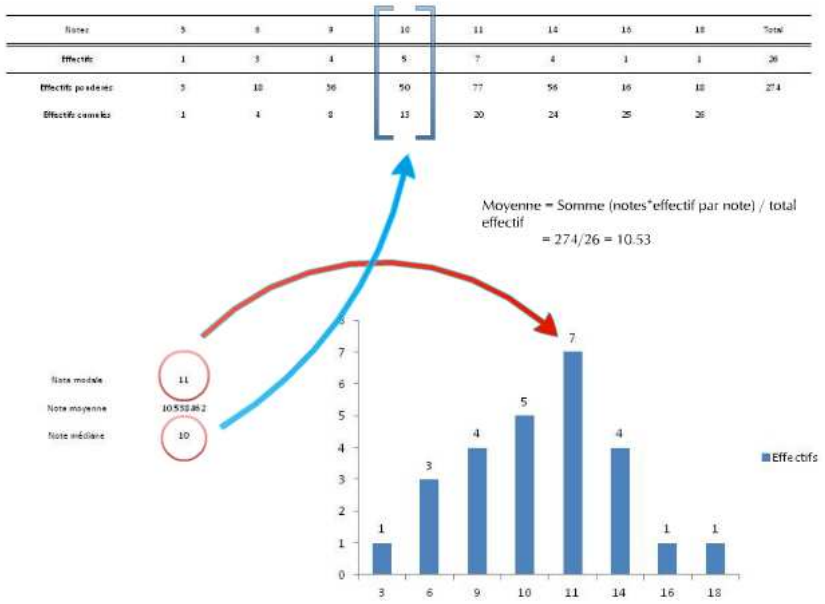
	Dordogne	Gironde	Landes	Lot-et-Garonne	Pyrénées Atlantiques	*Total*
- de 2000 habitants	17%	8%	0%	8%	67%	*100%*
Entre 2000 et 10000 habitants	0%	33%	33%	33%	0%	*100%*
+ de 10000 habitants	0%	60%	0%	20%	20%	*100%*
Total	*10%*	*25%*	*5%*	*15%*	*45%*	*100%*

Les mesures de tendance centrale

Le mode est le score qui apparaît le plus fréquemment pour une variable donnée

La médiane est la valeur qui divise en deux parties égales un ensemble ordonné de scores

La moyenne est la valeur que l'on obtient en rapport le nombre de scores pour une variable sur le nombre total de scores



obtient en rapport le nombre de scores pour une variable

Notes	3	6	9	10	11	14	16	18	Total
Effectifs	1	3	4	5	7	4	1	1	26
Effectifs pondérés	3	18	36	50	77	56	16	18	274
Effectifs cumulés	1	4	8	13	20	24	25	26	

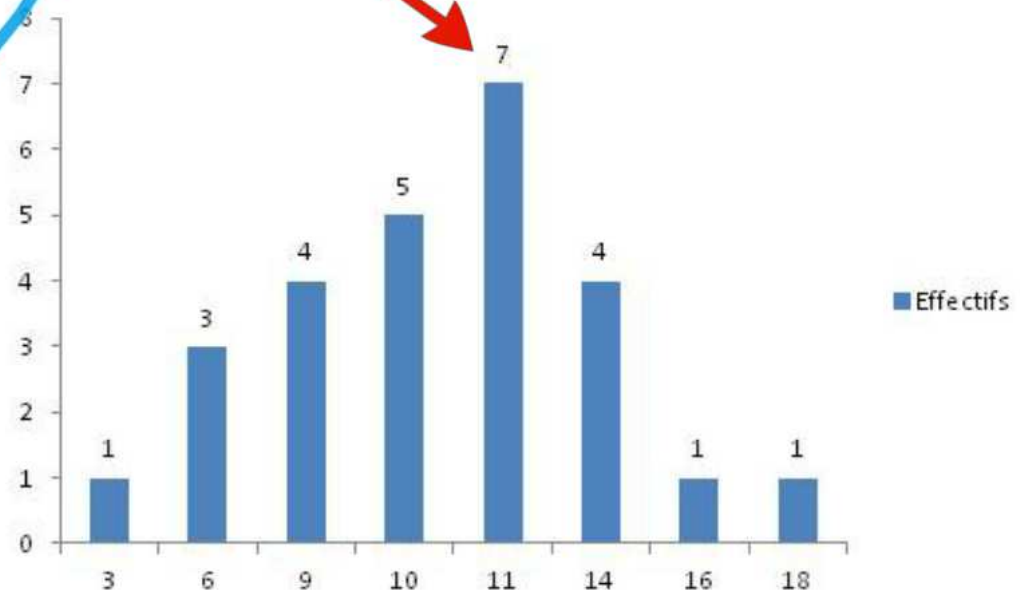
Moyenne = Somme (notes*effectif par note) / total effectif

= 274/26 = 10.53

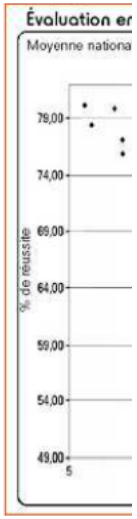
Note modale 11

Note moyenne 10538462

Note médiane 10



Moyenne, mode ou médiane ?



Lorsque la distribution des effectifs présente un profil plutôt symétrique, l'usage de la moyenne est fortement conseillé car elle est simple à mettre en oeuvre, simple à interprêter et qu'elle est la seule mesure qui incorpore la totalité des scores.

Si la distribution est fortement asymétrique ou s'il existe des valeurs extrêmes, la moyenne fausse la réalité de la situation. L'usage de ma médiane est alors préférable car elle lisse la distribution.

La médiane est également préférée lorsque les variables sont ordinales comme le niveau de diplôme par exemple ou la PCS

Pyrénées Atlantiques	Total
8	22
6	8
1	5
9	20

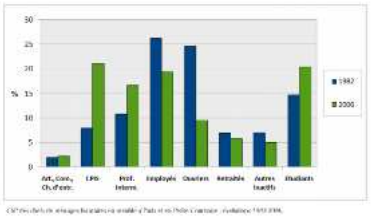
ne	Pyrénées Atlantiques	Total
0%	88%	60%
0%	0%	15%
0%	12%	25%
0%	100%	100%

e	Pyrénées Atlantiques	Total
	67%	100%
	0%	100%
	20%	100%
	45%	100%

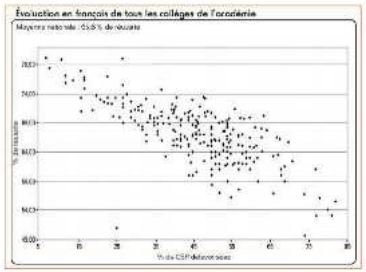
La visualisation des données

Graphique en secteurs (camembert)

Histogrammes



Nuage de points



Courbe d'évolution dans le temps (timeline)

Cartographie

...nne, mode ou médiane ?

...fs présente un profil plutôt symétrique, l'usage de la ... car elle est simple à mettre en oeuvre, simple à ... nesure qui incorpore la totalité des scores.

...métrique ou s'il existe des valeurs extrêmes, la moyenne ...sage de ma médiane est alors préférable car elle lisse la ...

...ée lorsque les variables sont ordinales comme le niveau ...S

Conclusion

Idées, réflexion et rigueur

Curiosité, bidouille et jeu

Mais attention !!

aux interprétations péremptoires

corrélation n'est pas causalité

attention aux corrélations cachées