Formation ElementR

Statistiques univariées et bivariées

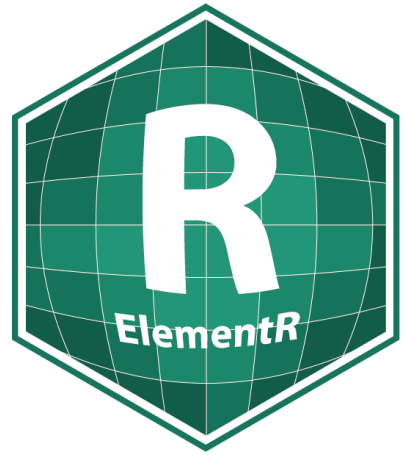
Sylvestre Duroudier

21 mars 2023

Contact : sylvestre.duroudier@univ-paris1.fr

Géographie-cités
UMR 8504

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

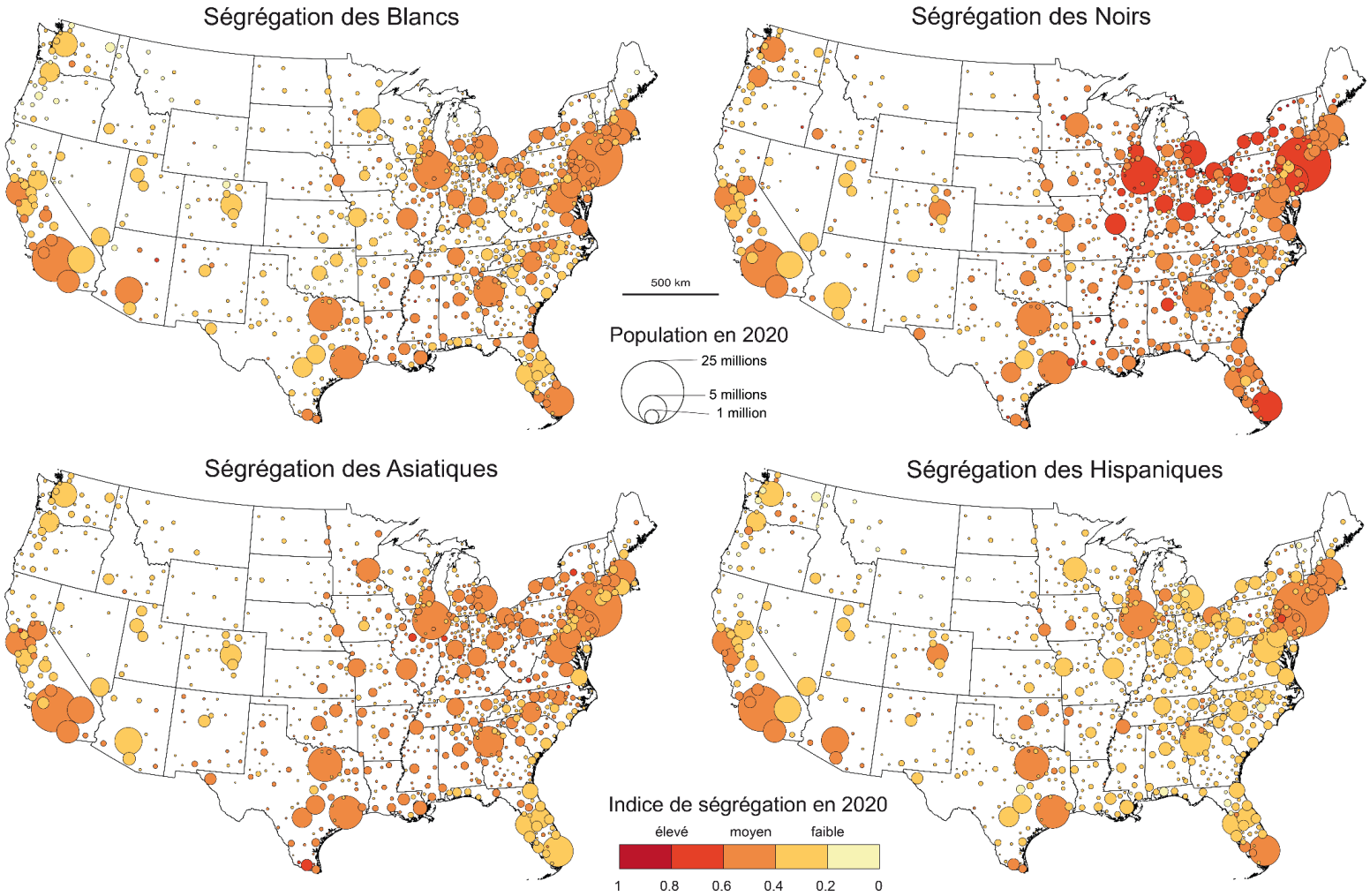


Des statistiques en géographie

Les statistiques en géographie sont l'étude des variations géographiques de phénomènes localisés.

Différents objectifs :

- Explorer des données
- Mesurer des phénomènes
- Caractériser des variations
- Identifier des formes et des régions
- Expliquer des phénomènes par d'autres



Exemple : Cartes des appartenances ethno-raciales des aires métropolitaines des Etats-Unis en 2020.

Des statistiques en géographie

Les statistiques comme un ensemble des techniques et des méthodes pour analyser une matrice d'information géographique.

Vocabulaire :

- Caractère : variable d'intérêt (colonnes)
- Individus : entités géographiques, unités spatiales (lignes)
- Valeur (ou modalité) : chiffre ou texte pris par un individu donné pour un caractère donné
- Statistiques univariées : étude de la variation d'un seul caractère.
- Statistiques bivariées : étude de la covariation de deux caractères.
- Approche descriptive : caractériser des caractères (par des indicateurs, des graphiques, des cartes).
- Approche explicative : prédire/modéliser une caractère par un autre.

CBSAA	CBSAN	State	POP20	WHITE	BLACK	ASIAN	HISPA	OTHER
10100	Aberdeen	SD	42287	36484	547	985	1720	2551
10140	Aberdeen	WA	75625	57056	972	1151	7833	8613
10180	Abilene	TX	175610	109578	13076	3142	41446	8368
10220	Ada	OK	38065	22703	861	385	2208	11908
10300	Adrian	MI	99424	83126	2437	484	8495	4882
10420	Akron	OH	702212	536305	86857	26093	16711	36246
10460	Alamogordo	NM	67838	30931	2299	1146	26151	7311
10500	Albany	GA	151797	62519	78973	1697	4067	4541
10540	Albany	OR	128609	104117	623	1714	12571	9584
10580	Albany-Schenectady-Troy	NY	895799	668079	72201	46260	53166	56093
10620	Albemarle	NC	62504	48645	7000	1149	3086	2624
10660	Albert Lea	MN	29324	23612	456	1122	3129	1005
10700	Albertville	AL	97612	74666	2293	707	15658	4288
10740	Albuquerque	NM	916530	349194	21562	22702	439141	83931
10780	Alexandria	LA	152199	93004	43323	1911	6424	7537
10820	Alexandria	MN	39006	36629	223	238	815	1101
10860	Alice	TX	38891	6963	180	155	30835	758
10900	Allentown-Bethlehem-Easton	PA-NJ	859645	596312	46964	26838	157312	32219
10940	Alma	MI	41761	34812	2195	189	3153	1412
10980	Alpena	MI	28907	27006	99	124	417	1261
11020	Altoona	PA	122815	113008	2464	880	1708	4755
11060	Altus	OK	24785	14883	1614	397	5937	1954
11100	Amarillo	TX	268689	153401	16623	8754	77790	12121
11140	Americus	GA	31049	13584	14730	367	1554	814
11180	Ames	IA	96834	78389	2980	5469	4962	5034
11220	Amsterdam	NY	49532	38237	987	488	7312	2508

Exemple : Tableau des appartenances ethno-raciales des aires métropolitaines des Etats-Unis en 2020 (extrait).

Des statistiques en géographie

Différents types de caractères

1. Quantitatif :

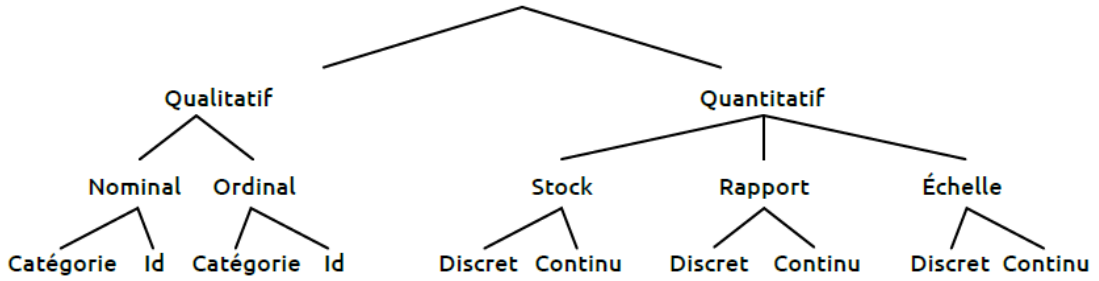
    - Stock : un effectif dont le 0 marque l'absence et la somme des individus a un sens.
    - Rapport : un ratio, un pourcentage, une quantité dont la somme des individus n'a pas de sens.
    - Echelle : la valeur 0 est un repérage conventionnel

2. Qualitatif :

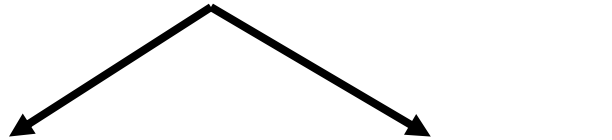
    - Nominal : il n'y a pas d'ordre entre les modalités
    - Ordinal : il y a une hiérarchie entre les modalités

Les méthodes statistiques dépendent de la nature des caractères…

Les natures des caractères



Les types d'analyses univariées



Les types d'analyses bivariées

	Variable quantitative	**Variable qualitative**
Variable quantitative	Corrélation, régression linéaire	Analyse de la variance (anova)
Variable qualitative	Analyse de la variance (anova)	Analyse du Chi2

Des statistiques en géographie

Du côté théorique

Au programme :

1. Statistiques univariées
2. Bivarié quantitatif
3. Bivarié qualitatif
4. Bivarié quali-quanti (?)

A chaque étape :

- Des mesures
- Des graphiques
- Des cartes
- Des exercices

Du côté R

Des packages :

- *tidyverse* : gestion de tableau
- *sf* : gestion de données spatiales
- *lmtest* : aide au bivarié
- *RColorBrewer* : palettes de couleur
- *ggplot2* : réalisation de graphiques
- *tmap* : cartographie

Des données : Appartenances ethno-raciales dans les aires métropolitaines des Etats-Unis en 2020.

Des fonds de cartes : aires métropolitaines et Etats fédérés.

Des exemples filés : les Hispaniques, les régions, les types de villes selon la taille

1. Statistiques univariées

Objectif principal : caractériser la distribution statistique d'un caractère.

summary()

1. Les bornes
   - Minimum — min()
   - Maximum — *max()*
2. Les valeurs centrales
   - Moyenne — *mean()*
   - Médiane — *median()*

	ALGORITHME	PROPRIÉTÉ
MOYENNE	Faire la somme des valeurs et diviser par l'effectif	Minimise la somme du carré des écarts entre toutes les valeurs et elle-même
MÉDIANE	Ordonner la série et trouver la valeur qui la découpe en deux ensembles d'eff. égaux	Minimise la somme de la valeur absolue des écarts entre toutes les valeurs et elle-même

```
> # 2.1 Statistiques univariées  ====
> # ........................................
>
> # Généralités  ====
> # ........................................
>
> # obtenir un résumé partiel d'une variable
> summary(db$HISPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    224    2125    6111   66004   20484 5835137
> ## la fonction reprend min(), max(), median(), mean(), quantile(,0.25), quantile(,0.75)
>
> # obtenir un résumé d'un ensemble de variables
> summary(db)
    CBSAA              CBSAN              State              Division       Region
 Length:909         Length:909         Length:909         Min.   :1.00   Min.   :1.000   L
 Class :character   Class :character   Class :character   1st Qu.:3.00   1st Qu.:2.000   C
 Mode  :character   Mode  :character   Mode  :character   Median :5.00   Median :3.000   M
                                                          Mean   :5.19   Mean   :2.674
                                                          3rd Qu.:7.00   3rd Qu.:3.000
                                                          Max.   :9.00   Max.   :4.000
     POP20              WHITE              BLACK              ASIAN              HISPA
 Min.   :   12456   Min.   :   796   Min.   :     9   Min.   :    16   Min.   :   224
 1st Qu.:   40105   1st Qu.: 30095   1st Qu.:   755   1st Qu.:   310   1st Qu.:  2125
 Median :   75922   Median : 56204   Median :  2985   Median :   921   Median :  6111
 Mean   :  339173   Mean   :192926   Mean   : 41887   Mean   : 20963   Mean   : 66004
 3rd Qu.:  186468   3rd Qu.:133178   3rd Qu.: 16623   3rd Qu.:  3823   3rd Qu.: 20484
 Max.   :20626380   Max.   :9033783   Max.   :3055666   Max.   :2477126   Max.   :5835137
```

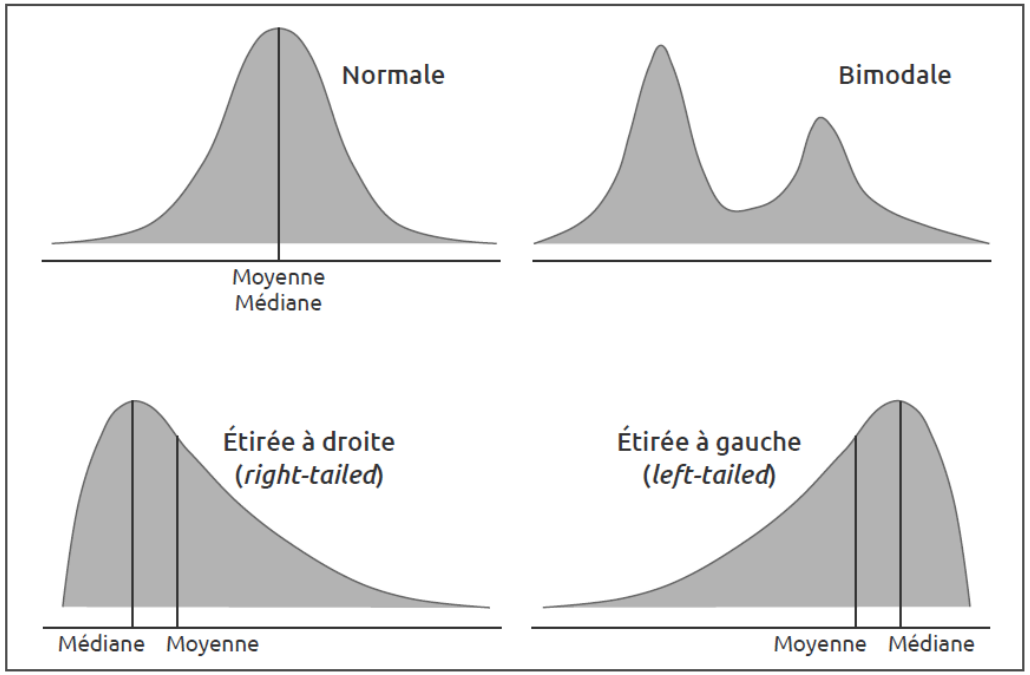
```
> # problème pour les variables qualitatives
> summary(as.factor(db$DEFI))
  L  LM   M  MS   S
  7   8  29   7 858
> summary(as.factor(db$Region))
  1   2   3   4
 91 282 368 168
>
> # juste les indices de ségrégation
> summary(db[,19:23], digits = 2)
    WHITEis         BLACKis         ASIANis         HISPAis         OTHERis
 Min.   :0.08   Min.   :0.14   Min.   :0.10   Min.   :0.09   Min.   :0.05
 1st Qu.:0.23   1st Qu.:0.36   1st Qu.:0.31   1st Qu.:0.22   1st Qu.:0.11
 Median :0.31   Median :0.43   Median :0.37   Median :0.28   Median :0.14
 Mean   :0.31   Mean   :0.43   Mean   :0.37   Mean   :0.29   Mean   :0.16
 3rd Qu.:0.38   3rd Qu.:0.50   3rd Qu.:0.43   3rd Qu.:0.35   3rd Qu.:0.17
 Max.   :0.70   Max.   :0.84   Max.   :0.67   Max.   :0.62   Max.   :0.80
> # Indicateurs  ====
> # ........................................
>
> min(db$HISPA) # minimum
[1] 224
> max(db$HISPA) # maximum
[1] 5835137
> mean(db$HISPA) # moyenne
[1] 66004.3
> median(db$HISPA) # mediane
[1] 6111
```

1. Statistiques univariées

Objectif principal : caractériser la distribution statistique d'un caractère.

3. Les paramètres de dispersion

- Etendue : max - min — range()
- Variance : moyenne des carrés des écarts à la moyenne (sans ordre de grandeur) — *var()*
- Ecart-type : racine carrée de la variance (dans l'ordre de grandeur) — *sd()*
- Coefficient de variation : écart-type divisé par la moyenne (entre 0 et +inf) — *sd() / mean()*
- Quantiles : valeurs prises pour des partitions en effectifs égaux (quartiles, quintiles, déciles…) — *quantile(var, ratio)*
- Intervalle inter-quantiles : par ex. intervalle interquartiles
  = quartile 3 – quartile 1 — *quantile(,0.75)-quantile(,0.25*




```
> range(db$HISPA) # étendue
[1]      224 5835137
> var(db$HISPA) # variance
[1] 1.13955e+11
> sd(db$HISPA) # ecart-type
[1] 337572.3
> sd(db$HISPA) / mean(db$HISPA) # coefficient de variation
[1] 5.114398
> quantile(db$HISPA, 0.25) # quartile 1
 25%
2125
> quantile(db$HISPA, 0.9) # décile 9
    90%
94001.6
```

1. Statistiques univariées

Objectif principal : caractériser la distribution statistique d'un caractère.

4. Pour les variables qualitatives

- Le mode : la modalité la plus fréquente

Obtenir un dénombrement :

fonction *table()*

Méthode plus riche :

package *questionr*,

fonction *freq(variable, valeurs cumulées, total)*

```
> # pour les variables qualitatives, le mode
> table(db$DEFI)

  L  LM   M  MS   S
  7   8  29   7 858
>
> ## une variante utile
> # install.packages(questionr)
> library(questionr)
Warning message:
le package 'questionr' a été compilé avec la version R 4.3.3
> defi <- freq(db$DEFI,
+              cum = FALSE,
+              total = TRUE)
> defi
        n     %  val%
L       7   0.8   0.8
LM      8   0.9   0.9
M      29   3.2   3.2
MS      7   0.8   0.8
S     858  94.4  94.4
Total 909 100.0 100.0
```

1. Statistiques univariées

Représentations graphiques univariées

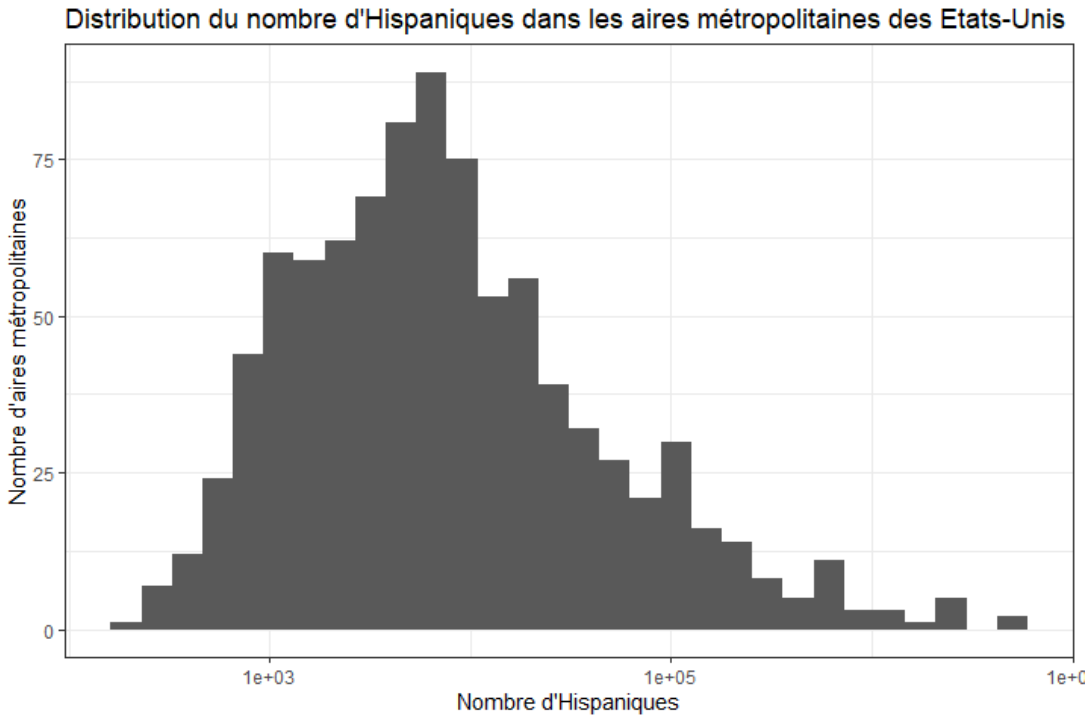
➔ Variables quantitatives : histogramme

➔ Utilisation de *ggplot() + geom_histogram()*

```r
# Graphiques ====
# ........................................

# Variable quantitative = histogramme
graph <- ggplot(db, aes(x = HISPA))+
  geom_histogram(binwidth = 100000) +
  theme_bw() +
  labs(
    title = "Distribution du nombre d'Hispaniques dans les aires métropolitaines des Etats-Unis",
    x = "Nombre d'Hispaniques",
    y = "Nombre d'aires métropolitaines"
  )
graph

graph <- ggplot(db, aes(x = HISPA)) +
  geom_histogram() +
  scale_x_log10() +
  theme_bw() +
  labs(
    title = "Distribution du nombre d'Hispaniques dans les aires métropolitaines des Etats-Unis",
    x = "Nombre d'Hispaniques",
    y = "Nombre d'aires métropolitaines"
  )
graph
```

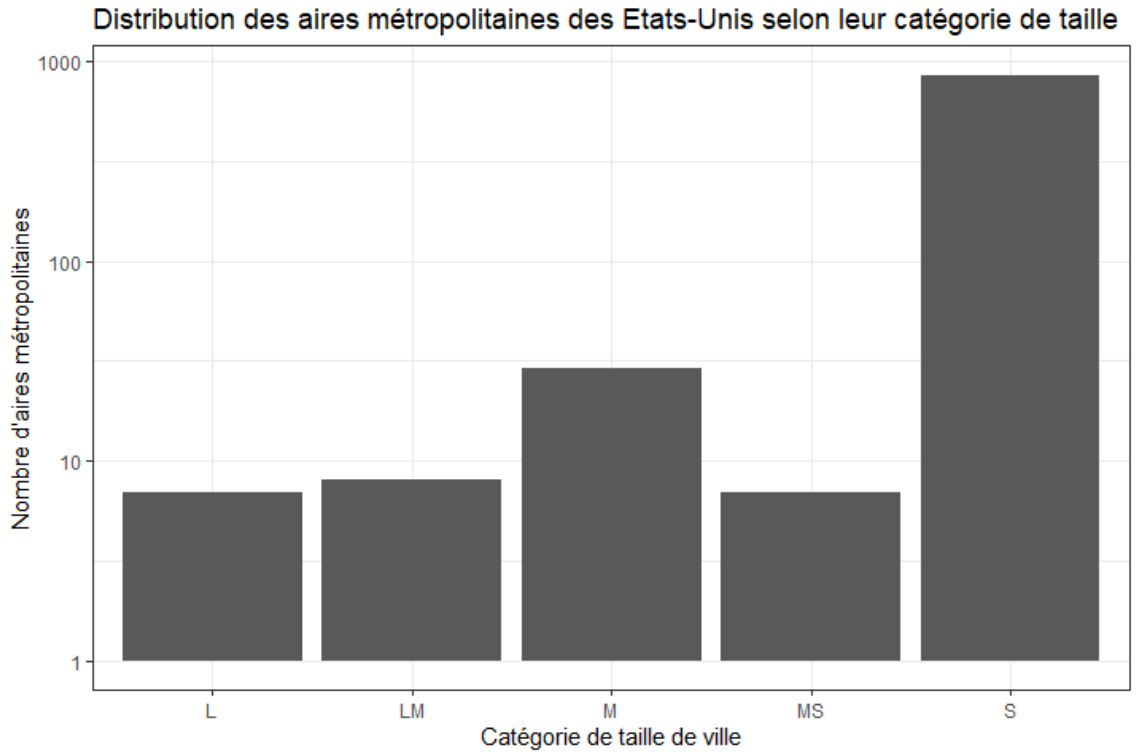


1. Statistiques univariées

Représentations graphiques univariées

➔ Variables qualitatives : diagramme en barres

➔ Utilisation de *ggplot() + geom_bar()*




```r
# variable qualitative = diagramme en barre
graph <- ggplot(db, aes(x = DEFI))+
  geom_bar() +
  theme_bw() +
  scale_y_log10() +
  labs(
    title = "Distribution des aires métropolitaines des Etats-Unis selon leur catégorie de taille",
    x = "Catégorie de taille de ville",
    y = "Nombre d'aires métropolitaines"
  )
graph
```

1. Statistiques univariées

Connaître la distribution pour cartographier : la discrétisation

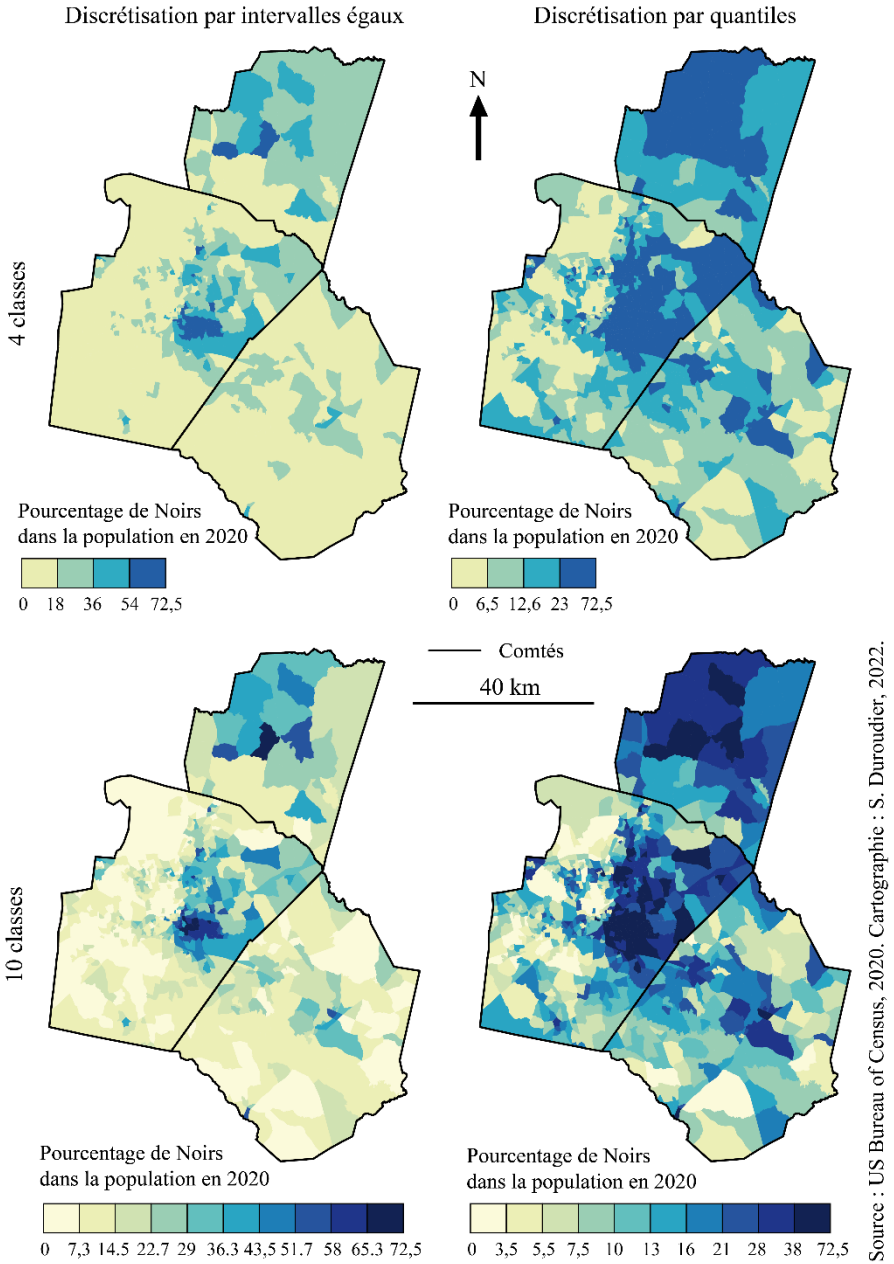
Discrétisation : transformation d'un caractère quantitatif continu en un caractère discret.

= découper une série en classes, chaque entité appartient à 1 seule classe

Combien de classes ? Par convention, entre 3 et 8.

Quelle méthode ?

- Seuils naturels
- Effectifs égaux (quantiles)
- Amplitudes égales (par ex. moyenne et écart-type)
- Progressions arithmétiques ou géométriques
- Algorithme de Jenks



Source : Duroudier, 2023, ISTE.

1. Statistiques univariées

Exercices : Par des calculs et des graphiques…

1. Identifier la répartition des villes selon les régions.
    - Quelle est la région modale ?
    - Quelle est la région comptant le moins de villes ?
2. Déterminer la forme de la distribution des Noirs dans les villes des Etats-Unis.
    - Quelle est l'étendue ?
    - Quel est le coefficient de variation ?
    - A quel pourcentage est le décile des villes ayant le moins de Noirs en poids relatif ?
3. Idem pour l'indice de ségrégation des Noirs.
    - Quelles sont les moyenne et médiane ?
    - Le coefficient de variation est-il supérieur ou inférieur à celui du pourcentage de Noirs ?

2. Bivarié quantitatif

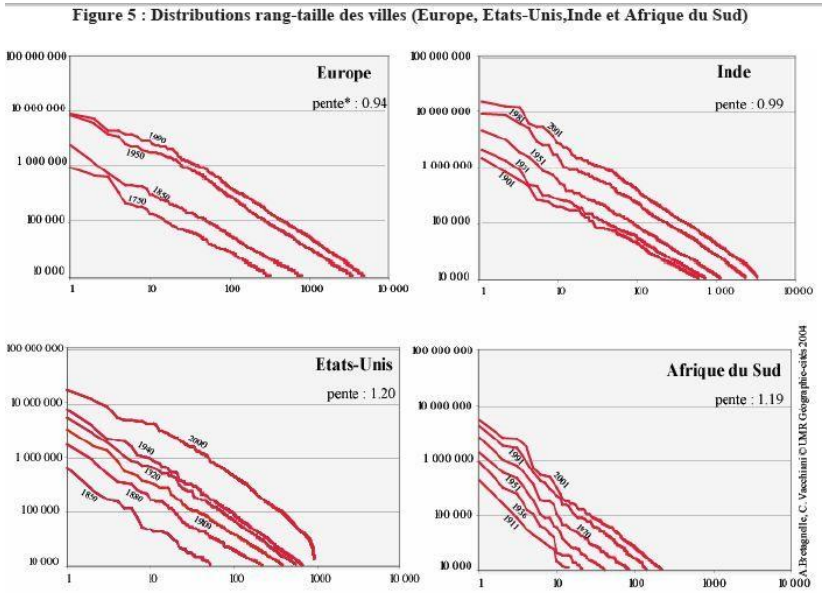
Objectifs :

- Analyser la covariation entre 2 caractères quantitatifs
- Expliquer un caractère quantitatif par un autre

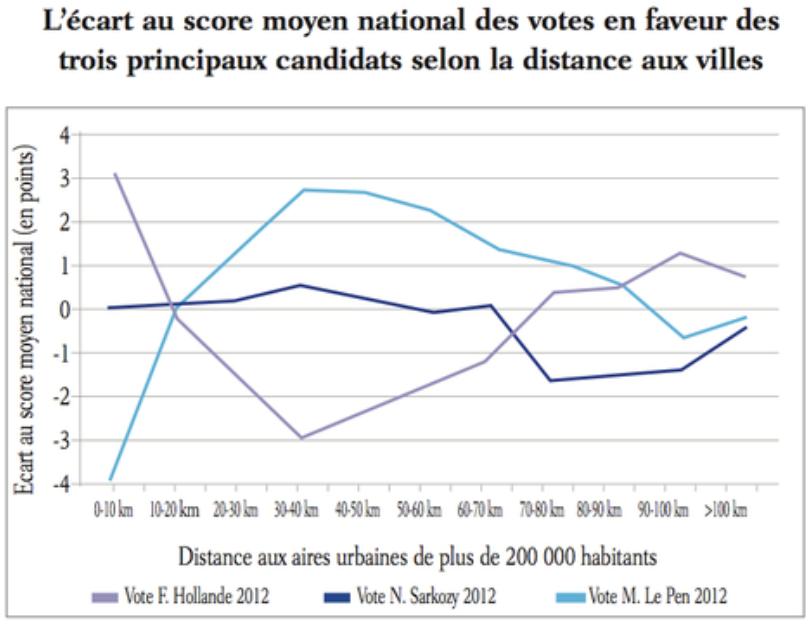
Notion d'« explication » en statistique :

- Expliquer une variation inconnue avec du connu, c'est-à-dire apporter de l'information sur un phénomène par des covariations.
- Explication n'est pas causalité !
- Distinction entre 2 caractères :
    - Variable « à expliquer », notée Y et axe des ordonnées
    - Variable « explicative », notée X et axe des abscisses

Exemple : l'effectif des Hispaniques dépend-il de la taille des villes en termes de population ?



Source : Pumain et al, 2005.



Source : Slate, 2012.

2. Bivarié quantitatif

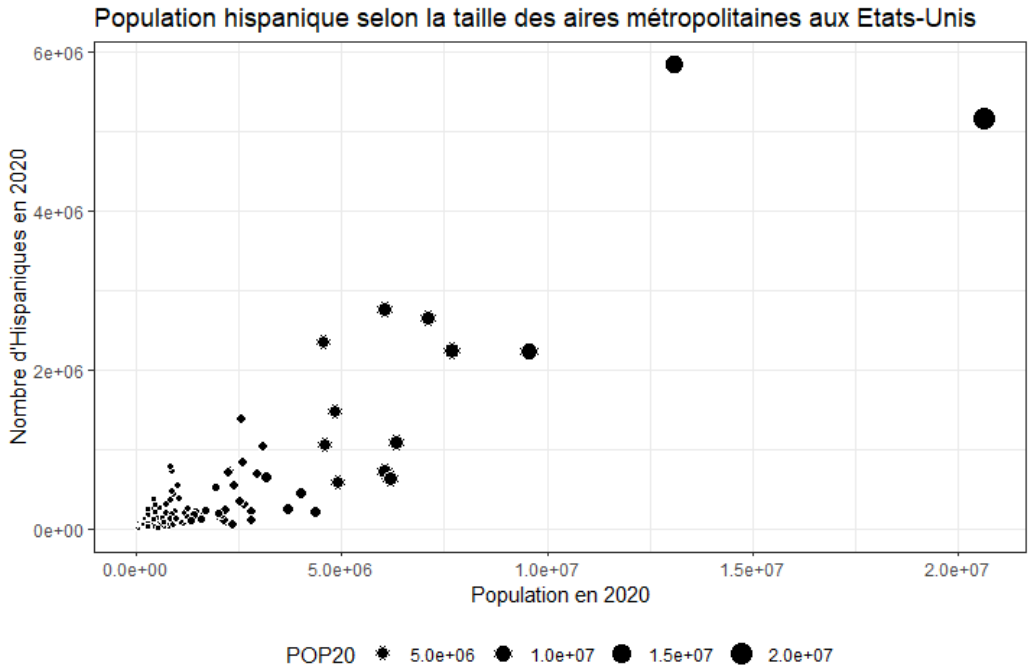
Approche descriptive par un graphique : nuage de points

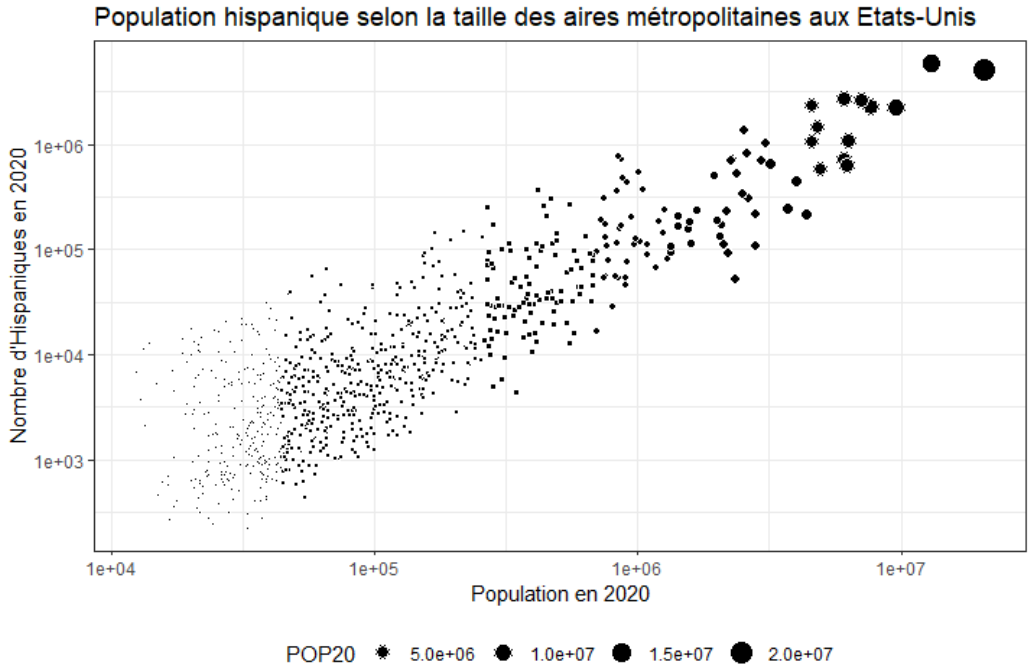
fonction ggplot() + geom_point()

```r
# Graphique ====
# ........................................

# Référentiel arithmétique
graph <- ggplot(db, aes(x = POP20, y = HISPA)) +
  geom_point(alpha = 1,
             shape = 21,
             color = "white",
             fill = "black",
             stroke = 1/20,
             aes(size = POP20)) +
  theme_bw() +
  labs(title ="Population hispanique selon la taille des aires métropolitaines aux Etats-Unis",
       x = "Population en 2020", y = "Nombre d'Hispaniques en 2020") +
  theme(legend.position = "bottom")
graph

# Référentiel logarithmique
graph <- ggplot(db, aes(x = POP20, y = HISPA)) +
  geom_point(alpha = 1,
             shape = 21,
             color = "white",
             fill = "black",
             stroke = 1/20,
             aes(size = POP20)) +
  scale_x_log10() +
  scale_y_log10() +
  theme_bw() +
  labs(title ="Population hispanique selon la taille des aires métropolitaines aux Etats-Unis",
       x = "Population en 2020", y = "Nombre d'Hispaniques en 2020") +
  theme(legend.position = "bottom")
graph
```





2. Bivarié quantitatif

Approche descriptive par une mesure : le coefficient de corrélation

Coef. de Bravais Pearson :

Principe : calculer la covariation statistique entre 2 caractères quantitatifs

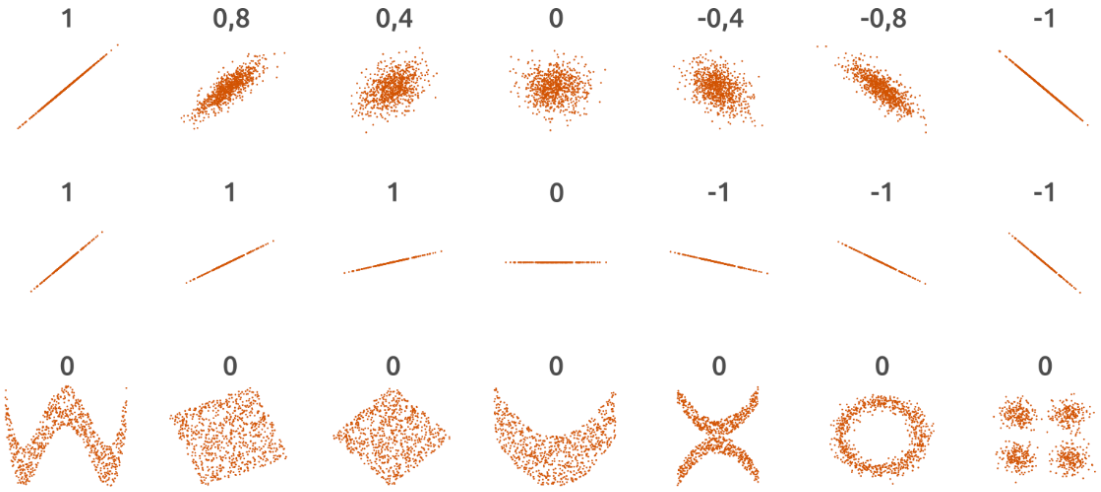
Calcul : R = covariance / produit des écart-types

$$R = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

Variables d'origine	x_i et y_i
Écart à la moyenne	$x_i - \bar{x}$ et $y_i - \bar{y}$
Produit des écarts	$(x_i - \bar{x})(y_i - \bar{y})$
Moyenne du produit des écarts (covariance)	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Propriétés du R de Pearson :

- Valeur standardisée : permet de comparer des relations entre caractères
- Variation entre :
  - -1 = corrélation négative : + en X et – en Y, ou inversement)
  - 1 = corrélation positive : + en X et + en Y).
- 0 signifie l'absence de relation statistique.



Source : Denis Boigelot, wikimedia.org

2. Bivarié quantitatif

Supplément : calcul de la covariance

Cov(x,y) = moyenne du produit des écarts à la moyenne de x et de y

$$R = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

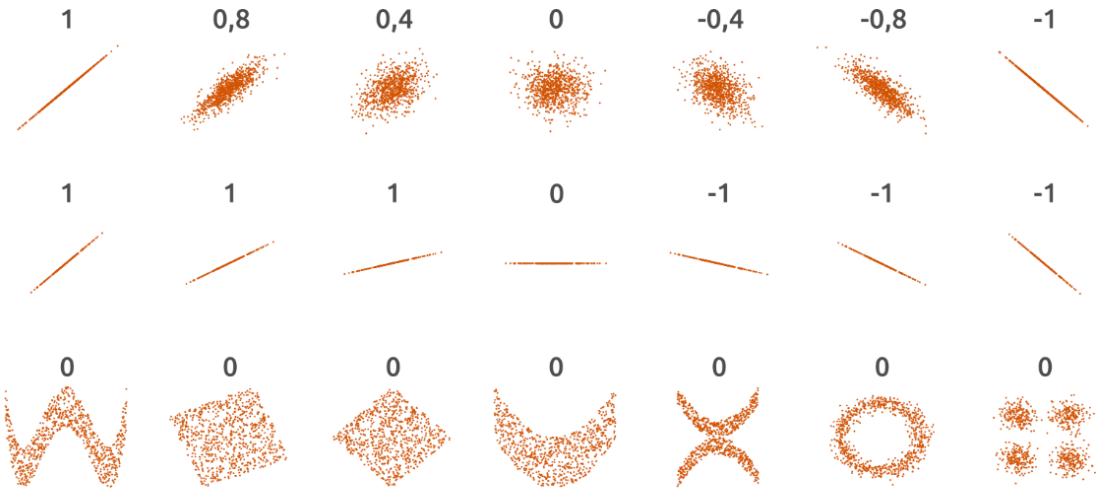
Variables d'origine	x_i et y_i
Écart à la moyenne	$x_i - \bar{x}$ et $y_i - \bar{y}$
Produit des écarts	$(x_i - \bar{x})(y_i - \bar{y})$
Moyenne du produit des écarts (covariance)	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Interprétation :

- Relation positive : covariance positive (- par – et + par +).
- Relation négative : covariance négative (- par + et + par -).

Propriétés de la covariance :

- Ne dépend pas de l'ordre de grandeur (produit)
- Ne dépend pas des effectifs (moyenne)
- Dépend de la dispersion autour de la moyenne (écarts) : on ne peut pas comparer deux résultats de covariance.



Source : Denis Boigelot, wikimedia.org

2. Bivarié quantitatif

Calcul dans R :

fonction *cor(x, y, methode, use)*

Variante : le coefficient de corrélation de Spearman

- Principe : covariation entre les rangs des individus
- Application : caractère quantitatif de stock (ex. populations)
- Dans R : *cor(method = "spearman")*

Opérations avancées dans R :

- Calcul d'une matrice des corrélations
- Générer un tableau
- Mettre en forme le tableau
- Exporter la matrice des corrélations

```
> # 3.2 Corrélations ====
> # ....................................
>
> # relation entre 2 variables
> cor(x = db$HISPA, y = db$POP20, method = "pearson", use = "all.obs")
[1] 0.9040224
> cor(x = db$HISPA, y = db$POP20, method = "spearman", use = "all.obs")
[1] 0.7400239
```

```
# relation entre plein de variables
## fonctions successives
cor(db[,9:23])
round(cor(db[,9:23]), digits = 1)
correl <- round(cor(db[,9:23]), digits = 1)
correl <- as.data.frame(round(cor(db[,9:23]), digits = 1))
## chainage
correl <- cor(db[,9:23]) %>%
  round(digits = 1) %>%
  as.data.frame() %>%
  mutate(CBSA = row.names(correl)) %>%
  select(CBSA, WHITE:OTHERis)
## Export
write_csv2(correl, "correlation.csv")
```

```
> round(cor(db[,9:23]), digits = 1)
        WHITE BLACK ASIAN HISPA OTHER WHITEp BLACKp ASIANp HISPAp OTHERp
WHITE     1.0   0.9   0.8   0.8   1.0   -0.2   0.1   0.5   0.1   0.0
BLACK     0.9   1.0   0.7   0.7   0.9   -0.2   0.2   0.3   0.1   0.0
ASIAN     0.8   0.7   1.0   0.9   0.9   -0.2   0.0   0.6   0.1   0.0
HISPA     0.8   0.7   0.9   1.0   0.8   -0.3   0.0   0.4   0.2   0.0
OTHER     1.0   0.9   0.9   0.8   1.0   -0.2   0.1   0.5   0.1   0.1
WHITEp   -0.2  -0.2  -0.2  -0.3  -0.2    1.0  -0.5  -0.3  -0.7  -0.1
BLACKp    0.1   0.2   0.0   0.0   0.1   -0.5   1.0   0.0  -0.2  -0.1
ASIANp    0.5   0.3   0.6   0.4   0.5   -0.3   0.0   1.0   0.2   0.0
HISPAp    0.1   0.1   0.1   0.2   0.1   -0.7  -0.2   0.2   1.0  -0.1
OTHERp    0.0   0.0   0.0   0.0   0.1   -0.1  -0.1   0.0  -0.1   1.0
```

2. Bivarié quantitatif

Approche explicative : la régression linéaire

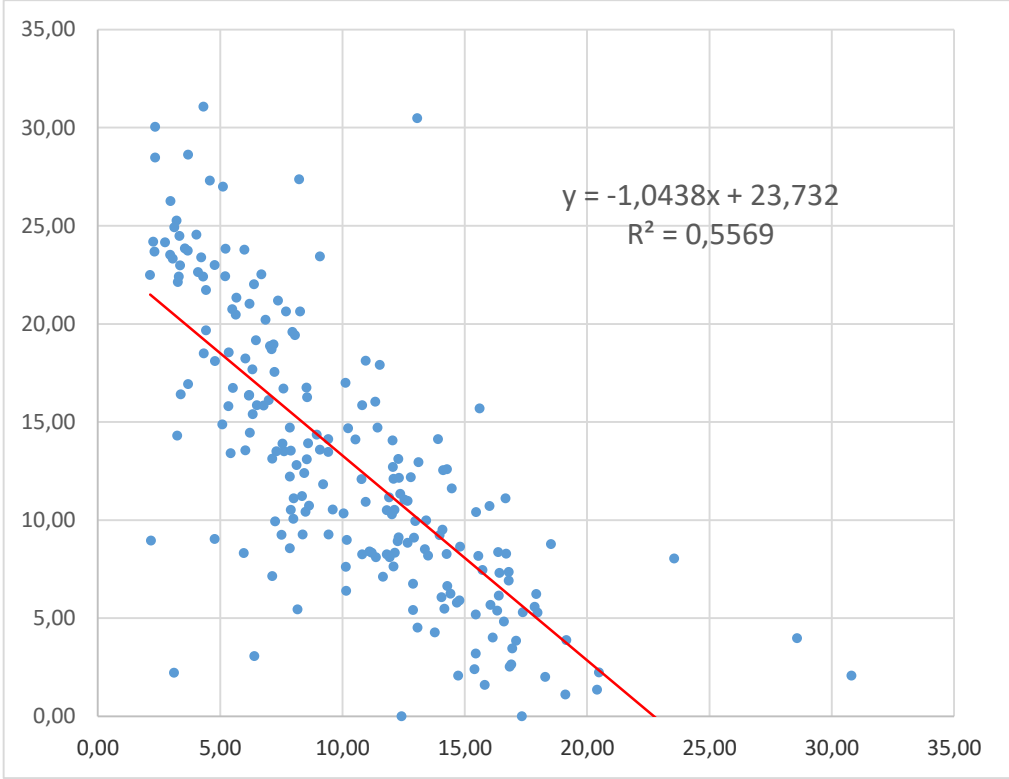
Principe : modéliser les valeurs d'un caractère Y à partir des valeurs d'un caractère X

- Equation : $Y = aX + b$
- Paramètres de l'équation :
  - Y = variable à expliquer
  - X = variable explicative
  - a = pente de la droite de régression
  - b = valeur de Y à l'origine, c'est-à-dire quand X = 0.

Coefficient de détermination (R^2) : permet de caractériser le pouvoir explicatif du modèle.

- Calcul : carré du coefficient de corrélation
- Interprétation : « la régression explique _% de la variation de Y

Relation entre le % de cadres et le % d'ouvriers dans l'unité urbaine de Grenoble



Source : Duroudier, 2022, Université Paris 1.

2. Bivarié quantitatif

Approche explicative : la régression linéaire

Application dans R :

lm(y ~ x)

Voir le modèle :

model1

Examiner l'objet créé :

str(model1)

```
> # 3.3 Modèle de régression linéaire ====
> # ........................................
>
> ## définition X et Y
> Y = db$HISPA
> X = db$POP20
>
> ## modele
> model1 <- lm(Y ~ X) # exemple de base
> model1 <- lm(data = db, HISPA ~ POP20) # variante dans un tableau
> model1 # appeler le modèle fournit l'équation

Call:
lm(formula = HISPA ~ POP20, data = db)

Coefficients:
(Intercept)        POP20
 -2.467e+04    2.673e-01
```

```
model1                List of  12
   $ coefficients : Named num [1:2] -2.47e+04 2.67e-01
    ..- attr(*, "names")= chr [1:2] "(Intercept)" "POP20"
   $ residuals    : Named num [1:909] 15088 12288 19170 16705 6587 ...
    ..- attr(*, "names")= chr [1:909] "1" "2" "3" "4" ...
   $ effects      : Named num [1:909] -1990005 9195786 18633 16121 6025 ...
    ..- attr(*, "names")= chr [1:909] "(Intercept)" "POP20" "" "" ...
   $ rank         : int 2
   $ fitted.values: Named num [1:909] -13368 -4455 22276 -14497 1908 ...
    ..- attr(*, "names")= chr [1:909] "1" "2" "3" "4" ...
   $ assign       : int [1:2] 0 1
   $ qr           :List of 5
    ..$ qr   : num [1:909, 1:2] -30.1496 0.0332 0.0332 0.0332 0.0332 ...
    .. ..- attr(*, "dimnames")=List of 2
    .. .. ..$ : chr [1:909] "1" "2" "3" "4" ...
    .. .. ..$ : chr [1:2] "(Intercept)" "POP20"
    .. ..- attr(*, "assign")= int [1:2] 0 1
    ..$ qraux: num [1:2] 1.03 1.01
    ..$ pivot: int [1:2] 1 2
    ..$ tol  : num 1e-07
    ..$ rank : int 2
    ..- attr(*, "class")= chr "qr"
   $ df.residual  : int 907
   $ xlevels      : Named list()
   $ call         : language lm(formula = HISPA ~ POP20, data = db)
   $ terms        :Classes 'terms', 'formula'  language HISPA ~ POP20
    .. ..- attr(*, "variables")= language list(HISPA, POP20)
```

2. Bivarié quantitatif

Approche explicative : la régression linéaire

Analyse des résultats du modèle :

summary(model1)

- Coef. de détermination R^2 : 0,82
- Significativité : très forte
- Estimate : pente et *intercept*

```
> summary(model1) # résultats généraux : equation, variabilité de

Call:
lm(formula = HISPA ~ POP20, data = db)

Residuals:
    Min      1Q  Median      3Q     Max
-999107   -3687   12936   18706 2357295

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.467e+04  4.996e+03  -4.938 9.37e-07 ***
POP20        2.673e-01  4.198e-03  63.689  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 144400 on 907 degrees of freedom
Multiple R-squared:  0.8173,    Adjusted R-squared:  0.8171
F-statistic:  4056 on 1 and 907 DF,  p-value: < 2.2e-16
```

Quelques tests supplémentaires :

- Analyse de variance des résidus
- Histogramme des résidus
- Test de Breusch-Pagan

```
> ## Jeter un oeil aux résidus
> anova(model1) # analyse de la variance des résultats (si hétéroscédasticité)
Analysis of Variance Table

Response: HISPA
           Df     Sum Sq    Mean Sq F value    Pr(>F)
POP20       1 8.4562e+13 8.4562e+13  4056.2 < 2.2e-16 ***
Residuals 907 1.8909e+13 2.0847e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> hist(model1$residuals) # histogramme rapide
> bptest(model1) # test de Breusch-Pagan sur la normalité des résidus

        studentized Breusch-Pagan test

data:  model1
BP = 288.19, df = 1, p-value < 2.2e-16
```

2. Bivarié quantitatif

Approche explicative : analyse des résidus de la régression linéaire

On appelle résidu l'information qui n'est pas expliquée par le modèle de régression.

➔ Intérêt majeur en géographie !

Résidus bruts :

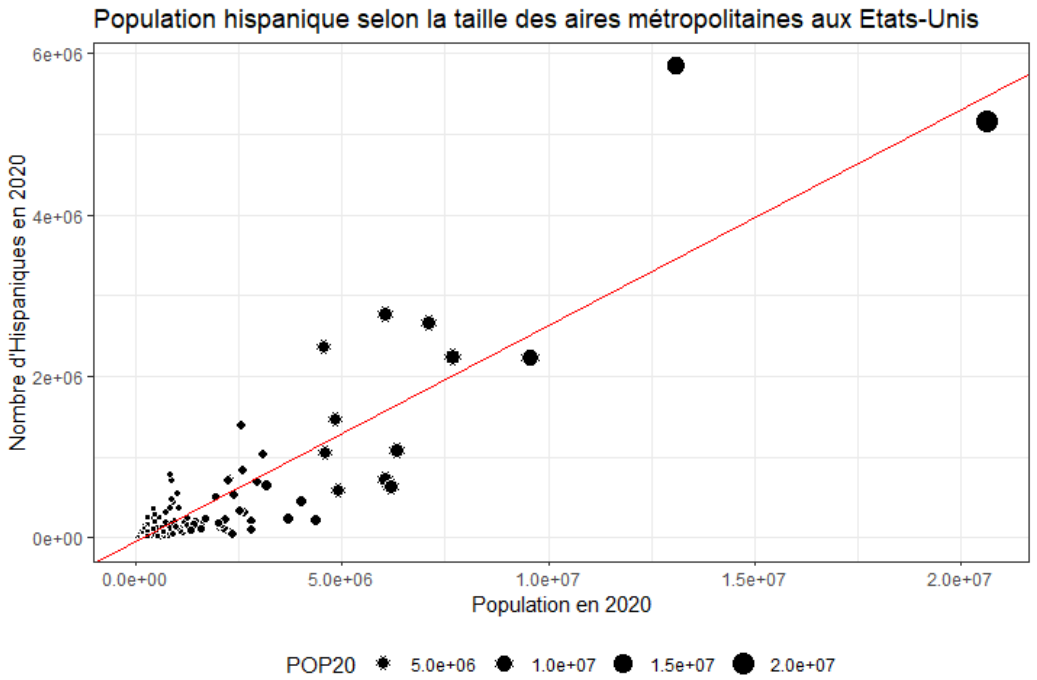
RB = Yobservé – Yestimé

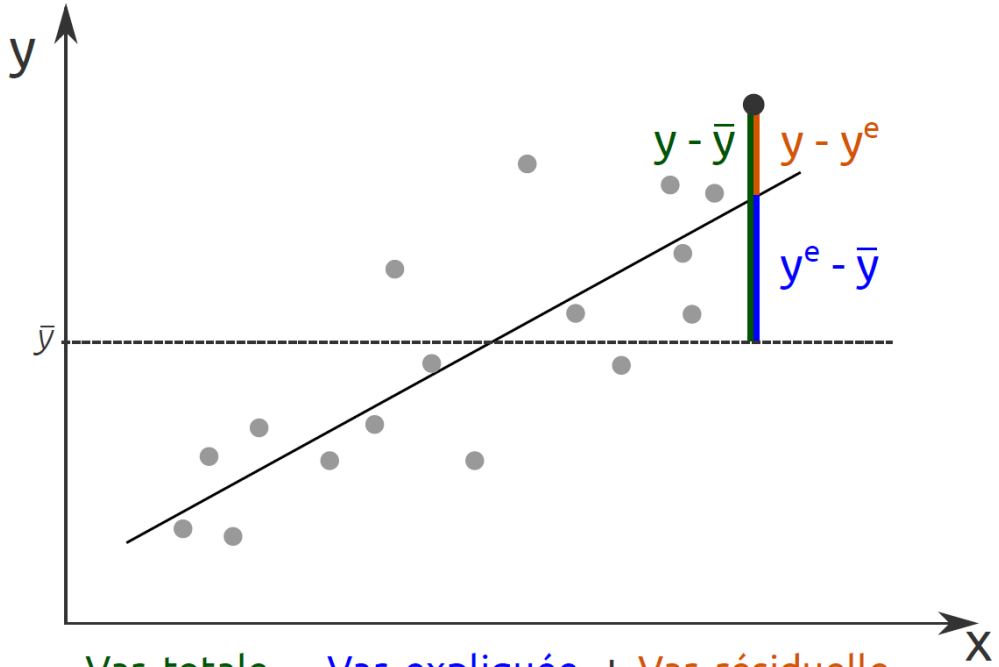
Propriétés : moyenne = 0, interprétation dans l'unité de la variable Y.

Résidus relatifs :

RR = RB / Yestimé

Propriétés : variation autour de 0, lissent l'effet de taille mais perdent l'ordre de grandeur.





2. Bivarié quantitatif

Approche explicative : analyse des résidus de la régression linéaire

Application dans R : extraire les résultats de lm()

- Valeurs estimées

  fitted(model1)

- Valeurs résiduelles

  resid(model1)

- Concaténation avec le tableau de données

  cbind(db, estimées, résidus)

- Calcul des résidus relatifs

- Visualiser les résultats :

  summary()

  ggplot() + geom_histogram()

```
> # 3.4 Analyse des résidus ====
> # ........................................
>
> # analyse des résidus
> ## exporter les résidus
> estim <- as.data.frame(fitted(model1))
> head(estim)
  fitted(model1)
1     -13367.779
2      -4454.920
3      22275.907
4     -14496.524
5       1907.704
6     163062.089
> res <- as.data.frame(resid(model1))
> head(res)
  resid(model1)
1     15087.779
2     12287.920
3     19170.093
4     16704.524
5      6587.296
6   -146351.089
> results <- cbind(db, estim, res)
> results <- results %>%
+   rename(estimate = `fitted(model1)`,
+          residus = `resid(model1)`)
> ## Calcul des résidus relatifs
> results <- results %>%
+   mutate(RR = (residus/estimate))
> # head(results)
> summary(results[,25:26])
    residus              RR
 Min.   :-999107   Min.   :-3776.909
 1st Qu.:  -3687   1st Qu.:   -1.313
 Median :  12936   Median :   -1.046
 Mean   :      0   Mean   :   -5.117
 3rd Qu.:  18706   3rd Qu.:   -0.442
 Max.   :2357295   Max.   :   26.990
>
> ## histogramme travaillé
> #### outils : ggplot2
> graph <- ggplot(results, aes(residus)) +
+   geom_histogram() +
+   scale_y_log10() +
+   theme_bw()
> graph
```

2. Bivarié quantitatif

Approche explicative : cartographie des résidus

Où sont les cas qui ne suivent pas le modèle ?
Est-ce qu'il existe une logique géographique ?

Quelques rappels pour l'application dans R :

- Package *tmap*
- Jointure des données à un fond de carte
- Représentation des résidus :

  tm_shape(tomap) +
  tm_polygons(col = "variable")

- Discrétisation :

  style = ""

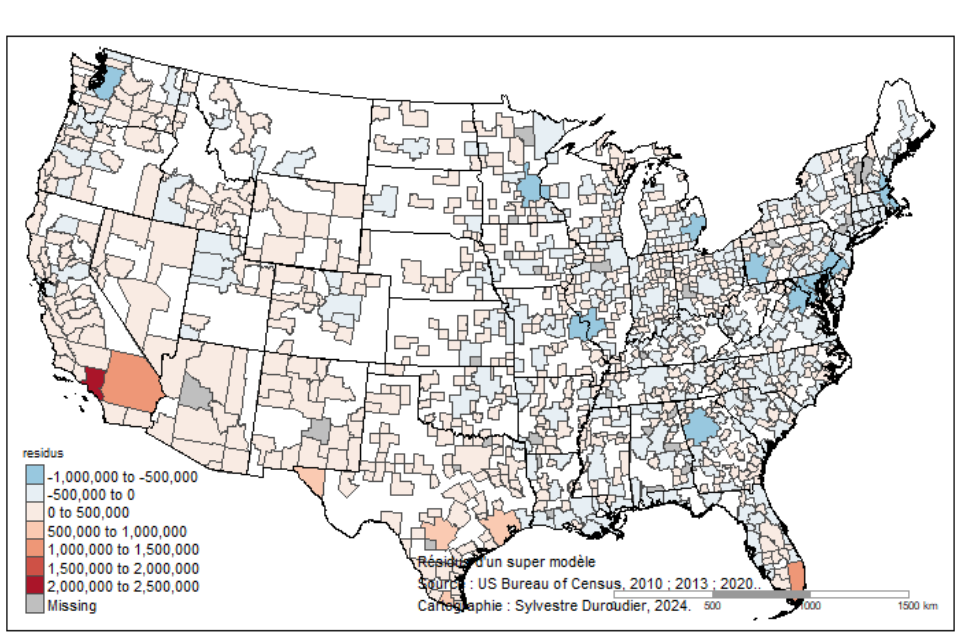
  ou *breaks = c()*

- Palette : double gamme inversée

```r
# 3.5 Cartographie des résidus ====
# ........................................

## Cartographie
#### outils : tmap, colorbrewer
tomap <- left_join(fdc_cities, results, by = c("GEOID" = "CBSAA"))

### carte de base
map <- tm_shape(fdc_states) +
  tm_polygons(col = "white", alpha = 0, border.col = "black", border.lwd = 1, border.alpha = 0) +
  tm_shape(tomap) +
  tm_polygons(col = "residus",
              style = "pretty",
              # breaks = c(-2, -1, -0.5, 0, 0.5, 1, 2, 3),
              palette = "-RdBu",
              alpha = 1,
              border.col = "grey30",
              border.lwd = 0.2,
              border.alpha = 1,
  ) +
  tm_shape(fdc_states) +
  tm_polygons(col = "white", alpha = 0, border.col = "black", border.lwd = 1, border.alpha = 1) +
  tm_legend(legend.position = c("left", "bottom"),
            legend.outside = T) +
  tm_credits(align = "left", size = 0.7, text = "Résidus d'un super modèle
Source : US Bureau of Census, 2010 ; 2013 ; 2020..
Cartographie : Sylvestre Duroudier, 2024.",
             position = c("left", "bottom")) +
  tm_layout(legend.title.size = 0.7,
  tm_scale_b
map
```



2. Bivarié quantitatif

Approche explicative : cartographie des résidus

- Cartographier selon la population des aires métropolitaines

tm_shape(tomap) +
tm_bubbles(col = "variable",
size = "variable de taille")

- Cartographier selon les résidus relatifs

```
### variante avec population des villes
map <- tm_shape(fdc_states) +
  tm_polygons(col = "white", alpha = 0, border.col = "black", border.lwd = 1, border.alpha = 1) +
  tm_shape(tomap) +
  tm_bubbles(size = "POP20",
             scale = 5,
             col = "residus",
             style = "pretty",
             # breaks = c(-2, -1, -0.5, 0, 0.5, 1, 2, 3),
             palette = "-RdBu",
             alpha = 1,
             border.col = "grey30",
             border.lwd = 0.2,
             border.alpha = 1,
  ) +
  tm_legend(legend.position = c("left", "bottom"),
            legend.outside = T) +
  tm_credits(align = "left", size = 0.7, text = "Résidus d'un super modèle
Source : US Bureau of Census, 2010 ; 2013 ; 2020..
Cartographie : Sylvestre Duroudier, 2024.",
             position = c("left", "bottom")) +
  tm_layout(legend.title.size = 0.7,
            legend.text.size = 0.7) +
  tm_scale_bar(color.dark = "gray60", position = c("right", "bottom"))
map
```