Workshop - Analyser la politique marocaine par les datas

Romain Ferrali

11-12 juillet 2018

TAFRA

1 Introduction: la science des données

1.1 Les statistiques, c'est quoi ?

Aujourd'hui, nous disposons de plus en plus en données sur toutes les dimensions de la vie sociale au Maroc. Celles-ci permettent de mieux répondre à nos questions, mais posent aussi de nouveaux défis : ces nouveaux types de données, disponibles en quantités plus importantes requièrent de nouvelles compétences d'analyse et de visualisation. Une science s'attache à "faire parler les données" : les statistiques. Cette journée d'étude en donne une brève introduction. Commençons par une définition.

> La statistique est l'étude d'un phénomène par la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. (Wikipedia)

1.2 Se mettre en jambes

La démarche statistique part d'une question que l'on pose aux données. En d'autres termes, il est bien plus simple de faire parler les données quand on sait ce qu'on veut leur demander. Lors de cette formation, nous allons nous poser la question suivante : *lors des élections communales de 2015, quand et pourquoi le PJD obient-t-il de meilleurs scores que le PAM ?*

Le statisticien analyse des données quantitatives. En d'autres termes, il analyse des données qui prennent une forme numérique. Aujourd'hui, nous allons utiliser deux jeux de données, appelés *datasets* en anglais :

- Résultats des élections communales de 2015, aggrégées au niveau commune.
- Données du Recensement Général de la Population et de l'Habitat (RGPH) de 2014, aggrégées au niveau commune.

Ces datasets ont la forme d'un tableau, où chaque ligne représente une *observation*, et chaque colonne une *variable*, une propriété de cette observation. L'ensemble des activités de cette journée d'étude peuvent être réalisées avec un simple logiciel de tableur, comme Excel ou Google Sheets. La démonstration sera effectuée avec R, un langage de programmation dédié aux statistiques, pour nous exposer à des outils d'analyse plus sophistiqués.

L'une ses tâches est bien souvent de transformer en indicateurs quantitatifs des concepts d'abord compris sous l'angle qualitatif. Quelques exemples de données quantitatives :

- **Données numériques** continues (ex. taux de participation) ou discrètes (ex. nombre de suffrages exprimés pour le PJD). Ce sont les données par excellence des statistiques.
- **Données catégorielles** cardinales (ex. statut marital) ou ordinales (ex. niveau d'éducation). Ces données sont généralement "converties" en données numériques en assignant un chiffre à chaque

catégorie, par exemple en assignant les nombres 1 à 3 aux réponses “éducation primaire”, “secondaire” et “supérieure”.

Activité. Ouvrir les deux datasets avec son logiciel de tableur. Donner d’autres exemples de variables numériques continues et discrètes, ainsi que de variables catégorielles cardinales et ordinales. Lesquelles de ces variables sont un aggrégat continu de variables catégorielles ?

Aller plus loin : de nouveaux types de données. Aujour’hui, les statisticiens sont de plus en plus confrontés à de nouveaux types de données, comme les réseaux (ex. le réseau de gouvernance des entreprises côtées à la bourse de Casablanca) ou le texte (ex. les questions posées par les parlementaires à la chambre des Représentants). Comme les données catégorielles, ces nouveaux types de données sont aussi convertis en donées numériques. Ces données requièrent des techniques d’analyse plus sophistiquées que nous ne couvrirons pas lors de cette formation.

1.3 Au programme

Les statistiques se divisent en deux grandes familles, que nous allons aborder tour à tour. Pour chacune de ces familles, nous aborderons les concepts mathématiques fondqmentaux et les techniques de visualisation les plus appropriées.

La première de ces familles est la statistique dite *descriptive,* dont l’objectif est de décrire les données dont on dispose. L’on abordera des concepts assez connus comme la moyenne, l’écart-type ou la médiane.

L’autre famille des statistiques, beaucoup plus vaste, s’appelle la statistique *inférentielle*. Son objectif est d’apprendre, à partir des données de *l’échantillon* dont on dispose, des caractéristiques de la population dont elles sont extraites. Nous verrons comment parler d’une ou de deux variables, en prenant bien soin de distinguer corrélation et causalité.

2 Statistiques descriptives : qui a gagné les élections ?

Les statisticiens manipulent beaucoup de données. L’ensemble des données qui décrivent la même chose est appelé une *série* statistique. Les statistiques descriptives permettent de décrire une série. Ici, nous travaillerons avec trois séries statistiques issues des résultats des élections communales :

- Nombre de suffrages exprimés pour le PAM
- Nombre de suffrages exprimés pour le PJD
- Nombre de bulletins valides

Activité. Construire les trois séries.

```
# chargement des packages utiles pour le workshop
library(tidyverse)
library(readxl)
library(corrgram)
# lecture du fichier
elections <- read_excel("elections.xlsx")
elections <- elections %>%
  mutate_all(funs(replace_na(., 0))) %>%
  mutate(nValide = rowSums(select(., AGD:USFP)))
```

2.1 En un chiffre: le PJD a battu le PAM

Activité. Obtenir le pourcentage de suffrages valides obtenus par le PAM et le PJD.

```
# Le score du PAM
sum(elections$PAM) / sum(elections$nValide)
```

```
## [1] 0.1811295
```

```r
# Le score du PJD
sum(elections$PJD) / sum(elections$nValide)
```

```
## [1] 0.2115978
```

2.2 En beaucoup de chiffres: le PAM a battu le PJD ?

2.2.1 Visualiser une distribution

L'analyse précédente prétendait que le Maroc ne consistait que d'une seule grande commune. Nous allons maintenant examiner les scores du PAM et du PJD dans chaque commune. Pour ce faire, nous allons avoir besoin de créer deux nouvelles séries statistiques : les scores du PAM et du PJD.

> **Activité.** Construire deux nouvelles séries statistiques : le score du PAM et du PJD dans chaque commune.

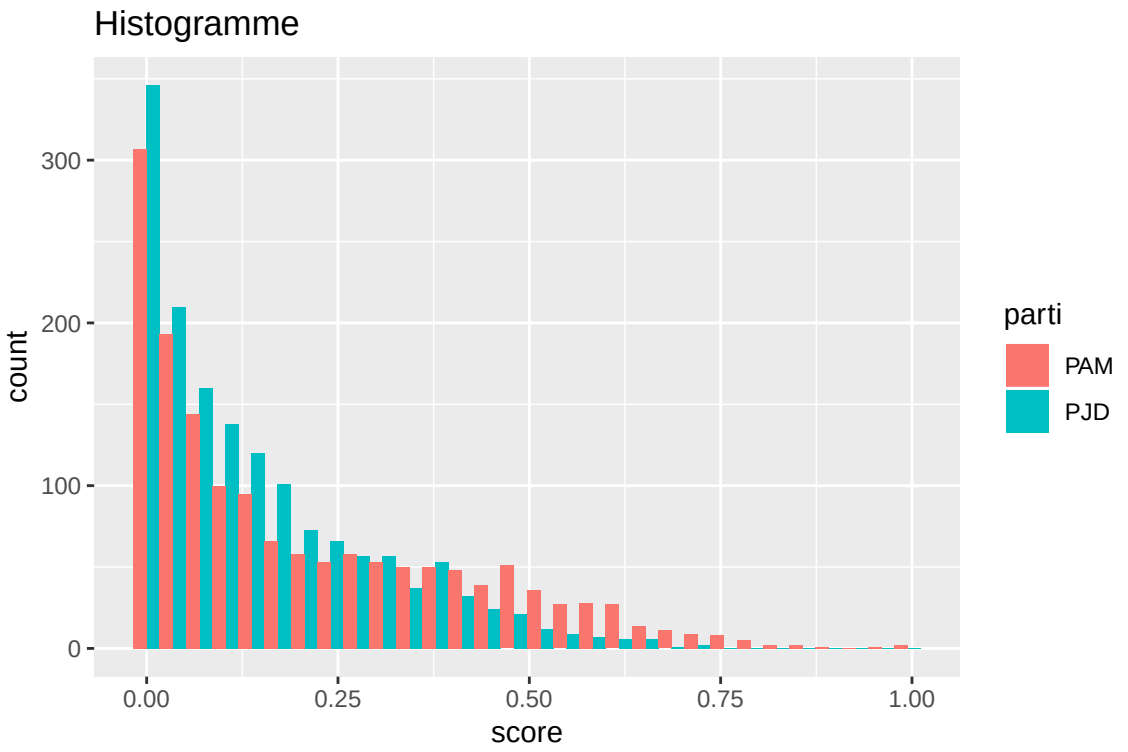
```r
elections <- elections %>%
  mutate(
    pctPJD = PJD / nValide,
    pctPAM = PAM / nValide
  )
```

Les statistiques descriptives fournissent les outils pour pouvoir parler des 1538 communes du Maroc dans leur ensemble. Le concept le plus important est celui de *distribution*. Intuitivement, une distribution décrit comment les nombres qui composent une série sont agencés. L'outil le plus commun pour visualiser une distribution est l'histogramme. Ce type de graphe représente la distribution en utilisant des rectangles verticaux pour signifier combien de nos observations sont comprises entre telle et telle valeur.

> **Activité.** Construire un histogramme des scores du PAM et du PJD. Lequel de ces deux partis réalise les meilleurs scores ?

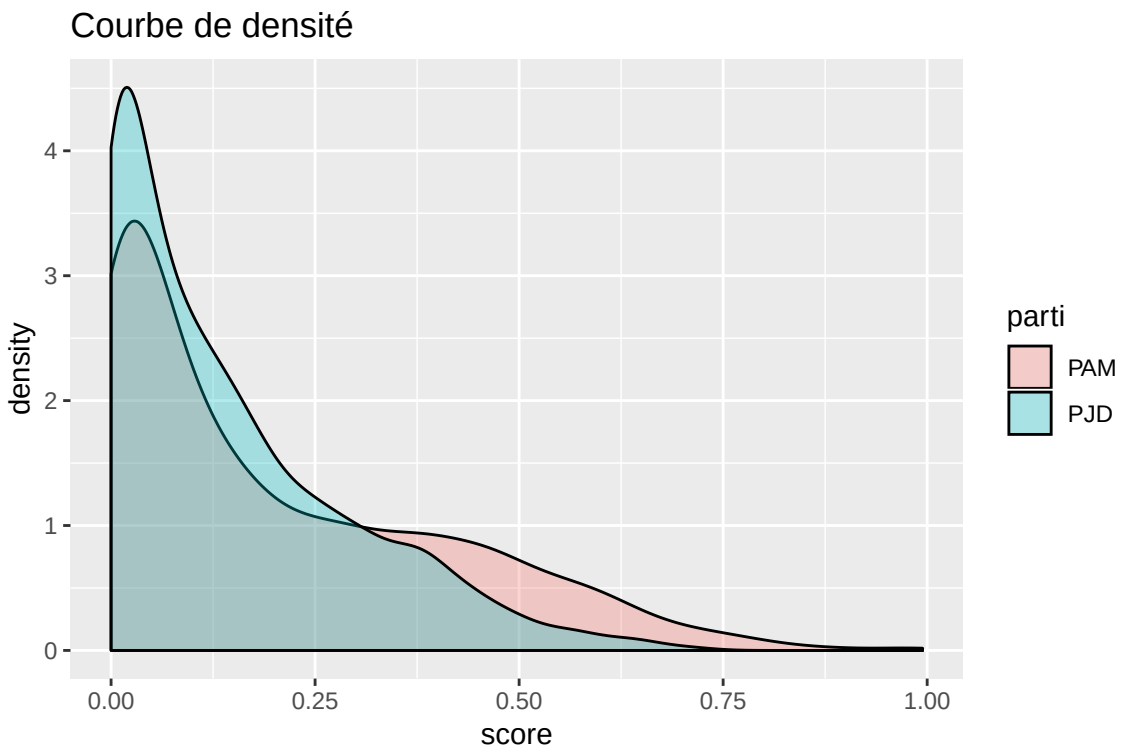
```r
plot1 <- elections %>%
  select(PJD = pctPJD, PAM = pctPAM) %>%
  gather(parti, score)

ggplot(plot1, aes(score, fill = parti)) +
  geom_histogram(position = "dodge") +
  labs(title = "Histogramme")
```



Aller plus loin : la courbe de densité. Pour construire un histogramme, on doit décider du nombre de rectangles que l'on va utiliser. Cette décision peut communiquer des intuitions erronnées, si l'on choisit trop ou trop peu de rectangles. La méthode de l'estimation par noyau, aussi appelée courbe de densité, apporte un lissage à l'histogramme.

```
ggplot(plot1, aes(score, fill = parti)) +
  geom_density(alpha = .3) +
  labs(title = "Courbe de densité")
```



2.2.2 Décrire une distribution

Des concepts statistiques permettent de transformer l'intuition visuelle en quantités numériques. Ces "statistiques" résument des propriétés fondamentales de la série en un seul chiffre.

Parmi les statistiques les plus importantes, la *moyenne* est la mesure la plus connue pour décrire la tendance centrale d'une série. Intuitivement, la moyenne représente la valeur type de la série. Mathématiquement, pour une série contenant les valeurs $x_1, ..., x_n$, la moyenne μ est

$$\mu = \frac{x_1 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

L'écart-type mesure l'écart moyen à la moyenne. L'écart-type indique combien l'observation typique est éloignée de la moyenne. Formellement, l'écart-type σ est la racine carrée de la variance σ^2 d'une série. En d'autres termes, $\sigma = \sqrt{\sigma^2}$, avec

$$\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2$$

Comme la moyenne, la *médiane* est une mesure de la tendance centrale d'une distribution. La médiane est la valeur qui divise la série statistique en deux groupes de taille égale : 50% des observations de la série ont une valeur inférieure à la médiane, et 50% ont une valeur supérieure à la médiane. La moyenne a l'avantage d'être un concept bien connu. Elle a aussi l'inconvénient d'être très sensible aux valeurs extrêmes. Par exemple, les revenus dans un pays sont souvent très inégaux, avec un petit nombre de riches individus gagnant parfois 1000 fois plus que les individus les plus pauvres. Le revenu moyen ne représente pas bien ce que gagne tout un chacun dans un pays, car il est gonflé par les revenus des millionaires. Le revenu médian capture mieux cette notion, car il est moins sensible à ces valeurs extrêmes. Formellement, soit $F(x)$ une fonction qui retourne le

pourcentage d'observations dans la série qui ont une valeur inférieure à x. La médiane est la valeur m telle que $F(m) = 0,5$.

Enfin, les *quantiles* généralisent la notion de médiane. Ils permettent d'obtenir plus de détails sur la distribution.

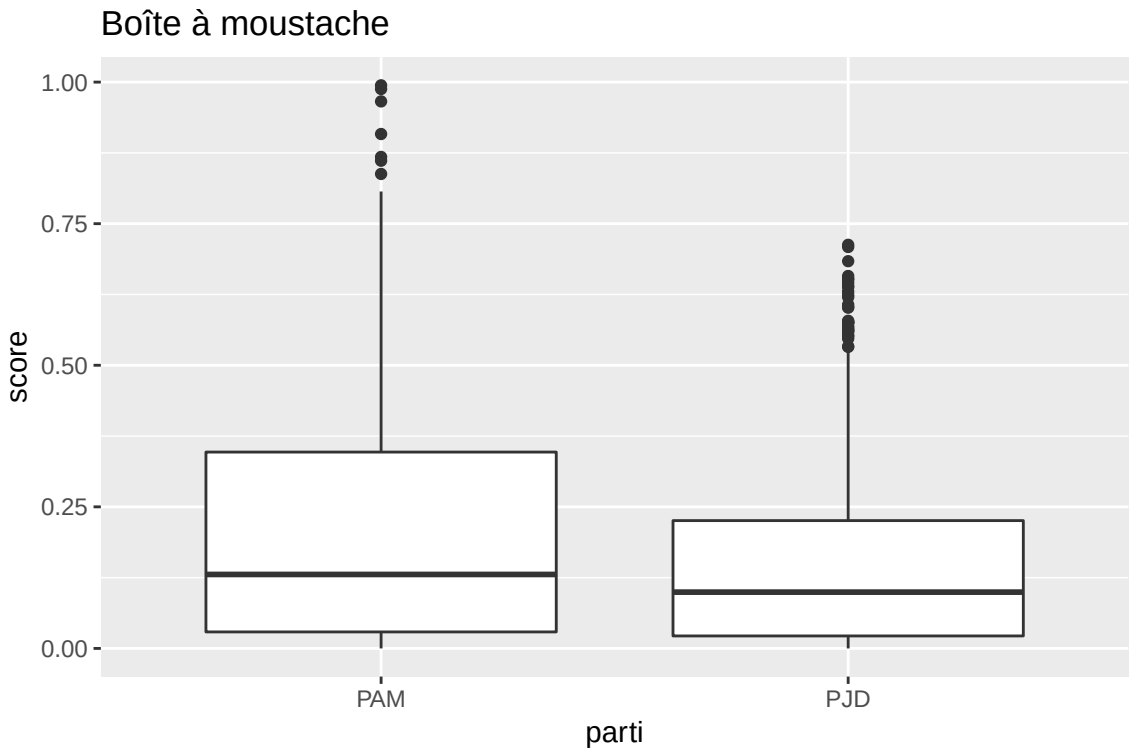
```
# PAM
c(
  "moyenne" = mean(elections$pctPAM),
  "ecart-type" = sd(elections$pctPAM),
  "10e pcentile" = quantile(elections$pctPAM, .1),
  "1er quartile" = quantile(elections$pctPAM, .25),
  "mediane" = median(elections$pctPAM),
  "3e quartile" = quantile(elections$pctPAM, .75),
  "90e pcentile" = quantile(elections$pctPAM, .9)
)
```

```
##          moyenne       ecart-type 10e pcentile.10% 1er quartile.25%
##       0.20277792       0.20566187       0.00000000       0.02913421
##          mediane  3e quartile.75% 90e pcentile.90%
##       0.13062099       0.34672418       0.51756984
```

```
# PJD
c(
  "moyenne" = mean(elections$pctPJD),
  "ecart-type" = sd(elections$pctPJD),
  "10e pcentile" = quantile(elections$pctPJD, .1),
  "1er quartile" = quantile(elections$pctPJD, .25),
  "mediane" = median(elections$pctPJD),
  "3e quartile" = quantile(elections$pctPJD, .75),
  "90e pcentile" = quantile(elections$pctPJD, .9)
)
```

```
##          moyenne       ecart-type 10e pcentile.10% 1er quartile.25%
##       0.14412552       0.14749323       0.00000000       0.02206033
##          mediane  3e quartile.75% 90e pcentile.90%
##       0.09941004       0.22569903       0.37186215
```

```
ggplot(plot1, aes(x = parti, y = score)) +
  geom_boxplot() +
  labs(title = "Boîte à moustache")
```



2.2.3 Pour finir : compétition politique

L'analyse précédente n'oppose pas directement le PAM au PJD. De deux choses l'une :

- soit PAM et PJD sont des *substituts* : quand le PJD réalise un score élevé, le PAM réalise un score faible, et vice versa ;
- soit PAM et PJD sont des *compléments* : quand le PJD réalise un score élevé, le PAM réalise aussi un score élevé.

Pour mieux répondre à notre interrogation de départ (le PAM a-t-il battu le PJD ?), nous allons examiner une nouvelle série : la marge de victoire du PJD sur le PAM. Celle-ci s'écrit :

$$\text{marge PJD} = \text{score PJD} - \text{score PAM}.$$

Activité. Construire la marge de victoire du PJD sur le PAM, puis calculer la marge de victoire moyenne, médiane, ainsi que les 1er et 3e quartiles. Construire l'histogramme de cette marge de victoire. Quel est le pourcentage de communes où le PJD a battu le PAM ?

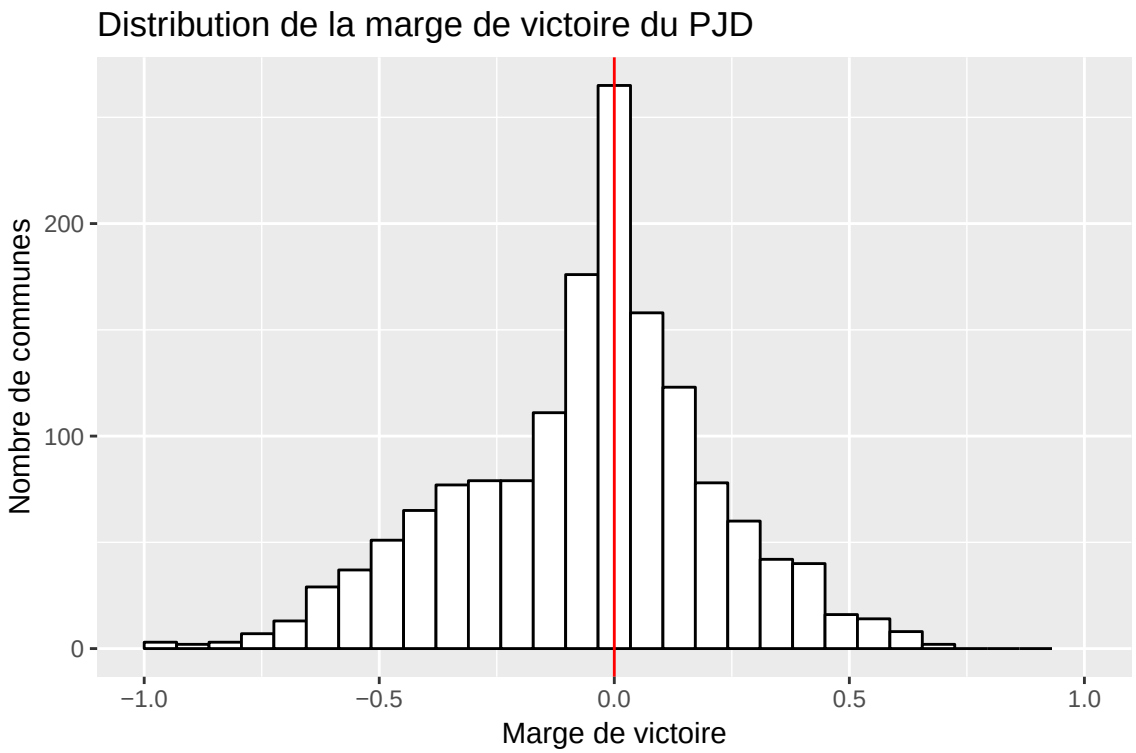
```
# construction de la serie
elections <- elections %>%
  mutate(diff = pctPJD - pctPAM)

# statistiques d'interet
summary(elections$diff)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.99436 -0.22728 -0.01834 -0.05865  0.10266  0.71272
```

```
# histogramme
ggplot(elections, aes(diff)) +
  geom_histogram(color = "black", fill = "white") +
```

```
  geom_vline(xintercept = 0, color = "red") +
  xlim(-1,1) +
  labs(title = "Distribution de la marge de victoire du PJD",
       x = "Marge de victoire", y = "Nombre de communes")
```




```
# pourcentage de communes ou le PJD a battu le PAM
mean(elections$diff > 0)
```

```
## [1] 0.4375813
```

3 Statistiques inférentielles : où est-ce que le PJD gagne ?

3.1 Les notions de base : population, échantillon, incertitude

Note. Cette section est avant tout explicative, et ne demande fera pas l'objet de beaucoup d'applications sur Excel ou Google Sheets.

La statistique inférentielle se pose le problème suivant : il existe une *population* dont le statisticien n'observe qu'un *échantillon*. Le statisticien cherche à connaître, en observant seulement l'échantillon, une quantité relative à la population. Cette quantité est appellée un *paramètre*. L'acte de trouver la valeur du paramètre s'appelle *l'inférence*.

Les statistiques inférentielles sont avantageuses pour deux raisons. D'abord, elles amènent une nouvelle notion : *l'incertitude*. Le concept d'incertitude permet de quantifier à quel point le statisticien mesure précisément ce qu'il souhaite mesurer. Ensuite, elles permettent de considérer des paramètres plus compliqués, généralement considérés conjointement dans ce qu'on appelle un *modèle* statistique. Nous verrons plus loin l'un de ces modèles, la *régression linéaire*.

Pour appréhender l'incertitude, le statisticien marie les données observées (c'est-à-dire l'échantillon) avec une idée de ce qu'il aurait pu observer (c'est-à-dire les échantillons qu'il aurait pu obtenir), en utilisant la *théorie des probabilités*.

Remarque : faire de la statistique inférentielle quand on travaille sur la population. Nous travaillons sur la population des communes du Maroc, et non sur un échantillon de ces communes. L'on peut tout de même faire de la statistique inférentielle dans ce cadre, en considérant que les communes observées font partie d'une population plus vaste, celle des communes potentielles. Pour donner un exemple plus parlant, on peut considérer un élève qui passe des examens. Cet élève a passé 2 examens et a une moyenne de 9.8/20. Même si l'on dispose de la population des examens passés par cet élève, il est intéressant de considérer que ces examens représentent un échantillon des examens qu'aurait pû passer cet élève. Mesurer l'incertitude permettra de décider si l'élève mérite réellement d'échouer la matière.

3.1.1 Quelques notions de théorie des probabilités

La théorie des probabilités considère ce que l'on appelle des *variables aléatoires*. Considérons, par exemple, le lancer d'un dé. On peut représenter par X le résultat d'un lancer. On dit que X est une *variable aléatoire*, et ses valeurs possibles sont $1, 2, 3, 4, 5, 6$. Notre dé est non-biaisé ; de ce fait, il tombe sur chaque face avec une probabilité de $\frac{1}{6}$. On note

$$\Pr(X = 1) = \Pr(X = ...) = \Pr(X = 6) = \frac{1}{6}.$$

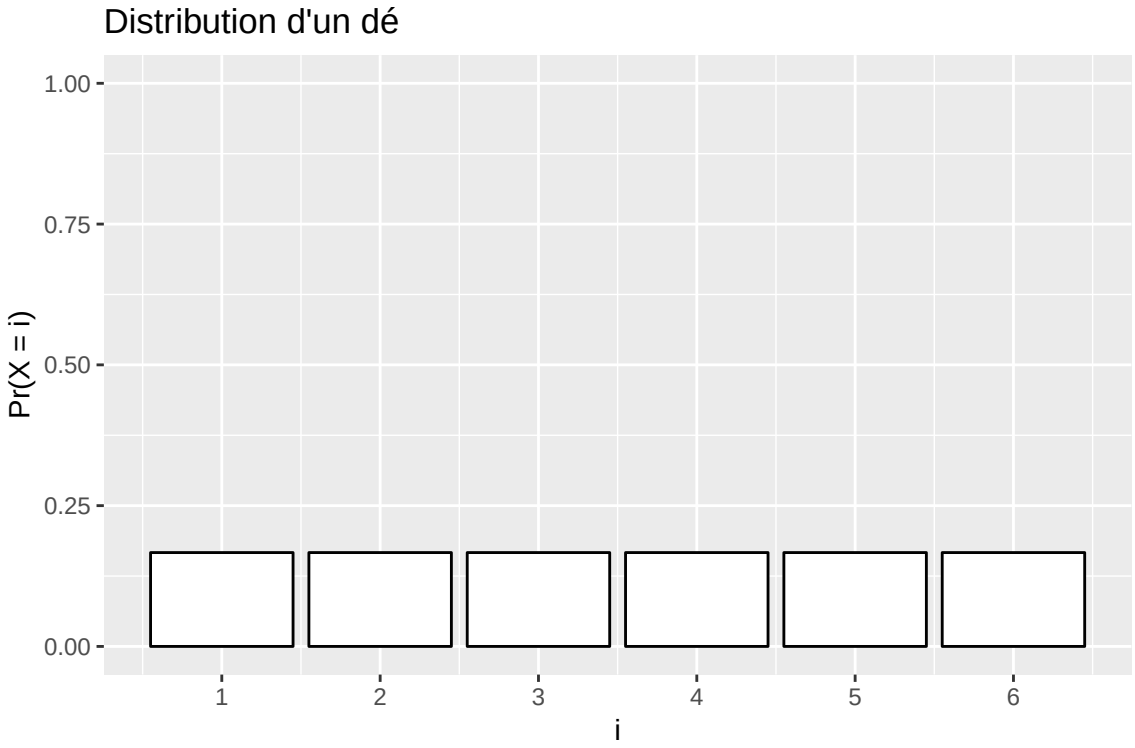
Les concepts que nous avons abordés plus tôt en statistique descriptive ont des analogues en théorie des probabilités. En particulier, l'analogue de la moyenne se nomme l'espérance. C'est la valeur "espérée" ou "moyenne" de notre variable aléatoire. On la calcule en faisant la somme de toutes les valeurs prises par notre variable aléatoire, pondérée par la probabilité d'obtenir cette valeur. Pour notre dé, l'espérance se note

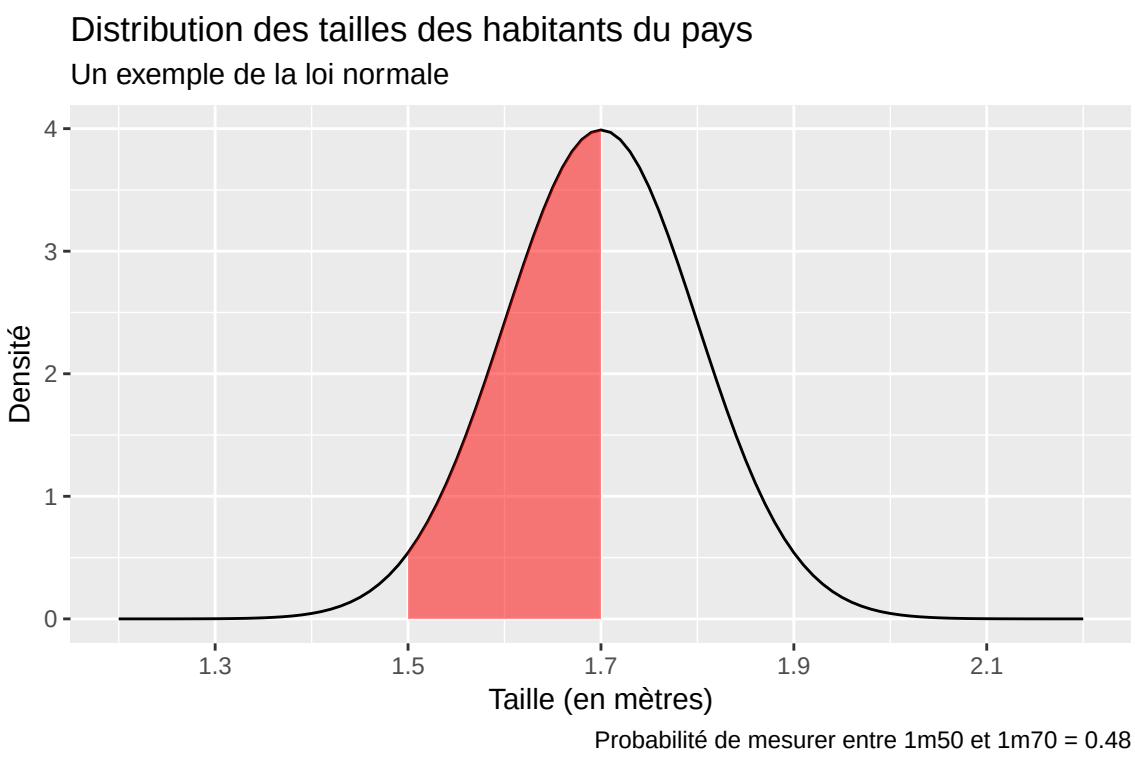
$$\mathbb{E}(X) = 1 \times \Pr(X = 1) + ... + 6 \times \Pr(X = 6) = \sum_{i=1}^{6} i \Pr(X = i) = 3.5$$

Les concepts de variance (généralement notée $\mathbb{V}(X)$) et d'écart-type ont eux aussi des analogues.

Enfin, le concept de distribution, que l'on a vu en statistique desscriptive, a lui aussi un analogue en théorie des probabilités, mais l'analogie est un peu plus compliquée. Elle varie selon que l'on considère une variable aléatoire discrète, comme notre dé, ou une variable aléatoire continue, comme la taille d'un individu choisi au hasard parmi les habitants du pays.

- *Variable discrète* : la distribution est simplement la probabilité d'obtenir l'une des valeurs de la distribution.
- *Variable continue* : la distribution est représentée par une courbe, comme la courbe de densité que nous avons vu plus haut. La probabilité d'obtenir un tirage compris entre deux valeurs est l'aire de la portion de la courbe délimitée par ces deux valeurs.





Activité. Reproduire la figure de la distribution des tailles en utilisant un histogramme et en réalisant 1000 tirages de la distribution normale avec moyenne 1.70 et écart-type 0.1. Ces tirages peuvent être réalisés en utilisant le site https://www.random.org/gaussian-distributions. Calculer la probabilité estimée d'avoir une taille entre 1m50 et 1m70.

Deux théorèmes de théorie des probabilités s'ajoutent aux concepts ci-dessus pour permettre de quantifier l'incertitude. Le premier s'appelle la loi des grands nombres. Ce théorème considère le problème suivant : le statisticien réalise un certain nombre de tirages d'une variable aléatoire, et calcule la moyenne de ces tirages. Le théorème dit que plus le nombre de tirages est important, plus la moyenne calculée par le statisticien sera, généralement, proche de l'espérance de cette variable aléatoire.

Le deuxième s'appelle le théorème central limite. Ce théorème précise la loi des grands nombre. En effet, il spécifie combien la moyenne des tirages sera proche de l'espérance de la variable aléatoire, en fonction du nombre de tirages.

Ces théorèmes ont une implication intuitive très simple : la loi des grands nombres nous dit que plus la taille de l'échantillon est importante, plus le paramètre est estimé avec précision. Le théorème central limite permet de quantifier exactement combien notre estimation est proche de la valeur réelle du paramètre, ce qui permet de mesurer l'incertitude.

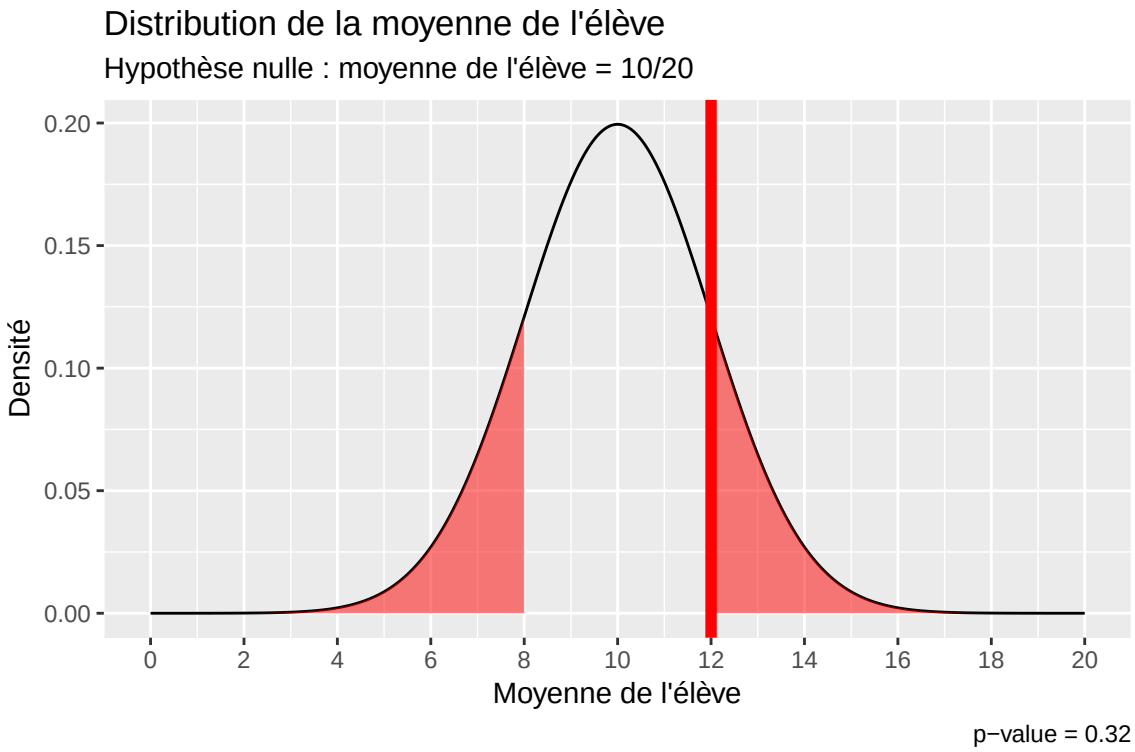
> **Attention !** Pour que la théorie fonctionne, il faut que le statisticien dispose d'un échantillon représentatif. La manière la plus simple de construire un échantillon représentatif est de sélectionner au hasard des membres de la population (échantillon *aléatoire simple*). Le théorème central limite nous dit qu'en général, un échantillon aléatoire simple de *1000 personnes* est suffisamment représentatif de la population. L'on peut aussi s'assurer que la distribution d'un certain nombre de charactéristiques de notre échantillon correspond à la population réelle (échantillon *stratifié*). Par exemple, pour construire un échantillon stratifié par sexe et statut urbain/rural, on s'assure que le pourcentage d'hommes ruraux, femmes rurales, hommes urbains, femmes urbains dans l'échantillon correspond à la population réelle, puis l'on choisit les représentants de ces quatres catégories au hasard. La stratification permet de s'assurer que l'échantillon sera parfaitement représentatif, même pour un échantillon de petite taille (moins de 300 personnes). Enfin, un échantillon peut *surreprésenter* certaines tranches de la population (par exemple, au Maroc, les personnes qui parlent français à la maison). La surreprésentation assure que l'échantillon contienne un nombre suffisant de ces tranches de la population pour pouvoir les étudier séparément. Pour que l'échantillon final soit représentatif de la population, on utilise alors une *pondération*. Par exemple, si l'on échantillone deux fois plus de francophones qu'il y en a au Maroc, on leur donnera deux fois moins de poids dans l'échantillon final.

3.1.2 Deux manières de représenter l'incertitude : intervalle de confiance et p-value

L'*intervalle de confiance* donne un intervalle - c'est-à-dire, une série de deux valeurs - qui représente l'incertitude avec laquelle le statisticien a mesuré une valeur. Plus l'intervalle est large, moins la quantité est estimée avec précision, et donc moins le statisticien a "confiance" en sa prédiction. Définir un intervalle de confiance nécessite de définir un *seuil de confiance*, noté en pourçentage. Une convention courante en sciences sociales est d'utiliser un seuil de 95%. L'intervalle de confiance a une interprétation simple mais (légèrement) erronnée : "je suis sûr à 95% que la vraie valeur est comprise entre x et y". L'interprétation correcte est la suivante : "si je construisais plein d'intervalles de confiance avec un échantillon de la même taille que celui dont je dispose, alors la vraie valeur serait comprise dans 95% de ces intervalles."

Une autre manière d'évaluer l'incertitude consiste à tester une hypothèse sur le paramètre, généralement appellée l'*hypothèse nulle*. L'hypothèse nulle fixe la valeur du paramètre. Un exemple d'hypothèse nulle : le paramètre "score moyen de l'élève" est de 10/20. L'on utilise ensuite nos deux théorèmes (loi des grands nombres, théorème central limite) pour dériver la distribution de ce paramètre sous l'hypothèse nulle. L'on regarde ensuite combien le paramètre estimé est typique si l'hypothèse nulle est vraie. Le statisticien considère généralement une quantité appellée *p-value*, qui correspond à la probabilité d'obtenir un paramètre plus extrême que le paramètre estimé sous l'hypothèse nulle. Plus la p-value est faible, plus le paramètre estimé est atypique, et moins il est probable que l'hypothèse nulle soit vraie.

La figure suivante représente la distribution de la moyenne d'un élève, sous l'hypothèse nulle qu'il a une moyenne de 10/20. La moyenne observée est de 12/20 (barre rouge). L'aire en rouge clair correspond à la p-value.



Enfin, il existe une relation entre p-value et nivau de confiance. Comme pour la définition d'un intervalle de confiance, le statisticien définit a priori un seuil de confiance α. Il rejette l'hypothèse nulle avec un degré de confiance α si la p-value est inférieure à $1-\alpha$. Sinon, on dit que le statisticien "ne peut pas rejeter l'hypothèse nulle."

> **Activité.** Considérer la figure ci-dessus. Peut-on rejeter avec un seuil de confiance de 95% l'hypothèse nulle que le niveau de l'élève est de 10/20 ?

3.1.3 L'application politique : le score du PJD est-il significativement différent de celui du PAM ?

Pour mettre en pratique les concepts vus plus haut, nous allons interpréter les résultats de deux tests statistiques. Le premier test considère le score moyen du PJD. Il évalue l'hypothèse nulle que ce score est de 50%. Si l'on peut rejeter l'hypothèse nulle, on dit que "le score moyen du PJD est *significativement* différent de 50%".

> **Activité.** Interpréter les résultats du test statistique suivant. Quel est l'intervalle de confiance autour du score moyen du PJD ? Quelle est la p-value attachée au test de l'hypothèse nulle "le score du PJD est de 50%" ? Peut-on rejeter l'hypothèse nulle avec un seuil de confiance de 95% ?

```
# test 1
t.test(elections$pctPJD, mu = .5)
```

```
##
##  One Sample t-test
##
## data:  elections$pctPJD
## t = -94.624, df = 1537, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.1367484 0.1515026
```

```
## sample estimates:
## mean of x
## 0.1441255
```

Le deuxième test compare le score moyen du PJD au score moyen du PAM. Il évalue l'hypothèse nulle "le score moyen du PJD est égal au score moyen du PAM." Notons que cette hypothèse est équivalente à l'hypothèse suivante : "la différence entre le score moyen du PJD et celui du PAM est de 0". Si l'on peut rejeter l'hypothèse nulle, on dit que "le score moyen du PJD est *significativement* différent de celui du PAM".

> **Activité.** Interpréter les résultats du test statistique suivant. A combien s'élève la différence entre le score moyen du PJD et celui du PAM ? Quel est l'intervalle de confiance autour de cette différence ? Quelle est la p-value attachée au test de l'hypothèse nulle "la différence entre le score moyen du PJD et celui du PAM est de 0" ? Peut-on rejeter l'hypothèse nulle avec un seuil de confiance de 95% ?

```
# test 2
t.test(elections$pctPJD, elections$pctPAM)
```

```
##
##  Welch Two Sample t-test
##
## data:  elections$pctPJD and elections$pctPAM
## t = -9.0887, df = 2787.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07130622 -0.04599858
## sample estimates:
## mean of x mean of y
## 0.1441255 0.2027779
```

3.2 Où est-ce que le PJD gagne ? Les statistiques multivariées

Comment considérer deux séries statistiques ensemble (ou plus) ? C'est l'objet de la suite de ce workshop. Pour répondre à cette question, nous allons observer comment le score du PJD varie avec les caractéristiques de la commune, tirées du recensement de 2014.

> **Activité.** Fusionner les résultats des élections communales avec les données du recensement.

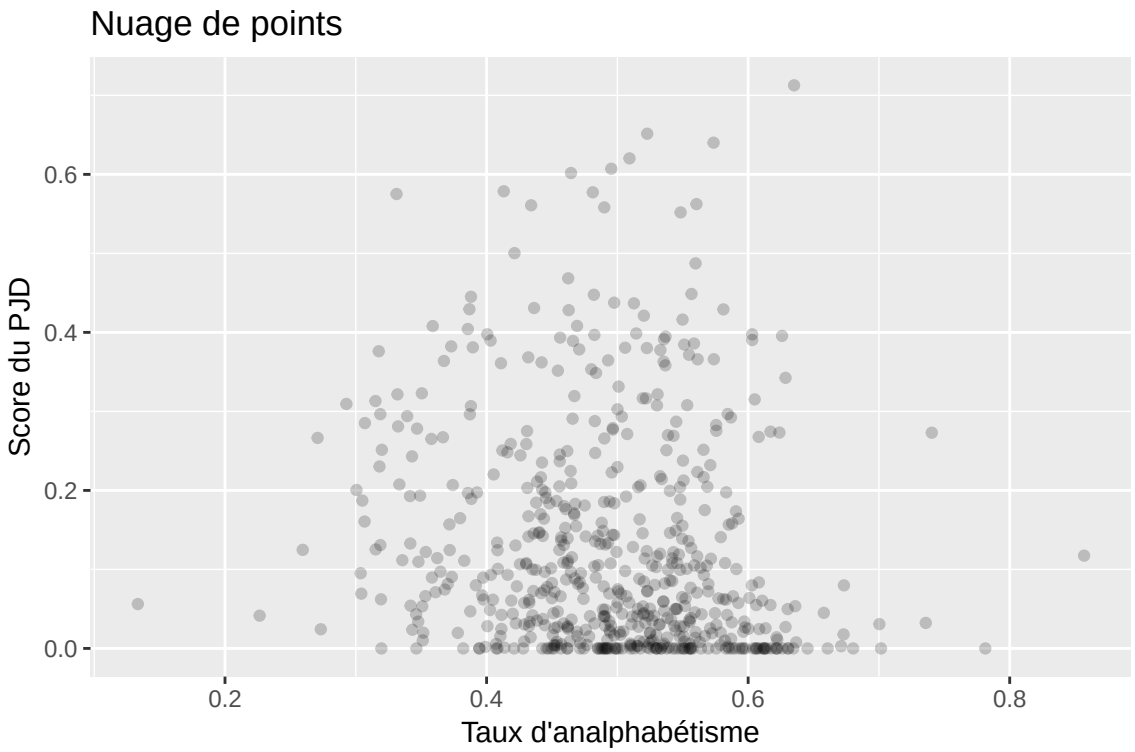
```
census <- read_excel("recensement.xlsx")
df <- inner_join(census, elections)
df <- na.omit(df)
```

3.2.1 Avec deux variables : le lien entre PJD et alphabétisation

La manière la plus simple de visualiser le lien entre deux variables est d'inspecter un nuage de points. Ce type de graphiques représente une variable sur l'axe des x, et une autre variable sur l'axe des y.

> **Activité.** Créer le nuage de point entre la variable `pctPJD` et la variable `txAnalphabetisme`. Quelle est la relation entre le score du PJD et le taux d'analphabétisme ?

```
ggplot(df, aes(x = txAnalphabetisme, y = pctPJD)) +
  geom_point(alpha = .2) +
  labs(x = "Taux d'analphabétisme",
       y = "Score du PJD",
       title = "Nuage de points")
```



Si l'analyse visuelle est un prérequis indispensable pour comprendre la relation entre deux variables, il faut souvent la complémenter d'une analyse quantitative. Le concept le plus simple est le *coefficient de corrélation*. Ce concept, généralement noté ρ, étend la notion de variance à la relation entre deux variables. Il résume la relation entre deux variables par un nombre compris entre -1 et 1. Quand $\rho = 0$, les deux variables ne sont pas corrélées ; en d'autres termes, le fait que l'une des variables augmente n'a aucun effet sur l'autre variable. Quand $\rho = 1$, il y a une parfaite corrélation *positive* entre les deux variables : une valeur élevée pour l'une des variables se traduit par une valeur élevée pour l'autre des variables. Au contraire, quand $\rho = -1$, il y a une parfaite corrélation *négative* entre les deux variables : une valeur élevée pour l'une des variables se traduit par une valeur faible pour l'autre des variables.

> **Activité.** Calculer le coefficient de corrélation entre les variables `pctPJD` et `txAnalphabetisme` (fonctions `CORREL` sur Excel et Google Sheet). Interpréter le résultat.

```
cor(df$pctPJD, df$txAnalphabetisme)
```

```
## [1] -0.1471174
```

Une technique plus sophistiquée pour analyser la relation entre deux variable est la *régression linéaire*, aussi appelée *moindres carrés ordinaires*. Cette technique résume la relation entre deux variables x et y par une droite. Mathématiquement, la technique dérive les "meilleurs" coefficients β_0 et β_1 tels que, pour chaque observation i

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

où y_i est appellée "variable dépendante," car elle "dépend" de x_i, que l'on appelle une "variable indépendante," et ϵ_i est ce que l'on appelle un "terme d'erreur." La régression linéaire construit la droite qui minimise l'erreur totale. Pour éviter que les erreurs positives ($\epsilon_i > 0$) ne compensent les erreurs négatives ($\epsilon_i < 0$), on minimise la somme des carrés de ces erreurs, d'où le nom "moindres carrés". Mathématiquement, les coefficients α et β résolvent le problème

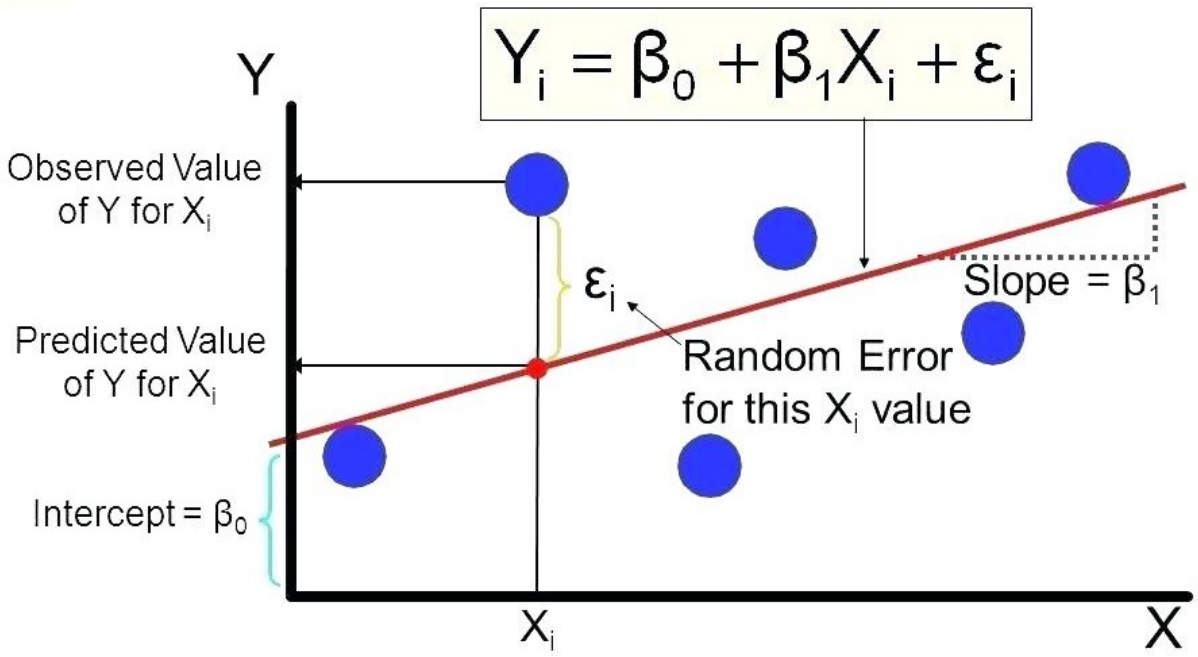


Figure 1: Illustration graphique de la régression linéaire.

$$\min_{\beta_0,\beta_1} \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

L'on peut ensuite utiliser les coefficients pour obtenir des prédictions, généralement notées $\hat{y}_i$; ces prédictions se calculent en enlevant le terme d'erreur :

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

La figure 1 donne une représentation graphique.

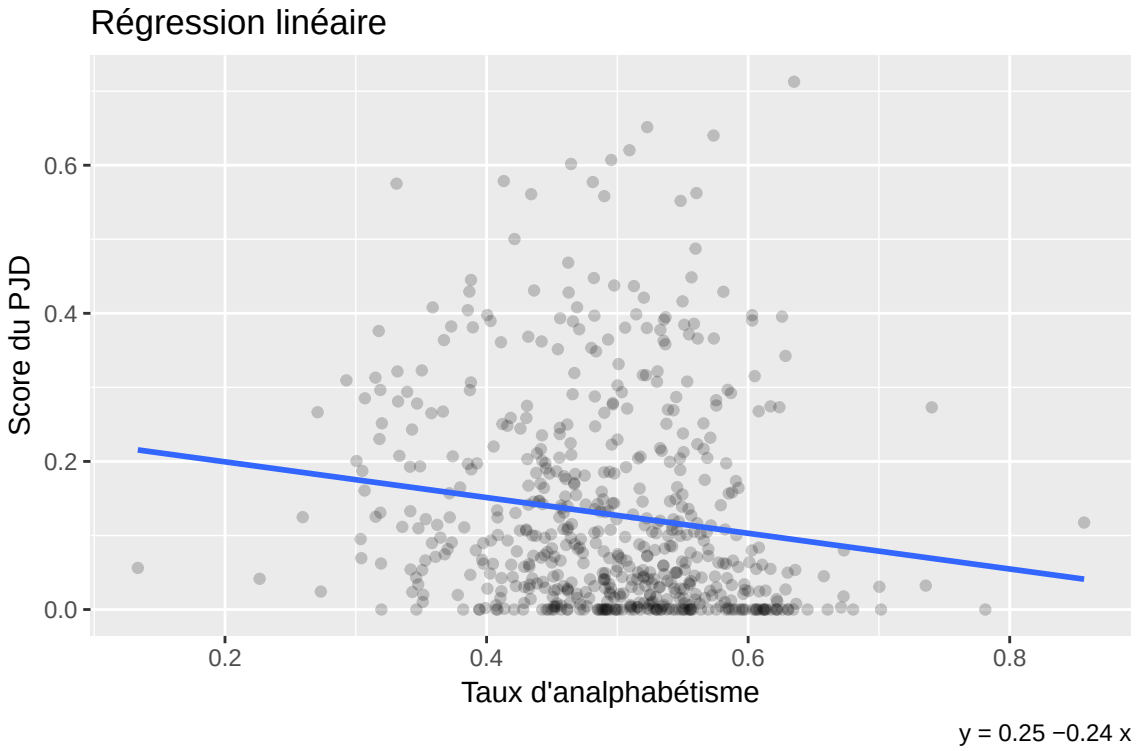
> **Activité.** Réaliser la régression de y, la variable `pctPJD` sur x, la variable `txAnalphabetisme`, et afficher les coefficients.

```
mod <- lm(pctPJD ~ txAnalphabetisme, data = df)
summary(mod) # afficher l'output detaille
```

```
##
## Call:
## lm(formula = pctPJD ~ txAnalphabetisme, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17061 -0.10198 -0.04877  0.05949  0.61821
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.24769    0.03356   7.381 5.41e-13 ***
## txAnalphabetisme -0.24118    0.06693  -3.604  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1397 on 587 degrees of freedom
## Multiple R-squared:  0.02164,    Adjusted R-squared:  0.01998
## F-statistic: 12.99 on 1 and 587 DF,  p-value: 0.0003405
```

```r
# afficher le output graphiquement sur notre nuage de points
ggplot(df, aes(x = txAnalphabetisme, y = pctPJD)) +
  geom_point(alpha = .2) +
  geom_smooth(method = "lm", se = F) +
  labs(x = "Taux d'analphabétisme",
       y = "Score du PJD",
       title = "Régression linéaire",
       caption = sprintf("y = %s %s x",
                         round(coef(mod)[1], 2),
                         round(coef(mod)[2], 2)))
```



Comment interpréter les coefficients ? L'*intercept* – c'est-à-dire le coefficient β_0 – correspond au score de base du PJD, net de l'effet de l'alphabétisation. La pente – le coefficient β_1 – est **l'effet marginal** de l'alphabétisation sur le score du PJD. En l'occurence, augmenter le taux d'analphabétisation de 0 à 1, soit passer de 0% d'analphabètes à 100% d'analphabètes réduit, *en moyenne*, le score du PJD de 42 points de pourcentage. D'autre part, l'output détaillé contient aussi une mesure de l'incertitude autour de chacun de nos paramètres : la dernière colonne, `Pr(>|t|)` est la p-value autour de l'hypothèse nulle que le coefficient est égal à 0. Cette hypothèse nulle est particulièrement importante, parce qu'elle stipule que la variable indépendante n'a aucun effet sur la variable dépendante. En d'autres termes, si le coefficient β_1 est 0, alors en moyenne, passer de 0% d'analphabètes à 100% d'analphabètes augmente le score du PJD de 0 points de pourcentage.

> **Aller plus loin.** Les coefficients d'une régression représentent *l'effet moyen* de x sur y. En effet, la valeur qui minimise l'erreur est la moyenne. Pour s'en convaincre, complétez la procédure suivante. Créez une nouvelle variable, nommée `alphabetise` qui est égale à 1 si le taux d'analphabétisme est inférieur à 50% et qui est égale à 0 si ce taux est supérieur à 50%. Calculez

le score moyen du PJD dans les villages alphabétisés (`alphabetise = 1`) et dans les villages non-alphabétisés (`alphabetise = 0`). Comparez ces valeurs aux valeurs prédites par la régression quand `alphabetise = 1` et quand `alphabetise = 0`.

```
df <- df %>%
  mutate(alphabetise = as.integer(txAnalphabetisme < .5))

tapply(df$pctPJD, df$alphabetise, mean) # score moyen PJD dans les villages alphabetises et non-alphabe
```

```
##         0         1
## 0.1125952 0.1439180
```

```
summary(lm(pctPJD ~ alphabetise, data = df)) # regression
```

```
##
## Call:
## lm(formula = pctPJD ~ alphabetise, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14392 -0.10802 -0.05023  0.06093  0.60012
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.112595   0.008259  13.633   <2e-16 ***
## alphabetise 0.031323   0.011573   2.707    0.007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1404 on 587 degrees of freedom
## Multiple R-squared:  0.01233,    Adjusted R-squared:  0.01064
## F-statistic: 7.326 on 1 and 587 DF,  p-value: 0.006996
```

3.2.2 Avec plusieurs variables : le lien entre PJD et alphabétisation

Cette section va plus loin. Elle nécessite l'usage d'un logiciel de statistiques spécialisé. Nous allons a présent voir comment le score du PJD varie en fonction de plusieurs variables. Les notions que nous voyons ici généralisent ce que nous venons de voir dans le cas de deux variables à trois variables ou plus. En l'occurrence, nous allons considérer les variables indépendantes suivantes :

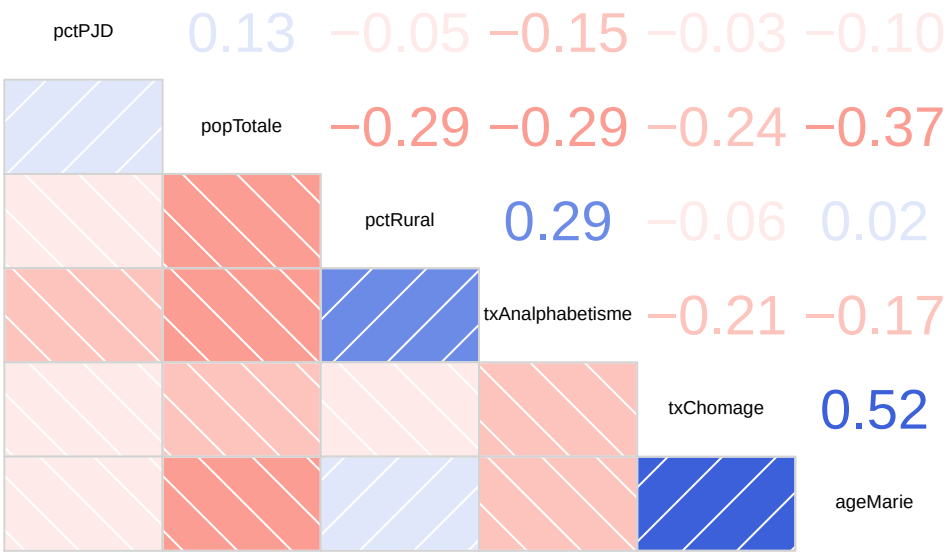
- `popTotale1000`: la population de la commune (en milliers, générée à partir des données brutes)
- `pctRural`: le pourcentage de ruraux dans la commune (généré à partir des données brutes)
- `txAnalphabetisme`: le pourcentage d'analphabêtes dans la commune (comme précédemment)
- `txChomage`: le taux de chômage dans la commune
- `ageMarie`: l'âge moyen du mariage dans la commune.

```
# creation des variables
df <- df %>%
  mutate(pctRural = popRurale / popTotale,
         popTotale1000 = popTotale / 1e3)
```

Nous avons vu le coefficient de corrélation. Le *corrélogramme* est une représentation graphique de plusieurs coefficients de corrélations. Il représente sous la forme d'un tableau le coefficient de corrélation entre chacunes de nos variables.

```
corrgram(
  x = df %>% select(pctPJD, popTotale, pctRural, txAnalphabetisme, txChomage, ageMarie),
```

```
  upper.panel = panel.cor
)
```



Activité. Interpréter le corrélogramme.

L'on voit que les variables les plus corrélées avec le score du PJD sont le taux d'analphabétisme (-.19) et la population totale (.16). Plus le taux d'analphabétisme est élevé, plus le score du PJD est faible. Plus la population totale est élevée, plus le score du PJD est élevé. Le problème est que ces variables peuvent en cacher d'autres. Par exemple, on voit aussi que la population totale est largement corrélée avec le taux d'analphabétisme (-.26) : les grandes villes sont plus alphabétisées. Alors, est-ce que le PJD tend à gagner dans les grandes communes parce qu'elles sont grandes ou parce qu'elles sont plus alphabétisées ? Nous avons atteint la limite de ce que la corrélation peut nous offir.

Pour résoudre ce problème, l'on peut recourir à la *régression linéaire multivariée*. Comme la régression linéaire univariée que nous avons vu plus haut, cette technique résume la relation entre la variable dépendante y et les variables indépendantes $x_1, x_2, ..., x_k$ par une droite. Mathématiquement, la technique dérive les "meilleurs" coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_k$ tels que, pour chaque observation i

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \epsilon_i,$$

A la différence du corrélogramme, la régression linéaire mesure l'effet moyen d'une variable indépendante sur la variable dépendante *toutes choses égales par ailleurs*. En d'autres termes, un coefficient quantifie l'effet de la variable indépendante auquel il correspond net de l'effet des autres variables incluses dans la régression, ce qui résout notre problème. Considérons la régression du score du PJD sur nos variables indépendantes :

Activité. Interpréter la régression linéaire multivariée ci-dessous.

```
mod2 <- lm(pctPJD ~ txAnalphabetisme + txChomage + popTotale1000 +  pctRural  + ageMarie, data = df)
summary(mod2)
```

```
##
```

```
## Call:
## lm(formula = pctPJD ~ txAnalphabetisme + txChomage + popTotale1000 +
##     pctRural + ageMarie, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17729 -0.10131 -0.04723  0.06238  0.62009
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.4211526  0.1119455   3.762 0.000186 ***
## txAnalphabetisme -0.2506384  0.0756557  -3.313 0.000981 ***
## txChomage         0.0108730  0.0621467   0.175 0.861174
## popTotale1000     0.0008444  0.0007539   1.120 0.263151
## pctRural          0.0163169  0.0511757   0.319 0.749961
## ageMarie         -0.0069266  0.0030905  -2.241 0.025384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1389 on 583 degrees of freedom
## Multiple R-squared:  0.04049,    Adjusted R-squared:  0.03226
## F-statistic: 4.921 on 5 and 583 DF,  p-value: 0.0002022
```

L'on voit que passer de 0 à 100% d'analphabètes réduit le score du PJD de 29 points de pourcentage. Augmenter la population de 1000 habitants augmente le score du PJD de .1 point de pourcentage. Augmenter l'âge moyen du mariage d'un an réduit le score du PJD de .6 points de pourcentage. Les autres variables (taux de chômage et pourcentage de ruraux) n'ont pas d'effet *significatif* sur le score du PJD. Par significatif, on entend que la p-value associée au test coefficient = 0 est trop élevée pour conclure que le coefficient est différent de 0. Le statisticien considère, pour cela, une série de seuils standards, symbolisés par des étoiles (à droite de la colonne `Pr(>|t|)`. La symbolique est détaillée au bas de la table. Par exemple, si la p-value est inférieure à 5% mais supérieure à 1%, on symbolise cela par une étoile. C'est le cas pour la variable `popTotale1000`. Si la p-value est supérieure à 10%, alors on considère qu'il est peu probable que le coefficient associé à la variable soit différente de 0, et on ne met pas d'étoile. C'est le cas pour les variables `txChomage` et `pctRural`.

> **Aller plus loin : l'analyse des résidus.** Pour certaines applications, analyser les résidus de la régression est très intéressant. La différence entre la prédiction générée par le modèle et les valeurs observées (les *résidus*) donnent une idée des cas où le PJD a sur-performé vu les variables incluses dans le modèle (résidu positif) et des cas où le PJD a sous-performé (résidu négatif). Les tables ci-dessous affichent ces cas. Il pourrait être intéressant de revenir plus en détail sur ces cas et, au besoin, de faire une étude de cas plus détaillée afin d'apprendre les raisons de ces succès et de ces échecs.

```
df$prediction <- predict(mod2)
df$residu <- df$pctPJD - df$prediction

df <- df %>% arrange(residu)
knitr::kable(df %>% select(region, province, commune, pctPJD, prediction, residu) %>% head(),
             digits = 4)
```

region	province	commune	pctPJD	prediction	residu
Casablanca - Settat	El-Jadida	My Abdellah	0.0541	0.2314	-0.1773
Oriental	Figuig	Ain Chouater	0.0000	0.1678	-0.1678
Casablanca - Settat	Berrechid	Oulad Ziyane	0.0058	0.1698	-0.1640
Casablanca - Settat	Benslimane	Oulad Yahya Louta	0.0000	0.1551	-0.1551

region	province	commune	pctPJD	prediction	residu
Casablanca - Settat	Benslimane	Oulad Ali Toualaa	0.0000	0.1528	-0.1528
Marrakech - Safi	Rehamna	Lamharra	0.0000	0.1513	-0.1513

```
knitr::kable(df %>% select(region, province, commune, pctPJD, prediction, residu) %>% tail(),
             digits = 4)
```

region	province	commune	pctPJD	prediction	residu
Marrakech - Safi	Al-Haouz	Azgour	0.6202	0.1381	0.4822
Souss - Massa	Tata	Kasbat Sidi Abdellah Ben M'Barek	0.6017	0.1079	0.4938
Souss - Massa	Tata	Tamanarte	0.6070	0.0988	0.5083
Marrakech - Safi	Rehamna	Bourrous	0.6401	0.1202	0.5199
Marrakech - Safi	Essaouira	Oulad M'Rabet	0.6514	0.1216	0.5299
Marrakech - Safi	Chichaoua	Timezgadiouine	0.7127	0.0926	0.6201

3.2.3 L'inférence causale

Pour conclure cette introduction, nous allons nous repencher sur la régression. Comme nous l'avons vu plus haut, l'avantage de la régression multivariée sur le corrélogramme est que la régression quantifie l'effet d'une variable indépendante sur la variable dépendante net de l'effet des autres variables indépendantes incluses dans le modèle. Considérons l'effet du taux d'analphabétisme. Dans notre régression univariée, il est de -.31. Dans notre régression multivariée, il est de -.29. L'effet devient moins important parce que désormais, cet effet est net des autres variables incluses dans le modèle qui sont corrélées avec le taux d'analphabétisme, notamment la population de la commune.

Mais alors, comment être sûr que l'on mesure correctement l'effet de l'analphabétisme sur le score du PJD ? En particulier, comment être sûr que l'on a pas oublié d'inclure dans notre modèle des variables importantes, qui seraient corrélées avec le taux d'analphabétisme ?

L'*inférence causale* est la branche des statistiques qui s'occupe de ce problème. L'inférence causale utilise une métaphore médicale. Elle a pour objet de mesurer l'effet d'un *traitement* sur un résultat d'intérêt, défini comme un résultat *contrefactuel*. Supposons que l'on s'intéresse à l'effet d'une nouvelle pilule contre la fièvre. Le résultat d'intérêt est la température du patient i, notée Y_i. Notons $T_i = 1$ si le patient i prend la pilule, et $T_i = 0$ si le patient ne prend pas la pilule. Notons maintenant $Y_i(T_i)$ la température du patient selon qu'il prenne la pilule ou non. En d'autres termes, si le patient prend la pilule, sa température est $Y_i(1)$. S'il ne la prend pas, sa température est $Y_i(0)$. On dit que $Y_i(1)$ et $Y_i(0)$ sont des résultats contrefactuels. L'effet du traitement sur notre patient i, noté τ_i, est sa différence de température selon qu'il prenne le traitement ou non :

$$\tau_i = Y_i(1) - Y_i(0)$$

Le problème fondamental est que l'on ne peut pas mesurer *l'effet individuel du traitement* (sur notre patient i), car on ne peut pas observer le patient ayant pris sa pilule et ce même patient n'ayant pas pris sa pilule. A la place, l'inférence causale se propose de mesurer *l'effet moyen du traitement* ; c'est-à-dire l'effet du traitement en moyenne sur une population. Pour cela, il faut administrer le traitement au hasard. On appelle le groupe qui a reçu le traitement le *groupe de traitement* et le groupe qui n'a pas reçu le traitement le *groupe de contrôle*. Si le traitement a été administré au hasard, alors les groupes de traitement et de contrôle sont statistiquement identiques. En d'autres termes, recevoir le traitement n'est corrélé à aucune des caractéristiques de la population. De ce fait, on peut mesurer l'effet du traitement tout simplement en faisant la différence des moyennes. Notons $\bar{\tau}$ l'effet moyen du traitement, $\bar{Y}_{T=1}$ la température moyenne de la population qui a reçu le traitement et $\bar{Y}_{T=0}$ la température moyenne de la population qui n'a pas reçu le traitement. L'effet moyen du traitement est

$$\bar{\tau} = \bar{Y}_{T=1} - \bar{Y}_{T=0}$$

Remarque : corrélation et causalité. Quand on interprète des données statistiques, il est important de distinguer corrélation et causalité. La relation que l'on a estimé entre le taux d'analphabétisme et le score du PJD n'est qu'une corrélation. Dire que nos résultats établissent que l'éducation *cause* le score du PJD est incorrect. Il est possible que l'éducation soit corrélée avec une autre variable que l'on n'a pas mesuré qui soit elle-même responsable du score du PJD.

Activité : méthodes expérimentales. L'inférence causale montre combien les expériences sont utiles pour quantifier l'effet causal d'un traitement. En effet, les expériences permettent à l'analyste d'administrer le traitement de manière aléatoire, et ainsi de mesurer de manière crédible son effet sur la population. Imaginer une expérience permettant de mesurer l'effet de l'éducation sur le score du PJD.

Une telle expérience pourrait être de sélectionner au hasard un certain nombre de communes un an avant les élections. Dans les communes de traitement, on distribue des bons pour accéder à un programme gratuit de cours du soir. Dans les communes de contrôle, on ne distribue pas ces bons. On compare ensuite les scores du PJD dans les communes de traitement et de contrôle après les élections.

4 Références

- Angrist, Joshua D; Pischke, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. Amazon. Un grand classique des statistiques, utilisé par de nombreuses universités. L'ouvrage se focalise sur des applications économiques et sur la théorie statistique plutôt que sur apprendre à maîtriser un logiciel.
- Wickham, Hadley; Grolemund, Garrett. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017. Amazon / Manuscrit en ligne. Une introduction très complète à R, écrite par l'un des contributeurs les plus importants à ce logiciel. Disponible gratuitement en ligne.
- Imai, Kosuke. *Quantitative social science, An Introduction*. Princeton University Press, 2018. Amazon. Un excellent ouvrage pour apprendre en même temps le code sur R et les statistiques, rempli d'applications aux sciences sociales et d'ouvertures vers de nouveaux types de données (réseaux, texte...)