

# Séance 2 – Premiers éléments de statistique appliquée

Pierre Pora

# L'objet de la séance

- ▶ Les économistes font largement appel à des arguments empiriques de nature statistique
- ▶ C'est au total en cela que consiste l'économétrie
- ▶ Nous allons commencer à voir ce que cela veut dire en pratique, sur des exemples récents
- ▶ En essayant de donner un aperçu du paysage des possibles

# Travailler sur des données ?

- ▶ L'économie n'est pas une discipline exclusivement théorique ou spéculative
- ▶ Elle est même de plus en plus nettement empirique
- ▶ Mais qu'est-ce que cela veut dire en pratique ?

# Taxonomie des données

- ▶ Une distinction usuelle : données quantitatives et données qualitatives
  - ▶ Données quantitatives : se présentent en général sous la forme de bases de données
    - ▶ avec des lignes (~ individus)
    - ▶ et des colonnes (~variables)
  - ▶ Données qualitatives : tout le reste
    - ▶ surtout utilisées dans d'autres sciences sociales
    - ▶ entretiens, observations etc.

# Taxonomie des données

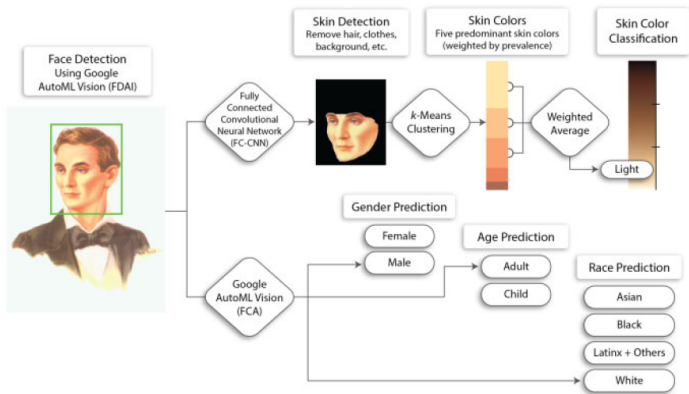
- ▶ Une distinction usuelle : données quantitatives et données qualitatives
  - ▶ Une frontière poreuse : développement de techniques permettant de traiter de façon quantitative des données non-structurées auparavant plus souvent traitées de façon qualitative
  - ▶ Economie historique : grosse activité consistant à créer des bases de données quantitatives à partir d'archives traitées auparavant de façon plus qualitative
  - ▶ Distinction un peu scolaire mais utile pour se repérer
  - ▶ Attention : parfois ces mots sont utilisés dans un autre contexte et veulent dire tout à fait autre chose !
    - ▶ Variables qualitatives vs. variables quantitatives → différents types de variables dans les données quantitatives

## Un exemple récent : transformer un corpus littéraire et iconographiques en données quantitatives

- ▶ Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha et Teodora Szasz. 2023. “What We Teach About Race and Gender : Representation in Images and Text of Children’s Books.” *The Quarterly Journal of Economics*, 138(4) : 2225–2285

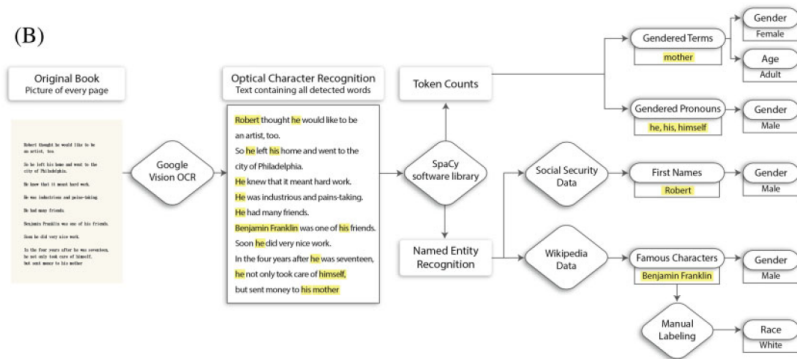
# Un exemple récent : transformer un corpus littéraire et iconographique en données quantitatives

(A)



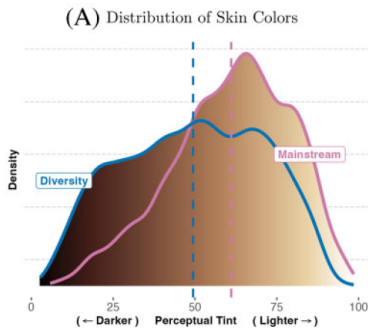
# Un exemple récent : transformer un corpus littéraire et iconographiques en données quantitatives

(B)

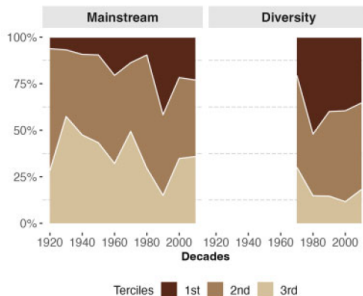




# Un exemple récent : transformer un corpus littéraire et iconographiques en données quantitatives



(B) Mean Proportion in Each Tercile, Over Time



# Taxonomie des données : branche des données quantitatives

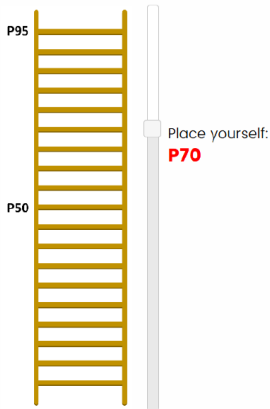
- ▶ Plein de façons de distinguer différents types de données !
- ▶ Une première distinction : qui a produit les données ?
  - ▶ Vous / les auteurs :
    - ▶ Cas fréquent en économie historique
    - ▶ Ou bien par exemple enquête *ad hoc* après un RCT
  - ▶ Données produites par la statistique publique
    - ▶ Par exemple l'enquête Emploi ou le recensement
  - ▶ Données détenues par un organisme public ou privé
    - ▶ Souvent plus difficiles d'accès

## Collecter soi-même ses données pour confronter perceptions et réalité

- ▶ Alesina, Alberto, Stefanie Stantcheva et Edoardo Teso. 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108 (2) : 521–54.
- ▶ Hvidberg, Kristoffer B, Claus T Kreiner et Stefanie Stantcheva. 2023. "Social Positions and Fairness Views on Inequality." *The Review of Economic Studies* 90(6) : 3083–3118

# Collecter soi-même ses données pour confronter perceptions et réalité

You previously reported that you had a yearly income in 2017 of 400000 DKK before tax. We will now ask you to report where you think this income placed you on the income ladder in 2017 for people who were born in 1970. Use the slider to select your position. Later, we will inform you about your true position.



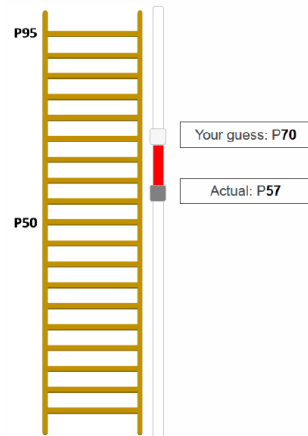
# Collecter soi-même ses données pour confronter perceptions et réalité

## Rank among all people born in 1970

You GUESSED that you were on position **P70**.

Based on the income you reported, your **TRUE** position is **P57**.

You are actually 13 positions lower on the ladder than you thought.

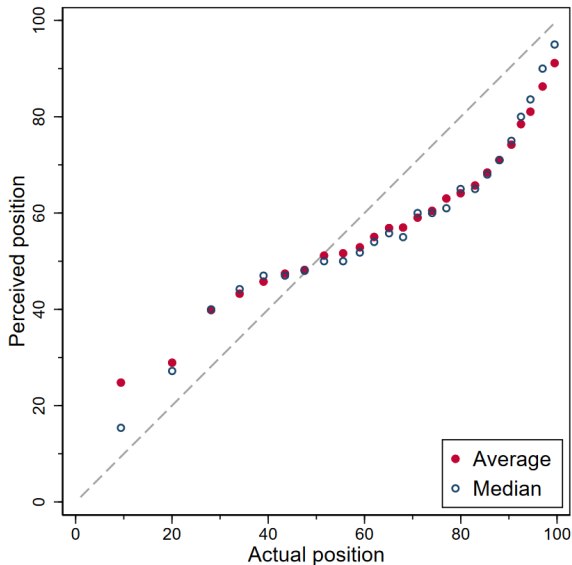


Collecter soi-même ses données pour confronter perceptions et réalité

On a scale from 1 to 7 where 1 is "Completely fair", 4 is "Neither fair or unfair" and 7 is "Completely unfair", indicate to what extent you think that is fair or unfair that there are differences in income among people born the same year as you **WITHIN** the following groups that you are yourself a part of?

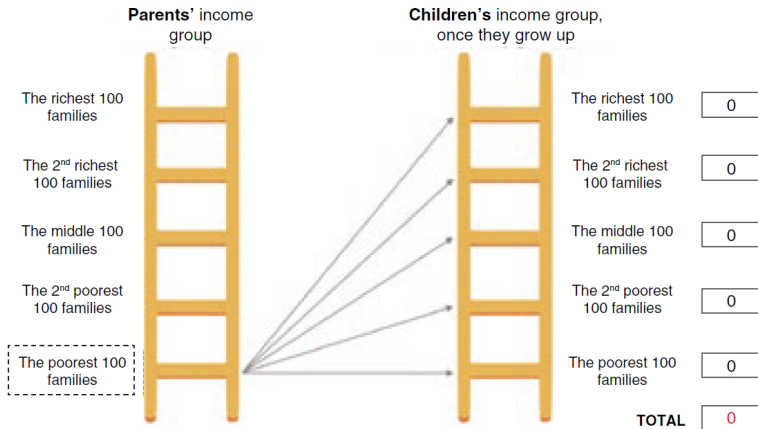
	Completely fair					Neither fair or unfair				Completely unfair	
	1	2	3	4	5	6	7				
Differences in income among people <b>born in 1970</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Differences in income among <b>men</b> born in 1970	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Differences in income among people, living in <b>København municipality</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Differences in income among people with the educational level <b>Master or PhD program</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				
Differences in income among people working in the sector <b>Finance and insurance</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>				

# Collecter soi-même ses données pour confronter perceptions et réalité



# Collecter soi-même ses données pour confronter perceptions et réalité

Here are **500 families** that represent the US population:





# Collecter soi-même ses données pour confronter perceptions et réalité

TABLE 2—PERCEIVED AND ACTUAL TRANSITION PROBABILITIES ACROSS COUNTRIES

	US		UK		France		Italy		Sweden		US versus EU	
	Actual	Perceived	Actual	Perceived	Actual	Perceived	Actual	Perceived	Actual	Perceived	Perceived	Perceived
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	US	EU
Q1 to Q5	7.8	11.7 (0.00)	11.4	10.0 (0.00)	11.2	9.1 (0.00)	10.4	10.1 (0.48)	11.1	9.2 (0.00)	11.7	9.6 (0.00)
Q1 to Q4	12.7	12.0 (0.00)	12.9	10.6 (0.00)	12.8	10.5 (0.00)	15.6	11.2 (0.00)	17.3	11.2 (0.00)	12.0	10.9 (0.00)
Q1 to Q3	18.7	22.3 (0.00)	19.9	19.4 (0.13)	23.0	21.5 (0.00)	21.0	21.9 (0.03)	21.0	24.5 (0.00)	22.3	21.6 (0.06)
Q1 to Q2	27.7	21.8 (0.00)	25.1	22.2 (0.00)	23.8	23.6 (0.55)	25.8	23.1 (0.00)	23.8	23.1 (0.09)	21.8	23.0 (0.00)
Q1 to Q1	33.1	32.2 (0.07)	30.6	37.8 (0.00)	29.2	35.3 (0.00)	27.3	33.6 (0.00)	26.7	32.0 (0.00)	32.2	34.9 (0.00)
Observations		2,170		1,290		1,297		1,242		881	2,170	4,710
<i>p</i> -value from joint test		0.00		0.00		0.00		0.00		0.00		0.00

# Collecter soi-même ses données : avantages et inconvénients

- ▶ **Avantage évident : adéquation parfaite des items mesurés à la question de recherche**
  - ▶ à condition quand même de travailler correctement
- ▶ **Inconvénients**
  - ▶ Collecte coûteuse qui demande beaucoup de moyens
    - ▶ y compris pour coller au cadre juridique de la gestion des données personnelles !
  - ▶ Difficulté de maîtriser l'échantillonnage avec une qualité comparable avec ce qui se fait dans la statistique publique
    - ▶ Les répondants sont-ils représentatifs de la population étudiée ?
  - ▶ Coordination des travaux de recherches entre eux → risque de saturation pour les enquêtés potentiels

# Collecter soi-même ses données : avantages et inconvénients

TABLE 1—SAMPLE CHARACTERISTICS

	United States		United Kingdom		France		Italy		Sweden	
	Sample	Pop	Sample	Pop	Sample	Pop	Sample	Pop	Sample	Pop
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Male	0.48	0.48	0.48	0.49	0.50	0.48	0.50	0.48	0.47	0.50
18–29 years old	0.26	0.27	0.26	0.24	0.23	0.21	0.19	0.19	0.21	0.24
30–39 years old	0.18	0.19	0.18	0.20	0.20	0.19	0.22	0.21	0.18	0.19
40–49 years old	0.19	0.21	0.21	0.21	0.21	0.20	0.23	0.24	0.19	0.21
50–59 years old	0.21	0.20	0.18	0.20	0.20	0.20	0.20	0.20	0.21	0.18
60–69 years old	0.16	0.14	0.16	0.16	0.16	0.19	0.17	0.17	0.21	0.18
Income bracket 1	0.16	0.18	0.31	0.31	0.31	0.32	0.27	0.27	0.33	0.33
Income bracket 2	0.22	0.20	0.35	0.35	0.30	0.30	0.28	0.28	0.26	0.29
Income bracket 3	0.23	0.22	0.11	0.11	0.14	0.14	0.18	0.19	0.22	0.22
Income bracket 4	0.39	0.39	0.23	0.23	0.25	0.24	0.27	0.26	0.18	0.17
Married	0.51	0.49	0.47	0.41	0.44	0.46	0.55	0.46	0.41	0.33
Native	0.94	0.85	0.89	0.87	0.94	0.85	0.97	0.92	0.91	0.82
Employed	0.62	0.58	0.65	0.61	0.63	0.47	0.64	0.45	0.66	0.67
Unemployed	0.08	0.08	0.05	0.03	0.12	0.05	0.11	0.06	0.07	0.05
College	0.42	0.28	0.37	0.42	0.30	0.25	0.38	0.15	0.33	0.36

# Une présentation rapide de la statistique publique

- ▶ Cadre institutionnel et juridique à part destiné à permettre et protéger la collecte d'information statistique utile à des fins d'information générale et d'aide à la décision publique
- ▶ En France le système statistique publique :
  - ▶ l'Insee
  - ▶ les SSM (services statistiques ministériels)
    - ▶ par exemple la Dares pour le ministère du Travail
  - ▶ encadrés par l'ASP (Autorité de la Statistique Publique), le CNIS (Conseil National de l'Information Statistique), le Comité du label, le Comité du secret statistique
    - ▶ instances qui encadrent l'indépendance professionnelle de la statistique publique, organisent la rencontre entre l'activité de la statistique publique et la demande sociale, garantissent la qualité des opérations statistiques et régulent l'accès à l'information statistique confidentielle

# Une présentation rapide de la statistique publique

- ▶ Activité qui couvre un domaine très large
  - ▶ statistique d'entreprise
  - ▶ statistique sociale
  - ▶ du tourisme à la santé en passant par les transports et la culture
- ▶ Garanties de qualité sur l'échantillonnage et les traitements post-collecte

# Une présentation rapide de la statistique publique

- ▶ Niveaux d'accessibilité des données dans le cas français
  - ▶ Données agrégées plus ou moins facilement accessibles *via* les sites de l'Insee et des différents SSM
  - ▶ Parfois des données individuelles accessibles directement
    - ▶ Fichiers détails
  - ▶ Données individuelles un peu retraitées pour limiter les risques de réidentification *via* la plateforme Quêtelet
    - ▶ Fichiers de Production et de Recherche (FPR)
    - ▶ Il faut faire une demande mais pas trop compliqué
- ▶ Données individuelles au CASD (Centre d'Accès Sécurisé aux Données) après passage au Comité du secret statistique
  - ▶ Plateforme d'accès aux données avec une identification individuelle des utilisateur-riche-s
  - ▶ Demande un peu lourde + accès coûteux

# Une utilisation astucieuse de l'enquête Emploi

- ▶ Maurin, Eric, and Julie Moschion. 2009. "The Social Multiplier and Labor Market Participation of Mothers." *American Economic Journal : Applied Economics* 1 (1) : 251–72.

# Une utilisation astucieuse de l'enquête Emploi

One key feature of the French LFS is that the basic sampling units actually consist of groups of about 20 adjacent households.<sup>3</sup> (these groups are called *aires*). More specifically, a typical LFS consists of a representative sample of about 3,500 *aires*. Each year, within each *aire*, all the households are surveyed and, within each household, all the persons aged 15 years old or older are surveyed. The French statistical office (INSEE) has chosen this sampling strategy in order to reduce the travelling expenses of the investigators who are in charge of the survey.



# Une utilisation astucieuse de l'enquête Emploi

For each woman in our sample, we observe, on average, four other women with two or more children living in the same small neighborhood.<sup>4</sup> Hence, for each woman in our sample, we can compute several variables describing the average characteristics of the other families with two or more children living in her *aire*, namely the proportion of families in which the two eldest children are “same sex,” the proportion of families whose second child was born at the end of the year, and the proportion of families where the mother participates in the labor market. Let us emphasize that, for each respondent, the different *aire*-level indicators are constructed using only the information on the neighbors, i.e., the respondent himself is excluded from these indicators.

# Une utilisation astucieuse de l'enquête Emploi

TABLE 1—IMPACT OF THE SEX COMPOSITION OF THE TWO ELDEST CHILDREN ON MOTHERS' FERTILITY AND LABOR MARKET PARTICIPATION

	Sex of the two eldest children						<i>Difference</i> (5) – (6)
	2 boys (1)	2 girls (2)	boy, girl (3)	girl, boy (4)	Same sex (5)	Different sex (6)	
Proportion in population	0.262 (0.002)	0.240 (0.002)	0.250 (0.002)	0.247 (0.002)	0.503 (0.002)	0.497 (0.002)	0.006 (0.003)
Proportion 3 children or more	0.313 (0.005)	0.313 (0.005)	0.270 (0.005)	0.280 (0.005)	0.313 (0.004)	0.275 (0.004)	0.038* (0.005)
Proportion participating in labor market	0.585 (0.006)	0.593 (0.006)	0.609 (0.006)	0.602 (0.006)	0.589 (0.004)	0.605 (0.004)	−0.017* (0.006)

# Une utilisation astucieuse de l'enquête Emploi

	Individual characteristics of mothers					
	Age	Age at 1st birth	French	Number of children	High school graduate	LM Part.
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel B: Neighbors' children sex-mix</i>						
(c) Proportion same sex > 0.75	31.14 (0.04)	23.11 (0.05)	0.906 (0.003)	2.373 (0.009)	0.681 (0.006)	0.591 (0.006)
(d) Proportion same sex < 0.25	31.15 (0.04)	23.04 (0.05)	0.908 (0.003)	2.378 (0.009)	0.685 (0.006)	0.611 (0.006)
<i>Difference</i> <i>(c) - (d)</i>	<i>0.00</i> <i>(0.06)</i>	<i>0.07</i> <i>(0.06)</i>	<i>-0.002</i> <i>(0.005)</i>	<i>-0.006</i> <i>(0.013)</i>	<i>-0.004</i> <i>(0.009)</i>	<i>-0.020*</i> <i>(0.009)</i>

# Utiliser les données de la statistique publique : avantages et inconvénients

- ▶ **Avantage** : garanties de qualité des données à de multiples niveaux
- ▶ **Inconvénient**
  - ▶ Données collectées indépendamment de la question de recherche particulière que l'on se pose
    - ▶ Mais possible d'intervenir dans la conception des questionnaires e.g. CNIS, comités scientifiques, comités d'utilisateurs
  - ▶ Pas maître-sse-s du processus de collecte et de traitement des données

## Ce que peuvent nous apprendre des données collectées par d'autres entités : un exemple

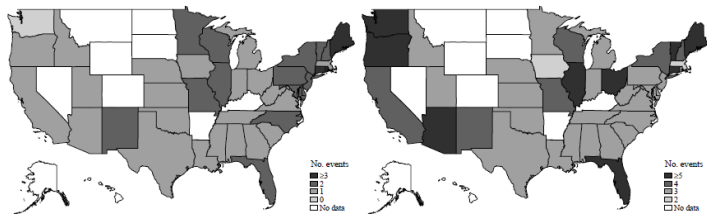
- ▶ Renkin, Tobias, Claire Montialoux et Michael Siegenthaler. 2022. "The Pass-Through of Minimum Wages into U.S. Retail Prices : Evidence from Supermarket Scanner Data." *The Review of Economics and Statistics* 104(5) : 890–908.

# Ce que peuvent nous apprendre des données collectées par d'autres entités : un exemple

**Retail scanner data.** Our empirical analysis is based on scanner data provided by the market research firm Symphony IRI. The dataset is described in detail in Bronnenberg et al. (2008) and Kruger and Pagni (2013). It contains weekly prices and quantities for 31 product categories sold at grocery and drug stores between January 2001 and December 2012. The estimation sample covers 2,493 distinct grocery and drug stores and contains their zip code location.<sup>6</sup> It provides information on an average of 60,600 products over this period. Products are identified by Unique Product Codes (UPC). As an example, a 12oz can and a 20oz bottle of Coca Cola Classic are treated as different products in our data. Stores are located in 530 counties, 41 states and belong to one of about 90 retail chains. The data covers 17% of US counties which are home to about 29% of the US

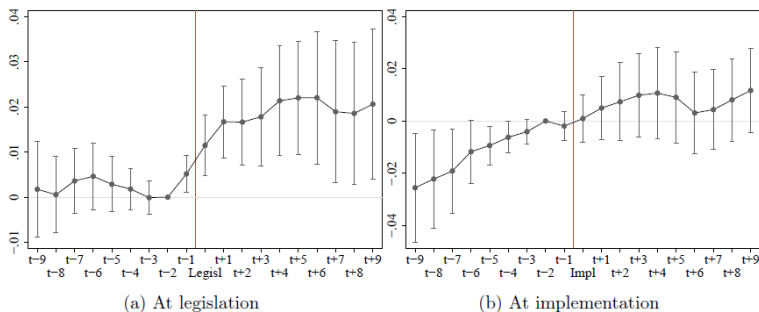
## Ce que peuvent nous apprendre des données collectées par d'autres entités : un exemple

Figure 2: Distribution of minimum wage hikes and legislative events over time and states



# Ce que peuvent nous apprendre des données collectées par d'autres entités : un exemple

Figure 3: Cumulative minimum wage elasticities of prices from separate estimation





## Utiliser des données collectées par d'autres entités : avantages et inconvénients

- ▶ **Avantage :** pour certaines questions, c'est la seule façon d'avoir accès à l'information qui permet de traiter la question empiriquement !
- ▶ **Inconvénients :**
  - ▶ Parfois difficile de s'assurer de la qualité des données
  - ▶ Question de l'accès à d'autres parties
    - ▶ Peut devenir problématique en matière de reproductibilité de la recherche empirique
    - ▶ Exigence croissante des revues de pouvoir accéder aux données pour d'autres chercheurs
    - ▶ Des solutions existent mais elles sont onéreuses

# Une autre façon de classer les données quantitatives

- ▶ Données d'enquêtes
  - ▶ Issues d'une opération statistique destinée à collecter une information sur une population définie à l'avance, à des fins de recherche et d'information générale
  - ▶ Définition d'une population, d'une stratégie d'échantillonnage, d'un questionnaire etc.
    - ▶ Tout cela se fait avec une question / un ensemble de questions de recherche en tête
  - ▶ Exemple dans ce qui précède : l'enquête Emploi

# Une autre façon de classer les données quantitatives

## ▶ Données administratives

- ▶ Issues d'un processus qui n'a pas pour finalité de produire une information statistique, mais souvent un traitement par une administration, une entreprise etc.
- ▶ Les quantités mesurées n'y sont pas forcément pertinente pour la recherche *a priori*, en tout cas pas pensées pour l'être
- ▶ La population concernée peut être difficile à définir
- ▶ Il faut bien comprendre les règles de gestion pour pouvoir les utiliser
- ▶ Exemple dans ce qui précède : les données de caisse

## Des cas intermédiaires à considérer

- ▶ Evolution récente dans la statistique publique : de plus en plus de données administratives disponibles
  - ▶ Lié à des évolutions législatives et réglementaires
- ▶ Objectif de limiter la charge d'enquête car les taux de réponse s'érodent
- ▶ Ne collecter que l'information nécessaire par enquête
- ▶ Et apparier avec des données administratives pour enrichir
- ▶ Exemple : enquête Familles (Insee) appariée aux données fiscales pour connaître le niveau de vie sans avoir à le demander

## Et comment cela se présente-il exactement ?

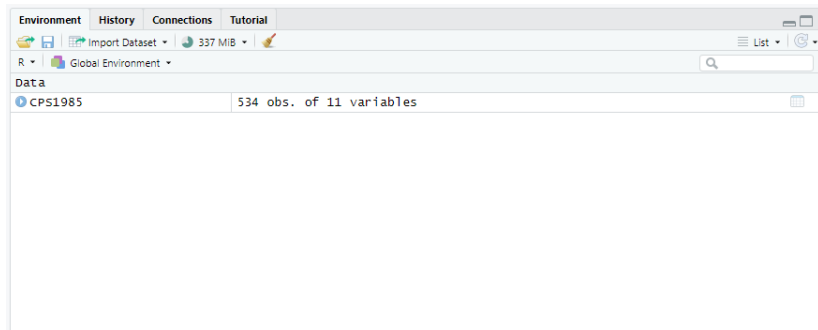
- ▶ En général, les données quantitatives se présentent sous la forme d'une ou plusieurs tables, avec des lignes qui correspondent aux unités enquêtées (les individus de la population) et des colonnes qui correspondent à des variables, les quantités mesurées pour chacune des unités enquêtées
- ▶ Nous allons le voir en regardant un extrait du Current Population Survey américain
  - ▶ Une enquête qui tient le milieu entre le recensement et l'enquête Emploi dans le cas français

## En pratique

```
library(AER)
```

```
data("CPS1985")
```

# En pratique



# En pratique

Environment	History	Connections	Tutorial
R ▾   Global Environment ▾   Import Dataset ▾   340 MIB ▾   🔍			
Data			
CPS1985		534 obs. of 11 variables	
\$ wage	: num	5.1 4.95 6.67 4 7.5 ...	
\$ education	: num	8 9 12 12 12 13 10 12 16 12 ...	
\$ experience	: num	21 42 1 4 17 9 27 9 11 9 ...	
\$ age	: num	35 57 19 22 35 28 43 27 33 27 ...	
\$ ethnicity	: Factor w/ 3 levels	"cauc","hispanic",...: 2 1 1 1 1 1 1 1 1 1 ...	
\$ region	: Factor w/ 2 levels	"south","other": 2 2 2 2 2 2 1 2 2 2 ...	
\$ gender	: Factor w/ 2 levels	"male","female": 2 2 1 1 1 1 1 1 1 1 ...	
\$ occupation	: Factor w/ 6 levels	"worker","technical",...: 1 1 1 1 1 1 1 1 1 1 ...	
\$ sector	: Factor w/ 3 levels	"manufacturing",...: 1 1 1 3 3 3 3 3 1 3 ...	
\$ union	: Factor w/ 2 levels	"no","yes": 1 1 1 1 1 2 1 1 1 1 ...	
\$ married	: Factor w/ 2 levels	"no","yes": 2 2 1 1 2 1 1 1 2 1 ...	



## En pratique

```
View(CPS1985)
```

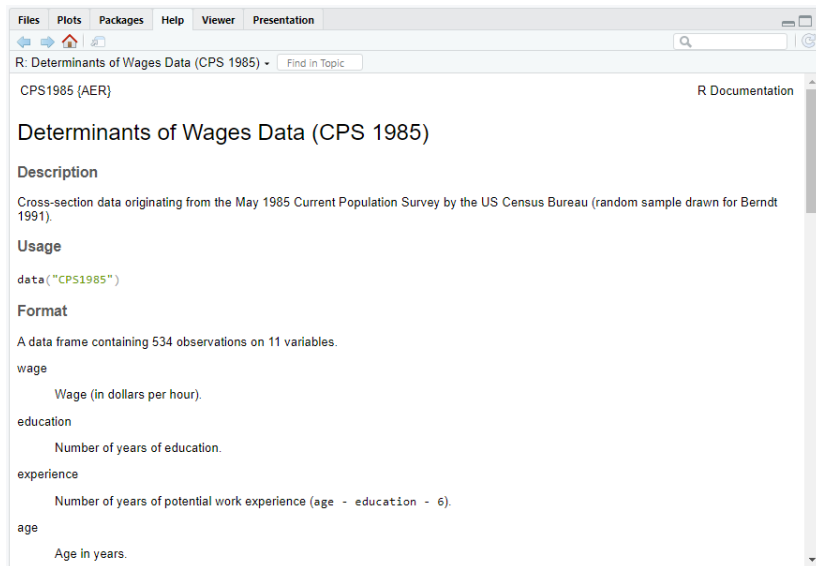
# En pratique

	⬆ wage ⬆	education ⬆	experience ⬆	age ⬆	ethnicity ⬆	region ⬆	gender ⬆	occupation ⬆	sector ⬆	union ⬆	married ⬆
1	5.10	8	21	35	hispanic	other	female	worker	manufacturing	no	yes
1100	4.95	9	42	57	cauc	other	female	worker	manufacturing	no	yes
2	6.67	12	1	19	cauc	other	male	worker	manufacturing	no	no
3	4.00	12	4	22	cauc	other	male	worker	other	no	no
4	7.50	12	17	35	cauc	other	male	worker	other	no	yes
5	13.07	13	9	28	cauc	other	male	worker	other	yes	no
6	4.45	10	27	43	cauc	south	male	worker	other	no	no
7	19.47	12	9	27	cauc	other	male	worker	other	no	no
8	13.28	16	11	33	cauc	other	male	worker	manufacturing	no	yes
9	8.75	12	9	27	cauc	other	male	worker	other	no	no
10	11.35	12	17	35	cauc	other	male	worker	other	yes	yes
11	11.50	12	19	37	cauc	other	male	worker	manufacturing	yes	no
12	6.50	8	27	41	cauc	south	male	worker	other	no	yes
13	6.25	9	30	45	cauc	south	male	worker	other	yes	no
14	19.98	9	29	44	cauc	south	male	worker	other	no	yes
15	7.30	12	37	55	cauc	other	male	worker	construction	no	yes
16	8.00	7	44	57	cauc	south	male	worker	other	no	yes
17	22.20	12	26	44	cauc	other	male	worker	manufacturing	yes	yes
18	3.65	11	16	33	cauc	other	male	worker	other	no	no
19	20.55	12	33	51	cauc	other	male	worker	other	no	yes

# Une documentation des données

```
help(CPS1985)
```

# Une documentation des données



The screenshot shows the RStudio application window. The top menu bar includes 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Presentation'. Below the menu is a toolbar with navigation icons and a search bar. The main pane displays the documentation for the 'CPS1985' dataset, which is part of the 'R: Determinants of Wages Data (CPS 1985)' package. The documentation is organized into sections: 'Description', 'Usage', and 'Format'. The 'Usage' section shows the command `data("CPS1985")`. The 'Format' section describes the data frame and lists variables: 'wage', 'education', 'experience', and 'age', each with a brief description.

Files Plots Packages Help Viewer Presentation

R: Determinants of Wages Data (CPS 1985) Find in Topic

CPS1985 {AER} R Documentation

## Determinants of Wages Data (CPS 1985)

### Description

Cross-section data originating from the May 1985 Current Population Survey by the US Census Bureau (random sample drawn for Berndt 1991).

### Usage

```
data("CPS1985")
```

### Format

A data frame containing 534 observations on 11 variables.

**wage**  
Wage (in dollars per hour).

**education**  
Number of years of education.

**experience**  
Number of years of potential work experience (`age - education - 6`).

**age**  
Age in years.

Chaque colonne peut aussi se voir comme un vecteur

```
head(CPS1985$wage)
```

```
[1] 5.10 4.95 6.67 4.00 7.50 13.07
```

# Premiers éléments de statistique descriptive univariée

```
summary(CPS1985$wage)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	5.250	7.780	9.024	11.250	44.500

# Premiers éléments de statistique descriptive univariée

- ▶ On récupère ainsi la moyenne empirique ainsi que certains quantiles de la variable wage
- ▶ On dispose aussi de commandes plus directes
- ▶ Attention : ici rien ne prend en compte l'éventuelle pondération !
  - ▶ Intérêt de la pondération : sera abordée dans la séance suivante

# Premier éléments de statistique descriptive univariée

```
mean(CPS1985$wage)
```

```
[1] 9.024064
```

```
quantile(CPS1985$wage, probs = 0:5 * 0.2)
```

0%	20%	40%	60%	80%	100%
1.000	4.980	6.598	9.000	12.482	44.500



# C'est tout pour aujourd'hui

- ▶ Séance encore assez introductive :
  - ▶ Balayage rapide de ce que peut être un travail empirique quantitatif en économie
  - ▶ Taxonomie des données quantitatives
  - ▶ Éléments très basiques de statistique descriptive univariée
- ▶ La semaine prochaine : les choses sérieuses commencent
  - ▶ Retour sur quelques éléments de probabilité nécessaire pour manipuler des concepts de statistique correctement construits
  - ▶ En particulier pour donner un sens à l'incertitude sur les mesures que nous faisons