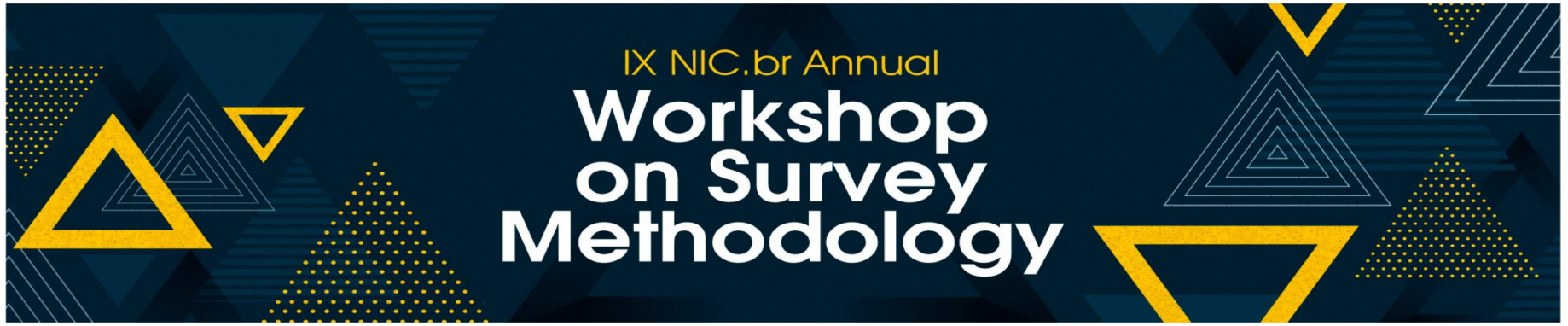


Data production for decision making

Iñigo Herguera
Universidad Complutense de Madrid
inigo@ccee.ucm.es

I shall talk about

1- two examples of a “traditional” data base (survey) exploitation to find out determinants of adopting a digital service

2- an example of crowdsourcing data and its bias

Example 1: determinants of adopting a digital service

DATA “traditional”: national survey on households for ICT services

- Survey on Equipment and Use of Information and Communication Technologies in Households **(ICT-H Survey) from 2007 to 2017**. **Spanish National Statistical Institute** (INE).
- Methodology of **Eurostat**.
- **196,974 observations; 73,651 individuals**.
- Includes elevation factor (weighting of individuals).

ICT variables 2016: equipment used and use of services

		variables	
HOGAR	Household characteristics	22	108
	Terminals/ gadgets used	10	
	Internet access	16	
	kids (10-15 y-o): use of PC and use of internet	60	
INDIVIDUO	Socio-economic variables	23	149
	Uses of internet	44	
	E-administration	10	
	Trust, security and privacy	17	
	Computer skills	10	
	E-commerce	45	
	Total	257	

Logistic regression

$$Prob(adopting\ a\ service\ k\ by\ household\ i / given\ socioeconomic\ variables\ of\ i)$$

$$Prob(Y_k = 1\ given\ x_i) = P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

Y_k= adopt/ do not (1/0)

E-commerce

E-banking

E-government

…….

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 x_i$$ when pooled data $\{i\}$

$$L_{i,t} = \ln\left(\frac{P_{i,t}}{1 - P_{i,t}}\right) = \beta_{0,t} + \beta_{1,t}\, x_{i,t}$$ when panel data structure $\{i, t\}$

odds- ratio (ratio of likelihood)

Household specific observed variables (socio-economic) x_i:

GENRE: 1 if male; 0= female

AGE: 6 groupings: 16-24; ……..

EDUCATION: 4 intervals (number of years with formal education)

PC_SKILLS: with computers, 3 levels

INTERNET_SKILLS: 4 levels

INTERNET_TRUST: 3 trust levels as declared by respondent

HH INCOME: 5 groups (net available income/ month)

The results presented here are part of several publications and research project conducted together with: *Teodosio Pérez, Teresa Garín, Angel Valarezo* and *Rafael López,* Universidad Complutense de Madrid.

household **income**, does it influence the adoption decision?

TABLE 2. Odds Ratios estimates of logistic regressions for adoption of each service (Spain, 2016)									
	MODEL 1. ECOMMERCE			MODEL 2. EBANKING			MODEL 3. EGOVERNMENT		
	Odds ratios	z	p value	Odds ratios	z	p value	Odds ratios	z	p value
INCOME			0.000			0.000			
<900									
900/1599	1.26	1.79		1.66	4.23		---	---	---
1600/2499	1.75	4.21		2.16	6.07				
2500/2999	2.17	5.16		2.96	7.03				
3000 or more	2.59	6.06		2.53	5.64				

Trust in internet, needed?

TABLE 2. Odds Ratios estimates of logistic regressions for adoption of each service (Spain, 2016)

	MODEL 1. ECOMMERCE			MODEL 2. EBANKING			MODEL 3. EGOVERNMENT		
	Odds ratios	z	p value	Odds ratios	z	p value	Odds ratios	z	p value
INTERNET_TRUST			0.000						
Low									
Medium	1.43	4.21		---	---	---	---	---	---
High	1.88	4.09							

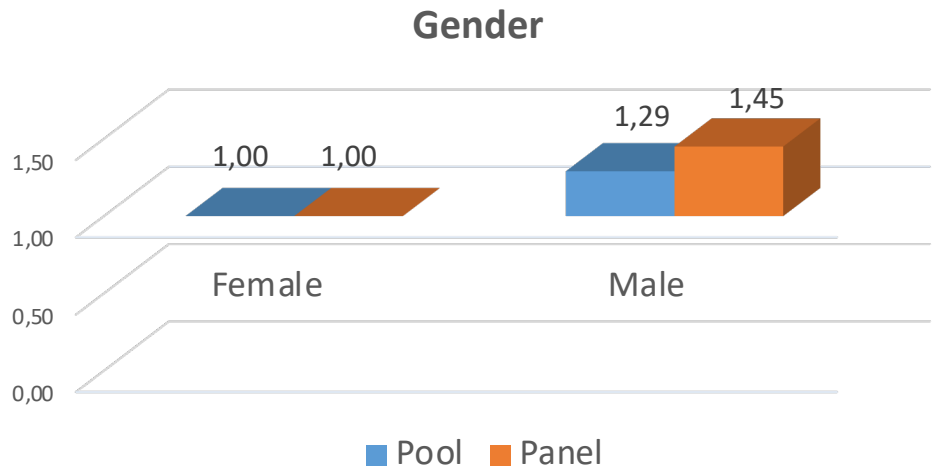
We have compared two estimation strategies for the adoption determinants of e-commerce in Spain:

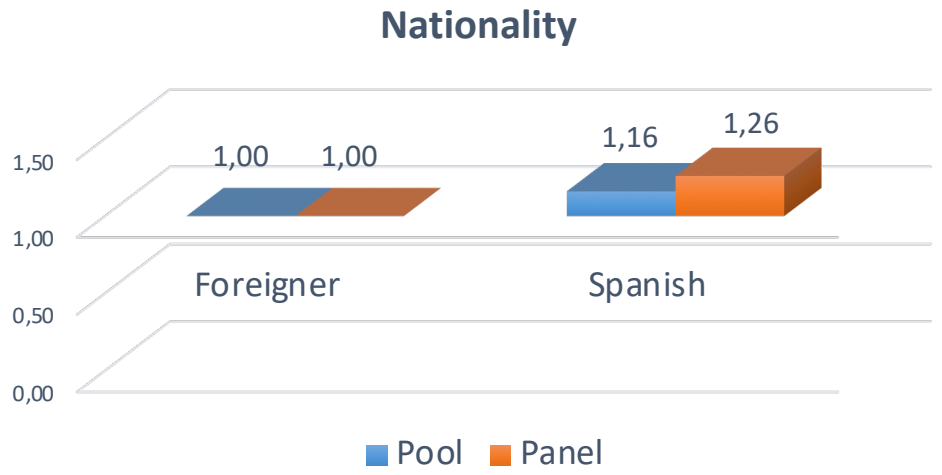
(1) **Pooled regression** (for each year)

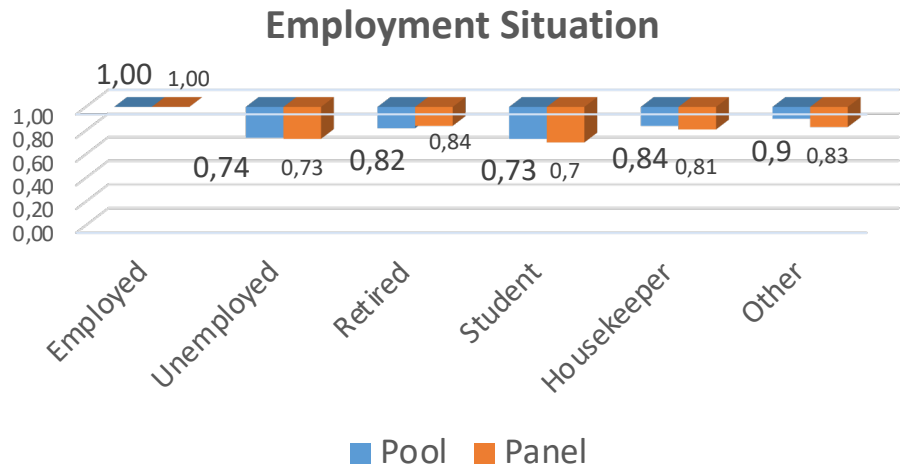
(2) **Panel regression**, 2007- 2017: to capture individual household decisions over time and changes in its socio- economic characteristics (dynamics)

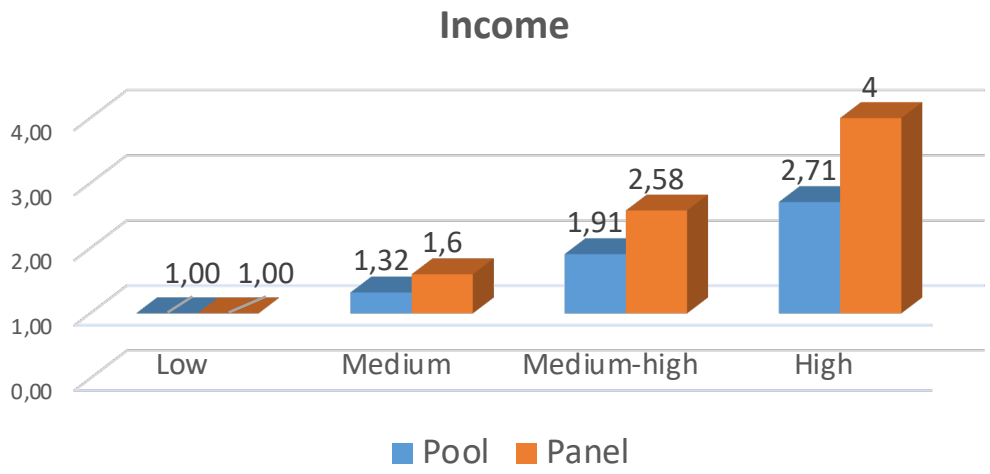
The results presented here are part of several publications and research project conducted together with: *Teodosio Pérez, Teresa Garín, Angel Valarezo* and *Rafael López,* Universidad Complutense de Madrid.

Odds ratios of Gender, Nationality, Employment Situation and Income on the decision of the **adoption of e-commerce**. **Pool** and **panel data** (2007–2017)

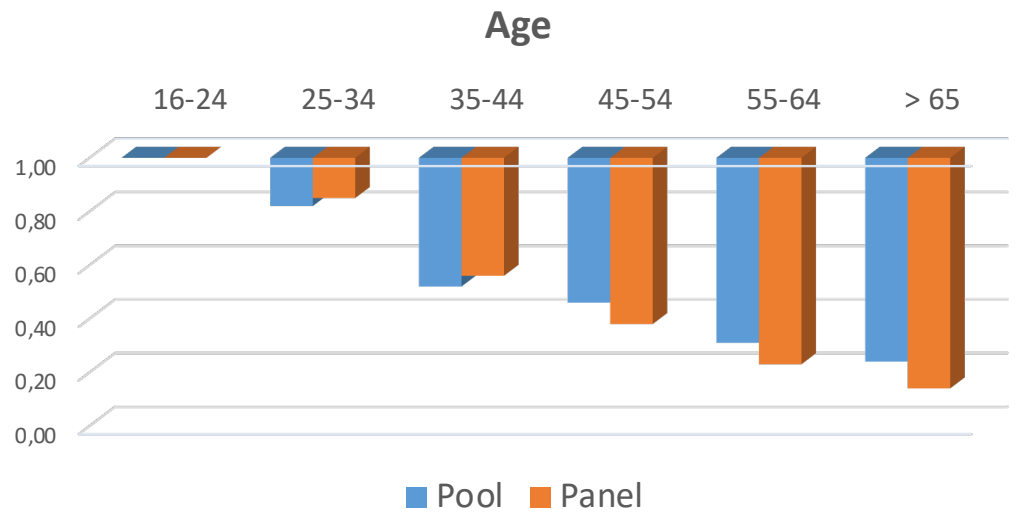


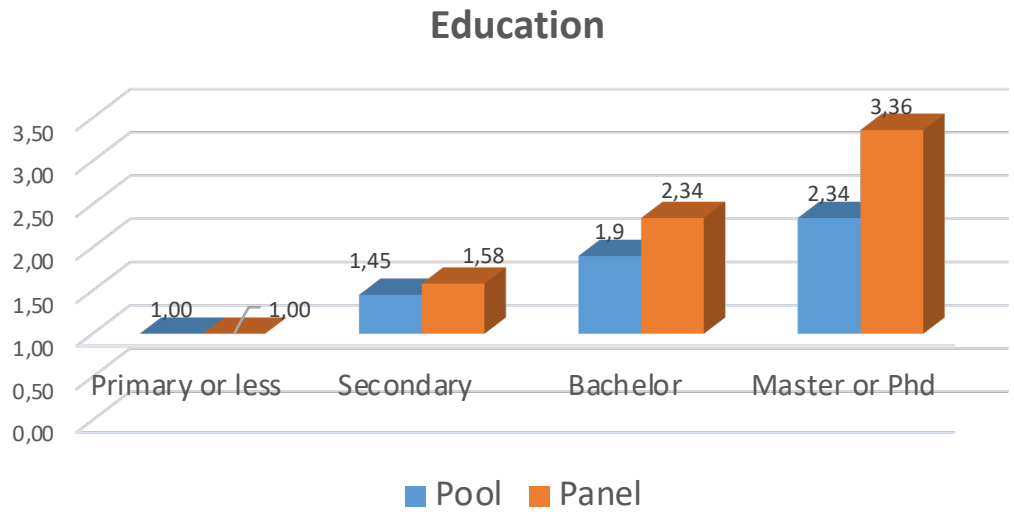


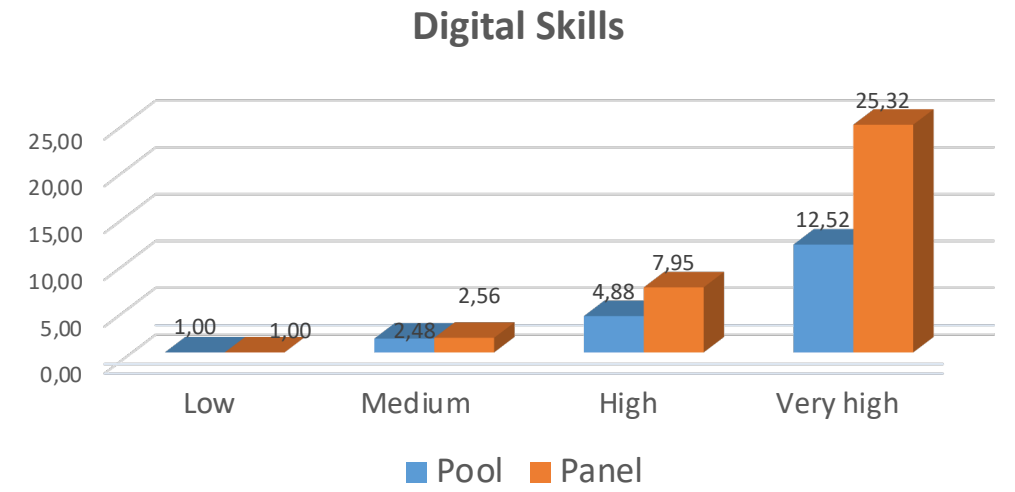


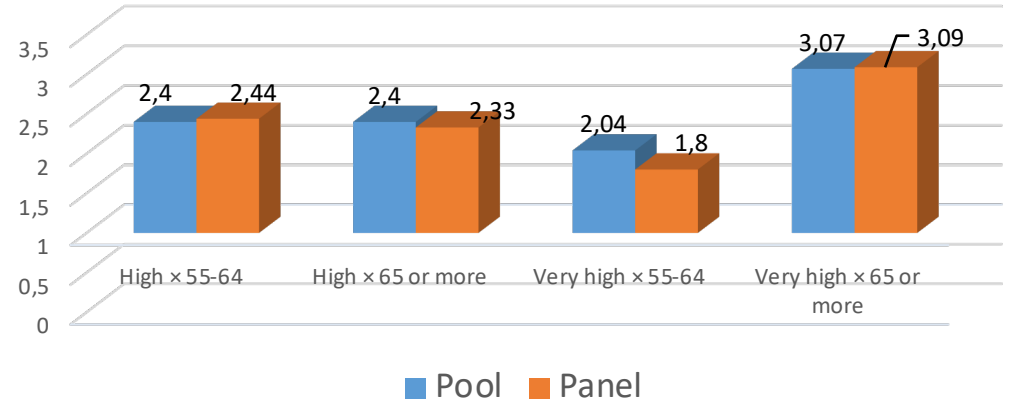


Odds ratios of Gender, Nationality, Employment Situation and Income on adoption of e-commerce. **Pool** and **panel data** (2007–2017)







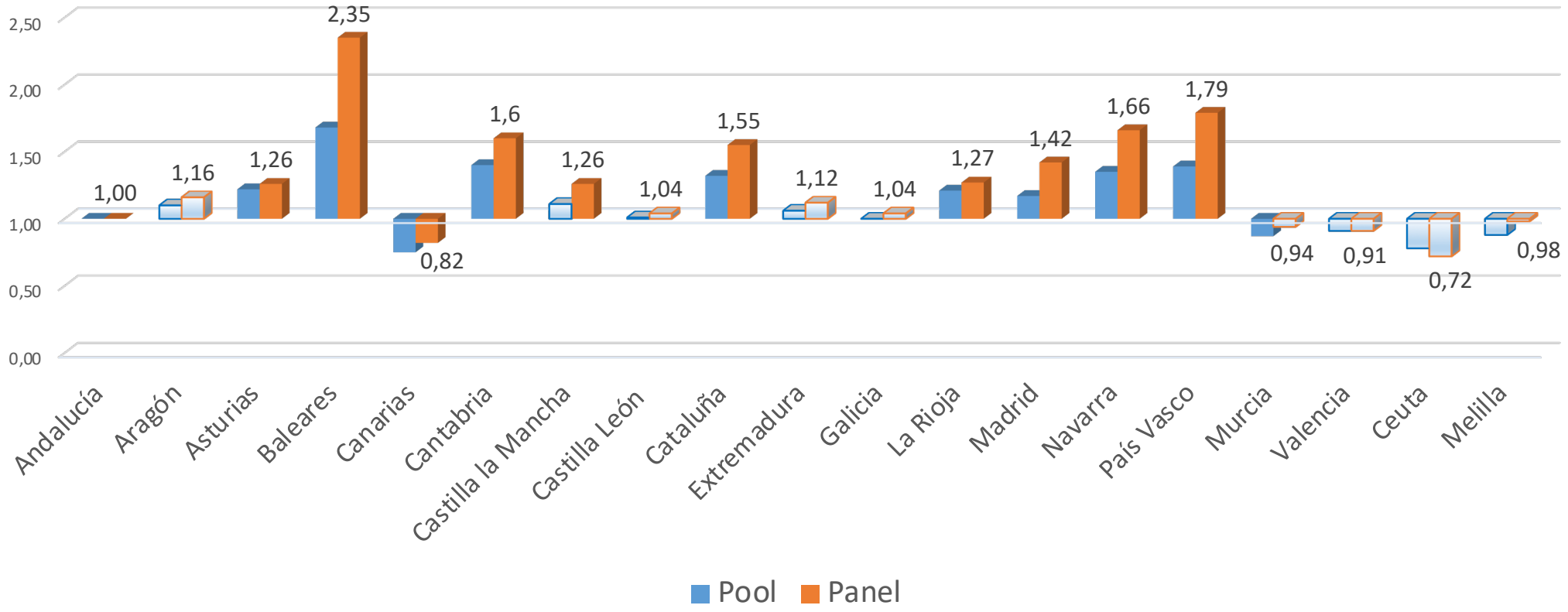


Thanks to introducing *dynamics* in this micro-data set we learned that:

-when estimating the adoption determinants in a panel data structure the main determinants for adoption **are even more relevant**: higher role for gender, income, age, digital skills and education

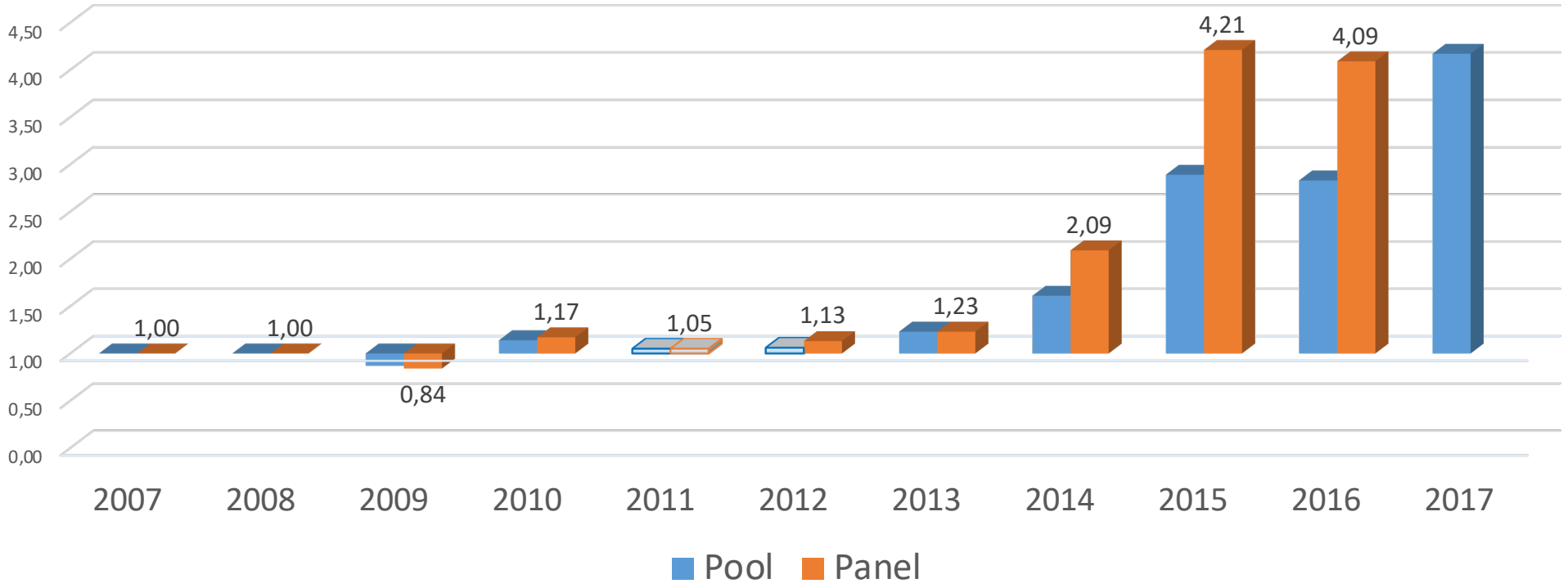
- **the more spread the adoption** of a digital service, ***the less relevant certain determinants appear to be***: age and gender seem of less important (hinting to network effects- not yet controlled for here!)

Odds ratios of Autonomous Communities on adoption of e-commerce. Panel data (2007–2016)- Reference region: Andalucia



Still over time big digital divides persist!!

Odds ratios of year dummies on the decision of the adoption of e-commerce. Panel data (2007-2016)



“supply side” effects do matter – even if not identified here!

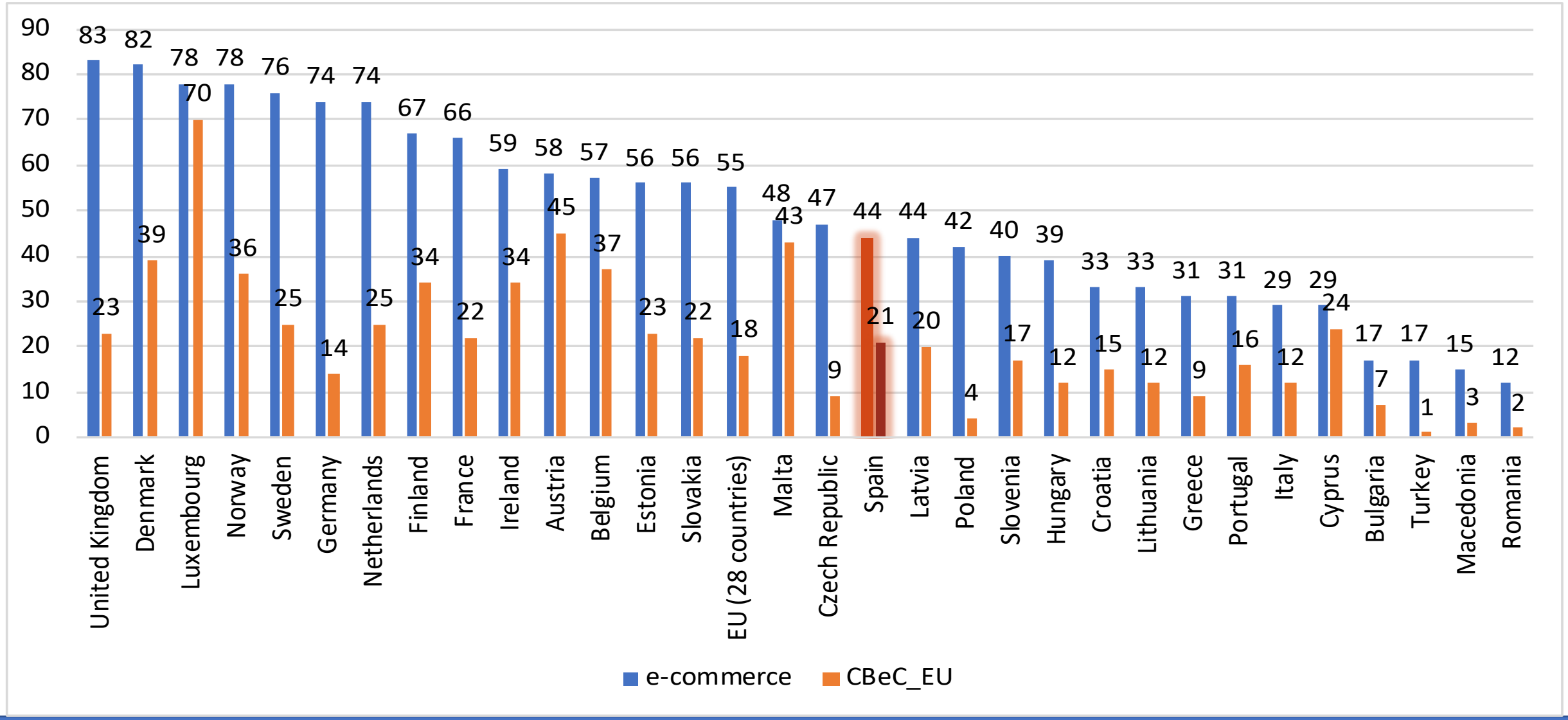
→ could other sources of data here be helpful (big/open data)?

Example 2: Trans-border e-commerce (in the European Union)

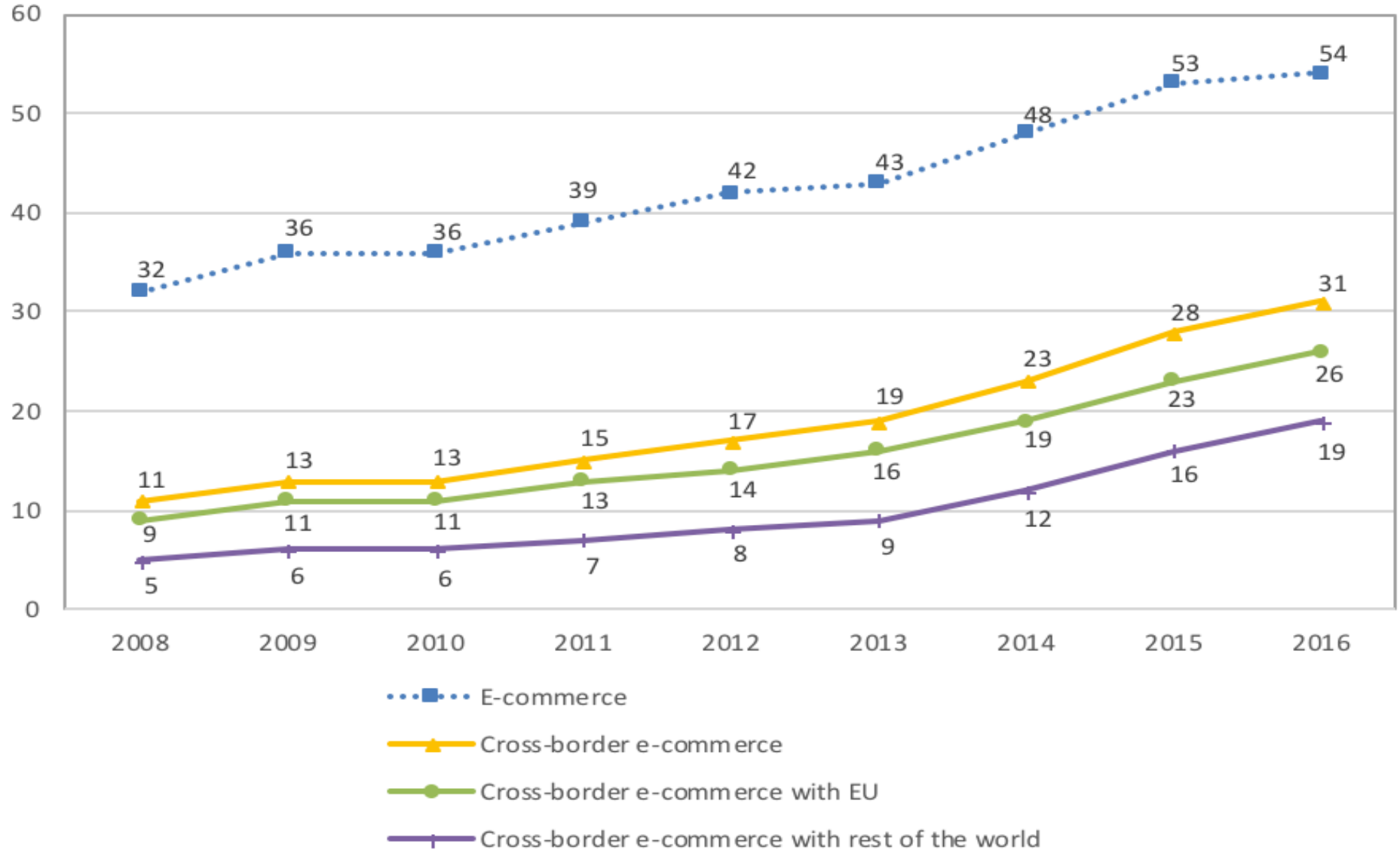
- In the EU 55% (or more) of citizens do e-commerce, but only 18% (on average) perform a transaction trans-border (with another EU country)

- How come?

Proportion of individuals (16- 74 y-o) that made e-commerce (total and with any other EU country)



Population 16- 74 y-o **that used internet and** used as well **e-commerce (domestic, in the EU or with Rest-of-the-World, RoW)** -descriptive statistics



Results

		Cross-border e-commerce		Cross border e-commerce with EU		Cross-border e-commerce with rest of the world	
		Odds ratios	z	Odds ratios	z	Odds ratios	z
GENDER	Female						
	Male	1.42	4.60	1.52	5.63	1.33	3.30
INTERNET TRUST	Low						
	Medium	1.20	2.06	--	--	--	--
	High	1.65	3.61	--	--	1.73	3.52
SEE REVIEWS	Hardly ever						
	Sometimes	1.65	3.75	1.90	4.58	1.41	1.98
	Always	1.97	5.78	2.12	6.15	2.10	4.81

N. observations	5,576	5,576	5,587
Wald χ^2	318.86 DF: 20	336.52 DF: 13	190.27 DF: 15
Pseudo R^2	0.0745	0.0817	0.0632
Correctly classified	63.77%	63.18%	69.09%

Results

- *age*, *income, education level* and *trust* in internet **do not seem to be significant determinants** in explaining volume of trans-border e-commerce

  - They have been already incorporated in previous decision ("having made e-commerce")

  - Some multicollinearity is possible between **age** and **income**

What is new?

- **“do reviews”** and the **accomplishment of information** of the sites where to buy matters for all types of cross-border e-commerce (as a substitute for “trust on the internet”?)

- the **nationality (of buyer) matters-** even if this was not indentifiable with simple statistics

- **skills with computers** and with **internet** matter (even if they were as well picked up in previous decision “make e-commerce”)

Example 3: Non- probabilistic samples

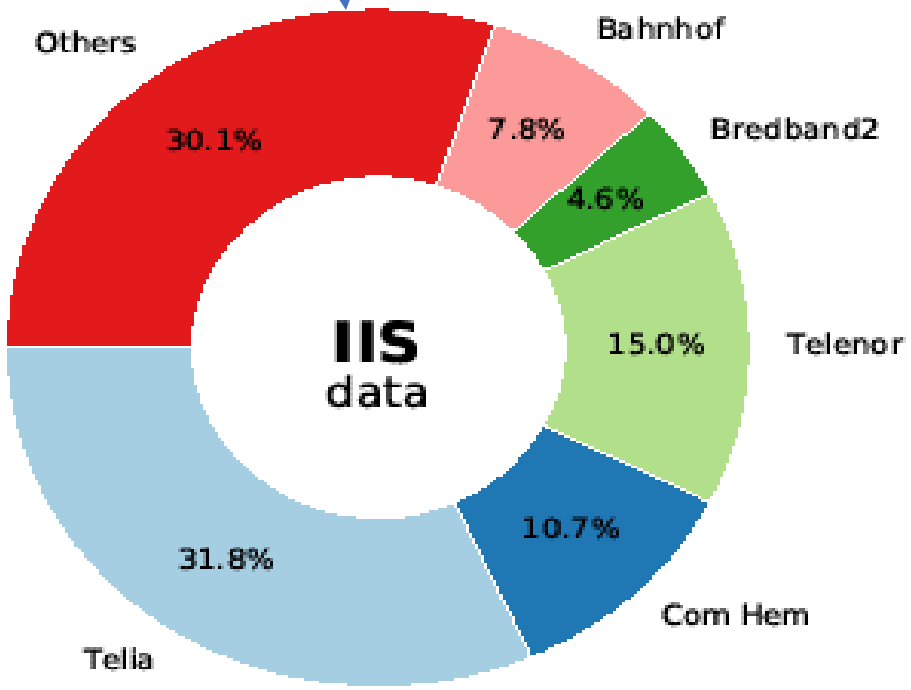
- Problems: ***selection bias*** and ***unequal participation***

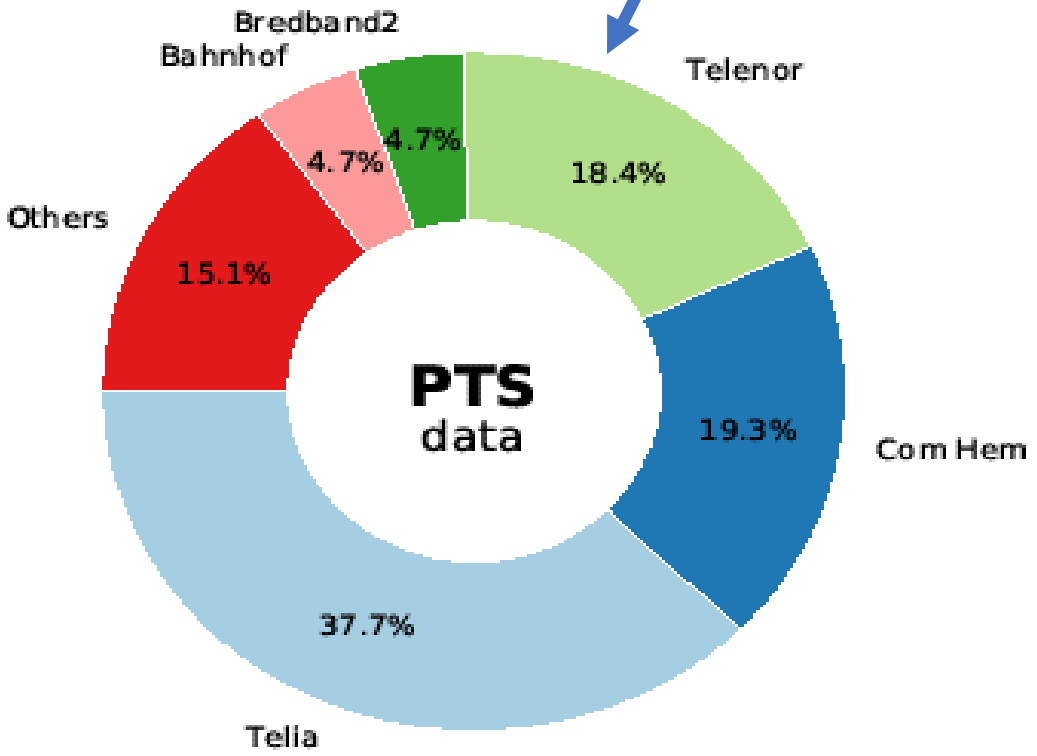
- Ways forward: (1) use of “ground truth” and calibrate test
  (2) Differences- in- Differences (DiD) approach

This example is based on the work conducted by **Ivan Vallejo** “Measuring real broadband speeds using crowdsourcing data from the Internet Foundation”, Master Thesis in the *Data Sciende Program*, Universidad Pompeu Fabra (Spain), June 2017

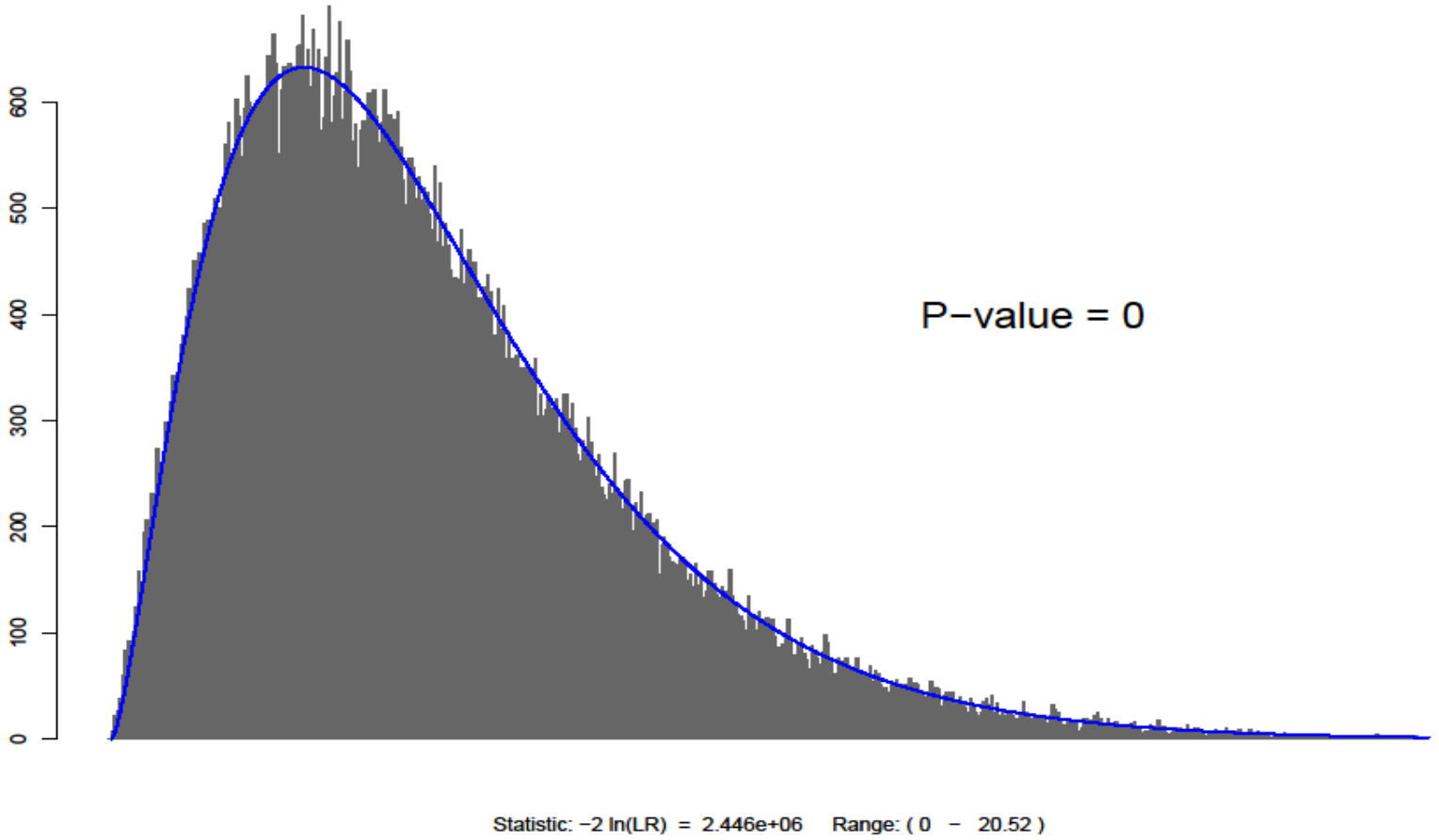
use of **observed (micro-data) from big data source** and the **“ground truth”**

(from regulatory authority): market shares of broadband lines in Sweden by operator





Selection bias: comparing the given distribution of market shares in raw data, with the “true” distribution given by administrative data (Regulatory Authority)



As a way of concluding…….

- Big data sources (non- probabilistic sampling) and “traditional” data sources (official) can be used as *complementary* methods

- Some issues (i.e., supply side effects) cannot be controlled for in demand- based surveys

- “traditional” data sets allow for conducting static and dynamic analysis (cross- section vs. panel estimation)

- Representativeness of sampled population is critical!