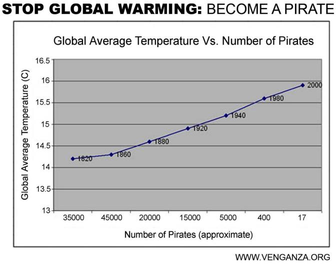
Correlation



Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 8

Outline

Today is almost entirely scatterplots and correlation coefficients. We'll make fun of "correlation looks like causation" issues in due time.

- Operations
- Patterns
- Issues

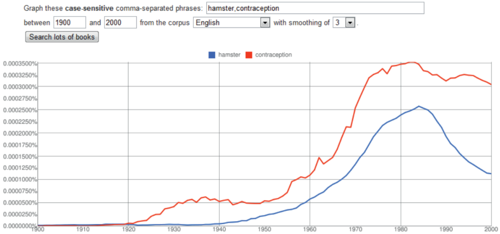


Figure 1: Frequencies of the words "hamster" and "contraception" in Google Books, 1900–2000

Getting there

- **Description**, through univariate statistics, establishes the distributional characteristics of your data.
  - Measures of central tendency and spread
  - Normality and variable transformations
- **Association**, through significance tests, establishes the basic framework of comparison between variables.
  - Comparison of groups (Chi-squared test)
  - Comparison of means (*t*-test) or proportions

Correlation is different in that it provides the strength and directionality of bivariate relationships:

- Correlation comes with a **correlation coefficient** that indicates how strong the correlation is.
- Correlation also comes with a **significance test** to indicate whether H_0 (no relationship) can be rejected.

Getting through

- **Visualize** bivariate relationships with scatterplots:
  - `sc` generates individual scatterplots
  - `gr mat` generates a scatterplot matrix
- **Assess** the the strength of visual relationships:
  - `pwcorr` provides a correlation matrix of 2+ variables
  - `corr` also computes a correlation coefficient but `pwcorr` also adjusts for missing data through pairwise case deletion (see `h corr` for help).
- **Interpret**
  - **Use a standard vocabulary:** "strong—weak" for strength, "positive/negative" for direction.
  - **Non-linear relationships can produce significant linear correlations.** If you reduce a non-linear pattern to a linear one (due to methodological constraints), mention it in your analysis.
  - **Do not assume causation.** You need theoretical grounds to support a correlation, however significant it is.

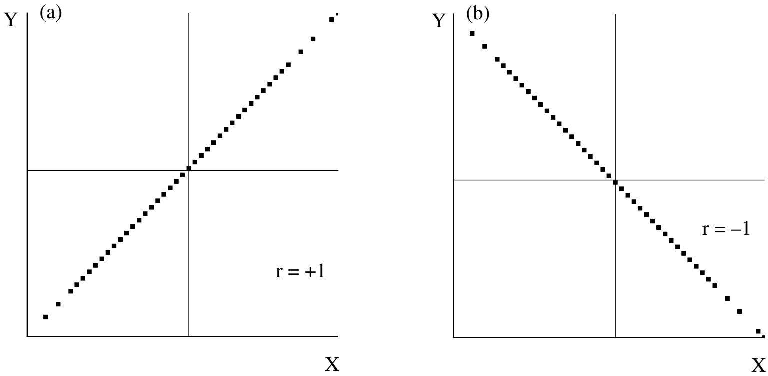
Getting the math

- **Pearson's r** determines the quality of a correlation:

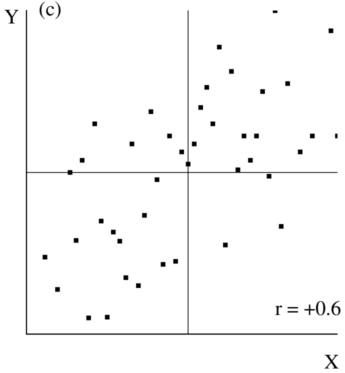
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

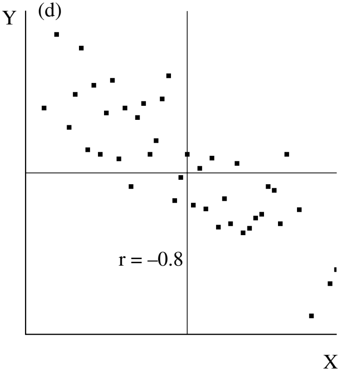
- **Reading guide:**
  - Pearson's r ranges from -1 to +1 and comes with a p-value.
  - -1 and +1 denote perfect negative and positive correlation.
  - The p-value for Pearson's r tests $H_0 : r = 0$ (no correlation).
- **Computation:**
  - $X_i - \bar{X}$ is the distance (or **residual**) between all observations $i_1, i_2, ..., i_n$ for variable X and the mean $\bar{X}$ of its distribution.
  - Pearson's r computes the **residual sum of squares** (RSS) and its product for variables X and Y.

Perfect positive/negative correlations

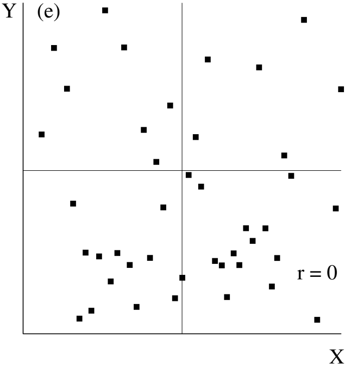


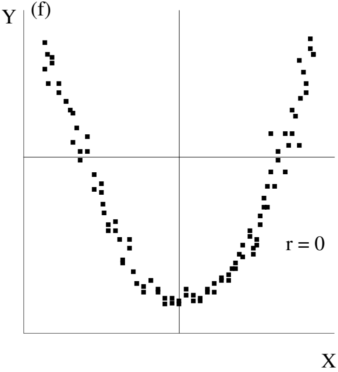
Significant (moderate–strong) correlations





Insignificant (weak/non-linear) correlations



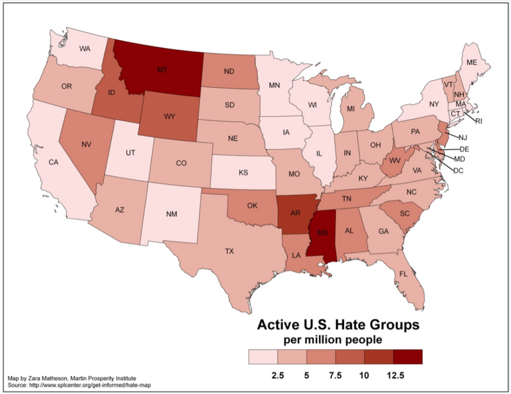


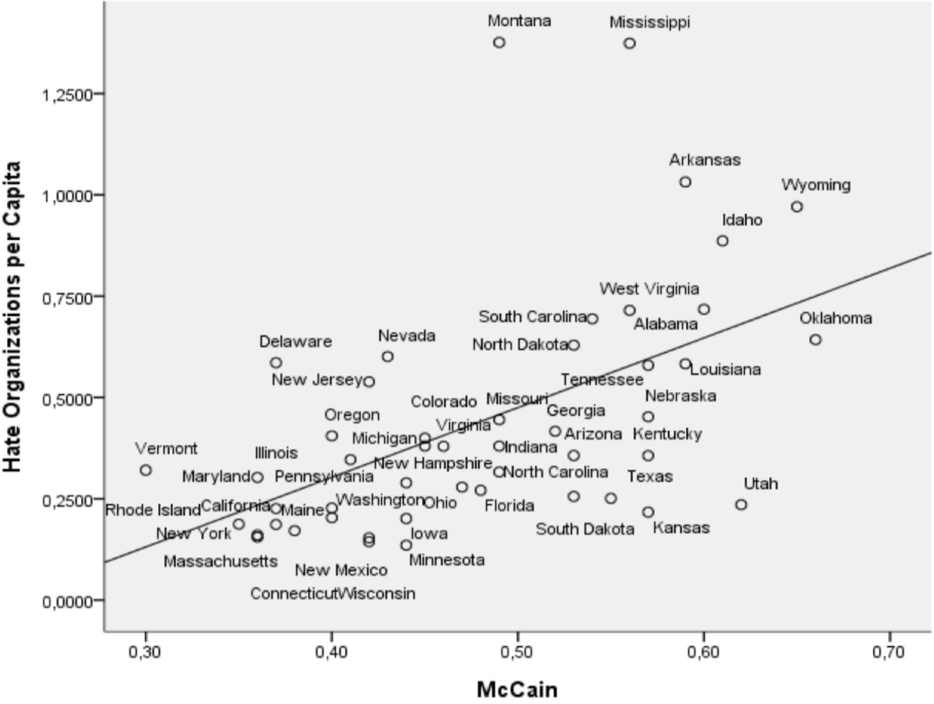
Issues

The main concern with correlation is whether you set it right in the first place: **what are your correlating, and why?**

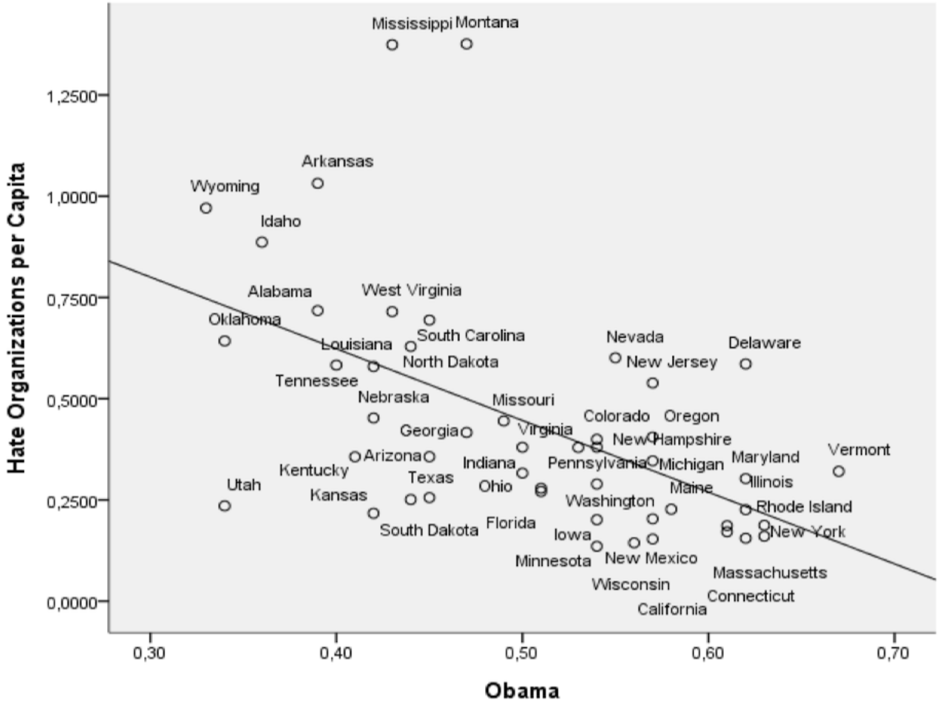
The next slides are from Richard Florida's "The Geography of Hate", *The Atlantic*, May 2011.

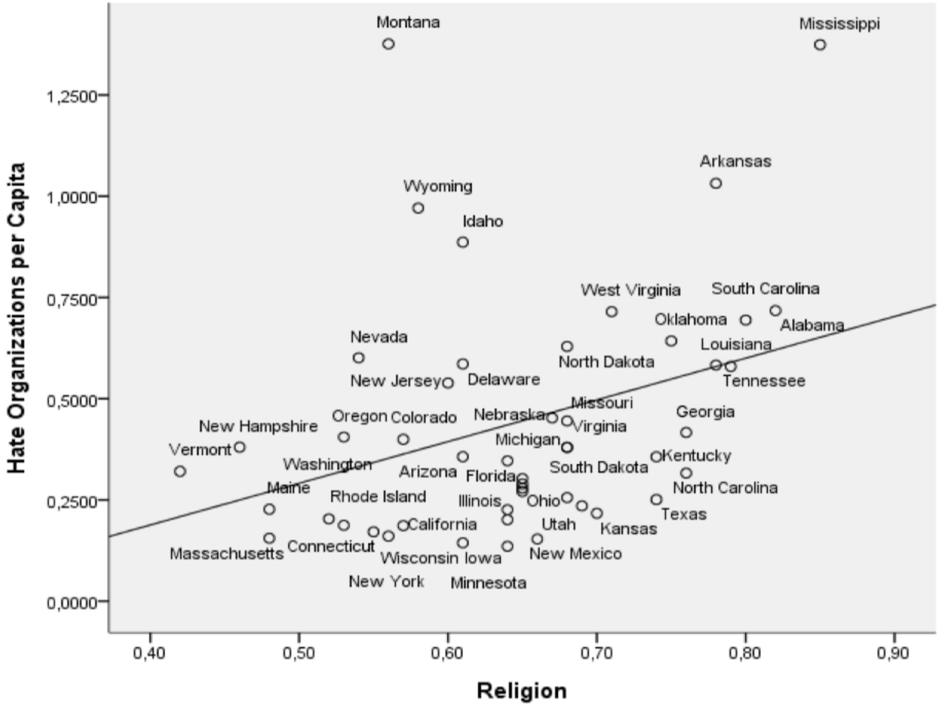
What do you **observe**?
What do you **infer**?
What do you **posit**?

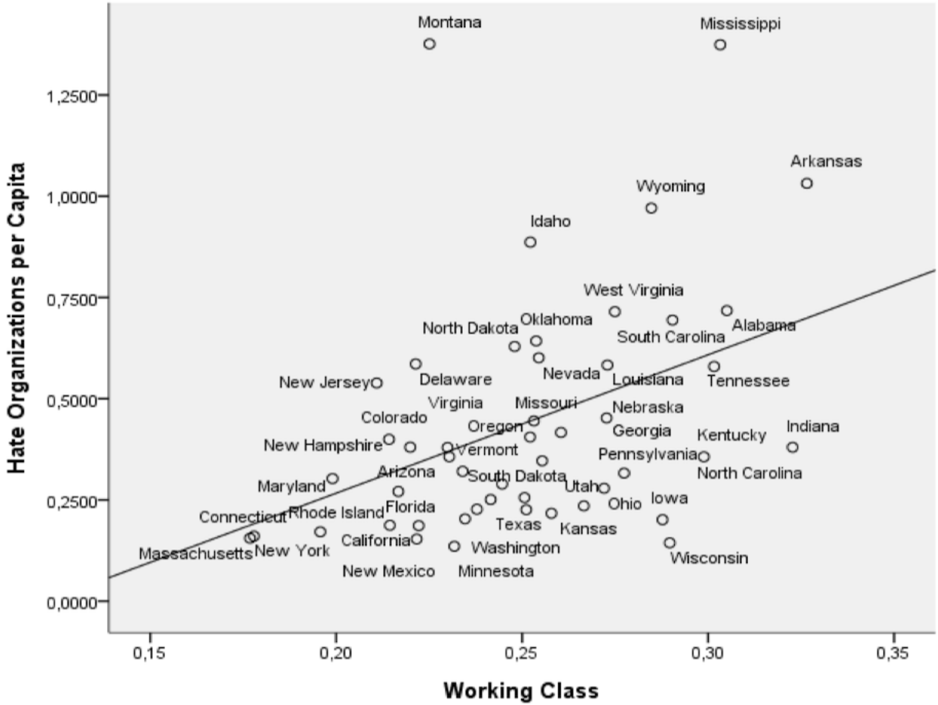




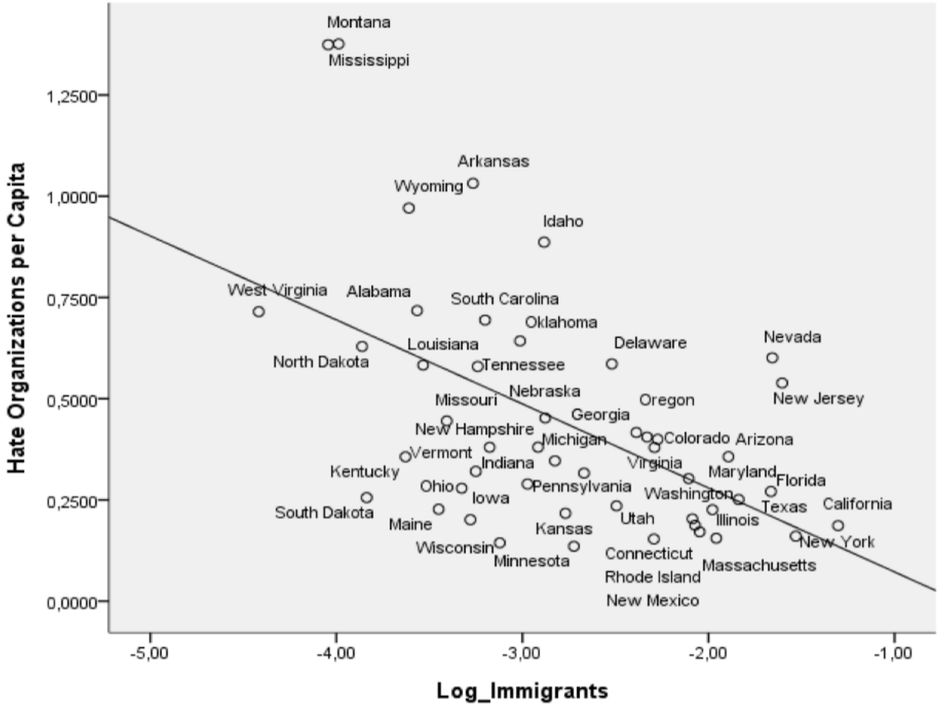
Hate Organizations per Capita
McCain
1,2500
1,0000
0,7500
0,5000
0,2500
0,0000
0,30
0,40
0,50
0,60
0,70
Montana
Mississippi
Arkansas
Wyoming
Idaho
West Virginia
South Carolina
Alabama
Oklahoma
Delaware
Nevada
North Dakota
Louisiana
New Jersey
Tennessee
Colorado
Missouri
Nebraska
Georgia
Oregon
Virginia
Arizona
Kentucky
Michigan
Vermont
Illinois
Indiana
New Hampshire
North Carolina
Maryland
Pennsylvania
Texas
Utah
Washington
Ohio
Florida
Rhode Island
California
Maine
South Dakota
Kansas
New York
Iowa
Massachusetts
Minnesota
New Mexico
Connecticut
Wisconsin


Hate Organizations per Capita
Obama
0,30
0,40
0,50
0,60
0,70
0,0000
0,2500
0,5000
0,7500
1,0000
1,2500
Mississippi
Montana
Arkansas
Wyoming
Idaho
Alabama
West Virginia
Oklahoma
South Carolina
Louisiana
North Dakota
Tennessee
Nevada
New Jersey
Delaware
Nebraska
Missouri
Colorado
Oregon
Georgia
Virginia
New Hampshire
Arizona
Kentucky
Indiana
Pennsylvania
Michigan
Maryland
Vermont
Utah
Texas
Ohio
Illinois
Kansas
Maine
Washington
Rhode Island
South Dakota
Florida
Iowa
New York
Minnesota
New Mexico
Massachusetts
Wisconsin
Connecticut
California


Hate Organizations per Capita
Religion
1,2500
1,0000
0,7500
0,5000
0,2500
0,0000
0,40
0,50
0,60
0,70
0,80
0,90
Montana
Mississippi
Arkansas
Wyoming
Idaho
West Virginia
South Carolina
Oklahoma
Alabama
Nevada
Louisiana
North Dakota
Delaware
New Jersey
Tennessee
Missouri
Oregon
Colorado
Nebraska
Georgia
New Hampshire
Virginia
Michigan
Vermont
Kentucky
Washington
Arizona
South Dakota
Florida
North Carolina
Maine
Rhode Island
Illinois
Ohio
Texas
California
Utah
Kansas
Massachusetts
Connecticut
Wisconsin
Iowa
New Mexico
New York
Minnesota



Hate Organizations per Capita
Working Class
0,0000
0,2500
0,5000
0,7500
1,0000
1,2500
0,15
0,20
0,25
0,30
0,35
Montana
Mississippi
Arkansas
Wyoming
Idaho
West Virginia
Oklahoma
North Dakota
South Carolina
Alabama
New Jersey
Delaware
Nevada
Louisiana
Tennessee
Virginia
Missouri
Nebraska
Colorado
Oregon
Georgia
Kentucky
Indiana
New Hampshire
Vermont
Pennsylvania
North Carolina
Arizona
South Dakota
Maryland
Utah
Iowa
Ohio
Connecticut
Rhode Island
Florida
Texas
Kansas
California
Massachusetts
New York
Washington
Wisconsin
New Mexico
Minnesota



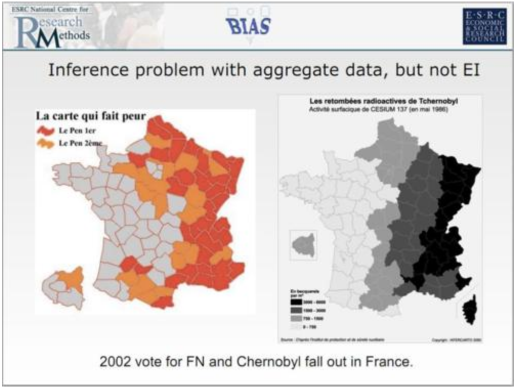
Montana
Mississippi
Arkansas
Wyoming
Idaho
West Virginia
Alabama
South Carolina
Oklahoma
North Dakota
Louisiana
Tennessee
Delaware
Nevada
New Jersey
Missouri
Nebraska
Oregon
Georgia
New Hampshire
Michigan
Colorado
Arizona
Vermont
Indiana
Virginia
Kentucky
Maryland
Florida
Ohio
Pennsylvania
Washington
Iowa
South Dakota
Texas
California
Maine
Utah
Illinois
Kansas
Wisconsin
New York
Minnesota
Connecticut
Massachusetts
Rhode Island
New Mexico
1,2500
1,0000
0,7500
0,5000
0,2500
0,0000
-5,00
-4,00
-3,00
-2,00
-1,00
Hate Organizations per Capita
Log_Immigrants

Inference

The main concern with correlation is **inference**: not everything that correlates is causally related at the right level of observation.

Your interpretive skills are put once more to the test.

What do you **observe**?
What do you **infer**?
What do you **posit**?



2002 vote for FN and Chernobyl fall out in France.