Dependence

1. Comparison
2. Chi-squared test
3. *t*-test
4. Correlation

Statistical comparison

Substantive hypotheses

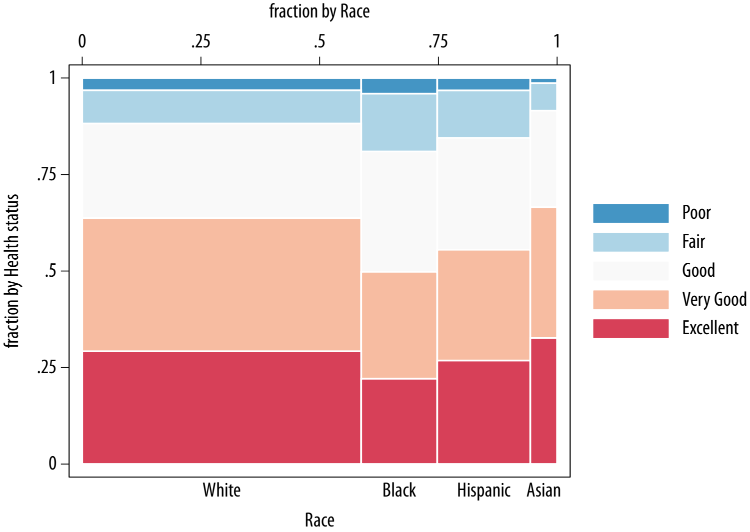
There is an association between X and Y, ...
There is a difference of X between groups of Y, ...

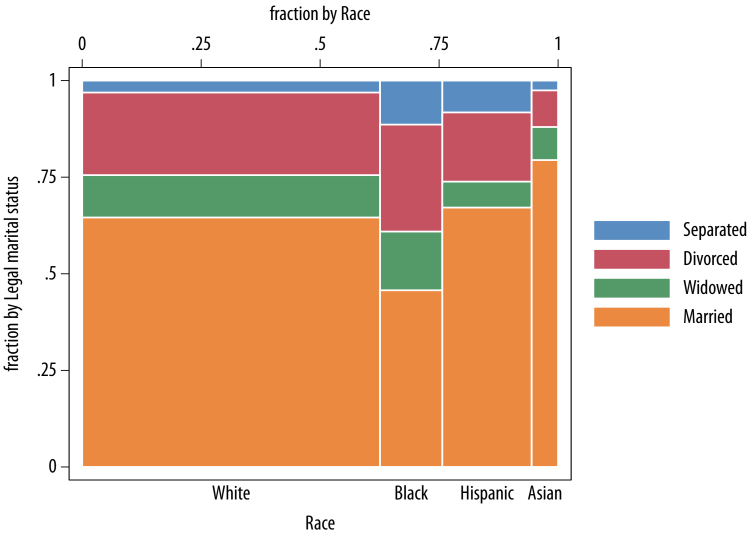
Null hypothesis tests

H_0: the association of X by Y is likely to be random.
H_0: the difference in X between groups of Y is likely to be random.

Rejecting the null

H_0 estimates the likelihood of an association or difference being attributable to sampling error under a certain level of confidence.


fraction by Race
0
.25
.5
.75
1
fraction by Health status
0
.25
.5
.75
1
White
Black
Hispanic
Asian
Race
Poor
Fair
Good
Very Good
Excellent


fraction by Race
0
.25
.5
.75
1
fraction by Legal marital status
1
.75
.5
.25
0
White
Black
Hispanic
Asian
Race
Separated
Divorced
Widowed
Married

Chi-squared test

The Chi-squared test is a nonparametric test of association that measures the deviation in orthogonality between groups:

- Null hypothesis H_0: $\chi^2 = 0$
- Test statistic: $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ (deviation between observed frequencies O_i and expected frequencies E_i for each table cell i)

```
tab v1 v2, exp chi2 V
```

- add `V` to measure the association with Cramér's V ($0 < V < 1$)
- use `tabchi` to inspect residuals and `tabodds` for odds ratios

use datasets/nhis2009, clear

- Variables: `d raceb marstat`
- Inspect frequencies (row and column, expected and observed)
- Run a Chi-squared test and analyze the residuals

```
. tab marstat raceb if marstat < 8, chi2 V
```

Legal marital status	Race: White	Black	Hispanic	Asian	Total
Married	7,151	1,059	2,231	780	11,221
Widowed	1,215	352	223	84	1,874
Divorced	2,367	641	595	93	3,696
Separated	343	264	274	25	906
Total	11,076	2,316	3,323	982	17,697

```
Pearson chi2(9) = 733.4437   Pr = 0.000
     Cramér's V =   0.1175
```

t-test

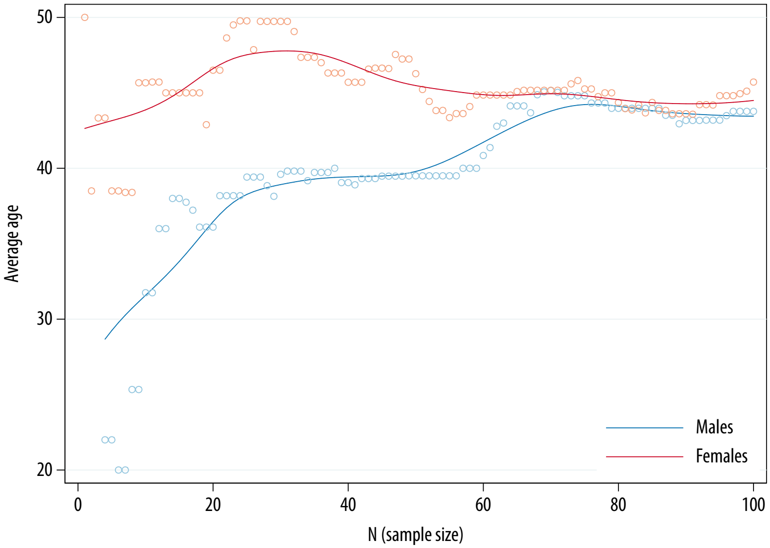
Measuring association as the difference in means between two groups:

- Population notation: $\delta = \mu_1 - \mu_2$
- Sample notation: $D = \bar{X}_1 - \bar{X}_2$

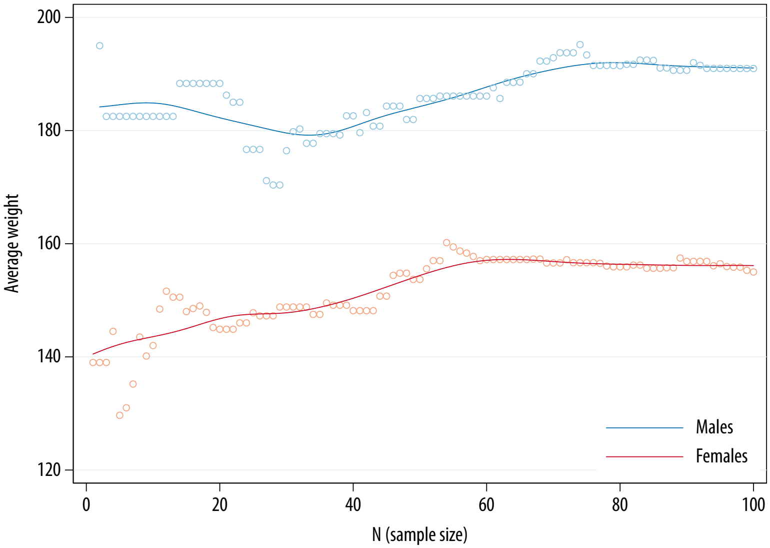
The t-test computes a 95% CI around the difference of their means and returns its p-value against the t-distribution.

- Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
- Test statistic: $t = \frac{D}{SE_D}$

Possible Type I errors



Possible Type II errors



Stata implementation

```
ttest v1, by(v2)
```

- v1 is continuous, v2 is a dummy; for two dummies, use `prtest`
- use `tab, gen()` to create dummies from categorical variables

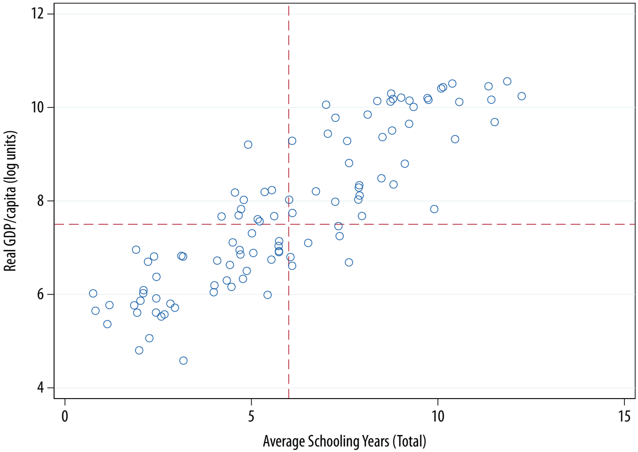
```
use datasets/qog2011, clear
```

- Run a proportions test: `prtest no_mes, by(gol_polreg)`
- Explore the variables and dissect the output.

```
use datasets/qog2011, clear
```

- Variables: `d gol_enep gol_est2`
- Create dummies and compare parties across electoral systems.

Correlation



Pearson correlation coefficient

Measuring association as the linear dependence of two variables:

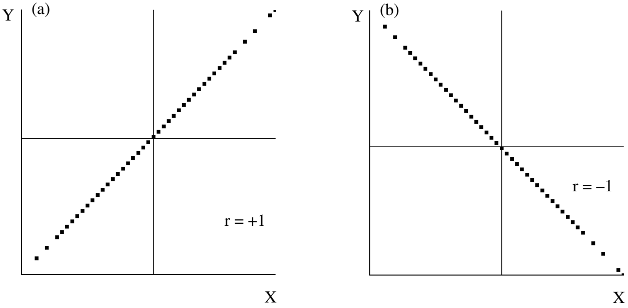
$$\text{Population notation} \quad \rho = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}_X \mathrm{Var}_Y}, \quad -1 \leq \rho \leq 1$$

$$\text{Sample notation} \quad r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

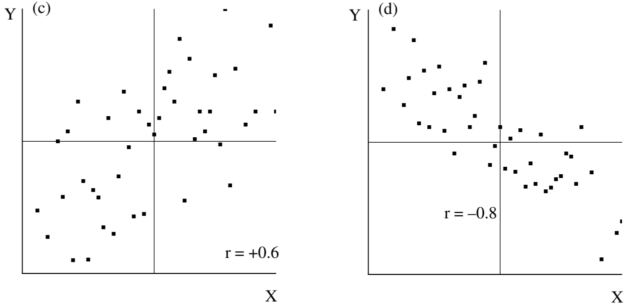
Detects linear correlation

- Uncorrelated $\neq$ unrelated
- Correlated $\neq$ unconfounded

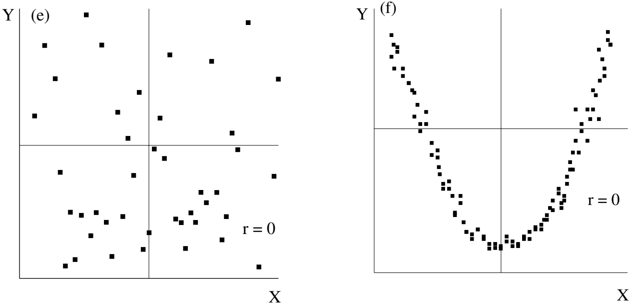
Perfect positive/negative correlations



Significant (moderate–strong) correlations



Insignificant (weak/non-linear) correlations



Pearson correlation coefficient

Significance test:

$$\text{Null hypothesis } H_0 \quad r = 0$$

$$\text{Test statistic} \quad T = r\sqrt{\frac{n-2}{1-r^2}}$$

Significance sanity check

- Uncorrelated $\neq$ independent
- Correlated $\neq$ causally related

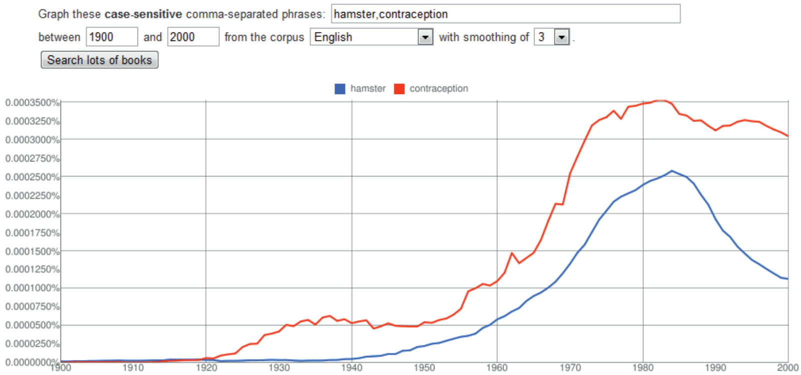


Figure 1: Frequencies of the words “hamster” and “contraception” in Google Books, 1900–2000

Source: Harkness, “Seduced by Stats?”, *Significance*, 2012.

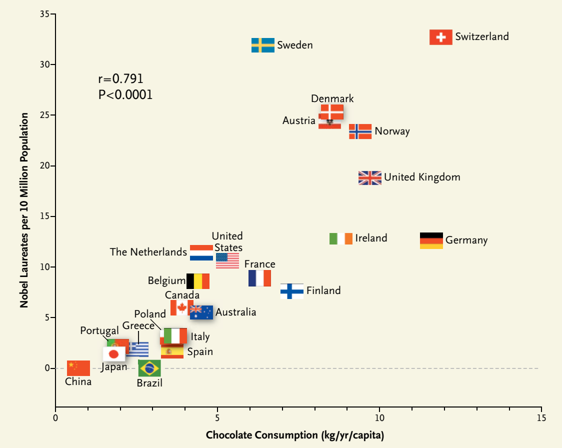


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Source: Messerli, "Chocolate Consumption, Cognitive Function, and Nobel Laureates", *New England Journal of Medicine*, 2012.

Stata implementation

`pwcorr [varlist], [obs sig]`

- `obs` shows the number of observations
- `sig` shows the coefficient's *p*-value

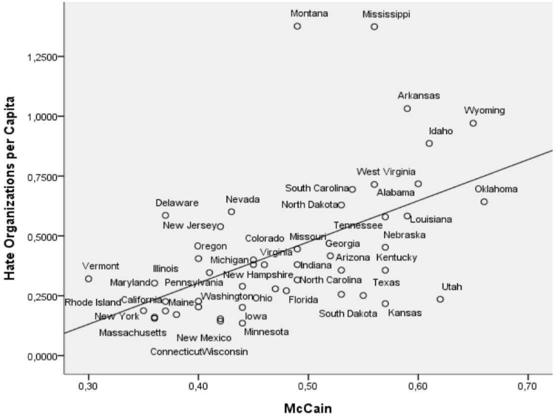
`gr mat [varlist], [half etc.]`

- `half` plots only half of all graphs (quicker)
- accepts scatterplot options (`jitter`, `mlab`, etc.)

```
use datasets/qog2011, clear
```

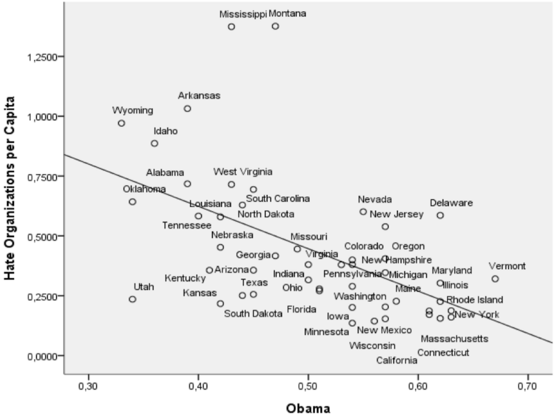
- Variables: `d wdi_brd wdi_mege wdi_pb2 wdi_the`
- Inspect and plot the correlation matrix.

One variable, many predictors



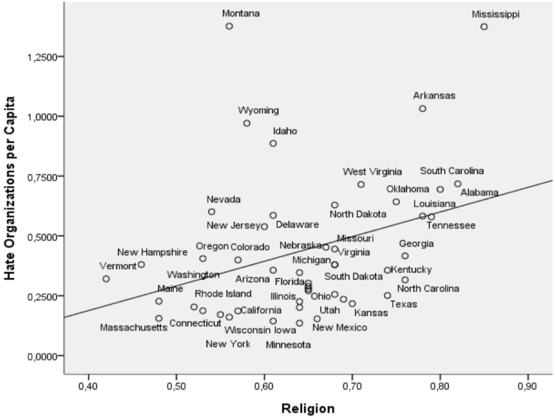
Source: Florida, “The Geography of Hate”, *The Atlantic*, 2011.

One variable, many predictors



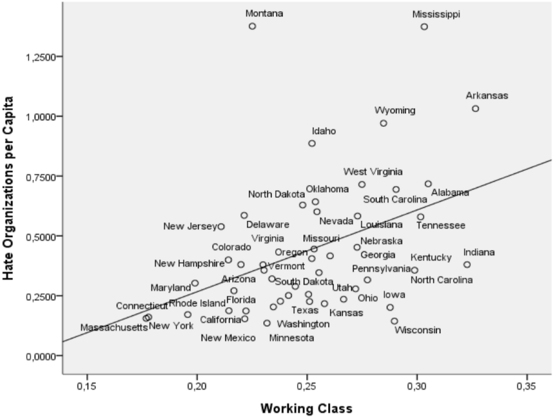
Source: Florida, “The Geography of Hate”, *The Atlantic*, 2011.

One variable, many predictors



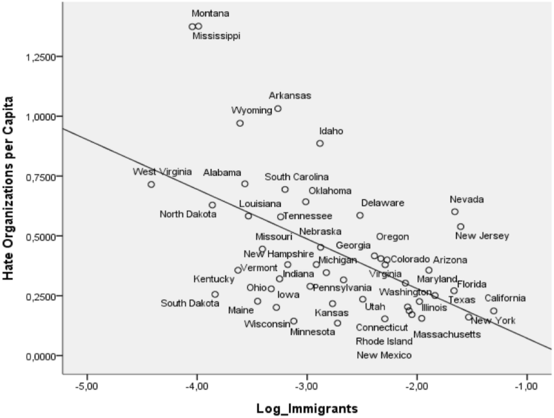
Source: Florida, "The Geography of Hate", *The Atlantic*, 2011.

One variable, many predictors

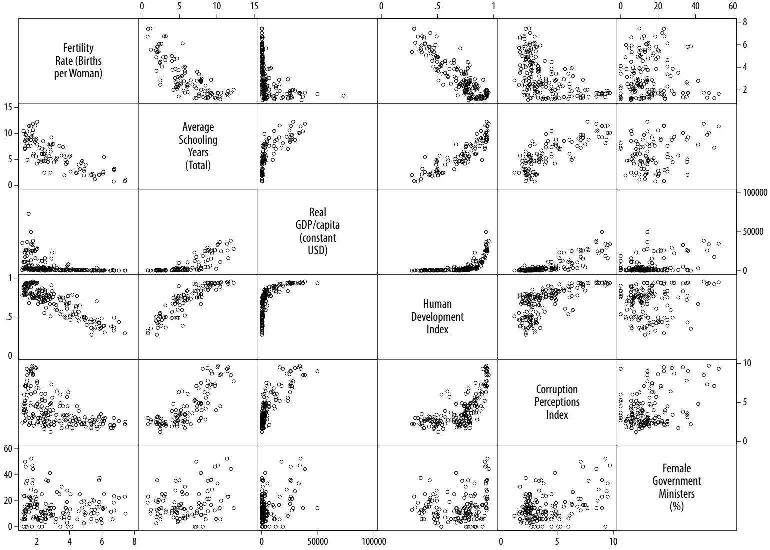


Source: Florida, "The Geography of Hate", *The Atlantic*, 2011.

One variable, many predictors



Source: Florida, "The Geography of Hate", *The Atlantic*, 2011.



Fertility Rate (Births per Woman)
Average Schooling Years (Total)
Real GDP/capita (constant USD)
Human Development Index
Corruption Perceptions Index
Female Government Ministers (%)

Table 4
Pearson pairwise correlations among the dependent and explanatory variables

	ETRC	ETRI	CAPINT	LEV	SIZE	POLCON1	POLCON2	MKBV	INVINT	ROA
ETRC	1									
ETRI	0.031*	1								
CAPINT	−0.033**	−0.044**	1							
LEV	−0.051*	−0.021	−0.041**	1						
SIZE	−0.124	−0.190	−0.163**	0.337**	1					
POLCON1	−0.023**	−0.047**	0.129	0.031**	0.146**	1				
POLCON2	−0.011*	−0.044*	−0.064	0.116	0.179**	0.138**	1			
MKBV	0.045	−0.036	−0.051	−0.035	−0.077**	−0.130	−0.026	1		
INVINT	0.020	−0.014	0.067**	−0.128**	−0.195**	0.193**	−0.005	−0.041	1	
ROA	0.073*	0.047*	0.067**	−0.038	0.073	0.049	0.012	0.053	−0.019	1

Variable definitions: ETRC = (Tax expenses − Deferred tax expenses)/(Operating cash flows); ETRI = (Tax expenses − Deferred tax expenses)/(Profit before interest and tax); POLCON1 = Percentage of government equity ownership; POLCON2 = 1 if the firm is connected with top politicians; 0 otherwise; SIZE = Natural log of total assets; LEV = (Total debt)/(Total assets); CAPINT = (Property, plant and equipment)/(Total assets); INVINT = (Inventory/Total assets); ROA = (Pre-tax profits)/(Total assets); MKBV = (Market price of share)/(Shareholders equity/Number of ordinary shares outstanding).

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Source: Adhikari *et al.*, *Journal of Accounting and Public Policy*, 2006.

Correlation matrixes

`mkcorr [varlist], lab num sig log(corr.txt) replace`

- `ssc install` the command if needed
- `lab num sig` add labels, numbers and *p*-values

Computer skills

- Import as a table in a spreadsheet editor.
- Convert from text to table in a rich text editor.

`use datasets/qog2011, clear`

- Variables: `d wdi_brd wdi_mege wdi_pb2 wdi_the`
- Export and import the correlation matrix.