

Association (III)

- 1 Statistical tests
- 2 t -test
- 3 Chi-squared test
- 4 Correlation

Statistical tests

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**.

t -test

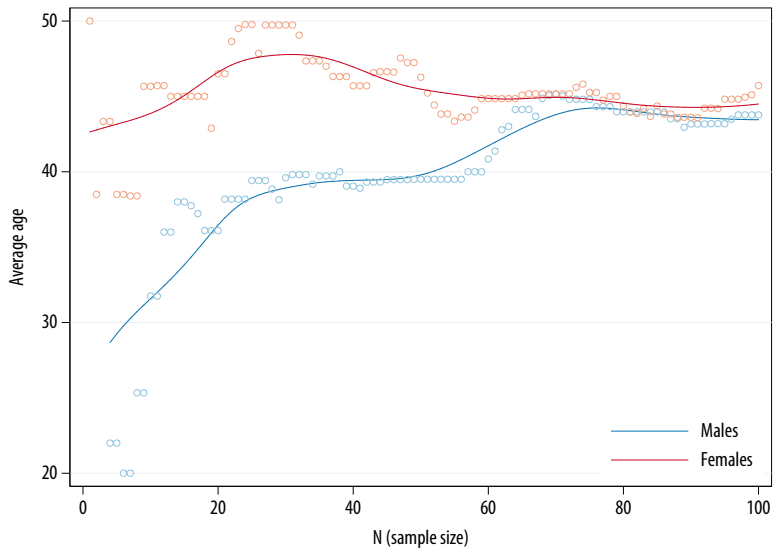
Measuring association as the difference in means between two groups of i.i.d. observations:

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

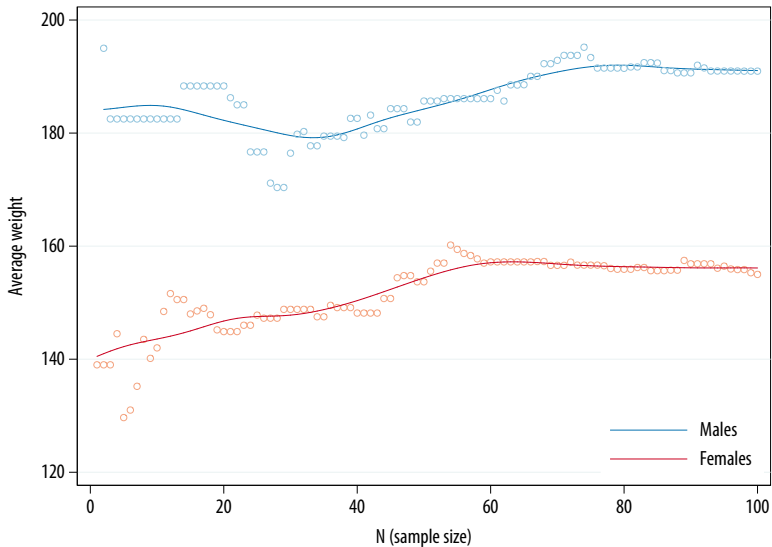
The t -test computes a 95% CI around the difference of their means and returns its p -value against the t -distribution.

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

Type I errors



Type II errors



Stata implementation

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

- v1 is continuous, v2 is a dummy
- use `prtest` if v1 is also a dummy (proportions test)
- use `tab`, `gen()` to create dummies from categorical variables

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

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

Stata implementation

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

Explore the variables and interpret the output below.

```
. prtest no_mes, by(gol_polreg)
```

Two-sample test of proportions

0. Democracy: Number of obs = 109

1. Dictators: Number of obs = 79

Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
0. Democracy	.293578	.0436195			.2080853 .3790706
1. Dictators	.2911392	.0511113			.1909628 .3913156
diff	.0024387	.067194			-.129259 .1341365
	under Ho:	.0672205	0.04	0.971	

diff = prop(0. Democracy) - prop(1. Dictators)

z = 0.0363

Ho: diff = 0

Ha: diff < 0

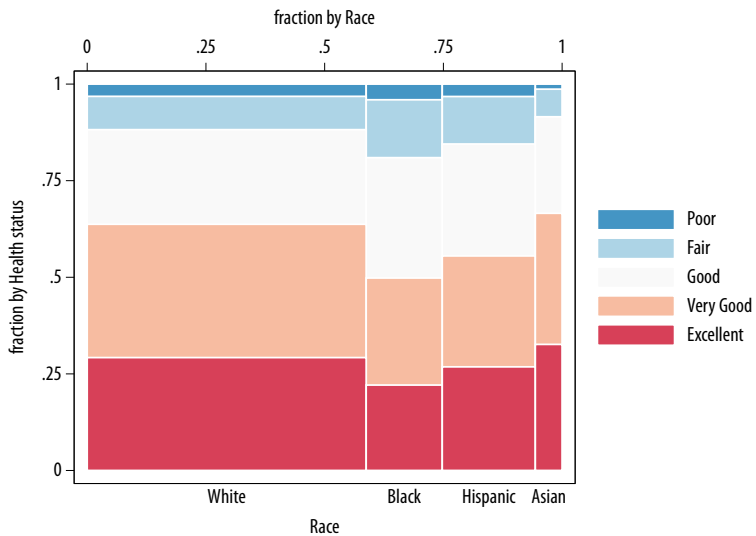
Pr(Z < z) = 0.5145

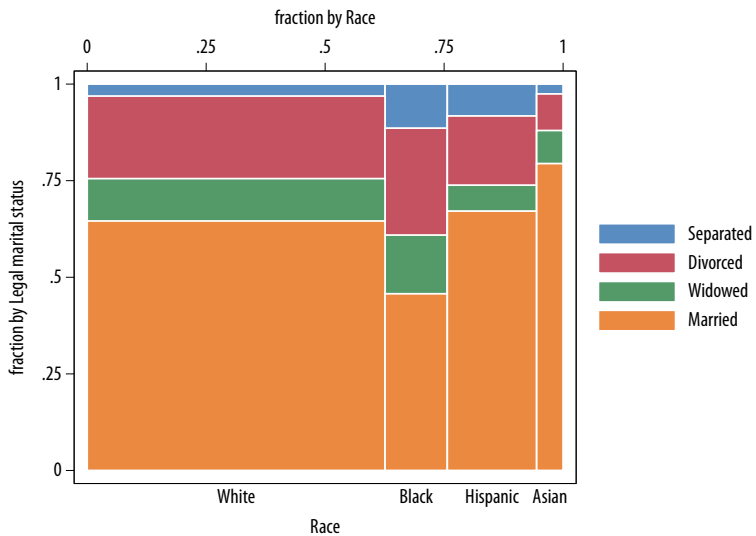
Ha: diff != 0

Pr(|Z| < |z|) = 0.9711

Ha: diff > 0

Pr(Z > z) = 0.4855





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, `tabodds` for odds ratios

Stata implementation

```
use datasets/nhis2009, clear
```

- Variables: d raceb marstat
- Analyze the frequencies and residuals with tabchi

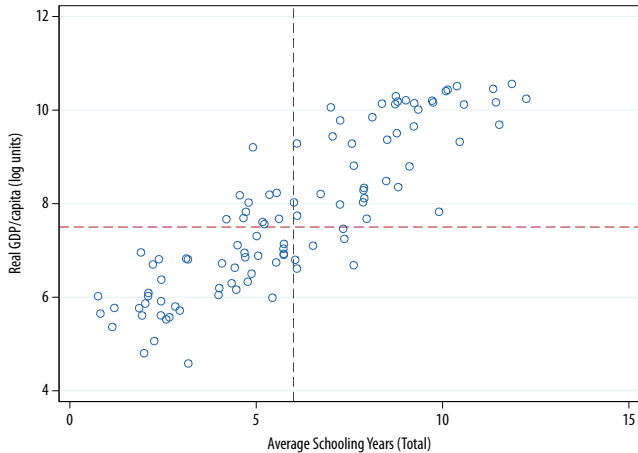
```
. tab marstat raceb if marstat < 8, chi2 V
```

Legal marital status	Race				Total
	White	Black	Hispanic	Asian	
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

Correlation



Pearson correlation coefficient

Measuring association as the linear dependence of two variables:

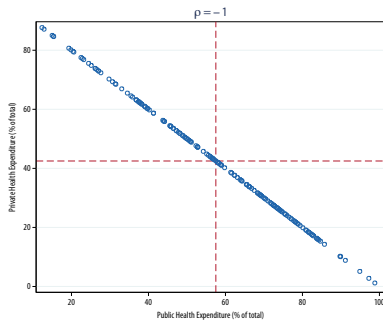
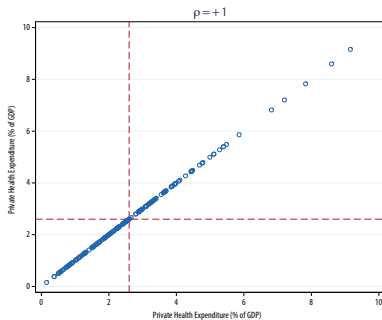
Population notation $\rho = \frac{\text{Cov}(X, Y)}{\text{Var}_X \text{Var}_Y}, \quad -1 \leq \rho \leq 1$

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

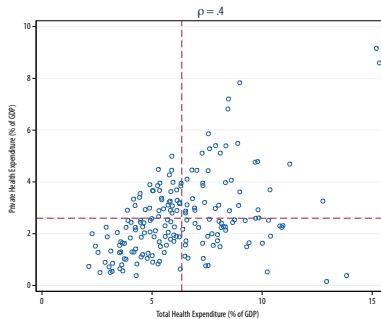
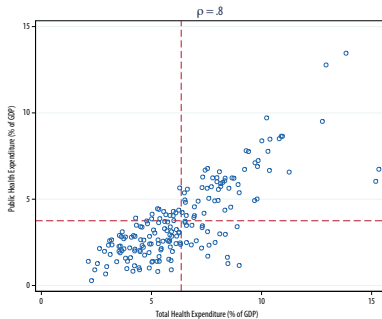
Detects linear correlation

- Uncorrelated \neq unrelated
- Correlated \neq unconfounded

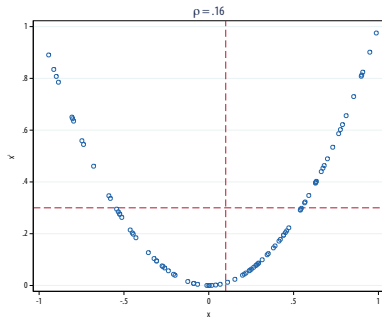
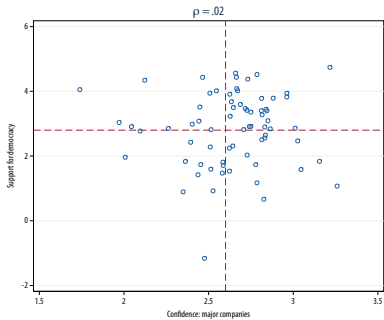
Perfect positive/negative correlation



Significant (moderate–strong) correlation



Insignificant (weak, non-linear) correlation



Pearson correlation coefficient

Significance test:

Null hypothesis $H_0 \quad r = 0$

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

Sanity check

- Uncorrelated \neq independent
- Correlated \neq causally related

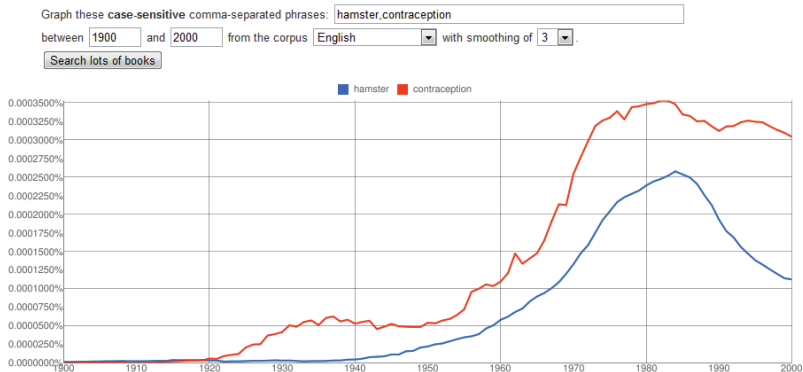
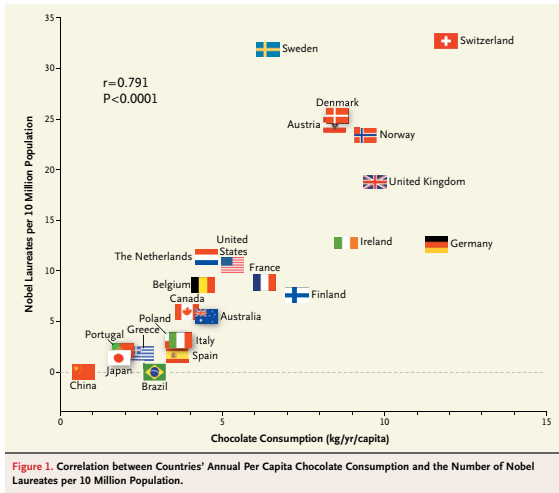


Figure 1: Frequencies of the words “hamster” and “contraception” in Google Books, 1900–2000

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



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

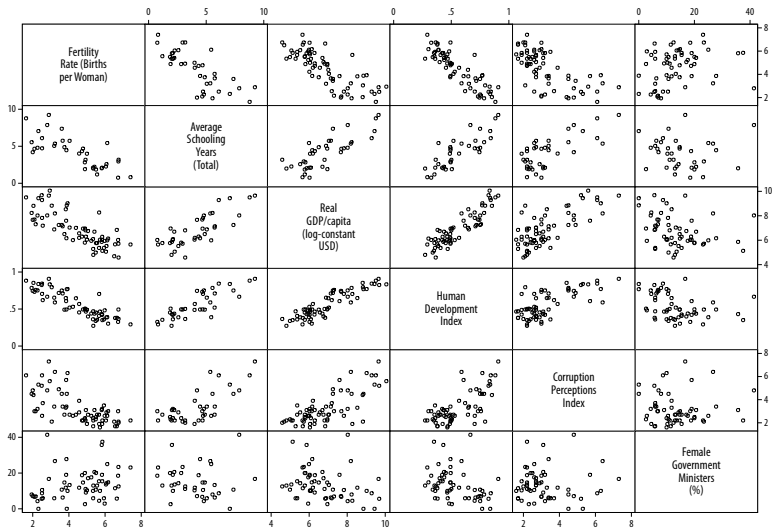
Stata implementation

```
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.)



Showing only Africa and the Middle East ($N = 68$).

Stata output

```
. pwcorr wdi_hiv wdi_hec wdi_prhe wdi_puhegdp, obs sig star(.05)
```

	wdi_hiv	wdi_hec	wdi_prhe	wdi_pu~p
wdi_hiv	1.0000			
	141			
wdi_hec	-0.1953* 0.0207	1.0000		
	140	187		
wdi_prhe	0.0979 0.2497	-0.0555 0.4509	1.0000	
	140	187	188	
wdi_puhegdp	-0.0607 0.4759	0.5490* 0.0000	-0.2099* 0.0038	1.0000
	140	187	188	188

coefficient

p-value

observations

Publishing standard

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.*, “Public Policy, Political Connections, and Effective Tax Rates: Longitudinal Evidence from Malaysia”, *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_puhegdp wdi_the wdi_prhe
- Visualize, compute, export and import the correlation matrix.

Thanks for your attention!

Project

- Start testing associations in your data
- Refine hypotheses and write draft findings

Readings

- *Stata Guide*, Sec. 9
- *Making History Count*, ch. 3

Practice

- Replicate do-file
- Exercises in slides