Quantitative methods

Week #10-11

Gergely Daróczi

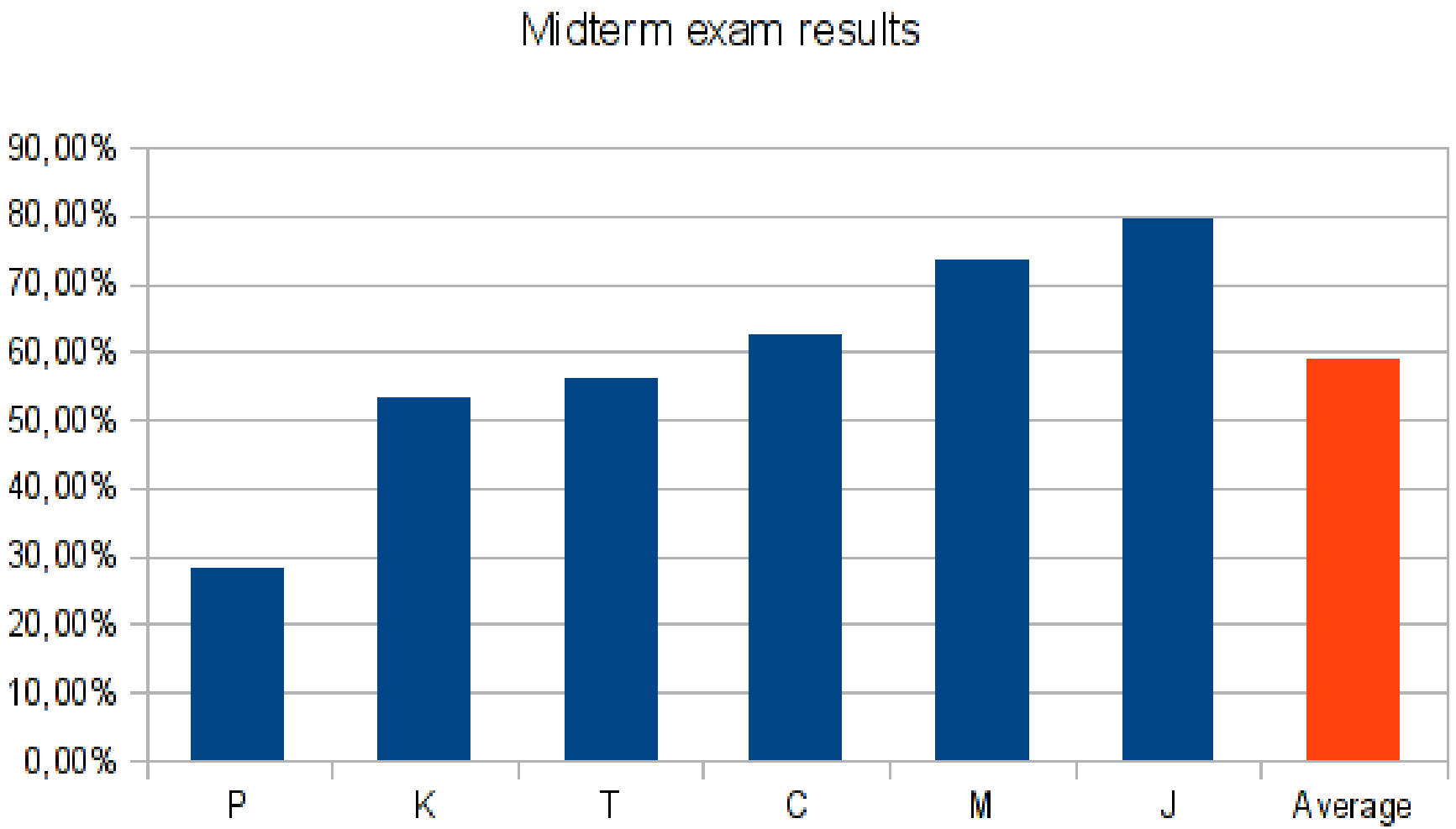
Corvinus University of Budapest, Hungary

12 April 2013

Outline

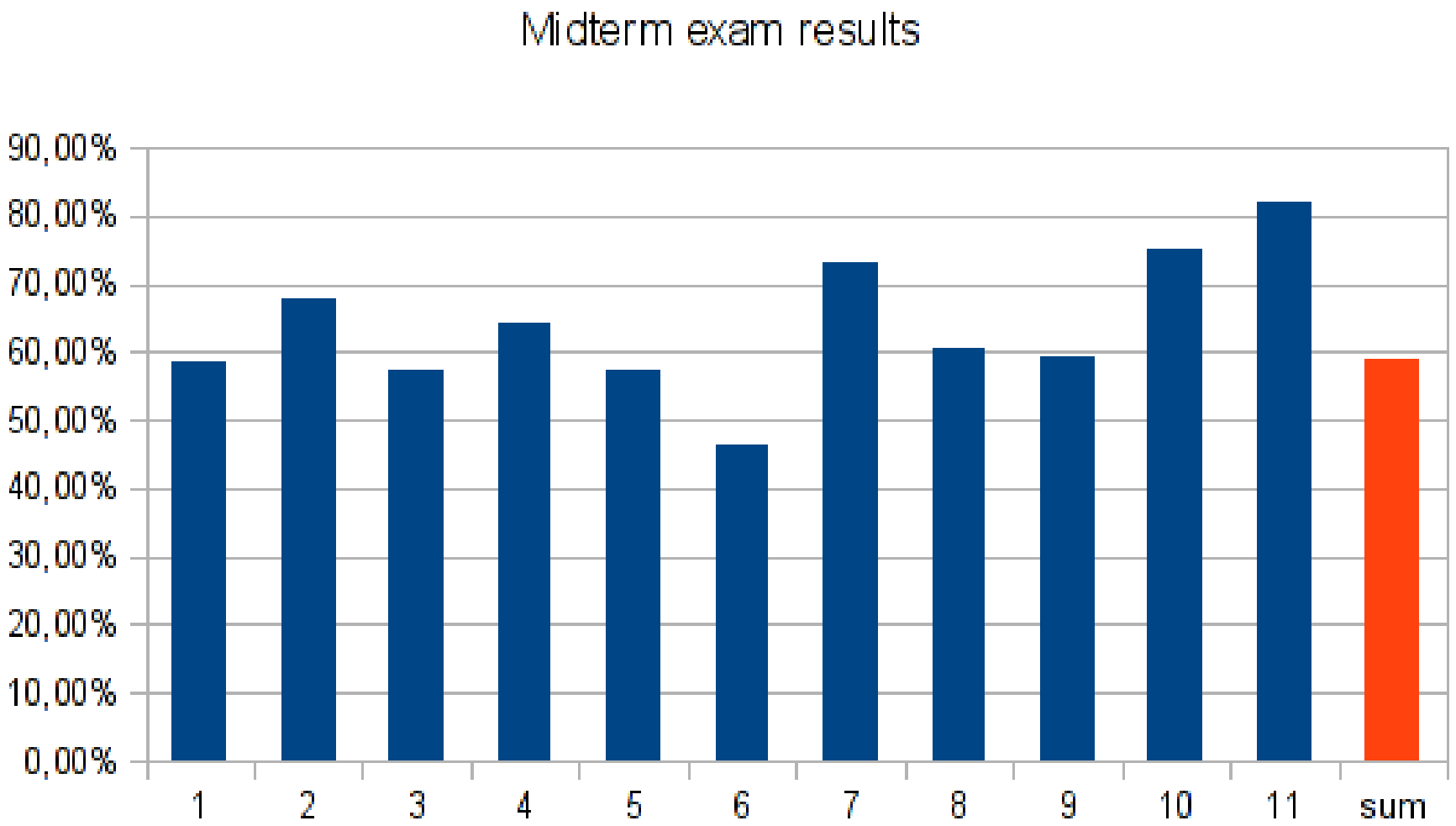
Midterm exam

Results



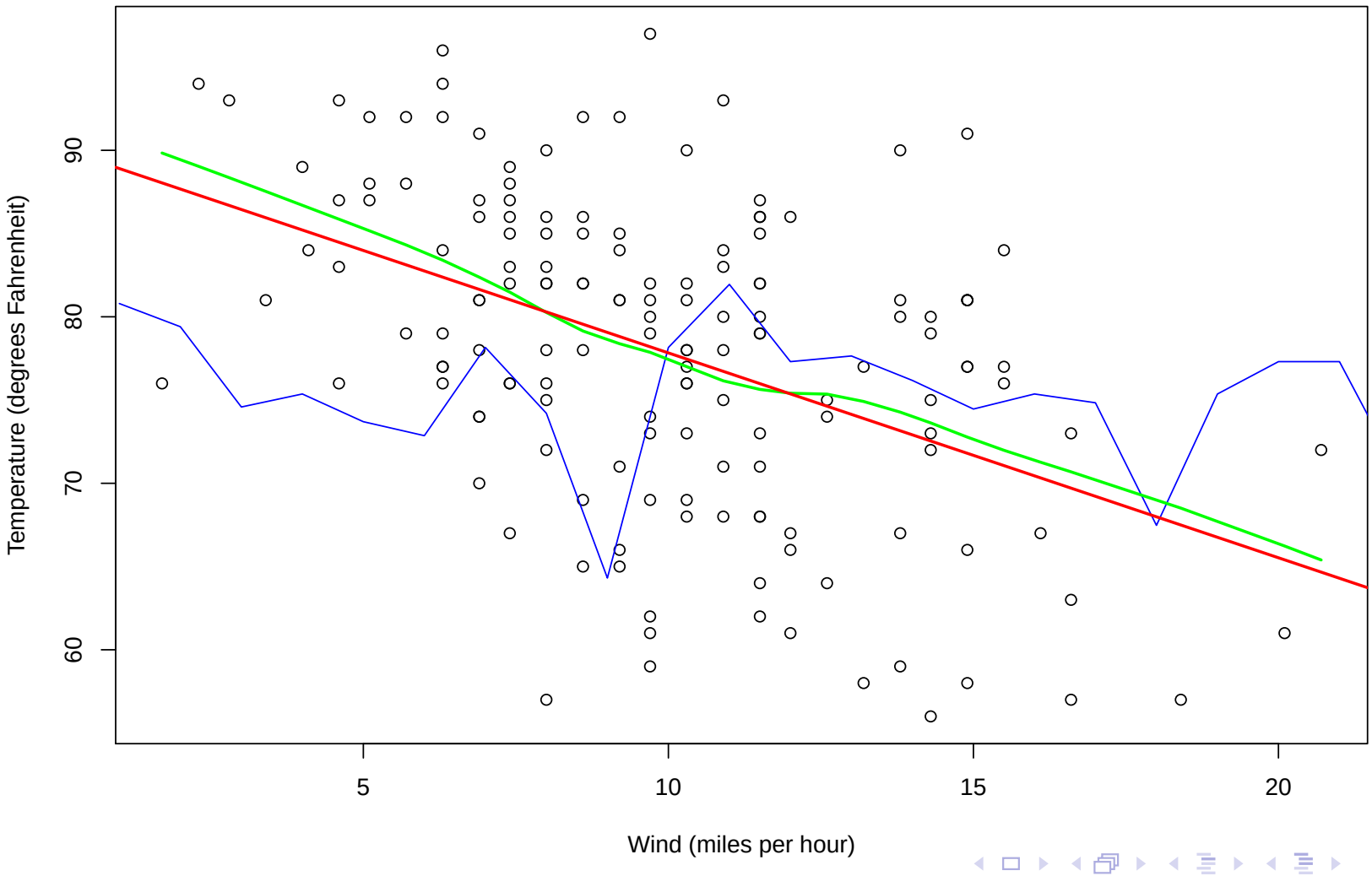
Midterm exam

Results



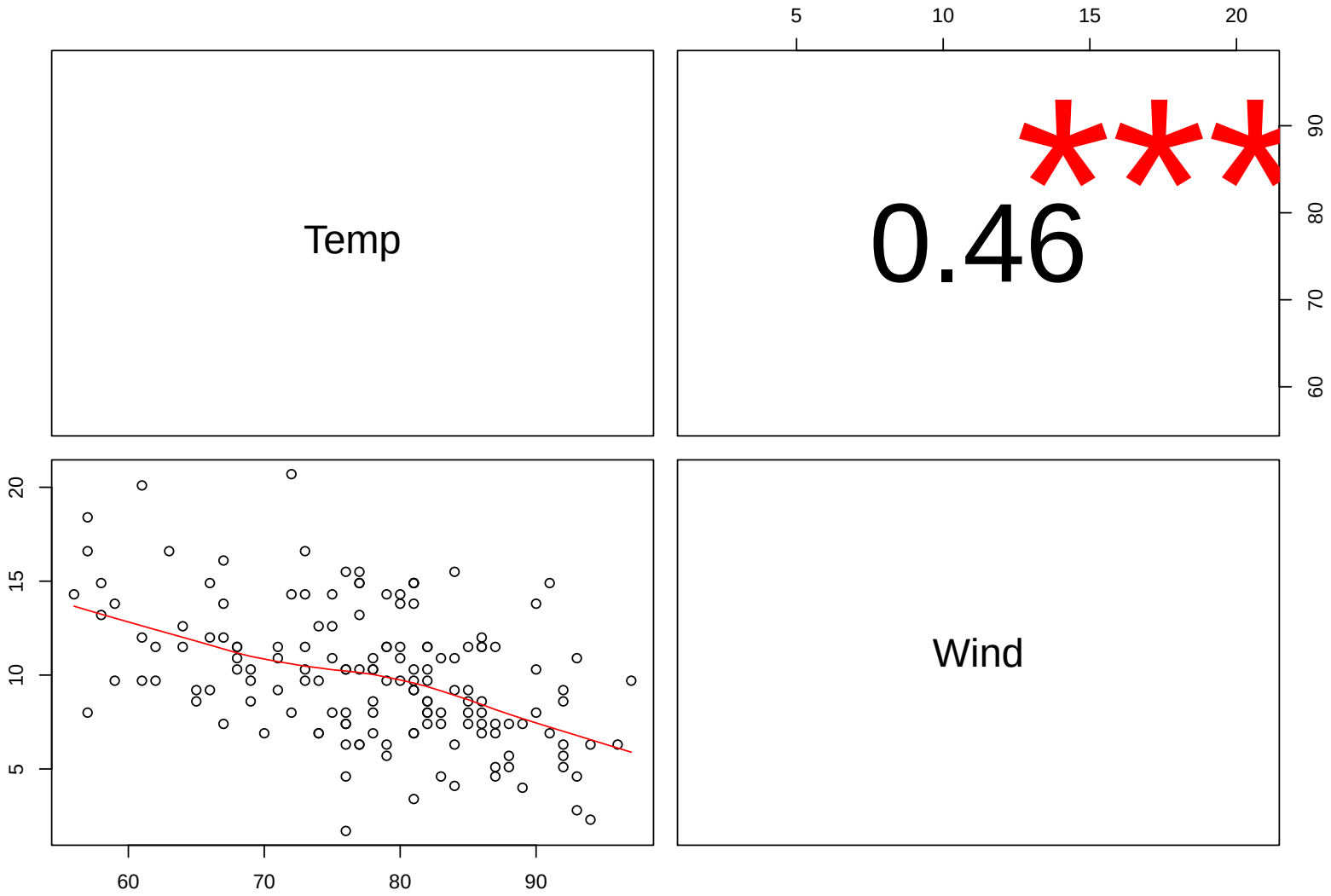
Correlation

Real association?



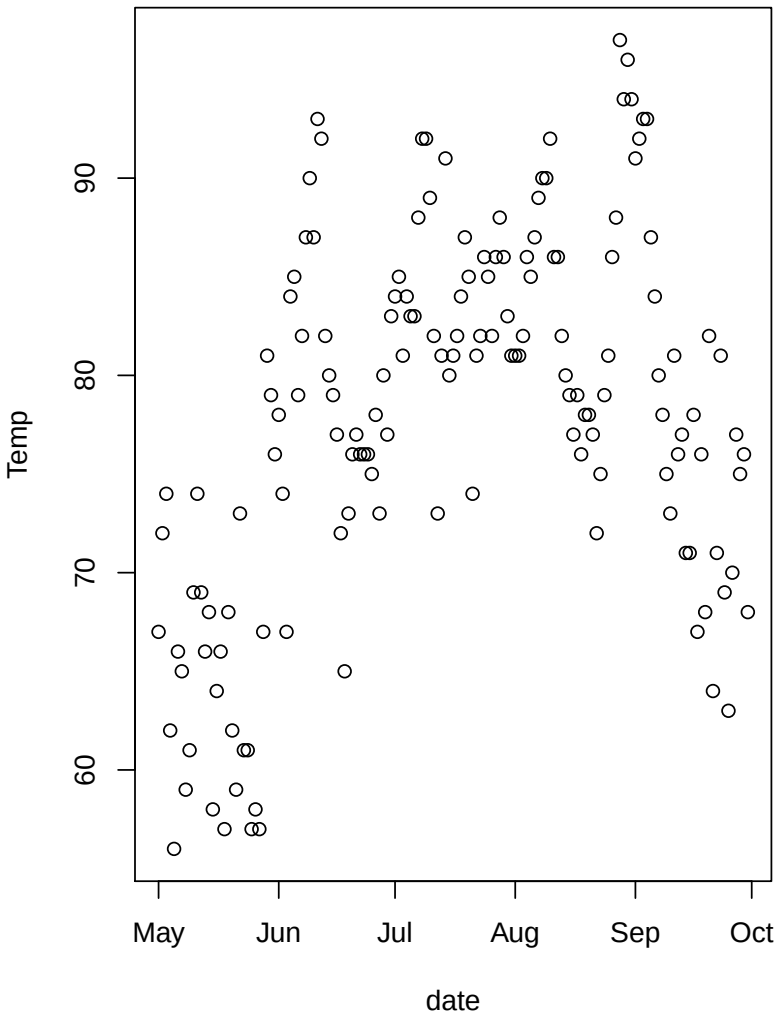
Correlation

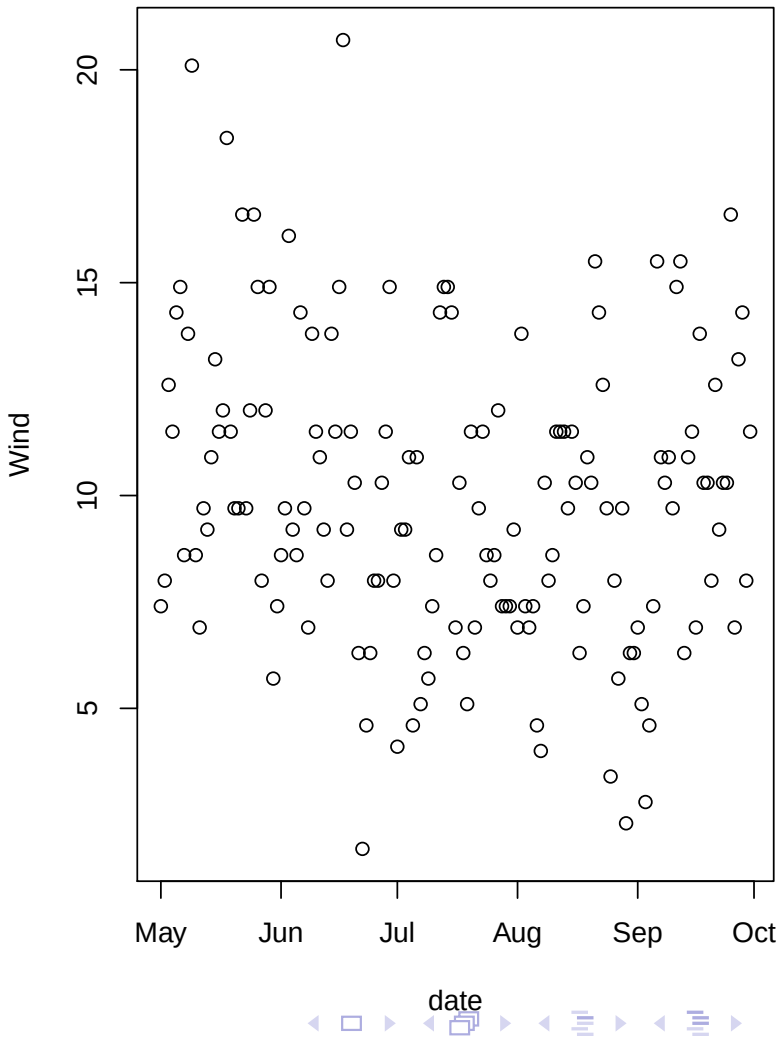
Real association?



Correlation

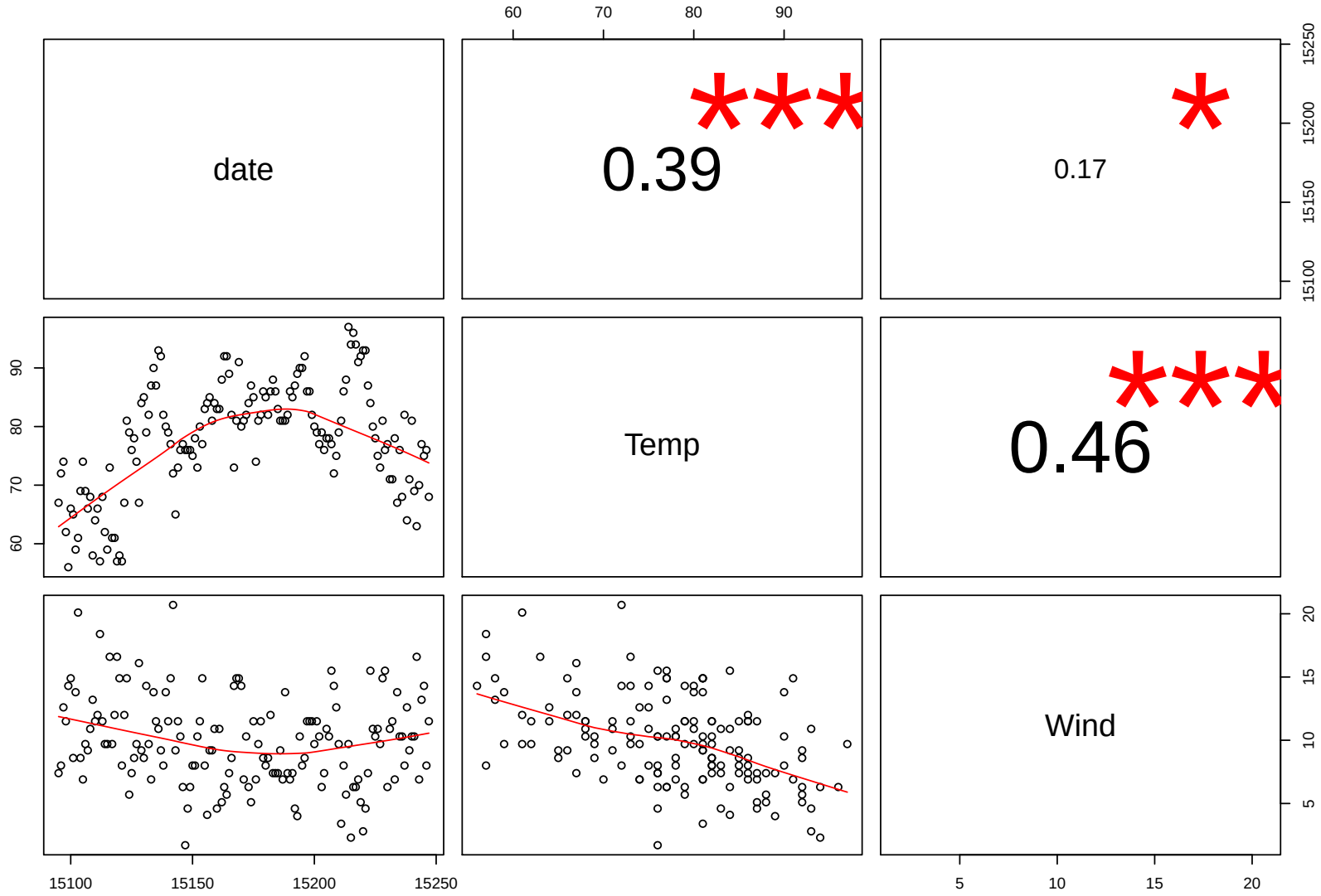
Real association?





Correlation

Real association?



Correlation exercise

Required formulas

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \quad COV(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$\hat{r}_{XY \cdot \mathbf{Z}} = \frac{r_{XY} - r_{X\mathbf{Z}} r_{Y\mathbf{Z}}}{\sqrt{(1 - r_{X\mathbf{Z}}^2)(1 - r_{Y\mathbf{Z}}^2)}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \hat{\beta} = \frac{COV(x,y)}{VAR(x)} \quad \hat{y} = \beta x + \alpha \quad \hat{y} = \alpha + \beta x$$

Compute the correlation and build linear models:

Grade	Monthly scholarship	Spent on books
3	22 000 Ft	4 000 Ft
4	24 000 Ft	3 000 Ft
5	27 000 Ft	2 500 Ft
3.5	24 000 Ft	3 500 Ft
2	23 000 Ft	2 000 Ft

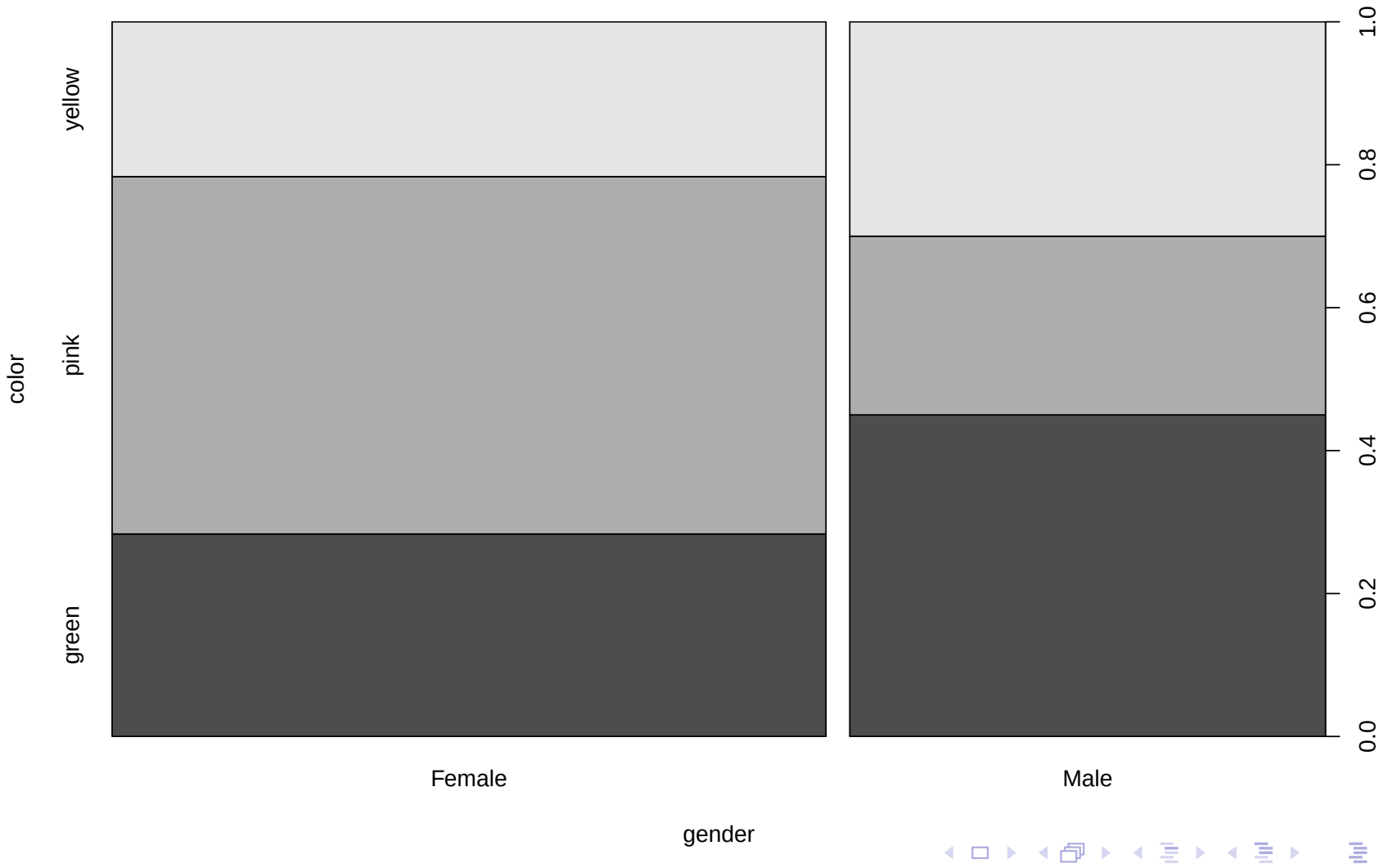
Crosstables

Discrete (qualitative) variables

ID	gender	color
1	Female	pink
2	Female	pink
3	Female	pink
4	Female	pink
5	Female	pink
6	Female	pink
	...	
95	Male	yellow
96	Male	yellow
97	Male	yellow
98	Male	yellow
99	Male	yellow
100	Male	yellow

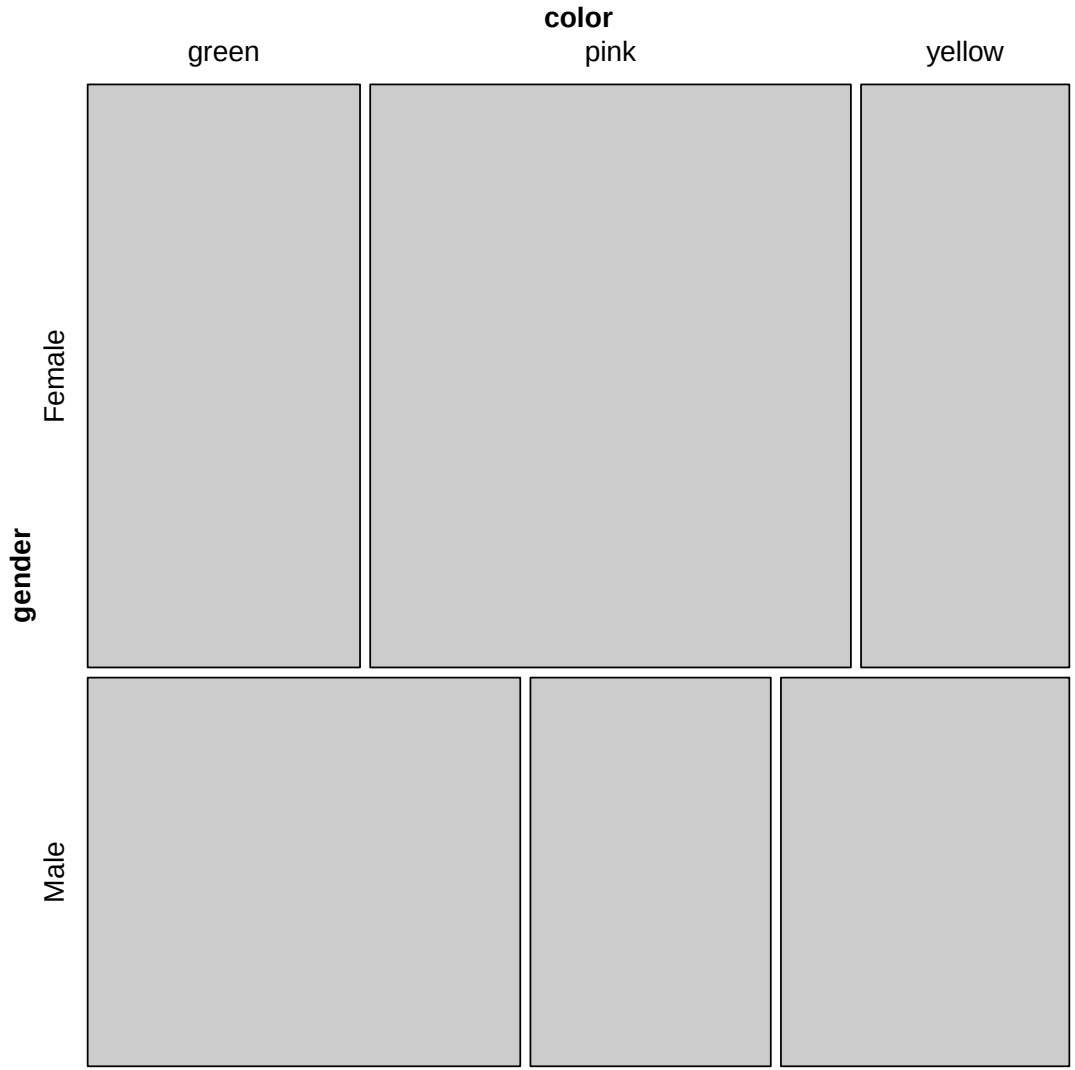
Crosstables

Discrete (qualitative) variables



Crosstables

Discrete (qualitative) variables



Crosstables

Discrete (qualitative) variables

	green	pink	yellow
Female	17	30	13
Male	18	10	12

Crosstables

Discrete (qualitative) variables

	green	pink	yellow	
Female	17	30	13	Marginals
Male	18	10	12	
	Marginals			N

Crosstables

Discrete (qualitative) variables

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	35	40	25	100

Crosstables

Percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	35	40	25	100

Table : Counted values

	green	pink	yellow	$\sum$
Female	17 %	30 %	13 %	60 %
Male	18 %	10 %	12 %	40 %
$\sum$	35 %	40 %	25 %	100 %

Table : Total percentages

Crosstables

Row percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	35	40	25	100

Table : Counted values

	green	pink	yellow	$\sum$
Female	28.3 %	50 %	21.7 %	100 %
Male	45 %	25 %	30 %	100 %
$\sum$	35 %	40 %	25 %	100 %

Table : Row percentages

Crosstables

Column percentages

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	35	40	25	100

Table : Counted values

	green	pink	yellow	$\sum$
Female	48.63 %	75 %	52 %	60 %
Male	51.4 %	25 %	48 %	40 %
$\sum$	100 %	100 %	100 %	100 %

Table : Column percentages

Crosstables

Expected values

	green	pink	yellow	$\sum$
Female	17	30	13	60
Male	18	10	12	40
$\sum$	35	40	25	100

Table : Counted values

	green	pink	yellow	$\sum$
Female	21	24	15	60
Male	14	16	10	40
$\sum$	35	40	25	100

Table : Expected values

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where:

- χ^2: Pearson's cumulative test statistic,
- O_i: an observed (counted) frequency,
- E_i: an expected (theoretical) frequency,
- n: the number of cells in the table.

H_0: observed and expected values are all the same

Requirements!

	green	pink	yellow	$\sum$
Female	$\frac{(17-21)^2}{21}$	$\frac{(30-24)^2}{24}$	$\frac{(13-15)^2}{15}$	-
Male	$\frac{(18-14)^2}{14}$	$\frac{(10-16)^2}{16}$	$\frac{(12-10)^2}{10}$	-
$\sum$	-	-	-	-

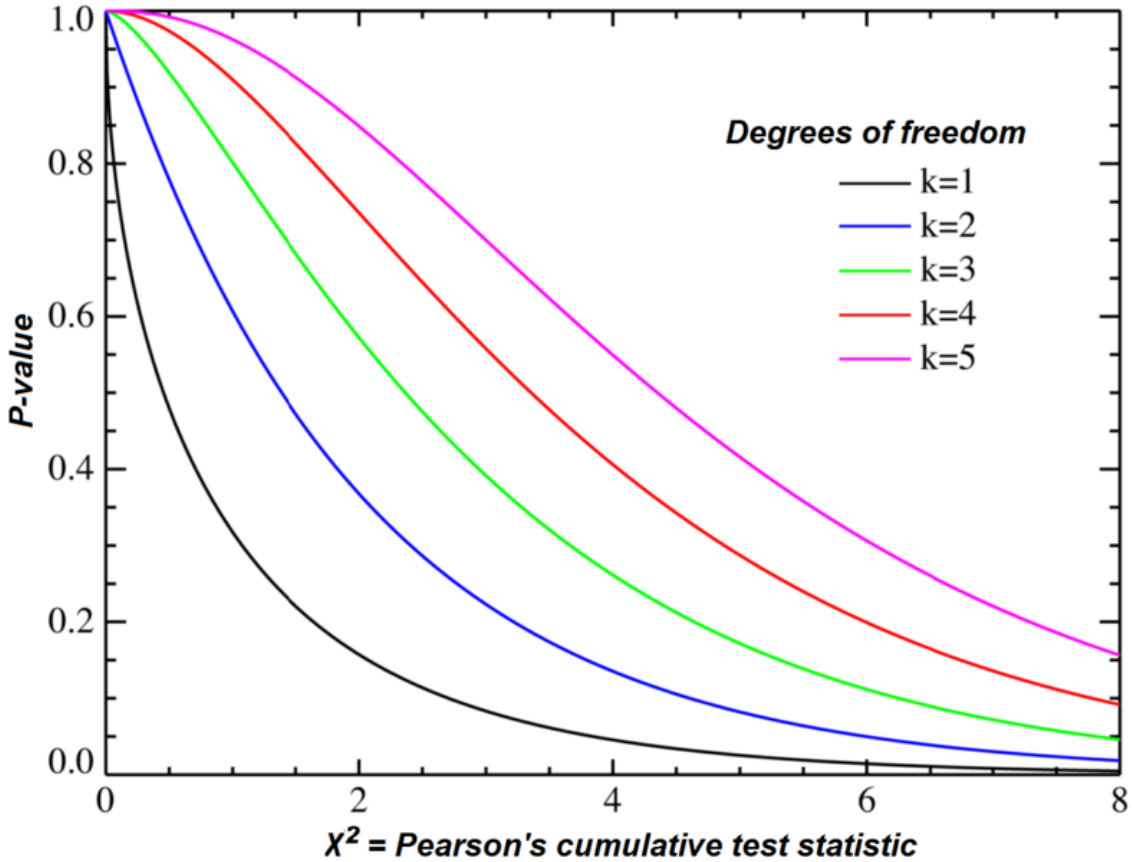
Table : Computed distances between observed and expected values

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 6.321429$$

degrees of freedom: $(2-1)(231) = 2$

Crosstables

Computed chi-square



$$\chi_c^2 = 2$$

$$\Rightarrow p = 0.04239545$$

Crosstables

Exercise

	Read required readings	Did not read required readings
3	15	5
4	20	10
5	45	5

$$E_{i,j} = \frac{M_{i.} \cdot M_{.j}}{N}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

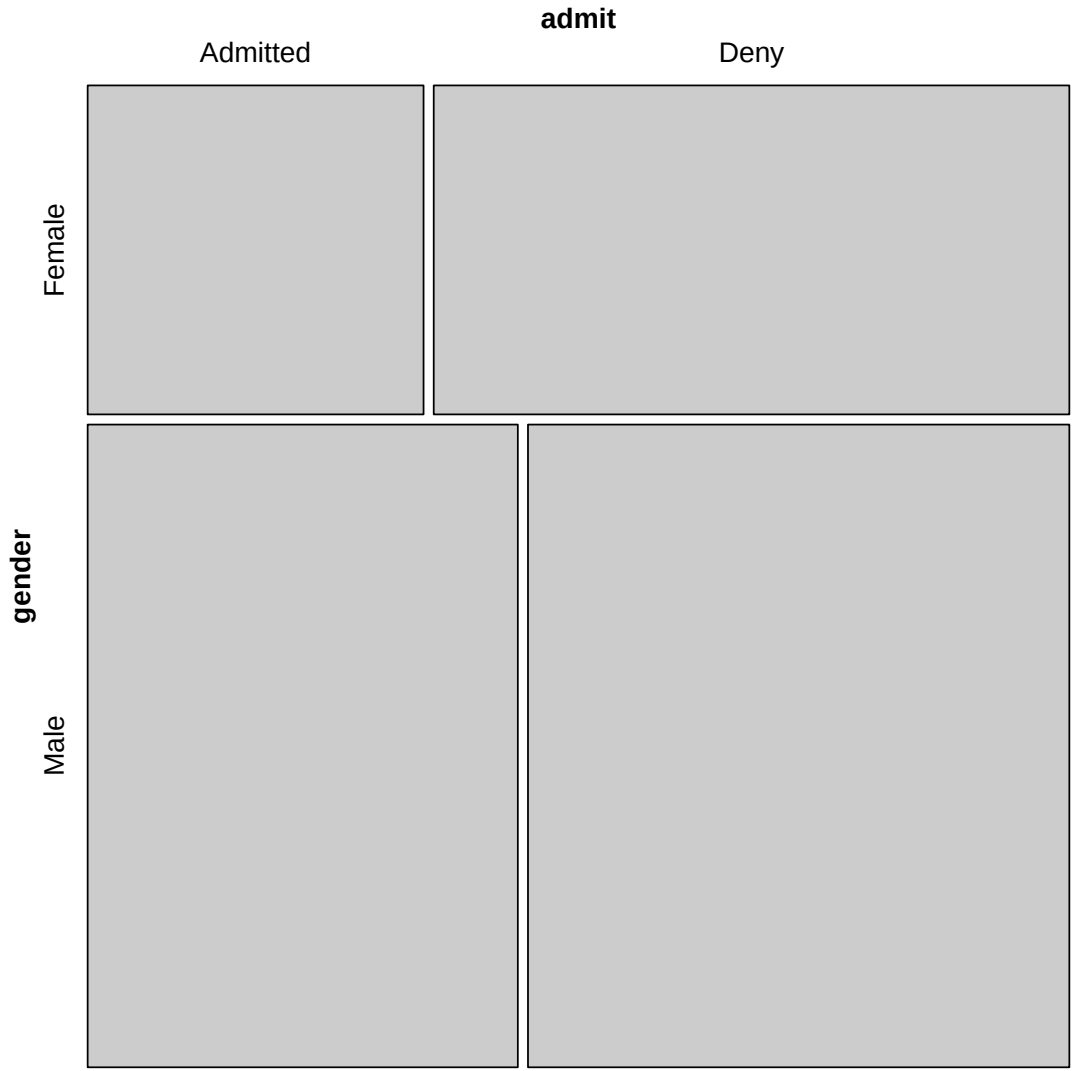
$$df = (3-1)(2-1) = 2$$

$$\chi_c^2 = 2$$

$$\phi = \frac{\chi}{N} \qquad V_c = \sqrt{\frac{\chi^2}{N \cdot (k-1)}}$$

Simpson's paradox

Berkeley sex bias case



Simpson's paradox

Berkeley sex bias case

	Admitted	Deny	$\sum$
Female	1494	2827	4321
Male	3738	4704	8442
$\sum$	5232	7531	12763

Table : Observed values

Simpson's paradox

Berkeley sex bias case

	Admitted	Deny	$\sum$
Female	1494	2827	4321
Male	3738	4704	8442
$\sum$	5232	7531	12763

Table : Observed values

	Admitted	Deny	$\sum$
Female	34.6 %	65.4 %	100 %
Male	44.3 %	55.7 %	100 %
$\sum$	41 %	59 %	100 %

Table : Row percentages

Simpson's paradox

Berkeley sex bias case

	Admitted	Deny	$\sum$
Female	1494	2827	4321
Male	3738	4704	8442
$\sum$	5232	7531	12763

Table : Observed values

	Admitted	Deny	$\sum$
Female	34.6 %	65.4 %	100 %
Male	44.3 %	55.7 %	100 %
$\sum$	41 %	59 %	100 %

Table : Row percentages

$$\chi^2 = 110.8489; d.f. = 1; p = 6.385628e-26$$

Simpson's paradox

Berkeley sex bias case

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

	Men		Women	
Departement	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Simpson's paradox

Batting averages in professional baseball

	1995		1996		Combined	
	Runs/Outs	%	Runs/Outs	%	Runs/Outs	%
Derek Jeter	12/48	25 %	183/582	31.4 %	195/630	**31** %
David Justice	104/411	**25.3** %	45/140	**32.1** %	149/551	27 %

Who is the better player?

It was a pleasure!

Gergely Daróczi
daroczi.gergely@btk.ppke.hu