Quantitative methods

Week #10-11

Gergely Daróczi

Corvinus University of Budapest, Hungary

27 April 2012



◆□▶ ◆□▶ ◆三▶ ◆三▶ ○□ ○○○

Outline

- Non-probability sampling
- Final examination questions
- 3 Correlation
 - Example
 - Theoretical background
 - Exercises
 - Limitations of the correlation coefficient
 - Exercises
- 4 Crosstables
 - Theoretical background
 - Simpson's paradox
- 5 Standardization and decomposition
 - Graphs

Nonprobability sampling:

- Accidental, Haphazard or Convenience Sampling,
- Ø Modal Instance Sampling,
- Expert (Judgmental) Sampling,
- Quota Sampling:
 - Proportional Quota Sampling,
 - 2 Nonproportional Quota Sampling.
- Heterogeneity Sampling (deviant cases),
- Snowball Sampling.

	POPULATION	Female	Male	Σ
	Economics	10	10	20
	Sociology	40	10	50
	Philoposphy	10	20	30
	Σ	60	40	100
Stratified	d Sampling:	Quot	ta Sam	pling:

?

Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

э

?

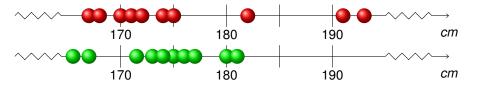
Singleton, R. A. Jr. and Bruce C. Straits (1999): Approaches to Social Research. Third Edition. Oxford University Press: New York/Oxford.

Questions:

- What is reliability? How do the main rules concerning the order of survey questions improve the reliability and validity of survey data? (pp. 113-117, 292-296)
- What is meant by probability sampling? How do stratification and multistage cluster sampling affect sampling errors? Why? (pp. 141-142, 145-156)
- What are the main types of non-probability sampling? Explain why these types do not meet the criteria of probability samples. (pp. 157-169)
- What factors affect the desired sample size? (pp. 163-169)

• • = • • = •

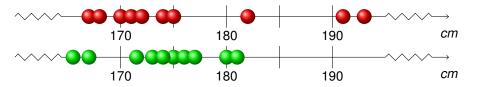
We have measured 10-10 students in two classrooms.

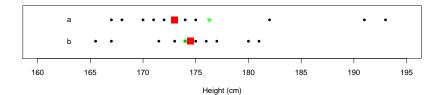


Which class has higher students based on this small sample? Think about averages as good estimates of populatioin parameters!



Averages





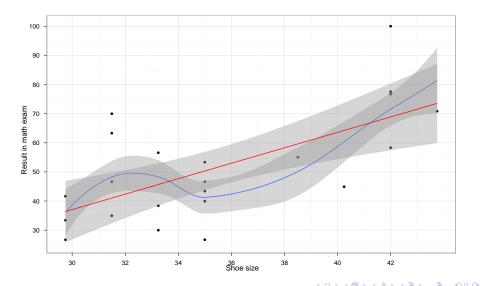
イロト イヨト イヨト イヨト

Big shoes and smart kids (example)

We made a small research on the age and shoe size of some students in an elementary shool, where we also conducted a math exam. See detailed results below:

	Shoe size	Math result
1	29.75	26.67
2	29.75	33.33
3	29.75	41.67
4	31.50	35.00
5	31.50	46.67
6	31.50	63.33
7	31.50	70.00
8	33.25	30.00
9	33.25	38.33
10	33.25	56.67
11	35.00	26.67
12	35.00	40.00
13	35.00	43.33
14	35.00	46.67
15	35.00	53.33
16	38.50	55.00
17	40.25	45.00
18	42.00	58.33
19	42.00	76.67
20	42.00	77.50
21	42.00	100.00
22	43.75	70.83

Big shoes and smart kids (example)

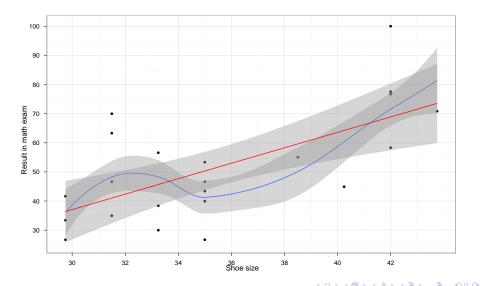


Big shoes and smart kids (example)

We made a small research on the age and shoe size of some students in an elementary shool, where we also conducted a math exam. See detailed results below:

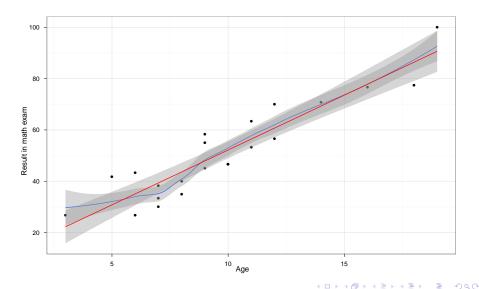
	Shoe size	Math result	Age
1	29.75	26.67	3
2	29.75	33.33	7
3	29.75	41.67	5
4	31.50	35.00	8
5	31.50	46.67	10
6	31.50	63.33	11
7	31.50	70.00	12
8	33.25	30.00	7.
9	33.25	38.33	7
10	33.25	56.67	12
11	35.00	26.67	6
12	35.00	40.00	8
13	35.00	43.33	6
14	35.00	46.67	10
15	35.00	53.33	11
16	38.50	55.00	9
17	40.25	45.00	9
18	42.00	58.33	9
19	42.00	76.67	16
20	42.00	77.50	18
21	42.00	100.00	19
22	43.75	70.83	14

Big shoes and smart kids (example)



27/4/2012 11 / 69

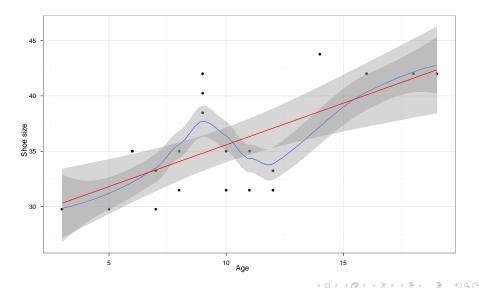
Big shoes and smart kids (example)



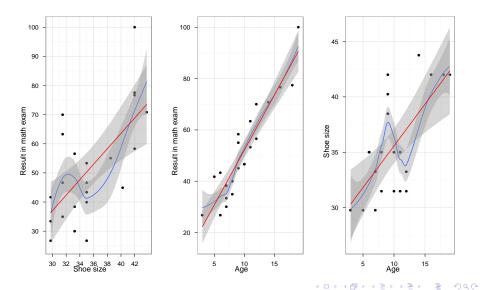
Gergely Daróczi (BCE)

27/4/2012 11 / 69

Big shoes and smart kids (example)



Big shoes and smart kids (example)

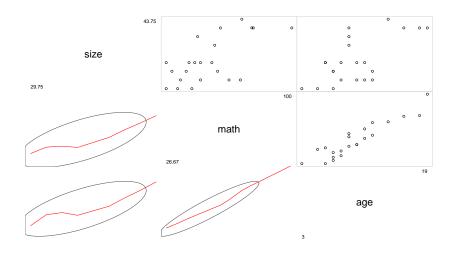


Gergely Daróczi (BCE)

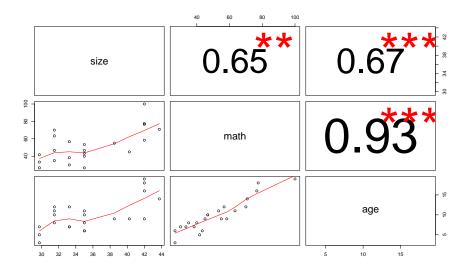
Quantitative methods, 10-11/14

27/4/2012 11 / 69

Big shoes and smart kids (example)



Big shoes and smart kids (example)



Gergely Daróczi (BCE)

27/4/2012 11 / 69

Big shoes and smart kids (example)

Partial correlation:

 $r_{math,size \cdot age} = 0.11$

 $r_{math,age\cdot size} = 0.87$

 $r_{size,age\cdot math} = 0.22$

Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

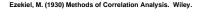
27/4/2012 11 / 69

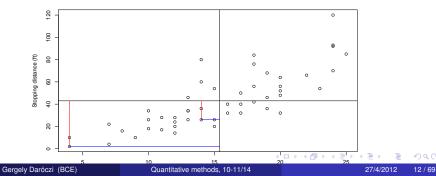
Covariation

For x and y variables the joint variability could be computed by :

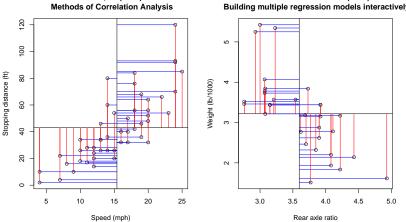
$$COV(xy) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

remember : $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$





Covariation



Ezekiel, M. (1930):

Henderson & Velleman (1981): Building multiple regression models interactively

< D > < B

▶ < 프 > < 프</p>

Correlation

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

$$\frac{1.0 \quad 0.8 \quad 0.4 \quad 0.0 \quad -0.4 \quad -0.8 \quad -1.0}{(1.0 \quad 1.0 \quad 1.0 \quad -1.0 \quad -1.0 \quad -1.0)}$$

$$\frac{0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0}{(0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0 \quad 0.0)}$$

27/4/2012 14 / 69

<ロト <回 > < 回 > < 三 > < 三

Partial correlation

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

so for three variables:

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

Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

27/4/2012 15 / 69

▶ < 토 ▶ < 토

Exercises

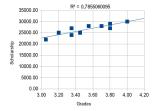
- What is correlation and partial correlation?
- Building upon your findings, compute the possible pairs of correlation coefficients on the below dataset!
- Also look for partial correlation and comment on your results!

Grade (mean)	Scholarship (in HUF)	Money spent on books (in HUF)
3.05	22000	3500
3.2	25000	3000
3.35	27000	2800
3.35	24000	3700
3.45	25000	2200
3.55	28000	3200
3.7	28000	3700
45	30000	4100
3.8	27000	4000
3.8	29000	3800

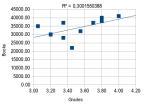
Exercises

Solution

Grade				$x_i - x$		$(x_{i} - \bar{x})^{2}$							
	Scholarship	Books	Grade	Scholarship	Books	Grade	Scholarship	Books		Grade-sch	Sch-books	Grade-books	
3.05	22000.00	35000.00	-0.48	-4500.00	1000.00	0.225625	20250000	1000000		2137.5	-4500000	-475	
3.20	25000.00	30000.00	-0.33	-1500.00	-4000.00	0.105625	2250000	16000000		487.5	6000000	1300	
3.35	27000.00	28000.00	-0.18	500.00	-6000.00	0.030625	250000	36000000		-87.5	-3000000	1050	
3.35	24000.00	37000.00	-0.18	-2500.00	3000.00	0.030625	6250000	9000000		437.5	-7500000	-525	
3.45	25000.00	22000.00	-0.07	-1500.00	-12000.00	0.005625	2250000	144000000		112.5	18000000	900	
3.55	28000.00	32000.00	0.02	1500.00	-2000.00	0.000625	2250000	4000000		37.5	-3000000	-50	
3.70	28000.00	37000.00	0.18	1500.00	3000.00	0.030625	2250000	9000000		262.5	4500000	525	
4.00	30000.00	41000.00	0.48	3500.00	7000.00	0.225625	12250000	49000000		1662.5	24500000	3325	
3.80	27000.00	40000.00	0.28	500.00	6000.00	0.075625	250000	36000000		137.5	3000000	1650	
3.80	29000.00	38000.00	0.28	2500.00	4000.00	0.075625	6250000	16000000		687.5	10000000	1100	
3.05	22000.00	22000.00			Sum	0.80625	54500000	320000000		5875	48000000	8800	
4.00	30000.00	41000.00			St. dev.	0.29930475	2460.80384	5962.84794	r	0.89	0.36	0.55	
0.95	8000.00	19000.00							r^2	0.79	0.13	0.30	
3.53	26500.00	34000.00						partial	corr	0.96	-0.24	0.46	
3.50	27000.00	36000.00											
	3.35 3.45 3.55 3.70 4.00 3.80 3.80 3.05 4.00 0.95 3.53	3.35 27000.00 3.35 24000.00 3.45 25000.00 3.70 28000.00 3.70 28000.00 3.80 27000.00 3.80 27000.00 3.80 29000.00 3.80 22000.00 4.00 3000.00 3.80 22000.00 4.00 30000.00 3.85 2600.00 3.85 26500.00	3.35 2700.00 28000.00 3.35 2400.00 3700.00 3.45 2500.00 22000.00 3.45 2500.00 3200.00 3.45 2600.00 3200.00 3.55 2600.00 3200.00 3.70 2800.00 3700.00 4.00 3000.00 41000.00 3.80 29000.00 32000.00 3.05 32000.00 32000.00 3.05 32000.00 32000.00 3.05 22000.00 32000.00 3.05 32000.00 34000.00 3.05 32000.00 41000.00 3.53 28500.00 14000.00	3.35 27000.00 2900.00 -0.16 3.35 24000.00 3700.00 -0.17 3.45 25000.00 2200.00 -0.07 3.45 26000.00 2200.00 0.02 3.55 26000.00 37000.00 0.02 3.70 28000.00 37000.00 0.18 3.80 2900.00 4000.00 0.28 3.80 29000.00 32000.00 -0.28 3.81 29000.00 20200.00 -0.28 3.05 22000.00 20200.00 -0.28 3.05 3000.00 41000.00 -0.28 3.05 3000.00 41000.00 -0.28 3.05 3000.00 41000.00 -0.28 3.53 26500.00 3000.00 -0.28	3.35 2700.00 2800.00 -0.18 500.00 3.35 2400.00 3700.00 -0.18 -2500.00 3.45 2500.00 2200.00 -0.18 -2500.00 3.45 2600.00 2200.00 -0.02 1500.00 3.55 2800.00 3200.00 0.12 1500.00 3.70 2800.00 3700.00 0.18 1500.00 3.00 3200.00 4000.00 0.28 500.00 3.80 2900.00 4000.00 0.28 500.00 3.80 22000.00 2200.00 0.28 500.00 3.05 22000.00 22000.00 -2800.00 -250.00 3.04 300.00 41000.00 -28 500.00 3.05 800.00 41000.00 -28 500.00 3.53 2650.00 34000.00 -55 800.00	3.35 2700.00 28000.00 -0.18 5000.00 6000.00 3.35 2400.00 37000.00 -0.18 -2500.00 2000.00 -0.18 -2500.00 3000.00 3000.00 3000.00 3000.00 3000.00 -0.18 -2500.00 -2000.00 -2000.00 -3000.	3.35 27000.00 28000.00 -0.18 500.00 -6000.00 0.030625 3.35 24000.00 3700.00 -0.18 -5200.00 3000.00 0.03625 3.45 25000.00 22000.00 -0.07 -1500.00 -12000.00 0.03625 3.45 28000.00 3200.00 -0.07 -1500.00 -1200.00 0.00625 3.55 28000.00 3700.00 0.04 1500.00 2000.00 0.00625 3.70 28000.00 4100.00 0.48 3500.00 0.000625 380 3.80 27000.00 40000.00 0.28 500.00 6000.00 0.075625 3.80 22000.00 3800.00 0.28 2500.00 4000.00 0.075625 3.05 22000.00 4100.00 -28 500.00 600.00 0.075625 3.05 22000.00 4100.00 -8 500.00 600.00 0.075625 3.05 28000.00 41000.00 -8 500.00 40002.00	3.35 27000.00 28000.00 -0.18 500.00 -6000.00 0.30625 250000 3.35 24000.00 37000.00 -0.18 -5200.00 3000.00 0.30625 225000.00 3.45 25000.00 32000.00 -0.16 -5200.00 3000.00 0.308625 225000.00 3.55 28000.00 32000.00 0.02 1500.00 -2000.00 0.006625 2250000 3.70 28000.00 37000.00 0.18 1500.00 -3000.00 0.006625 2250000 3.00 27000.00 40000.00 0.48 5500.00 6000.00 0.225625 2250000 3.80 27000.00 3000.00 0.28 500.00 6000.00 0.76525 250000 3.80 22000.00 22000.00 228 250.00 4000.00 0.76525 2550000 3.05 22000.00 4000.00 0.75625 2550000 3000.00 1076625 54500000 0.03 30000.00 4000.00.00	3.35 27000.00 28000.00 -0.18 550.00 -6000.00 0.030625 250000 3600000 3.45 24000.00 37000.00 -0.18 25000.00 0.030625 250000 9000000 3.45 25000.00 22000.00 -0.07 -1500.00 -12000.00 0.030625 2250000 44000000 3.55 2600.00 32000.00 -0.02 1500.00 -2000.00 0.00625 2250000 4000000 3.70 28000.00 37000.00 0.18 1500.00 3000.00 0.028625 2250000 4900000 3.80 27000.00 4000.00 0.28 550.00 6000.00 0.076525 250000 4600000 3.80 22000.00 22001.00 0.28 500.00 6000.00 0.076525 250000 4600000 3.80 22000.00 22001.00 2.850.00 550000 0.076525 250000 36000000 3.05 22000.00 22001.00 Streetteeteteteteteeteeteeteeteetetetete	3.35 27000.00 28000.00 -0.18 5500.00 -0.000.00 0.302625 250000 3000000 3.45 24000.00 3700.00 -0.18 -25000.00 0.302625 2250000 144000000 3.45 25000.00 22000.00 -0.16 -25000.00 0.302625 2250000 144000000 3.45 26000.00 32000.00 -0.02 1500.00 -2000.00 0.000625 2250000 14000000 3.55 26000.00 3700.00 0.02 1550.00 -3000.00 0.302625 2250000 9000000 3.70 28000.00 4700.00 0.48 550.00 6000.00 0.226525 2250000 9000000 3.80 27000.00 40000.00 0.28 550.00 6000.00 0.076525 2500000 3000000 3.81 22000.00 22000.00 22000.00 22000.00 3000.00 305 225000.01 3000.00 3000.00 3000.00 3000.00 3000.00 3000.00 3000.00 3	3.35 27000.00 2900.00 -0.18 550.00 -06000.00 0.306625 250000 3600.0000 -47.5 3.35 24000.00 3700.00 -0.18 25000.00 0.036625 226000.00 -47.5 3.45 26000.00 3700.00 -0.18 25000.00 0.036625 226000.00 -437.5 3.45 26000.00 3200.00 -0.02 1500.00 -0.0000 0.036625 226000.00 -0.00000 37.5 3.55 2600.00 3700.00 0.02 1500.00 -0.0000 0.006625 2250000 9000000 37.5 3.00 2000.00 4100.000 0.148 1500.00 0.006625 2250000 9000000 1662.5 3.00 27000.00 4000.00 0.28 520000 1600000 687.5 3.02 22000.00 22000.00 0.28 550.00 3000.00 326625 54500000 32000000 687.5 3.05 22000.00 22000.00 5450.000 3200.0	3.35 27000.00 26000.00 -0.18 500.00 0.030625 250000 36000000 -47.5 -3000000 3.45 24000.00 3700.00 -0.18 -25000.00 0.030625 250000 9000000 -47.5 -3000000 3.45 25000.00 20200.00 -0.18 -25000.00 0.030625 250000 9000000 447.5 -3000000 3.45 26000.00 22000.00 -0.07 -1500.00 12000.00 0.036625 2250000 144000000 17.5 3000000 3.70 28000.00 3700.00 0.02 1500.00 0.00625 2256000 9000000 462.5 4500000 3.00 2000.00 0.48 1500.00 0.02625 2250000 9000000 1682.5 4500000 3.80 27000.00 0.48 500.00 0.075625 250000 16000000 687.5 3000000 3.81 22000.00 282501 0.075625 250000 16000000 687.5 440000000 <t< td=""><td>3.35 27000.00 28000.00 -0.16 5000.00 0.000025 250000 96000000 437.5 -3000000 1050 3.35 2400.00 37000.00 -0.16 -25000.00 30000.000 437.5 -7500000 -525 3.45 25000.00 22000.00 -0.16 -2500.00 3000.000 437.5 -7500000 -525 3.45 25000.00 22000.00 -0.07 -1500.00 1000.00 2000.00 370.00 -50 3.55 28000.00 32000.00 -0.02 1500.00 -2000.00 0.000625 2250000 4000000 137.5 -3000000 -50 3.70 28000.00 41000.00 0.48 3500.00 0.000625 2250000 9000000 1662.5 4500000 3255 3.80 27000.00 4000.00 0.282552 5250000 9000000 1662.5 4500000 10000.00 3255 3.80 22000.00 2800.00 6000.00 0.076252 5250000 90000</td></t<>	3.35 27000.00 28000.00 -0.16 5000.00 0.000025 250000 96000000 437.5 -3000000 1050 3.35 2400.00 37000.00 -0.16 -25000.00 30000.000 437.5 -7500000 -525 3.45 25000.00 22000.00 -0.16 -2500.00 3000.000 437.5 -7500000 -525 3.45 25000.00 22000.00 -0.07 -1500.00 1000.00 2000.00 370.00 -50 3.55 28000.00 32000.00 -0.02 1500.00 -2000.00 0.000625 2250000 4000000 137.5 -3000000 -50 3.70 28000.00 41000.00 0.48 3500.00 0.000625 2250000 9000000 1662.5 4500000 3255 3.80 27000.00 4000.00 0.282552 5250000 9000000 1662.5 4500000 10000.00 3255 3.80 22000.00 2800.00 6000.00 0.076252 5250000 90000





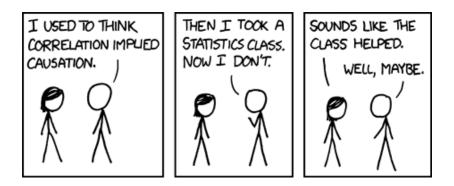


イロト イヨト イヨト イヨト

- Correlation and causality
- Lazarsfeld paradigm
- Correlation and linearity

Limitations of the correlation coefficient

Correlation does not imply causation!



Source: http://xkcd.com/552

< ロ ト < 同

▶ < 프 ▶ < 프 ▶</p>

Correlation does not imply causation! - Theoretical background

Aristotle: logic, syllogism – if $(A \rightarrow B)\&(B \rightarrow C) \Rightarrow A \rightarrow C$

David Hume: scepticism

- "only correlation can actually be perceived [not causality]"
- see: our belief that the sun will rise tomorrow
- see: "If I see a billiard ball moving towards another, on a smooth table, I can easily conceive to stop upon contact."

Popper: falsification

Pearl, J. - *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000

Lazarsfeld paradigm

Stouffer: The American Soldier

Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. Lazarsfeld paradigm

Stouffer: The American Soldier

 H_0 : Soldiers in branches with higher promotion rates are happier than soldiers in branches with lower rates of promotion. **BUT:**

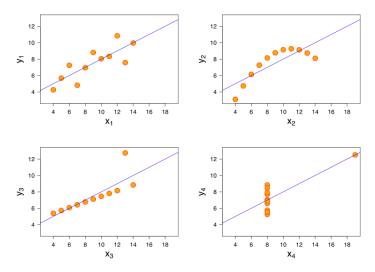
"Soldiers in branches with higher promotion rates were more pessimistic about their own chances of being promoted than soldiers in branches with lower rates of promotion."

Keywords: reference group, relative deprivation

Gergely Daróczi (BCE)

Limitations of the correlation coefficient

Correlation and linearity - Variations of the Same Theme



Source: Anscombe, F. J. (1973) Graphs in statistical analysis. American Statistician, 27, 17-21, C

Gergely Daróczi (BCE)

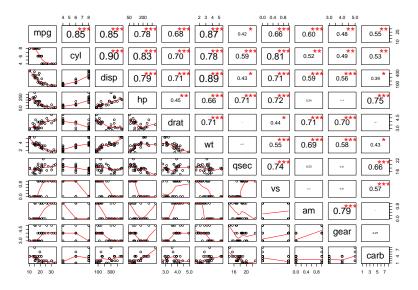
Exercise

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- o drat: Rear axle ratio
- wt: Weight (lb/1000)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Source: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411.

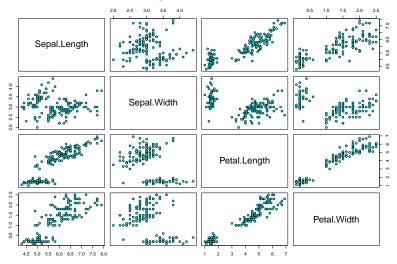
• • **E** • • **E** •



Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411= 🛌 😑

Gergely Daróczi (BCE)

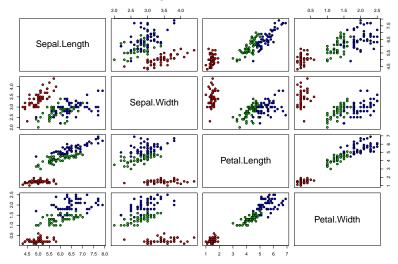
Edgar Anderson's Iris Data



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society 59, 2-5.

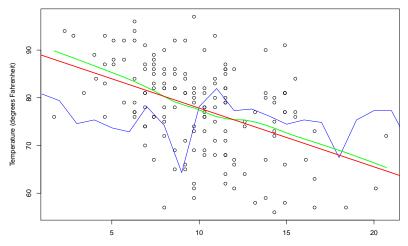
Gergely Daróczi (BCE)

Edgar Anderson's Iris Data



Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2-5.

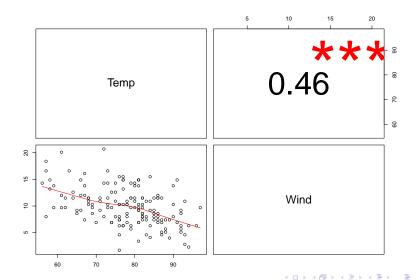
Gergely Daróczi (BCE)

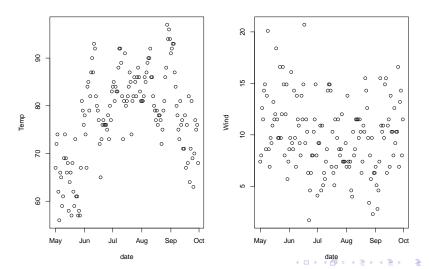


Wind (miles per hour)

Gergely Daróczi (BCE)

27/4/2012 28 / 69



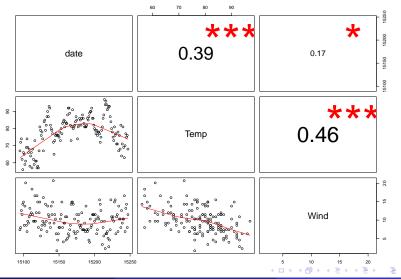


Gergely Daróczi (BCE)

27/4/2012 28 / 69

Exercise #3

Real association?



Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

27/4/2012 28 / 69

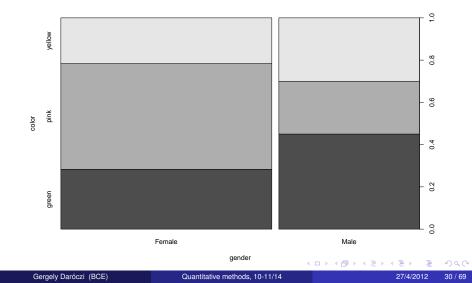
Discrete (qualitative) variables

_

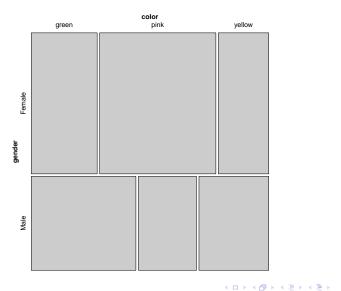
ID	gender	color
1	Female	pink
2	Female	pink
3	Female	pink
4	Female	pink
5	Female	pink
6	Female	pink
95	Male	yellow
96	Male	yellow
97	Male	yellow
98	Male	yellow
99	Male	yellow
100	Male	yellow

<ロ> (四) (四) (三) (三) (三)

Discrete (qualitative) variables



Discrete (qualitative) variables



Gergely Daróczi (BCE)

Discrete (qualitative) variables

	green	pink	yellow
Female	17	30	13
Male	18	10	12

ヘロア 人間 アメヨア 人間 アー

Discrete (qualitative) variables

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

ヘロン 人間 とくほ とくほ とう

Discrete (qualitative) variables

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

ヘロン 人間 とくほ とくほ とう

Percentages

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

Table: Counted values

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

Table: Total percentages

Row percentages

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

Table: Counted values

	green	pink	yellow	Σ
Female	28.3 %	50 %	21.7 %	100 %
Male	45 %	25 %	30 %	100 %
Σ	35 %	40 %	25 %	100 %

Table: Row percentages

Column percentages

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

Table: Counted values

	green	pink	yellow	Σ
Female	48.63 %	75 %	52 %	60 %
Male	51.4 %	25 %	48 %	40 %
Σ	100 %	100 %	100 %	100 %

Table: Column percentages

Expected values

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

Table: Counted values

	green	pink	yellow	Σ
Female	21	24	15	60
Male	14	16	10	40
Σ	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!

Computed chi-square

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

Table: Computed distances between observed and expected values

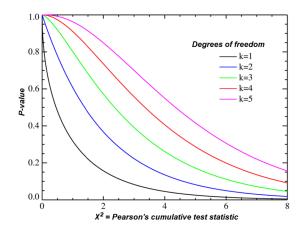
$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = 6.321429$$

degrees of freedom:
$$(3-1)(2-1) = 2$$

▶ < 프 > < 프</p>

< □ > < 🗇

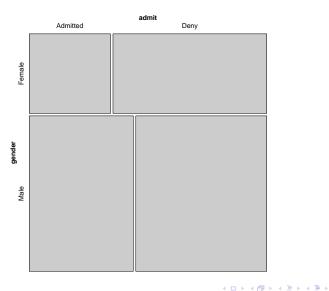
Computed chi-square



 $\Rightarrow p = 0.04239545$

Simpson's paradox

Berkeley sex bias case



Simpson's paradox

Berkeley sex bias case

	Admitted	Deny	Σ
Female	1494	2827	4321
Male	3738	4704	8442
Σ	5232	7531	12763

Table: Observed values

	Admitted	Deny	Σ
Female	34.6 %	65.4 %	100 %
Male	44.3 %	55.7 %	100 %
Σ	41 %	59 %	100 %

Table: Row percentages

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

Gergely Daróczi (BCE)

Berkeley sex bias case

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

-	Me	en	Women		
Departement	Applicants	Admitted	Applicants	Admitted	
A	825	62%	108	82%	
В	560	63%	25	68%	
С	325	37%	593	34%	
D	417	33%	375	35%	
E	191	28%	393	24%	
F	272	6%	341	7%	

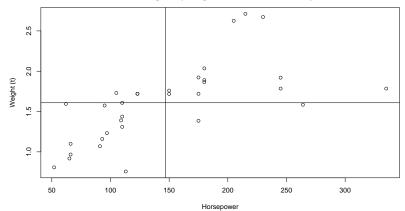
イロト 人間 とくほ とくほ とう

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?

A basic example

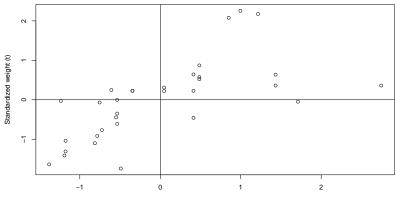


Henderson & Velleman (1981): Building multiple regression models interactively

Gergely Daróczi (BCE)

ъ

A basic example



Henderson & Velleman (1981): Building multiple regression models interactively

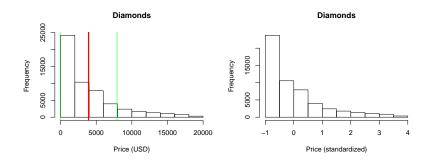
Standardized horsepower

ъ

Basic theory of normalization

Standard score (z-values, z-scores, normal scores, standardized variables) indicates how many standard deviations an observation is above or below the mean:

$$z = \frac{x-\mu}{\sigma}$$



Decomposition

		Miami			Alaska			U.S.	
Age	Pop.	Deaths	Rate*	Pop.	Deaths	Rate*	Pop.+	Deaths ⁺	Rate*
									_
< 15	114,350	136	1.19	37,164	59	1.59	23,961	32	1.34
15-24	80,259	57	0.71	20,036	18	0.90	15,420	9	0.58
25-44	133,440	208	1.56	32,693	37	1.13	21,353	30	1.40
45-64	142,670	1,016	7.12	14,947	90	6.02	19,609	140	7.14
65+	92,168	3,605	39.11	2,077	81	39.00	10,685	529	49.51
	562,887	5,022		106,917	285		91,028	740	
Crude death rate*			8.92			2.67			8.13

Population and Deaths by Age in 1970 for White Females in Miami, Alaska, and the U.S.

* Deaths per 1,000 population + in thousands

Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

27/4/2012 47 / 69

Direct standardization

Definition

In direct standardization the stratum-specific rates of study populations are applied to the age distribution of a standard population.

Directly standardized rate =
$$\frac{\sum stratum specific rates \times standard weights}{\sum standard weights}$$

$$\begin{aligned} \text{Miami} &= \frac{(1.19x23,961) + \dots + (39.11x10,685)}{91,208} = 6.92 \ \text{deaths/thousand} \\ \text{Alaska} &= \frac{(1.59x23,961) + \dots + (39x10,685)}{91,208} = 6.71 \ \text{deaths/thousand} \end{aligned}$$

Indirect standardization

Definition

In indirect standardization, the standard population provides the rates and the study population provides the weights.

Indirectly standardized rate = $\frac{\sum observed \ values}{\sum expected \ values}$

Expected values = Stratum specific rates from the study population \times stratum sizes from the study population

		Study population	Standard population		
Directly-standardized rate		Rates	Weights		
Indirectly-standardized rate		Weights	Rates		
		4	ロト・日本・モン・モージへ		
Gergely Daróczi (BCE)	Gergely Daróczi (BCE) Qu		27/4/2012 49 / 6		



27/4/2012 50 / 69



i



Visits from search engines

Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

E ► E ∽ Q C 27/4/2012 51 / 69

◆□▶ ◆圖▶ ◆厘▶ ◆厘≯

Graphs Line

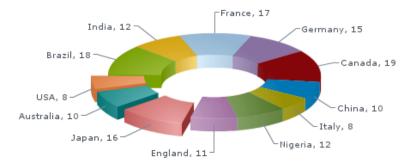


Gergely Daróczi (BCE)

≣ ► ≣ •ी २.० 27/4/2012 52 / 69



Industrial Growth Rate (Country)



27/4/2012 53 / 69

Graphs

Area



Gergely Daróczi (BCE)

Quantitative methods, 10-11/14

27/4/2012 54 / 69

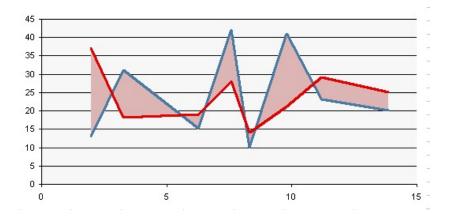




≣ ► ≣ •ी २.० 27/4/2012 55 / 69

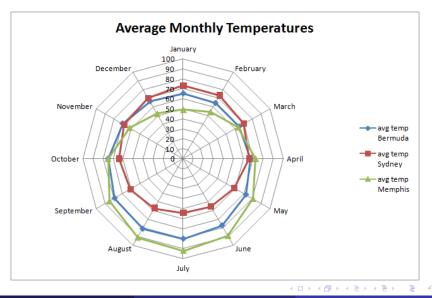
◆□▶ ◆圖▶ ◆厘▶ ◆厘≯

Graphs Combo



≣ ► ≣ •ी २.० 27/4/2012 56 / 69

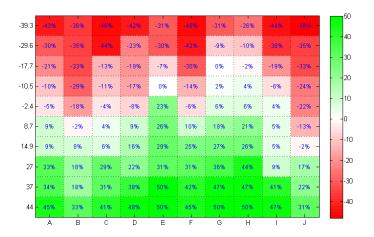




Gergely Daróczi (BCE)

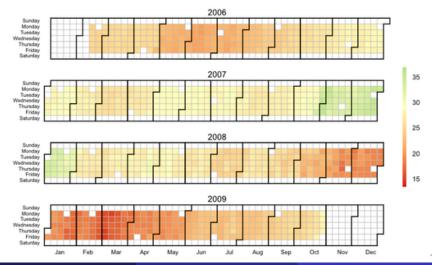
27/4/2012 57 / 69

Graphs Heatmap



≣ ► ≣ •ी २.० 27/4/2012 58 / 69

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・



Calendar Heat Map of MSFT Adjusted Close

Gergely Daróczi (BCE)





27/4/2012 60 / 69

◆□▶ ◆圖▶ ◆理▶ ◆理▶ ─ 理

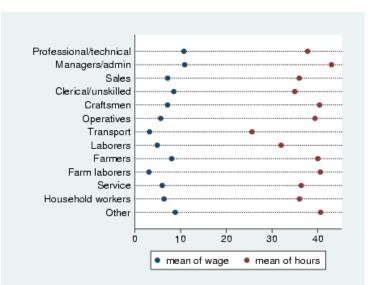
Graphs Dot plot

Gas Milage for Car Models grouped by cylinder

4									
÷.	Toyota Corolla					• • • •			
	Fiat 128					• • • • • • • • • • • • • • • • • • • •			
	Lotus Europa					• • • • • • • • • • • • • • • • • • • •			
	Honda Civic					••••••			
	Fiat X1-9				•••••				
	Porsche 914-2				•••••				
	Merc 240D				•••••				
	Merc 230				•				
	Datsun 710				•				
	Toyota Corona			•••••					
	Volvo 142E			• • • • • • • • • • • • • • • • • • • •					
6									
	Hornet 4 Drive			• • • • • • • • • • • • • • • • • • • •					
	Mazda RX4 Wag			• • • • • • • • • • • • • • • • • • • •					
	Mazda RX4			• • • • • • • • • • • • • • • • • • • •					
	Ferrari Dino			• • • • • • • • • • • • • • • • • • • •					
	Merc 280			•					
	Valiant		• • •						
	Merc 280C		• • •						
8									
	Pontiac Firebird			• •					
	Hornet Sportabout			•					
	Merc 450SL		• • • • • • • • • • • • • • • • • • • •						
	Merc 450SE		• • • • • • • • • • • • • • • • • • • •						
	Ford Pantera L		• • • • • • • • • • • • • • • • • • • •						
	Dodge Challenger		• • • • • • • • • • • • • • • • • • • •						
	AMC Javelin		••••						
	Merc 450SLC		••••						
	Maserati Bora		• •						
	Chrysler Imperial		• • • • • • • • • • • • • • • • • • • •						
	Duster 360		•						
	Camaro Z28	• • • •							
	Lincoln Continental								
	Cadillac Fleetwood	•							
		· · · · · · · · · · · · · · · · · · ·	1	1	1	1			
		10	15	20	25	30			
				Mileo Der	Callon				
		Miles Per Gallon							

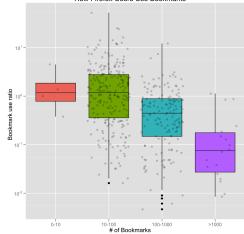
≣ ► ≣ •ी २.० 27/4/2012 61 / 69

ヘロト 人間 とくほ とくほ とう



27/4/2012 62 / 69



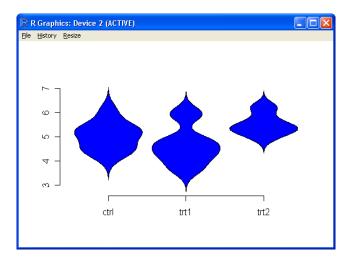


How Firefox Users Use Bookmarks

< D > < B

27/4/2012 63 / 69

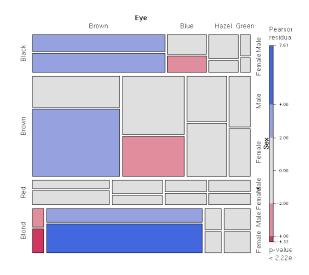
★ E > ★ E



≣ ► ≣ •ी २.० 27/4/2012 64 / 69

<ロ> (日) (日) (日) (日) (日)

Graphs Mosaic chart



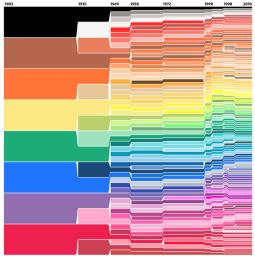
Gergely Daróczi (BCE)

E ► E ∽ Q C 27/4/2012 65 / 69

<ロ> (日) (日) (日) (日) (日)



Graphs "Crayola Color Chart, 1903-2010"



DATA POINTED datapointed.n

・ロト ・聞 と ・ 国 と ・ 国 と

Gergely Daróczi (BCE)

≣ ► ≣ ∽ ९ ९ 27/4/2012 67 / 69

- http://www.visual-literacy.org/periodic_table/periodic_table.html
- http://www.edwardtufte.com/tufte/
- http://www.perceptualedge.com/
- http://www.visualcomplexity.com/vc/
- http://flowingdata.com/
- http://infosthetics.com/
- http://chartsgraphs.wordpress.com/
- http://www.informationisbeautiful.net/
- http://chartporn.org/

It was a pleasure!

Daróczi Gergely daroczi.gergely@btk.ppke.hu

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへで