# Statistik ist ein Segen für die Menschheit

Günter Pilz

7.11.2019

# Statistics can do much more than people think!

- Usually people think: statistics = counting and computing means
- But statistics can also find „the reason(s) why“
- Key: (Statistical) Regression

You suspect that some „factors“ might have an influence on some value of your interest.

# A simple example, 1

Take a patient with unknown factors which trigger an allergy, where the usual diagnostic measures did not yield a satisfactory result. Suppose that the patient and the doctor suspect that 3 more factors $x_1$, $x_2$, $x_3$ might explain the allergy, e.g.,

- $x_1$ = exhaust air of the vacuum cleaner (measured in minutes of exposure)
- $x_2$ = intake of certain candies (measured in pieces), …
- $x_3$ = level of stress (measured by the blood pressure)

Then a test might expose the patient for 3 minutes to the vacuum cleaner, give him 5 candies, and measure his blood pressure. After - say - one hour, the patient ranks the degree *y* of allergy on a scale of up to 10. For this test, we might note

$$(3, 5, 145;\ 7)$$

# A simple example, 2

After the patient gets back to normal, a second test might yield

(0, 2, 150; 5),

and so on.

Our dream would – at the end – be the information if and how $x_1, x_2, x_3$ contribute to the allergy level $y$.

Regression fulfills this dream by producing a „formula“

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

A possible result might be

$$y = 2.5 + 1.2x_1 - 0.002x_2 + 0.01x_3$$

How should that be interpreted?

# Questions

- Which tests make sense?
- How many tests are needed?
- Is the „model“

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

„allowed“? What if the $x_i$ are not „independent“?

# More examples

- Agriculture: how do fertilizers $x_1$, $x_2$ , … influence the harvest yield $y$ ?
- How does $x_1$ = traffic,… influence global warming $y$ ?
- Which actions will reduce tropical diseases, and how much?
- What are the reasons $x_1$, $x_2$ , … for a rare disease?

Rare diseases ,…

# First summary

Regression can give you a „formula“ for (up to now) unknown connections.

So regression is like a license to print money !!!

# A personal regression

- $x_1$, $x_2$ , … = food components (magnesium, potassium, carbohydrates, …)
- $y$ = gain / loss of power after the intake
- Result:

  y = -0.5+8*(sodium in g) – 5*(potassium in g)

Example: 1 Burger brings

$$y = -0.5 + 8 * 1 - 5 * 0.4 = 5.5 \quad \text{(kg)}$$

more power!

# What if 2 factors are dependent?

In this case, we should test all single and all combinations of 2 factors.

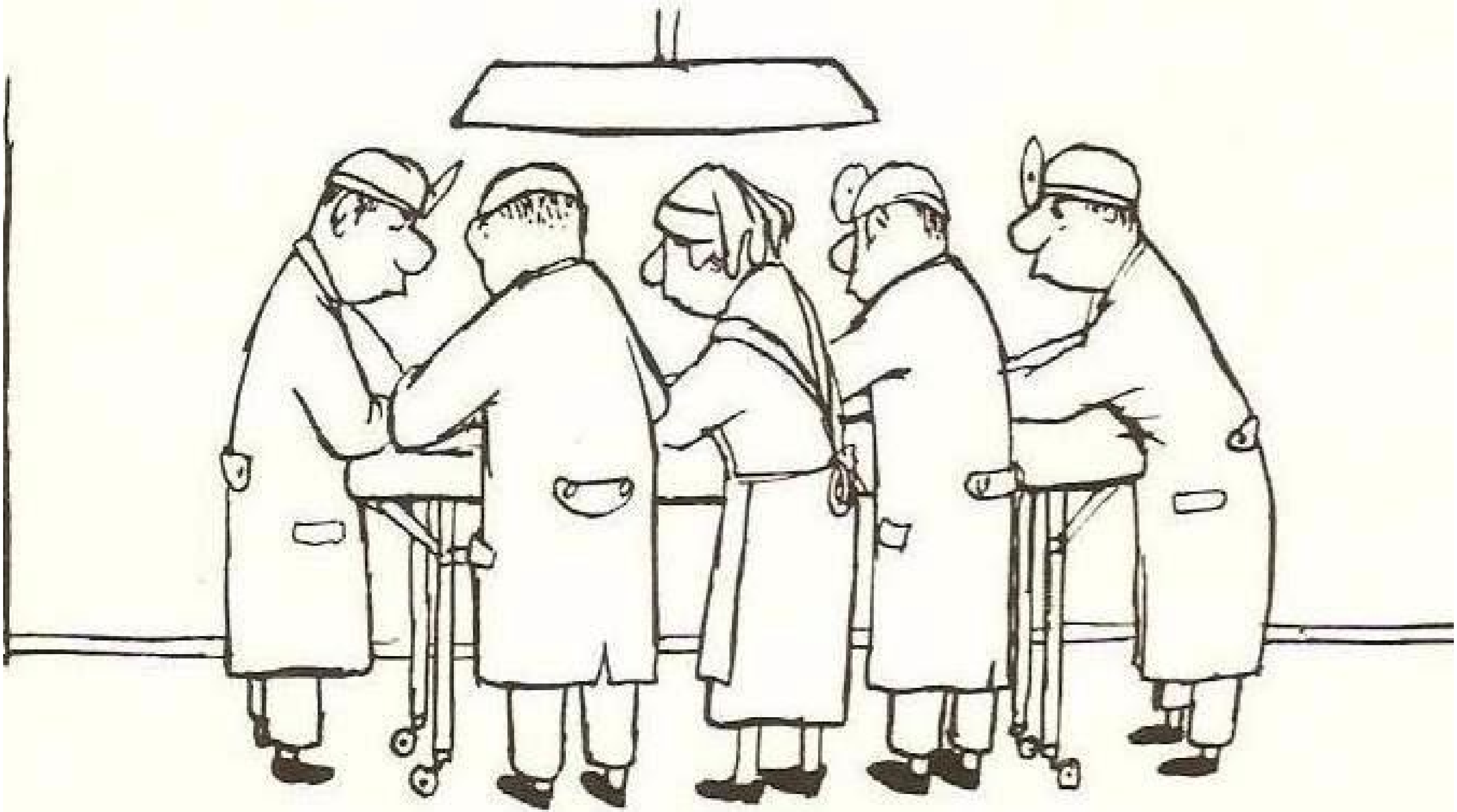
For 10 factors: 10+45 = 55 tests!

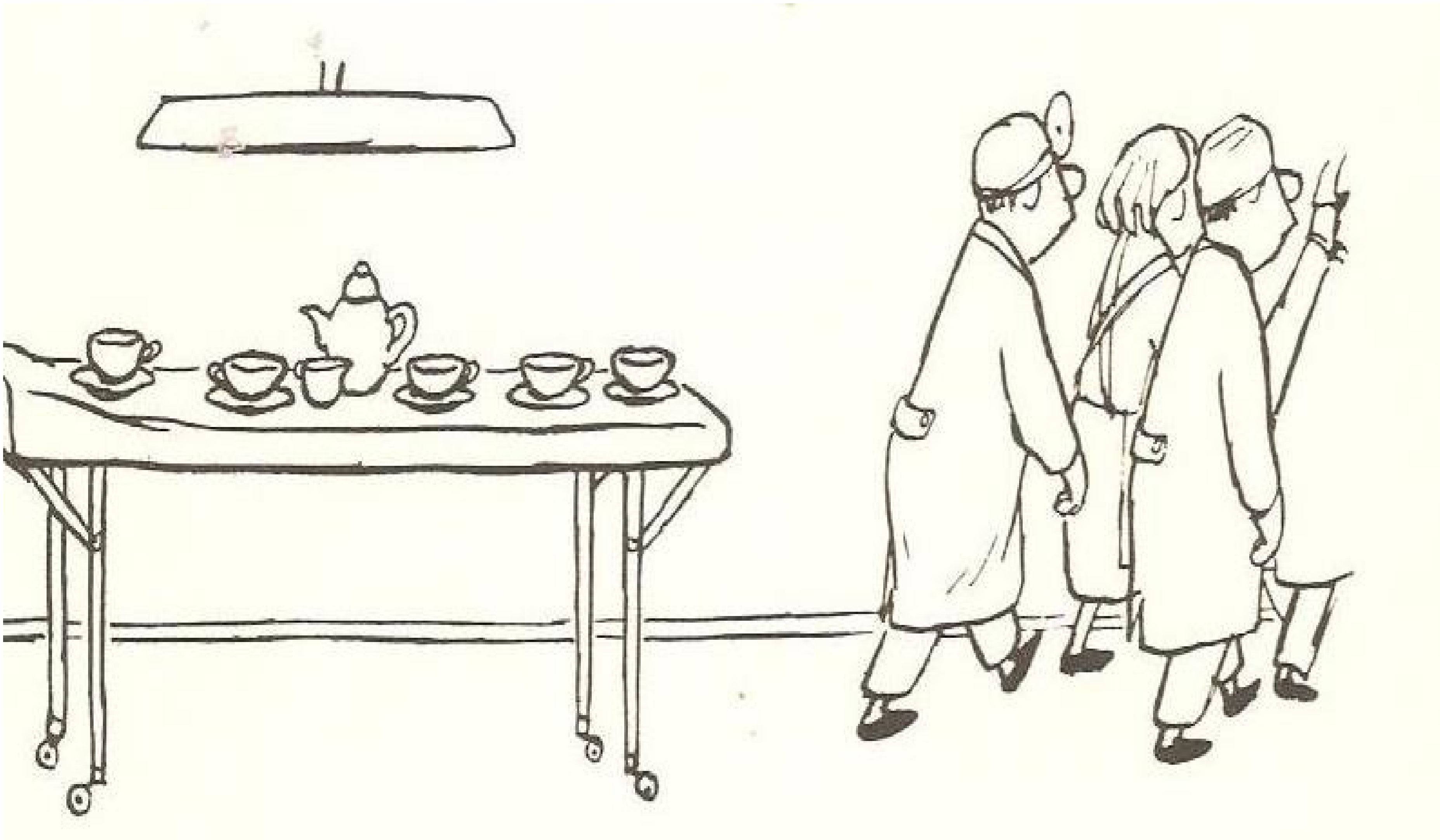
Much better:

Test all single factors the same number (=r) times, AND

Test all combinations of of 2 factors the same number (= $\lambda$ ) times.

How to do this ??? – Use operation tables!





# Incorrect computations help (!)

| + | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 0 |
| 2 | 2 | 3 | 4 | 5 | 6 | 0 | 1 |
| 3 | 3 | 4 | 5 | 6 | 0 | 1 | 2 |
| 4 | 4 | 5 | 6 | 0 | 1 | 2 | 3 |
| 5 | 5 | 6 | 0 | 1 | 2 | 3 | 4 |
| 6 | 6 | 0 | 1 | 2 | 3 | 4 | 5 |

| $\bullet_3$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 1 | 4 | 4 | 2 |
| 2 | 0 | 2 | 4 | 2 | 1 | 1 | 4 |
| 3 | 0 | 3 | 6 | 3 | 5 | 5 | 6 |
| 4 | 0 | 4 | 1 | 4 | 2 | 2 | 1 |
| 5 | 0 | 5 | 3 | 5 | 6 | 6 | 3 |
| 6 | 0 | 6 | 5 | 6 | 3 | 3 | 5 |

„Blocks“ arise:

**block 1: 1,2,4**
**block 2: 2,3,5**
**block 3: 3,4,6**
**block 4: 4,5,0**
**…………….**
**block 7: 0,1,3**

**block 8: 3,5,6**
**block 9: 4,6,0**
**block 10: 5,0,1**
**block 11: 6,1,2**
**………………**
**Block 14: 2,4,5**

**block 1: 1,2,4. So our first test should test factors no. 1, 2, and 4**
**block 2: 2,3,5. So the next test should test factors no. 2, 3, and 5**
**and randomize!**

| Tests Fact. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | x | x | x | | | | | x | x | | x | | |
| 1 | x | | x | | | x | | x | x | x | | | | | |
| 2 | x | x | | | x | x | | | | | x | x | | | |
| 3 | | x | | | x | | x | x | | x | | | | x | |
| 4 | x | | | x | | | x | x | | | | x | x | | |
| 5 | | x | x | | | | | | x | | | x | x | x | |
| 6 | | | | x | | x | x | | x | | x | | | x | |
| | | | | | | | | | | | | | | | |

# Design with results:

| Tests<br>Fact. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | x | x | x | | | | | x | x | | x | | |
| 1 | x | | x | | | x | | x | x | x | | | | | |
| 2 | x | x | | | x | x | | | | | x | x | | | |
| 3 | | x | | | x | | x | x | | x | | | | x | |
| 4 | x | | | x | | | x | x | | | | x | x | | |
| 5 | | x | x | | | | | | x | | | x | x | x | |
| 6 | | | | x | | x | x | | x | | x | | | x | |
| Res. | 69 | 18 | -28 | 3 | 54 | -1 | 51 | 98 | -31 | 49 | -28 | -35 | -25 | 22 | 3 |

Regression gives the best estimates according to as

$$y = 3 + 51x_4 + 19x_5 - 41x_6 \quad \text{.... Model 1}$$

If one also uses interaction terms („synergies“), one gets instead

$$y = 2 + 47x_4 - 31x_6 + 58x_2x_5 \quad \text{.... Model 2}$$

Now we can compare the actual results with the predicted ones using these two models:

| Test | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| real | 49 | -2 | -28 | 3 | 54 | -1 | 51 | 98 | -31 | 69 | 18 | -35 | -25 | 22 | 3 |
| Mod. 1 | 54 | 3 | -38 | 19 | 54 | 3 | 73 | 73 | -38 | 22 | 13 | -19 | -19 | 13 | 3 |
| Mod. 2 | 49 | -2 | -29 | 2 | 49 | 2 | 49 | 107 | -29 | 60 | 18 | -29 | -29 | 18 | 2 |

## Second summary:

So two factors can be dependent; a „synergy“ is much more than just an additive overlay of two factors!

Example: **Food-dependent exercise-induced anaphylaxis:** The contact with some allergens might be harmless, physical exercise can help a lot, while the combination can be disastrous. So one factor is neutral for the patient, the other one positive, but the combination is really negative!