Kapitel VI – Teil 2

Datenauswahl

Two Examples of flawed datasets

- **Classic flawed big data sets**

  – the Literary Digest Poll of 1936
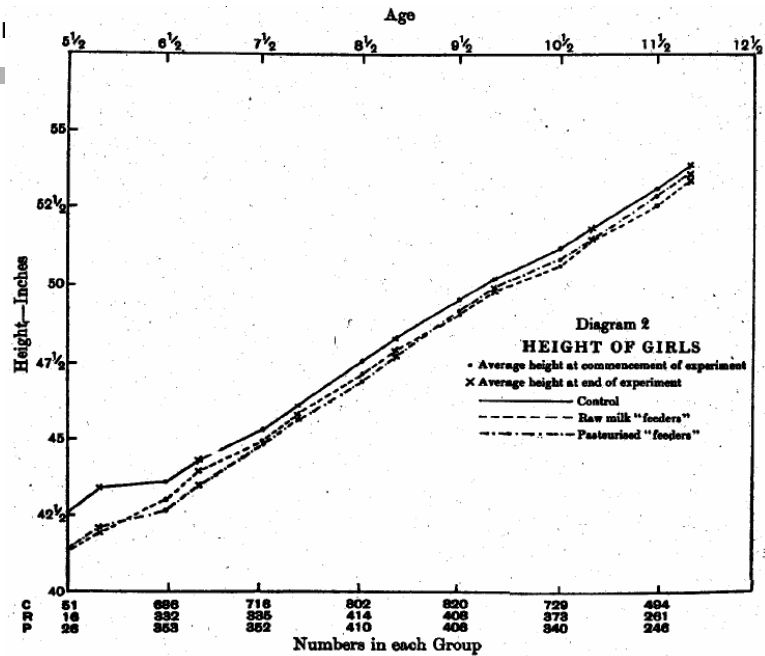
  – the Lanarkshire Milk Experiment of 1930

  Such flaws may be discovered using the techniques described before and/or by checking with common sense/background knowledge!

- **Typical modern flawed big data sets**

  – voluntary surveys by magazines

  – customer data bases ignoring competitor data

Thanks to Antony Unwin for providing these examples.

Literary Digest Poll 1936

Poll for the US elections 1936

- 10 million car owners and telephone subscribers mailed
  - 2.376 million responded:
  - 57% for Alf Landon, 43% for Franklin D. Roosevelt

- Gallup polled 50000, and predicted FDR to win

- Result: FDR 62%, Alf Landon 38%
- Reasons: Biased sample, voluntary response

Lanarkshire Milk Experiment of 1930

- For four months from February to June 1930, in the Scottish county of Lanarkshire 20,000 children, aged between 5 and 12 years, from 67 schools took part in an experiment:

  – 5,000 got raw milk
  – 5,000 got pasteurised milk
  – 10,000 got no milk

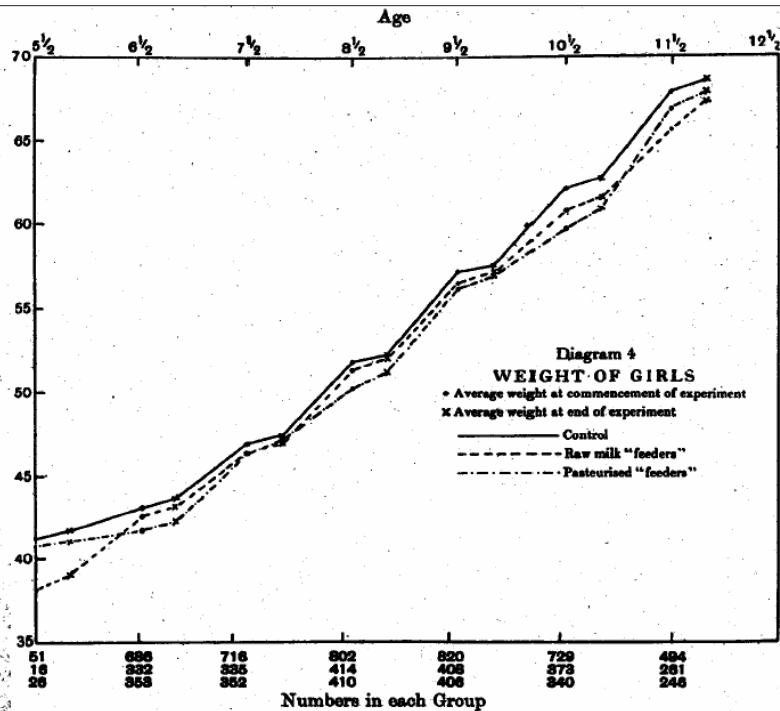- Did milk help growing and, if so, which kind was better?

Diagram 2
HEIGHT OF GIRLS

Vorlesung Knowledge Discovery 5

Lanarkshire Milk Experiment of 1930

Problems with the experiment

- No school got both types of milk

- Allocation by ballot or alphabetically BUT then the teachers could reallocate "to obtain a more level selection"

- Weighed in February (with heavier clothes) and in June (with lighter clothes)

- Controls were analysed as one group

> W.S. Gosset pointed out that a study of the identical twins amongst the group could have been much better controlled and would have given much more reliable results.

→ Small, well-planned studies are often better than large, hard to control ones.

Vorlesung Knowledge Discovery 7

Diagram 4
WEIGHT OF GIRLS

Data Quality

Some aspects influencing the data quality:

- Quality of variables, of definition of variables, of measurement and recording of variables

- Quality of sampling definition, of sampling procedure (choosing, locating, enrolling)

- Quality of representation

- Quality of data checks and balances

- Quality of control of potential influences

Vorlesung Knowledge Discovery 8

Data Quality

Data quality criteria at the Lanarkshire Milk Experiment :

- Weight the best measure of growth? Definition with clothes? Accuracy?

- Schools sampled? (Only big schools)
- Pupils chosen by teachers.
- Allocation to groups ‚corrected‘ by teachers

- Dropouts? Illnesses?

- Height as a check on weight

- Getting milk at home