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

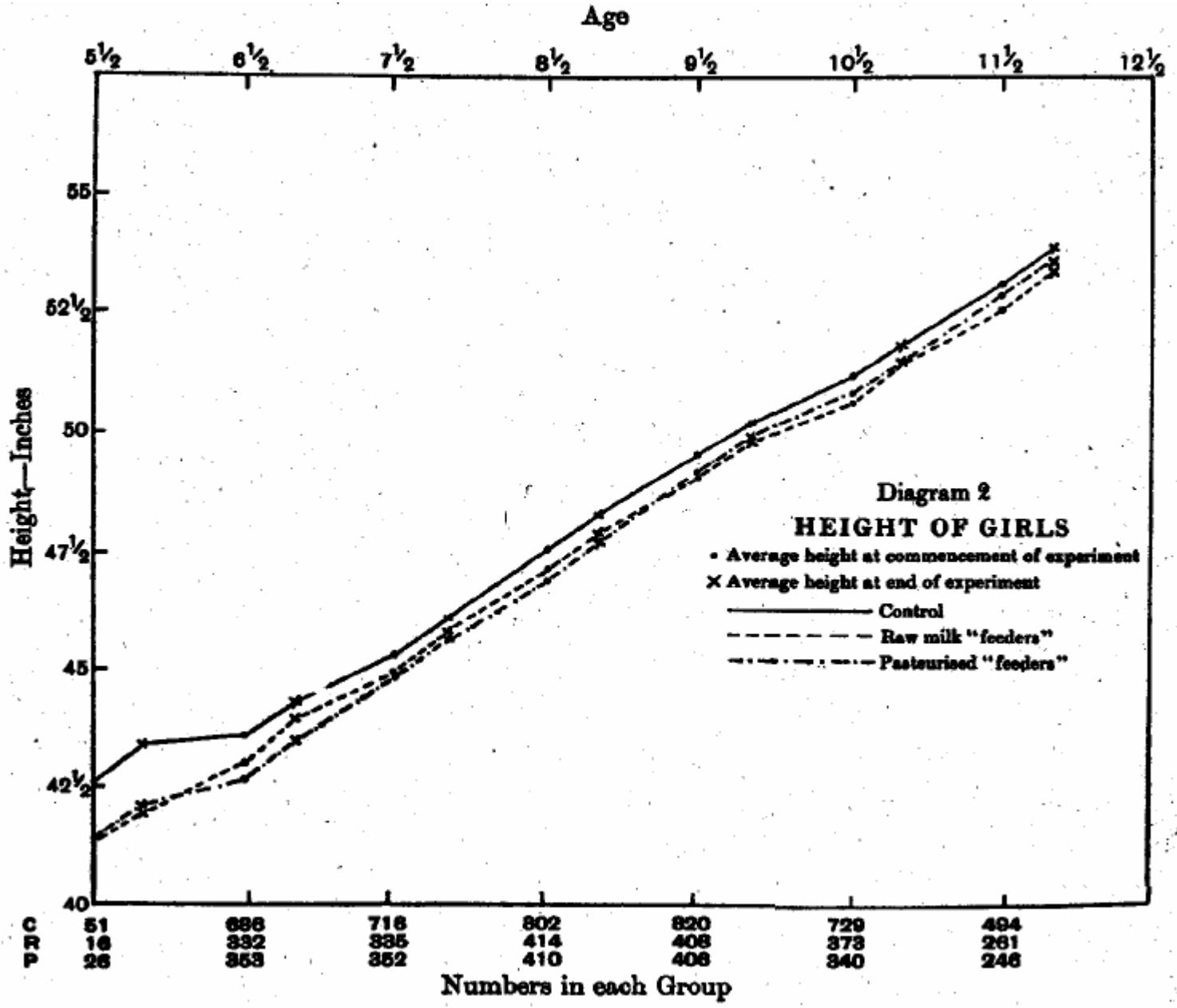
- 10 million car owners and telephone subscribers mailed
  - 2.376 million responded:
  - 57% for Alf Landon, 43% for Franklin D. Roosevelt

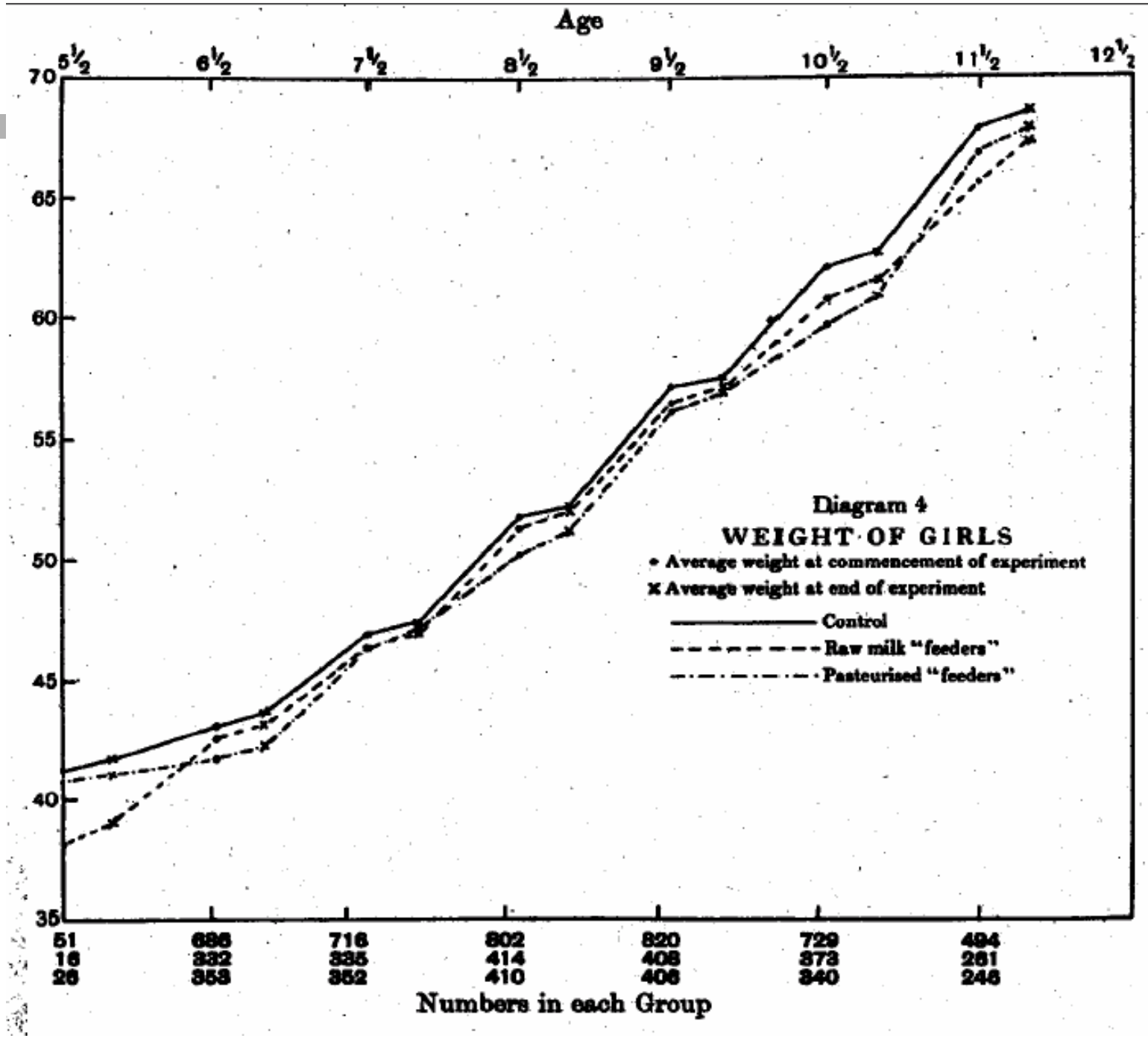
- Gallup polled 50000, and predicted FDR to win

- 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?


Age
5½
6½
7½
8½
9½
10½
11½
12½
55
52½
50
47½
45
42½
40
Height—Inches
Diagram 2
HEIGHT OF GIRLS
• Average height at commencement of experiment
× Average height at end of experiment
Control
Raw milk "feeders"
Pasteurised "feeders"
C 51 686 716 802 820 729 494
R 16 332 335 414 408 373 261
P 26 353 352 410 406 340 246
Numbers in each Group


Age
5½
6½
7½
8½
9½
10½
11½
12½
70
65
60
55
50
45
40
35
Diagram 4
WEIGHT OF GIRLS
• Average weight at commencement of experiment
× Average weight at end of experiment
Control
Raw milk "feeders"
Pasteurised "feeders"
51 686 716 802 820 729 494
16 332 335 414 408 373 261
26 353 352 410 406 340 246
Numbers in each Group

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.

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

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

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