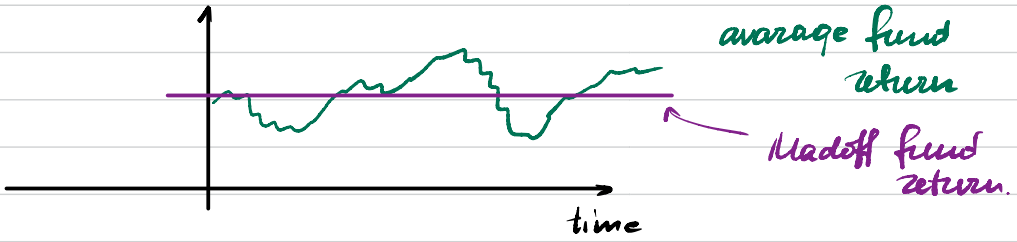
Issues in Applied Stat. & Real Data Analysis

03/09

(1) Data issues

Fabricated data:

- Bernie Madoff, financial fraud
  Investment fund, paid high interest for many years, so many people invested.
  Use new funds provided by new investor to pay dividents.



- Diederik Stapel, Dutch social psycologist
  "the biggest con man in academic science".

In 2011 it was revealed that he had been fabricating data for years, and had published studies based on completely made-up data.
data were "too good to be true": too clean & tidy, no outliers or unusual values (highly unlikely in a real-word dataset).

Reporting issues:

- GlaxoSmithKline company and its Paxil drug.

Data cleaning & Missing Data

	income	height	age	- - -
John Sm.			NN	

outliers

↑ some data missing

a) Full case analysis — keep only complete rows
   i) decrease sample size
   ii) may create bias (there may be a reason for missing data)

b) Imputation
   a) predict missing value in terms of known covariance
   b) replace missing value by the column median
   c) nearest neighbor method

② Confounding effect: association is not causation.

linear model: $y = \beta_0 + \beta_1 x + \varepsilon$. Does x influence y?

Exp.

	lung cancer	no lung cancer
smoking	134	28
non smoking	12	289

Conclude: smoking causes lung cancer.

Exp. Fires

	major damage	minor damage
many trucks	134	28
few trucks	12	289

When many fire tracts were sent, there was major damage, and when few were sent, there was minor damage.

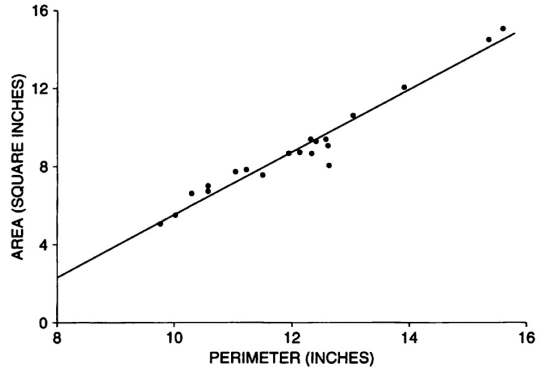
Conclude: sending many fire trucks causes fire damage?

fire site
↙ lots of damage
↘ many truck sent

Exp. Coffee drinkers have a higher proportion of heart attacts.
Can we conclude: coffee causes heart disease?
No: Coffee drinkers are more likely to be smokers

Exp Area & perimeter of a rectangle.

Figure 6. Scatter diagram of area against perimeter for 20 rectangles; the regression line is shown too.



area = (1.60 inches) × (perimeter) − 10.51 square inches

Does the regression make sense?

Statistics, p. 211;
D. Freedman, R. Pisani, R. Purves

Exp. Group of patients with some medical condition are given some medication to help their condition.

Adherance: Degree to which they followed the medication regime. Those that adehered did better then those that did not.

Conclude: The medication helps?
Groups are self selecting.
Those that adhere care more about their health, and do things differently from those in the other group.

Clofibrate trial (p. 13; D. Freedman, R. Pisani, R. Purves; Statistics, 2007)

(3) Design of experiments

Salk vaccine trial. (Gold standart) (for Polio, 1950's) Jonas Salk
treatment group & control group should be identital in every way but treatment.

[D. Freedman, R. Pisani, R. Purves; Statistics, 2007; p. 3]

- divide the group that agree using a fair coin
- placebo administrated in control group (placebo effect)
- patients are "blinded" to treatment
- double blinded – diagnostician was also blinded

Table 1. The results of the Salk vaccine trial of 1954. Size of groups and rate of polio cases per 100,000 in each group. The numbers are rounded.

The randomized controlled double-blind experiment			*The NFIP study*		
	Size	*Rate*		*Size*	*Rate*
Treatment	200,000	28	Grade 2 (vaccine)	225,000	25
Control	200,000	71	Grades 1 and 3 (control)	725,000	54
No consent	350,000	46	Grade 2 (no consent)	125,000	44

consent

Source: Thomas Francis, Jr., "An evaluation of the 1954 poliomyelitis vaccine trials—summary report," *American Journal of Public Health* vol. 45 (1955) pp. 1–63.

How "no concent" group was "different"? (compared to control group)

Table 1 also shows how the NFIP study was biased against the vaccine. In the randomized controlled experiment, the vaccine cut the polio rate from 71 to 28 per hundred thousand. The reduction in the NFIP study, from 54 to 25 per hundred thousand, is quite a bit less. The main source of the bias was confounding. The NFIP treatment group included only children whose parents consented to vaccination. However, the control group also included children whose parents would not have consented. The control group was not comparable to the treatment group.

(4) Sampling Bias

- Presidential election 1936 (Roosevelt vs Landon)

poll done by Literary Digest

	predicted	actual	sample size
Landon	57%	38%	2,400,000
Roosevelt	43%	62%	

How did the literary digest get their sample?

To find out where the *Digest* went wrong, you have to ask how they picked their sample. A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross section of the public. A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*. The *Digest*'s procedure was to mail questionnaires to 10 million people. The names and addresses of these 10 million people came from sources like telephone books and club membership lists. That tended to screen out the poor, who were unlikely to belong to clubs or have telephones. (At the time, for example, only one household in four had a telephone.) So there was a very strong bias against the poor in the *Digest*'s sampling procedure. Prior to 1936, this bias may not have affected the predictions very much, because rich and poor voted along similar lines. But in 1936, the political split followed economic lines more closely. The poor voted overwhelmingly for Roosevelt, the rich were for Landon. One reason for the magnitude of the *Digest*'s error was selection bias.

> When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.

The *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey

The *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey

[D. Freedman, R. Pisani, R. Purves; Statistics, 2007; p. 334]

organization still has to get their opinions. This is harder than it looks. If a large number of those selected for the sample do not in fact respond to the questionnaire or the interview, *non-response bias* is likely.

The non-respondents differ from the respondents in one obvious way: they did not respond. Experience shows they tend to differ in other important ways as well.[5] For example, the *Digest* made a special survey in 1936, with qucstionnaires mailed to every third registered voter in Chicago. About 20% responded, and of those who responded over half favored Landon. But in the election Chicago went for Roosevelt, by a two-to-one margin.

> Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.

In the main *Digest* poll, only 2.4 million people bothered to reply, out of the 10 million who got the questionnaire. These 2.4 million respondents do not even represent the 10 million people who were polled, let alone the population of all voters. The *Digest* poll was spoiled both by selection bias and non-response bias.[6]

Special surveys have been carried out to measure the difference between respondents and non-respondents. It turns out that lower-income and upper-income people tend not to respond to questionnaires, so the middle class is over-represented among respondents. For these reasons, modern survey organizations prefer to use personal interviews rather than mailed questionnaires. A typical response rate for personal interviews is 65%, compared to 25% for mailed questionnaires.[7] However, the problem of non-response bias still remains, even with personal interviews. Those who are not at home when the interviewer calls may be quite different from those who are at home, with respect to working hours, family ties, social background, and therefore with respect to attitudes. Good survey organizations keep this problem in mind, and have ingenious methods for dealing with it (section 6).

> Some samples are really bad. To find out whether a sample is any good, ask how it was chosen. Was there selection bias? non-response bias? You may not be able to answer these questions just by looking at the data.

- survival bias

Exp. Aircraft returned after bombing runs in WW2

Area	average # of bullet holds	
→ Engine	1.11	
Fusalage	1.72	} should add extra protection here?
Fuel system	1.55	
Rest plane	1.8	

(5) Simpson's paradox

Exp. Gender bias in grad. admission in UC Berkeley 1973

[D. Freedman, R. Pisani, R. Purves, Statistics, 2007; p. 17]

Applicants to graduate school		acceptence rate
8,442	men	44%
4 321	women	35%

difference 11%, gender-bias??

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

	Men		*Women*	
Major	*Number of applicants*	*Percent admitted*	*Number of applicants*	*Percent 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	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.