
8. Case studies from 'Group C'

- The case studies of 'Group C' are research-oriented case studies where it is hard to pick up the statistical (analytics) problem but where you have to do a lot of thinking and the answer may or may not be well known.
 - ~> These case studies require multivariate data analysis methods.
 - ~> Several stages of analysis may also be needed to obtain suitable results.
 - ~> There may not necessarily be an 'answer' to the statistical problem.
- ◇ In what follows, we present six case studies.

C.1 Plastic explosives detection

- One of the most effective devices for detecting plastic explosives is a type of X-ray scanner that produces a profile of the chemical composition of a small area inside, for example, a suitcase.

~> If the profile shows a pattern similar to one of the known explosives then the suitcase is classified as a 'bomb'.

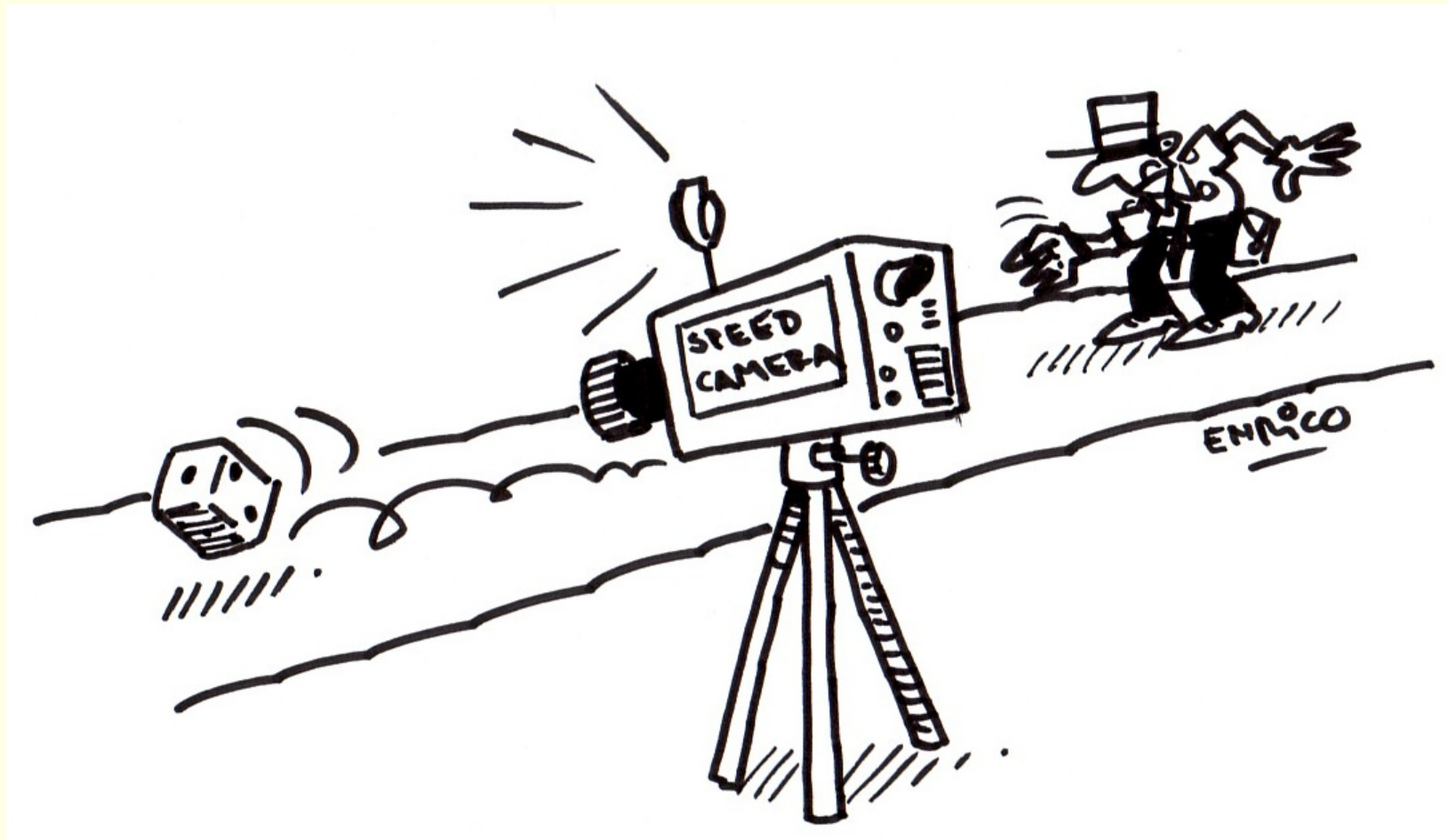
~> The emphasis in this case study is on the reliability of plastic explosive detection based on an early X-ray machine prototype.

~> How well can plastic explosive be detected? What type of decision rule should be used?

Data

- The client performed an experiment where 3'053 profiles were obtained, of which 1'108 corresponded to explosive substances and the remaining 1'945 were from typical substances found in suitcases.
- Source:

```
www.statoo.com/DATA/AC/Cexplosives.xls
```
- Size: 3'053 rows (*i.e.* suitcases), 24 columns (*i.e.* variables).
- Variables:
 - Y: is 1 if the suitcase has a 'bomb', and 0 otherwise;
 - X1, ..., X23: signal profile — each profile is a vector of 23 numbers which are a summary of the signal absorbed by the material.



C.2 Head injuries

- There is much controversy about the use of 'Computed Tomography' (CT) for patients with 'minor' head injury.
 - Note that head injury is classified as 'minor' if the so-called 'Glasgow Coma Scale' (GCS) is ≥ 13 (or $= 13$ herein) — the GCS being a neurological scale used to assess level of consciousness after head injury.
- ↪ The lowest possible GCS is 3 (deep coma or death), while the highest is 15 (fully awake person).

-
- A client wants to develop a highly sensitive clinical decision rule for use of CT in patients with ‘minor’ head injuries.
 - ~> Such a rule could have the potential to significantly standardise and improve the emergency management of patients with ‘minor’ head injury.
 - ~> To do so, the client has data on the clinical characteristics of a sample of 3’121 individuals who suffered ‘minor’ head injuries.
 - ~> You are asked to predict occurrence of clinically important brain injury as revealed on CT, given the other clinical characteristics.
 - ~> What type of decision rule could be used? What are the ‘high-risk’ factors?

Data

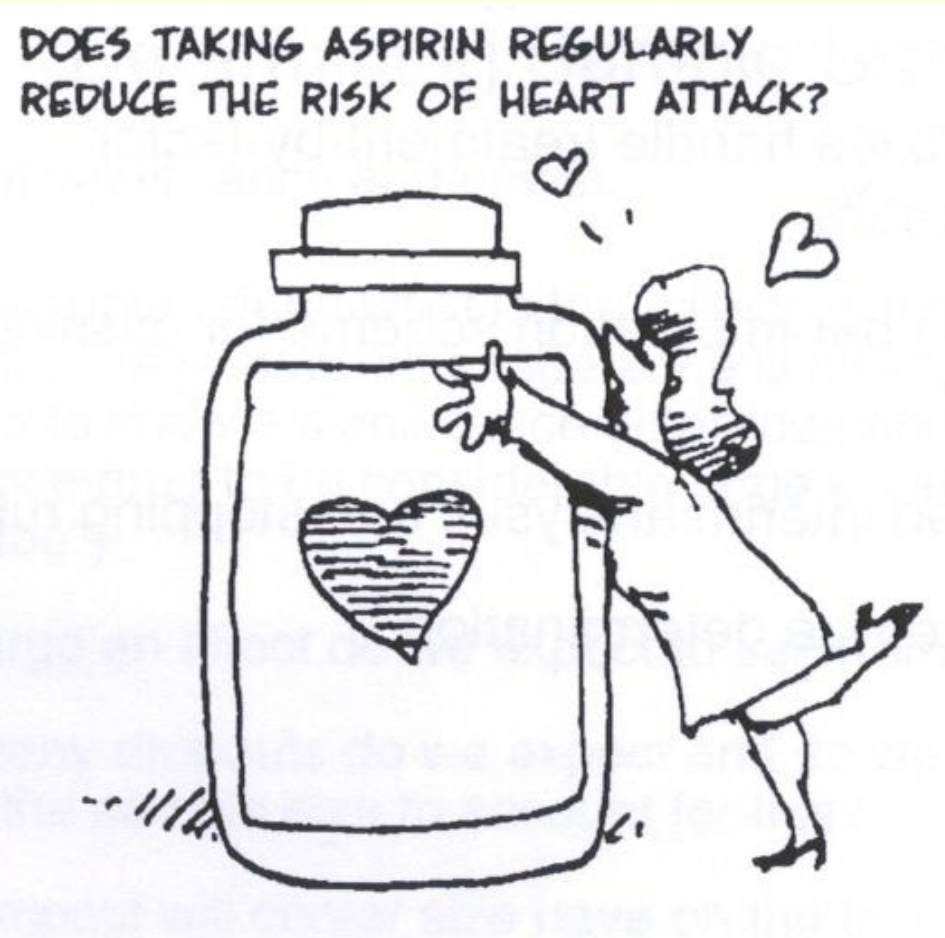
- Source:

`www.statoo.com/DATA/AC/Cheadinjury.xls`

- Size: 3'121 rows (*i.e.* individuals), 11 columns (*i.e.* variables).

- Variables:

- `age.65`: age factor (0 = 'under 65', 1 = 'over 65');
- `amnesia.before`: amnesia before impact (0 = 'less than 30 minutes', 1 = 'more than 30 minutes');
- `basal.skull.fracture`: (0 = 'no fracture', 1 = 'fracture');
- `GCS.decrease`: GCS decrease (0 = 'no deterioration', 1 = 'deterioration');
- `GCS.13`: initial GCS (0 = 'not 13', 1 = '13');
- `GCS.15.2hours`: GCS after 2 hours (0 = 'not 15', 1 = '15');
- `high.risk`: assessed by clinician as high risk for neurological intervention (0 = 'not high risk', 1 = 'high risk');
- `loss.of.consciousness`: 0 = 'conscious', 1 = 'loss of consciousness';
- `open.skull.fracture`: 0 = 'no fracture', 1 = 'fracture';
- `vomiting`: 0 = 'no vomiting', 1 = 'vomiting';
- `clinically.important.brain.injury`: any acute brain finding revealed on CT (0 = 'not present', 1 = 'present').



C.3 AIDS study

- Measuring the 'cell count' of particular cells provides an effective means of monitoring patients who are affected by the AIDS virus, or have diseases such as cancer or hepatitis.
- For someone who is HIV-positive, two important diagnostics are their CD4 and CD8 'cell counts'.

◇ CD4 are white blood cells that the AIDS virus uses as a host to reproduce itself.

↪ Hence CD4 cell counts provide a key indicator of a person's immune status:

- below 200: full-blown AIDS;
- 200 to 500: intermediate stage;
- above 500: sound functioning immune system.

◇ CD8 cells help suppress the infectiousness of the virus by killing cells the body decides are foreign. Unfortunately, the AIDS virus itself evades detection by residing within the CD4 cell.

↪ However, the CD8 cell count provides a measure of the person's ability to fight off other infections.

◇ Another measure of the viral 'load' carried by a person is their RNA count.

↪ This is a single strand of DNA which the AIDS virus uses to reproduce itself.

-
- The purpose of this case study is to see if these three measures (*i.e.* CD4, CD8 and RNA counts) provide ‘discrimination’ between two groups of couples classified as ‘Discordinant’ (*DP*: only one partner HIV-positive) and ‘Concordinant’ (*CP*: both HIV-positive).

⇒ Only one partner from each couple was included in the study, with the infected partner being measured in the *DP* group.

⇒ This provided a more homogeneous cohort, and to eliminate confounding effects, drug users and nonmonogamous couples were excluded.

Data

- Source:

`www.statoo.com/DATA/AC/Caids.xls`

- Size: 278 rows (*i.e.* patients), 5 columns (*i.e.* variables).

- Variables:

- sex: gender of patient (f = 'female', m = 'male');
- type: 'CP' if both patient and their partner are infected, 'DP' if patient's partner is not infected;
- cd4, cd8, rna: cell counts of patient.



C.4 Credit rating

- A banking client provides you with data containing information concerning the credit rating of 4'017 individuals.

~> The client's primary question is to know whether it is possible to predict 'credit rating' (*i.e.* whether or not an individual is a 'bad risk' customer) using the other information in the data.

~> How well can credit rating be detected? What type of decision rule should be used?

Data

- Source:

`www.statoo.com/DATA/AC/Crisk.xls`

- Size: 4'017 rows (*i.e.* individuals), 10 columns (*i.e.* variables).

- Variables:

- AGE: age in years;
- INCOME: annual income (in dollars);
- GENDER: gender (f = 'female', m = 'male');
- MARITAL: marital status ('single', 'married' or 'divsepwid');
- NUMKIDS: number of kids;
- NUMCARDS: number of cards;
- HOWPAID: how paid ('monthly' or 'quarterly');
- MORTAGE: mortgage (y = 'yes', n = 'no');
- LOANS: number of existing loans;
- RISK: indication whether or not an individual is a 'bad risk' customer ('bad risk' or 'good risk').



C.5 Assessing telecommunications churn risk

- Your client is a telecommunications company who provides you with data for 5'000 customers.
 - ↪ The client's primary question is to know whether it is possible to predict 'customer churn' (also known as 'customer attrition', *i.e.* whether or not customers discontinue doing business with the telecommunications company) using the other information in the data.
 - ↪ How well can customer churn be detected? What are the most important 'drivers' for churn?

Data

- Source:

`www.statoo.com/DATA/AC/Cchurn.xls`

- Size: 5'000 rows (*i.e.* customers), 18 columns (*i.e.* variables).

- Variables:

- `account_length`: duration of the customer relationship (in weeks);
- `international_plan`: international plan ('yes' or 'no');
- `voice_mail_plan`: voice mail plan ('yes' or 'no');
- `number_vmail_messages`: number of voice mail messages;
- `total_day_minutes`: total number of day-time national call minutes;
- `total_day_calls`: total number of day-time national calls;
- `total_day_charge`: total number of chargeable day-time national call minutes;

-
- `total_eve_minutes`: total number of evening national call minutes;
 - `total_eve_calls`: total number of evening national calls;
 - `total_eve_charge`: total number of chargeable evening national call minutes;
 - `total_night_minutes`: total number of night-time national call minutes;
 - `total_night_calls`: total number of night-time national calls;
 - `total_night_charge`: total number of chargeable night-time national call minutes;
 - `total_intl_minutes`: total number of international call minutes;
 - `total_intl_calls`: total number of international calls;
 - `total_intl_charge`: total number of international chargeable call minutes;
 - `number_customer_service_calls`: total number of calls to the customer service;
 - `churn`: indication whether or not the customer churned ('yes' or 'no').



C.6 Sales of orthopedic equipment

- An orthopedic equipment manufacturer provides you with data containing information from 4'703 hospitals concerning the sales of their orthopedic material to these hospitals.

~> The client's objective is to find ways to increase sales of orthopedic material from their company to hospitals.

Data

- Source:

`www.statoo.com/DATA/AC/Cortho.xls`

- Size: 4'703 rows (*i.e.* hospitals), 15 columns (*i.e.* variables).

- Variables:

- BEDS: number of hospital beds;
- RBEDS: number of rehabilitation beds;
- OUTV: number of outpatient visits;
- ADM: administrative costs (in 1'000's dollars per year);
- SIR: revenue from inpatient;
- SALESY: sales (in 1'000's dollars) of rehabilitation equipment for the current year;
- SALES12: sales (in 1'000's dollars) of rehabilitation equipment for the previous year;
- PRHIP: number of hip operations (total for the previous year);
- PRKNEE: number of knee operations (total for the previous year);
- TH: teaching hospital (0 = 'no', 1 = 'yes');
- TRAUMA: presence of a trauma unit in the hospital (0 = 'no', 1 = 'yes');
- REHAB: presence of a rehabilitation unit in the hospital (0 = 'no', 1 = 'yes');
- HIP: number of hip operations (total for the current year);
- KNEE: number of knee operations (total for the current year);
- FEMUR: number of femur operations (total for the current year).

