Data organisation

Heike, Hofmann, Hadley Wickham

Outline

- Normalisation
- Practice

Data Tables

- Records in rows
- Variables in columns (put id variables first)
- Good idea to use freeze panes in big spreadsheets

Name	Date	Weight	Sex
Jane Doe	1 Jan	160	Female
Jane Doe	1 Feb	150	Male
Jane Doe	1 Mar	140	Female

Name	Date	Weight	Sex
Jane Doe	1 Jan	160	Female
Jane Doe	1 Feb	150	Male
Jane Doe	1 Mar	140	Female

Normalisation

- Minimise redundancy and inconsistency
- Each “fact” stored only once
- May not be optimal for entry or analysis
- Multiple datasets, each containing information about one “entity”

Normal forms

- 5 different levels of normalisation
- We'll look at first three

Less duplication

More consistent

Normalisation

Small pieces joined together

Harder to edit

Harder to view

Tables must be rectangular

1st normal form

- Data is in table
- Each record has to be unique / identifiable
- each entry has to be atomic (number, string, date, ...)

1st normal form

- Data is in table
- Each record has to be unique / identifiable
- each entry has to be atomic (number, string, date, ...)

Keys

- Key: set of fields that identify a record (so must be unique)
- Like indices on a random variable
- May be single or composite
- Fixed by design of experiment (known in advance)
- As compared to measurements/measured variables

Identify the keys

- 100 patients randomly assigned treatment for heart attack, measured 5 different clinical outcomes.

Identify the keys

- 100 patients randomly assigned treatment for heart attack, measured 5 different clinical outcomes.

Identify the keys

- Randomised complete block trial with four fields, four different types of fertiliser, over four years. Recorded total corn yield, and fertiliser run off

Identify the keys

- Randomised complete block trial with four fields, four different types of fertiliser, over four years. Recorded total corn yield, and fertiliser run off

Identify the keys

- Cluster sample of twenty students in thirty different schools. For each school, recorded distance from ice rink. For each student, asked how often they go ice skating, and whether or not their parents like ice skating

Identify the keys

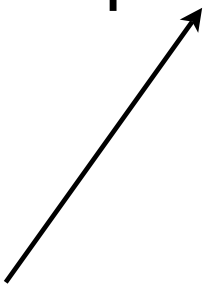
- Cluster sample of twenty students in thirty different schools. For each school, recorded distance from ice rink. For each student, asked how often they go ice skating, and whether or not their parents like ice skating

Identify the keys

- For each person, recorded age, sex, height and target weight, and then at multiple times recorded their weight

Identify the keys

- For each person, recorded age, sex, height and target weight, and then at multiple times recorded their weight



Need a composite key!

Relationships

- Keys identify an object
- And allow us to see relationships in the data
- 1-to-1, 1-to-many, many-to-many

The key, the whole key and nothing but the key

2nd normal form

- table has to be in 1st normal form and
- **whole key** necessary to describe non-key field in table

- Violated when: Fact is about a subset of a key (in composite keys)
- To fix: create another table

Person	**Date**	Weight	Sex
James	1 Jan	205	Male
James	1 Feb	195	Male
James	1 Mar	190	Male

Person	**Date**	Weight	Sex
James	1 Jan	205	Male
James	1 Feb	195	Male
James	1 Mar	190	Female

Person	**Date**	Weight	Age
James	1 Jan	205	20
James	1 Feb	195	20
James	1 Mar	190	20

3rd normal form

- tables are in 2nd normal form and
- non-key fields are not facts about other non-key fields

Person	Zip Code	State
Hadley Wickham	50014	Iowa
John Smith	90210	California
...	...	...

Practice

Download file data-organisation-practice.xls from
streaming.stat.iastate.edu/~stat480
and transfer into 2nd normal form.

Make notes, which fields are violating the 3rd normal form