Cannlytics

Cannabis Data Science

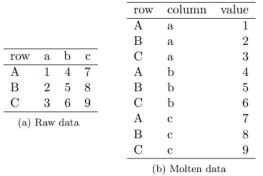
Saturday Morning Statistics #16

March 19th, 2022

The State of Data Science

Avenues for advancing **data tidying**:

- Correcting character encodings; ✓
- Parsing dates and numbers; ✓
- Identifying missing values; ✓
- Matching similar but not identical values;
- Filling in structural missing values;
- Model-based data cleaning.



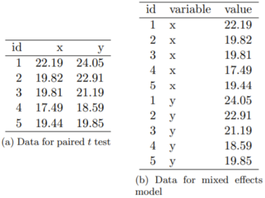
row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Melting data.



id	x	y
1	22.19	24.05
2	19.82	22.91
3	19.81	21.19
4	17.49	18.59
5	19.44	19.85

(a) Data for paired t test

id	variable	value
1	x	22.19
2	x	19.82
3	x	19.81
4	x	17.49
5	x	19.44
1	y	24.05
2	y	22.91
3	y	21.19
4	y	18.59
5	y	19.85

(b) Data for mixed effects model

Pairing data.

Reference: Tidy Data, Hadley Wickham, Journal of Statistical Software (2014).

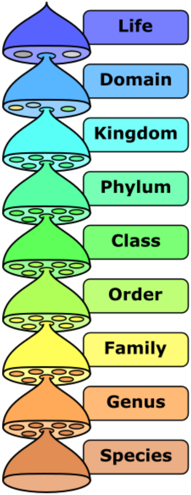
A Peek at Scientific History

Taxonomy is the scientific study of naming, defining, and classifying groups of biological organisms based on shared characteristics.

Taxonomic characteristics used to differentiate **taxa** include:

- Morphological;
- Physiological;
- Molecular;
- Behavioral;
- Ecological;
- Geographic.

A **strain** is a genetic variant, a subtype or a culture within a biological species.



Application to cannabis research

The Indica / Sativa Dichotomy

Jean-Baptiste de **Lamarck** (1744 - 1829)
Notable naturalist, biologist, and taxonomer.
Collector of rare plants.

- Named *Cannabis indica* (the 2nd cannabis species).
  - Hindu Kush mountain range;
  - Temperate climates;
  - The *botanical defence* (1970s).

- **Claim:** *Cannabis indica* strains tend to have higher THC content than *Cannabis sativa* strains (Fischedick et. al 2010, Hillig and Mahlberg 2004).

- **Claim:** Known indica strains include *Kush*, *Northern Lights*, and *Purple Kush*.

New born Cannabis plants (2017).
Author: Mar11, License: CC BY-SA 4.0
https://creativecommons.org/licenses/by-sa/4.0

Question and Hypothesis

Question of the day.

- Can we build a model to predict a cannabis product's **strain** given a readily available factors, such as:
  - If the product is a *Kush*.
  - If the product is purple.
  - The THC concentration of the product, perhaps relative to the CBD concentration.

Methodology: Probit Models

Given a latent variable representation of the **probit model**:

$$z_i = x_i\beta + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1),$$

$$y_i = \begin{cases} 1 \text{ if } z_i > 0 \\ 0 \text{ if } z_i \leq 0 \end{cases}$$

You can estimate the parameters using the **likelihood function**

$$L(\beta) = \prod_{i=1}^{n} \Phi(x_i\beta)^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}.$$

🙏 Thank you for coming.

Lesson of the Day

- Names are powerful.