Golem of Prague

2022-01-05

Creatures:

- Golem
- Owls
- Dogs/DAGs

The Golem of Prague

- Created to defend the Jews, but actually did not performed.
- Golem as a metaphor of statistical modeling:
  - Clay robots: Computer based, joke with **sylicon**
  - Powerful: good to make what people can't
  - No wisdom or foresight: rather dumb
  - Dangerous: must be interpreted
- A heuristic chart perform well on industrial frameworks, but the statistical tests are designed to reject null hypotheses instead of research hypotheses.
- Karl Popper, science of falsificationism: the falsification is on the research hypotheses, not on the null hypotheses
- There is no clear relationship between research hypotheses and tests, for they are designed in relation to null hypotheses

An example: evolutionary genetics

(Fig 1.2)

- Hypotheses
  - H0: "Evolution is Neutral"
  - H1: "Selection Matters"

The hypotheses are vague, so one must process it in a scientific model, that has logical causation, instantiated entities etc.

- Processes models:
  - P0A Neutral equilibrium: no selection, stable population size
  - P0B Neutral, non-equilibrium: no selection, unstable population size
  - P1A Contant selection: a trait is good and always good
  - P1B Fluctuating selection: a trait is good sometimes

But P0A and P1B generate statistical distributions similar between them, so the distribution of alleles are not good to differentiate them.

- Statistical Models:
  - MI
  - MII: ex. distribution of alleles
  - MIII

Null models rarely unique

- Null phylogeny?
- Null ecological community?
- Null social network?

In these cases there is no unique null!

Hypotheses and Models

- We need more than tiny null robots, precise models and models (procedure, golems) justified by implications of process models and questions (estimand).

Owls

Lots of steps to drawn an owl, but the basic and final given.

We will code, in detail, step by step.

```
p_grid <- seq(from=0, to=1, length.out=1000)
prob_p <- rep(1,1000)
prob_data <- dbinom(6, size=9, prob=p_grid)
posterior <- prob_data * prob_p
posterior <- posterior / sum(posterior)
```

- Three modes:
  1. Undestand what you are doing
  2. Document your work, reduce error, reuse
  3. Respectable scientific workflow

Drawing the Bayesian Owl

1. Theoretical estimand: what you are trying to do?
2. Scientific (causal) model(s): models that can produce data and syntactic observations
3. Use 1 & 2 to build statistical model(s): that can get the estimand or let us know it is even possible
4. Simulate from 2 to validate whether 3 yields 1
5. Analyse real data: to come back home! How to back up if data was messed?

What is a Bayesian Owl?

It is a flexible approach. Galileo's telescope was bad so Saturn looked like oOo (ball with ears). What generates the blurry data? How Saturn looks like.

- The Bayesian approach is permissive, flexible, it does not care if the uncertainty is by sampling variation or light scattering
- Express uncertainty at all levels
- Direct solutions for measurement error and missing data
- Focus on scientific modeling: you should not spend time wondering about the statistical estimator, the only estimator is the posterior.

DAGs

- Bayes vs. Frequentism does not matter, what matters is the causal inference.
- Bayes is better, but who cares.
- Science before statistics:

For **statistical models** to produce scientific insight, they require additional **scientific (causal) models**

The **reasons** for a statistical analysis are not found in the data themselves, but rather in the **causes** of the data

The **causes** of the data cannot be extracted from the data alone. *No causes in, no causes out*

- Causal inference, description and designed: All the same task

  Even when the goal is **descriptive**, need causal model

  The **sample** differs from the **population**; describing the population requires causal thinking

What is causal inference?

Causal inference is the attempt do understand the causal model using data

It requires more than association between variables

Causal inference is **prediction** of intervention

Causal inference is **imputation** of missing observations

- Prediction: knowing a **cause** means being able to predict the **consequences** of an **intervention**; "What if I do this?"
  - The wind causes a tree to move, but only by predicting what would happen if we intervene on a given variable in this system that is possible to infer the right causation.
- Imputations: knowing a **cause** means being able to construct unobserved **counterfactual outcomes**: "What if I had done something else?"

Direct Acyclic Graphs

- heuristic causal models
- clarify scientific thinking
- analyze to deduce appropriate statistical models
- much more as the course develops

Imagine we want to relate a variable X to an outcome Y. The relation $X \to Y$ is not enough, because:

- $A \to X$
- $B \to Y$

So it is not clear which variable put on the model. Besides,

- $C \to X$
- $C \to Y$,

So C is a cofound, but also:

- $A \to C$
- $B \to C$

Why DAGs?

- Different queries, different Models
- Which control variables?
- Not safe to add everything – **BAD CONTROLS**
- How to test the causal model?
- With more scientific knowledge, can do more.

Golems, Owls, DAGs

- Golems: brainless, powerful statistical models
- Owls: documented, objective procedures
- DAGs: transparent scientific assumptions to:
  - **justify** scientific effort
  - **expose** it to useful critique
  - **connect** theories to golems