εstimand

Estimation and hypothesis testing

Dr Gianluca Campanella

12th May 2016

Contents

Where we are

1. Define the **research question**
2. **Get** the data
3. **Explore** the data
   - (Re)format, clean, merge, stratify...
   - Identify trends and outliers
4. **Model** the data
   - Select and build model(s)
   - Evaluate and refine model(s)
5. **Summarise** the results
   - Summarise findings
   - Describe assumptions and limitations
   - Identify follow-up research questions

Statistical problem solving

1. Formulate the problem (**research question**)
2. Define **population** and **sample**
3. **Collect** the data
4. **Describe** the data
    - Graphs and tables
    - Descriptive measures
5. Make **inferences**
    - Point and interval estimation
    - Hypothesis testing
6. **Report** results

Inference

Desire to generalise

from a random sample to a population
(from which the sample was selected)

- Estimation (including uncertainty quantification)
- Hypothesis testing

Estimation

Population and sample

Population

The **entire collection of units** possessing one or more characteristics we wish to understand (depends on the research question)

Sample

A **representative subset** of units for which we collect information (known as **observations**) that is then used to **estimate** characteristics of the whole population

Estimation

Point estimation
One value summarises the characteristic of interest

Interval estimation
Two values (an interval), usually together with a point estimate, summarise the characteristic of interest and the **uncertainty** around the estimate

Exercise: sampling

If we draw two samples from the same population, will we always reach the same conclusions?

Exercise: sampling

If we draw two samples from the same population, will we always reach the same conclusions?

No!

- Sampling variability introduces **uncertainty** in our estimates
- What happens if we repeat the experiment over and over again?

Quantifying uncertainty: confidence intervals

- **Observed** (may change from sample to sample)
- Defined such that, **were the sampling repeated multiple times**, the proportion of CIs that contain the population-level value would match a certain frequency known as confidence level
  (Note that there is no such thing as the 'probability of containing the population-level value' within any given confidence interval)
- 95% or 99% confidence levels are typical

Example: confidence intervals

We will ‘verify’ that confidence intervals do what they say on the tin by simulation

Simulating CIs from the normal distribution

- Define normal population with known μ and σ
- Repeat:
    - Sample from population without replacement
    - Compute confidence interval for the mean
    - If μ falls within the CI, increment counter

Hypothesis testing

Exercise: A/B testing

1. Divide into groups
2. Using data on the following slide, **discuss**: does the nicer packaging make customers less likely to cancel their subscriptions?

Exercise: A/B testing

	Cancelled	Didn't cancel	Total
New packaging	168	271	439
Old packaging	175	267	442
Total	343	538	881

New 168 out of 439 (38.27%) cancelled

Old 175 out of 442 (39.59%) cancelled

$\rightarrow$ Probability of cancelling given the new packaging is 3.34% lower

Exercise: A/B testing

1. Divide into groups
2. Using data on the following slide, **discuss**: does the nicer packaging make customers less likely to cancel their subscriptions?
3. **Read** the blog post at `https://www.candyjapan.com/results-from-box-design-ab-test`

Hypothesis testing

How do we decide
in the face of uncertainty?

Remember that we already know
how to **quantify** uncertainty…

Hypothesis testing

1. Simplify the question into two competing claims:
    - Null hypothesis H_0
    - Alternative hypothesis H_1
2. Outcome of hypothesis testing is either:
    - ‘Reject H_0’ (in favour of H_1)
    - ‘Do not reject H_0’

H_0 is usually the hypothesis we wish to **disprove**, and the test is set up so that it cannot be rejected unless there is **sufficient evidence against it**

Absence of evidence is not evidence of absence

If we conclude ‘do not reject H_0’, does it mean H_0 is true (and thus H_1 is false)?

Absence of evidence is not evidence of absence

If we conclude ‘do not reject H_0’, does it mean H_0 is true (and thus H_1 is false)?

No!

It only means that there isn’t sufficient evidence against H_0, and thus that H_0 **may** be true

Exercise: lady tasting tea

- Rothamsted, early 1920s
- Given a cup of tea, a lady claims she can tell whether milk or tea was first added to the cup
- To test her claim, Sir Fisher prepares eight cups of tea, four of which have the milk added first, and four of which have the tea added first

Question

How many cups does she have to correctly identify to convince us of her ability?

Exercise: lady tasting tea

The lady performs the experiment by selecting 4 cups (e.g. those she believes had tea poured first)

Questions

- How many ways are there to choose 4 cups out of 8? (Hint: check `scipy.misc.comb` or `sympy.binomial`)
- How many correspond to correctly identifying...
  - All 4 cups?
  - 3 cups only?

Exercise: lady tasting tea

Question

The lady correctly identifies all 4 cups. What can Sir Fisher conclude?

- She has no ability, and has chosen the correct 4 cups purely by chance
- She has the discriminatory ability she claims

Choosing correctly is unlikely in the first case (1 in 70), so Sir Fisher **rejected** this conclusion in favour of the second

Hypothesis testing step-by-step

1. Choose an appropriate **statistical test** based on H_0, H_1, and assumptions about the sample
2. Select a **significance level** α, i.e. a probability threshold below which H_0 will be rejected
3. Compute the **test statistic** from the observations and calculate the p-**value**, i.e. the probability of observing it under H_0
4. If $p < \alpha$, conclude 'H_0 is rejected at significance level α' (result is **statistically significant**)

What is the significance level α?

A probability threshold below which:

- The test statistic will be deemed 'too large' to have occurred by chance (i.e. under H_0)
- H_0 will be deemed unlikely given the data, and will thus be rejected

What is the significance level α?

	State of nature	
	H_0 is false	H_0 is true
Reject H_0	✓ True positive	✗ False positive
Do not reject H_0	✗ False negative	✓ True negative

α also corresponds to the probability of a '**type I error**' (false positive) that we are willing to accept

Multiple comparisons

Question

Assume you are conducting n independent tests at some significance level α. What is the probability of at least one false positive finding?

- The probability of a FP in any one test is α

Multiple comparisons

Question

Assume you are conducting n independent tests at some significance level α. What is the probability of at least one false positive finding?

- The probability of a FP in any one test is α
- The probability of **no** FP in any one test is $1 - \alpha$

Multiple comparisons

Question

Assume you are conducting n independent tests at some significance level α. What is the probability of at least one false positive finding?

- The probability of a FP in any one test is α
- The probability of **no** FP in any one test is $1 - \alpha$
- The probability of **no** FPs is $(1 - \alpha)^n$

Multiple comparisons

Question

Assume you are conducting n independent tests at some significance level α. What is the probability of at least one false positive finding?

- The probability of a FP in any one test is α
- The probability of **no** FP in any one test is $1 - \alpha$
- The probability of **no** FPs is $(1 - \alpha)^n$
- The probability of **at least one** FP is $1 - (1 - \alpha)^n$

Example: multiple comparisons

Question

For $\alpha = 5\%$ and $n = 100$ tests, what is the probability of at least one FP?

Example: multiple comparisons

Question

For $\alpha = 5\%$ and $n = 100$ tests, what is the probability of at least one FP?

Using the previous formula...

$$1 - (1 - 0.05)^{100} \approx 0.994,$$

which means we are **99.4% likely to have at least one FP**!

Bonferroni correction

- **Idea**: require more evidence to reject H_0
- Using $\alpha' = \alpha/n$, the 'overall' significance level (**family-wise error rate**) is what we intended

In the previous example…

$$\alpha' = 0.05/100 = 0.0005$$

Substituting back…

$$1 - (1 - 0.0005)^{100} \approx 0.05$$