Estimation and hypothesis testing

Gianluca Campanella

Contents

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 one or more characteristics of the whole population

Sampling

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

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?

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

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

Hypothesis testing

Lady tasting tea

Scenario

- Rothamsted, early 1920s
- Given a cup of tea, a lady claims she can tell whether milk or tea was first added to the cup

Question

How would you design an experiment to test her claim?

Lady tasting tea

Scenario

- To test her claim, Sir Fisher prepares eight cups of tea:
  - Four have the milk added first
  - Four have the tea added first
- The lady performs the experiment by selecting 4 cups (e.g. those she believes had tea poured first)

Question

How many cups does she have to correctly identify to convince you?

Lady tasting tea

Questions

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

Lady tasting tea

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

- She has no ability, and has chosen the 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

A/B testing

	Cancelled		Total
Old packaging	175	39.59%	442
New packaging	168	38.27%	439

Question

Does the new, nicer, more expensive packaging make customers less likely to cancel their subscriptions?

A/B testing

Read the blog post at

`https://www.candyjapan.com/behind-the-scenes/results-from-box-design-ab-test`

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
- 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?

Absence of evidence is not evidence of absence

If we conclude ‘do not reject H_0’, does it mean H_0 is true?

No!

- It only means that there isn’t sufficient evidence against H_0

$\rightarrow$ The study is inconclusive

Hypothesis testing step-by-step

1. Choose an appropriate statistical test
2. Select a significance level α
   (i.e. the probability below which you will reject H_0)
3. Conduct the experiment and record its outcome
4. Calculate the p-value
   (i.e. the probability of observing something as or more extreme than the outcome supposing that H_0 is true)
5. If $p < \alpha$, conclude: 'H_0 is rejected at significance level α'
   (the result is 'statistically significant')

What is the significance level α?

A probability threshold below which:

- The outcome of the test will be deemed ‘too large’ to have occurred under H_0 (i.e. by chance)
- H_0 will be deemed unlikely given the data

$\rightarrow$ H_0 will 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

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

Multiple comparisons

Question

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

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

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 overall is $(1 - \alpha)^n$

Multiple comparisons

Question

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 overall is $(1 - \alpha)^n$
- The probability of at least one FP is $1 - (1 - \alpha)^n$

Multiple comparisons

Question

For $\alpha = 5\%$ and $n = 100$ tests, what is the probability of FP ≥ 1?

Multiple comparisons

Question

For $\alpha = 5\%$ and $n = 100$ tests, what is the probability of FP ≥ 1?

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 approximately what we intended

In the previous example…

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

Substituting back…

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