RICE | D2K
Data to Knowledge Lab

Welcome to DSCI 101

Introduction to Data Science

Week 9 Recap

- Probability: theoretical foundation for Statistics

- Population vs. sample and parameters vs. statistics
  - empirical distribution of a random sample
  - sampling distribution of a statistic

- Statistical inference (quantify uncertainty): 3 ways
  - generate many random samples from population
  - generate Bootstrap resamples from one original random sample
  - use math to figure out the sampling distribution

Week 10 Preview

- Hypothesis testing – apply statistical inference to decision making
  - general framework and rational

- One sample test – assessing a model
  - Is the sample from a known population (the model)?
  - does the data support my model assumption?

- Two sample test – permutation test
  - are the two samples “the same” (from the same population)?

Tests of hypotheses

- You have two hypotheses about the population distribution:
  - “the population distribution is xxx.”
  - “no it’s not.”

- You have some data

- Use the data to test the hypotheses:
  - which of the two hypotheses is better supported by the data

What’s the big deal?

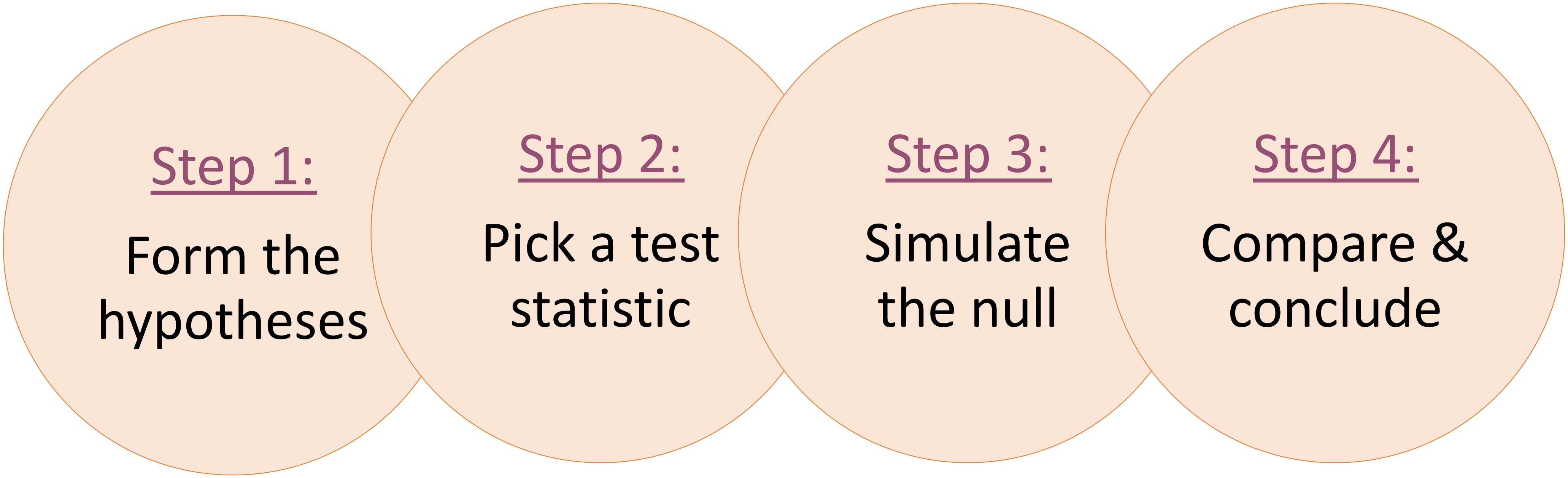
- Why can’t we just look at the data and make a decision?

uncertainty!!!

- If this hypothesis is true, how likely I will get a sample data like the one I have observed?

- How likely, how unlikely, where is the cut-off?

General framework



Step 1:

Form the hypotheses

Step 2:

Pick a test statistic

Step 3:

Simulate the null

Step 4:

Compare & conclude

Examples

Form the hypotheses

- Null hypothesis: data is generated by a specified probability model
  - “the sample is from a Normal(mean=100, sd=15).”
  - “the two groups are from the same population.”
  - “the jury panel was selected at random from eligible jurors”
  - “the drug is not working”

- Alternative hypothesis: no, it’s not
  - “the sample is NOT from a Normal(mean=100, sd=15).”
  - “the two groups are NOT from the same population.”
  - “the jury panel consists too few black people”
  - “the drug has some effects”

- When the null is rejected, usually lead to a discovery

Pick a test statistic

- A test statistic is a one number summary of your data
  - remember what a statistic is?

- A statistic that can help distinguish the two hypotheses
  - either large value or small value supports the alternative hypothesis
  - can be easily calculated from data – observed stat

- Choice is usually not unique
  - some conventions exits

Simulate under the null

- How would this test statistic be like if the null hypothesis is true?

- Because null model is specified, we can simulate data from it!
  - simulate data under the null model
  - calculate the test statistic
  - repeat – simulated stat

- Sampling distribution of the test statistic under the null hypothesis

Compare and conclude

- Simulated statistic vs. observed statistic
  - Sampling distribution vs. one value

- Does the observed statistic look like the simulated ones?
  - extreme value from tails → unlikely → reject null
  - not extreme value → likely → fail to reject null

- Need a “cut-off” for tails: the significance level

Example: Jury Selection

- Swain v.s. Alabama 1965
  - Talladega County, Alabama
  - Robert Swain, a black man convicted of crime
  - appeal: one factor was all white jury
  - 26% of population in the county were black
  - Swain's jury panel consisted of 100 men
  - 8 men on the panel were black

- Supreme Court wrote:
  - "…the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of …"
  - appeal denied

Discussion

- Step 1: form
  - Null: 100 jury panel is a random sample from population
  - Alternative: no it’s not, too few blacks
- Step 2: pick
  - observed data: demographic of 100 jury panel
  - a test Statistic: number of black
- Step 3: simulate
  - generate many random samples of 100 from the population
  - calculate simulated statistic for each random sample
- Step 4: compare
  - what is the chance of getting 8 or fewer blacks if null is true?

Define the “cut-off”

- Different alternative hypotheses:
  - which tail is supporting alternative (against the null)?

- Data: results of 100 coin flips
- 1-sided test: “fair” vs. “biased towards tails”

# of H: small value

# of T: large value

- 2-sided test: “fair” vs. “not fair”
  - # of H or # of T???

→ |# of H - # of T|: large value

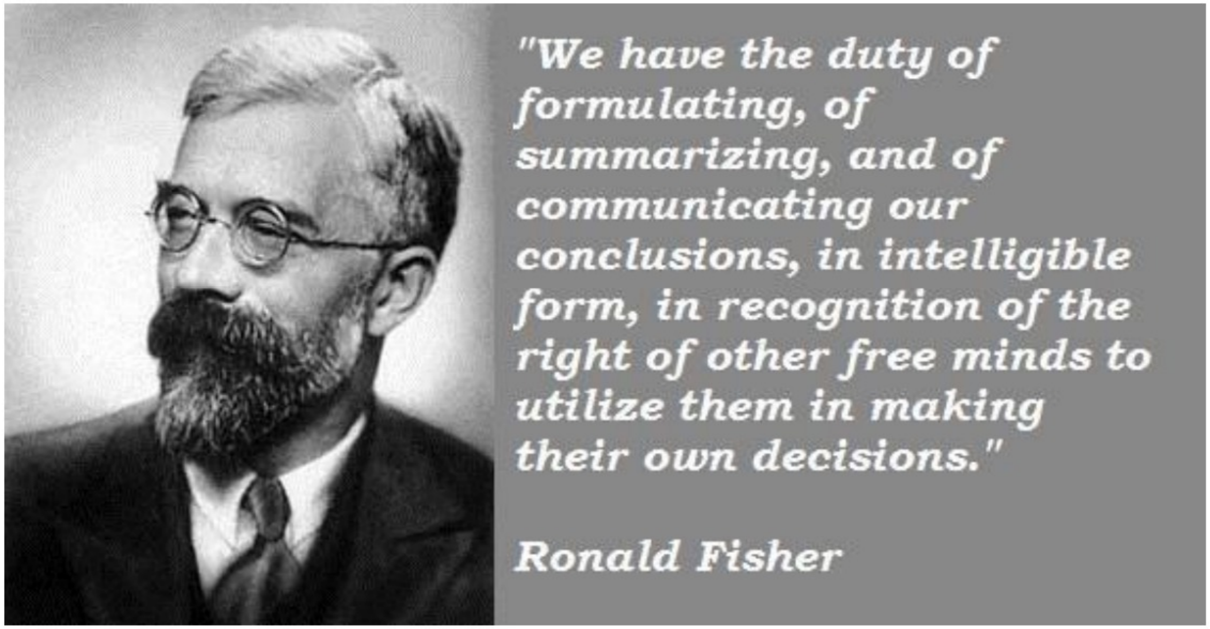
Statistical Significance

- IF the null is true
  - observed stat should be from that sampling distribution (simulated stat)
  - a “typical” value from the middle of the histogram

- Convention of cut-off: 5% or 1%
  - **chance** $\leq$ 5%, very unlikely to observe this data IF null is true→ **reject null**
  - **chance** $>$ 5%, somehow likely → **fail to reject null**

- This cut-off is called statistical significance level

Origin of the convention



"*We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions.*"

Ronald Fisher

" If one in twenty does not seem high enough odds, we may, if we prefer, to draw the line at one in fifty, or one in a hundred…"

P-values

- The p-value is a probability
  - but it's **not** the probability that null hypothesis is true!!!

- What does a p-value mean?
  - the chance (probability) of seeing the observed stat or something even more extreme (supporting the alternative) IF the null is true.

- What does a small p-value mean?
  - seeing the observed statistic is very unlikely IF the null is true → reject null!

How to calculate the p-value?

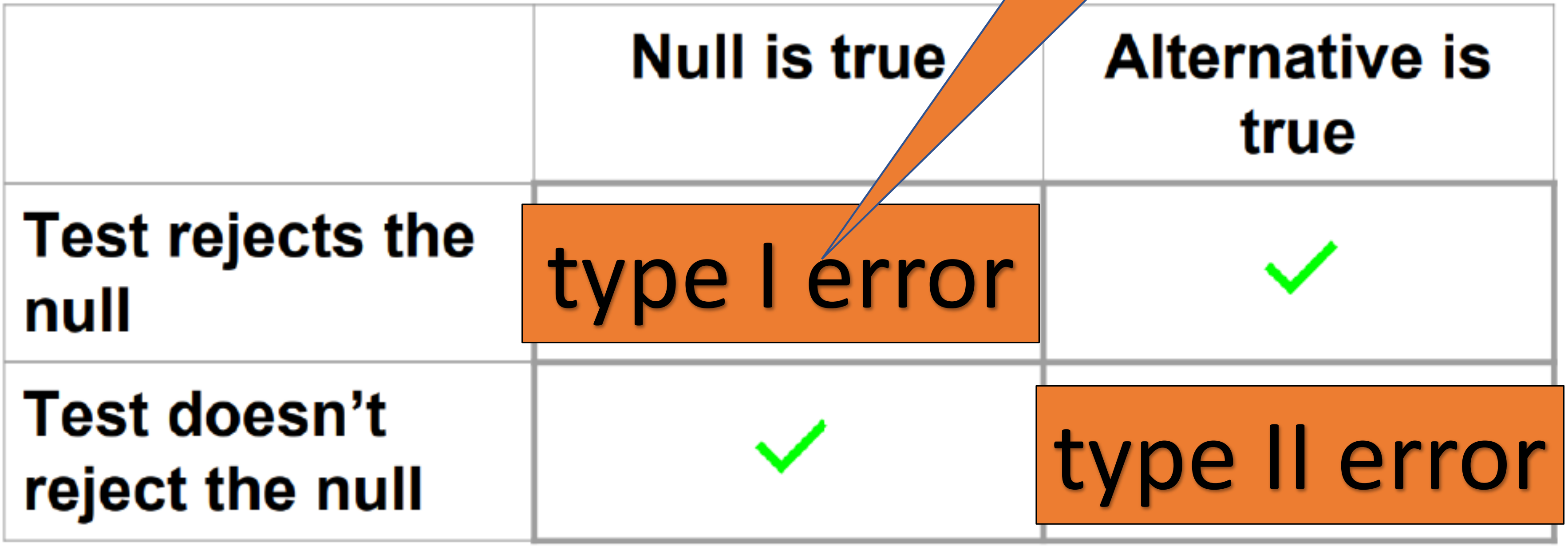
- Rational:
  - if the null is true, this is the sampling distribution of the test statistic
  - how likely is the observed statistic from that sampling distribution?

- Simulate the test statistic
  - plot observed stat on the histogram of the simulated stat
  - tail towards the direction that support alternative!
  - calculate the tail area probability
- P-value < significance level → reject the null

What does the p-value really mean?

- An error probability!
  - what errors?

p-value = type I error prob < 5%
why?



	Null is true	Alternative is true
Test rejects the null	type I error	✓
Test doesn’t reject the null	✓	type II error

A/B testing: comparing two samples

- A categorical variable divide samples into Group A and B
- A numerical variable X for both Group A and B
- Null hypothesis: the value X for Group A and B are from the same population distribution
- Alternative hypothesis: the value X for Group A and B are statistically different
  - do the treatment and control group have the same outcome?
  - do students from two schools have the same standardized test results?
  - do newborn babies from smoker and non-smoker mothers have same weights?

Discussion

- How do we simulate data from the null hypothesis?

- If the null is true, there is one population distribution out there
  - but we don’t really know what the distribution is

- Idea:
  - if null is true, group labels should NOT matter
  - simulate how large the difference could have been if group labels are randomly shuffled

Example: baby weights

- Hypotheses:
  - Null: newborn baby weights are the same for two groups
  - Alternative: babies from smoker mothers have lower weights

- Test statistic:
  - Group A non-smoker average – Group B smoker average
  - large values favor the alternative → 1-sided test
  - if alternative has no specified direction → 2-sided test
    - use absolute difference

Demo

Example: GAI’s defense

- Stats101 is divided into 12 sections of about 30 students in each
- Same course material / assignments/ exams, taught by different GAI
- After midterm exam, section 3 has the lowest average grade

- Are section 3 grades really “lower”?
- Are section 3 grades like a random sample from the entire population distribution?
  - what is the population distribution here?

Demo

Take away message

- Hypothesis testing
  - apply statistical inference in decision making
- What uncertainty is being quantified?
  - how different could a random sample be if null is true
  - quantify sampling distribution of the test statistic under the null
- Formal hypothesis testing
  - the sampling distribution of the test statistic: z-test, t-test, F-test, χ^2 test…
  - understand the framework and the rational is the most important