Statistical Inference 101. Pitfalls

Section Outline

- **Population vs Samples**
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Population

- Set of objects/events of interest for a question

Population vs Sample

- Sometimes we don’t have access to / we don’t know the population
- We only have access to a **sample**: a subset of the population
  - We can create/collect the sample
  - We may be given the sample
- We want to understand parameters of the population using the sample

Inferential Statistics

- Draw conclusions of the *population* from a *sample* of data
  - Each conclusion we reach will have an associated **sample error**
  - A lot of inferential statistics is about characterizing sample error
- First question: What kind of conclusions can we draw?

Population – Sample Mismatch

- Overgeneralization
  - We use the sample to claim something about a *broader* population than the one the sample represents

Population – Sample Mismatch

- Overgeneralization
  - We use the sample to claim something about a *broader* population than the one the sample represents
- Bias
  - We fail to obtain a *representative* sample of the population

Population – Sample Mismatch

- Overgeneralization
  - We use the sample to claim something about a *broader* population than the one the sample represents
- Bias
  - We fail to obtain a *representative* sample of the population
- Faulty generalization
  - Anecdotal evidence
  - Sample is too small

Population – Sample Mismatch

- Overgeneralization
  - We use the sample to claim something about a *broader* population than the one the sample represents
- Bias
  - We fail to obtain a *representative* sample of the population
- Faulty generalization
  - Anecdotal evidence
  - Sample is too small
- Correlation is not Causation

Section Outline

- Population vs Samples
- **Quick recap of Statistical Inference**
- Hypothesis testing, p-values
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Goal of statistical inference

- To understand and quantify uncertainty of parameter estimates
  - Parameter: what we are interested in learning (population)
    - Average, proportion, etc.
- *Sample, obtain point estimate, assume point estimate comes from a distribution so we can characterize its quality*

More terminology

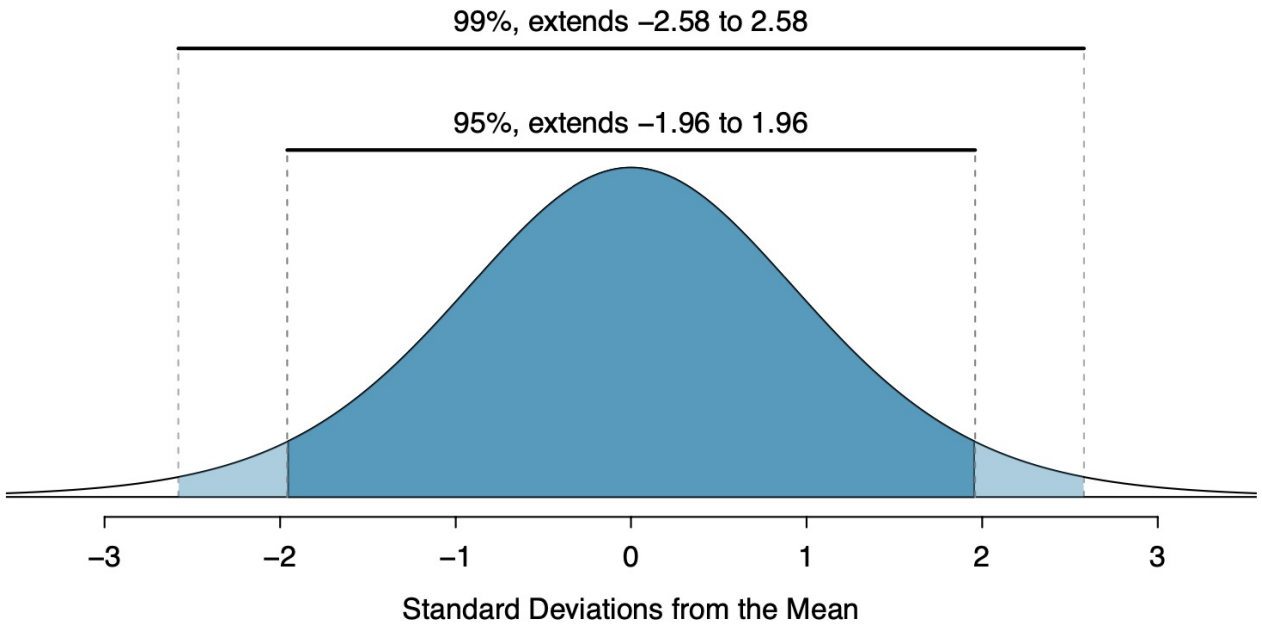
- **Population**: set of objects/events of interest for a question
- **Sample**: a subset of objects/events from the population
- **Parameter**: the statistic of interest computed over the population
- **Point estimate**: the statistic of interest computed over a sample of the population
- **Error**: The difference between estimate and ground truth
- **Sampling error**: How much estimate changes across samples
  - There will be some natural variation
  - Our goal is to characterize and understand this sampling error

Inferential Stats 101

- **Confidence intervals**
  - Sampling distribution
  - CLT
  - Interpreting confidence intervals
- Hypothesis Testing
  - Spell out null and alternative hypothesis
  - Can’t we reject the null?
    - Two outcomes: either we reject H0, or we fail to reject H0

Example building confidence interval

- We construct a normal with mean = point estimate
- X% Confidence interval is the range that encompasses X% of the distribution
- In the case of 95% that's 1.96 standard deviations around the mean



Example

- The proportion of American adults who support solar energy is 0.887 based on a sample of size n=1000
  - Is it a random sample?
- Is the sample sufficiently large
  - Success-failure condition, does CLT apply?
- Margin of error (ME) = 1.96 * 0.01
  - 95% Confidence interval: 0.887 +- ME
    - (0.867, 0.906)
- **Interpret**: We are 95% confident that the population proportion of American adults that support solar expansion is between 86.7% and 90.6%

Inferential Stats 101

- Confidence intervals
  - Sampling distribution
  - CLT
  - Interpreting confidence intervals
- **Hypothesis Testing**
  - Spell out null and alternative hypothesis
  - Can’t we reject the null?
    - Two outcomes: either we reject H0, or we fail to reject H0

Hypothesis Testing using Confidence Intervals

- The evidence (sample) will give us a proportion
- Build a confidence interval around that proportion
- Check if the null hypothesis fall inside or outside the interval
  - Conclude if we have enough evidence to reject it
  - If we don’t have enough evidence, all we can say is:
    - The null hypothesis is not implausible. We fail to reject H0

Hypothesis Testing using Confidence Intervals

- Random guessing would lead to 33.3%
- Random sample of 50 college-educated students
  - Note the sample determines for what population are we testing Ha
- 24% of students got the response correct
- Is the deviation between 24% and 33.3% due to sampling error?
  - Construct confidence interval around 24%
    - 12% to 35%
  - 33% falls within the confidence interval so we H0 is not implausible
- This sample does not provide evidence to reject the idea that students do better than random guessing. We cannot reject H0.

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- **Hypothesis testing, p-values**
- Errors
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

p-values

- p-value: quantifies the strength of the evidence against the null hypothesis and in favor of the alternative hypothesis
- p-value: probability of observing data at least as favorable to the alternative hypothesis as the current evidence if the null hypothesis were true
- We use a summary statistic of the data to compute the p-value

Coal usage 1/6

- Do you support increased usage of coal to produce energy?
  - Sample: 1000 American adults.

Coal usage 2/6

- Do you support increased usage of coal to produce energy?
  - Sample: 1000 American adults.
- H0: 50% support it. **Null value:** p0 = 0.5
- Ha: Significantly more/less than half support it

Coal usage 3/6

- Do you support increased usage of coal to produce energy?
  - Sample: 1000 American adults.
- H0: 50% support it. **Null value:** p0 = 0.5
- Ha: Significantly more/less than half support it
- 37% support increased usage of coal

Coal usage 4/6

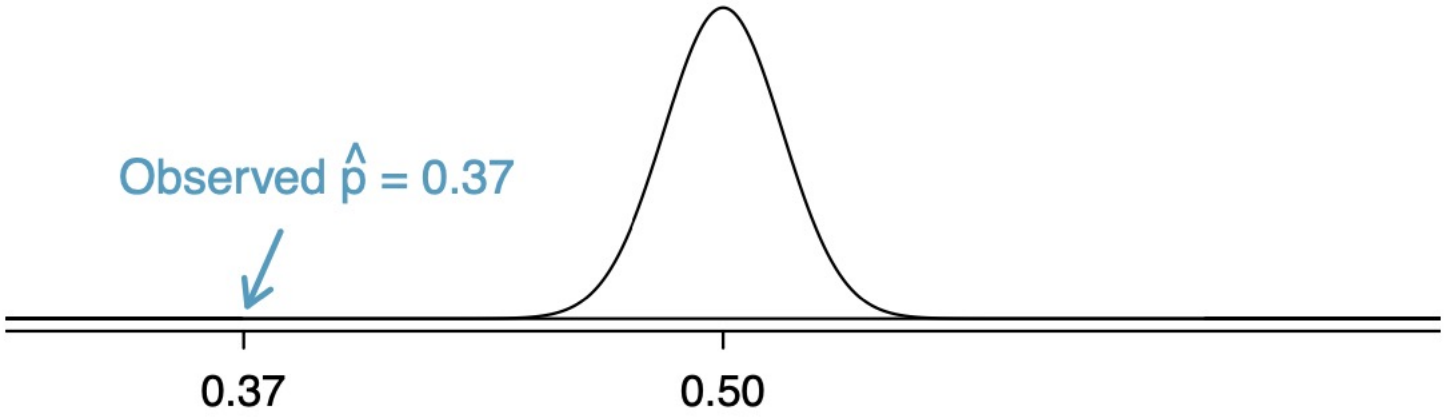
- Does 37% represent a real difference with respect to 50%? Or is it just sampling error?

Coal usage 5/6

- Does 37% represent a real difference with respect to 50%? Or is it just sampling error?
  - What would the sampling distribution of p look like if H0 were true?
    - If H0 is true then population proportion is p0 = 0.5
    - Is the sampling distribution normal? Independent sample
      - We check success failure condition using p0 (we are assuming H0 is true).
      - We compute the standard error
  - If H0 is true, distribution follows a normal with mean 0.5 and se = 0.016

Coal usage 6/6

- Now we know the shape of the distribution (called **null distribution**) we can place the point estimate we have



- The p-value represents the probability of observing p_hat, if the null hypothesis were true.
  - If the value is smaller than the chosen significance level, we reject H0

Constructing p-value for proportion 1/2

- Our mental model is: the null hypothesis is true, and we want to check if the evidence rejects it or not
- Assuming it's true, we check if the null hypothesis comes from a normal distribution:
  - Need to check conditions of independence and success-failure ratio
  - This is called the null distribution
- We now check the sample proportion with respect to the null hypothesis distribution
- p-value represents the probability of sample proportion (or a more extreme one) on the null distribution

Constructing p-value for proportion 2/2

- We construct the null distribution, we then find the tail area given by our sample proportion
- p-value represents the probability of observing the sample proportion by chance if the null hypothesis were true
- We compare p-value to alpha. If p-value < alpha we reject H0
  - We say data provide strong evidence against H0
- Finally, describe conclusion in the context of the data

Interpreting hypothesis tests

- There are two outcomes after conducting a hypothesis test:
  - We reject the null hypothesis
  - We do not reject the null hypothesis
- Note that the following are not valid outcomes:
  - We accept the null hypothesis
  - We prove the null hypothesis
- “You can never prove the null hypothesis”
- When explaining p-values it’s ok to use double negative

Choosing significance level

- This is the probability of false positive
  - Probability of accepting Ha when H0 is true
  - False discovery!
- “It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not”. Fisher

Practical vs Statistical Significance

- Statistical significance, p < alpha
  - We can increase statistical significance by using larger samples
- What is practical significance?
  - This depends on the application. Sometimes, it is more subjective
- Statistical Power: probability of true positive
  - P(reject H0) when H1 is true.
  - This depends on the test you use

Other tests

- Different statistics of interest have different sampling distributions
  - There are different tests and different ways of computing confidence intervals for those
- The principles are the same. Hypothesis testing is a framework

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- **Errors**
- Multiple comparisons
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Errors

- Type 1 error / False positive
  - Reject null hypothesis when it’s true
- Type 2 error/ False negative
  - Fail to reject null hypothesis when alternative is true

Errors

- Type 1 error / False positive
  - Reject null hypothesis when it’s true
- Type 2 error/ False negative
  - Fail to reject null hypothesis when alternative is true
- Consider how changing confidence interval affects these errors
  - When do we consider we have enough evidence to reject H0
    - The threshold we choose is the **significance level**

Errors

- Type 1 error / False positive
  - Reject null hypothesis when it’s true
- Type 2 error/ False negative
  - Fail to reject null hypothesis when alternative is true
- Consider how changing confidence interval affects these errors
  - When do we consider we have enough evidence to reject H0
    - The threshold we choose is the **significance level**
- How can we make sure we don’t have false positives?
  - By never rejecting the null hypothesis

Errors

- Type 1 error / False positive
  - Reject null hypothesis when it’s true
- Type 2 error/ False negative
  - Fail to reject null hypothesis when alternative is true
- Consider how changing confidence interval affects these errors
  - When do we consider we have enough evidence to reject H0
    - The threshold we choose is the **significance level**
- How can we make sure we don’t have false positives?
  - By never rejecting the null hypothesis
    - Does this make sense? How many false negatives will we have then?

Section Outline

- Population vs Samples
- Quick recap of Statistical Inference
- Hypothesis testing, p-values
- Errors
- **Multiple comparisons**
- Bonferroni Correction
- Data Dredging, p-hacking
- Publication bias

Remember

- A *p*-value says there's *p* chance of getting the observed result *if* the null hypothesis were true
  - It **does not** mean there's a *p* chance the null hypothesis is true

Predicting the Stock Market

Predicting the Stock Market

- alpha = 5%
- 0.5^10 ~= 0.0009
- p-value = 0.0009 << 0.05
  - We conclude we should trust and pay the money to get their predictions!
    - Or not?

Terminology

- Multiple comparison
- Multiple test problem
- Look-elsewhere effect

Multiple comparisons problem

- Test multiple hypothesis on same dataset
  - False positives add to each other. Probability of false positive (false discovery) increases
- Test same hypothesis on multiple datasets
  - Same as above
- Assume H0 is true. alpha=5%. 20 hypotheses to test
  - What's the probability of obtaining 1 discovery (rejecting H0)?
    - 1 – P(no significant results)
    - P(no significant results) = (1 - alpha)^num_hypotheses
    - $1 – (1 – (0.05)^{20})$ -> 64% of making a 'discovery' (even if no individual hypothesis is true)

Predicting the Stock Market

- The email we received on day 1 was received by 999999 others
- On day 2 by 499999 others
- …
- After 10 days, about 1K people have received a sequence of 10 correct predictions

Multiple comparisons problem

- Test multiple hypothesis on same dataset
  - False positives add to each other. Probability of false positive (false discovery) increases
- Test same hypothesis on multiple datasets
  - Same as above
- *Multiple comparisons must be corrected*
  - Bonferroni correction
    - Make each of m tests to have significance level alpha/m
  - False Discovery Rate techniques

Multiple Hypothesis 2/2

- Perform 2 tests on same data
- Suppose null hypothesis is true (ground truth)
- Test 1 has some probability of false positives
- Test 2 has some probability of false positives
- Think about the samples that would lead to false positives in test 1 and those that would lead to false positives in test 2.
  - Are those samples the same? No
- So we consider the joint probability of false positive
  - Familywise-error rate: probability that at least one sample will give a type 1 error for at least one of the hypothesis

Bonferroni Method

- *If the sum of the Type I error rates for different tests is less than α, then the overall Type I error rate (FWER) for the combined tests will be at most* α.
- The Bonferroni method is conservative.
  - Bonferroni's conservativeness means it reduces statistical power
    - i.e., it reduces the probability of true positives
  - In practice: use the Bonferroni-Holm adjustment
  - In practice: but how do you choose/define the *family* of tests?

False Discovery Rate Methods

- Alternative: control false discovery rate
  - Proportion of discoveries that are false positives
- Set the maximum allowed # of false positives (Q)
  - You choose this based on the application, like alpha
- Sort p-values from low to high, rank i=0 to i=m, for m tests
  - Compare each p-value to (i/m)Q
- Find largest p-value, p^* s.t. $p^* < (i/m)Q$
  - p^* and any p s.t. $p < p^*$ are significant

False Discovery Rate Methods 2/2

- How to choose Q?
  - What's the cost of an additional experiment? And of a false negative?
    - If low and high, then you should tend to choose a higher Q
- FDR is less sensitive to the test family than Bonferroni
- Are tests really independent?
  - There are more advanced methods for when there's some dependence

P-Hacking, Data Dredging, Data Snooping

- This is what happens when we disregard the multiple comparisons problem
- Identify statistically significant patterns (discoveries) while increasing and understating the risk of false positives.
  - Run many statistical tests and only report those with significant results

”Most Published Research Findings are False”. John Ioannidis, 2005

”Most Published Research Findings are False”. John Ioannidis, 2005

- The importance of reproducibility!

Testing Hypothesis Suggested by the Data

- Look at data, identify hypothesis, test hypothesis on same data
- Exploratory data analysis is great to identify hypothesis
  - But those hypothesis must be tested in different datasets
    - Different samples
    - Otherwise test may pass due to false positives
- *Big data, automated hypothesis generation, and new tooling*