Hypothesis Testing

Another way to understand Lecture 06 from DSC 80

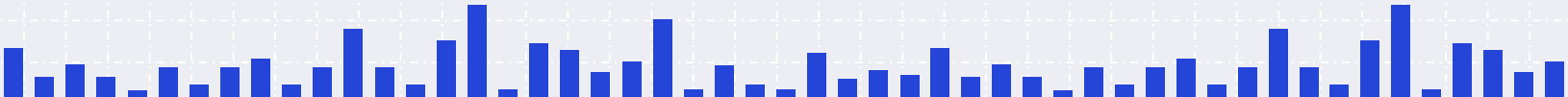

Table of contents

01

Definitions

What is hypothesis testing?

What is hypothesis testing? Why do we need it?

From lecture: “Hypothesis testing allows us to determine whether an observation is ‘significant.’"

- How likely is the appearance of our observation in this population?
- Is our observation unusual enough to be considered “too rare”?

From lecture: “Showing something is not true is usually easier than showing something is true!”

- I’m not sure which population(s) our observation belongs to, but I can prove that it’s not this one.
- We only have evidence to reject or fail to reject the assumption that our observation comes from a specific population.

Parts of a hypothesis test

Null Hypothesis

The default assumption that we compare our observation to. Usually claims no difference between the observation and population.

Alternative Hypothesis

What you suspect is “really” true, instead of the null. Claims a difference between the observation and population.

Test Statistic

A mathematical value used to compare the observation to the null hypothesis. Choosing (e.g. TVD, mean difference, K-S) depends on the alternative hypothesis.

Observed Test Statistic

The value calculated from the test statistic on our observed data.

Empirical Distribution

The distribution (or histogram) of the test statistic calculated from many random samples under the null hypothesis.

P-Value

The probability of seeing our observed test statistic (or more extreme), based on the empirical distribution.

More parts of a hypothesis test

Population

All individuals or items that share a common characteristic with the observed data. Samples are drawn from here.

Sample

A smaller group of individuals or items that accurately represent the population. Allows analysis to be generalized to the larger population.

Alpha

The significance level, chosen by researchers, that determines the threshold p-value for rejecting the null hypothesis. Often 0.05.

Type I Error

The mistake of rejecting the null hypothesis when it is actually true (false positive).

Type II Error

The mistake of failing to reject the null hypothesis when it is actually false (false negative).

Permutation Test

A type of hypothesis test that determines if two samples are from the same (unknown) population by comparing the observed data to many random rearrangements (permutations) of that data.

02

Analogy

Hypothesis testing vs. Court cases

Analogy - Setup

Hypothesis Testing

Utilize evidence to reject or fail to reject the null hypothesis.

Null Hypothesis:
Alternative Hypothesis:

The null hypothesis is assumed true until there is significance evidence to reject it.

Court Case

Utilize evidence to prove or disprove a person's innocence.

The person is innocent.
The person is guilty.

A person is assumed innocent until there is enough evidence to prove them guilty.

Hypotheses should always be clear! Avoid using terms like 'about' or 'approximately' that suggest ambiguity

Analogy - Evidence

Hypothesis Testing

Choose an appropriate test statistic that accurately compares the observed data to the sample data.

The researcher chooses a significance value to determine the p-value threshold.

Court Case

Each side presents truthful evidence that support their case. This evidence is compared.

The judge/jury decide what evidence is considered significant enough to deem the person guilty.

Analogy - Decision and Error

Hypothesis Testing	Court Case
Fail to reject the null	Not enough evidence to prove guilty
Reject the null	There is enough evidence to prove guilty
Type I Error (false positive)	The person was deemed guilty when they are actually innocent
Type II Error (false negative)	The person was deemed innocent when they are actually guilty

Aside: Choosing the “best” test statistic

In this course, we cover 4 different test statistics:

- total variation distance (TVD)
- mean/median difference between groups
- absolute mean/median difference between groups
- Kolmogorov-Smirnov (K-S) statistic

However, there are infinitely many different test statistics, and each of them can be equally great for specific hypotheses.

Choosing the “best” test statistic is the ability to recognize what test statistic(s) can be used for a specific set of hypotheses.

For instance, if my hypotheses require a two sided test between two quantitative variables, I should know to use the absolute mean difference rather than just the mean difference, because my alternative hypothesis does not specify a direction.

However, I am not restricted to only the absolute mean difference. I might instead use the absolute median difference.

Aside: Choosing the "best" test statistic

Total Variation Distance (TVD)

```
def tvd(dist1, dist2):
    return np.abs(dist1 - dist2).sum() / 2
```

- Describes the distance between two **categorical** distributions
- Absolute value → only **two-sided** tests

Kolmogorov-Smirnov (K-S) Statistic

- Roughly defined as the largest (absolute*) difference between two **quantitative** distributions (CDFs)
- Both **one-sided** and **two-sided*** tests
  - Two-sided is more common
  - See documentation

Absolute Mean/Median Difference

`np.abs(m1 - m2)`

- **Quantitative** variables
- Absolute value → only **two-sided** tests

Mean/Median Difference

`(m1 - m2)`

- **Quantitative** variables
- No absolute value → only **one-sided** tests

03

Interactive

Is this coin fair?

Simulation: Is this coin fair?

Check out this virtual coin tosser. We'll use it to create our own sample of fair coins and test an unfair coin as our observation. That way, we know our results will show that the coin is not fair (reject the null) and understand why that is!

Null hypothesis: The coin is fair. It lands on heads and tails equally. P(H) = P(T) = 0.5

Alternative hypothesis: The coin is not fair. The probability of landing on heads is not 0.5.
*note that we did not specify a direction, so this is a two sided test.

Check out this GitHub repository. We'll use the notebook to input our data in code and visualize the results.

Test Statistic: Total variance distance (TVD)
*coin flips are categorical!
**note that we could also use the number of heads/tails, like we did in the pre-lecture

Significance Level: 0.05 or your choice

04

Guided Coding

Try it yourself!