Hypothesis Testing

Diyi Yang
Stanford CS
Lecture Overview

❖ Hypotheses
❖ Significance
## Hypotheses

<table>
<thead>
<tr>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The average income in two sub-populations is different</td>
</tr>
<tr>
<td>Web design A leads to higher CTR than web design B</td>
</tr>
<tr>
<td>Self-reported location on Twitter is predictive of political preference</td>
</tr>
<tr>
<td>Male and female literary characters become more similar over time</td>
</tr>
</tbody>
</table>
Hypotheses

The first step is formalizing a question into a testable hypothesis.

This bestseller’s book cover has changed a lot since 1998.
Voters in big cities prefer Clinton.
Email marketing pitch A is better than the pitch B.
Null Hypothesis

A claim, assumed to be true, that we’d like to test (because we think it’s wrong)

<table>
<thead>
<tr>
<th>Hypothesis “area”</th>
<th>H_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>The average income in two sub-populations is different</td>
<td>The incomes are the <strong>same</strong></td>
</tr>
<tr>
<td>Web design A leads to higher CTR than web design B</td>
<td>The CTR are the <strong>same</strong></td>
</tr>
<tr>
<td>Self-reported location on Twitter is predictive of political preference</td>
<td>Location has <strong>no</strong> relationship with political preference</td>
</tr>
<tr>
<td>Male and female literary characters become more similar over time</td>
<td>There is <strong>no</strong> difference in M/F characters over time</td>
</tr>
</tbody>
</table>
Hypothesis Testing

If the null hypothesis were true, how likely is it that you’d see the data you see?
Example

**Hypothesis:** Palo Alto residents tend to be politically liberal

\( H_0: \) Among all \( N \) registered \{Democrat, Republican\} primary voters, there are an equal number of Democrats and Republicans in Palo Alto

\[
\frac{N_{dem}}{N} = \frac{N_{Rep}}{N} = 0.5
\]
Hypothesis Testing

Hypothesis testing measures our confidence in what we can say about a null from a sample.
Example

At what point is a sample statistic unusual enough to reject the null hypothesis?
Example

The form we assume for the null hypothesis lets us quantify that level of surprise.

We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size $n$;

For large $n$, we can often make a normal approximation.
Z Score

\[ Z = \frac{X - \mu}{\sigma / \sqrt{n}} \]

For normal distributions, transform into standard normal (mean=0, standard deviation=1)

\[ Z = \frac{Y - np}{\sqrt{np(1 - p)}} \]

For Binomial distributions, normal approximation (for large n)
Z Score Example

- 510 democrats = z score 0.63
- 580 democrats = z score 5.06
Test and Significance Level

Decide on the level of significance $\alpha$, \{0.05, 0.01\}

Testing is evaluating whether the sample statistic falls in the rejection region defined by $\alpha$
Tails

**Two-tailed tests:** whether the observed statistic is different (in either direction)

**One-tailed tests:** difference in a specific direction

All differ in where the rejection region is located: 0.05 for all
P-Value

The p-value, or calculated probability, is the estimated probability of rejecting the null hypothesis $H_0$ when that hypothesis is true.

Or the probability that the observed statistic occurred by chance alone.
P-Value

Two-tailed test \[ p\text{-value}(z) = 2 \times P(Z \leq -|z|) \]

Lower-tailed test \[ p\text{-value}(z) = P(Z \leq z) \]

Upper-tailed test \[ p\text{-value}(z) = 1 - P(Z \leq z) \]
**Errors are possible! Error Types**

**Type I Error:** we reject the null hypothesis but we shouldn’t have

**Type II Error:** we don’t reject the null, but we should have

<table>
<thead>
<tr>
<th>Researcher’s Decision</th>
<th>Actual Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>No Effect = H₀ True</strong></td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I error (α)</td>
</tr>
<tr>
<td>Fail to reject H₀</td>
<td>Decision correct</td>
</tr>
</tbody>
</table>
Errors

For any significance level $\alpha$ and $n$ hypothesis tests,
we can expect $\alpha \times n$ type I errors

$\alpha=0.01$, $n=1000$ will lead to 10 “significant” results simply by chance

When would this occur in practice?
Multiple Hypothesis Corrections

Bonferroni correction

For family-wise significance level $\alpha_0$ with $n$ hypothesis tests

\[ \alpha \leftarrow \frac{\alpha_0}{n} \]

- Very strict; controls the probability of at least one type I error
- False discovery rate

Slides Credit to David Bamman
Hypothesis Testing Summary

**Step 1:** State hypotheses and select alpha level

**Step 2:** Collect data; compute the test statistic

**Step 3:** Make a probability-based decision about $H_0$. Reject $H_0$ if the test statistic is unlikely when $H_0$ is true ("statistically significant")
Reporting Significant Effect

A result is significant or statistically significant if it is very unlikely to occur when the null hypothesis is true, that is rejecting $H_0$

- Report that you found a significant effect
- Report value of test statistic
- Report the p-value of your test statistic
Non-parametric Tests

Many hypothesis tests rely on parametric assumptions (e.g., normality)

Alternatives that don’t rely on those assumptions
  Permutation test
  The Bootstrap
Significance of Coefficients

- A $\beta_i$ value of 0 means that a feature $x_i$ has no effect on the prediction of $y$
- How great does a $\beta_i$ value have to be for us to say that its effect probably doesn’t arise by chance?
- People often use parametric tests (coefficients are drawn from a normal distribution) to assess this for logistic regression, but we can use it to illustrate another more robust test
Permutation Test

Non-parametric way of creating a null distribution for testing the difference in two populations A and B

For example, the median height of men (=A) and women (=B)

We shuffle the labels of the data under the null assumption that the labels don’t matter (the null is that A=B)
<table>
<thead>
<tr>
<th>x</th>
<th>Score</th>
<th>True Labels</th>
<th>Perm 1</th>
<th>Perm 2</th>
<th>Perm 3</th>
<th>Perm 4</th>
<th>Perm 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>62.8</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>man</td>
</tr>
<tr>
<td>x2</td>
<td>66.2</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
</tr>
<tr>
<td>x3</td>
<td>65.1</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>man</td>
</tr>
<tr>
<td>x4</td>
<td>68.0</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
</tr>
<tr>
<td>x5</td>
<td>61.0</td>
<td>woman</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>man</td>
<td>man</td>
</tr>
<tr>
<td>x6</td>
<td>73.1</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
</tr>
<tr>
<td>x7</td>
<td>67.0</td>
<td>man</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>man</td>
</tr>
<tr>
<td>x8</td>
<td>71.2</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
<td>man</td>
<td>man</td>
</tr>
<tr>
<td>x9</td>
<td>68.4</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
</tr>
<tr>
<td>x10</td>
<td>70.9</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
</tr>
</tbody>
</table>
How many times is the difference in medians between the permuted groups greater than the observed differences?

Observed true difference in medians: -5.5

<table>
<thead>
<tr>
<th></th>
<th>true</th>
<th>perm 1</th>
<th>perm 2</th>
<th>perm 3</th>
<th>perm 4</th>
<th>perm 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>62.8</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>woman</td>
<td>man</td>
</tr>
<tr>
<td>x2</td>
<td>66.2</td>
<td>woman</td>
<td>man</td>
<td>man</td>
<td>man</td>
<td>woman</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>x9</td>
<td>68.4</td>
<td>man</td>
<td>woman</td>
<td>man</td>
<td>woman</td>
<td>man</td>
</tr>
<tr>
<td>x10</td>
<td>70.9</td>
<td>man</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
<td>woman</td>
</tr>
</tbody>
</table>

Difference in medians: 4.7 5.8 1.4 2.9 3.3
Permutation Test

The p-value is the number of times the permuted test statistic $t_p$ is more extreme than the observed test $t$

$$
\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I[abs(t) < abs(t_p)]
$$
Permutation Test

To test whether the coefficients have a statistically significant effect (i.e., they are not 0), we can conduct a permutation test where, for B trials, we:

1. Shuffle the class labels in the training data
2. Train logistic regression on the new permuted dataset
3. Tally whether the absolute value of $\beta$ learned on permuted data is greater than the absolute value of $\beta$ learned on the true data
Permutation Test

The p-value is the number of times the permuted test statistic $\beta_p$ is more extreme than the observed test $\beta$

\[
\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I[\text{abs}(\beta_i) < \text{abs}(\beta_p)]
\]
Bootstrap

Randomly sampling with replacement

1. Resample a data set \( x^* \) \( B \) times with replacement
2. Evaluate the bootstrap statistic \( t(\cdot) \) each time
3. Approximate significance level via

\[
\frac{t(x^{*b}) \geq t(x)}{B}
\]