Hypothesis testing

22 April 2009

Research Methods for Empirical Computer Science
CMPSCI 691DD
R.A. Fisher and ‘The lady tasting tea’
A value of a statistic is statistically significant if it (or a more extreme value) is unlikely to occur under $H_0$.

$$\alpha = p(\text{Reject } H_0 \mid H_0 \text{ True}) = p(\text{Type I Error})$$
Sampling distributions

Hypothetical Population (for which $H_0$ is true) → All Possible Samples → Derived Statistic Values → Sampling Distribution
Hypothesis testing strategy

• Formulate a null (and alternative) hypothesis
  • $H_0 : \mu_A = \mu_B$
  • $H_1 : \mu_A \neq \mu_B$

• Gather data

• Calculate a sample statistic (e.g., $\bar{x}$)

• Estimate the sampling distribution for that statistic given $H_0$

• Use the sampling distribution to calculate $p(\bar{x}|H_0)$ (probability of obtaining $\bar{x}$ given $H_0$)

• If the probability is low, reject $H_0$ in favor of $H_1$
How good was Fisher’s lady (Miss Bristol)?

\[
p(\text{outcome} \mid H_0) = \frac{\text{equivalent or more extreme outcomes}}{\text{outcomes}}
\]

<table>
<thead>
<tr>
<th>Cup</th>
<th>Actual</th>
<th>Guess2</th>
<th>3</th>
<th>…</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>…</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>…</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>++</td>
<td>–</td>
<td>…</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>…</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>…</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>…</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Requirements for permutation tests

• What is needed to apply this type of test?
  • Well-defined space of outcomes
    (e.g., permutations of guesses)
  • Probability distribution over those outcomes
  • Method for sampling from that space
  • Method for ranking outcomes (so each outcome is either equivalent, more, or less extreme than the experimentally obtained outcome)

• Do we need to generate all possible outcomes from the space?
  • No, it is enough to sample randomly, assuming we know the probability distribution
Parametric Tests
Example: Evaluating IR

• The online service Google evaluates performance of two search engines over 30-day periods.
• The new system achieves a mean per-day performance of 5603 (var=150)
• The old system achieves a mean per-day performance of 5476 (var=143)
• How can we test the null hypothesis that the mean per-day performance is drawn from the same distribution?
Difference between means

- Two-sample t-test
  - Assumptions
    - Population is normally distributed
    - Variance of two populations are equal
    - Samples are independent, random draws from the population

- Two-sample Z-test
  - Assumptions
    - $N_1$ and $N_2$ are sufficiently large
    - Samples are independent, random draws from the population
**Paired t-test**

- Boost the power of a two-sample test by controlling for variance

- Conventional two-sample t-test
  - Measure system A for 30 day period
  - Measure system B for **different** 30 day period
  - Use pooled variance \( \sigma_A^2 + \sigma_B^2 \)

- Paired t-test
  - Measure system A for 30 day period
  - Measure system B for **same** 30 day period
  - Use variance of differences \( \sigma_{A-B}^2 \)
Non-Parametric Tests
Non-parametric methods

- Compare relative locations of probability distributions rather than specific parameters of the populations
  - Many use **relative ranks** of sample observations rather than numerical values

- Weaknesses
  - Lower statistical power
  - Loss of precision (ranks vs. scores)
  - Not robust to all violations of assumptions
McNemar’s test

- Data consist of paired observations of labels
  - Test of difference in proportions is not applicable because samples are dependent

- Example:
  - Classification algorithms A and B applied to the same test set
  - Ignore CC and II cases
  - $H_0$: CI is as likely IC
  - Use binomial distribution

<table>
<thead>
<tr>
<th>Alg B</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>CC</td>
<td>CI</td>
</tr>
<tr>
<td>Incorrect</td>
<td>IC</td>
<td>II</td>
</tr>
</tbody>
</table>
Selecting hypothesis tests
Criteria to consider

- Sample information
  - One or two samples?
  - Paired or unpaired?
- Data type information
  - Discrete or continuous?
  - Parametric or non-parametric?
- Statistical power of test
- Robustness of test
## “Cookbook” advice

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th>Non-parametric</th>
<th>Discrete</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unpaired</strong></td>
<td>t-test</td>
<td>Wilcoxon test</td>
<td>Chi-square or binomial test</td>
</tr>
<tr>
<td><strong>Paired</strong></td>
<td>Paired t-test</td>
<td>Wilcoxon test</td>
<td>McNemar’s test</td>
</tr>
<tr>
<td><strong>Two samples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unpaired</strong></td>
<td></td>
<td>Mann-Whitney test</td>
<td></td>
</tr>
<tr>
<td><strong>Paired</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Use table for suggestions not rules

- Data type categorization may depend on...
  - Hypothesis being examined
  - External information about the data
- May be able to collect different data
  - One sample vs. two-sample
  - Paired vs. unpaired
- Check test assumptions
  - Use initial test of normality, equal variance, etc.
- Try low-power tests first, because often...
  - Easier to run
  - More robust
  - Require fewer assumptions

*(Vellman and Wilkinson 1993)*