### Permutation tests and ROC curves

Frédéric Schütz

**EMBnet course**  
20 January 2009

---

**Hypothesis Truth vs. Decision**

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
<th>not rejected</th>
<th>rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>true H</td>
<td>specificity</td>
<td>🎁</td>
<td>✗ Type I error (False +) α</td>
</tr>
<tr>
<td>false H</td>
<td>✗ Type II error (False -) β</td>
<td>🎁 Power 1 - β; sensitivity</td>
<td></td>
</tr>
</tbody>
</table>

---

**Swiss Institute of Bioinformatics**
Steps in hypothesis testing

1. Identify the population parameter being tested (e.g. population mean)

2. Formulate the NULL (H) and ALT (A) hypotheses

3. Compute an appropriate TS (Test Statistic)

4. Compute the p-value

5. (Optional) Decision Rule: REJECT H if the p-value ≤ α

Computing the p-value

- For a parametric hypothesis test, we start with some assumptions to derive the sampling distribution of the TS assuming that the NULL hypothesis is true

- Example: If our samples come from a normal (Gaussian) distribution with a known SD, the sampling distribution of the (standardized) sample mean is also normal

- Use this sampling distribution to get the p-value: chance of obtaining a TS as or more extreme than the one we got, ASSUMING THE NULL IS TRUE
Permutation test

- A type of *nonparametric hypothesis test*
- Also called *randomization test, rerandomization test, exact test*
- Very widely applicable class of tests
- Introduced in the 1930s (that’s right, before everyone had a desktop/laptop computer!)
- Usually require only a few weak assumptions
- These tests often have good power

5 Steps to a permutation test

1. Analyze the problem: identify the NULL and ALT hypotheses
2. Choose a test statistic (TS)
3. Compute the TS for the original labeling of the observations
4. *** Rearrange (permute) the labels and recompute the TS for the rearranged labels (do for all possible permutations) ***
5. Decide whether to reject NULL based on this *permutation distribution*
Permutations

- A permutation is a reordering of the numbers 1, ..., n
- Example: What are some permutations of the numbers 1, 2, 3, 4 ??
- The NULL specifies that the permutations are all equally likely
- The sampling distribution of the TS under the NULL is computed by forming all permutations, calculating the TS for each and considering these values all equally likely

Example

- How could we carry out a permutation test to test the NULL hypothesis of no difference between treated and untreated ??

<table>
<thead>
<tr>
<th>Treated</th>
<th>121</th>
<th>118</th>
<th>110</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>95</td>
<td>34</td>
<td>22</td>
<td>12</td>
</tr>
</tbody>
</table>
Steps 3 and 4: reference NULL

- This is an example of an *unpaired 2-sample test*
- Example of test-statistic:
  sum of measures in the treated group
- Observed value: $121 + 118 + 110 + 90 = 439$
- Here, we have to find all of the combinations (since order within each group doesn’t matter)
- These are ...

Step 4: reference NULL

- There are $\binom{8}{4} = 70$ possible ways of choosing 4 objects among 8.
- Calculate the test statistic for all possible 70 combinations
- Look at the distribution of the TS.
Step 5: calculate the p-value

- There are two combinations which yield a TS as or more extreme than the one we got:
  - $121 + 118 + 110 + 90 = 439$
  - $121 + 118 + 110 + 95 = 444$

- Out of 70 possible combinations, the probability of obtaining a TS as or more extreme (p-value) is $2/70 = 0.03$
Advantages

- Can get a permutation test for any TS, even if its sampling distribution is unknown
- This gives more freedom in choosing a TS
- Can use on unbalanced designs
- Can combine dependent tests on mixtures of different data types (e.g. with numerical and categorical data)

Limitations

- Assumption that the observations are \textit{exchangeable under the NULL}
- This allows us to randomly move observations between the groups
- For example, when testing for a difference in 2 group means you would need to assume that the distributions in both groups have the same shape and spread
- Cannot use for testing hypotheses in a single population, or to compare groups that are different under the NULL
Sampling permutations

- What if the total number of possible permutations is too large for complete enumeration?
- Use *Monte Carlo sampling*: that is, randomly select some of the permutations
- Monte Carlo methods are based on the use of random numbers and probability to investigate problems
- The number of permutations to sample depends on desired accuracy

Introduction to ROC curves

- **ROC** = Receiver Operating Characteristic
- Started in electronic signal detection theory (1940s - 1950s)
- Has become very popular in biomedical applications, particularly radiology and imaging
- Also used in machine learning applications to assess classifiers
- Can be used to compare tests/procedures
ROC curves: simplest case

- Consider diagnostic test for a disease
- Test has 2 possible outcomes:
  - ‘positive’ = suggesting presence of disease
  - ‘negative’
- An individual can test either positive or negative for the disease
- In practice, the test yields a numerical score, and we classify the patient as “positive” or “negative” if the test is higher or lower than a given threshold.

Specific Example

Patients without the disease

Patients with disease

Test Result
Threshold

Call these patients “negative”  Call these patients “positive”

Test Result

Some definitions ...

Call these patients “negative”  Call these patients “positive”

Test Result

without the disease  with the disease

EMBnet course
Basel 20 January 2009
Test Result

Call these patients “negative”

Call these patients “positive”

without the disease

with the disease

False Positives

True negatives
Call these patients “negative”
False negatives
Call these patients “positive”

Test Result

without the disease
with the disease

Moving the Threshold: right

without the disease
with the disease

EMBnet course
Basel 20 January 2009
Moving the Threshold: left

Test Result

without the disease
with the disease

True Positive Rate (sensitivity)

False Positive Rate (1-specificity)

ROC curve
ROC curve comparison

A good test:

![ROC curve of a good test](image1)

A poor test:

![ROC curve of a poor test](image2)

ROC curve extremes

Best Test:

![ROC curve of the best test](image3)

Worst test:

![ROC curve of the worst test](image4)

The distributions don't overlap at all

The distributions completely overlap
Area under ROC curve (AUC)

- **Overall measure** of test performance
- **Comparisons** between two tests based on differences between (estimated) AUC
- For continuous data, AUC equivalent to *Mann-Whitney U-statistic* (nonparametric test of difference in location between two populations)

AUC for ROC curves

- AUC = 100%
- AUC = 90%
- AUC = 50%
- AUC = 65%
Interpretation of AUC

- AUC can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen nondiseased individual: 
  \[ P(X_i \geq X_j \mid D_i = 1, D_j = 0) \]
- So can think of this as a nonparametric distance between disease/nondisease test results

Problems with AUC

- No clinically relevant meaning
- A lot of the area is coming from the range of large false positive values, no one cares what’s going on in that region (need to examine restricted regions)
- The curves might cross, so that there might be a meaningful difference in performance that is not picked up by AUC
Examples using ROC analysis

- Threshold selection for ‘tuning’ an already trained classifier (e.g. neural nets)
- Defining signal thresholds in DNA microarrays (Bilban et al.)
- Comparing test statistics for identifying differentially expressed genes in replicated microarray data (Lönnstedt and Speed)
- Assessing performance of different protein prediction algorithms (Tang et al.)
- Inferring protein homology (Karwath and King)