Bootstrap, Jackknife and other resampling methods
Part V: Permutation tests

Rozenn Dahyot

Room 128, Department of Statistics
Trinity College Dublin, Ireland
dahyot@mee.tcd.ie

2005
So far

The resampling methods are:

- **Bootstrap resampling**: generate samples with the same size $n$ as $\mathbf{x}$ with replacement.

- **Jackknife subsampling**: generate samples with a smaller size than $\mathbf{x}$ without replacement.

Used for:

- Compute accuracy measures (standard error, bias, etc.) of a statistic $\hat{\theta}$ from one set $\mathbf{x} = (x_1, \ldots, x_n)$.

- Compare two sets of observations: the example of the mouse data
Example on the mouse data

<table>
<thead>
<tr>
<th>Data (Treatment group)</th>
<th>94; 197; 16; 38; 99; 141; 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data (Control group)</td>
<td>52; 104; 146; 10; 51; 30; 40; 27; 46</td>
</tr>
</tbody>
</table>

Table: The mouse data [Efron]. 16 mice divided assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery. Did the treatment prolong survival?
Example on the mouse data

1. Compute $B$ bootstrap samples for each group

- $\mathbf{x}^{*(b)}_{Treat} = (x^{*(b)}_{Treat\ 1}, \cdots, x^{*(b)}_{Treat\ 7})$
- $\mathbf{x}^{*(b)}_{Cont} = (x^{*(b)}_{Cont\ 1}, \cdots, x^{*(b)}_{Cont\ 9})$

2. $B$ bootstrap replications are computed: $\hat{\theta}^*(b) = \bar{x}^{*(b)}_{Treat} - \bar{x}^{*(b)}_{Cont}$

3. you can approximate the p.d.f. of the replications by a histogram.
Example on the mouse data

Figure: P.d.f. $P(\hat{\theta}^*)$ (histogram) of the replication $\hat{\theta}^*$ ($\hat{\theta} = 30.63$ and $\hat{se}_B = 26.85$).
Introduction

- Two sample problem: definitions

- Parametric solution

- Non parametric solution:
  - permutation test
  - randomization test
  - bootstrap test
The two sample problem

Two independent random sample are observed $x_a$ and $x_b$ drawn from possibly different probability density functions:

$$F_a \sim x_a = \{x_{a,1}, \cdots, x_{a,n}\}$$

$$F_b \sim x_b = \{x_{b,1}, \cdots, x_{b,m}\}$$

**Definition**

The **null hypothesis** $H_0$ assumes that there is no difference in between the density function $F_a = F_b$. 
Hypothesis test and Achieved significance level (ASL)

**Definition**

A **hypothesis test** is a way of deciding whether or not the data decisively reject the hypothesis $H_0$.

**Definition**

The **achieved significance level** of the test (ASL) is defined as:

\[
\text{ASL} = P(\hat{\theta}^* \geq \hat{\theta} | H_0)
\]

\[
= \int_{\hat{\theta}}^{+\infty} P(\hat{\theta}^* | H_0) \, d\hat{\theta}^*
\]

The smaller ASL, the stronger is the evidence of $H_0$ false. The notation star differentiates between an hypothetical value $\hat{\theta}^*$ generated according to $H_0$, and the actual observation $\hat{\theta}$. 

R. Dahyot (TCD) 453 Modern statistical methods 2005 8 / 22
A traditional way is to consider some hypotheses: $F_a \sim \mathcal{N}(\mu_a, \sigma^2)$ and $F_b \sim \mathcal{N}(\mu_b, \sigma^2)$, and the null hypothesis becomes $\mu_a = \mu_b$.

Under $H_0$, the statistic $\hat{\theta} = \bar{x}_a - \bar{x}_b$ can be modelled as a normal distribution with mean 0 and variance $\sigma_{\hat{\theta}}^2 = \sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right)$.

The ASL is then computed:

$$\text{ASL} = \int_{\hat{\theta}}^{+\infty} \frac{-(\hat{\theta}^* - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2} \frac{e^{-\frac{(\hat{\theta}^* - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2}}}{\sqrt{2\pi}\sigma_{\hat{\theta}}} \ d\hat{\theta}^*$$
Parametric test

- $\sigma$ is unknown and has to be estimated from the data:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_{ai} - \bar{x}_a)^2 + \sum_{i=1}^{m}(x_{bi} - \bar{x}_b)^2}{m + n - 2}$$

- For the mouse data $ASL = .131$ : the null hypothesis cannot be rejected.

- However, this (parametric) method relies on the hypotheses made while calculating the ASL.
Permutation tests

- *Permutation tests* are a computer-intensive statistical technique that predates computers.

- This idea was introduced by R.A. Fisher in the 1930’s.

- The main application of permutation tests is the two-sample problem.
Computation of the two sample permutation test statistic

Notation $m$ number of values in observation $x_{\text{Treat}}$, $n$ number of values in observation $x_{\text{Cont}}$.

If $\mathcal{H}_0$ is true, then:

1. We can combine the values from both observations in one of size $m + n = N$: $x = \{x_{\text{Treat}}, x_{\text{Cont}}\}$.

2. Take a subsample $x^*_{\text{Treat}}$ from $x$ of size $m$. The remaining $n$ values constitute the subsample $x^*_{\text{Cont}}$.

3. Compute the replication $\bar{x}^*_{\text{Treat}}$ and $\bar{x}^*_{\text{Cont}}$ on $x^*_{\text{Treat}}$ and $x^*_{\text{Cont}}$ respectively.

4. Compute the replication of the difference $\hat{\theta}^* = \bar{x}^*_{\text{Treat}} - \bar{x}^*_{\text{Cont}}$. 
Example on the mouse data

Figure: Histogram of the permutation replications $P(\hat{\theta}^* | \mathcal{H}_0)$. ASL is the red surface ($ASL_{perm} = 0.14$).

If the original difference $\hat{\theta} = d = \bar{x}_{Treat} - \bar{x}_{Cont}$ falls outside the 95% of the distribution of the permutation replication (i.e. $ASL_{perm} < 0.05$), then the null hypothesis is rejected.
Computation of the two sample permutation test statistic

1. \( x = \{ x_a; x_b \} \) of size \( n + m = N \).

2. Compute all:

   - \( \binom{N}{n} \) permutation samples \( x^* \). Select the \( n \) first values to define \( x_a^* \) and the last \( m \) ones to define \( x_b^* \)

   - \( \binom{N}{n} \) replications \( \hat{\theta}^*(b) = x_a^* - x_b^* \)

3. Approximate \( \text{ASL}_{perm} \) by:

\[
\hat{\text{ASL}}_{perm} = \frac{\# \{ \hat{\theta}^* \geq \hat{\theta} \}}{\binom{N}{n}}
\]
Remark on the permutation test

- The histogram of the permutation replications $\hat{\theta}^*$ approximates $P(\hat{\theta}^*|H_0)$.

- The resamples are not really permutations but more combinations.

- $\binom{N}{n}$ can be huge so in practice, ASL$_{perm}$ is approximated by Monte Carlo methods.
Computation of the two sample randomization test statistic

1. \( \mathbf{x} = \{ \mathbf{x}_a; \mathbf{x}_b \} \) of size \( n + m = N \).

2. Compute \( B \) times:
   - Randomly selected permutation samples \( \mathbf{x}^* \). Select the \( n \) first values to define \( \mathbf{x}^*_a \) and the last \( m \) ones to define \( \mathbf{x}^*_b \)
   - Compute the replications \( \hat{\theta}^*(b) = \bar{x}^*_a - \bar{x}^*_b \)

3. Approximate \( \text{ASL}_{perm} \) by:

\[
\widehat{\text{ASL}}_{perm} = \frac{\# \{ \hat{\theta}^* \geq \hat{\theta} \}}{B}
\]
Figure: Histograms of the bootstrap replications $\mathcal{P}(\hat{\theta}^*)$ (blue), and the permutation replications $\mathcal{P}(\hat{\theta}^* | \mathcal{H}_0)$ (red).
Remarks

- Permutation replications are computed without replacement.

- The distribution of permutation replications approximates $P(\theta^* | H_0)$.

- The bootstrap replications presented in the introduction are computed on resamples with replacements. The distribution of those bootstrap replications defines $P(\theta^*)$.

- Is there a way to get $P(\theta^* | H_0)$ using a bootstrap method?
Computation of the two sample bootstrap test statistics

1. \( \mathbf{x} = \{\mathbf{x}_a; \mathbf{x}_b\} \) of size \( n + m = N \).

2. Compute \( B \) times:
   - Bootstrap samples from \( \mathbf{x} \). Select the \( n \) first values to define \( \mathbf{x}_a^* \) and the last \( m \) ones to define \( \mathbf{x}_b^* \).
   - Compute the replications \( \hat{\theta}^*(b) = \bar{x}_a^* - \bar{x}_b^* \)

3. Approximate \( \text{ASL}_{\text{boot}} \) by:

\[
\hat{\text{ASL}}_{\text{boot}} = \frac{\#\{\hat{\theta}^*(b) \geq \hat{\theta}\}}{B}
\]
Example on the mouse data

Figure: Histogram of the bootstrap replications in the two sample test $\mathcal{P}(\hat{\theta}^*|\mathcal{H}_0)$. ASL is the green surface ($\text{ASL}_{\text{boot}} = 0.13$).
Relationship between the permutation test and the bootstrap test

- Very similar results in between the permutation test and the bootstrap test.

- $\text{ASL}_{perm}$ is the exact probability.

- $\text{ASL}_{boot}$ is not an exact probability but is guaranteed to be accurate as an estimate of the ASL, as the sample size goes to infinity.

- In the two-sample problem, the permutation test can only test the null hypothesis $F_a = F_b$ while the bootstrap can perform other hypothesis testing.
Hypothesis testing has been introduced, involving the computation of a probability ASL.

Permutation, Randomization and bootstrap tests have been introduced as alternative to parametric tests.

Again the main difference in between those nonparametric tests, is the way the resamples are computed (with or without replacements).