# Resampling Methods for Statistical Estimation and Hypothesis Testing

Philip Sabes

April 20, 2005

The presentation in this lecture draws from two good (and comprehensive) references on resampling techniques:

- **E&T:** *An introduction to the Bootstrap,* Bradley Efron and Robert Tibshirani. CRC Press, 1993.
- **Good:** *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses,* Phillip Good. Springer, 2004 (2nd Ed).

## 1 Resampling and the Bootstrap

*Estimation:* derive a guess for the value of some statistic of a distribution, $q(F)$, given random samples from that distribution, $x_i \sim F(x)$, $X \equiv \{x_i\}_{i=1}^N$.

Not only do we want to know $q(X)$, but we want to know how well we know $q$!

### 1.1 Example: Standard Error

- Given a sampling $\{x_i\}_{i=1}^N$ from a distribution $F(x)$, how do we estimate the standard error of the mean?

$$s_{\bar{x}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- But where does this come from?
- If you had infinite time and resources, how would you determine the *true* std err?

### 1.2 Empirical Distribution and Resampling

- Empirical Distribution:

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i)$$

- Resampling with replacement from $\hat{F}(x)$ is equivalent to repeated experiments on the Empirical Distribution.
- Matlab makes this VERY easy. If X is a sample vector, Nx1:

```
Xboot = X( ceil(N*rand(N,1)) );
```

## 1.3 The Bootstrap Estimate of Standard Error

- Say you want to know the expected standard deviation of some statistic $q(X)$. Follow this simple recipe:

    A. Get a bootstrap resampling of $X$, $X_{\text{boot}}$
    B. Compute $q_{\text{boot}} = q(X_{\text{boot}})$
    C. Repeat steps A and B many (a few hundred, 1000) times, and save the values of $q_{\text{boot}}$.
    D. Compute the standard deviation of the $q_{\text{boot}}$, $s_q$.

- Matlab makes this VERY easy. If X is a sample vector, Nx1:

```
Qboot = zeros(B,1);
for b=1:B,
      Xboot = X( ceil(N*rand(N,1)) );
      Qboot(b) = QFUNC( Xboot );
end
sQ = std(Qboot);
```

- So Why bother?

    - Standard error for non-Gaussian distributions?
      Except: Central Limit Theorem!
      See Example. The bootstrap estimate has a downward bias for low N!
    - So if $q$ is the mean, we can usually rely on standard eqn. However here $q$ can be *anything*.
      e.g. energy in theta band of LFP recorded in repeated trials
      e.g. variance in reach endpoints (variance in variance estimate is a *very* common use of the bootstrap).
    - $q$ could even be a vector of statistics, and we could compute the covariance across the statistics!
      e.g.????

- "Plug-in principle" and empirical distributions

## 1.4 More General Data Structures

The same approach can be used to get error bars on any statistic from almost any set of data. Here we consider several more complex data structures.

- Two-sample problems

  Given a sampling $\{x_i\}_{i=1}^{N_x}$ drawn i.i.d from $F(x)$ and a sampling $\{y_i\}_{i=1}^{N_y}$ drawn independently and i.i.d from $G(x)$, can we put error bars on the estimate of $z = \bar{x} - \bar{y}$?

  $z$ is statistic of $\{X, Y\}$, so by the plug-in principle we just need synthetic data $\{X_{\text{boot}}, Y_{\text{boot}}\}$.

  Resample from $\hat{F}(x)$ and $\hat{G}(y)$ independently.

- The Regression Model

  The statistical model underlying linear regression is:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon)$$

  In a "generative" spirit, we write

$$\alpha, \beta, \{\epsilon_i\}, \{x_i\} \Rightarrow \{y_i\}$$

"Estimation", in this case linear regression, can be written:

$$\{x_i, y_i\} \Rightarrow \hat{\alpha}, \hat{\beta}$$

We would like to know the standard deviation of our estimates $\hat{\alpha}$ and $\hat{\beta}$, given the generative model. This is well worked out for the linear-Gaussian case (see, for example, Zar). But we can also take a resampling approach.

A. Fit $\hat{\alpha}$ and $\hat{\beta}$ from $\{x_i, y_i\}$.

B. Compute the set of *estimated* residuals, $\hat{\epsilon}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$

C. Resample with replacement from the $\{\hat{\epsilon}_i\}$ to obtain $\{\hat{\epsilon}_{b\,i}\}$

D. Create bootstrap output data, $y_{b\,i} = \hat{\alpha} + \hat{\beta} x_i + \hat{\epsilon}_{b\,i}$.

E. Fit $\hat{\alpha}_b$ and $\hat{\beta}_b$ from $\{x_i, y_{b\,i}\}$.

F. Repeat steps C-E many (1000?) times.

G. Compute the standard deviation of the set of $\alpha_b$ and $\beta_b$ (or covariance, in the case of multi-dimensional inputs).

- General model+residual approach
- Time Series

  Very nice example in E&T (lutenizing hormone, p.92):

  $$x_t = \beta x_{t-1} + \epsilon_t$$

  It is typically very difficult or impossible to derive simple closed-form equations for the variability in the parameter estimates in models such as these.

## 1.5 Confidence Intervals

- Bootstrap gives you a value for $s_q$. If you think the $q_{\text{boot}}$ are normally distributed, then you can just look up confidence intervals using the Normal distribution:

  $$\text{With confidence } 100(1-\alpha)\%, \quad q \in \left[\hat{q} - s_q\, z^{(1-\frac{\alpha}{2})}, \hat{q} - s_q\, z^{(\frac{\alpha}{2})}\right],$$

  where $z^{(\alpha)}$ is the $100\alpha$-th percentile point for a standard normal distribution, $N(0,1)$.

- *The percentile interval.* In fact, the $q_{\text{boot}}$ are likely *not* to be Normally distributed, especially if $q(X)$ is a non-linear function of $X$. (Remember CLT). Instead we look at the percentiles of $q_{\text{boot}}$.

  Let $\hat{G}$ be the (empirical) cdf for $q_{\text{boot}}$. Then we use confidence intervals:

  $$\text{With confidence } 100(1-\alpha)\%, \quad q \in \left[\hat{G}^{-1}(\frac{\alpha}{2}), \hat{G}^{-1}(\frac{1-\alpha}{2})\right]$$

  Again, Matlab makes this very easy. For a 95% confidence interval, all you need is

  CLQ = prctile(Qboot,[2.5 97.5])

- *t-Test.* Given confidence limits, one could imagine performing a Bootstrap-based t-Test. This is possible (see E&T, Chps. 12-13). But better to do Permutation Test (see below).

## 2 Permutation Test

The Bootstrap is best for *estimation*. For *hypothesis testing*, I recommend the Permutation test.

### 2.1 Example: Two-Sample t-Test

Given a sampling $\{x_i\}_{i=1}^{N_x}$ drawn i.i.d from $F(x)$ and a sampling $\{y_i\}_{i=1}^{N_y}$ drawn independently and i.i.d from $G(x)$, can we determine whether $\mu_x = \mu_y$?

- Standard approach: t-Test. Compare difference in means $d = \bar{x} - \bar{y}$ to expected variability in that difference, given $s_{\bar{X}}$ and $s_{\bar{Y}}$:
$$s_d = \sqrt{s_x^2/N_x + s_y^2/N_y}.$$
- What is Null Hypothesis, $H_0$?
$$H_0 : \mu_x = \mu_y$$
- How can we permute/jumble/randomize the dataset in a way which shouldn't matter under $H_0$?

  Simply randomize the $x$ and $y$ labels of the data:
$$z_i = \left\{ \begin{array}{ll} x_i, & i = [1 : N_x] \\ y_{i-N_x}, & i = [(N_x + 1) : (N + x + N_y)] \end{array} \right. .$$
  Under $H_0$, the means of the first $N_x$ and second $N_y$ data entries in $z$ should be the same, on average across repeated trials, even if the vector $z$ is randomly permuted.
- A Two-Sample Permutation Test difference of means:
  A. From samples $X$ and $Y$, create composite vector $Z$.
  B. Permute the elements of $Z$ to get $Z_{\text{perm}}$.
  C. Get $X_{\text{perm}}$ and $Y_{\text{perm}}$ by selecting the appropriate entries in $Z_{\text{perm}}$.
  D. Compute the difference in means, $d_{\text{perm}} = \bar{X}_{\text{perm}} - \bar{Y}_{\text{perm}}$.
  E. Repeat steps B-D many (1000?) times, and save the values of $d_{\text{perm}}$.
  F. Percentage of $d_{\text{perm}}$ with absolute value greater than $|d|$ is a measure of *p-Value*.
- Once again, Matlab makes this easy:

```
D = mean(X)-mean(Y);
Dperm = zeros(R,1);
Z = [X;Y];
for r=1:R,
    [tmp,i]=sort(rand(Nx+Ny,1));
    Zperm = Z(i,:);
    Dperm(r) = mean(Zperm(1:Nx,:))-mean(Zperm(Nx+[1:Ny],:));
end
p = mean( abs(Dperm) > abs(D) );
```

- Very easy to extend!
  - Other statistics, e.g. difference in median values, or percentiles
  - **Differences in variance**

– Multivariate comparisons:
What is the equivalent of $D$?
Hotelling's $T^2$:

$$V_{j,k} = \frac{1}{N_x + N_y - 2}\left[\sum_{i=1}^{N_x}(X_{i,j} - \bar{X}_j)(X_{i,k} - \bar{X}_k) + \sum_{i=1}^{N_y}(Y_{i,j} - \bar{Y}_j)(Y_{i,k} - \bar{Y}_k)\right]$$

$$T^2 = (\bar{X} - \bar{Y})' V^{-1} (\bar{X} - \bar{Y})$$

Mahalnobis distance...

– Paired t-Test: $x_i$ and $y_i$ are paired.
The Null Hypothesis $H_0$ is the same as for the unpaired case above.
How do you randomize the data while preserving $H_0$?

## 2.2 General Formulation

It is difficult to write down a general formulation for devising a Permutation Test. Learning by example is best, and cleverness is often required. However, the general principle is always the same.

A. Determine what $H_0$ and $H_1$ are.

B. Devise a statistic $q$ which, say, is excted to be lower if $H_0$ is true and higher if $H_1$ is true.

C. Devise some way to permute your data so that *on average* it will have no effect on $q$ if $H_0$ really is true, but it will cause $q$ to fall to its $H_0$ range if $H_1$ is true.

D. Compute the true $q$ and a large set of $q_{\text{perm}}$ using the permutation scheme from C.

E. Let $p$ be the percentage of $q_{\text{perm}}$ that are greater $q$. If $p$ is small, we are unlikely to have obtained our dataset under $H_0$, and so we reject $H_0$ in favor of $H_1$ with confidence level $100 - p$.

## 2.3 Regression

What is a hypothesis test in a regression model?

$$\begin{aligned} H_0 : \quad & y_i = \alpha + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon) \\ H_1 : \quad & y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon) \end{aligned}$$

How do you randomize the data while preserving $H_0$? Randomize the inputs, $x_i$, and see how it affect our ability to fit $y_i$.

A. Fit $\hat{\alpha}$ and $\hat{\beta}$ from $\{X, Y\}$.

B. Compute $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

C. Compute some measure of "Goodness of Fit", typically $R^2$ or the Sum-Squared-Error (SSE),

$$SSE(Y, \hat{Y}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

.

D. Permute the elements of $X$ to obtain $X_{\text{perm}}$.

E. Fit $\hat{\alpha}_{\text{perm}}$ and $\hat{\beta}_{\text{perm}}$ from $\{X_{\text{perm}}, Y\}$ and compute the goodness of fit, e.g. $SSE_{\text{perm}} = SSE(Y, \hat{Y}_{\text{perm}})$.

F. Repeat steps D-E many (1000?) times.

G. Percentage of $SSE_{\text{perm}}$ with a value better than (less than) the true SSE is a measure of *p-Value*, i.e. the probability of $\{X, Y\}$ given $H_0$.

Easy to generalize to multiple regression and model-selection!

## 2.4 Other Examples

- Temporal patterns: does firing rate increase/decrease significantly during presentation of a stimulus, or does the neuron fire homogeneously during the stimulus presentation?

  e.g. you have $\{r_{i,t}\}$, $i \in [1, N], t \in [1, T]$, for $N$ repeated trials, each with $T$ time bins.

  - What is null hypothesis, $H_0$?
  - How do you randomize the data while preserving $H_0$?

- Regression. Is slope $\beta$ different from some known value, $\beta_o$?

  - What is null hypothesis, $H_0$?
  - How do you randomize the data while preserving $H_0$?