Permutation Test & Monte Carlo Sampling

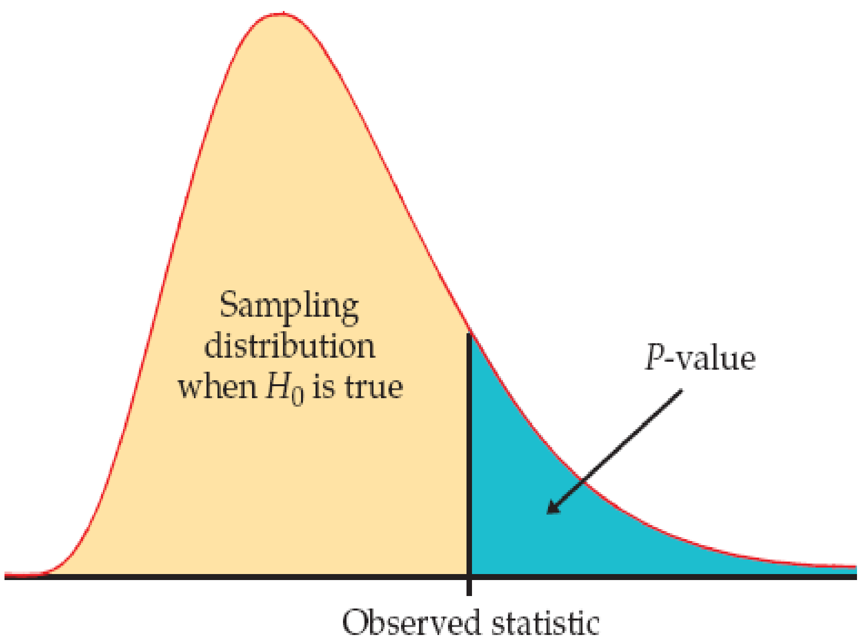
by MA, Jianqiang

March 18th, 2009

Outline

- Introduction to Permutation Test
- Permutation Test in Linguistics : Measuring Syntactic Differences
- Brief View of Monte Carlo Method
- Monte Carlo in Linguistics: An Simple Example
- Conclusion

Hypothesis Test



- Define H0,H1..
- Choose Test (t, Z, F, etc) *then we know test statistic distribution under H0*
- Compute Test Statistic
- Make Statistical Decision *by looking the observed statistic in the distribution*

- P-value: that probability that we would observe a statistic value as extreme or more extreme than the one we did observe

Assumption for a z-test, t-test or F-test

- When conducting a z-test or a t-test, we are actually assuming that the data (or the random errors) follow a normal distribution.
- Based on this assumption, we know the distribution of the test statistic (T.S.) under the null hypothesis.
- Based on the distribution (z-distribution, t-distribution or F-distribution), we get a p-value for each observed T.S..
- This can be referred to as “parametric approaches”.

What if the distributional assumption does not hold?

- If the normal assumption does not hold for the data and the sample size is small, the results of z-test, t- or F-test are not reliable.
- What can we do?

1) Transformation of data to make the data normal

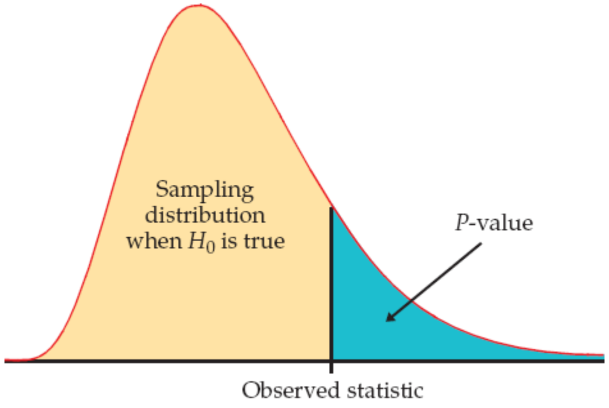
2) Choose some tests that do not make such distributional assumptions – “nonparametric approaches”

Permutation Test

- Permutation Test (randomization tests) can be used without the normal assumption for the distribution of data.
- Permutation Test is a resampling test (like bootstrapping)
- Permutation Test is an Exact Test
- Monte Carlo Sampling: makes testing on large data possible

Idea of permutation test

- Under H_0 (the null hypothesis), some of the data are exchangeable.
- We permute (rearrange) the data by shuffling their labels of treatments, and then calculate our T.S. on each permutation. The collection of T.S. from the permuted data constructs the distribution under H_0 .



An example of Permutation

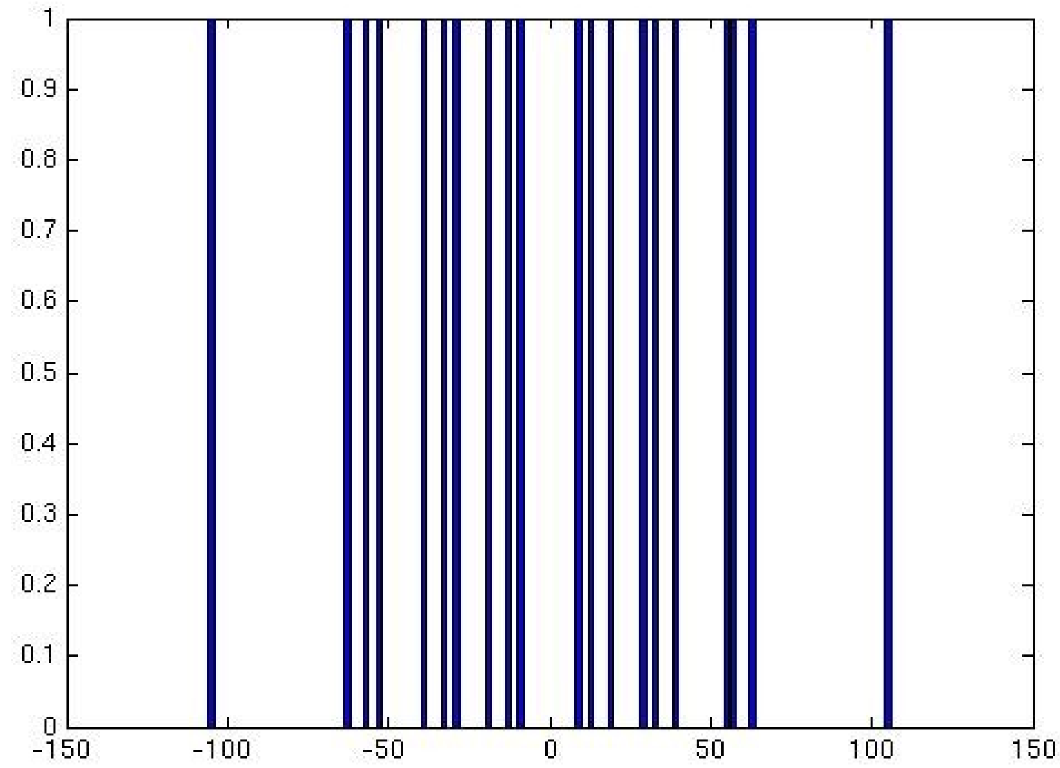
- Two groups of participants, score of a linguistics test:

  Group A: 55 58 60   Group B: 12 22 34

- Statistic= $X_A - X_B$, In the observation=173-68=105
- Rearrange the observations and compute corresponding T.S.
- Compare the T.S. from original observation with the ones from re-arranged data.
- In this case, TS(observation) is the biggest, thus the p-value is 1/20=0.05

Distribution of $X_A - X_B$

Order	Group 1	Group 2
1	55 58 60	12 22 34
2	55 58 12	60 22 34
3	55 58 22	12 60 34
4	55 58 34	12 22 34
5	55 12 60	58 22 34
6	55 22 60	12 58 34
7	55 34 60	12 22 58
8	12 58 60	55 22 34
9	22 58 60	12 55 34
10	34 58 60	12 22 55
11	12 22 60	55 58 34
12	12 58 22	55 60 34
13	55 12 22	12 55 58
14	12 34 60	55 58 34
15	12 58 34	55 22 60
16	55 12 34	12 58 60
17	22 34 60	55 58 34
18	22 58 34	55 22 60
19	55 22 34	12 58 60
20	12 22 34	55 58 60



Application: Measuring Syntactic Distance

- By John Nerbonne and Wybo Wiersema 2006
- Measure linguistic contamination

  mobility, multilinguality
- Languages in contact influence one another

  first languages influence second languages, vise versa
- What are the factors, how important are they?

  experience, attitude, instruction, relations of languages
- Differences between varieties of a language

The Idea

- Goal: detect lots of syntactic differences
- Material: Corpora of language use in contact situations (e.g. 2 corpus of Finnish Australian Immigrants, of adults and kids respectively)
- Mark syntactic categories of words with *Part-of-speech (POS) tags*
- Collect and analyse trigrams of tags

How to measure? Indirectly!

- We aim to observe differences in syntactic use
  - including overuse and underuse, not just “errors”
- Indirect, since it's difficult to model syntactic difference
- Lexical categories mirror syntactic analysis

- We assume that syntactic differences correlate strongly with the distribution of POS tag-trigrams

Trigram Vectors and their Differences

- Finnish people who emigrated to Australia
- Two groups of participants, got two sub-corpus

  Kids (< 17) — 30 interviews & Adults (>=17) — 60 interviews
- Frequency Vectors containing the counts of 13,784 different POS trigrams, one for each of the sub-corpus
- Measure Vector Differences

  Using cosine, R/Rsq comparing two vectors

Statistical Significance

- Aarts & Granger examined tag-trigrams, but did not subject their collections to statistical analysis
- We do not have general distribution of these trigrams or distribution of syntactic differences
- We have:13,784 trigrams actually occurred
- Solution: permutation test, with Monte Carlo techniques

Normalization Problem in this case

- we need to permute sentences, not trigrams to avoid measuring only the effect of syntactic coherence
- Normalization for sentences length

  Since average sentence length differs in two sub-corpus (24 wd/sent. vs. 16 wd/sent.), number of trigrams will differ across permutation as well → numbers of trigrams in each group will vary if no normalization is applied.

Normalization in Detail (1)

- Initially: a series of counts of all the trigrams of vectors the young group vs. the older group.

1. Sums no. of trigrams for each vector

$$\mathbf{c}^{\mathrm{y}} = < c_1^y, c_2^y, ..., c_n^y > \qquad N^y = \sum_{i=1}^n c_i^y$$
$$\mathbf{c}^{\mathrm{o}} = < c_1^o, c_2^o, ..., c_n^o > \qquad N^o = \sum_{i=1}^n c_i^o$$
$$N(= N^y + N^o)$$

2. compute the frequencies based on counts and sums.

$$\mathrm{f}^{\mathrm{y}} = < ..., f_i^y(= c_i^y/N^y), ... > \qquad \sum_{i=1}^n f_i^y = 1$$
$$\mathrm{f}^{\mathrm{o}} = < ..., f_i^o(= c_i^o/N^o), ... > \qquad \sum_{i=1}^n f_i^o = 1$$

Normalization in Detail (2)

3. weight these frequencies on the basis of the distributions in the aggregated categories

$$\begin{aligned} \mathbf{p^y} &= < ..., p_i^y (= f_i^y / (f_i^y + f_i^o)), ... > \\ \mathbf{p^o} &= < ..., p_i^o (= f_i^o / (f_i^y + f_i^o)), ... > \end{aligned}$$

4. compute final elements of vectors (here, $c_i = c_i^y + c_i^o$)

$$\begin{aligned} \mathbf{w^y} &= < ..., p_i^y \cdot c_i, ... > \\ \mathbf{w^o} &= < ..., p_i^o \cdot c_i, ... > \end{aligned}$$

- Another Normalization is skipped here, anyway, we can see from this case normalization is useful for deal with real data in which is not perfectly “exchangeable”

Apply Permutation Test

1. Determine difference between 2 vectors of trigrams, which is our test statistic
2. Permute a pair of sentences from two sub-corpus, compare the differences of resulting two vectors of trigrams (compute test statistics for this permutation)
3. Repeat step (3) e.g. 10,000 times, each time, we pick pairs of sentences randomly.
4. Estimation of stat. significance, the probability that the original samples were due to chance (p-value).

Findings

- Relative difference between young and old emigrants significant ($P<0.001$)
- Some striking patterns:

’	it	’s	very	low	tax	in	here
PAUSE	PRON	COP	INTNS	ADJ	N-COM	PREP	ADV
a	boat	and	I	was	professional	fisherman	
ART	N-COM	CONJ	PRO	COP	ADJ	N-COM	

- Problems caused by tagger (elided here)

So, where is Monte Carlo?

--what is Monte Carlo (sampling)?

“3. Repeat step (3) e.g. 10,000 times, each time, we pick pairs of sentences randomly. ”

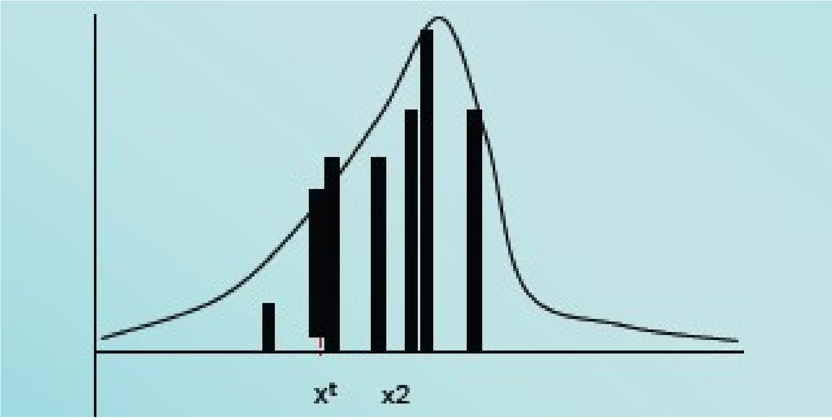
--why bothering?

In permutation test, there may be too many possible orderings of the data to conveniently allow complete enumeration

This is done by *generating the reference distribution by* Monte Carlo sampling, which takes a relatively small random sample of the possible replicates

Monte Carlo principle

- Given a very large set X and a distribution p(x) over it
- Draw N samples randomly from the distribution
- Approximate the distribution using these samples



$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x^{(i)} = x) \underset{N\to\infty}{\longrightarrow} p(x)$$

- Can also approximate expectation

$$E_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \underset{N\to\infty}{\longrightarrow} E(f) = \sum_x f(x)p(x)$$

Monte Carlo: a simple example

- Find out the probability that, out of a group of 30 people, 2 people share a birthday

1. Pick 30 random numbers in the range [1,365]. Each number represents one day of the year.
2. Check to see if any of the thirty are equal.
3. Go back to step 1 and repeat 10,000 times.
4. Report the fraction of trials that have matching days.

--Results: 0.7129, which is very close to exact result

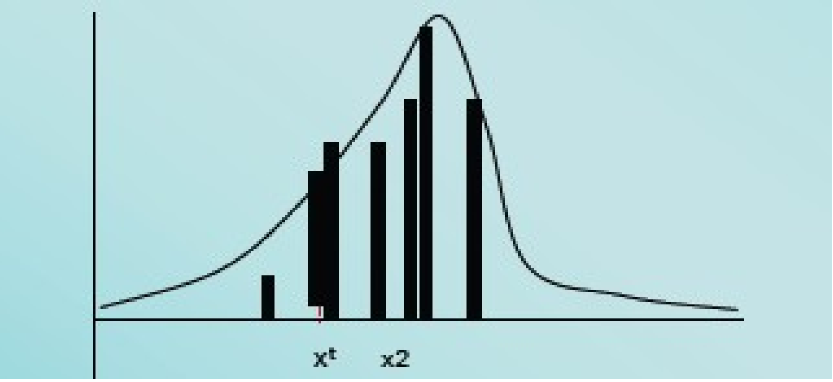
Another example: calculating pi:

http://www.eveandersson.com/pi/monte-carlo-circle

Features of Monte Carlo in general

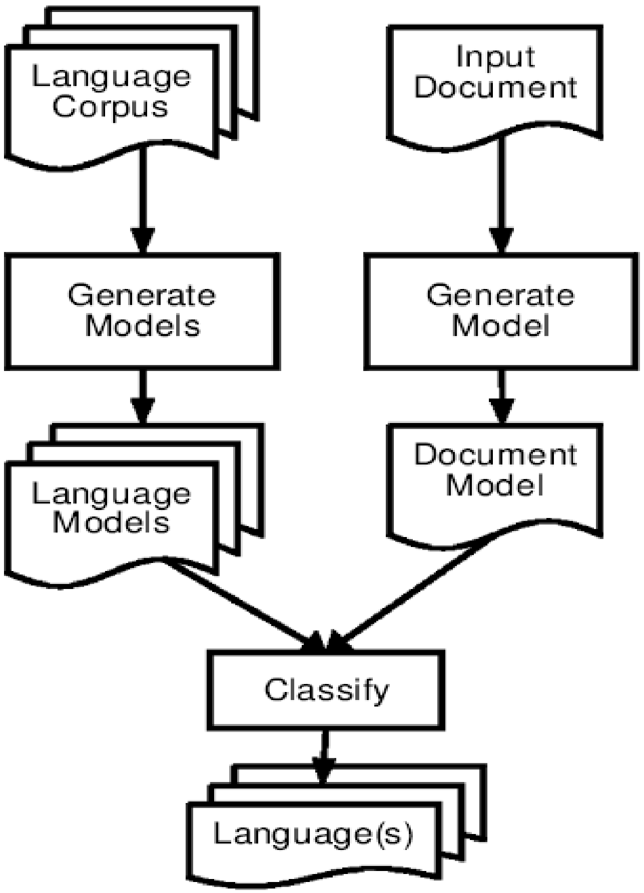
- A domain of possible inputs
- Random number generating and sampling

  rejection, metropolis and exact sampling...



- Error estimation

An application: Identifying Language



- Language Model:

  most frequent words/most frequent N-grams of alphabets

- Document Model: similar features as in language model

- Classification Methods:

  rank order statistic, mutual information statistics, Monte Carlo Method

Identifying Language by Monte Carlo (1)

By Arjen Poutsma

- Find the most probable language given a certain document, i.e. maximize $P(L|D)$
- Apply Bayesian Law:

$$\begin{aligned} \max P(L|D) &= \max \frac{P(L) \cdot P(D|L)}{P(D)} \\ &\approx \max P(D|L) \end{aligned}$$

- As both language and documents are features:

$$\max P(L|D) = \max \sum_{f \in D} P(f|L)$$

Identifying Language by Monte Carlo (2)

- we can determine the language of this document to be the language which results most often from these random features.

  Monte Carlo Approach:

  1. Generating random number and sample one feature from all features of the document

  2. Check which language(s) also have this feature

  3. Repeat 1~2 for N times

Results of Monte Carlo Method

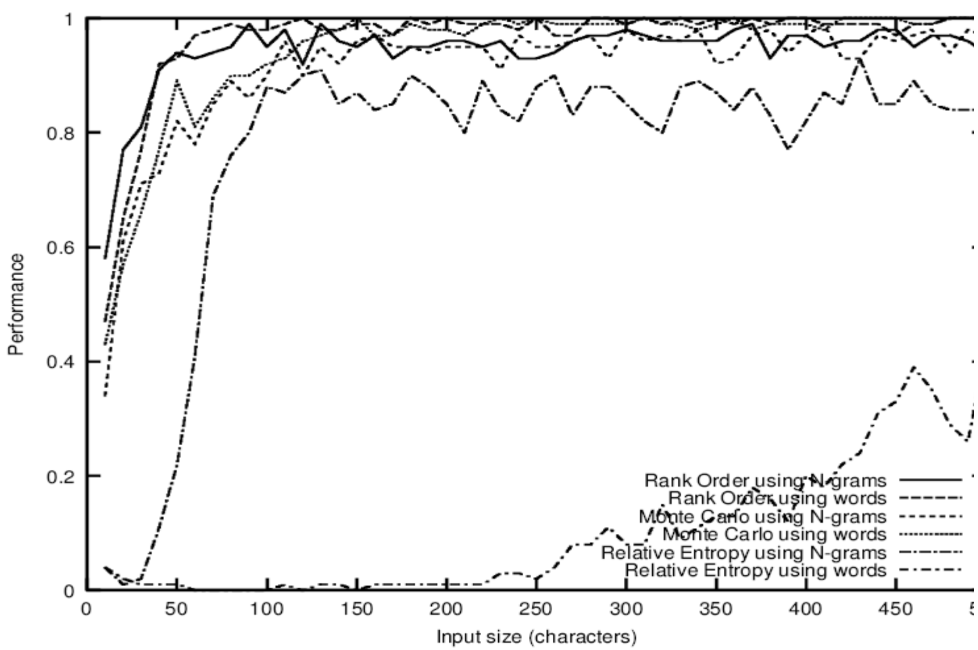


Figure 3: Performance score for six Language Identification methods.

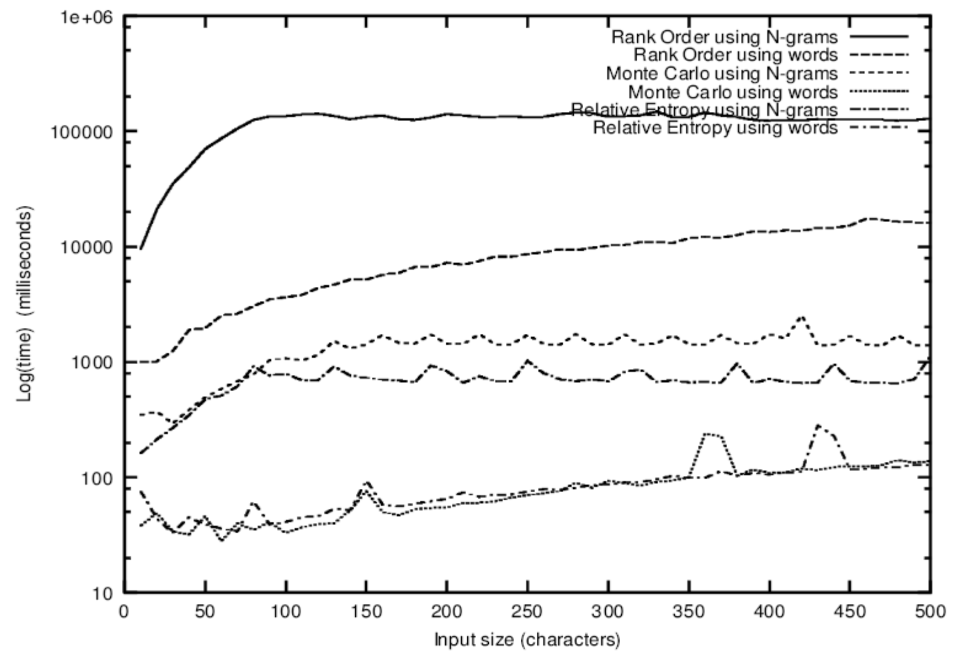


Figure 4: Time required for three Language Identification methods.

- Performance is close to the best
- Time complexity is much lower than the best

Conclusion

- Permutation Test is a good choice for hypothesis test of unknown distribution.

  It works regardless of the shape and size of the population
  gives exact p value

- Monte Carlo Sampling is introduced to permutation test when it is impossible to complete enumeration the data.
- Monte Carlo Method can well approximate the distribution using random samples