Pattern Analysis. Lab 3.

Konstantin Tretyakov

November 9, 2006

Hypothesis Testing

Is your data "special" or not? Does it originate from system A or B? Is your assumption true or false? — these are all examples of *hypothesis testing*, which is a common issue in data analysis.

In all these cases we start by presenting two mutually exclusive statements, denoted H_0 and H_1. Statement H_0 is called the *null hypothesis* and describes the distribution of data if it were "uninteresting" (i.e. our assumption about it were not true). Statement H_1 is known as *alternative hypothesis* and describes the property of interest. After fixing the two hypotheses we design a *statistic* $T(X)$, that should, in some sense, correspond to H_1. At last, we examine the value $T(\mathbf{x})$ on the data item $\mathbf{x}$ and calculate the probability of getting this value equal to or "worse" than $T(\mathbf{x})$ if $\mathbf{x}$ were distributed according to H_0. This probability is known as *p-value* and it's quite often calculated as:

$$p = \mathbf{P}(T(\underline{X}) \geq T(\mathbf{x})).$$

If p is lower than some predefined *confidence threshold* (that is, it's too improbable to see such $T(\mathbf{x})$ in "usual" data), we *reject* H_0 *in favor of* H_1. Otherwise, we state that *we can't reject* H_0. In this case we don't reject H_1, though, because the fact of p-value not being low enough is mostly not significant enough to disprove H_1.

Exercise 1: Does a low p-value *prove* hypothesis H_1?

Exercise 2: We shall now practice hypothesis testing on a small pseudo-bioinformatical example. The file `dna1.txt` contains a string of characters `A`, `C`, `T`, `G`, representing a DNA sequence. We suppose that this might originate from a certain gene region, that is known to contain many occurences of the substring `GCACC`. We'd like to test this hypothesis. Formulate H_0 and H_1. Specify $T(X)$. Evaluate $T(\mathbf{x})$ on given data. How probable would it be to obtain such $T(\mathbf{x})$ if H_0 were true?

Hints:

- Loading the string from the file:
  ```
  s = mgetl("dna1.txt");
  ```
- Finding all occurences of `GCACC` in `s`:
  ```
  n = strindex(s, "GCACC");
  ```

You must have noted that it is complicated to derive an analytical expression for the p-value in the previous exercise. So we'll have to estimate it analytically, using a *randomization test*. The idea is simple: we generate random samples from the distribution H_0, evaluate the statistic on these, and examine the distribution of obtained values.

Exercise 3: Generate 100 uniformly random strings of length 1000. For each of them evaluate $T(X)$, (i.e. count the number of occurences of `GCACC`). Plot the histogram of obtained values. Estimate the p-value of the previous exercise.
Hints:

- Generating a random string:
  ```
  i = grand(1, 1000, 'uin', 1, 4);
  s = part('ATCG', i);
  ```
- Appending a value `n` to a vector `v`:
  ```
  v($+1) = n;
  ```
- Counting how many values in the vector `v` exceed `n`:
  ```
  count = length(find(v > n));
  ```
- Plotting a histogram of values in `v` (nicer than `histplot`):
  ```
  t = tabul(v);
  bar(t(:,1), t(:,2));
  ```

Exercise 4: Now we've also got this second piece of DNA (`dna2.txt`). We think it might come from some other interesting gene region, containing some repeating substring of length 5, but we don't know what substring exactly should be repeating. Formulate H_0 and H_1, specify $T(X)$. Of course, $T(X)$ will count the number of occurences of the most frequently repeating substring of length 5. Check it on the data. Is it significant?
Hints:

- Selecting a substring of `s` of length 5 at position `i`:
  ```
  subs = part(s, i:(i+4));
  ```

- Counting how many times each item is present in vector `v`:
  `t = tabul(v);`

- Finding the index `w` and value `m` of the maximal entry of `t(2)`:
  `[m, w] = max(t(2));`

Exercise 5: How to determine whether our finding in the previous exercise is significant? It's even more difficult to present an analytical expression for the probability to have *the most frequent string* repeat itself n times, so we'll have to refrain to randomization testing again. Generate 100 random strings, and evaluate $T(X)$ for each of them. Examine the distribution and evalate the p-value of interest. Make conclusions about the significance of your finding in Exercise 4. Compare it to the first example in this section. Explain the differences.