# T-61.5020 Statistical Natural Language Processing

Exercises 1 — Basics of probability calculus
Version 1.0

1. Following probabilities might apply to English:

   $$P(\text{ word is abbreviation } | \text{ word has three letters }) = 0.8$$

   $$P(\text{ word has three letters }) = 0.0003$$

   What is the probability that an observed word is a three letter abbreviation?

2. Let us observe a stemming program for Finnish. By using the context, it can conduct whether the stem for word-form *"siitä"* is *"se"* (it) or *"siittää"* (conceive). For a word-form of *"se"* the program can determine the right stem for probability of 0.95. The same holds for the word-forms of *"siittää"*. Because the stem *"se"* is much more common, only one of a thousand *"siitä"* should be conducted to the stem *"siittää"*.

   The program tells that for an occurred word-form *"siitä"*, the corresponding stem is *"siittää"*. What is the probability that the program is correct?

3. So called Zipf's law is often referred when simple statistics are calculated from a language. Let the words be tabulated so that the most common word comes first ($r = 1$), and the rest follow by the order of frequency ($r = 2, 3, \ldots$). Next to the word is written how many times it occurred it the text ($f$). Zipf alleges that

   $$f \propto \frac{1}{r}$$

   i.e. $f$ times $r$ remains constant.

   Does this apply to a randomly generated language, that has 30 letters including the word boundary?

4. a) A 101-sided dice is thrown. The sides have numbers 0 – 100. Calculate expectation value and variance for the result. Sketch a graph for probability p(X) when X is the result.

   b) Two similar 101-sided dices are thrown, and the sum of results is divided by two. Calculate expectation value and variance. Sketch a graph for probability p(X) when X is the sum divided by 2.

   c) Now ten dices are thrown. Again, draw a graph for the sum of results divided by the number of dices.

   d) Let's use all the dices of the world ($n \to \infty$). What kind of distribution do we have? Draw a graph.

5. *Minimum Description Length principle* (MDL) is a method for finding best model for an observed data. In MDL we try to find a description with which the data can be compressed to least number of bits. In so-called *Ideal MDL* the goal is to find the shortest possible Turing machine (i.e. computer program) that prints the desired data sequence, and then halts.

   Ideal MDL is of little practical use, since it has been shown that it is impossible to design an algorithm that finds the shortest possible Turing machine. Therefore there exists several variations, of which one quite common is so called *two-part coding scheme*.

   In two-part coding scheme we first select a model class that can describe some data given the model parameters $\theta$. The goal is to describe and send a set of data, $x$, that is assumed to be generated by a model of the decided class, with the minimum number of bits possible. As the receiver does not know the parameters that we choose, also they must be sent. Denote $L(\theta)$ as the description length needed to encode the parameters, and $L(x \mid \theta)$ as the description length needed to encode the data when the parameters are known. Thus we need to minimize the total code length $L(x, \theta) = L(\theta) + L(x \mid \theta)$.

   In statistical modeling, the coding of the model class corresponds to probability distribution $p(X \mid \theta)$, and the coding of parameters to distribution $p(\theta)$. From information theory we know that if probability of a message is $p(i)$, the minimum code length for the message is $-\log p(i)$ bits. Show that the optimal selection of the parameters in two-part coding scheme equals to *Maximum A Posteriori* estimation.