INF5830, 2015, 24 Sept, Applying statistical tools in NLP

In this session we will apply some of the statistical tools from the lectures. We will also consider some additional tools. The lab session will be a mixture of terminal work and discussions with the use of the white board. You are advised to work on the practical parts before the lab session, and to start thinking about the theoretical parts. You should also have worked through the sheet on NumPy and SciPy before the session.

Exercise 1 – Estimation

a) We will look at sentence lengths in the Brown corpus and how it corresponds to genre. First select a random sample of 100 sentences from the 'news' genre. Use this to estimate the mean sentence length of the news genre in general.

b) Repeat with a different sample of 100 sentences form the same genre. How different are the two estimations? Are they compatible?

c) Then select a sample from the same genre of 900 sentences. What is the estimate this time?

d) Select a sample of 100 sentences from the 'fiction' category. Use this to estimate the average sentence length in fiction.

e) Could this sample come from the 'news' genre, in other words are the two genres different with respect to sentence length?

Exercise 2 – Binomial vs. normal distribution

We have claimed that the normal distribution is a good approximation to the binomial distribution when *n* is sufficiently large. But how large is that? Some books say we may use the normal distribution when both $p\sqrt{n}$ and $(1-p)\sqrt{n}$ are ≥ 5. Other books say that we may use the approximation when both these numbers are $\geq$10. We will inspect how good the approximation is for the first value, 5. To do this we will consider the binomial distribution for n=10 and p=0.5.

a) Which normal distribution corresponds to this, i.e. what is the μ and σ?

b) To compare the two, we may compare the probability mass function (pmf) for the binomial to the probability density function (pdf) for the normal distribution. Do this for j = 0, …, 10, i.e. calculate b(j; 10, 0.5) and compare to the corresponding value for the normal distribution.

c) In statistical tests, one uses the cumulative distribution function (cdf). Calculate the cdf for the two distributions for the same values as in (b). Do you register anything peculiar?

Exercise 3 – Evaluation and matched pairs

We will use this exercise to learn something more on hypothesis testing in general and application to evaluation in particular. The section 7.3 in ch. 7 of Jurafsky and Martin, 3.ed. is rather terse and we will try to get a better understanding.

a) Say, you are constructing a parser. You are comparing your parser to a reference parser. While the reference parser parses 90% of all sentences correctly, your parser parses 91.25% of all sentences correctly. You are very pleased with the results, but you remember something about chance and significance and want to make sure that this result is not a result of pure luck.

You established the numbers as follows. You parsed 1600 sentences from a reference corpus with the reference parser, parser A, and 1440 of them were correctly parsed. Then you parsed 1600 sentences from the reference corpus with your own parser, parser B, and 1460 of them were parsed correctly.

To compare the two, we will first use the two-sided t-test. What is the p-value you get when you compare the two with the two-sample t-test?

$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_A^2}{N} + \frac{s_B^2}{N}}}$$

Would you conclude that parser B is better than parser A?

b) Luckily, you have some more information. You did not test the two parsers on different test sets but on the same set and you kept records for each sentence. Some sentences seem hard for both parsers and the joint results may be summarized as follows.

- There are 1415 sentences where both parsers succeeded.
- There are 25 sentences where A succeeded and B failed.
- There are 45 sentences where B succeeded and A failed.
- There are 115 sentences where both parsers failed.

We may define a new random variable $X_C = X_B - X_A$. The null hypothesis is that $\mu_C = 0$. We may perform a T-test of the null hypothesis. What is the result of the test? Can you now conclude that parser B is better than parser A?

c) We will then compare to the method from J&M, 3.ed., sec. 7.3.