

# Statistical Inference

Statistical Algorithms in Language Technology

Joakim Nivre

## Introduction

- ▶ Topics in this lecture:
  1. Sampling
  2. Estimation
  3. Hypothesis testing
- ▶ Basic elements of statistical learning

## Sampling

- ▶ Let  $X$  be a stochastic variable.
  1. A vector  $(X_1, \dots, X_n)$  of independent variables  $X_i$  with the same distribution as  $X$  is said to be a random sample of  $X$ .
  2. A value vector  $(x_1, \dots, x_n)$  such that  $X_1 = x_1, \dots, X_n = x_n$  in a particular experiment is called a statistical material.
- ▶ Example:
  - ▶ Consider a corpus  $C$  consisting of words  $(w_1, \dots, w_n)$ .
  - ▶ Can we regard  $C$  as a statistical material resulting from a sample  $(W_1, \dots, W_n)$  of the word variable  $W$ ?
  - ▶ Why (not)?

## Sample Variables

- ▶ Given a random sample of a variable  $X$ , we can define new stochastic variables that are functions of the sample, called sample variables:
  1. The sample mean:  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
  2. The sample variance:  $s_n^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- ▶ These variables are called sample variables to distinguish them from the expectation and (true) variance of  $X$ , which are called population variables or model parameters.

## Statistical Inference

- ▶ Statistical inference is the science of making predictions or inferences from finite sets of observations (samples) to (potentially infinite) sets of new observations (populations).
- ▶ Two main kinds of statistical inference:
  1. Estimation: Use samples and sample variables to predict population variables.
  2. Hypothesis testing: Use samples and sample variables to test hypotheses about populations and population variables.

## Estimation

- ▶ Two kinds of estimation:
  1. Point estimation: Use sample variable  $f(X_1, \dots, X_n)$  to estimate parameter  $\phi$ .
  2. Interval estimation: Use sample variables  $f_1(X_1, \dots, X_n)$  and  $f_2(X_1, \dots, X_n)$  to construct an interval such that  $P(f_1(X_1, \dots, X_n) < \phi < f_2(X_1, \dots, X_n)) = p$ , where  $p$  is the confidence level adopted.

## Maximum Likelihood Estimation (MLE)

- ▶ Given a statistical material  $x_1, \dots, x_n$  and a set of parameters  $\theta$ , the likelihood function  $L$  is:

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n P_{\theta}(x_i)$$

where  $P_{\theta}(x_i)$  is the probability that the variable  $X_i$  assumes the value  $x_i$  given a set of values for the parameters in  $\theta$ .

- ▶ Maximum likelihood estimation means choosing  $\theta$  so that the likelihood function is maximized:

$$\max_{\theta} L(x_1, \dots, x_n, \theta)$$

## MLE: Example 1

- ▶ Given a random sample  $(X_1, \dots, X_n)$  of a numerical variable  $X$ , the sample mean  $\bar{X}_n$  is a maximum likelihood estimate of the expectation  $E[X]$ .
- ▶ The average sentence length  $X$  in a certain type of text can be estimated with the mean sentence length in a representative sample:

$$\hat{E}[X] = \bar{X}_n$$

## MLE: Example 2

- ▶ Given a random sample  $(X_1, \dots, X_n)$  of a categorical variable  $X$ , the relative frequency of the value  $x$ ,  $f_n(x)$ , is a maximum likelihood estimate of the probability  $P(X = x)$ .
- ▶ The probability of an arbitrary word from a text being a noun can be estimated with the relative frequency of nouns in a suitable corpus of texts:

$$\hat{P}(\text{noun}) = f_n(\text{noun})$$

## The Rationale of MLE

- ▶ We want to choose the most probable model given the data:

$$P(\theta|x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n|\theta)P(\theta)}{P(x_1, \dots, x_n)}$$

$$\arg \max_{\theta} P(\theta|x_1, \dots, x_n) = \arg \max_{\theta} P(x_1, \dots, x_n|\theta)P(\theta)$$

- ▶ If we assume a uniform distribution for  $P(\theta)$ , then

$$\arg \max_{\theta} P(\theta|x_1, \dots, x_n) = \arg \max_{\theta} P(x_1, \dots, x_n|\theta)$$

- ▶ The status of  $P(\theta)$  is controversial in statistical theory (Bayesians vs. Frequentists)

## Motivating Example (MLE)

- ▶ Assume two sets of two dice:
  1. Two ordinary dice
  2. One die with 5 on all sides; one with 2 on 2 sides, 6 on 4 sides
- ▶ Sums from 5 throws:  $\langle 7, 11, 11, 11, 7 \rangle$  (data)
- ▶ Which set of dice was used (model)?
- ▶ Likelihood values:
  1.  $P(\text{data}|\text{set 1}) = 1/209952 \approx 0.000005$
  2.  $P(\text{data}|\text{set 2}) = 8/243 \approx 0.033$

## Practical Considerations (MLE)

- ▶ MLE is a good solution to the estimation problem if the statistical material is large enough. In practice, MLE is often suboptimal because of sparse data.
- ▶ Another problem is caused by the fact that the data sets used in practical applications, such as language corpora, seldom satisfy the conditions for being a true random sample.
- ▶ Practical solutions to the estimation problem often use the MLE as a starting point, applying more or less sophisticated smoothing techniques in an attempt to improve the quality of the estimate.

## Interval Estimation

- ▶ In general, we can derive a 95% confidence interval for our maximum likelihood estimate  $\hat{\phi}$  of a mean as follows:

$$\hat{\phi} \pm 1.96 \frac{s}{\sqrt{n}}$$

where  $s$  is the standard deviation of the underlying variable and  $n$  is the number of observations in our statistical material.

- ▶ Examples:

1. Sentence length:  $\hat{E}[X] = \bar{X}_n \pm 1.96 \frac{s}{\sqrt{n}}$
2. Noun probability:  $\hat{P}(\textit{noun}) = f_n(\textit{noun}) \pm 1.96 \frac{s}{\sqrt{n}}$

## More on Interval Estimation

- ▶ The previous formulas for 95% confidence intervals are based on the assumption that the estimated parameter has a normal distribution, an assumption that is theoretically provable in the limit but which is met to varying degrees in the samples used in practice.
- ▶ The formulas also presuppose that the true standard deviance (variance) of the variable is known. In practice, this parameter has to be estimated from the sample as well.

## Hypothesis Testing

1. Choose a test statistic  $t$  whose distribution is known when the null hypothesis is true.
2. Use  $t$  to calculate the probability  $p$  of observing the data given that the null hypothesis is true.
3. If  $p < \alpha$ , reject the null hypothesis, where  $\alpha$  is the significance level adopted.

## Example: Hypothesis Testing

- ▶ In a random sample of 1000 words from written texts, there were 154 nouns. In a corresponding sample from transcribed spoken dialogues, there were only 128 nouns. Are nouns more frequent in written language?
- ▶ If there is no difference, we should expect  $(154 - 128) / 2 = 141$  nouns in a 1000 word sample (the null hypothesis).
- ▶  $\chi^2$  test:
  1.  $\chi^2 = \sum \frac{(O-E)^2}{E} = 2.79$
  2.  $P(\chi^2 = 2.79)$  with 1 df  $\approx 0.1$

# Summary

- ▶ Samples and sample variables
- ▶ Estimation
  1. Point estimation (MLE)
  2. Interval estimation
- ▶ Hypothesis testing: Probability of data given null hypothesis