

The Linguist's Guide to Statistics

DON'T PANIC

Brigitte Krenn

Universität des Saarlandes
FR 8.7, Computerlinguistik
Postfach 1150
D-66041 Saarbrücken
Germany
krenn@coli.uni-sb.de

Christer Samuelsson

Bell Laboratories
600 Mountain Ave
Room 2D-339
Murray Hill, NJ 07974
USA
christer@research.bell-labs.com

<http://www.coli.uni-sb.de/{~krenn,~christer}>

December 19, 1997

© Brigitte Krenn and Christer Samuelsson, 1994, 1995, 1996, 1997.

In a review of Eugene Charniak's book "Statistical Language Learning" in Computational Linguistics Vol. 21, No. 1 of March, 1995, David Magerman writes:

The \$64,000 question in computational linguistics these days is: What should I read to learn about statistical natural language processing? I have been asked this question over and over, and each time I have given basically the same reply: there is no text that addresses this topic directly, and the best one can do is find a good probability-theory textbook and a good information-theory textbook, and supplement those texts with an assortment of conference papers and journal articles.

...

The overriding concern should be to learn (and teach) the mathematical underpinnings of the statistical techniques used in this field. The field of statistical NLP is very young, and the foundations are still being laid. Deep knowledge of the basic machinery is far more valuable than the details of the most recent unproven ideas.

So let's get down to it!

Contents

1	Basic Statistics	1
1.1	The Stability of the Relative Frequency	1
1.2	Elementary Probability Theory	1
1.2.1	Sample Space	1
1.2.2	Probability Measures	2
1.2.3	Independence	3
1.2.4	Conditional Probabilities	3
1.2.5	Bayesian Inversion	4
1.2.6	Partitions	5
1.2.7	Combinatorics	6
1.3	Stochastic Variables	8
1.3.1	Distribution Function	8
1.3.2	Discrete and Continuous Stochastic Variables	9
1.3.3	Frequency Function	10
1.3.4	Expectation Value	12
1.3.5	Variance	13
1.3.6	Moments and Moment-Generating Functions	14
1.4	Two-dimensional Stochastic Variables	14
1.4.1	Distribution Function	14
1.4.2	Frequency Function	15
1.4.3	Independence	16
1.4.4	Functions of Stochastic Variables	16
1.4.5	Higher Dimensions	18
1.5	Selected Probability Distributions	18
1.5.1	Binomial Distribution	18
1.5.2	Normal Distribution	20
1.5.3	Other Distributions	23
1.5.4	Distribution Tables	24
1.6	Some Theoretical Results	25
1.7	Estimation	27
1.7.1	Random Samples	27
1.7.2	Estimators	28
1.7.3	Maximum-Likelihood Estimators	29
1.7.4	Sufficient Statistic	31
1.7.5	Confidence Intervals	32
1.7.6	Hypothesis Testing and Significance	35
1.8	Further Reading	37

2	Applied Statistics	39
2.1	Markov Models	39
2.1.1	Stochastic Processes	39
2.1.2	Markov Chains and the Markov Property	40
2.1.3	Markov Models	42
2.1.4	Hidden Markov Models	43
2.1.5	Calculating $P(\mathbf{O})$	44
2.1.6	Finding the Optimal State Sequence	47
2.1.7	Parameter Estimation for HMMs	48
2.1.8	Further reading	50
2.2	Elementary Information Theory	50
2.2.1	Entropy	50
2.2.2	Related Information Measures	52
2.2.3	Noiseless Coding	55
2.2.4	More on Information Theory	59
2.3	Multivariate Analysis	59
2.4	Sparse Data	59
3	Basic Corpus Linguistics	61
3.1	Empirical Evaluation	61
3.1.1	Contingency Tables	61
3.1.2	Important Measures on Contingency Tables	62
3.1.3	Extended Contingency Table Model	63
3.1.4	Measures, Extended	63
3.1.5	Loose Attempts on Measuring System Efficiency	64
3.1.6	Empirical Evaluation of Part-of-Speech Taggers: A Case Study	64
3.2	Corpora	65
3.2.1	Types of Corpora	66
3.2.2	Test Suites versus Corpora	67
3.2.3	Tokenization	68
3.2.4	Training and Testing	69
3.2.5	Tagsets	69
4	Stochastic Grammars	73
4.1	Some Formal Language Theory	73
4.1.1	Context-free Grammars	73
4.1.2	Derivations	74
4.1.3	Trees	74
4.1.4	Parse Trees	75
4.2	Stochastic Context-free Grammars	76
4.3	A Parser for SCFGs	78
4.4	Parameter Estimation for SCFGs	80
4.4.1	The Inside and Outside Variables	81
4.4.2	Deriving the Reestimation Equations	82
4.5	Adding Probabilistic Context to SCFGs	84
4.6	Theoretical Probability Losses	86
4.7	Stochastic Tree-Substitution Grammars	88
4.8	Stochastic History-Based Grammars	89
4.9	Lexicalization of Stochastic Grammars	90
4.9.1	Stochastic Dependency Grammar and Related Approaches	90
4.10	Probabilistic LR Parsing	92
4.10.1	Basic LR Parsing	92

4.10.2	LR-Parsed Example	93
4.10.3	LR-Table Compilation	95
4.10.4	Generalized LR Parsing	96
4.10.5	Adding Probabilities	97
4.10.6	Probabilistic GLR Parsing	98
4.11	Scoring	99
5	Selected Topics in Statistical NLP	101
5.1	Statistical Part-of-Speech Tagging	101
5.1.1	In short	101
5.1.2	Linguistic Background	101
5.1.3	Basic Statistical PoS tagging	101
5.1.4	Suggested Reading	103
5.2	Statistical Machine Translation	103
5.2.1	In short	103
5.2.2	Suggested Reading	104
5.3	Statistical Language Learning	104
5.4	Structural Ambiguity and Semantic Classes	104
5.4.1	Linguistic Background	104
5.4.2	Association Models	105
5.4.3	Suggested Reading	108
5.5	Word Sense Disambiguation	108
5.5.1	Phenomena	108
5.5.2	Parameter Estimation	108
5.5.3	Disambiguation Model	108
5.5.4	Suggested Reading	109
5.6	Lexical Knowledge Acquisition	109
A	Desiderata	111
B	Tools	113
B.1	Simple Unix Commands	113
B.2	Split up a text using tr	114
B.3	Sort word list: sort, uniq	114
B.4	Merge counts for upper and lower case: tr, sort, uniq	115
B.5	Count lines, words, characters: wc	115
B.6	Display the first n lines of a file: sed	115
B.7	Find lines: grep, egrep	116
B.8	n-grams: tail, paste	116
B.9	Manipulation of lines and fields: awk	116
C	Some Calculus	119
C.1	Numbers	119
C.1.1	Natural Numbers	119
C.1.2	Integers	119
C.1.3	Rational Numbers	119
C.1.4	Real Numbers	120
C.1.5	Complex Numbers	121
C.1.6	Algebraic and Transcendental Numbers	121
C.2	The Very Basics	122
C.2.1	Exponentials	122
C.2.2	Roots	123
C.2.3	The Exponential Function	124
C.2.4	Logarithms for Beginners	125

C.2.5	Factorial	127
C.2.6	Sequences	128
C.2.7	Series	129
C.3	On the Number e , the Exponential Function and the Natural Logarithm	132
D	Tagsets	133
D.1	Word Level Tagsets	133
D.1.1	Representation of Word Level Tags	133
D.1.2	Mapping between Linguistic Descriptions and Tagsets	134
D.1.3	Tagsets and the Representation of Ambiguity	134
D.1.4	Minimal Criteria for the Development of PoS-Tagsets	134
D.2	Tagsets for English Text Corpora	135
D.2.1	The Susanne Tagset	135
D.2.2	The Penn Treebank	141
D.2.3	The Constraint Grammar Tagset	146
D.3	Tagsets for German Text Corpora	153
D.3.1	The Stuttgart-Tübingen Tagset	153
E	Optimization Theory, Wild-West Style	157
E.1	Introduction	157
E.2	Constrained Optimization in R^n	158
E.3	Numerical Analysis	160

Preface

This compendium is the result of the authors' attempts at teaching courses on statistical approaches in Computational Linguistics and Natural Language Processing, and it is continuously evolving and undergoing revision. Although we have already put considerable effort into writing, correcting and updating this compendium, there are numerous errors and omissions in it that we hope to deal with in the nearest future. The most recent release of this compendium, in the form of a uuencoded gzipped PostScript file, `stat_cl.ps.gz.uu`, or as an ordinary PostScript file, `stat_cl.ps`, can be retrieved from one of the authors' WWW homepage at

`http://www.coli.uni-sb.de/{~christer,~krenn}`

The ambition is to cover most statistical, stochastic and probabilistic approaches in the field. As will be obvious by inspecting the table of contents, this is by no means yet the case.

The first three chapters have a distinct text-book character: Chapter 1 provides the necessary prerequisites in Probability Theory and Statistics, Chapter 2 describes some statistical models that are much in use in the field, and Chapter 3 constitutes an introduction to Corpus Linguistics. Chapters 4 and 5 discuss how statistical models and techniques are applied to various task in Computational Linguistics and Natural Language Processing, relating the material presented in the previous chapters to relevant scientific articles.

Feel free to use this compendium, but please do acknowledge the source. Also, any comments or suggestions to improvements are more than welcome, and are most likely to enhance future versions of the compendium. We are already greatly indebted for this to Rens Bod, Bob Carpenter, John Carroll, Ted Dunning, Jussi Karlgren, Kimmo Koskenniemi, David Magerman, David Milward, Joakim Nivre, Khalil Sima'an, Atro Voutilainen and numerous students at the University of the Saarland, Uppsala University, the University Helsinki and to course participants at the ESSLLI-97 summer school in Aix-en-Provence. Special credit is due to Thorsten Brants, who wrote the first version of the section on Hidden Markov Models, and to our online mathematician Åke H. Samuelsson. Parts of the compendium are used in a web-based introductory course on statistical natural language processing which is set up by Joakim Nivre at Göteborg University. It can be accessed via `http://www.ling.gu.se/~nivre/kurser/wwwstat/`.

Saarbrücken

New York

December 1997

Brigitte Krenn

Christer Samuelsson

Chapter 1

Basic Statistics

1.1 The Stability of the Relative Frequency

Although an elegant theory in its own right, the main reason that statistics has grown as popular as it has is something that is known as “The stability of the relative frequency”. This is the empirical observation that there is some structure also in random processes. If we for example flip a coin a large number of times, we will note that in approximately half the cases the outcome is “heads” and in approximately half the cases it is “tails”. If we flip a coin a small number of times, say only once, twice or three times, this is not consistently the case.

The proportion of times a certain outcome occurs is called the *relative frequency* of the outcome. If n_u is the number of times the outcome u occurs in n trials, then $\frac{n_u}{n}$ is the relative frequency of u . The relative frequency is often denoted f_n^u .

Empirically, there seems to be some number which the relative frequency stabilizes around after a large number of trials. A fundamental assumption in statistics is that such numbers exist. These numbers are called *probabilities*.

1.2 Elementary Probability Theory

Introductory presentations of statistics often disguise probability theory in set theory, and this is no exception.

1.2.1 Sample Space

The *sample space* is a set of *elementary outcomes*. An *event* is a subset of the sample space. Sample spaces are often denoted Ω , and events are often called A , B , C , etc. Let’s get this grounded with an example:

Example: For a normal die¹, the six elementary outcomes are One, Two, Three, Four, Five and Six, and thus the sample space Ω is the set $\{\text{One, Two, Three, Four, Five, Six}\}$. The events are various subsets of this set, e.g., “the outcome is One”, $\{\text{One}\}$; “the outcome is less than four”, $\{\text{One, Two, Three}\}$; the outcome is odd, $\{\text{One, Three, Five}\}$; etc. In fact, there are $2^6 = 64$

¹ “Die” is the singular form of “dice”, a sort of mechanical random generators used in games like Craps, Monopoly and Fia-med-knuff.

different subsets of Ω , i.e., there are 64 distinct events in Ω , including the empty set \emptyset and Ω itself, see Section 1.2.7.

Statisticians are a kind of mathematicians, and like to distinguish between for example the outcome One, which is a basic element, and the event $\{\text{One}\}$, which is a set consisting of one element. We will try to keep this up for a while.

1.2.2 Probability Measures

A *probability measure* P is a function from events in the sample space Ω , i.e., from the set of subsets of Ω ,² to the set of real numbers in $[0, 1]$ that has the following properties:

- 1) $0 \leq P(A) \leq 1$ for each event $A \subseteq \Omega$
- 2) $P(\Omega) = 1$
- 3) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Intuitively, the total mass of 1 is distributed throughout the set Ω by the function P . This will assign some particular mass to each subset A of Ω . This mass is the probability of event A , denoted $P(A)$. $A \cap B = \emptyset$ means that A and B are disjoint, i.e., that they have no common element.

Example: For a fair (unbiased) die, where as we recall the sample space is the set $\{\text{One}, \text{Two}, \text{Three}, \text{Four}, \text{Five}, \text{Six}\}$, the mass 1 is evenly and justly divided among the six different singleton sets. Thus $P(\{\text{One}\}) = P(\{\text{Two}\}) = P(\{\text{Three}\}) = P(\{\text{Four}\}) = P(\{\text{Five}\}) = P(\{\text{Six}\}) = \frac{1}{6}$. If A is the event of the outcome being divisible by two, i.e., the subset $\{\text{Two}, \text{Four}, \text{Six}\}$, and if B is the event of the outcome being divisible by three, i.e., the subset $\{\text{Three}, \text{Six}\}$, then $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{3}$.

A loaded die, used by cheaters, would not have the probability mass as evenly distributed, and could for example assign $P(\{\text{Six}\})$ a substantially larger value than $\frac{1}{6}$.

Some immediate corollaries that are worth remembering fall out from the definition of a probability measure:

- a) $P(B \setminus A) = P(B) - P(A \cap B)$
- b) $A \subseteq B \Rightarrow P(A) \leq P(B)$
- c) $P(\bar{A}) = 1 - P(A)$
- d) $P(\emptyset) = 0$
- e) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$B \setminus A$ denotes the difference set B minus A , i.e., the set of elements in B that are not members of A . \bar{A} denotes the complement of A , i.e., $\Omega \setminus A$.

Proofs:

$$\begin{aligned} \text{a) } B &= (B \setminus A) \cup (A \cap B) ; (B \setminus A) \cap (A \cap B) = \emptyset \Rightarrow \\ P(B) &= P((B \setminus A) \cup (A \cap B)) = P(B \setminus A) + P(A \cap B) \end{aligned}$$

²This is called the *power set* of Ω and is denoted 2^Ω .

neously can be established directly from the individual probabilities of A and B .

Example: To continue the example of the fair die, with the events A of the outcome being divisible by two and B of the outcome being divisible by three, we note that $P(A \cap B) = P(\{\text{Six}\}) = \frac{1}{6}$, and that $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. Thus A and B are independent.

Let C be the event of the outcome being divisible by four and note that $P(C) = P(\{\text{Four}\}) = \frac{1}{6}$. Intuitively, the property of being divisible by four is related to the property of being divisible by two, and if we calculate on the one hand $P(A \cap C) = P(\{\text{Four}\}) = \frac{1}{6}$, and on the other hand $P(A) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$, we see that A and C are not independent.

1.2.4 Conditional Probabilities

$P(A | B)$ is a so-called *conditional probability*, namely the probability of event A given that event B has occurred, and is defined as

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (1.2)$$

This is the updated probability of A once we have learned that B has occurred. $P(A)$ is often called the *prior probability* of A since we have no

³ “Iff” means if and only if.