Many things many linguists should know about the creation, evaluation, and use of corpora*

* But sometimes don’t bother to ask

Felix Bildhauer and Roland Schäfer

In memory of Adam Kilgarriff

Contents

Preface

Acknowledgments

Abbreviations and symbols

Abbreviations

ANOVA	analysis of variance
CDF	cumulative distribution function
CLT	central limit theorem
cp.	ceteris paribus (all other things being equal)
iid.	independent and identically distributed
LM	linear model
LMM	linear mixed model
GLM	linear mixed model
GLMM	generalised linear mixed model
PDF	probability density function
VCOV	variance-covariance matrix

Symbols

Symbols are overloaded ad-hoc to denote either a (possibly indexed) value such as $s_x = 1$ (for "the population mean of variable x is 1") or a function such as $s(x) = 1$ where applicable.

	Mathematical symbols
$x \sim D$	x *follows* D (x a variable, D a distribution)
$\bar{x}$	sample arithmetic mean of x
$\tilde{x}$	sample median of x
$\hat{x}$	predicted value of x

	Letter-like symbols
α	alpha level
α_i	intercept i
β	beta level
β_i	first-level coefficient i

df	degrees of freedom
e	Euler constant
ϵ	observation-level error
f	frequency
F	F statistic (see ANOVA)
γ_i	second-level coefficient i
H	Kruskal-Wallis statistic
H_0	null hypothesis
H_A	alternative hypothesis
M_M	main hypothesis
IQR	inter-quartile range
$\mathscr{L}$	Likelihood
μ	population mean
μ_i	mean of modeled effect i
n	sample size
N	population size
O	Odds
p	proportion
P_i	the i-th percentile
Pr	probability
φ	dispersion parameter
Q_i	i-th quartile
r	sample covariance coefficient
r^2	coefficient of determination
R^2	multifactorial coefficient of determination
ρ	population covariance coefficient
s	sample standard deviation of x
s^2	sample variance of x
SE	standard error
SS	sum of squares
σ	population standard deviation
σ^2	population variance
U	Mann-Whitney statistic
χ^2	chi square statistic

Random distributions are denoted by bold-printed abbreviated names instead of the incoherent symbols sometimes used.

Bern	Bernoulli distribution
Exp	exponential distribution
F	F distribution
Norm	normal (Gaussian) distribution
t	t distribution
Unif	uniform distribution
Chisq	χ^2 distribution

1 Sampling and corpus composition

2 Linguistic annotation

3 Web as corpus

4 Classifying documents

5 Practical topic modeling

6 Corpus comparison

7 Designing corpus studies and making queries

8 Collo-phenomena

9 Corpora and statistcal inference

10 Advanced statistical modeling for corpus studies

Senn (2011)

References

Senn, Stepen J. 2011. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets, and Morals* 2. 48–66.