

UNIVERSITY OF LIÈGE

Université
de Liège



Statistics – Theory

PROF. DR. DR. K. VAN STEEN

July 2012

Contents

1	Introduction	2
1.1	The Birth of Probability and Statistics	2
1.2	Statistical Modeling under Uncertainties: From Data to Knowledge	3
1.3	Course Outline	4
1.4	Motivating Examples	5
2	Samples, random sampling and sample geometry	8
2.1	Introduction: No statistics without Data!	8
2.2	Populations and samples	9
2.3	Sampling schemes	12
2.3.1	Non-probability sampling	12
2.3.2	Probability sampling	13
2.4	Sampling Challenges	14
2.5	Distribution of a sample	15
2.6	Statistics	15
2.7	Sample Geometry	18
2.7.1	The geometry of the sample	18
2.7.1.1	The mean	18
2.7.1.2	Variance and correlation	19
2.7.2	Expected values of the sample mean and the covariance matrix	20
2.7.3	Generalized variance	21
2.8	Resampling	27
2.9	The Importance of Study Design	28
3	Exploratory Data Analysis	30
3.1	Typical data format and the types of EDA	30
3.2	Univariate non-graphical EDA	31
3.2.1	Categorical data	32
3.2.2	Characteristics of quantitative data	32
3.2.3	Central tendency	33
3.2.4	Spread	35
3.2.5	Skewness and kurtosis	37
3.3	Univariate graphical EDA	38
3.3.1	Histograms	38
3.3.2	Stem-and-leaf plots	44
3.3.3	Boxplots	44
3.3.4	Quantile-normal plots	47

3.4	Multivariate non-graphical EDA	50
3.4.1	Cross-tabulation	51
3.4.2	Correlation for categorical data	52
3.4.3	Univariate statistics by category	53
3.4.4	Correlation and covariance	53
3.4.5	Covariance and correlation matrices	55
3.5	Multivariate graphical EDA	56
3.5.1	Univariate graphs by category	56
3.5.2	Scatterplots	56
3.6	A note on degrees of freedom	58
4	Estimation	59
4.1	Introduction	59
4.2	Statistical philosophies	60
4.3	The frequentist approach to estimation	62
4.4	Estimation by the method of moments	63
4.4.1	Traditional methods of moments	64
4.4.2	Generalized methods of moments	67
4.5	Properties of an estimator	67
4.5.1	Unbiasedness	67
4.5.2	Trading off Bias and Variance	67
4.5.2.1	Mean-Squared Error	67
4.5.2.2	Minimum-Variance Unbiased	69
4.5.3	Efficiency	73
4.5.4	Consistency	73
4.5.5	Loss and Risk Functions	74
4.6	Sufficiency	75
4.7	The Likelihood approach	77
4.7.1	Maximum likelihood estimation	77
4.7.2	Properties of MLE	79
4.7.3	The Invariance principle	80
4.8	Properties of Sample Mean and Sample Variance	90
4.9	Multi-parameter Estimation	92
4.10	Newton-Raphson optimization	95
4.10.1	One-paramter scenario	95
4.10.2	Two-paramter scenario	96
4.10.3	Initial values	96
4.10.4	Fisher's method of scoring	97
4.10.5	The method of profiling	97
4.10.6	Reparameterization	98
4.10.7	The step-halving scheme	99
4.11	Bayesian estimation	99
4.11.1	Bayes' theorem for random variables	99
4.11.2	Post 'is' prior \times likelihood	100
5	Confidence intervals	103
5.1	Introduction	103

5.2	Exact confidence intervals	106
5.3	Pivotal quantities for use with normal data	110
5.4	Approximate confidence intervals	114
5.5	Bootstrap confidence intervals	115
5.5.1	The empirical cumulative distribution function	115
6	The Theory of hypothesis testing	120
6.1	Introduction	120
6.2	Terminology and notation	122
6.2.1	Hypotheses	122
6.2.2	Tests of hypotheses	122
6.2.3	Size and power of tests	123
6.3	Examples	123
6.4	One-sided and two-sided Tests	126
6.4.1	Case (a): Alternative is one-sided	127
6.4.2	Case (b): Two-sided Alternative	128
6.4.3	Two approaches to hypothesis testing	129
6.5	Two-sample problems	131
6.6	Connection between hypothesis testing and CI's	133
6.7	Summary	134
6.8	Non-parametric hypothesis testing	135
6.8.1	Kolmogorov-Smirnov (KS)	136
6.8.2	Asymptotic distribution	137
6.8.3	Bootstrap Hypothesis tests	138
6.9	The general testing problem	140
6.10	Hypothesis testing for normal data	141
6.11	Generally applicable test procedures	146
6.12	The Neyman-Pearson lemma	148
6.13	Goodness of fit tests	151
6.14	The χ^2 test for contingency tables	153
7	Chi-square Distribution	155
7.1	Distribution of S^2	155
7.2	Chi-Square Distribution	157
7.3	Independence of \bar{X} and S^2	162
7.4	Confidence intervals for σ^2	162
7.5	Testing hypotheses about σ^2	164
7.6	χ^2 and $\text{Inv-}\chi^2$ distributions in Bayesian inference	166
7.6.1	Non-informative priors	166
7.7	The posterior distribution of the Normal variance	167
7.7.1	Inverse Chi-squared distribution	168
7.8	Relationship between χ^2_ν and $\text{Inv-}\chi^2_\nu$	168
7.8.1	Gamma and Inverse Gamma	168
7.8.2	Chi-squared and Inverse Chi-squared	168
7.8.3	Simulating Inverse Gamma and $\text{Inv-}\chi^2$ random variables	169
8	Analysis of Count Data	171