□ 应用统计学丛书

Extremes in Random Fields

A Theory and Its Applications

随机域中的极值统计学

理论及应用（英文版）

Benjamin Yakir 著

高等教育出版社
HIGHER EDUCATION PRESS

□ 应用统计学丛书

Extremes in Random Fields

A Theory and Its Applications

随机域中的极值统计学

理论及应用（英文版）

Benjamin Yakir 著

SUIJIYU ZHONG DE JIZHI TONGJIXUE
LILUN JI YINGYONG

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

Author
Benjamin Yakir
Department of Statistics, Mount Scopus
The Hebrew University of Jerusalem
Jerusalem 91905, Israel

图书在版编目（CIP）数据

随机域中的极值统计学：理论及应用 = Extremes in random fields : a theory and its applications : 英文 /（以）亚基尔（Yakir, B.）著. -- 北京：高等教育出版社，2013. 9
ISBN 978-7-04-037817-7

Ⅰ. ①随… Ⅱ. ①亚… Ⅲ. ①统计学 - 研究 - 英文
Ⅳ. ① C8

中国版本图书馆 CIP 数据核字（2013）第 187655 号

策划编辑 王丽萍　　责任编辑 李华英　　封面设计 姜 磊　　责任印制 田 甜

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街4号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	北京铭成印刷有限公司	网上订购	http://www.landraco.com
开　　本	787 mm × 1092 mm　1/16		http://www.landraco.com.cn
印　　张	15.25	版　　次	2013 年 9 月第 1 版
字　　数	320 千字	印　　次	2013 年 9 月第 1 次印刷
购书热线	010-58581118	定　　价	59.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物 料 号　37817-00

Extremes in Random Fields

应用统计学丛书

书号	书名	著译者
9787040378177	随机域中的极值统计学：理论及应用（英文版）	Benjamin Yakir 著
9787040322927	金融工程中的蒙特卡罗方法	Paul Glasserman 著 范韶华、孙武军 译
9787040348309	大维统计分析	白志东、郑术蓉、姜丹丹
9787040348286	结构方程模型：Mplus与应用（英文版）	王济川、王小倩 著
9787040348262	生存分析：模型与应用(英文版)	刘宪
9787040345407	MINITAB软件入门：最易学实用的统计分析教程	吴令云 等 编著
9787040321883	结构方程模型：方法与应用	王济川、王小倩、姜宝法 著
9787040319682	结构方程模型：贝叶斯方法	李锡钦 著 蔡敬衡、潘俊豪、周影辉 译
9787040315370	随机环境中的马尔可夫过程	胡迪鹤 著
9787040256390	统计诊断	韦博成、林金官、解锋昌 编著
9787040250626	R语言与统计分析	汤银才 主编
9787040247510	属性数据分析引论（第二版）	Alan Agresti 著 张淑梅、王睿、曾莉 译
9787040182934	金融市场中的统计模型和方法	黎子良、邢海鹏 著 姚沛沛 译

网上购书：academic.hep.com.cn,
www.china-pub.com, 卓越, 当当
其他订购办法：
各使用单位可向高等教育出版社读者服务部汇款订购。书款通过邮局汇款或银行转账均可。购书免邮费，发票随后寄出。
单位地址：北京西城区德外大街4号
电　　话：010-58581118/7/6/5/4
传　　真：010-58581113

通过邮局汇款：
地　　址：北京西城区德外大街4号
户　　名：高等教育出版社销售部综合业务部
通过银行转账：
户　　名：高等教育出版社有限公司
开 户 行：交通银行北京马甸支行
银行账号：110060437018010037603

To David

Preface

This text started as class notes for a course that I gave in the Mathematical Sciences Center (MSC) in Tsinghua University, Beijing, that got overblown and became a book. I was enjoying a sabbatical leave in the Department of Statistics and Applied Probability (DSAP) of the National University of Singapore when I was given an offer to teach a summer course in China. Of course I accepted. How could I resist the opportunity to fulfil a childhood dream of visiting China?

After accepting the proposal I had to decide what to teach. I decided to fulfil yet another dream, the dream of summarizing and unifying a subject I was writing about all my career, even before I knew what the subject was. The subject is the distribution of extremes in random fields and the analysis of statistical problems that can be formulated in relation to such extremes. Immediately after obtaining my PhD, and as a continuation of my PhD thesis, I was interested in the investigation of the average run length of the Shiryaev–Roberts change-point detection rule. Therefore, I found it natural to try to address a challenge that was presented to me by David Siegmund during a barbecue meal that he prepared for me in his yard. The challenge was to develop a simpler method for analyzing this average run length. In an attempt to attack this problem I began experimenting with the likelihood ratio identity, one of David's favorite techniques, and followed the road that eventually led me to writing this book.

The original problem was the investigation of average run length in a sequential change-point detection problem.[1] However, the basic technique that was developed turned out to be useful for the investigation of a relatively wide array of different statistical problems that involve the distribution of maxima.[2] Among other things, David and I used the method in order to investigate the significance level of sequence alignment, for the computation of the false detection rate in

[1] Yakir B., Pollak M. A new representation for the renewal-theoretic constant appearing in asymptotic approximations of large deviations. Ann. Appl. Probab. **8**, 749-774 (1998).

[2] Grossman S., Yakir B. Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignment. Bernoulli **5**, 829–845 (2004).

Seigmund D.O., Yakir B. Approximate p-values for local sequence alignments. Ann. Statist. **28**, 657–680 (2000).

Seigmund D.O., Yakir B. Statistical analysis of direct identity by descent mapping. Ann. Hum. Genet. **67**, 464–470 (2003).

Seigmund D.O., Yakir B. Correction note: Approximate p-values for local sequence alignments. Ann. Statist. **31**, 1027–1031 (2003).

scanning statistic, for producing more efficient ways of simulation, etc. Each application required this modification or that trick in order to apply the basic principle. However, after 20 years of repeating the same argument even I was able to identify the pattern. The thrust of this book is a description of the pattern and the demonstration of its usefulness in the analysis of nontrivial statistical problems.

The basic argument relies on a likelihood ratio identity that uses a sum of likelihood ratios. This identity translates the original problem that involves the approximation of a vanishingly small probability to a problem that calls for the summation of approximations of expectations. The expectations are with respect to alternative distributions in which the event in question is much more likely to occur. Moreover, by carefully selecting the alternative distributions one may separate the leading term in the probability from the expectations that form the sum, enabling the investigation to concentrate on finer effects.

The method is useful since it does not rely on the ordering of the parameter set and it does not require the normal distribution. In many applications, some of them are presented in the book, a natural formulation of the model calls for the use of collections of random variables that are parameterized not by subsets of the real line. Frequently, the normal assumption may fit the limit in a central limit formulation but may not fit as a description of the extreme tail. In all such cases an alternative to the methods that are usually advocated in the literature are required. The method we present is such an alternative which we felt others may benefit from by knowing about.

This is why we wrote the book. But who is the target audience? This is a tough call. Even if I may state otherwise, the book requires a relatively advanced knowledge in probability as background, perhaps at the level of Durrett's book.[3] Prior knowledge in statistics is an advantage. Indeed, there is an appendix that lists theorems and results and can be used as reference for the statements that are made in the book. Still, I guess that this book is not an easy read even for experts, and much less so for students.

With this warning in mind, I hope that the effort that is required in reading the book will be rewarding. Definitely, for an expert who wants to add yet another method to his toolbox but also for a student who wants to become an expert.

Seigmund D.O., Yakir B. Significance level in interval mapping. In Development of Modern Statistics and Related Topics, Series in Biostatistics, Volume 1. World Scientific Publishing, River Edge, NV, 10–19 (2003).

Shi J., Siegmund D.O., Yakir B. Importance sampling for estimating p-values in linkage analysis. JASA **102**, 929–937 (2007).

Yakir B. On the average run length to false alarm in surveillance problems which possess an invariance structure. Ann. Statist. **26**, 1198–1214 (1998).

Yakir B. Approximation of the p-value in a multipoint linkage analysis using grandparent grandchild pairs and partially informative markers. Nonlinear Anal. **47**, 1973–1984 (2001).

Yakir B. Discussion on "Is average run length to false alarm always an informative criterion?" by Yajun Mei. Sequential Analysis **27**, 406–410 (2008).

[3] Durrett R. Probability: Theory and Examples (2nd Edition). Duxbury Press, Belmont, CA (1995).

For such students, the book can be used as a basis for an advanced seminar. Reading chapters of the book can be used as a primer for a student who is then required to analyze a new problem that was not digested for him/her in the book. This is how I intend to use this book with my students.

The teacher can start such a course by discussing Chapters 1–4 that give the basic background and demonstrate the technique. Chapter 5 is more technical and can be skipped, unless the main interest is in the mathematical details. From the second part of the book it is probably recommended to go over Chapter 6, which is of an intermediate level of difficulty, and then read some or all of Chapters 7–10 depending on the interests of the teacher and the students and on the time constraints.

Acknowledgments

My first acknowledgments are to environments, and especially the people who enabled these environments. The first half of the book was written mainly in DSAP. I know of no better place to do this type of scientific work. I will always be grateful. The second place is MSC. Without them I do not know when, if at all, this book would have been written. Next, I would like to recognize the financial support that I got from the Israel Science Foundation (Grant No. 325/09) and from the US–Israel Binational Science Foundation (Grant No. 2006101). This support was instrumental for the development of the original work that led to the applications that are presented in the second part of the book.

Some of the people that gave me a helping hand I would like to mention by name. Unfortunately, I cannot give the names of the anonymous reviewers who made very useful suggestions on the first draft of the book and helped me improve it. But I can give the name of the editor from Higher Education Press, Liping Wang. Thanks to her this is a book and not just class notes. Also I would like to thank Yuval Nardi, Moshe Pollak, Ton Dieker and Nancy Zhang who coauthored with David Siegmund and myself some of the works that are related directly to the content of the book.

And finally there is David Siegmund. The work presented in this book is basically our joint work. The only reason that we do not share authorship is the fact that I wanted to dedicate this book to him as my modest contribution to the celebration of his career and his accomplishments and as an appreciation for what he gave me. It is not appropriate for a book to be dedicated to one of its authors. So here it is: this is for you, David.

February 2013 Benjamin Yakir, Jerusalem, Israel

Contents

Part I

THEORY

1

Introduction

1.1 Distribution of extremes in random fields

The aim of this book is to present a method for analyzing the tail distribution of extreme values in random fields. A random field can be considered as a collection of random variables $\{X_t : t \in T\}$, indexed by a set of parameters T. The index set T may be quite complex. However, in the applications that we will analyze in this book it will typically turn out that T is a 'nice' subset of $\mathbb{R}^d$, the d-dimensional space of real numbers.

In some statistical applications one is interested in probabilities such as:

$$\mathrm{P}\left(\sup_{t\in T} X_t \geq x\right),$$

the probability that the maximum of the random field exceeds a threshold x, for large values of x. There are only a few special cases in which the problem of computing such probabilities has an exact solution. In all other cases one is forced to use numerical methods, such as simulations, or to apply asymptotic approximations in order to evaluate the probability. This book concentrates on the application of the proposed method for producing asymptotic analytical expansions of the probability. Nonetheless, some elements in the method may, and have been, applied in order to simulate numerical evaluations more efficiently. An application that illustrates the usefulness of the method in the context of simulations is presented in the second part of the book.

As a motivating example consider scanning statistics. Scanning statistics are used in order to detect rare signals in an environment contaminated by random noise. For example, let us assume measurements that are taken in a one-dimensional environment. Each measurement is associated with a point in the environment and the points are equally spaced. For the most part, the expected values of the observations are fixed at some baseline level throughout

Extremes in Random Fields: A Theory and Its Applications, First Edition. Benjamin Yakir.