

SPRINGER BRIEFS IN STATISTICS

George A. F. Seber

Statistical Models for Proportions and Probabilities



Springer

SpringerBriefs in Statistics

For further volumes:
<http://www.springer.com/series/8921>

George A. F. Seber

Statistical Models for Proportions and Probabilities

 Springer

George A. F. Seber
Department of Statistics
The University of Auckland
Auckland
New Zealand

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-642-39040-1 ISBN 978-3-642-39041-8 (eBook)
DOI 10.1007/978-3-642-39041-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013942014

© The Author(s) 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Most elementary statistics books discuss inference for proportions and probabilities, and the primary readership for this monograph is the student of statistics, either at an advanced undergraduate or graduate level. As some of the recommended so-called “large-sample” rules in textbooks have been found to be inappropriate, this monograph endeavors to provide more up-to-date information on these topics. I have also included a number of related topics not generally found in textbooks. The emphasis is on model building and the estimation of parameters from the models.

It is assumed that the reader has a background in statistical theory and inference and is familiar with standard univariate and multivariate distributions, including conditional distributions. This monograph may also be helpful for the statistics practitioner who is involved with statistical consulting in this area, particularly with regard to inference for one and two proportions or probabilities.

[Chapter 1](#) looks at the difference between a proportion and probability. It focuses on a proportion leading to the Hypergeometric model and its Binomial approximation, along with inference for the proportion. Inverse sampling is also considered. [Chapter 2](#) focuses on estimating a probability and considers the Binomial distribution in detail as well as inverse sampling. Exact and approximate inferences for a probability are considered. In [Chap. 3](#), the main focus is on comparing two proportions or two probabilities and related quantities such as the relative risk and the odds ratio from the same or different populations using the Multi-hypergeometric or Multinomial distributions. Simultaneous confidence intervals for several parameters are also considered. The Multinomial distribution is the basis for a number of hypothesis and goodness of fit tests, and these are discussed in [Chap. 4](#) with particular attention given to 2×2 tables and matched data. In [Chap. 5](#), we look briefly at two logarithmic models for discrete data, namely the log linear and the logistic models.

I would like to thank two reviewers for their very helpful comments on a previous draft.

Auckland, New Zealand, June 2012

George A. F. Seber

Contents

1	Single Proportion	1
1.1	Distribution Theory	1
1.2	Inverse Sampling	4
1.3	Application to Capture-Recapture	5
1.4	Inference for a Proportion	6
	References	8
2	Single Probability	9
2.1	Binomial Distribution	9
2.1.1	Estimation	9
2.1.2	Likelihood-Ratio Test	10
2.1.3	Some Properties of the Binomial Distribution	10
2.1.4	Poisson Approximation	12
2.2	Inverse Sampling	12
2.3	Inference for a Probability	13
2.3.1	Exact Intervals	13
2.3.2	Exact Hypothesis Test	14
2.3.3	Approximate Confidence Intervals	15
	References	17
3	Several Proportions or Probabilities	19
3.1	Multi-Hypergeometric Distribution	19
3.2	Comparing Two Proportions from the Same Population	20
3.2.1	Nonoverlapping Proportions	20
3.2.2	Dependent Proportions	21
3.2.3	Two Independent Proportions	23
3.3	Comparing Two Probabilities from Independent Binomial Distributions	23
3.3.1	Difference of Two Probabilities	23
3.3.2	Relative Risk	26
3.3.3	Odds Ratio	27

3.4	Multinomial Distribution	28
3.4.1	Maximum Likelihood Estimation	28
3.4.2	Comparing Two Probabilities from the Same Population	30
3.5	Asymptotic Multivariate Normality	30
3.5.1	Simultaneous Confidence Intervals	32
3.5.2	Bonferroni Confidence Intervals	34
3.6	Animal Population Applications	34
3.6.1	Random Distribution of Animals	34
3.6.2	Multiple-Recapture Methods.	35
3.7	Appendix: Delta Method	37
3.7.1	General Theory.	37
3.7.2	Application to the Multinomial Distribution.	38
3.7.3	Asymptotic Normality	39
	References	39
4	Multivariate Hypothesis Tests.	41
4.1	Multinomial Test Statistics	41
4.1.1	Likelihood-Ratio Test for $p = p_0$	41
4.1.2	Wald Test	42
4.1.3	Score Test	43
4.1.4	Deviance	44
4.2	A More General Hypothesis	44
4.2.1	Freedom Equation Specification of H_0	45
4.2.2	General Likelihood-Ratio Test	45
4.3	Contingency Tables	46
4.3.1	Test for Independence in a Two-Way Table.	47
4.3.2	Several Multinomial Distributions.	49
4.4	2×2 Contingency Tables	50
4.4.1	Examples	50
4.4.2	Chi-Square Test	51
4.4.3	Fisher's Exact Test	53
4.4.4	Correlated Data.	54
	References	56
5	Logarithmic Models.	59
5.1	Log-Linear Models	59
5.1.1	Contingency Tables.	59
5.1.2	Test Statistics	61
5.1.3	Application of Log Linear Models to Epidemiology	62
5.2	Logistic Models.	64
5.2.1	Independent Binomial Distributions.	65
5.2.2	Logistic Multinomial Regression Model.	67
	References	69

Chapter 1

Single Proportion

Abstract This chapter focusses on the problem of estimating a population proportion using random sampling with or without replacement, or inverse sampling. Exact and approximate confidence intervals are discussed using the Hypergeometric distribution. Applications to capture-recapture models are given.

Keywords Proportion · Hypergeometric distribution · Simple random sample · Sampling fraction · Negative-Hypergeometric distribution · Binomial distribution · Confidence intervals for a proportion · Single capture-recapture model

1.1 Distribution Theory

Since this monograph is about modelling proportions and probabilities, I want to begin by comparing the two concepts, proportion and probability, as these two ideas are sometimes confused. If we have a population of N people and M are male, then the proportion of males in the population is $p = M/N$. Suppose we now carry out a random experiment and choose a person at random from the population. What we mean by this is that we choose a person in such a way that every person is equally likely to be chosen. If the population is small we could achieve this by putting the names of everyone in a container, shuffling the names by rotating the container, and drawing one name out. This kind of manual process is used in lottery games. For example in New Zealand we have Lotto in which 40 numbered balls are tossed around in a container until one eventually drops out.

For a large population of people we could number everyone and then choose a number at random using a computer. In this case we can obtain the probability of getting a male using the law of probability relating to equally likely events, namely if we have N equally likely outcomes of an experiment (the so-called sample space) and M of these have a given characteristic, then the probability of choosing a member with the given characteristic is simply M/N or p again. This means that for the