The Ratio-Plot in the Practice of Capture-Recapture Studies and Analysis

Dankmar Böhning

Professor in Medical Statistics
School of Mathematics & S3RI
University of Southampton

Medical Statistics Seminar
November 15, 2011

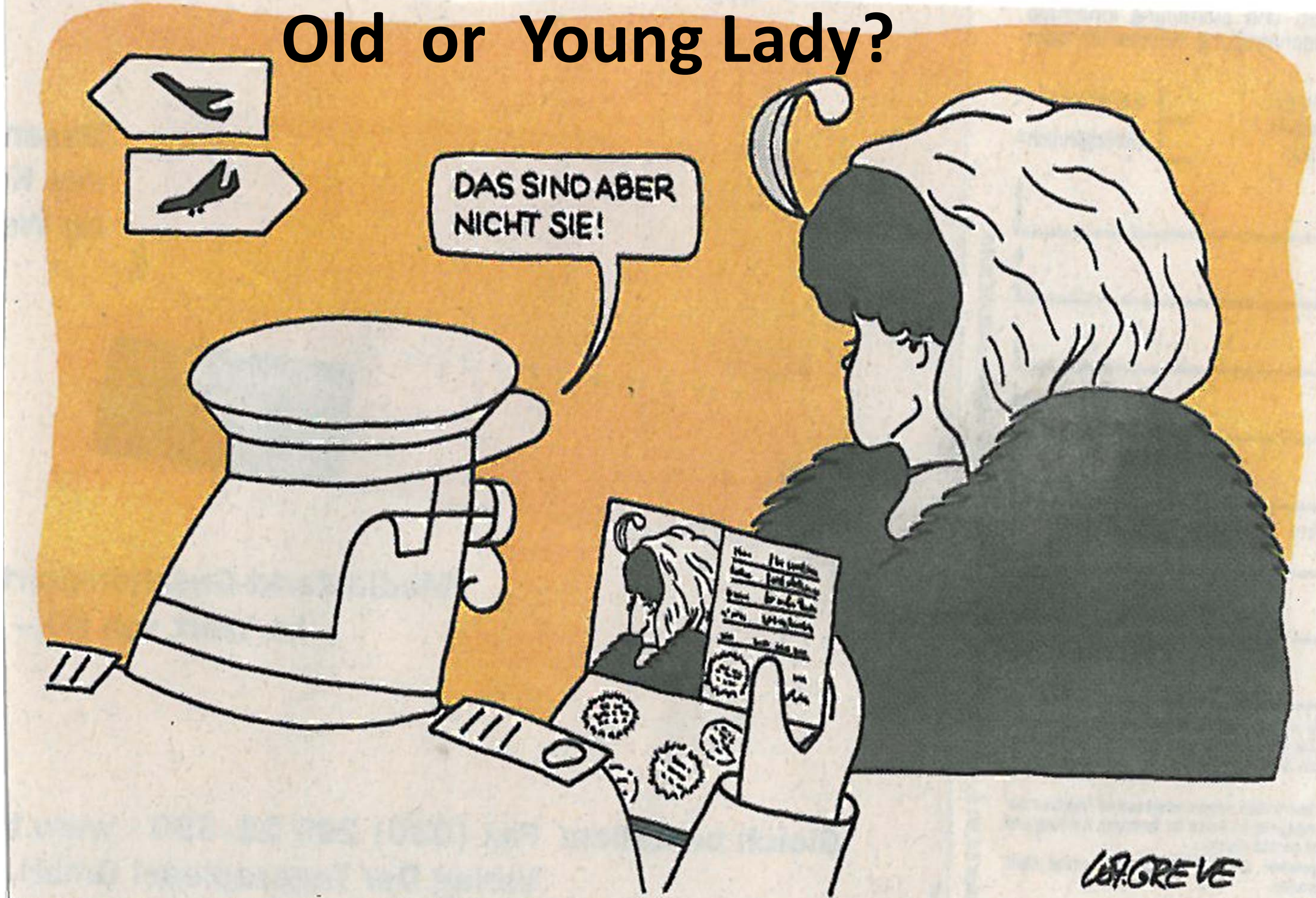
purpose of the talk

present and discuss a graphical device for investigating a distributional structure

areas involved

- graphical statistics
- robust statistics
- capture-recapture in life science applications

Old or Young Lady?



Verwirrung am Schalter. *Alte Frau oder junge Frau?*

Cartoon: Katharina Greve

The idea of capture-recapture

- objective is to determine the size N of an elusive target population
- some mechanism (life trapping, register, surveillance system) identifies a unit **repeatingly**
- this repeated identification (recapturing) works either
  - in **time**
  - in **clusters**
  - in **space**
- there is a count X informing about the number of identifications of each unit in the target population

sample

available: sample

$$X_1, X_2, ...X_N$$

also the frequencies

$$\begin{aligned}
f_0 &= \text{frequency of units captured zero times} \\
f_1 &= \text{frequency of units captured exactly 1 times} \\
f_2 &= \text{frequency of units captured exactly 2 times} \\
... &= ... \\
f_x &= \text{frequency of units captured exactly } x \text{ times} \\
... &= ... \\
f_m &= \text{frequency of units captured exactly } m \text{ times}
\end{aligned}$$

problem

if $X_j = 0$ unit is **not observed** leading to a reduced observable sample

$$X_1, X_2, ...X_n$$

where – w.l.g. – we assume that

$$X_{n+1} = X_{n+2} = ... = X_N = 0$$

hence

$$f_0 = N - n \text{ is } \textbf{unknown}$$

Grizzly bears in the Yellowstone ecosystem

A case study for illustration

grizzly bears in the Yellowstone ecosystem

Boyce et al. (2001) and Keating et al. (2002) recorded the sighting frequencies of female grizzly bears with cubs-of-the-year in the Yellowstone ecosystem; the data for three different observational periods are provided in the table below:

Table: *Female Grizzly Bears in the Yellowstone ecosystem*

Year	f_1	f_2	f_3	f_4	f_5	f_6	f_7	n
1996	15	10	2	1	0	0	0	28
1997	13	7	4	1	3	0	1	29
1998	11	13	5	1	1	0	2	33

Hser's Data on Estimating Hidden Intravenous Drug Users in Los Angeles 1989

- intravenous drug users in L.A. county were entered into the California Drug Abuse Data System (CAL-DADS)
- the data below refer to the frequency distribution of the episode count per drug user in 1989

the frequency distribution of the **episode count per drug user** for the year 1989:

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Del Rio Vilas's Data on Estimating Hidden Scrapie in Great Britain 2005

- sheep is kept in holdings in Great Britain (and elsewhere)
- the occurrence of scrapie is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) summarizing abbatoir survey, stock survey and the statutory reporting of clinical cases
- CSFS established since 2004

the frequency distribution of the **scrapie count within each holding** for the year 2005:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Finding errors in software

Table: *Illustration with Case Data from Software Inspection (Wohlin et al. 1995)*

Error i	*Reviewers* R1	R2	...	R22	Marginal Y_i
1	1	0	...	1	2
2	1	1	...	0	4
3	0	0	...	1	2
4	0	0	...	0	**0**
5	0	1	...	0	1
...	...	...	...	...	...
38	1	1	...	0	7

Hidden errors in software

Table: Zero-truncated count distribution of software errors

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	5	1	5	1	3	2	0	5	4	2

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}	n
3	1	0	2	0	1	0	0	0	1	36

a problem from text analysis

- given published text, how many words does an author know, but not use?
- can be approached as a capture–recapture problem

Biometrika (1976), 63, 3, *pp.* 435–47
With 3 text-figures
Printed in Great Britain

Estimating the number of unseen species: How many words did Shakespeare know?

By BRADLEY EFRON and RONALD THISTED
Department of Statistics, Stanford University, California

Summary

Shakespeare wrote 31 534 different words, of which 14 376 appear only once, 4343 twice, etc. The question considered is how many words he knew but did not use. A parametric empirical Bayes model due to Fisher and a nonparametric model due to Good & Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that the models are equivalent to supposing that Shakespeare knew at least 35 000 more words.

Some key words: Empirical Bayes; Euler transformation; Linear programming; Negative binomial; Vocabulary.

1. Introduction

Estimating the number of unseen species is a familiar problem in ecological studies. In this paper the unseen species are words Shakespeare knew but did not use. Shakespeare's known works comprise 884 647 total words, of which 14 376 are types appearing just one time, 4343 are types appearing twice, etc. These counts are based on Spevack's (1968) concordance and on the summary appearing in an unpublished report by J. Gani & I. Saunders. Table 1 summarizes Shakespeare's word type counts, where n_x is the number report of the same title, available from the authors on request.

Table 1. *Shakespeare's word type frequencies*

x	1	2	3	4	5	6	7	8	9	10	Row total
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78

Entry x is n_x, the number of word types used exactly x times. There are 846 word types which appear more than 100 times, for a total of 31 534 word types.

2. The basic model

We use the species trapping terminology of Fisher's paper. Suppose that there exist S species and that after trapping for one unit of time we have captured x_s members of species s. Of course we only observe those values x_s which are greater than zero. The basic distributional assumption is that members of each species s enter the trap according to a Poisson process, the process for species s having expectation λ_s per unit time, so that x_s has a Poisson distribution of mean λ_s ($s = 1, \ldots, S$). Most of the calculations in this paper do not require the S individual Poisson processes to be independent of one another. Whenever indepen-

How many words did Shakespeare know?

- Efron and Thisted (1987, *Biometrika*): How many words did Shakespeare know, but not use?
- important question in text analysis and estimation of language knowledge

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	…	n
-	14,376	4,343	2,292	1,463	1,043	837	638	..	31,534

Application Areas

- Epidemiology and Medicine
- Biology and Agriculture
- Social Science and Criminology
- Research on Terrorism
- Systems Engineering
- Text and Language Analysis

Assumptions

crucial

- closed population
  - no deaths
  - no births
  - no migration
- independence between subjects
  - no dependence between different subjects

Assumptions

less crucial

- independence between captures
  - lists or sources identify independently
  - repeated identification occurs independently
- homogeneity of capture probability
  - different members of population are equally likely to be captured

ways to a solution: modelling repeated capturing

$$
\begin{aligned}
p_0 &= \text{ probability for never capturing the unit} \\
p_1 &= \text{ probability for capturing the unit 1 time} \\
p_2 &= \text{ probability for capturing the unit 2 times} \\
\ldots &= \ldots
\end{aligned}
$$

in general

$$p_x = p_x(\lambda) = \text{ probability for capturing the unit } x \text{ times}$$

the idea for a solution

use a model for $p_x = p_(\lambda)$ such as the Poisson

$$p_x = \exp(-\lambda)\lambda^x / x$$

estimate λ by some method to yield $\hat{\lambda}$, and, since

$$E(n) = N(1 - p_0)$$

we get

$$\hat{N} = \frac{n}{1 - p_0(\hat{\lambda})} = \frac{n}{1 - \exp(-\hat{\lambda})}$$

a good estimator under homogeneity:

write

$$p_0 = \exp(-\lambda) = \frac{\exp(-\lambda)\lambda}{\lambda} = \frac{p_1}{E(X)}$$

which can be estimated by

$$\hat{p}_0 = \frac{f_1/N}{S/N} = \frac{f_1}{S}$$

where $S = 0f_0 + 1f_1 + ... + mf_m$ which is always known, leading to

$$\hat{N}_T = \frac{n}{1 - f_1/S}$$

the **Good-Turing** estimate of N (Good 1953)

Diagnostic Device for Homogeneity: The Ratio Plot

Poisson homogeneity can be supported by means of the **ratio plot**

$$x \to r_x = \frac{(x+1)p_{x+1}}{p_x}$$

for a Poisson

$$p_x = \exp(-\lambda)\lambda^x/x!$$

so the ratio

$$r_x = \frac{(x+1)p_{x+1}}{p_x} = \lambda$$

is a **horizontal line**

The Ratio Plot for Poisson Homogeneity

the **ratio plot**

$$x \rightarrow r_x = \frac{(x+1)p_{x+1}}{p_x}$$

can be estimated by the **empirical ratio plot**

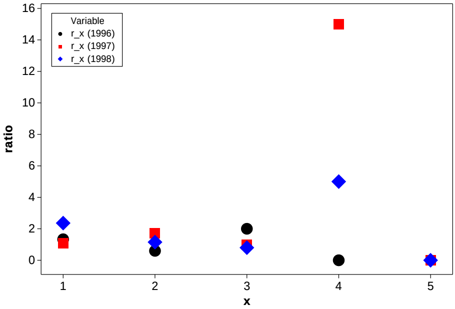
$$x \rightarrow \hat{r}_x = \frac{(x+1)f_{x+1}}{f_x}$$

which should show a specific pattern: **a horizontal line**

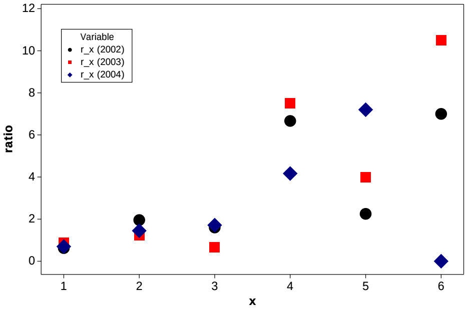
Ratio plot in reality

- we look at two examples:
- Grizzle bears in the Yellowstone eco system
- hidden scrapie in Great Britain 2002, 2003, 2004

Grizzly bears in the Yellowstone ecosystem



Scrapie in Great Britain based upon the SND



Effects of heterogeneity on the Turing estimator $\hat{N}$

Table: *Simulation using* $X \sim (1-\alpha)Po(0.5) + \alpha Po(\mu)$ *for* $N = 1000$ *and* $\alpha = 0.2$*; replication size* 1000

	contamination				
μ	0.5	1	2	3	9
$E(\hat{N})$	1002.7	938.61	776.54	694.70	579.17
$Var(\hat{N})$	85.3	63.47	38.21	28.31	18.39

The Decontaminated Turing Estimator

$$p_0 = \exp(-\lambda) = \frac{\exp(-\lambda)\lambda}{\lambda} = \frac{p_1}{E(X)}$$

we had estimated by

$$\hat{p}_0 = \frac{f_1/N}{S/N} = \frac{f_1}{S}$$

where $S = 0f_0 + 1f_1 + \ldots + mf_m$ which will be **too large** if there are **contaminations** and

$$\hat{N}_T = \frac{n}{1 - f_1/S}$$

will be **biased (much too small)**

The Decontaminated Turing Estimator

$$p_0 = \exp(-\lambda) = \frac{\exp(-\lambda)\lambda}{\lambda} = \frac{p_1}{E(X)}$$

we had estimated by

$$\hat{p}_0 = \frac{f_1/N}{S/N} = \frac{f_1/N}{\widehat{E(X)}}$$

use **robust** estimators for $E(X) = \lambda$:

- $\hat{\lambda}_1 = \frac{2f_2}{f_1} \longrightarrow \frac{2p_2}{p_1} = \lambda$
- $\hat{\lambda}_2 = \frac{2f_2+3f_3}{f_1+f_2} \longrightarrow \frac{2p_2+3p_3}{p_1+p_2} = \lambda$
- $\hat{\lambda}_3 = \frac{2f_2+3f_3+4f_4}{f_1+f_2+f_3} \longrightarrow \frac{2p_2+3p_3+4p_4}{p_1+p_2+p_3} = \lambda$
- ...

The Decontaminated Turing Estimator

then use $\hat{\lambda}_j$ instead of S/N

$$\hat{p}_0 = \frac{f_1/N}{S/N} \underbrace{=}_{\text{replace}} \frac{f_1/N}{\hat{\lambda}_j}$$

leading to

$$\hat{N} = \frac{n}{1 - (f_1/\hat{N})/\hat{\lambda}_j}$$

$$\hat{N} - f_1/\hat{\lambda}_j = n$$

so that the **decontaminated Turing** estimator arises

$$\hat{N}_{DT} = n + f_1/\hat{\lambda}_j$$

The Decontaminated Turing Estimator

the **decontaminated Turing** estimator

$$\hat{N}_{DT} = n + f_1/\hat{\lambda}_j$$

- $\hat{\lambda}_1 = \frac{2f_2}{f_1} \rightarrow \hat{N}_{DT} = n + f_1/(2f_2/f_1) = \overbrace{n + f_1^2/(2f_2)}^{\text{robust}} = \hat{N}_C$(Chao)
- $\hat{\lambda}_2 = \frac{2f_2+3f_3}{f_1+f_2} \rightarrow \hat{N}_{DT} = n + \frac{f_1(f_1+f_2)}{2f_2+3f_3}$
- ...
- $\hat{\lambda}_3 = \frac{2f_2+3f_3+...mf_m}{f_1+f_2+...f_{m-1}} \rightarrow \hat{N}_{DT} = \underbrace{n + \frac{f_1(f_1+f_2+...f_{m-1})}{2f_2+3f_3+...mf_m}}_{\text{efficient}}$

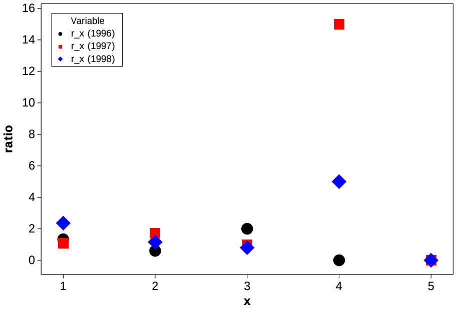
The Decontaminated Turing Estimator

the general question remains for

$$\hat{N}_{DT} = n + f_1/\hat{\lambda}_j$$

- which $\hat{\lambda}_j$?
- again the ratio plot can help!

Grizzly bears in the Yellowstone ecosystem



measuring goodness-of-fit

$$\chi^2(k) = \sum_{x=1}^{k+1} \frac{[f_x - n_k Po_+(x;\hat{\lambda}_k)]^2}{n_k Po_+(x;\hat{\lambda}_k)} \tag{1}$$

where

- $Po_+(x;\lambda) = Po(x;\lambda)/[Po(1;\lambda) + ... + Po(k+1;\lambda)]$
- and $n_k = f_1 + ... + f_{k+1}$.

choosing k in the female Grizzle Bears data

Table: Estimates of the number of Female Grizzly Bears in the Yellowstone ecosystem for 1997 for different values of the robustified Turing estimate

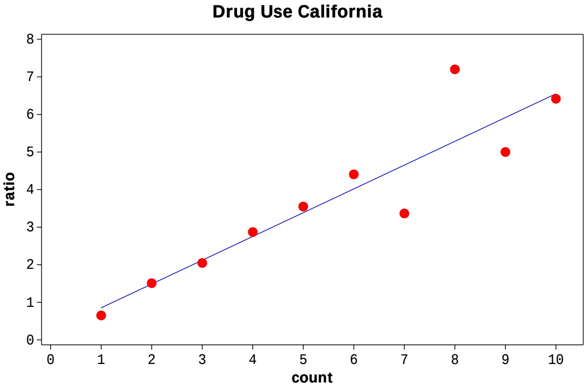
k	$\chi^2(k)$	p-value	$\hat{\lambda}_k$	$\hat{N}_k$
1	0.000	1.000	1.08	41.1
2	0.241	0.623	1.30	39.0
3	0.264	0.876	1.25	39.4
4	7.627	0.054	1.80	36.2
5	10.473	0.033	1.61	37.1

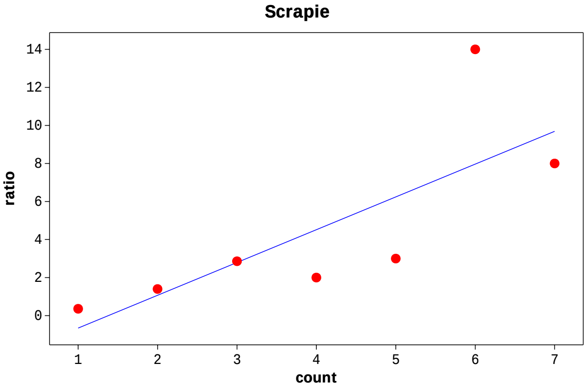
Structured Heterogeneity

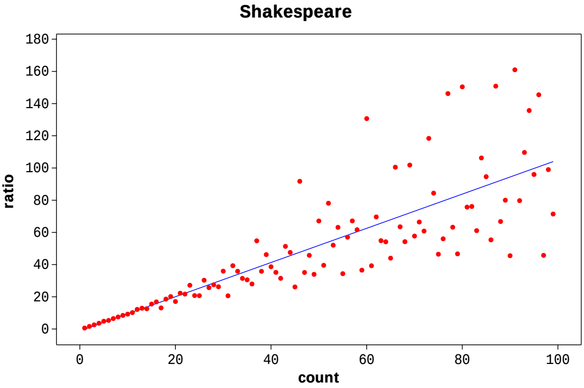
frequently, a certain structure for the heterogeneity distribution can be identified by means of the **ratio plot**

$$x \to \frac{(x+1)f_{x+1}}{f_x}$$

showing a specific pattern: **a straight line**


Drug Use California
ratio
count
0
1
2
3
4
5
6
7
8
9
10


Scrapie
ratio
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
count


Shakespeare
ratio
180
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
count

Structured Heterogeneity

if

$$x \to \frac{(x+1)p_{x+1}}{p_x}$$

with

$$p_x = \int_0^\infty \frac{\exp(-t)t^x}{x!} \lambda(t) dt$$

is a **straight line**

- what does it tell us about $\lambda(t)$?

structured heterogeneity: Gamma–density

suppose $\lambda(t) = \frac{1}{\theta} t^{k-1} \exp(-t/\theta)/\Gamma(k)$ is the Γ-density with parameters θ and k; then

$$p_x = \int_0^\infty \exp(-t) t^x / x! \lambda(t) dt = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} p^k (1-p)^x$$

the **negative binomial density** with event parameter $p = \frac{1}{1+\theta}$ and shape parameter k

structured heterogeneity: Gamma–density

suppose $\lambda(t)$ is the Γ-density with parameters p and k; then

$$\frac{(x+1)p_{x+1}}{p_x} = (x+k)(1-p)$$

the **straight line** with slope $(1-p)$ and intercept $k(1-p)$

- this indicates that it is reasonable to assume a **Gamma-distribution** for the heterogeneity distribution of the Poisson parameter

structured heterogeneity: the NB

under the negative binomial

$$p_x = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} p^k (1-p)^x$$

we are interested in $p_0 = p^k$ and have that

- $p_1 = kp^k(1-p)$
- $E(X) = k\frac{1-p}{p}$

hence

$$p_0 = p^k \text{ and } \frac{p_1}{E(X)} = \frac{kp^k(1-p)}{k\frac{1-p}{p}} = p^{k+1}$$

structured heterogeneity: the NB

$$p_0 = p^k \text{ and } \frac{p_1}{E(X)} = \frac{kp^k(1-p)}{k\frac{1-p}{p}} = p^{k+1}$$

and, finally,

$$p_0 = \left(\frac{p_1}{E(X)}\right)^{\frac{k}{k+1}}$$

this leads to the **generalised Turing** estimator

$$\hat{N}_{GT} = \frac{n}{1 - \left(\frac{f_1}{S}\right)^{\frac{k}{k+1}}}$$

estimating k with a regression approach:

since for a NB

$$x \to \frac{(x+1)p_{x+1}}{p_x} = (1-p)(x+k) = (1-p)k + (1-p)x = \alpha + \beta x$$

it seems reasonable to explore the regression model

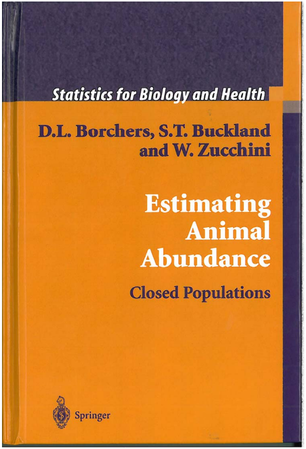
$$\frac{(x+1)f_{x+1}}{f_x} = \alpha + \beta x + \epsilon$$

using weights inversely related to the variance of $\frac{(x+1)f_{x+1}}{f_x}$) and the estimate for k is obtained from $k = (1-p)k/(1-p) = \alpha/\beta$

$$\hat{k} = \hat{\alpha}/\hat{\beta}$$

Example with N known: Golf Tees Study

- 250 clusters of golf tees were placed
- in an area of 1,680 m^2
- surveyed by students of the University of St. Andrews
- details are as follows:



Key idea: choose the most likely value as the estimate, given what was observed.

Key notation:

N:	population size (abundance)
$\hat{N}$:	estimator of population size
n:	number of animals detected (sample size)
p:	probability of detecting an animal

2.1 An example problem

It is often easier to understand how abundance estimation methods worl if we can check our estimates against the true population after estimatin abundance, to see how well we did. This is impractical with real popula tions, so we will be using examples with artificial populations for illustra tion.

One such population, used repeatedly in this book, is the one introduce in Chapter 1. The data are actually from independent surveys by eigh

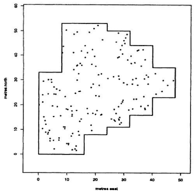


Figure 2.1. Example data, detected animals. Each dot represents a detected ani mal within the survey region. In all, $n = 162$ animals were detected.

different observers of a population of 250 groups (760 individuals) of gol tees, not plants, contrary to what we said in Chapter 1. The tees, of tw colours, were placed in groups of between 1 and 8 in a survey region o 1,680 m^2, either exposed above the surrounding grass, or at least partl hidden by it. They were surveyed by the 1999 statistics honours class a the University of St Andrews,[1] Scotland, so while golf tees are clearly no animals (or plants), the survey was real, not simulated. We treat each grou of golf tees as a single "animal", with size equal to the number of tees i the group; yellow tees are "male", green are "female"; tees exposed abov the surrounding grass are classified as exposed ("exposure=1"), others a unexposed ("exposure=0").

Other populations presented later in the book were generated with the R library WiSP and only ever existed inside a computer. In all cases, we refe to them as animal populations, and to their members as animals.

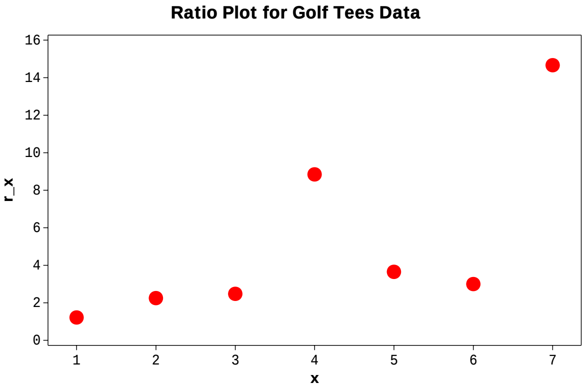
Figure 2.1 shows the locations of the animals detected by at least on observer on a survey of our first example population. A total of $n = 16$ animals were seen, but an unknown number were missed. We would like t use what was seen to answer the question: How many animals are there?

[1] We are grateful to Miguel Bernal for making these data available to us. They wer collected by him as part of a Masters project at the University of St Andrews. St Andrew is known as "the home of golf", so tees seemed an appropriate target object.

Example with N known

Table: *The Golf Tees data of Borchers, Buckland and Zucchini (2002): true number* N *of golf tees is* 250

f_1	46
f_2	28
f_3	21
f_4	13
f_5	23
f_6	14
f_7	6
f_8	11
n	162


Ratio Plot for Golf Tees Data
r_x
x
0
2
4
6
8
10
12
14
16
1
2
3
4
5
6
7

Results of Estimation: True $N = 250$

estimator	value
$\hat{k}$	1.07
$\hat{N}_T$	177
$\hat{N}_{GT}$	224
$\hat{N}_C$	200

related and recent papers

- Rocchetti, I., Bunge, J., Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics* **5**, 1512-1533.
- Böhning, D., Baksh, M.F., Lerdsuwansri, R., Gallagher, J. (2011). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics* (in press).
- Rocchetti, I., Alfò, M., Böhning, D. (2011). A regression-type estimator for beta-binomial capture-recapture data. *Biostatistics* (submitted).