The Ratio Plot for Capture-Recapture Estimation

Dankmar Böhning

Applied Statistics, School of Biological Sciences
University of Reading

June 10, 2010

Acknowledgment

joint work with:

- Fazil Baksh, Applied Statistics, UoR
- James Gallagher, Statistical Services Center, UoR
- Victor Del Rio Vilas, DEFRA

The idea of capture-recapture

- some mechanism (life trapping, register, surveillance system) identifies a unit **repeatingly**
- this repeated identification (recapturing) works either
  - in **time**
  - in **clusters**
- 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

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

An example for repeated sampling in time

multiple detections in time

- suppose there is an observation period in which each member of the target population can be detected several times
- let count X_i denote the number of times unit i is detected in the observational period where $i = 1, 2, ..., N$
- note that $X_i = 0$ is **not** observed
- also let $f_0, f_1, ..., f_m$ denote the frequency of times a unit has been detected, more precisely, f_y is the frequency of units (animals, human) exactly detected y times

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	S
1996	15	10	2	1	0	0	0	28	45
1997	13	7	4	1	3	0	1	29	65
1998	11	13	5	1	1	0	2	33	75

An example for repeated sampling within a cluster

multiple detections within a cluster

- suppose there is a mechanism in place which detects (potentially multiple) cases (of a disease) within a cluster (herd, village, household) in a target population of clusters
- let count X_i denote the number of cases which are detected in unit (cluster) $i$where $i = 1, 2, ..., N$
- note that $X_i = 0$ is **not** observed
- also let $f_0, f_1, ..., f_m$ denote the frequency of times a unit has been detected, more precisely, f_x is the frequency of units (animals, human) exactly detected x times

A case study for illustration

Scrapie in Great Britain

Böhning and Del Rio Vilas (2008) look at scrapie occurrence based upon the Scrapie Notifications Database (SND):

- X is the case count per herd
- repetition occurs within a **herd (cluster)**

Scrapie in Great Britain: symptoms

- chronic diseases of the immune system
- behavioral symptoms including typical moving patterns
- scrapie apparently causes an itching sensation in the animals from which the name scrapie is derived as one of the clinical signs is that affected animals will compulsively **scrape** off their fleece against rocks, trees or fences

A case study for illustration

Scrapie in Great Britain

Böhning and Del Rio Vilas (2008) look at scrapie occurrence based upon the Scrapie Notifications Database (SND):

- X is the case count per herd
- repetition occurs within a **herd (cluster)**

Table: *Scrapie surveillance in Great Britain based upon the SND*

Year	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_{8+}	n
2002	74	23	15	6	8	3	3	12	144
2003	66	29	12	2	3	2	3	17	134
2004	83	29	14	6	5	6	0	8	151

problem

if p_0 would be known, then

$$N = Np_0 + \underbrace{N(1-p_0)}_{\text{estimated by } n}$$

hence

$$\hat{N} = \frac{n}{1-p_0}$$

the **Horvitz-Thompson** estimator of N

idea: to model repeated capturing

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

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

the idea for a solution

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

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

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

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

under heterogeneity

more realistic to assume population heterogeneity in the Poisson parameter with **heterogeneity distribution** $\lambda(t)$

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

dealing with the heterogeneity distribution

- parametric
- nonparametric
- estimation

where do we stand with this?

- under homogeneity (**solved**)
- under homogeneity with contaminations (**solved**)
- under structured heterogeneity (**solution under progress**)
- under unstructured heterogeneity (**largely unsolved**)

Estimation under Homogeneity and the Ratio Plot

a good estimator under homogeneity: Turing

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)

What are the problems?

- Turing (and others such as MLE) is only appropriate under homogeneity
- Chao (1987) is appropriate under heterogeneity, but remains to be biased
- how can homogeneity be supported?

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

$$r_x = \frac{(x+1)\exp(-\lambda)\lambda^{x+1}/(x+1)!}{\exp(-\lambda)\lambda^x/x!} = \lambda$$

is a **horizontal line** if p_x is a Poisson (Hoaglin 1980; Gart 1970)

The Ratio Plot for Poisson Homogeneity

the **ratio plot**

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

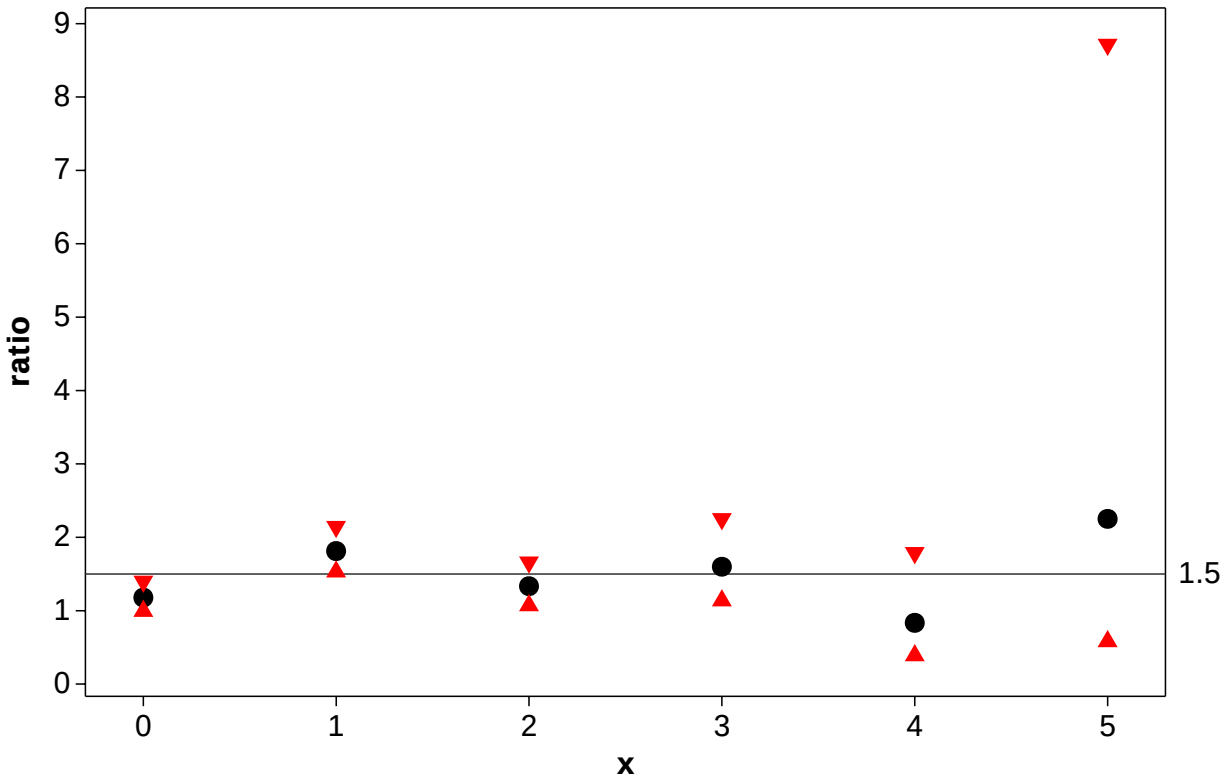
can be estimated by the **empirical ratio plot**

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

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

Ratio plot in an ideal case of homogeneity

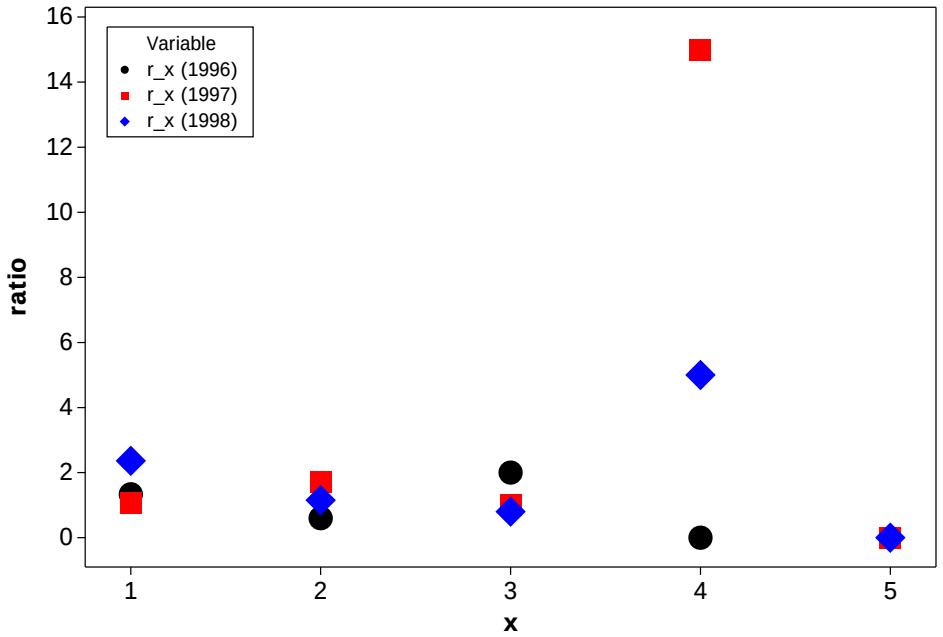
- ratio plot in an ideal case of homogeneity:
- $X_i \sim Po(1.5)$, $i = 1, ..., 1000$
- look at: $x \longrightarrow \hat{r}_x = (x+1)f_{x+1}/f_x$


ratio
x
1.5
0
1
2
3
4
5
6
7
8
9

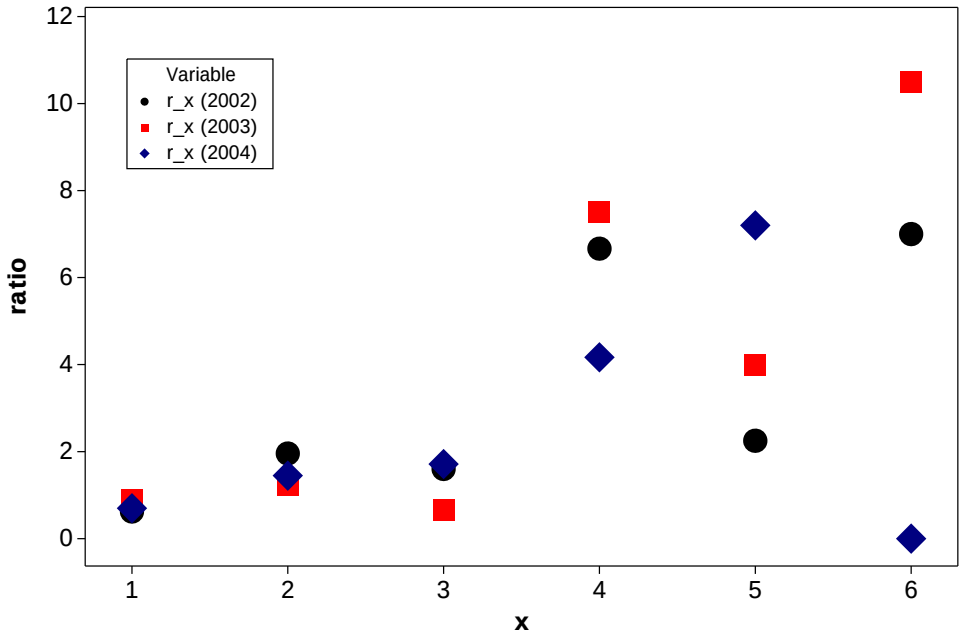
Ratio plot in reality

- 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



The Ratio Plot for Contaminated Homogeneity and 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 $S = 0f_0 + 1f_1 + ... + mf_m$ 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 **robustified** or **decontaminated** estimators for $E(X) = \lambda$:

- $\hat{\lambda}_1 = \frac{2f_2}{f_1} \longrightarrow \frac{2p_2}{p_1} = \lambda$
- $\hat{\lambda}_2 = \frac{2f_2+3f_3}{f_1+f2} \longrightarrow \frac{2p_2+3p_3}{p_1+p_2} = \lambda$
- $\hat{\lambda}_3 = \frac{2f_2+3f_3+4f_4}{f_1+f2+f_3} \longrightarrow \frac{2p_2+3p_3+4p_4}{p_1+p_2+p_3} = \lambda$
- ...

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

$$\hat{N} = \frac{n}{1 - (f_0/\hat{N})/\hat{\lambda}_j}$$

$$\hat{N} - f_0/\hat{\lambda}_j = n$$

so that the **decontaminated Turing** estimator arises

$$\hat{N}_j = n + f_0/\hat{\lambda}_j$$

An optimality property of the Decontaminated Turing estimator

optimality

The beneficial behavior of $\hat{N}_j$ can be seen in the following result for a **simple contamination model** in which the Poisson distribution is contaminated by a second Poisson component with weight α.

Theorem

Let $p_j = (1-\alpha)Po(j;\lambda) + \alpha Po(j;\mu)$, where $Po(j;\lambda) = e^{-\lambda}\lambda^j/j!$ and $E(X) = \sum_{j=0}^{\infty} jp_j$. Then, for any $2 \leq k \leq m$

$$lim_{N\to\infty}\hat{N}_k/N = (1-p_0) + p_1\frac{p_1 + ... + p_{k-1}}{2p_2 + ... + kp_k}$$

$$\to 1 \text{ for } \mu \to \infty.$$

In addition, for the Turing estimator we have

$$lim_{N\to\infty}\hat{N}/N = lim_{N\to\infty}\frac{n}{1-f_1/S}/N = \frac{(1-p_0)}{1-p_1/E(X)}$$

$$\to [1-(1-\alpha)\exp(-\lambda)] \text{ for } \mu \to \infty.$$

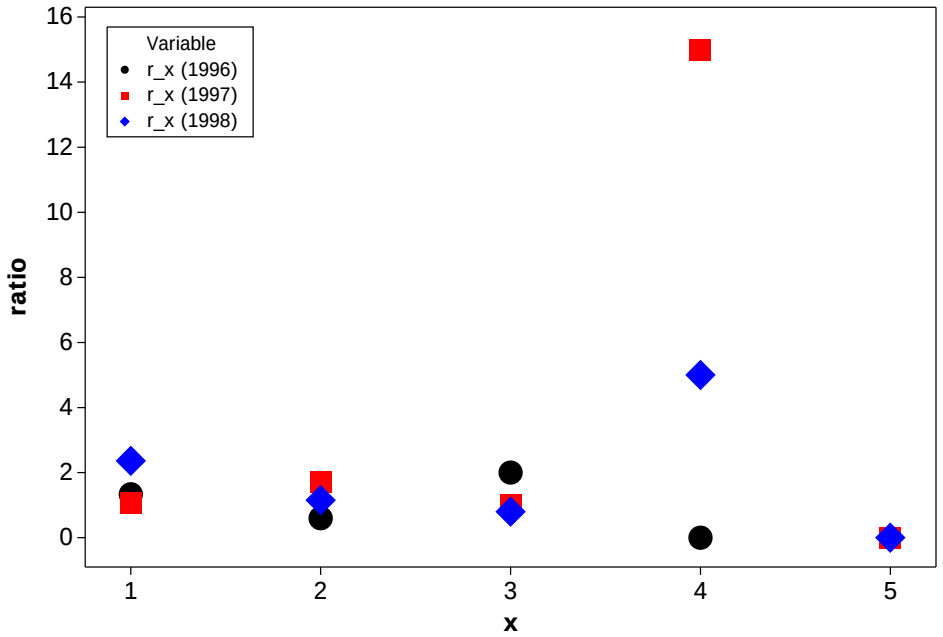
The Decontaminated Turing Estimator

the question remains for

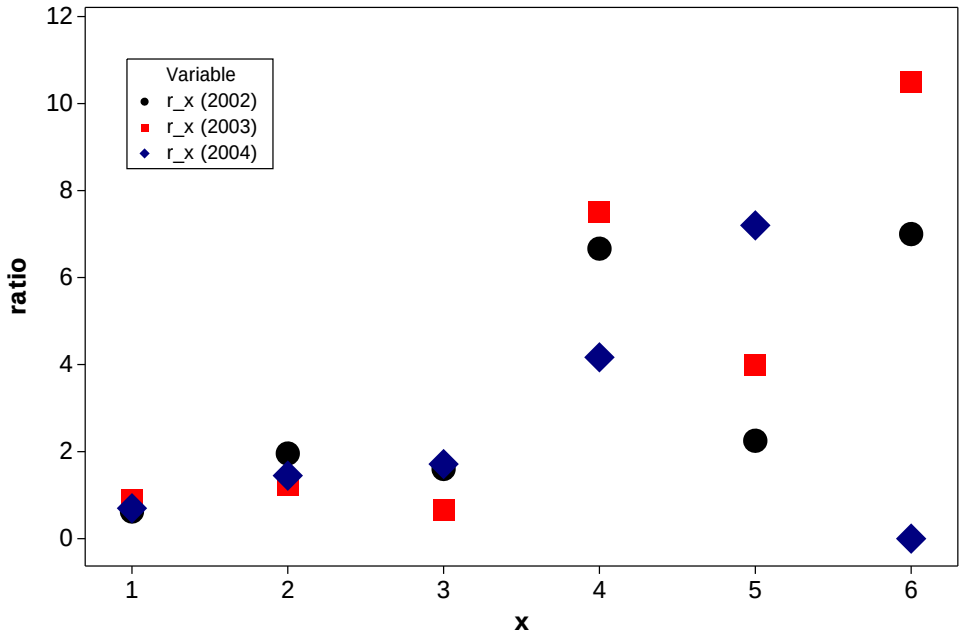
$$\hat{N}_j = n + f_0/\hat{\lambda}_j$$

- which $\hat{\lambda}_j$?
- here again the ratio plot helps:

Grizzly bears in the Yellowstone ecosystem



Scrapie in Great Britain based upon the SND



Example: estimation of dystrophin density

Background

- The example is from a study (Cullen et al. (1990)) on dystrophin density in human muscle. Dystrophin, a gene product of possible importance in muscular dystrophies, may be located within muscle fibers using an electron microscope (see also Matthews and Appleton 1993)
- Units (epitops) of Dystrophin cannot be detected until they have been labelled by a suitable electron-dense substance; gold-conjugated antibodies which adhere to the dystrophin was used.
- Not all units can be labelled and more than one anti-body molecule may attach to a dystrophin unit. To achieve an unbiased estimate of the dystrophin density, it is important to account for all labelled and unlabelled units.

Zero-truncated Count Data for Dystrophin Density

Shown in the Table below is the observed count of the the number of antibody molecules on each dystrophin unit within the muscle fibres of biopsy specimens taken from normal patients.

Table: Distribution of Antibody Counts attached to Dystrophin Units

f_0	f_1	f_2	f_3	f_4	f_5	n
-	122	50	18	4	4	198

Evidently, interest is in f_0, the number of unlabeled or unobserved dystrophin units.

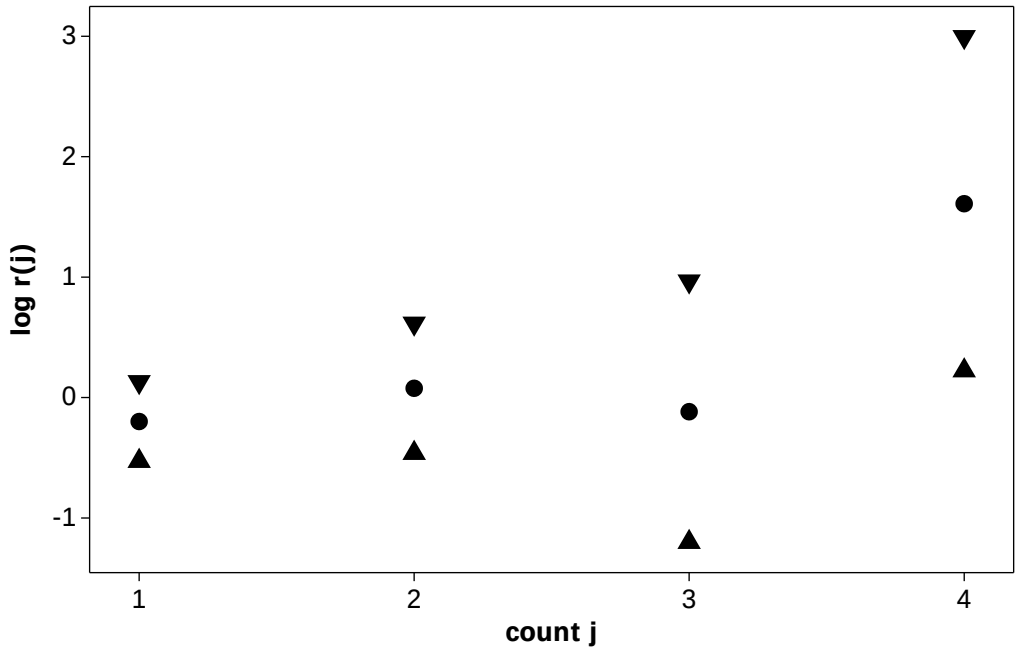


Figure: Ratio plot (on log-scale) for dystrophin data

Choosing cut-off k

- how to choose the cut-off k in $2 \leq k \leq m$
- suggestion: use χ^2 based upon the **truncated distribution**:

$$\chi^2(k) = \sum_{j=1}^{k} [f_j - n_k \times Po_+(j; \hat{\lambda}_k)]^2 / n_k \times Po_+(j; \hat{\lambda}_k)$$

where $n_k = f_1 + ... + f_k$ and

$$\lambda_k = \frac{2f_2 + 3f_3 + ... + kf_k}{f_1 + f_2 + ... + f_{k-1}}$$

- and $Po_+(j; \lambda) = Po(j; \lambda)/[Po(1; \lambda) + ... + Po(k; \lambda)]$

Choosing cut-off k: Dystrophin Data

k	$\chi^2(k)$	p-value	$\hat{\lambda}_j$	$\hat{N}_j$
2	0	1	0.82	347
3	0.53	0.47	0.90	334
4	0.58	0.75	0.89	334
5	12.00	0.01	0.98	323

other	$\hat{\lambda}$	$\hat{N}$
Turing	-	325
Plackett	0.96	321
MLE	0.99	315
MVUE	0.99	313

The ratio plot and missing data

- the ratio plot is also useful for missing data situations!

Cholera Outbreak in East-Pakistan

- A large study was undertaken in East Pakistan to monitor endemic cholera outbreaks, as it occurs with annual epidemic waves during the dry season (Mosley *et al.* 1972).
- Cholera is a serious disease that can lead to hospitalisation and death, but it can also occur as a mild infection.
- These inapparent infections occur with a high frequency (4-5 times more than clinical disease with classic cholera, and 20-40 times more with the El Tor strain according to Mosley *et al.* 1972) and people with this form are largely responsible for the spread of the disease given that people with the more serious form are generally quarantined.

Cholera Outbreak in East-Pakistan

- The number of hospitalizations can be used to measure the serious infections, but the mild infections are the unobserved population. The purpose of the study was to ultimately try to control cholera outbreaks in East Pakistan.
- Therefore as a method of controlling the disease, it was looked at establishing treatment centres.

Cholera Outbreak in East-Pakistan

- The Pakistan SEATO Cholera Research Laboratory set up a surveillance program. We focus here on the data from 1963 to 1966 when the study surveyed 132 villages, which contained 110,000 people.
- note that the repetition in this case arises in the village (cluster)

Table: Distribution of Observed Cholera Cases for Mosley-Study in East-Pakistan; the given number of $f_0 = 57$ was ignored since it is believed to contain a proportion of cholera-affect villages

f_0	f_1	f_2	f_3	f_4	$f_5 - f_9$	f_{10+}	n
-	20	21	8	7	11	8	75

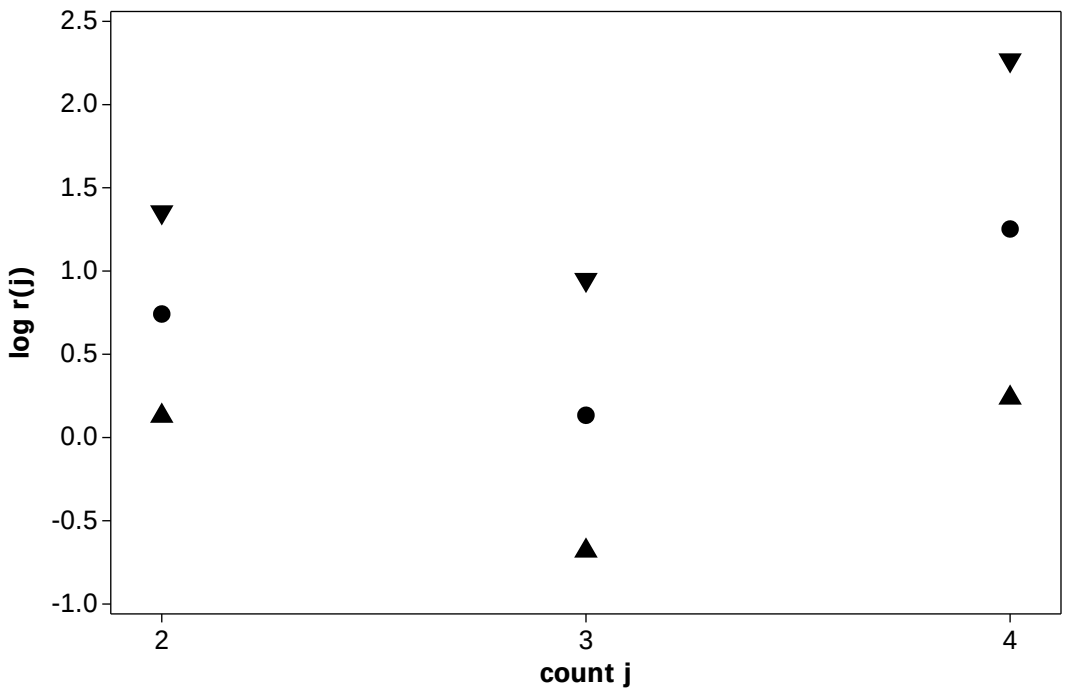


Figure: Ratio plot (on log-scale) for cholera data

The ratio plot and structured heterogeneity

- does the ratio plot provide any other important information?

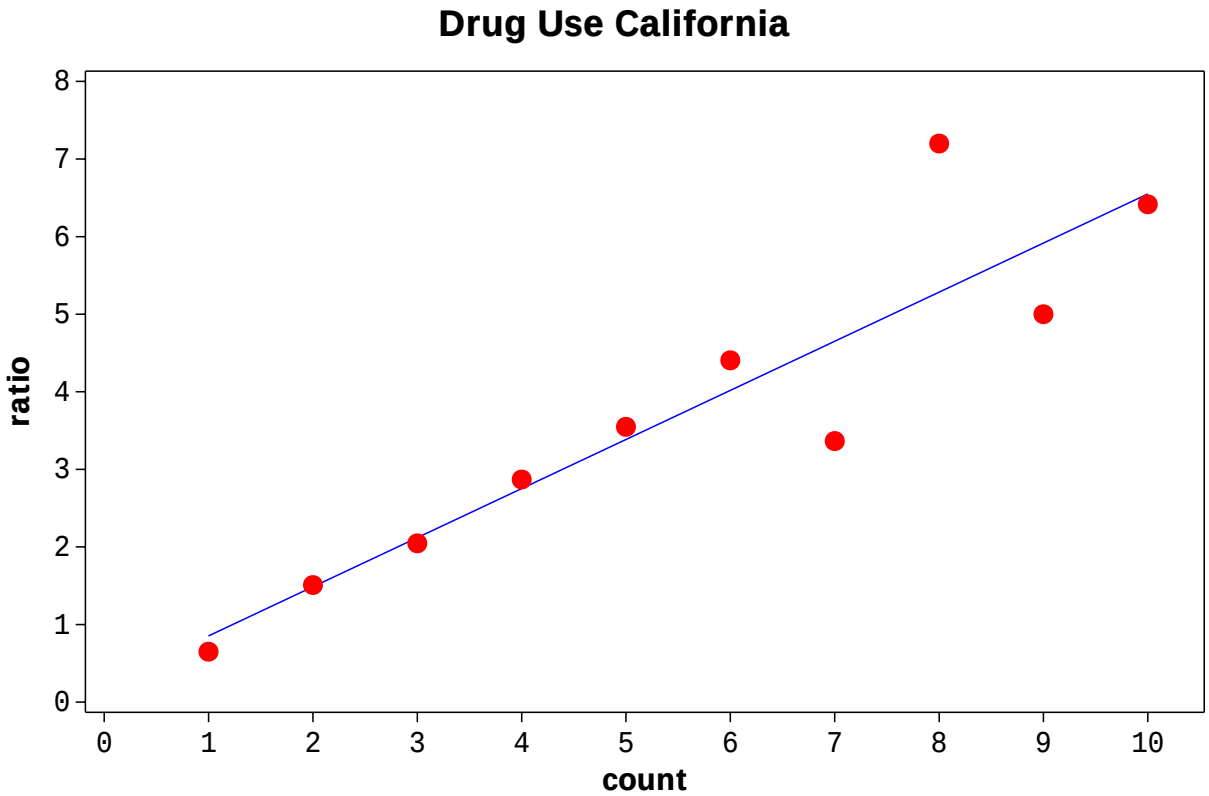
Hser's Data on Estimating Hidden Intravenous Drug Users in Los Angeles 1989

- 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


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

Structured Heterogeneity

if

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

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 $\lambda(t)$ is the Γ-density with parameters p 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 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: Gamma–density

if

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

indicates a pattern compatible with a **straight line**

- this indicates that it is reasonable to assume a **Gamma-distribution** for the heterogeneity distribution $\lambda(t)$ of the Poisson parameter
- remarkable since $\lambda(t)$ is a **latent** variable distribution and data from it are not directly observable

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 have that

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

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

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

structured heterogeneity: the generalised Turing

the **generalised Turing** estimator

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

is **consistent** since

$$\left(\frac{f_1}{S}\right) \rightarrow p^{k+1}$$

and

$$N_{GT} \rightarrow N(1 - p^k)/(1 - \left(p^{k+1}\right)^{\frac{k}{k+1}}) = N$$

Examples with N known

- a number of studies with known N
- are mentioned in a recent book

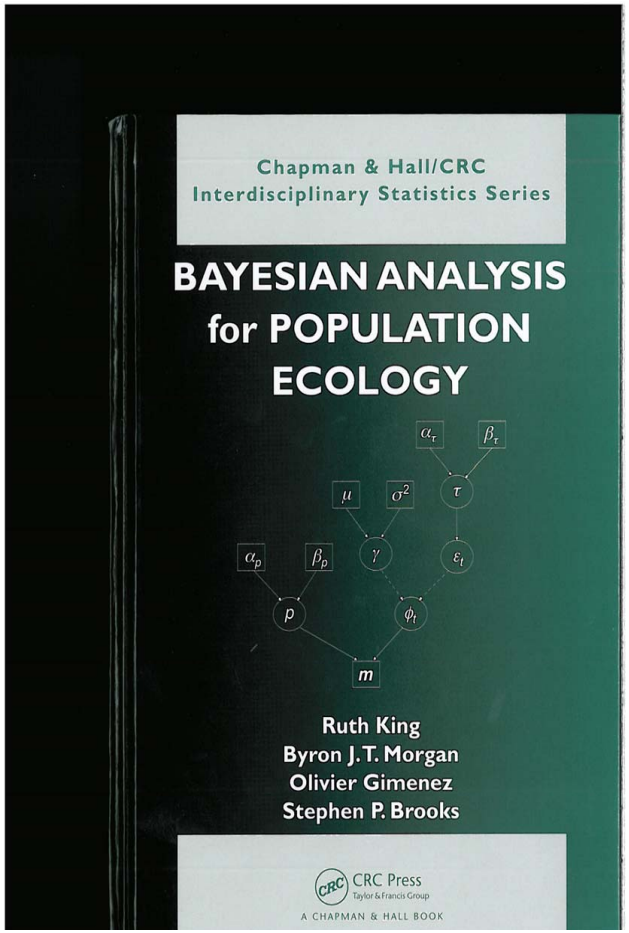
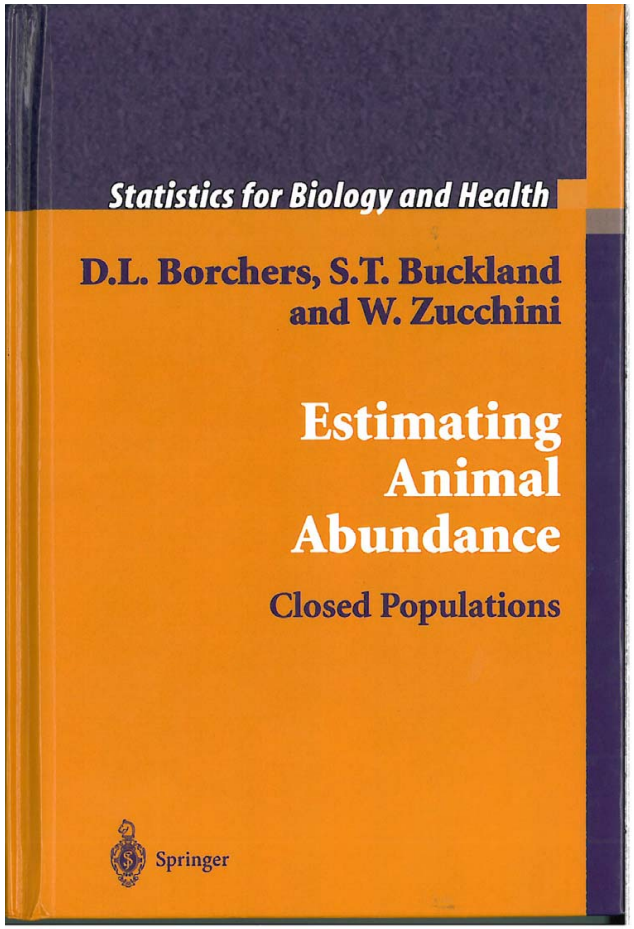


Table 2.1 Four Examples of Real Data Resulting from the Schnabel Census

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
Hares	25	22	13	5	1	2	-	-	-	-
Voles	18	15	8	6	5	-	-	-	-	-
Golf tees	46	28	21	13	23	14	6	11	-	-
Taxi-cab	142	81	49	7	3	1	0	0	0	0

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 work if we can check our estimates against the true population after estimating abundance, to see how well we did. This is impractical with real populations, so we will be using examples with artificial populations for illustration.

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

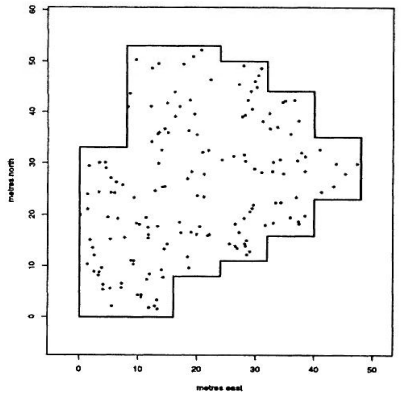


Figure 2.1. Example data, detected animals. Each dot represents a detected animal within the survey region. In all, $n = 162$ animals were detected.

different observers of a population of 250 groups (760 individuals) of golf tees, not plants, contrary to what we said in Chapter 1. The tees, of two colours, were placed in groups of between 1 and 8 in a survey region of 1,680 m^2, either exposed above the surrounding grass, or at least partly hidden by it. They were surveyed by the 1999 statistics honours class at the University of St Andrews,[1] Scotland, so while golf tees are clearly not animals (or plants), the survey was real, not simulated. We treat each group of golf tees as a single "animal", with size equal to the number of tees in the group; yellow tees are "male", green are "female"; tees exposed above the surrounding grass are classified as exposed ("exposure=1"), others as 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 refer to them as animal populations, and to their members as animals.

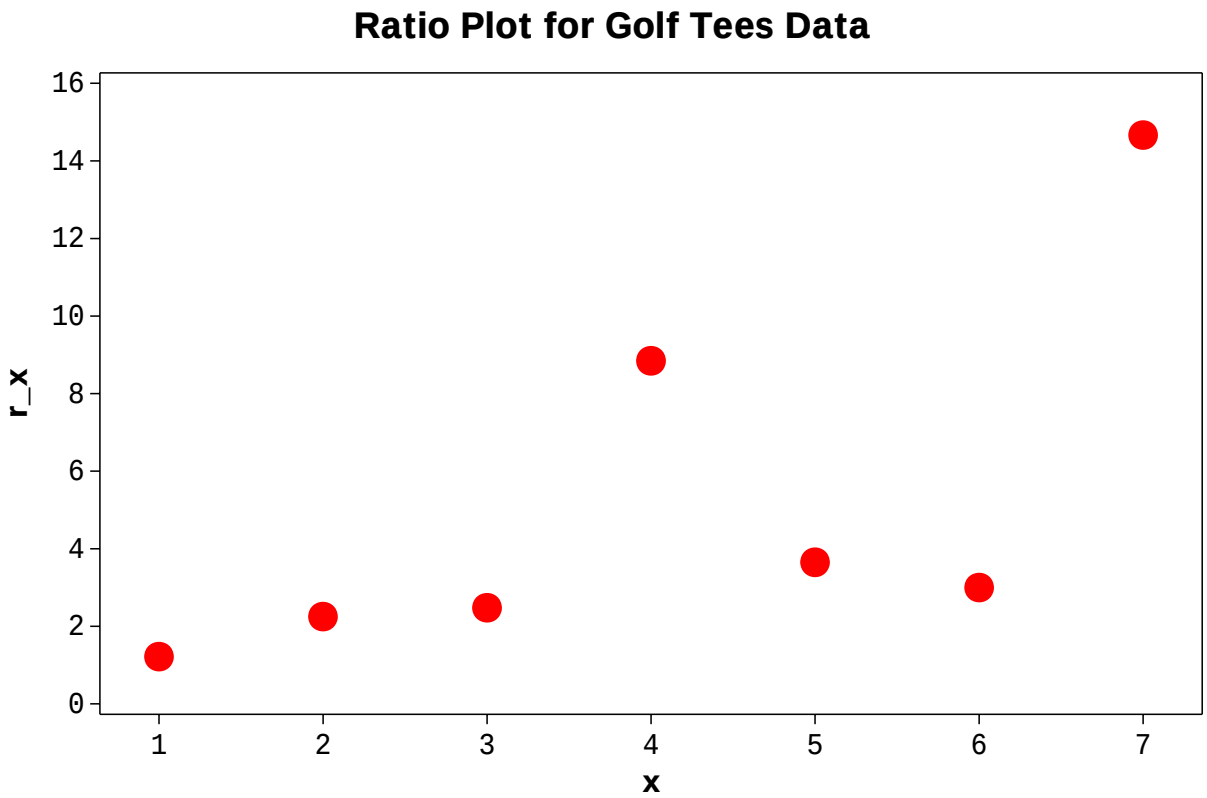
Figure 2.1 shows the locations of the animals detected by at least one observer on a survey of our first example population. A total of $n = 162$ animals were seen, but an unknown number were missed. We would like to use what was seen to answer the question: How many animals are there?

[1] We are grateful to Miguel Bernal for making these data available to us. They were collected by him as part of a Masters project at the University of St Andrews. St Andrews 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


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
$\widehat{k/(k+1)}$	0.52
$\hat{N}_{GT}$	224
$\hat{N}_T$	177
$\hat{N}_1(\hat{\lambda}_1)$	200 (1.22)
$\hat{N}_2(\hat{\lambda}_2)$	191 (1.61)
$\hat{N}_3(\hat{\lambda}_1)$	188 (1.80)

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

$$\text{decontaminated Turing: } \hat{N}_j = n + \frac{f_1}{\hat{\lambda}_j}$$

final comment

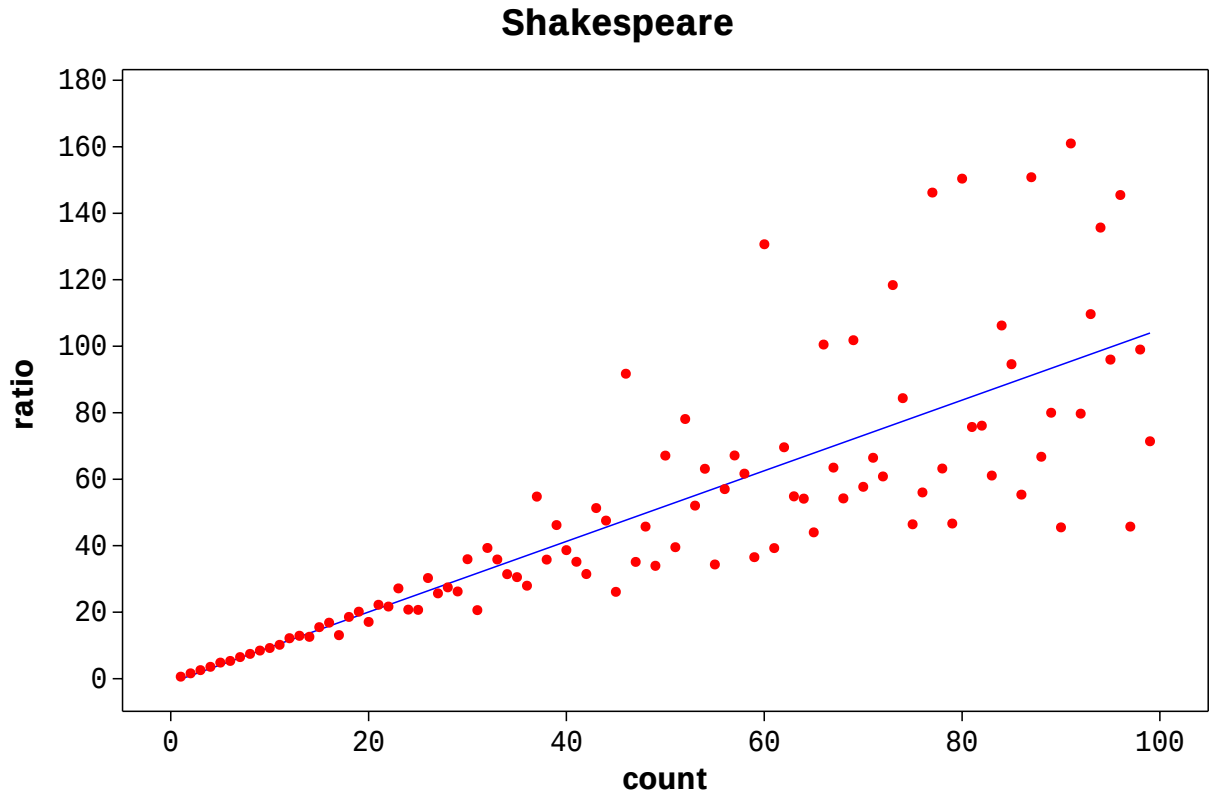
structured heterogeneity appears to be occur in in a variety of areas

- text analysis and language studies
- terroristic and criminal activity analysis

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


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

Drakos' Data on Estimating Hidden Transnational Terrorist Activity

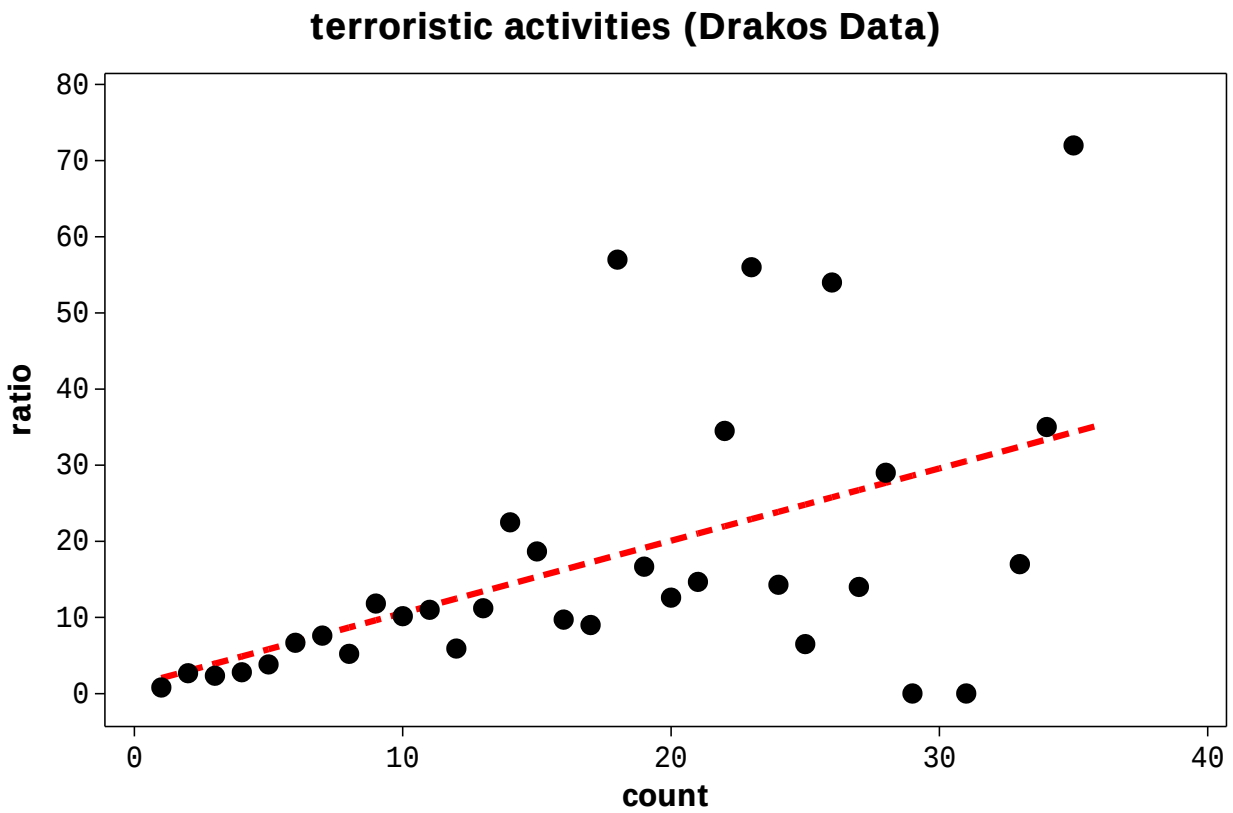
- data on terrorism are provided by various databases including RAND terrorism chronology, the terrorism indictment and DFI International research on terrorist organizations on 153 countries and the period 1985-1998
- terrorism is violence or the threat of violence, calculated to create an atmosphere of fear and alarm

Drakos' Data on Estimating Hidden Transnational Terrorist Activity

the frequency distribution (Drakos 2007) of the **count of transnational terrorist activity** Y_{it} in country i and year t:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	...	f_{136}	n
-	286	114	101	59	33	21	20	19	11	...	1	785

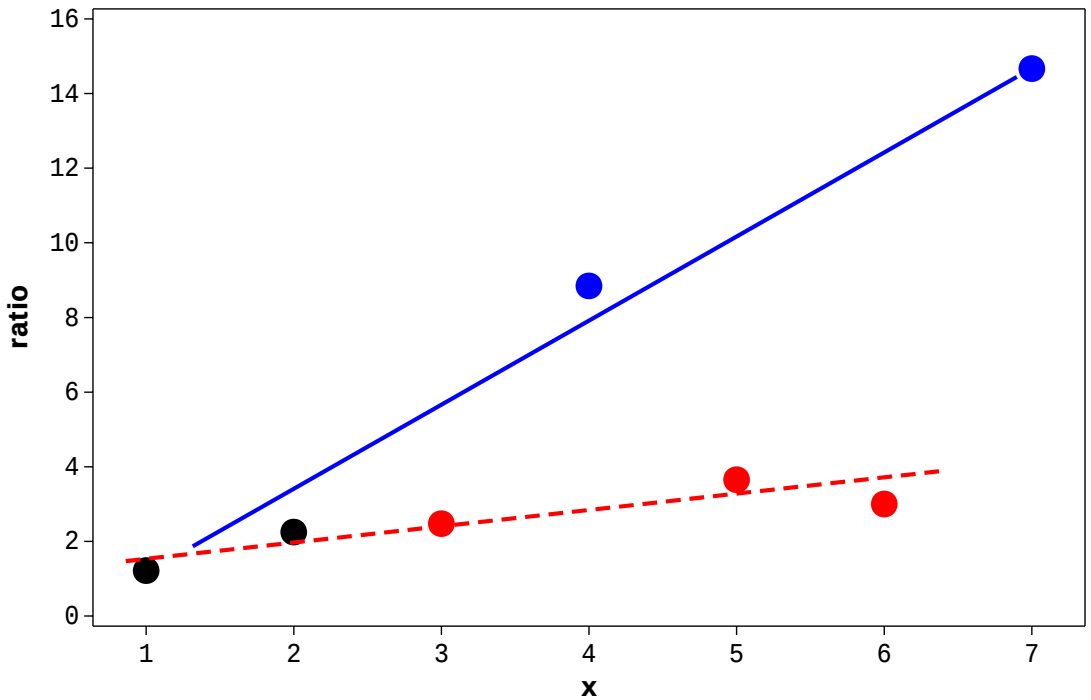
- similar to other data sets there is an $f_0 = 1,357$
- however it is thought that there is a hidden number of terrorist activities which is of interest to be estimated
- estimate f_0 the frequency of periods and countries **with unrecorded terrorist activities**


terroristic activities (Drakos Data)
ratio
count
0
10
20
30
40
0
10
20
30
40
50
60
70
80

Beyond Structured Heterogeneity

- after fitting **structured heterogeneity**:
- is there any **residual heterogeneity** left?

Ratio-Plot for the Golf Tees Study



beyond structured heterogeneity: generalizing Chao

let p_x be given by any **mixed power series**

$$p_x = \int a_x t^x \mu(t) \lambda(t) dt$$

where $\mu(t)$ is the normalizing function and a_x are the coefficients of the power series; recall we had

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

so that we are now looking at

$$p_x = \int_0^1 \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)} p^k (1-p)^x \lambda(p) dp$$

beyond structured heterogeneity: generalizing Chao

then, by Cauchy-Schwartz inequality

$$\frac{p_1/a_1}{p_0/a_0} \leq \frac{p_2/a_2}{p_1/a_1} \leq \frac{p_3/a_3}{p_2/a_2} \leq \dots$$

so that in particular

$$a_0 \frac{p_1^2/a_1^2}{p_2/a_2} \leq p_0$$

and **Chao's lower bound estimator** follows:

$$n + \frac{f_1^2/a_1^2}{f_2/a_2}.$$

Chao (1987, 1989) developed it for the **Poisson case:** $n + \frac{f_1^2}{2f_2}$.

structured heterogeneity: generalizing Chao

for the NB

$$a_x = \frac{\Gamma(k+x)}{\Gamma(x+1)\Gamma(k)}$$

and, in particular,

$$a_1 = \frac{\Gamma(k+1)}{\Gamma(2)\Gamma(k)} = k$$

$$a_2 = \frac{\Gamma(k+2)}{\Gamma(3)\Gamma(k)} = \frac{\Gamma(k+1)(k+1)}{2\Gamma(k)} = k(k+1)/2$$

so that

$$\frac{a_2}{a_1^2} = \frac{k(k+1)}{2k^2} = \frac{(k+1)}{2k}$$

structured heterogeneity: generalizing Chao

for the mixed power series Chao's lower bound is given by

$$n + \frac{f_1^2/a_1^2}{f_2/a_2} = n + \frac{a_2 f_1^2}{a_1^2 f_2}$$

and, in particular for the NB, using

$$\frac{a_2}{a_1^2} = \frac{k(k+1)}{2k^2} = \frac{(k+1)}{2k}$$

$$N_{GC} = n + \frac{k+1}{k}\frac{f_1^2}{2f_2}$$

is the **generalized Chao lower bound** in this case

structured heterogeneity: summary

note that both estimators involve

$$k' = k/(k+1)$$

- generalised Turing: $\hat{N}_{GT} = \frac{n}{1-\left(\frac{f_1}{s}\right)^{k'}}$
- generalised Chao: $\hat{N}_{GC} = n + \frac{1}{k'}\frac{f_1^2}{2f_2}$

Table: *Simulation using* $X \sim NB(p, k)$ *for* $N = 5,000$, $p = 0.5$, $k = 1$ *so that* $E(X) = 1$; *replication size* 1000

$N = 5,000$			
estimator	mean	trimmed mean	SD
$\hat{k}$	1.	1.	0.2
$\hat{N}_{GT}$	5035.7	5018.3	357.2
$\hat{N}_{GC}$	5042.8	5027.0	355.2
$\hat{N}_{T}$	3760.6	3759.9	81.4
$\hat{N}_{C}$	4085.9	4084.3	130.0

Table: *Simulation using* $X \sim NB(p, k)$ *for* $N = 20,000$, $p = 0.5$, $k = 1$ *so that* $E(X) = 1$; *replication size* 1000

$N = 20,000$			
estimator	mean	trimmed mean	SD
$\hat{k}$	1.0	1.0	0.1
$\hat{N}_{GT}$	19989	19974	690
$\hat{N}_{GC}$	19990	19979	670
$\hat{N}_{T}$	13333	13333	115
$\hat{N}_{C}$	14994	14995	194

Concluding with the Golf Tees Study: True $N = 250$

estimator	value
$\hat{k}$	1.07
$\widehat{k/(k+1)}$	0.52
$\hat{N}_{GT}$	224
$\hat{N}_{GC}$	235
$\hat{N}_{T}$	177
$\hat{N}_{C}$	200
$\hat{N}_1(\hat{\lambda}_1)$	200 (1.22)
$\hat{N}_2(\hat{\lambda}_2)$	191 (1.61)
$\hat{N}_3(\hat{\lambda}_1)$	188 (1.80)