Höhle, Held:

Bayesian Estimation of the Size of a Population

Sonderforschungsbereich 386, Paper 499 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Bayesian Estimation of the Size of a Population

Michael Höhle*
Department of Statistics
University of Munich
Germany

Leonhard Held
Biostatistics Unit, ISPM
University of Zurich
Switzerland

Abstract

We consider the following problem: estimate the size of a population marked with serial numbers after only a sample of the serial numbers has been observed. Its simplicity in formulation and the inviting possibilities of application make this estimation well suited for an undergraduate level probability course. Our contribution consists in a Bayesian treatment of the problem. For an improper uniform prior distribution, we show that the posterior mean and variance have nice closed form expressions and we demonstrate how to compute highest posterior density intervals. Maple and R code is provided on the authors' web-page to allow students to verify the theoretical results and experiment with data.

Keywords: Bayesian inference, Combinatorics, Hypergeometric functions, Maple, R.

1 INTRODUCTION

Assume we have a population of unknown size N which is labeled using the serial numbers $\{1, \ldots, N\}$, e.g. the number of participants in a marathon, taxis in a city or serial markings of a production. A random sample $(X_1, \ldots, X_n)$ of size $n \leq N$ is observed without replacement. Let $X = \max(X_1, \ldots, X_n)$ be the maximum of this sample. The task is to make inference for the population size N based on the observed value of X; throughout the text we shall call this the *SNP-problem*.

A fascinating historic report about the application of this type of population size estimation in economical intelligence during World War II is given by Ruggles and

*Department of Statistics, University of Munich, Ludwigstr. 33, 80539 München, Germany, Email: hoehle@stat.uni-muenchen.de

Brodie (1947). The corresponding statistical treatment of the topic in a frequentist perspective is given by Goodman (1952, 1954) who derive unbiased minimum variance estimators and confidence intervals. Johnson (1994) treats the SNP-problem in a pedagogical setting and a search on courses pages on the Internet reveals quite a few pages containing relevant lab exercises. This characteristic as a nice application of statistics is also reflected by popular science literature (Matthews 1998) giving mention to the problem.

Our contribution is to cast the SNP-problem into a Bayesian framework. Based on prior information about N the posterior distribution of N given the observed value $X = x$ is calculated using Bayes formula. This has to some degree been done earlier by Roberts (1967), however his derivations are done in the context of stopping rules where the sample size n is not fixed in advance. Our derivations are more straightforward – a thorough Bayesian treatment is given including computation of highest posterior density intervals and different cases of prior distributions. All derivations are exact, which avoids the usual problems (e.g. convergence, time-usage) of simulation based Bayesian inference.

From a pedagogical point of view the application has, similarly to a Bayesian version of the capture-recapture experiment, the advantage that all involved distributions are discrete thus not requiring any knowledge about continuous distributions. In an undergraduate probability course context it would be possible to let the derivations be part of an exercise on working with symbolic algebra packages like Maple (MapleSoft 2004) or an exercise in implementing simple solutions in e.g. R (R Development Core Team 2006).

2 MATHEMATICAL SETTING

Methods to make inference about N discussed in the literature cover maximum likelihood estimation, finding an unbiased estimator, bounding the probability to overshoot the maximum or calculating posterior densities (Goodman 1952, 1954; Roberts 1967; Gum et al. 2000). Independent of the framework, the likelihood for observing $X = x$ in a sample of size n plays a central role. It is given as

$$P(x|N) = P(X = x|N) = \frac{\binom{x-1}{n-1}}{\binom{N}{n}}, \text{ if } \quad n \leq x \leq N, \text{and 0 otherwise.} \quad (1)$$

In Goodman (1952) the unbiased estimator of N with minimum variance is shown to be

$$\hat{N} = \frac{n+1}{n}x - 1, \quad (2)$$

and a $1-\alpha$ confidence interval for N is given as $[x, x+1, \ldots, N_u]$, where N_u is the largest integer satisfying $(x)_n/(N_u)_n \geq \alpha$. Here we have used $(x)_n$ to denote the so called *falling factorial*, i.e.

$$(x)_n = x!/(x-n)! = x(x-1)\cdots(x-n+1).$$

Turning to a Bayesian framework, we assume a suitable prior distribution for N, after which the posterior distribution can be calculated via Bayes' theorem, i.e.

$$P(N|x) = \frac{P(x|N)P(N)}{P(x)} = \frac{P(x|N)P(N)}{\sum_{N'=x}^{\infty} P(x|N')P(N')} \tag{3}$$

for $x \leq N < \infty$ and zero otherwise. Note that (1) implies, that only values of N with $N \geq x$ have positive posterior probability. We therefore let the sum in the denominator start at x, although the prior may be positive also for smaller values of N. In particular, it is irrelevant if the support of the prior distribution starts at 0 or at 1. For ease of presentation, we let the prior start at 0.

Various choices can be imagined as prior distribution for N:

- An improper uniform prior on all positive integers, i.e. $P(N) \propto 1$ for $N = 0, \ldots, \infty$.
- A proper uniform distribution with an upper limit k for N, i.e. $P(N) = 1/(k+1)$ for $0 \leq N \leq k$ and zero otherwise.
- A Geometric, Poisson or Negative Binomial distribution.

3 POSTERIOR PROPERTIES UNDER AN IMPROPER UNIFORM PRIOR

The case of an improper uniform prior is of particular interest. Using theory about infinite binomial series and hypergeometric functions it can be shown that, for $n > 1$, the posterior distribution is then given as the discrete distribution

$$P(N|x) = \frac{n-1}{x}\binom{x}{n}\binom{N}{n}^{-1}, \text{ if } N = x, x+1, x+2, \ldots, \tag{4}$$

and 0 otherwise. The above distribution is thus a shifted *factorial distribution* (Marlow 1965), i.e. $N-x$ follows a factorial distribution with parameters x and n. Note that the posterior distribution is improper for $n = 1$, which we will show below.

It is easy to show that the Maximum Likelihood estimate, i.e. the posterior mode under the improper uniform prior, is $\hat{N}_{ML} = x$, the smallest N-value of

the posterior distribution with positive posterior probability. However, it is perhaps less known that simple formulae also exist for the posterior mean and variance:

$$E(N|x) = \frac{n-1}{n-2} \cdot (x-1) \quad \text{for } n > 2 \text{ and} \tag{5}$$

$$\mathrm{Var}(N|x) = \frac{(n-1)(x-1)(x-n+1)}{(n-2)^2(n-3)} \quad \text{for } n > 3. \tag{6}$$

In the following these results are derived analytically. Results for the posterior in case of various proper priors are given in Section 5.

3.1 Analytic derivations using binomial sums

Under the assumption of an improper uniform prior, the posterior distribution (3) simplifies to

$$P(N|x) = \frac{\binom{N}{n}^{-1}}{\sum_{N'=x}^{\infty} \binom{N'}{n}^{-1}} \tag{7}$$

for $x \leq N < \infty$ and zero otherwise. The main problem here is to simplify the denominator, which is an infinite sum of inverse binomial coefficients $\binom{n}{k}$. This is a case for classic combinatorial theory, but usually the forms treated in the literature are of the type where n is fixed and k varies, e.g. sums like $\sum_{k=0}^{\infty} \binom{n}{k}^{-1}$. Classical combinatorial theory has used various ad-hoc, problem-specific and ingenious ways to come up with the right solution to finite and infinite binomial sums. A unified framework for handling binomial sums which has become increasingly popular in the literature are hypergeometric functions.

Barne's extended hypergeometric function form is given as

$${}_pF_q\left[\, a_1, \ldots, a_p; b_1, \ldots, b_q; z \right] = \sum_{i=0}^{\infty} \frac{(a_1)^i \cdots (a_p)^i}{(b_1)^i \cdots (b_q)^i} \frac{z^i}{i!}, \tag{8}$$

where $(x)^i$ are the *rising factorials* (also called *Pochhammer symbols*) defined as

$$(x)^i = \frac{(x+i-1)!}{(x-1)!} = \frac{\Gamma(x+i)}{x} = (x+i-1) \cdot \ldots \cdot x.$$

Note that x can be negative in the factorials and thus a general definition of the factorial and the Gamma function is required. Examples of the above definitions are $(1)^z = z!$ and $\exp(z) = {}_0F_0[z]$.

There are various results about hypergeometric functions, the most important with respect to our problem is *Gauss' hypergeometric function*,

$${}_2F_1\left[\, a, b; c; 1 \right] = \sum_{i=0}^{\infty} \frac{(a)^i (b)^i}{(c)^i} \frac{1}{i!} = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)},$$

and originates as the solution to the so called hypergeometric differential equation (Wolfram Research 2004b).

To compute the sum in the denominator of (7), we now proceed as follows:

$$\begin{aligned}
\sum_{N'=x}^{\infty} \binom{N'}{n}^{-1} &= \sum_{i=0}^{\infty} \binom{x+i}{n}^{-1} \\
&= \sum_{i=0}^{\infty} \frac{n!(x+i-n)!}{(x+i)!} \\
&= \binom{x}{n}^{-1} \sum_{i=0}^{\infty} \frac{x!}{(x+i)!} \frac{(x+i-n)!}{(x-n)!} \\
&= \binom{x}{n}^{-1} \sum_{i=0}^{\infty} \frac{(x-n+1)^i}{(x+1)^i} \\
&= \binom{x}{n}^{-1} \sum_{i=0}^{\infty} \frac{(x-n+1)^i (1)^i}{(x+1)^i} \frac{1}{i!} \\
&= \binom{x}{n}^{-1} {}_2F_1\left[1+x-n,\ 1; x+1; 1\right] \\
&= \binom{x}{n}^{-1} \frac{\Gamma(x+1)\Gamma(n-1)}{\Gamma(n)\Gamma(x)} \\
&= \binom{x}{n}^{-1} \frac{x}{n-1}, \qquad (9)
\end{aligned}$$

where we have used $\Gamma(x+1)/\Gamma(x) = x$ at the very end. The form of the posterior distribution in (4) follows immediately.

Clearly this derivation can only be valid for $n \geq 2$. In fact, it is easy to see that for $n = 1$ the posterior is improper, because the sum in the denominator of (7)

$$\sum_{N'=x}^{\infty} \binom{N'}{1}^{-1} = \frac{1}{x} + \frac{1}{x+1} + \frac{1}{x+2} + \ldots$$

is infinite.

To determine the posterior mean $E(N|x)$ and the posterior variance $\mathrm{Var}(N|x)$, we exploit results on the factorial distribution. A discrete random variable Z follows a $\mathrm{Fact}(n, m)$ distribution (factorial distribution with parameters n and m), if its probability function is given by

$$P(Z = z) = (n-1)\frac{(m-1)!}{(m-n)!}\frac{(m+z-n)!}{(m+z)!}, \quad z = 0, 1, 2, \ldots$$

Marlow (1965) showed that the mean and variance of Z are

$$\begin{aligned} E(Z) &= \frac{m-n+1}{n-2}, \quad n>2 \\ \mathrm{Var}(Z) &= \frac{(n-1)\,(m-1)\,(m-n+1)}{(n-3)\,(n-2)^2}, \quad n>3. \end{aligned}$$

Equation (4) shows that that $N-x$ is $\mathrm{Fact}(n,x)$-distributed, hence equations (5) and (6) follow immediately.

3.2 Posterior quantiles and HPD-intervals

We now turn to the problem of calculating posterior quantiles. Let N_q be the q-quantile of the posterior, i.e. the smallest integer that fulfills

$$\sum_{N'=x}^{N_q} P(N'|x) = \sum_{N'=x}^{N_q} \frac{n-1}{x}\binom{x}{n}\binom{N'}{n}^{-1} \geq q$$

An equivalent definition is via

$$\sum_{N'=N_q+1}^{\infty} P(N'|x) = \sum_{N'=N_q+1}^{\infty} \frac{n-1}{x}\binom{x}{n}\binom{N'}{n}^{-1} \leq 1-q \tag{10}$$

Note that the infinite sum in (10) can be calculated similar to above, since

$$\sum_{N'=N_q+1}^{\infty} \binom{N'}{n}^{-1} = \binom{N_q+1}{n}^{-1} \frac{N_q+1}{n-1},$$

compare with (9). It follows that

$$\sum_{N'=N_q+1}^{\infty} P(N'|x) = \frac{(x-1)!}{(x-n)!}\frac{(N_q-n+1)!}{N_q!}$$

This gives us a way to calculate the median and any posterior quantile directly, without explicitly summing up the posterior distribution. The polynomial to be solved for real $\bar{N}_q$ is

$$\left(\bar{N}_q \cdot (\bar{N}_q-1) \cdot \ldots \cdot (\bar{N}_q-n+2)\right) - \frac{(x-1)!}{(1-q)(x-n)!} = 0, \tag{11}$$

and $N_q = \lceil \bar{N}_q \rceil$, the smallest integer larger than $\bar{N}_q$. Because the mode of the posterior distribution is always at x and the posterior distribution is monotone decreasing for increasing N, the computation of highest posterior density (HPD) intervals

reduces to the computation of quantiles of the posterior, i.e. the HPD-interval of level q equals $[x, x+1, \ldots, N_q]$.

For $n = 2$, the solution to (10) reduces to a linear form, i.e.

$$N_q = \lceil (x-1)/(1-q) \rceil,$$

very simple to compute. In particular the posterior median is simply $2(x-1)$.

To obtain the q-quantile N_q in case of arbitrary n, Equation (11) is recognized as a polynomial in $\bar{N}_q$ of degree $n-1$ with coefficients $c_0, \ldots, c_{n-1}$, where

$$c_0 = -\frac{(x-1)!}{(1-q)(x-n)!}$$

and $c_1, \ldots, c_{n-1}$ are determined as the coefficients of the falling factorial $(\bar{N}_q)_{n-1}$. An exact method to find the roots of this falling factorial polynomial is to exploit the following combinatorial identity (Wolfram Research 2004c)

$$(x)_n = \sum_{k=0}^{\infty} \begin{bmatrix} n \\ k \end{bmatrix} (-1)^{n-k} x^k.$$

Here, $\begin{bmatrix} n \\ k \end{bmatrix}$ is the *unsigned Stirling number of the first kind*, i.e. the number of permutations of the set $\{1, \ldots, n\}$ having k cycles. The six permutations when $n = 3$ are $\{1,2,3\}, \{1,3,2\}, \{2,1,3\}, \{2,3,1\}, \{3,1,2\}, \{3,2,1\}$ and the corresponding cyclic decompositions are $(1)(2)(3), (1)(23), (3)(12), (123), (132), (2)(13)$, thus $\begin{bmatrix} 3 \\ 3 \end{bmatrix} = 1, \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 3, \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 2$. Stirling numbers of the first kind can also be computed by the following recursion ($k, n \geq 1$),

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1, \ \begin{bmatrix} 0 \\ k \end{bmatrix} = 0, \ \begin{bmatrix} n \\ 0 \end{bmatrix} = 0, \text{ and } \begin{bmatrix} n+1 \\ k \end{bmatrix} = n \begin{bmatrix} n \\ k \end{bmatrix} + \begin{bmatrix} n \\ k-1 \end{bmatrix}.$$

The above identity provides the coefficients of $c_1, \ldots, c_{n-1}$ and thus allows us to solve the polynomial in (11) by any standard polynomial root finding routine. The N_q solution of interest is then the ceiling of the largest real root.

Alternatively, one obtains an approximate solution by using the following approximation to the falling factorial

$$a \cdot (a-1) \cdot \ldots \cdot (a-n+2) \approx \left(a - \frac{n-2}{2} \right)^{n-1},$$

and hence

$$N_q \approx \left\lceil \left(\frac{(x-1)!}{(x-n)!} / (1-q) \right)^{\frac{1}{n-1}} + \frac{n-2}{2} \right\rceil,$$

which can be further approximated to

$$N_q \approx \left\lceil (x - \frac{n}{2})(1-q)^{-\frac{1}{n-1}} + \frac{n-2}{2} \right\rceil,$$

which shows that N_q is approximately linear in x. Both formulae are exact for $n = 2$. For higher n the approximative method is surprisingly good whereas the polynomial root finding routing for the exact method can run into numerical problems once n is large ($n \approx 50$).

4 EXAMPLES

In July 1885 the Irish Munster Bank collapsed and historically interested economists are discussing today whether the Bank of Ireland at that time should have exercised its responsibility as lender of last resort and saved the bank (Gráda 2001). As part of the investigation by Gráda (2001) the size of the 41 individual branches of the bank needed to be estimated at the time of closure. The only data available today for this purpose are lists of the account numbers of depositors with dividend payments still unclaimed in October 1888. For the head office in Cork, 48 account holders (with account numbers ranging from 63 to 1812) were still owned money. Using $n = 48$ and $x = 1812$ we shall try to estimate the total number of customers at the head office in Cork.

Using the frequentist approach of Goodman (1952) we obtain $\hat{N} = 1848.75$ with an exact 95% confidence interval of $[1812, \ldots, 1926]$. By assuming an improper uniform prior and inserting into the derived formulas we obtain a posterior mean of 1850.37, a posterior median of 1838, and a 95% HPD interval of $[1812, \ldots, 1929]$.

Figure 1 shows in (a) the posterior distribution of $N|x$. Note the characteristic shape of the shifted factorial distribution: values below x are zero, the mode is at x and the probabilities are strictly decreasing after x. Part (b) of the figure shows the posterior mean $E(N|x)$ and a 95% HPD as a function of the observed maximum x. Here, the linearity in x becomes obvious.

The above estimations and the study of historic sources lead Gráda (2001) to the conclusion that the collapse of the Muster Bank was caused by a combination of over-generous credit, insider lending and plain fraud. Therefore, the Bank of Ireland did fulfill its responsibility by not reacting.

Another example of the SNP-Problem is the study of production numbers and capacity of an industrial production. For example a commercial company might be interested in assessing relevant numbers of a competitor. An example of this is the study of German production numbers during World War II: Using statistical

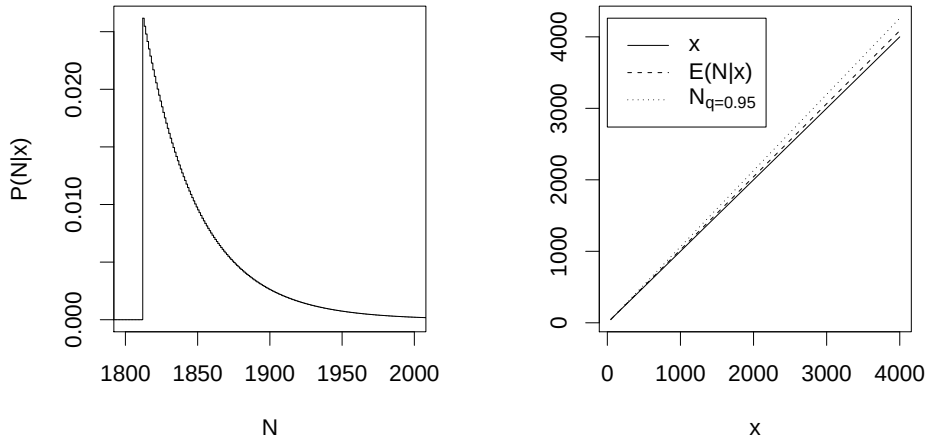

Figure 1: (a) Posterior distribution of $N|x$ and (b) posterior mean $E(N|x)$ together with upper and lower limit of an 95% HPD interval as a function of the observed maximum x.

techniques the Economic Warfare Division of the American Embassy in London was able to provide much more accurate estimates of the German Mark I-IV tanks than the often wildly exaggerated numbers by allied intelligence sources (Ruggles and Brodie 1947).

Goodman (1954) gives an example of him counting the serial numbered office furniture (desks, bookcases, etc.) at the Division of Social Sciences, University of Chicago. Here, the maximum serial number of 31 random selected items was 2787. With these numbers our methods obtain that the posterior mean in case of an improper uniform prior is 2882.07, the posterior median is 2851 and an exact 95% HPD interval is $[2787, 3078]$. As an aside: Goodman reports that, through hard work by his secretary, Mrs. Denny, he was able to get hold of records showing that the true number of items serially numbered was 2885.

5 DISCUSSION

We have given a Bayesian treatment of SNP-Problem: the task of inferring the population size in a sequentially numbered population. When using an improper uniform distribution nice formulas were obtained for the posterior expectation and variance, similarly it was easy to compute highest posterior density intervals. The application has the nice feature that only discrete distributions need to be known

while still making allowance for a full Bayesian treatment of an estimation problem. The SNP-problem offers many opportunities for letting the students perform their own derivations and experiments. On the page `http://www.stat.uni-muenchen.de/~hoehle/software/bayespopsize/` a selection of R and Maple source code can be found which might provide inspiration.

For the more involved priors the resulting equations are not as nice as for the improper prior. For example, if a proper uniform prior is used on the interval $1, \ldots, k$ the posterior together with its mean and variance are available in closed form expression through the use of Gosper's algorithm (Wolfram Research 2004a). The above mentioned Maple code contains the necessary derivations.

In case of a negative binomial prior with mean $r(1-p)/p$ and variance $r(1-p)/p^2$ the posterior, its mean and variance can also be derived using hypergeometric functions. We only state the posterior distribution and refer to the Maple code for further information.

$$P(N|x) = \frac{\binom{N+r-1}{r-1}\binom{x}{n}(1-p)^{N-x}}{\binom{N}{n}\binom{x+r-1}{r-1}{}_3F_2\left[1, 1+x-n, x+r; 1+x, 1+x; 1-p\right]}$$

Returning to the example of the Munster Bank, assume that one of the former employees at the head office in Cork was quoted to believe that the number of customers with dividend payments still unclaimed to be in the neighbourhood of 2,000. We use this information in a prior elicitation: let the parameters of a negative binomial prior be such that $E(N) = 2000$ and $P(N \leq 2500) = 0.95$. This results in $r = 48.52$ and $p = 0.02368$. Figure 2 illustrates the prior distribution and the corresponding posterior with mean 1851.72 and variance 1621.61 when $x = 1812$ and $n = 48$. Compared to the posterior mean and variance of 1850.37 and 1237.13 under an improper uniform prior, it appears that the prior distribution has little influence on the posterior distribution.

Several extensions of the estimation problem can be imagined. For example the case where the serial numbers are known to lie in the interval $\{M, \ldots, N\}$ with both M and N unknown. Goodman (1952, 1954) and Roberts (1967) contain a discussion of this topic in a frequentist or Bayesian setting. Goodman (1954) also contains a discussion on performing approximative inference for the SNP-Problem in case of large N by considering the sampling to occur uniformly distributed from the continuous interval $[0, N]$.

6 REFERENCES

Goodman, L. A. (1952), "Serial number analysis," *Journal of the American Statistical Association*, 47, 622–634.

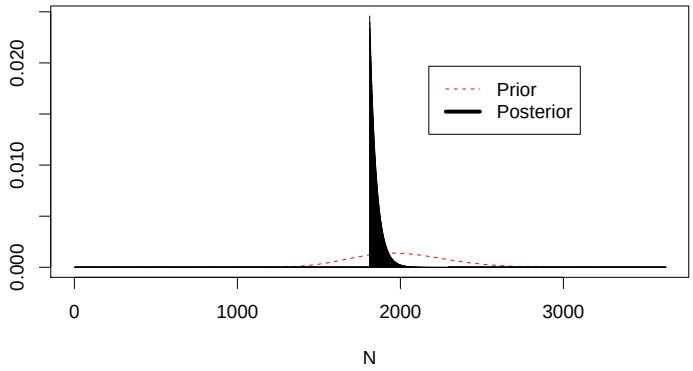

Figure 2: Elicitated negative binomial prior and corresponding posterior distribution in the Munster Bank example.

Goodman, L. A. (1954), "Some Practical Techniques in Serial Number Analysis," *Journal of the American Statistical Association*, 49, 97–112.

Gráda, C. O. (2001), "Should the Munster Bank have been saved?" Tech. Rep. WP01/15, Department of Economics, University of Dublin.

Gum, B., Lipton, R., LaPaugh, A., and Fich, F. (2000), "Estimating the maximum," in *Proceedings of the Tenth SIAM Conference on Discrete Mathematics*.

Johnson, R. (1994), "Estimating the size of a population," *Teaching Statistics*, 16, 50–52.

MapleSoft (2004), "Maple v9.5," `http://www.maplesoft.com/products/maple/`.

Marlow, W. (1965), "Factorial Distributions," *The Annals of Mathematical Statistics*, 36, 1066–1068.

Matthews, R. (1998), "Hidden truths," *New Scientist*, 158, 28.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Roberts, H. V. (1967), "Informative Stopping Rules and Inferences about Population Size," *Journal of the American Statistical Association*, 62, 763–775.

Ruggles, R. and Brodie, H. (1947), "An empirical approach to economic intelligence in World War II," *Journal of the American Statistical Association*, 42, 72–91.

Wolfram Research (2004a), "Gosper's Algorithm," http://mathworld.wolfram.com/HypergeometricDifferentialEquation.html.

— (2004b), "Hypergeometric Differential Equation," http://mathworld.wolfram.com/HypergeometricDifferentialEquation.html.

— (2004c), "Stirling Number of the First Kind," http://mathworld.wolfram.com/StirlingNumberoftheFirstKind.html.