The difference between the hypergeometric and the binomial distribution.

H. R. Künsch
ETH Zurich, Seminar for Statistics
CH-8092 Zurich, Switzerland

May 1998

1 Results

Consider drawing a sample of size n from an urn containing K white and $N-K$ black balls. The number of white balls in the sample has a hypergeometric distribution denoted $Q(n;K,N)$ if we sample without replacement and a binomial distribution denoted $P(n;p)$ with $p=K/N$ if we sample with replacement. Since it is easier to work with the binomial distribution than with the hypergeometric, we would like to replace latter by the former. It is straightforward to show that for fixed n

$$\|Q(n;K,N)-P(n;K/N)\| \longrightarrow 0 \quad (N\to\infty,\ K/N\to p)$$

in any norm. Here we are interested in obtaining explicit bounds for the total variation norm

$$\|Q(n;K,N)-P(n;p)\|_{TV}=\sup_{A\subseteq\{0,1,\dots n\}}|Q(n;K,N)(A)-P(n;p)(A)|\,.$$

Using elementary arguments, Freedman (1977) gave tight bounds from above and below for the total variation norm between sampling with and without replacement. His results show that this norm is small iff $\frac{n^2}{N}$ is small. But in many typical situations from survey sampling, $\frac{n^2}{N}$ is not small. On the other hand, we are interested only in events involving the number of white balls and so the difference can be substantially smaller.

Using Stein's method (see Barbour et al., 1992), Ehm (1991) obtained the following inequalities

$$\frac{1}{124}\frac{n-1}{N-1}\le\|Q(n;K,N)-P(n;p)\|_{TV}\le\frac{n-1}{N-1}$$

provided $npq\ge 1$ ($q=1-p$). This result does not seem to be widely known. For instance it is not quoted in Johnson et al. (1992) who list various approximations and bounds for the hypergeometric distribution in Section 6.5. Ehm's (1991) result is based on a result of Vatutin and Mikhailov (1982, Lemma 1) which says that a hypergeometric random variable can be represented as a sum of independent, but not identically distributed indicators.

The purpose of this note is to show that we can obtain Ehm's (1991) result without using the representation of Vatutin and Mikhailov (1982). We work with the obvious representation by a sum of identically distibuted, but dependent indicators and use the coupling method (see Barbour et al., 1992, Chapter 2). In this way the proofs become a

little simpler and we can slightly improve the lower bound. After completing this work, we became aware of the paper by Soon (1996) where the same coupling has been used. But he does not give a lower bound.

Theorem If $npq \geq 1$

$$\frac{1}{28}\frac{n-1}{N-1} \leq \|Q(n;K,N) - P(n;p)\|_{TV} \leq \frac{n-1}{N-1}.$$

Unfortunately the gap between the two bounds is still too large to decide for instance whether the rule of thumb $\frac{n}{N} < 0.1$ (see e.g. Johnson et al. (1992), p. 257) is justified or not. We have calculated the ratio

$$\frac{N-1}{n-1}\|Q(n;K,N) - P(n;p)\|_{TV}$$

for $(n,N) = (1000, 10'000),\ (1000, 5000),\ (500, 5000)$ and $p = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$ with the S-Plus software (Becker, Chambers, Wilks, 1988). In order to check the software we also did the calculations with R (Ihaka and Gentleman, 1996). The results were the same and this ratio is practically constant and equal to 0.26 (it varied between 0.255 and 0.27). Even in the extreme case $N = 2500, n = 50, p = 0.02$ the ratio is still 0.28. The following result brings the theoretical lower bound closer to the empirical bound at least in the limit when n is large.

Theorem If $n \to \infty$, $N \to \infty$ and $K/N \to p \in (0,1)$, then

$$\liminf \frac{N-1}{n-1}\|Q(n;K,N) - P(n;p)\|_{TV} \geq 0.115.$$

2 Proofs

We begin by introducing some notation. Let f denote a function defined on $\{0, 1, \ldots n\}$. We denote the expectation of f with respect to $Q(n;K,N)$ by $Q(n;K,N)[f]$ (and similarly for expectations with respect to $P(n;p)$). Furthermore we introduce the following two operators

$$\Delta f(k) = f(k+1) - f(k) \quad (k = 0, 1, \ldots n-1),$$
$$Tf(k) = p(n-k)f(k) - qkf(k-1) \quad (k = 0, 1, \ldots n).$$

Note that our definition of Δf differs from the one used by Ehm (1991) and that we do not need to define f at $k = -1$ to obtain $Tf(0)$. The following Lemma is the key

Lemma For any f

$$\begin{aligned} P(n;p)[Tf] &= 0, \\ Q(n;K,N)[Tf] &= Q(n-2;K-1,N-2)[\Delta f]\frac{n-1}{N-1}npq. \end{aligned}$$

Proof of Lemma 1: The first equality follows from the well known recursion

$$q(k+1)P(n;p)[k+1] = p(n-k)P(n;p)[k].$$

For the second equality, define the random variables

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ ball is white} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \dots n$ and let

$$S = \sum_{i=1}^{n} X_i, \quad S_i = \sum_{j \neq i} X_j.$$

Then

$$\begin{aligned} Q(n;K,N)[Tf] &= E[p(n-S)f(S) - qSf(S-1)] \\ &= p\sum_{i=1}^{n} E[(1-X_i)f(S_i+X_i)] - q\sum_{i=1}^{n} E[X_i f(S_i+X_i-1)] \\ &= pq\sum_{i=1}^{n}(E[f(S_i)|X_i=0] - E[f(S_i)|X_i=1]). \end{aligned}$$

But conditional on $X_i = 0$, the distribution of S_i is $Q(n-1;K,N-1)$, and conditional on $X_i = 1$, it is $Q(n-1;K-1,N-1)$. Hence we can realize these two conditional distributions simultaneously on the same probability space as follows: Draw a sample of size $n-1$ without replacement from an urn with $K-1$ white, one grey and $N-K-1$ black balls. Then the number of white balls in the sample is $Q(n-1;K-1,N-1)$-distributed and the number of white and grey balls is $Q(n-1;K,N-1)$-distributed. Denoting the event that the sample contains exactly k white and the grey ball by B_k, we thus obtain

$$\begin{aligned} Q(n;K,N)[Tf] &= npq\sum_{k=0}^{n-2}(f(k+1)-f(k))P[B_k] \\ &= npq\sum_{k=0}^{n-2}\Delta f(k)\binom{K-1}{k}\binom{N-K-1}{n-2-k}\binom{N-1}{n-1}^{-1} \\ &= npqQ(n-2;K-1,N-2)[\Delta f]\binom{N-2}{n-2}\binom{N-1}{n-1}^{-1} \\ &= Q(n-2;K-1,N-2)[\Delta f]\frac{n-1}{N-1}npq. \end{aligned}$$

□

Proof of Theorem 1: The upper bound follows directly from Stein's method: To any $A \subset \{0, \dots n\}$ there is a unique $f = f_{A,n,p}$ such that

$$\begin{aligned} Tf(k) &= 1_A(k) - P(n;p)[A], \\ \max_{0 \le k \le n-2} \Delta f(k) &\le \frac{1}{npq}, \end{aligned}$$

see Formula (10) and Lemma 1 in Ehm (1991).

For the lower bound we use the fact that for any f

$$Q(n;K,N)[Tf] - P(n;p)[Tf] \le 2\,\|Tf\|_\infty\,\|Q(n;K,N) - P(n;p)\|_{TV}$$

or by Lemma 1

$$\|Q(n;K,N)-P(n;p)\|_{TV} \geq npq\frac{Q(n-2;K-1,N-2)[\Delta f]}{2\|Tf\|_\infty}\frac{n-1}{N-1}.$$

We choose $f(k)=g(k-np)$ where

$$g(z)=\begin{cases} 0 & (|z|>2a) \\ -2a-z & (-2a\leq z<-a) \\ z & (-a\leq z\leq a) \\ 2a-z & (a<z\leq 2a) \end{cases}$$

and a is a constant to be chosen later. Then g satisfies

$$\begin{aligned} &\|g\|_\infty = a, \\ &0\leq zg(z)\leq a^2, \\ &0\leq g(z)/z\leq 1, \\ &-1\leq \Delta g(z)\leq 1, \\ &\Delta g(z)=1 \quad (-a\leq z\leq a-1). \end{aligned}$$

Arguing as on p.14 and 15 of Ehm (1991) we obtain

$$-npq-a^2-qa\leq Tf(k)\leq npq+\frac{1}{4}.$$

Let S have distribution $Q(n-2;K-1,N-2)$ and set $n'=n-2$, $p'=(K-1)/(N-2)$ and $\delta=np-n'p'$. Then $0\leq\delta\leq 1$ if $p\leq 0.5$ which we can assume without loss of generality. By the properties of g we obtain

$$\begin{aligned} \Delta f(S) &= g(S-n'p'-\delta+1)-g(S-n'p'-\delta) \\ &\geq 1-\frac{2}{a^2}(S-n'p')^2. \end{aligned}$$

Therefore

$$\begin{aligned} Q(n-2;K-1,N-2)[\Delta f]=E[\Delta f(S)] &\geq 1-\frac{2}{a^2}\operatorname{Var}(S) \\ &= 1-\frac{2}{a^2}(n-2)p'q'\frac{N-n}{N-3} \\ &\geq 1-\frac{2}{a^2}npq. \end{aligned}$$

Combining these results, we obtain

$$\|Q(n;K,N)-P(n;p)\|_{TV}\geq\frac{n-1}{N-1}\ \frac{1-2npq/a^2}{2(1+a^2/(npq)+a/(npq))}.$$

Choosing $a^2=4npq$ completes the proof. □

Proof of Theorem 2: We use the same function f as in the proof of Theorem 1, but we approximate $E[\Delta f(S)]$ with the central limit theorem instead of the Chebyshev inequality. Choosing $a^2=z^2npq$ we have

$$E[\Delta f(S)]\geq 1-2P[|S-n'p'|\geq z(npq)^{1/2}]\sim 1-4(1-\Phi(z)).$$

Moreover $\|Tf\|_\infty \sim npq(1+z^2)$. Together we thus have

$$\liminf \frac{N-1}{n-1} \|Q(n;K,N) - P(n;p)\|_{TV} \geq \frac{1-4(1-\Phi(z))}{2(1+z^2)}.$$

Choosing $z = 1.38$ completes the proof. □

Remarks:

1. We have tried to use different slopes for the increasing and decreasing parts of g, but there seems to be no gain in doing so.
2. Our proof is simpler than the one in Ehm (1991) because the term $2\left\|\Delta^2 f\right\|_\infty$ in his formula (21) is absent in our lower bound for $\|Q(n;K,N) - P(n;p)\|_{TV}$.

References

Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language*. Wadsworth & Brooks/ Cole, Pacific Grove, California.

Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statist. Probab. Lett.* **11**, 7-16.

Freedman, D (1977). A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association* **72**, 681.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second edition. Wiley, New York.

Soon, S. Y. T. (1996). Binomial approximation for dependent indicators. *Statistica Sinica* **6**, 703–714.

Vatutin, V. A. and Mikhailov, V.G. (1982). Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory Probab. Appl.* **27**, 734-743.