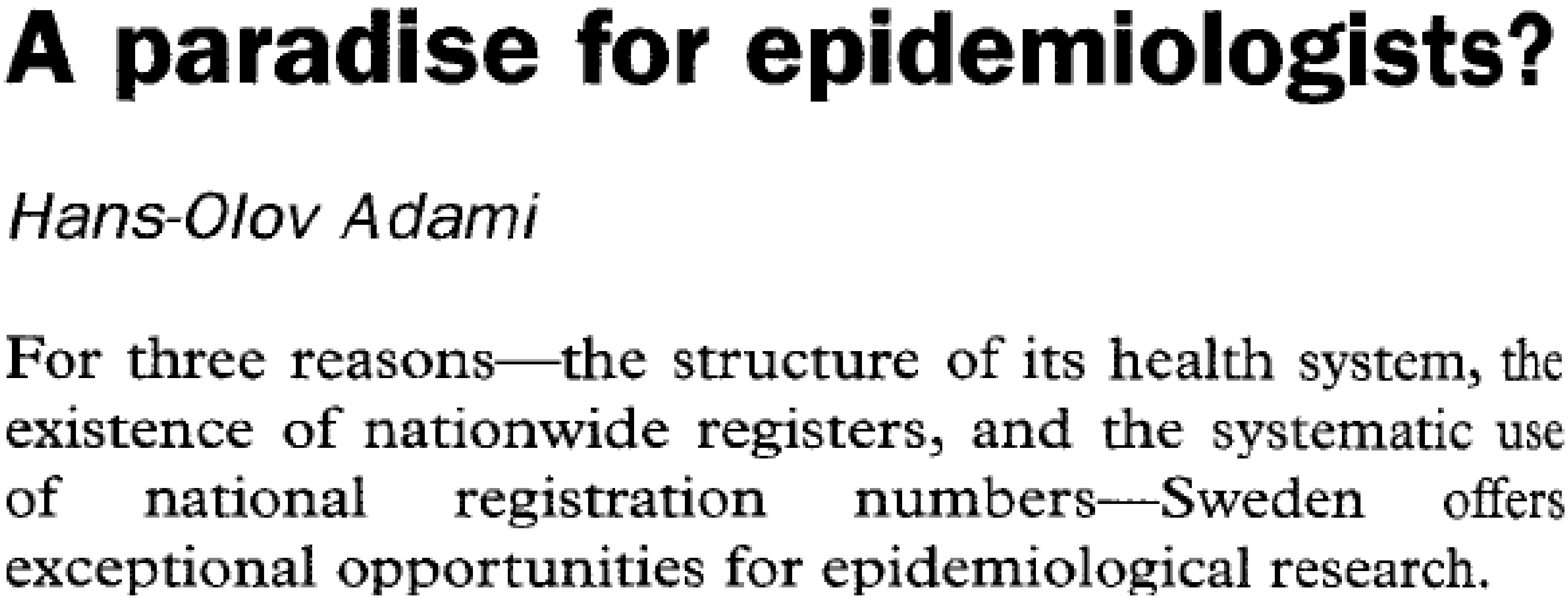
# Should we present measures of uncertainty when we have enumerated the entire population?

Paul Dickman
Sandra Eloranta
Therese Andersson
Department of Medical Epidemiology and Biostatistics (MEB)
Karolinska Institutet, Stockholm, Sweden

# Outline

- About me and my research.
- Register-based research (with a focus on Sweden).
- How we, in practice, view random error in register-based research.
- Why we present measures of uncertainty when we have enumerated the entire population.

# The Lancet 1996:347;588-589

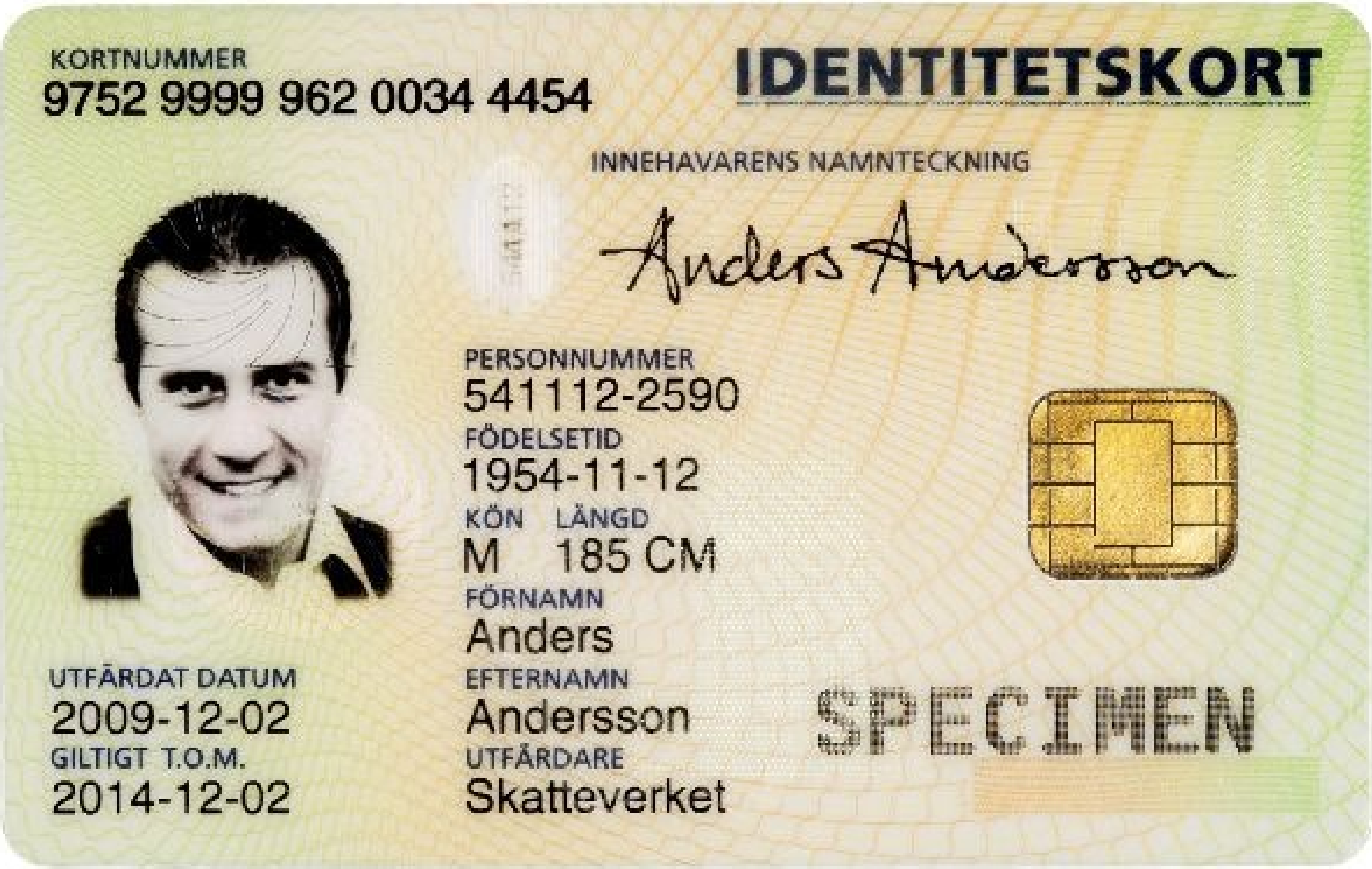


## A paradise for epidemiologists?

*Hans-Olov Adami*

For three reasons—the structure of its health system, the existence of nationwide registers, and the systematic use of national registration numbers—Sweden offers exceptional opportunities for epidemiological research.

See also (about Denmark):
When an entire country is a cohort. *Science* 2000;287:2398-9.



KORTNUMMER
9752 9999 962 0034 4454

IDENTITETSKORT

INNEHAVARENS NAMNTECKNING
Anders Andersson

PERSONNUMMER
541112-2590

FÖDELSETID
1954-11-12

KÖN LÄNGD
M 185 CM

FÖRNAMN
Anders

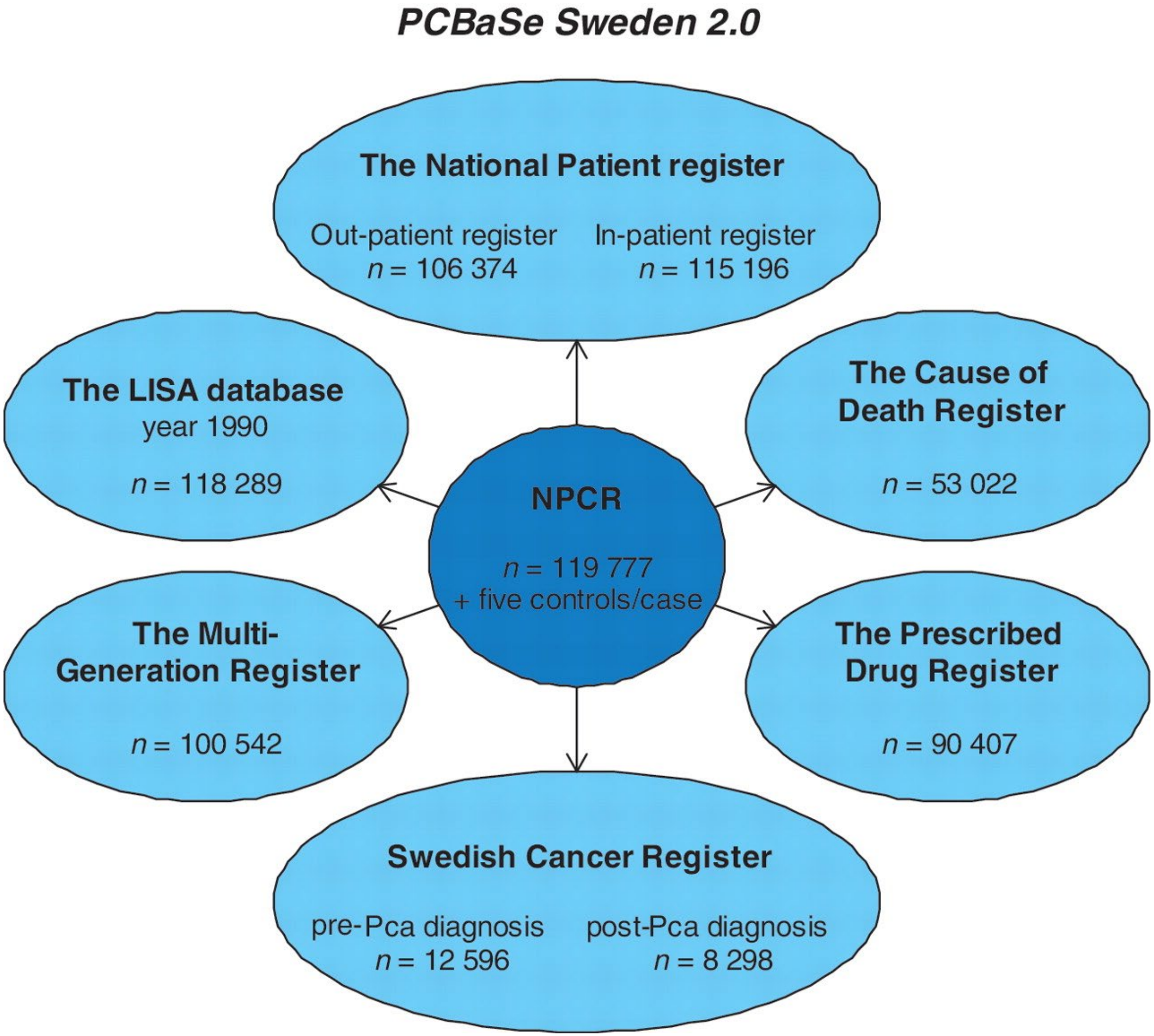
EFTERNAMN
Andersson

UTFÄRDARE
Skatteverket

UTFÄRDAT DATUM
2009-12-02

GILTIGT T.O.M.
2014-12-02

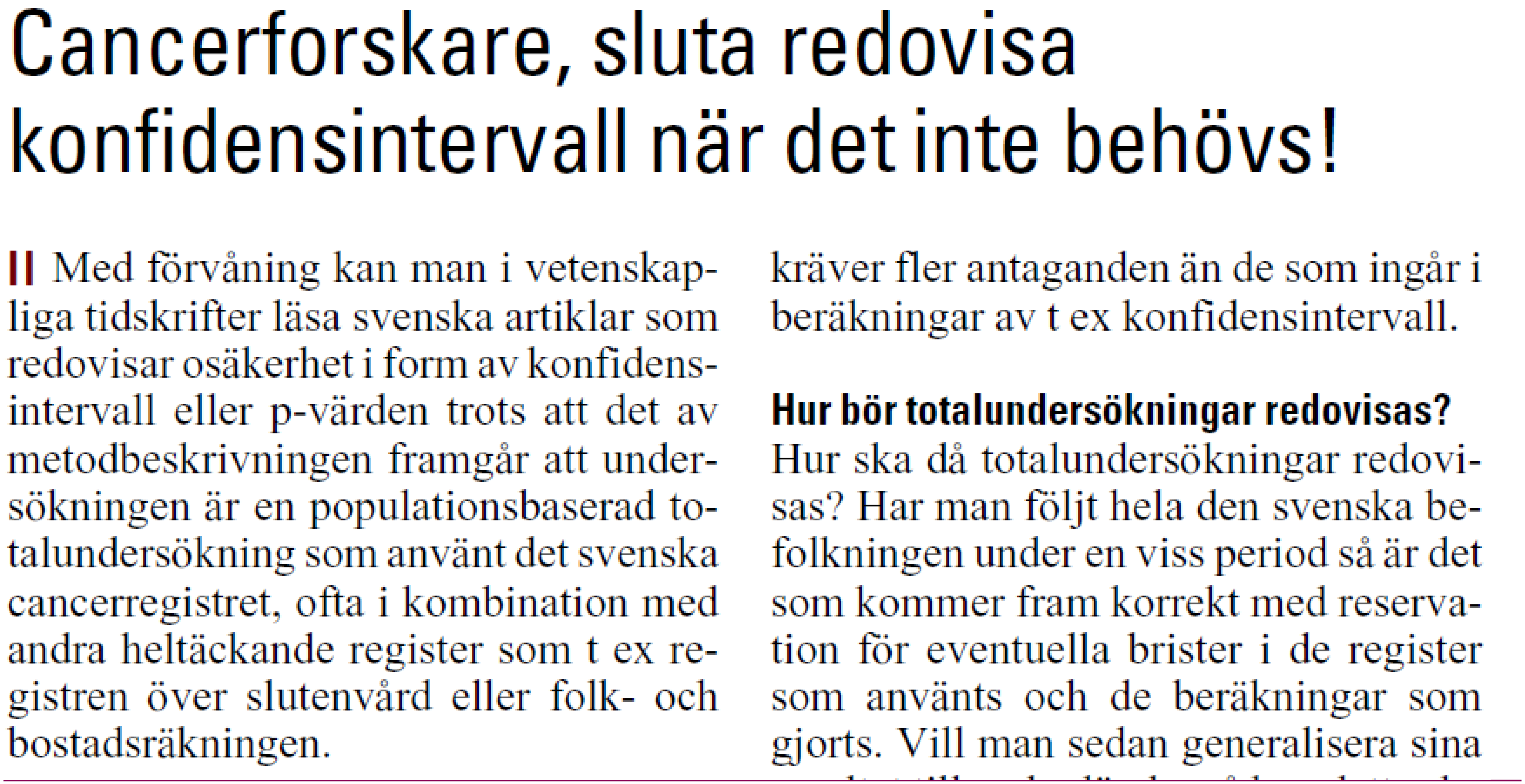
SPECIMEN



Karolinska Institutet

**Van Hemelrijck et al. *Int. J. Epidemiol.* 2012**

# Läkartidningen (Swedish Medical Journal) 2004



## Cancerforskare, sluta redovisa konfidensintervall när det inte behövs!

Med förvåning kan man i vetenskapliga tidskrifter läsa svenska artiklar som redovisar osäkerhet i form av konfidensintervall eller p-värden trots att det av metodbeskrivningen framgår att undersökningen är en populationsbaserad totalundersökning som använt det svenska cancerregistret, ofta i kombination med andra heltäckande register som t ex registren över slutenvård eller folk- och bostadsräkningen.

kräver fler antaganden än de som ingår i beräkningar av t ex konfidensintervall.

**Hur bör totalundersökningar redovisas?**

Hur ska då totalundersökningar redovisas? Har man följt hela den svenska befolkningen under en viss period så är det som kommer fram korrekt med reservation för eventuella brister i de register som använts och de beräkningar som gjorts. Vill man sedan generalisera sina

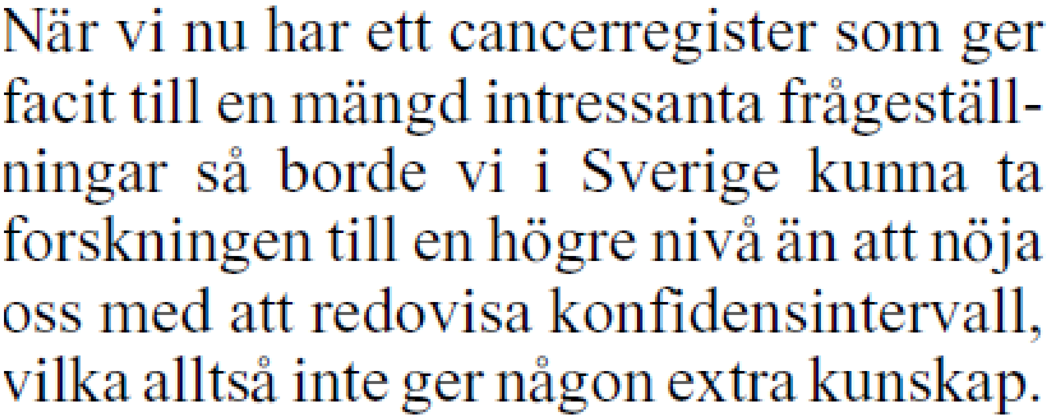
# Cancer researcher, stop presenting confidence intervals when they are not needed!

“We are surprised to read Swedish articles in scientific journals that report uncertainty using confidence intervals or p-values despite the methods section clearly describing that the study is based on the entire Swedish population using the nationwide Swedish Cancer Registry.”

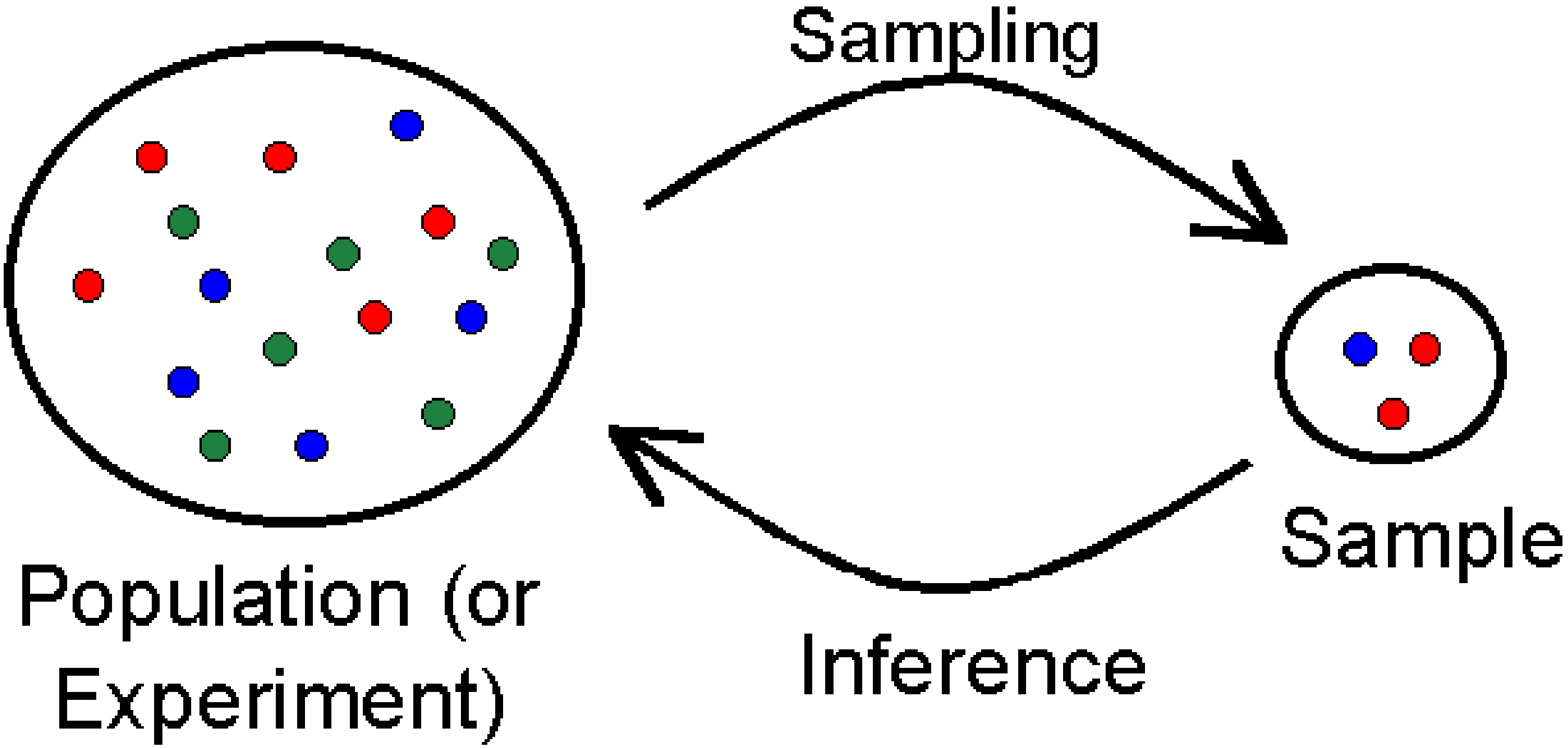
“There is no statistical uncertainty due to sampling and therefore no reason for presenting confidence intervals”

# Final sentence

“Now that we have a cancer registry that provides answers to all manner of interesting research questions, we Swedish researchers should aim higher than reporting confidence intervals that do not provide additional knowledge.”



När vi nu har ett cancerregister som ger facit till en mängd intressanta frågeställningar så borde vi i Sverige kunna ta forskningen till en högre nivå än att nöja oss med att redovisa konfidensintervall, vilka alltså inte ger någon extra kunskap.

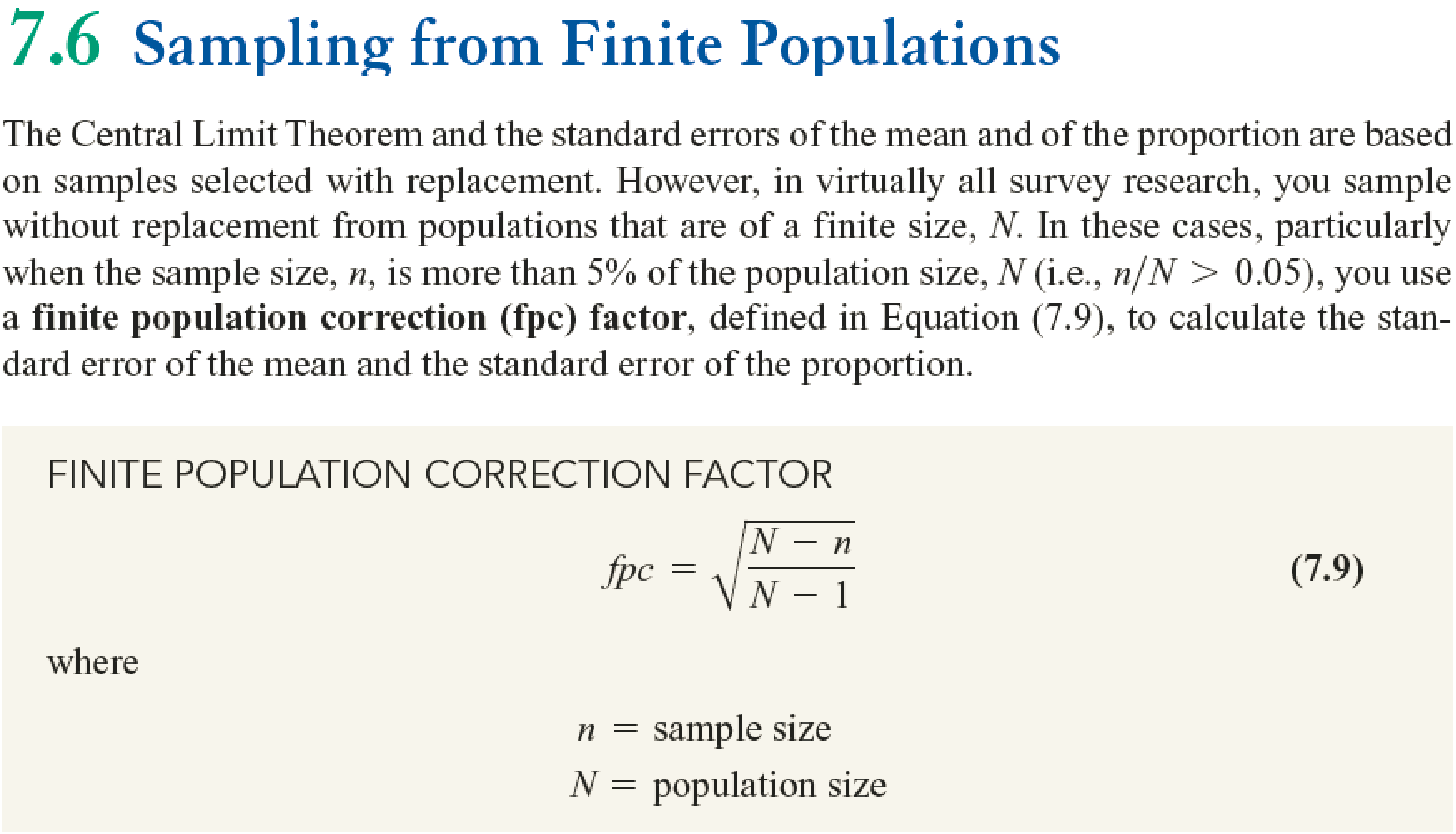

Sampling
Population (or Experiment)
Sample
Inference

# Finite population correction
## (http://en.wikipedia.org/wiki/Margin_of_error)

- The formulae for the variance assume an infinitely large population and thus do not depend on the size of the population of interest. According to sampling theory, this assumption is reasonable when the sampling fraction is small (< 5%).
- In cases where the sampling fraction exceeds 5%, analysts can adjust the margin of error using a "finite population correction", (FPC) to account for the added precision gained by sampling close to a larger percentage of the population.

$$\mathrm{FPC} = \sqrt{\frac{N-n}{N-1}}.$$

- Multiply the variance by the FPC.
- The FPC approaches zero as the sample size (*n*) approaches the population size (*N*), which has the effect of eliminating the margin of error entirely. This makes intuitive sense because when *N* = *n*, the sample becomes a census and sampling error becomes moot.



# 7.6 Sampling from Finite Populations

The Central Limit Theorem and the standard errors of the mean and of the proportion are based on samples selected with replacement. However, in virtually all survey research, you sample without replacement from populations that are of a finite size, $N$. In these cases, particularly when the sample size, $n$, is more than 5% of the population size, $N$ (i.e., $n/N > 0.05$), you use a **finite population correction (fpc) factor**, defined in Equation (7.9), to calculate the standard error of the mean and the standard error of the proportion.

FINITE POPULATION CORRECTION FACTOR

$$fpc = \sqrt{\frac{N - n}{N - 1}} \qquad \textbf{(7.9)}$$

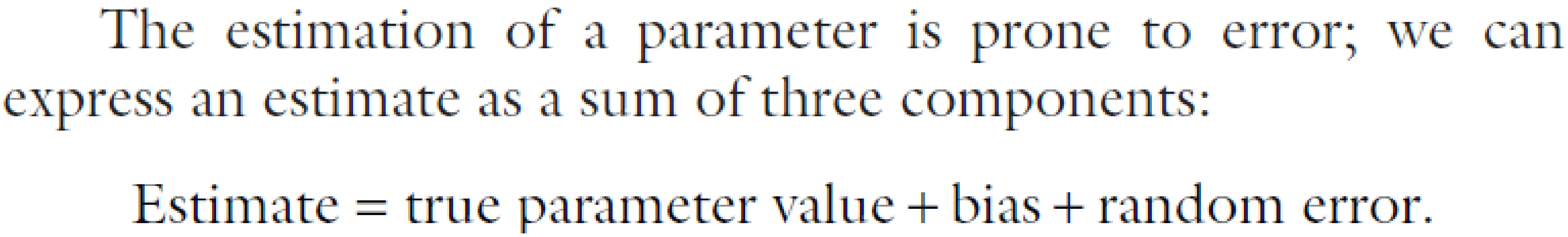
where

$$n = \text{sample size}$$
$$N = \text{population size}$$

# Possible explanations for an observed association in an observational study

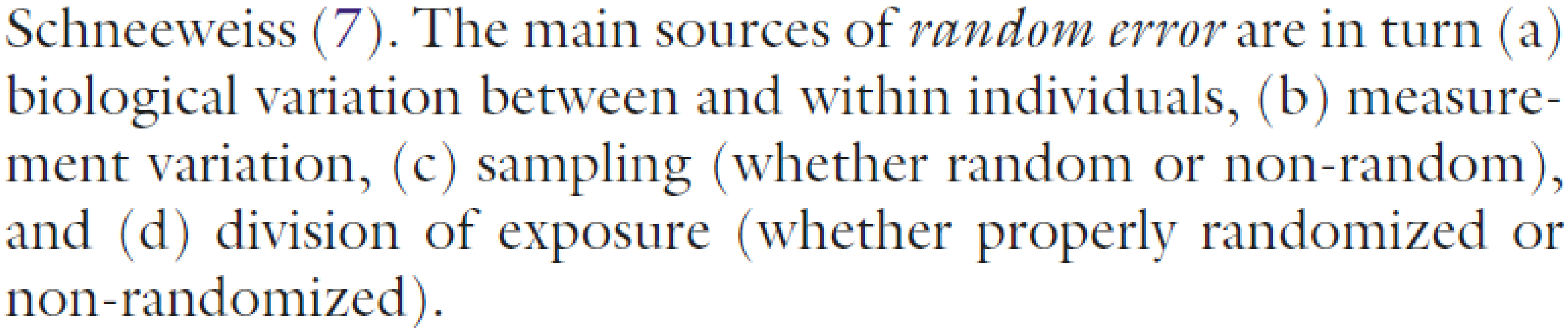
- Bias
- Confounding
- Random error
- True association between exposure and outcome

# Läärä (2011) in *Methods in Biobanking*



The estimation of a parameter is prone to error; we can express an estimate as a sum of three components:

Estimate = true parameter value + bias + random error.



Schneeweiss (7). The main sources of *random error* are in turn (a) biological variation between and within individuals, (b) measurement variation, (c) sampling (whether random or non-random), and (d) division of exposure (whether properly randomized or non-randomized).

- Also uncertainty in model selection

# From our letter to the editor in response

- In 2002, 95 new cases of tongue cancer were reported among Swedish males, while among females there were only 70 cases.
- Comparing the absolute numbers 95 and 70 does not have much scientific value. It is, however, meaningful to ask if the underlying rate of tongue cancer differs between males and females.
- Even if the numbers come from a complete census of the population, they are scientifically interesting only if considered a realisation of a random process.

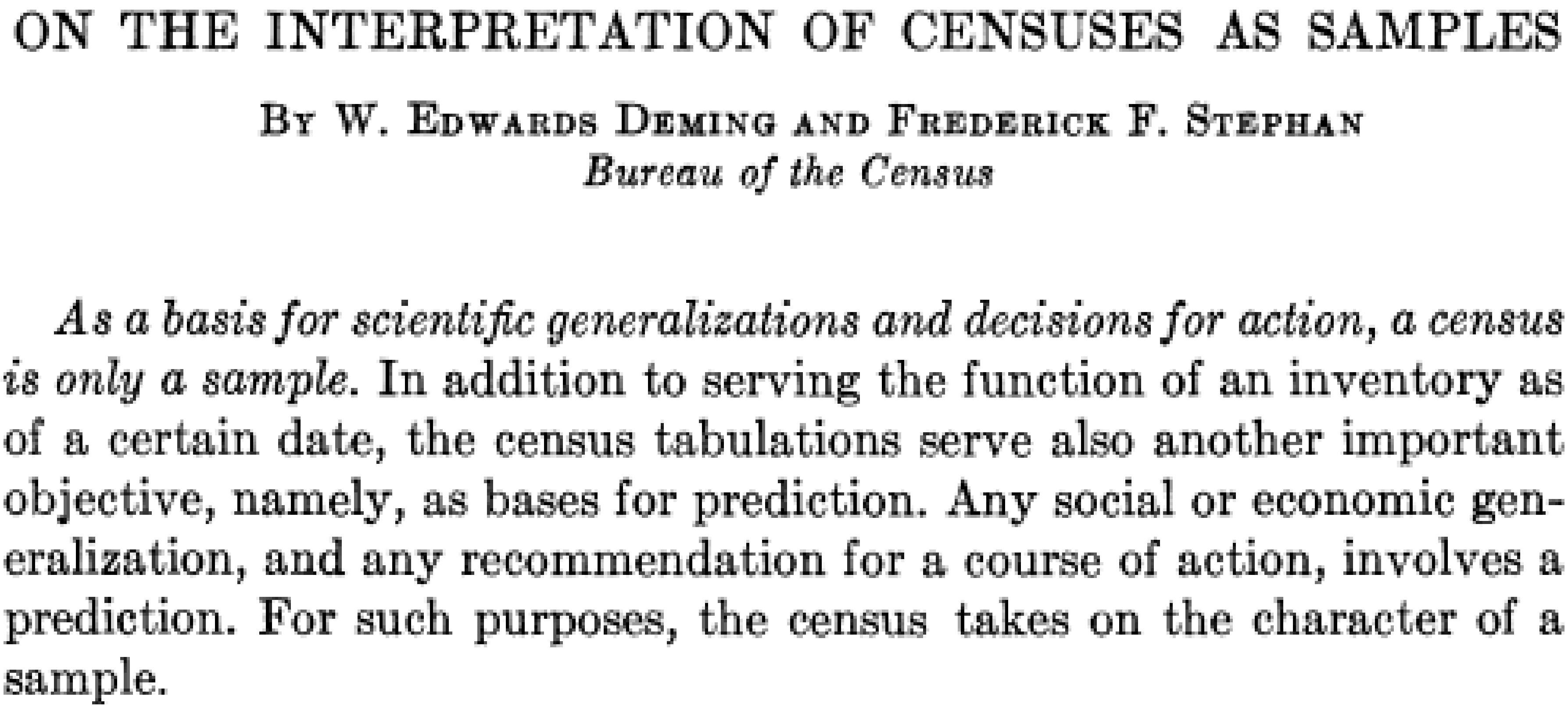
# Should we present measures of uncertainty when we've enumerated the population?

- We feel that for purposes of scientific inference the answer is unequivocally yes.
- If one is reporting the figures for administrative purposes then CIs are not required.
- We are certain that CIs are required but not certain of:

1. The statistical framework that justifies this;
2. The variance formulae.

- We will present an argument that, as a basis for generalisations and decisions, a census should be viewed as a sample from a superpopulation.

# Journal of the American Statistical Association 1941;36:45-49

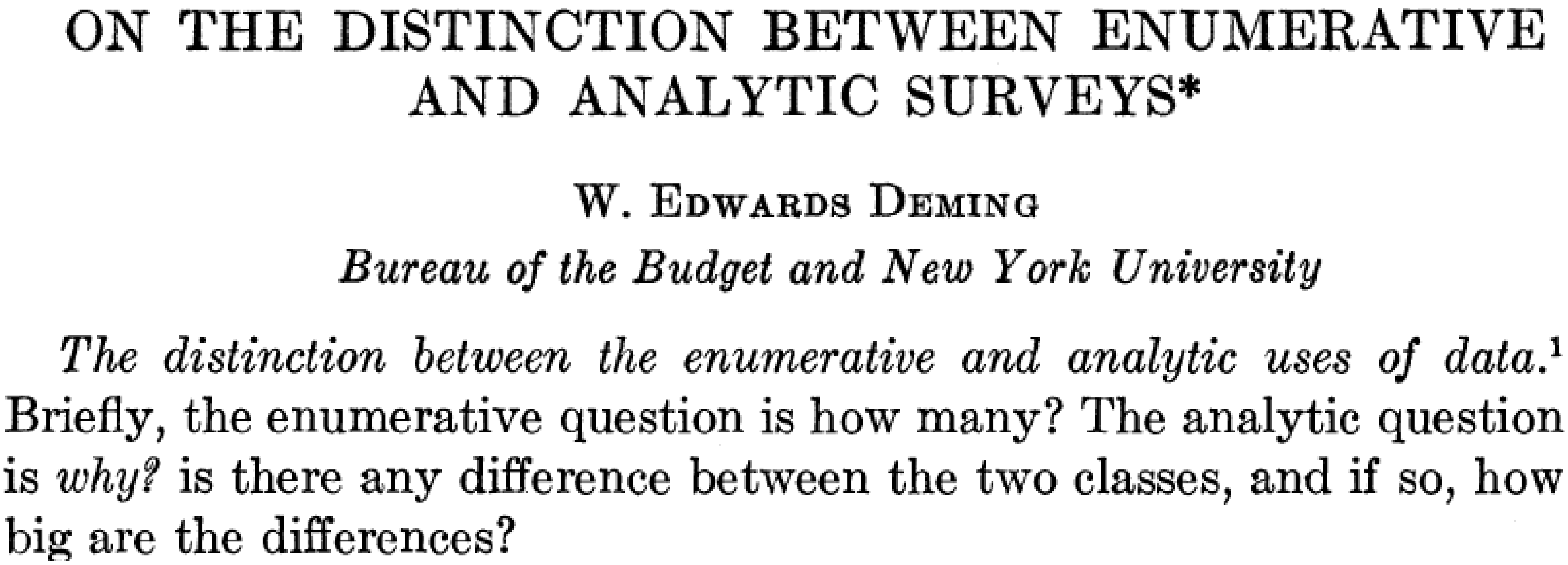


## ON THE INTERPRETATION OF CENSUSES AS SAMPLES

By W. Edwards Deming and Frederick F. Stephan
*Bureau of the Census*

*As a basis for scientific generalizations and decisions for action, a census is only a sample.* In addition to serving the function of an inventory as of a certain date, the census tabulations serve also another important objective, namely, as bases for prediction. Any social or economic generalization, and any recommendation for a course of action, involves a prediction. For such purposes, the census takes on the character of a sample.

# Journal of the American Statistical Association 1953;48:244-255

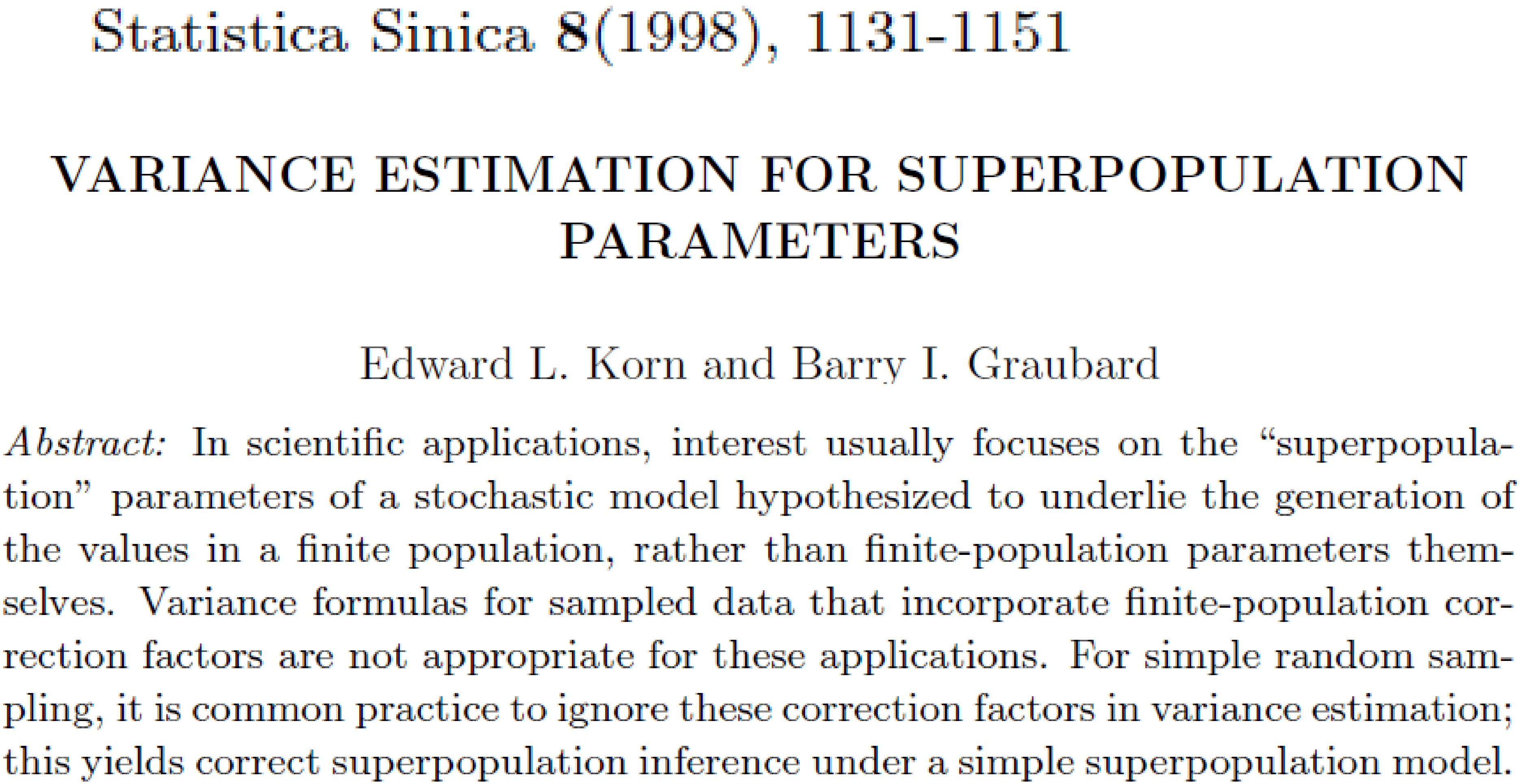


## ON THE DISTINCTION BETWEEN ENUMERATIVE AND ANALYTIC SURVEYS*

W. Edwards Deming

*Bureau of the Budget and New York University*

*The distinction between the enumerative and analytic uses of data.*[1] Briefly, the enumerative question is how many? The analytic question is *why?* is there any difference between the two classes, and if so, how big are the differences?



Statistica Sinica 8(1998), 1131-1151

# VARIANCE ESTIMATION FOR SUPERPOPULATION PARAMETERS

Edward L. Korn and Barry I. Graubard

*Abstract:* In scientific applications, interest usually focuses on the "superpopulation" parameters of a stochastic model hypothesized to underlie the generation of the values in a finite population, rather than finite-population parameters themselves. Variance formulas for sampled data that incorporate finite-population correction factors are not appropriate for these applications. For simple random sampling, it is common practice to ignore these correction factors in variance estimation; this yields correct superpopulation inference under a simple superpopulation model.