Center for Studies in Demography and Ecology

An Evaluation of the One Percent Clustered Sample of the 1990 Census of China

by

William M. Mason
University of California – Los Angeles

William Lavely
University of Washington

An Evaluation of the One Percent Clustered Sample of the 1990 Census of China*

February 4, 2004

Version 2.0

William M. Mason
Sociology Department and
California Center for Population Research
University of California—Los Angeles
masonwm@ucla.edu

William Lavely
Sociology Department and
Center for Studies in Demography and Ecology
University of Washington
lavely@u.washington.edu

*We gratefully acknowledge the valuable advice of Griffith Feeney, the generous assistance of G. William Skinner, and the support of the Henry Luce Foundation.

1. Introduction

Two micro-samples of the 1990 Chinese Census have circulated in China and abroad. The first, in order of creation, is a one percent sample of rural administrative villages and urban neighborhoods. (Note 1) The second is a one percent sample of households. We refer to the former, the subject of this article, as the "one percent clustered sample," and refer to the latter as the "one percent household sample." These data sets are not public use micro samples (PUMS) in the sense understood by users of, for example, U.S. Census data. The data are not automatically available to all who are willing to pay a posted, standard transaction fee; and no official printed documentation has been released that is specific to either sample. In particular, there is no documentation or evaluation of the method and procedures used to draw the samples.

China's National Bureau of Statistics (formerly the State Statistical Bureau) has provided the one percent household sample to several researchers, but to our knowledge has never released the corresponding mortality data. A version of the one percent sample is listed as available on the website of the China Population Information and Research Center, also without the corresponding mortality data, but this data set is provided only on a selective basis. (Note 2) The website of the Minnesota Population Center's IPUMS-International lists PUMS files for the censuses of 1982, 1990, and 2000. However, IPUMS-I has only the file for 1982 available to subscribers for download. (Note 3) We obtained the clustered sample and the corresponding mortality data separately, from unofficial sources.

This paper evaluates the one percent clustered sample, using direct inspection as well as a series of comparisons with published data drawn from the complete

enumeration. We first discuss the nature of the clustering, and report what we know about the sampling of clusters. We then note the existence of duplicate cases, and report corrected totals when duplicates are dropped. Subsequent sections describe geographic coverage of the one percent clustered sample, compare sample to published 100 percent tabulations for basic descriptors, and report selected comparisons between sample and 100 percent enumeration data at the county level.

2. Clustering

Lacking any documentation on the method used in sampling, our description of the clustered sample relies on inference and a bit of hearsay. We have been told that the clustered sample was drawn as a way to provide a timely preview of census results, before final tabulations based on the complete enumeration could be prepared. We suspect, but cannot verify, that the sample is a systematic selection of all of the living persons in every hundredth administrative village (or urban neighborhood) within each province, and of all deaths in the sampled places that occurred in the 18 months leading up to the official July 1, 1990 date of the census. (Note 4) Like the 1990 Census itself (except for national totals published in selected documents, e.g., State Council Population Census Office (1993)), the clustered sample contains civilians only.

Administrative villages and urban neighborhoods lend themselves as sampling units because they also serve as census districts, and census returns are sent up the line bundled by village or neighborhood committees. Sampling and data entry presumably took place in provincial offices. These circumstances may explain some anomalous aspects of the sample.

3. Duplicates

Duplicate cases are one such anomaly. Duplicates appeared in approximately half the provinces, and in all cases entire sample villages were duplicated. We considered the possibility that the duplicates were inserted intentionally, perhaps as a weighting scheme. But because the sample is far more faithful to the 100 percent tabulations (Population Census Office 1993) when the duplicates are omitted, it appears more likely that the duplicates resulted from errors in data processing. In their discussion of the one percent clustered sample Li and Zhu (2000:228) also conclude that the duplicates are due to processing errors. We have removed the duplicates from all computations reported in this paper.

Without duplicate records the sample consists of 8,518 administrative villages or urban neighborhoods containing 11,475,104 enumerated persons, which averages to 1,347 persons per sample unit. In addition there are 99,196 records of persons who died in the 18 months prior to the census.

4. Geographic Coverage

The clustered sample contains data in all 30 provinces and regions covered by the 1990 census. We consider first the extent to which the clustered sample reproduces the distribution of population across provinces and the major cities with provincial status in 1990 (Beijing, Tianjin, and Shanghai). Table 1 shows the percent distribution of population by province in the census, and the ratio of the sample to census percent in each province. The ratio varies from a low of .81 in Ningxia to a high of 1.28 in Tianjin

Municipality, two provincial level units with small populations. Because the sample units are clusters, the sampling variability tends to be greater than one would expect for a simple random sample of individuals, and the extreme ratios occur in provinces with very small proportions of the population. The ratio in larger provinces generally varies between .95 and 1.05.

Table 1 here

There is also broad geographic coverage within provinces. Approximately 91 percent of China's 2,845 county-level units contain at least one sample administrative village or urban neighborhood. Among the 2,600 county-level units with coverage, there is an average of 3.2 sample units per county. Sample coverage is shown in the accompanying map of China's counties (Figure 1). Counties containing at least one village unit in the sample are shown in gray or black, while counties with no coverage are in white. Coverage is quite regular in China Proper and Manchuria, as contrasted with the sparser coverage in the Inner Asian regions of Inner Mongolia, Xinjiang, and Tibet, and parts of Gansu and Qinghai provinces. As may be seen in Table 1, none of these provinces is under-sampled. The sparseness of populations in these areas appears to explain the lack of coverage. Tibet, however, is an exception.

Figure 1 here [map]

The sample for Tibet lacks cities and towns and is thus entirely rural. According to the official 100 percent tabulations for 1990 (Tibet Autonomous Region Population Census Office 1992), the Tibet Autonomous Region is 11.5 percent urban. The Tibet sample consists of 23 villages distributed over 10 of 78 possible counties. Thirteen villages are in a single county (Dingri, the site of Mount Everest), two are in one county,

and the remaining eight are distributed one per county. It is likely that the sample villages not in Dingri are composites of county sub-samples. This is suggested by the large size of these units (approximately seven times as large as the Dingri sample villages), and by their sample code numbers, each of which is "1." The Tibet sample thus appears to have been constructed according to different principles from the rest of the one percent clustered sample. Nonetheless, the sample data accord well with the 100 percent tabulations for rural Tibet (Tibet Autonomous Region Population Census Office 1992). For example, the 100 percent tabulation for Tibet shows that 76.3 percent of the rural population age 15 and above is illiterate. In the one percent sample, the corresponding figure is 77.4 percent. There is also a close correspondence with the rural age distribution, the distribution of rural women by their number of live births, and on other characteristics. The Tibetan sample may thus be useful for some purposes.

5. National Comparisons of Sample and Census

A series of comparisons at the national level (Tables 2-3 and Figures 2-3) reveals a reasonable concordance between the sample and the underlying complete census data as derived from published tabulations (State Council Population Census Office 1993). When the total of persons in the sample is multiplied by the reciprocal of the sampling fraction (i.e., 100) and divided by the census total, the resultant ratio is 1.02. The one percent clustered sample thus overstates the census population by two percent (see Table 2). Births are similarly overstated, while the death sample (after an adjustment to account for excluded counties, discussed in the next section) understates deaths by .4 percent.

Table 2 here

Having established a fair concordance between census and sample for total population, births, and deaths, we now consider the concordance of distributions of populations across various categories listed in Table 2. Most measures, such as percent male, percent rural, the sex ratio at birth, and deaths by semester, are within two percent of the census value. The distributions of population by occupation and by marital status are similarly close. There are two exceptions. The sample over-states the percent university by 10 percent, perhaps as a consequence of the over-sample of the provincial level cities Shanghai and Tianjin that can be observed in Table 1. There is also an over-sampling of births in 1990 relative to 1989, for which we have no explanation.

Figure 2 presents a sample to census comparison of the sex-specific age distributions of those alive at the time of enumeration. The sample distribution of females by age is quite close to that for the census, varying within .5 percent at every age below 80. The male sample distribution is less regular. It contains an excess of males at ages 20-29, and a deficit of males age 60-75. Even so, these deviations are within one percent of the census value. The greater variability of males may be due to the greater concentration of males in sparse collective households and related institutional concentrations.

Figure 2 here

Figure 3, based on those who died in the 18-month period between January 1, 1989 and June 30, 1990, is constructed identically to Figure 1, but is based on age at death. The deviations of sample from census are more dramatic for the deceased. There is a notable dearth of dead males at ages 5-14—approximately 8 percent fewer than the corresponding census percentage. There is a similar dearth of females at ages 25-34.

There are too many sample male death cases at ages 35-39, and too few sample female death cases at ages 45-49, approximately 10 percent fewer than expected. We have no explanation for these irregularities other than sampling variation.

Figure 3 here

Because infant mortality is of particular interest, in Table 3 we further compare deaths at age 0 conditional on sex and semester of birth with the corresponding figures from published census tabulations. Official sources do not document the calculation of infant mortality *rates* with detail sufficient to sustain independent replication. For this reason our analysis is restricted to comparison of sample *frequencies* of death with those derived from the 100 percent enumeration tables (State Council Population Census Office 1993). The interior cells of Table 3 display ratios of sample deaths to complete enumeration deaths conditional on sex and semester of birth. The row and column margins contain ratios separately for sex and semester. Because of gaps in death coverage at the county level, which are discussed in the next section, the "total" ratio is less than one (.960), which suggests that the clustered sample undercounts infant deaths. However, upon adjusting the total ratio for coverage gaps under the assumption that infant deaths were missed with probability identical to that for deaths to older individuals, the total ratio becomes .996. This result suggests that infant deaths are not specifically undersampled in the one percent clustered sample. The sex-semester specific ratios in Table 3 should thus be considered to be downwardly biased owing to the absence of adjustment for nonreporting of deaths in particular counties. Of greater concern is the apparent undersampling of male relative to female infant deaths in every semester, as well as the inconsistency over semesters in the relative undersampling. (Note 5) We have

no explanation for this variability, but note that at a minimum it complicates the conclusions that can be drawn from individual level analyses of infant mortality.

Table 3 here

6. Comparisons with County-Level Data

For total population and for infant deaths we carried out sample to census comparisons at the county level. If the sample data are unbiased at the county level, a regression at this level of sample data on census data for an identically defined variable should yield a coefficient of .01. For total population we found a slope of .0096 (N=2,312). Fitting a cubic polynomial spline to the data revealed modest departures from linearity. If a handful of counties is excluded, the regression coefficient becomes .01.

The sample to census comparison of infant deaths per county is limited to 1,357 counties for which 1990 county (complete enumeration) census reports on infant mortality are available. (Note 6) Although this subsample of counties may be biased, that possibility should not affect the sample-census relationship. At the county level, for sample infant deaths regressed on complete enumeration infant deaths the slope is .0089. There appears to be under-sampling in counties with higher numbers of infant deaths.

Our examination of the clustered sample death data detected a problem. Of the 2,600 county-level units in the clustered sample, 97 contain no death data. For a subset of these 97 counties, the absence of deaths is probably due to procedures followed at the local level, rather than to the inherent variability occasioned by probability sampling. In three contiguous prefectures in Henan (Shangqiu, Zhoukou, and Zhumadian), and two contiguous prefectures in Sichuan (Wanxian and Fuling), there are no mortality data.

These five prefectures account for 43 of the 97 county-level units for which there are no deaths. The county-level sample sizes in these prefectures range from a minimum of 1,600 to a maximum of 15,441. Given the size and contiguity of the areas, it is clear that the lack of mortality data in these counties is due to some aspect of procedure and not sampling variability. These five prefectures should be excluded from any analysis of mortality.

The remaining 54 zero-death county-level units are geographically scattered, although many pertain to urban units in Heilongjiang and Anhui. Because it is at least theoretically possible that the sampled villages in fact recorded no mortality in the 18 months prior to the census, we took a statistical approach to the problem of including zero-death counties. The procedure involved generating two series of county-based log-odds of death—one based on 100 percent census data, the other on the corresponding sample log odds for these counties with at least one death. We regressed the sample logits of death on the census-based logits. Zero-death counties were then supplemented with a single death (so that a log-odds could be estimated), and included in a second regression. Zero-death counties with imputed odds of death more than two standard deviations from the predicted value were marked for exclusion from mortality analysis. This led to the exclusion of an additional 45 counties, while retaining nine out of the original 97 zero death counties. (Note 7) The counties so marked for exclusion, portrayed in the map (Figure 1) with black shading, contain 3.8 percent of census deaths. (Note 8) Excluded counties are listed in Appendix 1.

7. Discussion

The one percent clustered sample appears to be a true one percent sample of the 1990 census. It reproduces the geographic distribution of population and major population components quite well. Although the clustering of the sample reduces precision, it permits contextual analyses based on multilevel methods of statistical analysis.

There are anomalies. The sample for Tibet lacks urban units and appears to use a different sampling procedure. The national distribution of deaths by age for males is irregular, and male infant mortality is somewhat under-sampled relative to the census. These deficiencies must be assessed for their relevance to specific analytic purposes. For example, there is mounting evidence (e.g., Ministry of Health 1999) that the 1990 census underreported infant mortality by a margin far wider than the gap between sample and census infant mortality. Against this kind of uncertainty, the sample can be useful. The results of the sample/enumeration comparisons we have presented suggest that the one percent clustered sample will be serviceable for many purposes.

Notes

1. An administrative village consists of one or several adjacent natural villages, and is the lowest level rural civil unit. In 1990, a neighborhood was the corresponding urban civil unit.

2. Instructions for obtaining the data, current as of 30 January 2004, may be found on this organization's website (http://www.cpirc.org.cn/en/eindex.htm).

3. Instructions for downloading data, current as of 30 January 2004, may be found on the IPUMS-I website (http://www.ipums.umn.edu/international/).

4. According to the codebook, "this dataset (1% sampling) was prepared taking villages as sampling unit …" (State Statistical Bureau 1994).

5. We employed Poisson and negative binomial models, and checked for over-dispersion, to reach these conclusions. The sex effect is significant at the .1 level. Of the semester contrasts, only that between the first and third semester is significant at the .05 level.

6. County census totals of numbers of infant deaths were compiled from 1990 county census volumes.

7. The logit regression based on counties with at least one death, and the corresponding regression in which zero-death counties are included with small imputed values, are each consistent with the hypothesis that the line has a slope of .01 and intercept of zero.

8. In Table 2, for the ratio of sample to census deaths, sample deaths are increased by 3.8 percent prior to calculation of the ratio.

References

Li Shuzhuo and Zhu Chuzhu. 2000. *Research and Community Practice on Gender Difference in Child Survival in China.* Beijing: China Population Publishing House.

Ministry of Health, Peoples Republic of China. 1999. Zhongguo weisheng tongji tiyao 1999 (Chinese Health Statistical Digest 1999). Beijing: Health Press.

State Council Population Census Office, and Department of Population Statistics, State Statistical Bureau, People's Republic of China. 1993. *Tabulation on the 1990 Population Census of the People's Republic of China.* Beijing: China Statistical Publishing House. Four volumes.

State Statistical Bureau. 1994. "Data Structure of the 1990 Population Census of China." [Unpublished codebook.]

Tibet Autonomous Region Population Census Office. 1992. *Tabulation on the 1990 Population Census of Tibet Autonomous Region.* Lhasa: Tibet People's Publishing House.

Table 1. Comparison of the Provincial Distribution of Population in the 1% Clustered Sample to that in 100% Census Tabulations, China, 1990

Province	Percent of Population	Ratio of Sample to Census
Beijing	0.96	1.03
Tianjin	0.78	1.29
Hebei	5.40	0.96
Shanxi	2.54	1.03
Inner Mongolia	1.90	1.02
Liaoning	3.49	1.00
Jilin	2.18	1.07
Heilongjiang	3.12	1.00
Shanghai	1.18	1.14
Jiangsu	5.93	0.99
Zhejiang	3.67	0.99
Anhui	4.97	0.95
Fujian	2.66	1.11
Jiangxi	3.34	1.04
Shandong	7.46	0.97
Henan	7.57	0.96
Hubei	4.77	1.03
Hunan	5.37	1.02
Guangdong	5.56	0.98
Guangxi	3.74	1.03
Hainan	0.58	1.09
Sichuan	9.48	0.96
Guizhou	2.87	0.95
Yunnan	3.27	0.99
Tibet	0.19	1.09
Shaanxi	2.91	1.01
Gansu	1.98	1.06
Qinghai	0.39	1.27
Ningxia	0.41	0.82
Xinjiang	1.34	0.98
Total	100.01%	1.015

Sources: Population Census Office of China (1993) and one percent clustered sample.

Table 2. Comparison of National Statistics Derived from the 1% Clustered Sample to Corresponding Statistics Derived from Complete Enumeration Tabulations of the 1990 Chinese Census

Item	100% Census Value	Ratio of Sample to Census
Total population	1130510638	1.02
Births (enumerated)	35110945	1.01
Deaths	10328899	1.00[a]
Population % male	51.47	1.00
Population % rural	73.80	0.99
Population % in collective households	2.89	1.02
Population % non-Han	8.08	1.00
Sex ratio of births (enumerated)	111.45	1.00
Sex ratio of births (mother reports)	114.18	1.00
Births by semester[b]		
Births 1989 first half %	32.25	0.98
Births 1989 second half %	37.88	0.98
Births 1990 first half %	29.87	1.04
Total	100.00	
Deaths total by semester[b]		
Deaths 1989 first half %	31.76	0.99
Deaths 1989 second half %	31.87	0.99
Deaths 1990 first half %	36.36	1.02
Total	99.99	
Educational Level[b]		
University %	0.62	1.10
Technical college %	0.97	1.04
Vocational high school %	1.74	0.97
Upper middle school %	7.30	1.01
Lower middle school %	26.50	1.00
Primary school %	42.27	1.00
Illiterate or semi-literate %	20.61	1.00
Total	100.01	
Occupation[b]		
Professional and Technical %	5.32	1.02
Cadres %	1.75	1.02
Administrative staff %	1.74	0.98
Commercial workers %	3.01	1.01
Service workers %	2.40	1.02
Agricultural workers %	70.61	0.99
Production workers %	15.17	1.02
Total	100.00	

Continued on next page

Table 2. Continued.

Marital Status[b]		
Unmarried %	25.13	1.00
Married %	68.18	1.00
Widowed %	6.10	1.00
Divorced %	0.59	1.00
Total	100.00	

Sources: Population Census Office of China (1993) and one percent clustered sample.

[a]Sample deaths are adjusted; see text for explanation.
[b]Percentage distribution sums to 100 percent, with deviations due to rounding.

Table 3. One Percent Cluster Sample Deaths at Age 0 (Multiplied by 100) Divided by Corresponding 100% Enumeration Deaths, Conditional on Sex and Semester of Death, China, 1990[a]

Semester	Male	Female	Total
1989: January-June	0.911 (120,406)	0.946 (118,893)	0.928 (239,299)
1989: July-December	0.959 (142,015)	0.962 (145,593)	0.961 (287,608)
1990: January-June	0.943 (177,418)	1.02 (185,177)	0.981 (362,595)
Total	0.941 (439,839)	0.980 (449,663)	0.960 (889,502)

Sources: Population Census Office of China (1993) and one percent clustered sample.
Note: Numbers in parentheses are death counts from 100 percent enumeration
[a]Sample values are unadjusted for lack of death coverage in certain counties; see text for discussion.

Appendix 1. Zero Death Counties Selected for Exclusion from Mortality Analysis in the One Percent Clustered Sample, China, 1990

GB Code	Province	Name
150103	Inner Mongolia	Huhehaote: Huimin qu
150122	Inner Mongolia	Tuoketuo xian
150402	Inner Mongolia	Chifeng: Hongshan qu
210802	Liaoning	Yingkou: Zhanqian qu
210803	Liaoning	Yingkou: Xishi qu
210811	Liaoning	Yingkou: Laobian qu
230402	Heilongjiang	Hegang: Xiangyang qu
230403	Heilongjiang	Hegang: Gongnong qu
230702	Heilongjiang	Yichun: Yichun qu
230705	Heilongjiang	Yichun: Xilin qu
230811	Heilongjiang	Jiamusi: Jiaoqu
230826	Heilongjiang	Huachuan xian
230834	Heilongjiang	Youyi xian
230881	Heilongjiang	Tongjiang shi
232603	Heilongjiang	Wudalianchi shi
320703	Jiangsu	Lianyungang: Lianyun qu
320704	Jiangsu	Lianyungang: Yuntai qu
320705	Jiangsu	Lianyungang: Xinpu qu
330921	Zhejiang	Daishan xian
340302	Anhui	Bengbu: Dong qu
340304	Anhui	Bengbu: Xi qu
340404	Anhui	Huainan: Xiejiaji qu
340503	Anhui	Ma`anshan: Huashan qu
340702	Anhui	Tongling: Tonggongshan qu
340803	Anhui	Anqing: Daguan qu
341002	Anhui	Huangshan shi CC: Tunxi qu
341003	Anhui	Huangshan: Huangshan qu
341004	Anhui	Huangshan: Huizhou qu
341023	Anhui	Yi xian
342101	Anhui	Fuyang shi
342530	Anhui	Jingde xian
350203	Fujian	Xiamen: Siming qu
360302	Jiangxi	Pingxiang: Chengguan qu
360311	Jiangxi	Pingxiang: Shangli qu
362124	Jiangxi	Dayu xian
362129	Jiangxi	Dingnan xian
410411	Henan	Pingdingshan: Jiaoqu
412321	Henan	Yucheng xian
412322	Henan	Shangqiu xian
412323	Henan	Minquan xian
412324	Henan	Ningling xian

Continued

Appendix 1, continued—p. 2

412325	Henan	Sui xian
412326	Henan	Xiayi xian
412327	Henan	Zhecheng xian
412328	Henan	Yongcheng xian
412701	Henan	Zhoukou shi
412721	Henan	Fugou xian
412722	Henan	Xihua xian
412723	Henan	Shangshui xian
412724	Henan	Taikang xian
412725	Henan	Luyi xian
412726	Henan	Dancheng xian
412727	Henan	Huaiyang xian
412728	Henan	Shenqiu xian
412729	Henan	Xiangcheng xian
412801	Henan	Zhumadian shi
412821	Henan	Queshan xian
412822	Henan	Biyang xian
412823	Henan	Suiping xian
412824	Henan	Xiping xian
412825	Henan	Shangcai xian
412826	Henan	Ru`nan xian
412827	Henan	Pingyu xian
412828	Henan	Xincai xian
412829	Henan	Zhengyang xian
450502	Guangxi	Beihai: Haicheng qu
512201	Sichuan	Wanxian shi
512221	Sichuan	Wan xian
512222	Sichuan	Kai xian
512223	Sichuan	Zhong xian
512224	Sichuan	Liangping xian
512225	Sichuan	Yunyang xian
512226	Sichuan	Fengjie xian
512227	Sichuan	Wushan xian
512228	Sichuan	Wuxi xian
512229	Sichuan	Chengkou xian
512301	Sichuan	Fuling shi
512322	Sichuan	Dianjiang xian
512323	Sichuan	Nanchuan xian
512324	Sichuan	Fengdu xian
512326	Sichuan	Wulong xian
513227	Sichuan	Xiaojin xian
533121	Yunnan	Luxi xian
610303	Shaanxi	Baoji: Jintai qu
620105	Gansu	Lanzhou: Anning qu
640121	Ningxia	Yongning xian

Continued

Appendix 1, continued—p. 3

654225	Xinjiang	Yumin xian
654226	Xinjiang	Hebukesai`er Mengguzu zizhixian