SIZE OF POPULATION AND VARIABILITY OF DEMOGRAPHIC DATA 17th and 18th Centuries
Author(s): BARBARA SPENCER
Source: *Genus*, Vol. 32, No. 3/4 (1976), pp. 11-42
Published by: Università degli Studi di Roma "La Sapienza"
Stable URL: http://www.jstor.org/stable/29788128

BARBARA SPENCER

SIZE OF POPULATION AND VARIABILITY OF DEMOGRAPHIC DATA 17th and 18th Centuries (*)

1. INTRODUCTION

In most historical demographic studies pertaining to the 17th and 18th centuries there is very little discussion of the appropriateness of the size of parish chosen or the number of years of data over which the observations should be averaged. Implicitly work is often based on a number of assumptions which may not be valid. In particular it is often assumed that studies based on small populations will not give results which are representative for any larger population because of the presumed high degree of variability resulting from the small size. Attempts are made by almost all authors (1) working with what they have considered to be small populations to compensate for the smallness of the sample either by aggregating over a number of parishes or by increasing the number of years over which the data is averaged. This is of concern, since for reasons which will soon become apparent, we believe that not much progress can be made unless much fuller use is made of disaggregated data (2).

The latter method of increasing the number of years over which the data is averaged, has of course been extremely popular in the historical demographic literature dealing with 17th and 18th century Europe. Henry and Fleury themselves took this approach in the two editions of their «Manuel de dépouillement...»: the seasonal fluctuations of baptisms and deaths are to be studied for periods of 100 years if the population is less than 500, for periods of 50 years if the population is between 500 and 1500 and for periods of 20 years if the population exceeds 1500.

(*) I wish to thank Professor Paul Deprez for having drawn my attention to the topic discussed in this paper and to thank Dr. Derek Hum for his help in processing the data. Both of these colleagues provided valuable discussion and comments in the process of writing this paper.

(1) Most recent French, Belgium and Italian studies suffer from that evil and a quick glance at the studies published in «Population» and the «Annales de démographie historique» will convince the reader of the fact that our statement is not exaggerated.

(2) M. FLEURY and L. HENRY, *Nouveau manuel de dépouillement et d'exploitation de l'état civil ancien*, 2nd édition, Paris, 1965, p.103.

Wrigley correctly viewed the pitfalls of such an approach: «... This is a self-defeating procedure since it makes it impossible to measure changes within the period. If these changes were important the result of lumping together families over a long period of time is to produce average figures which may not have been true for any of the subperiods» (3). Although we do not think that anyone will disagree with the above statement, it is rather disheartening to observe the number of authors who follow the procedure of averaging over a long period especially when the population under study is small. When two variables so different from each other as size and time-span are integrated in order to make up for a possible statistical insufficiency in one of the variables (in this case the spatial variable), misleading results may easily occur because of fundamental changes over time. Furthermore, as will be shown below, aggregation over time may not result in the hoped for decrease in variability defeating the purpose of the exercise.

The above discussion does not of course refer to the type of study dealing with one single parish or village which may have been chosen according to the personal preferences of the author (his place of birth or residence). Since in this case, it is the village or parish that is of interest in itself there should be no reason for the author to try to compensate for the small village size by aggregating over time (even though this is occasionally attempted).

The alternative method used to compensate for small parish size is to aggregate over a number of parishes. It is argued by some researchers that there is no point in studying small parishes individually, since if they are judged to be too small to be statistically representative, no general conclusions can be made. The solution, it is argued, is to aggregate over space increasing the broadness of the sample. An example of this type of approach is the Enquete of the I.N.E.D. (4), the study of the French Canadian population etc. (5). This approach commonly makes use of a small sample from a large number of villages. For instance the Enquete of the I.N.E.D. study on the French population since Louis XIV is based on a 1% sample judged to be representative (6).

Our main objection to this approach is that when a sample is considered representative it is assumed to be representative of a larger population which is supposed to be identical in structure and in relative importance of various homogeneous characteristics. While such an assumption is acceptable in a theoretical context one can question to what extent the average characteristics obtained from many studies do correspond to real characteristics of the population. As numerous

(3) E.A. WRIGLEY, *Family Reconstitution vs. Introduction to English Historical Demography* (ed. E.A. Wrigley), New York, s.d., p. 104,

(4) M. FLEURY and L. HENRY, *Pour connaître la population de la France depuis Louis XIV*, «Population», XIII, 4, 1958, pp. 633-686.

(5) H. CHARBONNEAU, J. LEGARÉ, R. DUROCHER, G. PAQUET, J.P. WALLOT, *La démographie historique au Canada*, «Recherches Sociographiques» VIII, 2, 1967, pp. 1-4.

(6) See note 4, pp. 666 et seg.

variables of a social, economic, religious and cultural nature will determine or at least strongly influence demographic characteristics, a sample will only be representative of a larger population for which the non-demographic variables are similar or comparable to the ones of the sample. If this condition is not fulfilled any discussion dealing with representativity becomes highly irrelevant. In most situations the chances of seeing the above condition fulfilled will be rather small and the total population to which the sample pertains will be very limited (7). Taking a large sample which is usually possible only by aggregating over a varied population will only yield averaged results. With this kind of data the study of causalities or relationships underlying demographic behaviour becomes extremely difficult since differences that may give an idea of the causes underlying a given situation will be averaged out and smoothed away. The only way that one can come to grips with causes and ultimately find an explanation for the phenomena observed, is to study small population entities at the level at which relationships between cause and effect can be established. Even so, we still have the problem of generalizing from results obtained from a sample such as a small village. Here the issue at stake basically reduces to one of properly assessing the degree of variability in the data since that variability will of course be influenced by the sample size.

In this paper, as a first step in studying this problem of sample size, we estimated the average number of births, marriages and deaths from a large number of villages and urban parishes in 17th and 18th century Belgium as well as some similar data from France. We examined the relationship between these averages and a measure of variability, the coefficient of variation, to obtain an estimate of the village or parish size for which the variability of these particular demographic characteristics could become an issue in interpreting and making generalizations from the data.

In fact, as expected from theoretical considerations (see Section III), as the mean number of births and marriages increases, the coefficient of variation declines in a clearly defined fashion (see descriptions in Sections IV and VI and the graphs). No clear relationship between the mean and the coefficient of variation for deaths could be found presumably because it is so subject to the effect of crises. From analysis of the data we found an estimate of the mean number of births or marriages after which the advantage of further increases in village or parish size in reducing the coefficient of variation is minimal. It follows that if the sample village or parish is larger than the critical size, the argument that a larger village or parish should have been used to reduce variability becomes weak. On the other hand, if the mean number of marriages is less than the critical size, the researcher may need to be cautious.

---

(7) We refer to an article by P. Deprez to be published in the series Studia Historica Gahoensia, (University of Ghent, Belgium).

Thus our analysis indicates that relationships obtained from relatively small populations should be just as reliable as those from larger populations for the purpose of inferring the existence of these relationships in other populations experiencing similar economic circumstances. We hope that with the information from this study researchers will have more confidence in demographic variables estimated from small villages or parishes so that rather more studies such as regression models relating economic and demographic variables will be undertaken.

As an extension of our study much work is needed to examine in detail the actual consistency of patterns of relationships in small villages in similar economic and social situations rather than looking at variability by itself. For example, a recent study (8) shows that the monthly pattern of births is very similar for small villages classified according to soil type. Unfortunately very few studies of this kind have been carried out possibly because of the pessimistic belief that the populations compared are too small. We may find that consistent results can be obtained from villages even smaller than the critical size we estimate.

Further, as mentioned above in examining individual small villages, despite the commonness of the practice, the effectiveness of trying to compensate for a small population by increasing the length of the time series has not been assessed. We have attempted to examine this by estimating our mean and coefficient of variation figures over 20, 30 and 50 year time spans. It should be useful for future investigations in this field to note that increasing the time span appears to be rather a useless procedure. Including more years of data increases the coefficient of variation from births, marriages or deaths (over the minimum of 20 observations that we used), and in some cases this may be sufficient to offset the reduced standard error of the mean number of births or marriages, obtained by averaging over more years.

In the following sections we substantiate these results in detail.

II. DATA

The major part of this study analysed birth, marriage and death data for Belgium over those parts of the period 1610-1795 for which data was available. The birth, marriage and death figures were obtained from 68, 46 and 75 rural villages respectively and urban results were obtained from Brugge and Ghent. To extend the generality of our findings we also examined birth and marriage figures for thirteen French villages from 1590-1809 (or part thereof) and for three parishes of Meulan France from 1590-1789.

(8) An example of such a study was B. SPENCER, D. HUM and P. DEPREZ, *Spectral Analysis and the Study of Seasonal Fluctuations in Historical Demography*, «Journal of European Economic History» in the press. The study shows that the seasonal pattern of fluctuations in births is consistent between even some very small villages and the larger ones if they were grouped according to soil type.

Since the 17th century births and marriages were found to be more variable for a given population size (possibly due to the greater prevalence of crises in the 17th century) we separated the 17th and 18th century figures and treated them separately. For the 17th century the data was broken into 20 year intervals starting 1610, concluding 1709. The corresponding 30 and 50 year intervals were from 1610-1699 and 1646-1695 respectively. For the 18th century, the 20, 30 and 50 year intervals ran from 1710 to 1789, 1700 to 1789, 1691 to 1795 respectively.

III. THEORETICAL RELATIONSHIPS

As explained in Section I, the aim of this study is to find out how the size of village or sample affects the variability of the data. In this section we shall set out the relationship between mean size of village or parish and the level of variability of births or marriages that might be expected under «ideal» conditions.

We use the coefficient of variation, that is the standard deviation of the births or marriages divided by its mean, as the measure of variability, for reasons which will soon become apparent. Observing the data (for example, see Graphs I, II, V and VI) we find that as the size of the population increases, the coefficient of variation for both births and marriages decreases at a decreasing rate. Since from these graphs the distribution of the observations of the coefficient of variation about the regression line seems to be normal, the coefficient of variation would seem to be an appropriate measure of variability for the purpose of finding the regression line. On the other hand if the mean number of births or marriages is plotted against the standard deviation we find, not unexpectedly, that the standard deviation increases with the larger size of population but that any line fitted to the observations has a heteroskedastic error term (Graph not shown). Thus there are complications in fitting a regression line directly. For further discussion see equation [3] below in this section.

The form of the curve to be fitted between the coefficient of variation and the mean is suggested by the following theoretical considerations.

Let us assume that each village or parish can be thought of as a random sample of size v from a common population where the probability of birth p remains constant. If the number of births as a proportion of the population of women at risk remains constant apart from chance deviations, the number of births per year is distributed as a binomial distribution.

A similar theoretical relationship would hold for marriages. This of course is unlikely unless the social situation, age and sex composition of the villages and parishes are the same. Nevertheless the theoretical relationship can be used to determine the form of the curve to fit to the data and is also useful as a base against which the actual data can be compared.

On the assumption that the number of births follows a binomial distribution the mean or expected number of births from a village or parish of size v is vp.

i.e. $$E(X_{v,t}) = vp$$

where $X_{v,t}$ is the actual number of births from a village or parish of size v. Also the expected variance is given by

$$E(S_v^2) = vp(1-p)$$

where S_v is the sample standard deviation from a village of size v.

In order to examine the effect of extending the sample by increasing the time period, means and standard deviations were calculated over intervals of 20, 30 and 50 years (see further discussion in Section V). In this context the level of births for each year, $X_{v,t}$, could be considered as one observation generated by the binomial distribution, so that the mean number of births from a village of size v calculated over a time interval is the average of the 20, 30 and 50 observations taken. Thus the expected value of the mean number of births from a village of size v calculated over a time interval would also be vp.

$$E(x) = E(X_{v,t}) = vp \quad \text{where} \quad x = \sum_{t=1}^{n} \frac{X_{v,t}}{n}$$

Also the probability limit of the standard deviation S, calculated over n years, is the underlying binomial standard deviation $\sqrt{vp(1-p)}$ (9). Thus, if the binomial situation holds the standard deviation should increase as the mean number of births increases. Our empirical data also followed this pattern.

$$\begin{aligned} plim\ y &= \frac{plim\ S}{plim\ x} = \frac{\sqrt{vp(1-p)}}{vp} \\ &= \frac{\sqrt{(1-p)}}{\sqrt{vp}} \\ &= \frac{\sqrt{(1-p)}}{\sqrt{E(x)}} \end{aligned} \tag{1}$$

Following this theoretical relationship the form of the regression chosen was

$$y = a + (b/\sqrt{x}) + e \tag{2}$$

where

y is the coefficient of variation
x is the mean number of births or marriages
e is the error term

Since the probabilities of birth or marriage are in fact not constant over

---

(9) The standard deviation S used to calculate the coefficient of variation, y is defined as:

$$S = \sqrt{\sum_{t=1}^{n}} \frac{(X_{vt} - x)^2}{n-1}$$

the period, we would expect greater variability in the actual data than that predicted by the theoretical model. This increased variability would be reflected in the coefficients a and b. To the extent that the increased variability is additive, the constant term, a, would be greater than zero, since the strict binomial formulation predicts that a should be zero. The coefficient b will also be affected and therefore not necessarily a good estimator of $\sqrt{(1-p)}$.

Also given that the relationship between the coefficient of variation $y = \frac{S}{x}$ and mean births or marriages x is of the form [2], we have

$$S = a\,x + (b\,\sqrt{x}) + e\,x \qquad [3]$$

Equation [2] shows that the error term would be heteroskedastic as was found in the actual data by plotting S against x (as mentioned above in this section).

IV. BIRTHS – CHARACTERIZATION OF THE DATA

We shall discuss the birth results for Belgium first. As explained in Section III theoretical considerations suggest that equation [2] may give a good fit to the data. This is indeed the case. Tables 1 and 2 give the equations obtained for the 17th and 18th centuries respectively. Each table shows rural and urban results separately and the effect of averaging the data over 20, 30 and 50 year periods. The coefficient of variation, which is less than 1, was in each case multiplied by 100 for the purposes of the regression. Each equation is significant at better than 0.1% probability. This was true even for the urban results of the 17th century where the proportion of the variability explained by the regression was around 45% since the F value required for significance at 1 and 38 degrees of freedom at the 0.1% level was only 12.71 (corresponding actual $F = 31$).

From examination of the data, it can be seen that there is a long tail of observations at roughly greater than 99 mean births which seem to follow an approximately linear form (see Graphs I and II) (10). A straight line regression

---

(note n. 9 follows)

where $$x = \sum_{t=1}^{n} \frac{X_{v,t}}{n}$$

$$E\,(S^2) = E\;\frac{\sum_{t=1}^{n} [(X_{vt} - vp) - (x - vp)]^2}{n-1}$$

$$= vpq$$

$$plim\;(S) = \sqrt{vpq}$$

(10) The choice of the mean number of births 99 is convenient since this captures most of the urban parishes with large populations. For purposes of comparison a straight line regression was calculated for mean births greater than 80 with very little difference.

TABLE 1

Comparison of Rural and Urban Birth Results—Belgium
Coefficient of Variation y Against Mean Number of Births x

						17th Century					
	$y = a + b/\sqrt{x}$	n	Years	sd.b	R^2	s.e.	F	$\bar{x}$	Sx	$\bar{y}$	Sy
Rural	$y = 5.84 + 96.78/\sqrt{x}$	177	20	4.97	0.68	5.70	379	43.56	32.70	24.34	10.11
	$y = 8.01 + 98.88/\sqrt{x}$	84	30	7.20	0.70	5.84	188	41.76	30.96	27.25	10.54
	$y = 9.36 + 87.00/\sqrt{x}$	28	50	10.61	0.72	3.48	67	46.31	29.32	24.42	6.46
Urban (a)	$y = 3.83 + 105.27/\sqrt{x}$	63	20	14.94	0.45	5.12	50	198.01	121.76	13.23	6.84
	$y = 5.34 + 92.19/\sqrt{x}$	40	30	16.66	0.45	4.72	31	192.42	122.47	13.79	6.26
	$y = 3.03 + 101.97/\sqrt{x}$	15	50	20.23	0.66	3.46	25	206.11	130.39	12.04	5.73
Rural & Urban (a)	$y = 4.92 + 100.42/\sqrt{x}$	240	20	4.01	0.72	5.55	626	84.10	96.29	21.43	10.56
	$y = 5.59 + 107.30/\sqrt{x}$	124	30	5.56	0.75	5.63	372	90.36	102.00	22.91	11.30
	$y = 4.43 + 109.15/\sqrt{x}$	43	50	8.55	0.80	3.89	163	102.05	110.27	20.10	8.57

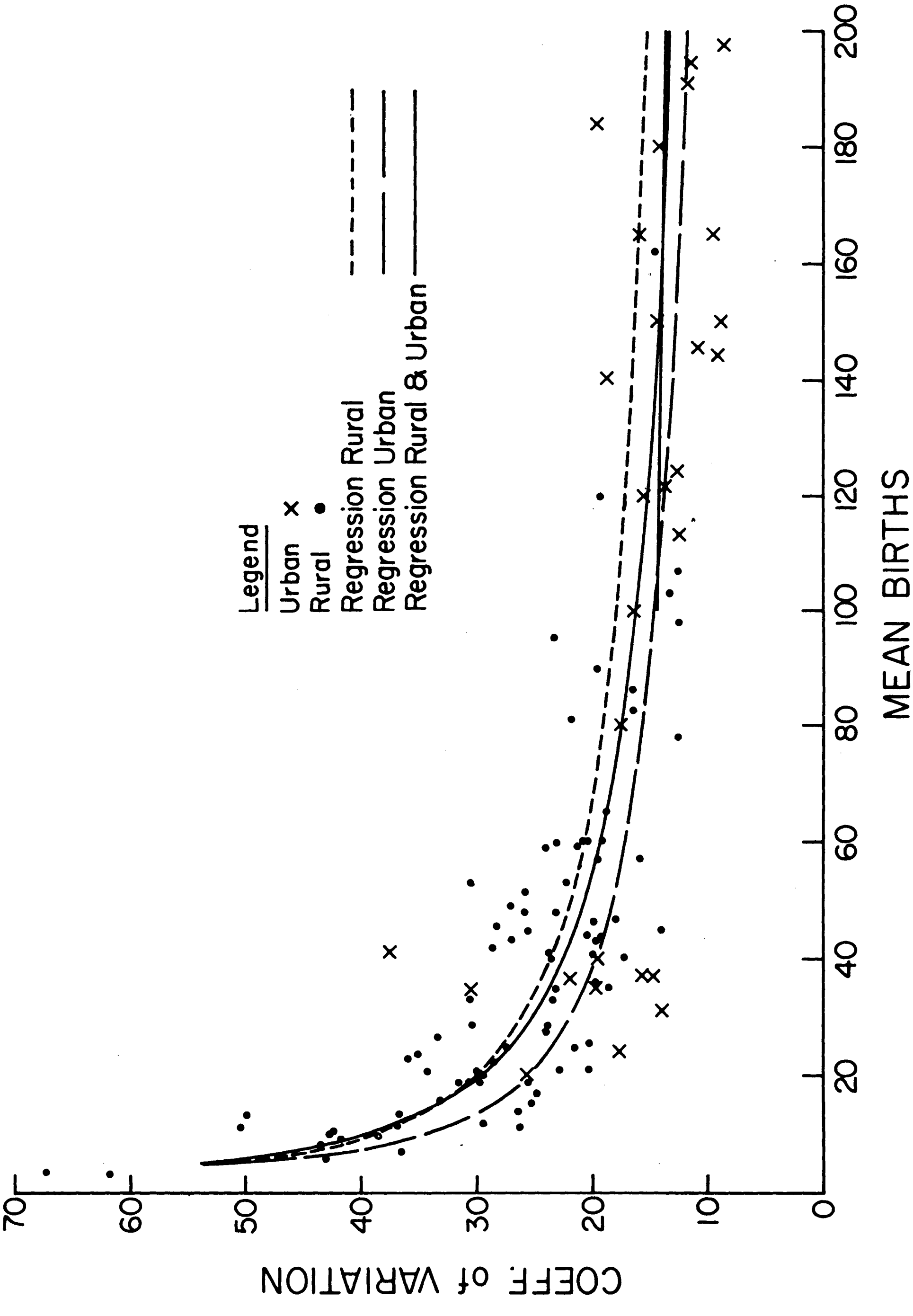
(a) The observation for Spanjaards Ghent 1646-95 is excluded from the estimation using a 50 year time interval. Births for Spanjaards fluctuate severely in this period.

TABLE 2

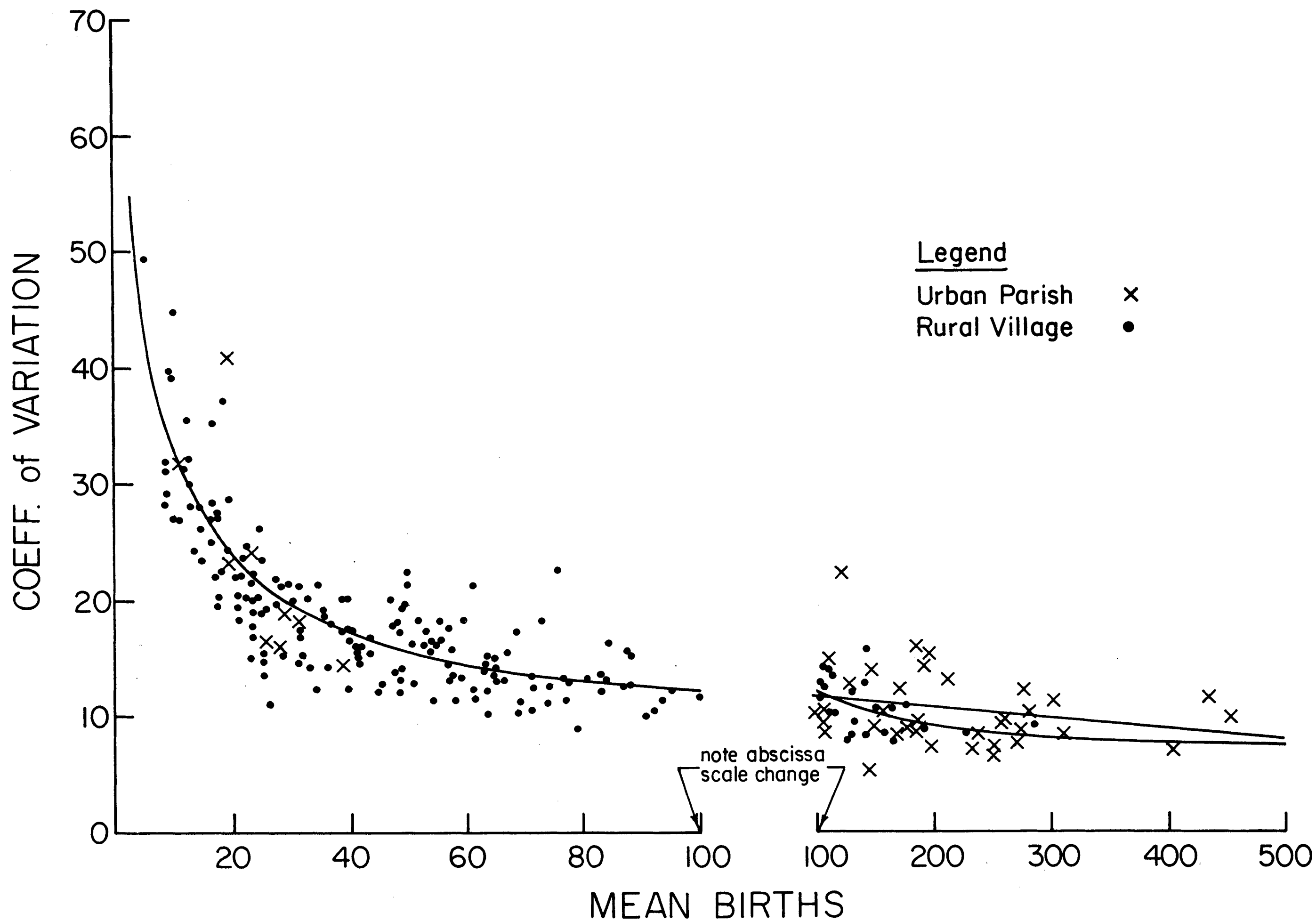
Comparison of Rural and Urban Birth Results–Belgium
Coefficient of Variation y Against Mean Number of Births x

	$y = a + b/\sqrt{x}$	n	Years	sd.b	R^2	$s.e.$	F	$\bar{x}$	Sx	$\bar{y}$	Sy
						18th Century					
Rural	$y = 2.78 + 84.78/\sqrt{x}$	257	20	3.45	0.70	3.56	605	55.46	42.57	16.71	6.53
	$y = 2.35 + 94.53/\sqrt{x}$	185	30	3.60	0.79	3.29	690	54.49	42.49	18.18	7.18
	$y = 5.75 + 85.77/\sqrt{x}$	117	50	5.55	0.68	3.77	239	56.41	43.19	19.74	6.59
Urban (a)	$y = -0.35 + 125.42/\sqrt{x}$	65	20	9.48	0.74	5.70	175	160.74	112.19	13.87	11.00
	$y = -2.64 + 171.80/\sqrt{x}$	48	30	22.27	0.56	10.47	60	165.95	113.88	16.10	15.69
	$y = 5.26 + 112.67/\sqrt{x}$	31	50	15.31	0.65	5.13	54	174.89	114.69	16.79	8.55
Rural & Urban (a)	$y = 2.01 + 91.69/\sqrt{x}$	322	20	3.44	0.69	4.30	711	76.71	75.80	16.14	7.71
	$y = 2.98 + 91.94/\sqrt{x}$	231	30	3.21	0.78	3.46	822	78.00	78.22	17.18	7.39
	$y = 6.30 + 85.25/\sqrt{x}$	148	50	5.18	0.65	4.22	270	81.22	80.60	19.12	7.11

(a) Spanjaards Ghent is excluded for the time periods 1696-1745, 1700-29, 1730-59, from the estimates using a 50 year time interval and a 30 year time interval respectively.



GRAPH I. 17*th* *Century Rural and Urban Births, Belgium – 30 Year Time Interval.*



GRAPH II. *18th Century Rural and Urban Births, Belgium – 30 Year Time Interval.*

TABLE 3

Comparison of Rural and Urban Birth Results – Straight Line
Relationship Between Coefficient of Variation y and Average Number of Births x

	$y = a + bx$	n	Years	sd.b	R^2	s.e.	F	$\bar{x}$	Sx	$\bar{y}$	Sy
				17th Century – Average Yearly Births Greater than 99							
Urban only	y = 13.10 – 0.01x	50	20	0.00	0.08	3.2	–	240	99	10.9	3.3
Urban only	y = 14.29 – 0.01x	31	30	0.01	0.13	3.1	–	238	99	11.5	3.3
Urban only	y = 11.97 – 0.01x	12	50	0.01	0.15	2.2	–	249	107	9.9	2.3
Rural & Urban	y = 14.13 – 0.01x	62	20	0.00	0.13	3.2	9.10	219	100	11.4	3.4
Rural & Urban	y = 15.02 – 0.01x	36	30	0.01	0.18	3.0	7.46	222	101	12.0	3.3
Rural & Urban	y = 13.42 – 0.01x	14	50	0.01	0.26	2.4	4.25	230	110	10.5	2.7
				18th Century – Average Yearly Births Greater than 99							
Urban only	y = 11.51 – 0.01x	48	20	0.00	0.12	2.7	6.09	209	88	9.2	2.8
Urban only	y = 12.86 – 0.01x	36	30	0.01	0.08	2.2	3.12	213	90	10.6	3.4
Urban only	y = 15.22 – 0.01x	24	50	0.01	0.03	3.8	0.76	219	91	13.6	3.7
Rural & Urban	y = 11.53 – 0.01x	79	20	0.00	0.12	2.5	5.60	184	81	9.4	2.7
Rural & Urban	y = 12.81 – 0.01x	59	30	0.00	0.10	2.8	6.33	186	83	10.7	2.9
Rural & Urban	y = 14.20	39	50	0.01	0.01	3.3	0.54	190	84	13.3	3.3

was fitted to the observations with mean births greater than 99 to check whether the relationship obtained from the binomial distribution, equation [2] Section III, gave a good fit even in the long tail. These straight line estimations are given in table 3 and those for the 30 year time interval are plotted on graphs I and II for the 17th and 18th centuries. As can be seen from these graphs equations of type [2] and the straight line give very similar estimates.

The separation of the 17th and 18th century data in the analysis seems to be necessary since the regression equations (see Tables 1 and 2 and Graphs I to IV) reveal that the more crisis prone 17th century generally has a higher coefficient of variation than the 18th century for any given level of mean births. One would expect that a time of crisis such as war or famine would increase the variability of vital statistics such as births, marriages and deaths. The average coefficient of variation $(\bar{y})$ as printed in tables 1 and 2 masks this relationship somewhat since for the urban parishes the average number of births is higher in the 17th century than the 18th century. For this urban group, the 17th century level of variability seems lower than that of the 18th century but this is explained by the fact that a larger number of births per year is associated with a lower coefficient of variation. When those villages and parishes with mean births between 0 and 20 and between 20 and 99 are separated out, the higher variability in the 17th century relative to the 18th century becomes obvious (see Table 4).

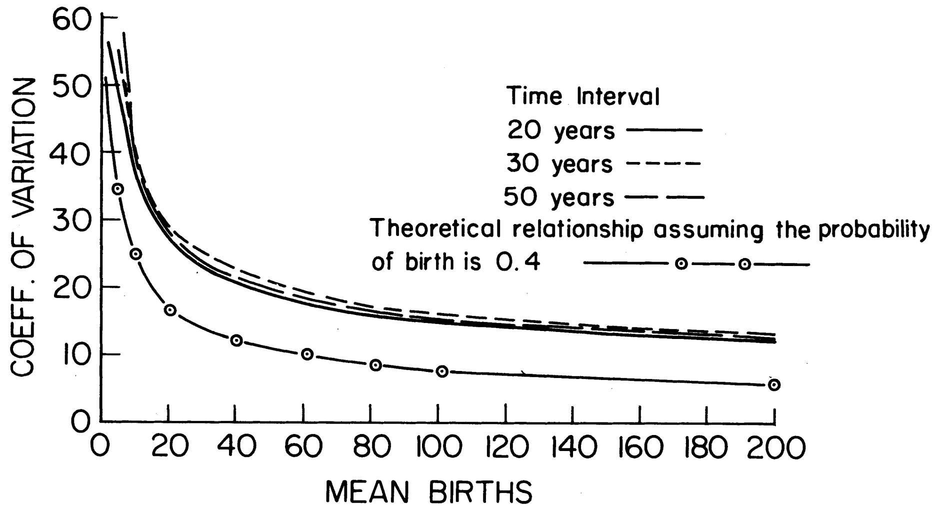
The relationship between the expected values of the coefficient of variation and the expected mean number of births assuming a strict binomial formulation with a probability of birth of 0.4 (that is equation [1] of Section III) is also plotted on graphs III and IV. There is a problem in determining the probability of birth especially for this pre-industrial period since it should be based on the number

TABLE 4

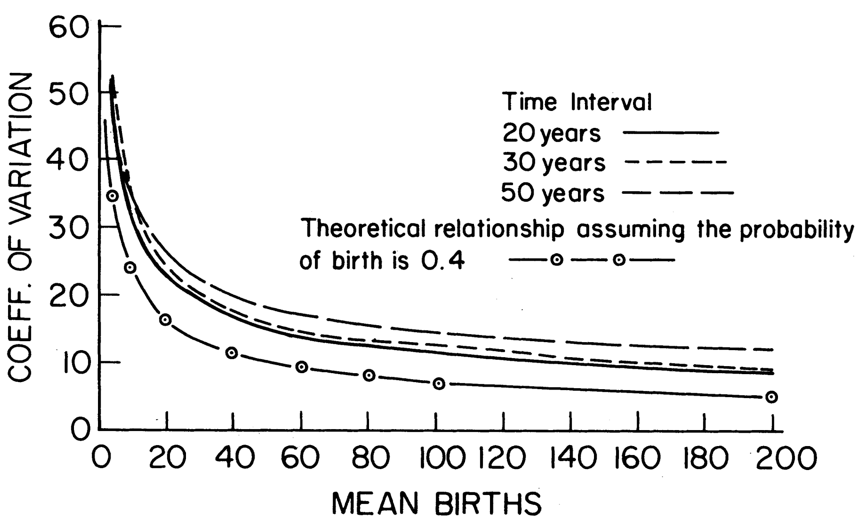
Average Coefficient of Variation $(\bar{y})$

		17th Century			18th Century		
Range (a)		20 yrs.	30 yrs.	50 yrs.	20 yrs.	30 yrs.	50 yrs.
$0 < X < 20$	Rural	34.82	38.18	33.88	26.71	29.99	29.62
$20 \leq X < 99$	Rural	21.35	23.70	23.15	15.61	17.18	18.44
$X > 99$	Rural & Urban	11.43	11.92	10.52	9.42	10.52	13.32
$X > 99$	Urban	10.85	11.50	9.92	9.21	10.24	13.68
$X > 99$	Rural	13.85	—	—	9.75	10.96	12.91

(a) X is mean number of births.



GRAPH III. *17th Century Rural and Urban Belgium Births.*



GRAPH IV. *18th Century Rural and Urban Belgium Births* of women at risk, which is not always known, and because the probability of birth varies with the age of the woman. The general fertility rate, which relates births to total number of women aged 15-49 would underestimate the probability of birth since many of the single women may not be at risk. The general legitimate fertility rate which relates legitimate births to married women aged 15 to 49 would give a better estimate, although it is also by no means an exact measure of this concept, since the probability of birth will vary over this age group. Our estimate of a probability of birth of 0.4 is based on estimates of this general legitimate fertility rate (11). In any case, the expected coefficient of variation is not highly sensitive to the choice of the probability of birth. For example, the assumption made by R.D. Lee (12) that the probability of birth is 0.5 reduces the expected coefficient of variation at 20 births per year to 15.82 rather than the 17.32 obtained by assuming the probability is 0.4. From graphs II and IV it is obvious that the empirical estimates of the coefficient of variation are above this theoretical curve. This greater variability most likely follows from that fact that the probability of birth is not constant over this period. The empirical estimates are much closer to the theoretical estimate using 18th century data, particularly when the data was averaged over only 20 years. The greater variability of the 17th century could possibly be explained by fluctuating birth rates during the crisis periods which were rather more common in this century.

---

(11) The general legitimate fertility rate is, *Total number of births (1 – illegitimate births as % of total births)/ Married women aged 15-49.* Data for Belgium indicate that for the rural parishes the number of illegitimate births is less than 2% of the total number of births. For urban parishes the incidence of illegitimate births is higher and may in some cases reach 7%. We use these figures together with estimates of the general legitimate fertility rate. The average general fertility rate calculated for 10 rural parishes and the total population of Brugge is 138 per 1000 with 200 and 90 being the upper and lower limits. The estimate of the average general legitimate fertility rate is then 321% with 386 and 195 being the upper and lower limits. Brugge has a rate of 311%. Since some of the women over 40 may not in fact be at risk and for other reasons suggested in footnote (12), we use a probability of birth of 0.4 in our estimates. I am indebted to Professor P. Deprez – University of Manitoba – for the information in this footnote.

(12) See R.D. LEE, *Methods and Models for Analysing Historical Series of Births, Deaths and Marriages.* R.D. LEE, R. EASTERLIN, P. LINDERT and E. van de WALLE, eds., *Population Patterns in the Past,* (New York, Academic Press, forthcoming 1977).

I am indebted to R.D. Lee for drawing my attention to this paper at the April 1976 Population Association of America meeting where I delivered an earlier draft of the current paper. Lee's estimate of the probability of birth of 0.5 is based on the Gautier and Henry estimate (see *La population de Crulai,* Paris, Presses Universitaires de France, 1958, p.115) of the general legitimate fertility rate of 0.4. Allowing for pregnancy and postpartum amenorrhea, Lee raises his estimate of the probability of birth to 0.5. Our admittedly rather rough estimate of the general legitimate fertility rate was less than Gautier and Henry's perhaps because their study, based on family reconstitution methods, referred only to those women who continued to reside in the same parish. It does not include those women at risk who moved in their child bearing years and who may have had a lower fertility.

It is interesting to note that for the 17th century the addition of urban observations to the rural observations generally reduces the estimates of the variability of the data. This is clearly shown for example in graph I where means and variances are calculated using a 30 year time interval. The regressions based on 50 year averages also show the same phenomenon althougt the difference between the rural and urban results with 17th century data based on 20 year averages is not so pronounced. On the other hand for the 18th century the regression equations are hardly affected by the addition of urban data. In discussing this point it must be remembered that the preponderance of urban parishes is in the range of mean births greater than 99, and vice versa for the villages. In fact in the range of mean births greater than 99,villages comprise 19, 14, 14 percent of the observations from the 17th century for the 20, 30 and 50 year intervals respectively and 39, 39, 38 percent of the observations from the 18th century from the same intervals respectively (see Table 3). Despite this, the difference between the average coefficient of variation for villages and urban parishes becomes apparent when the rural and urban data is estimated separately in this range of mean births. For example, using a twenty year time interval the average coefficient of variation from the 12 rural parishes in the range with mean births greater than 99 for the 17th century was 13.95 relative to only 10.85 for the 50 urban observations in this range (13). On the other hand, in the less crisis prone 18th century, for the same time interval and mean range of births the average coefficient of variation for the 31 rural villages was 9.75 only slightly above the 9.21 figure for the urban parishes (see Table 4). The interesting question that is presented here is the possibility that urban parishes were less susceptible to crises than their rural counterparts. That is rather hard to believe on the basis of other historical evidence although we can make the initial suggestion that the wider supply lines of the urban centers reduced the risk of an extreme fall-off in food supply. It must also be remembered that this analysis does not take migration into account. If in periods of crisis there was migration to the cities, this might account for the lower variability of births in the city.

---

(13) Of course the fact that the urban parishes are on average larger than the rural ones will over-emphasise the difference. However the straight line regression equations (see Table 3) confirm that there is a difference.

V. VARIABILITY OF BIRTHS; EFFECTS OF INCREASING THE SAMPLE SIZE BY EXTENDING OBSERVATIONS OVER TIME OR OVER SPACE.

As discussed in the introduction, it is important to discover whether the approach of increasing the number of years of observations in an attempt to compensate for a small village or parish population, that is a small number of observations at a point of time, can be justified by its results. The impetus behind this approach no doubt lies in the fact that the standard deviation of a mean is the standard deviation of the observations divided by the square root of the number of observations. Thus everything else constant, an average based on a larger number of observations will give a better prediction of the true mean. As is shown in footnote (14) the expected size of the confidence interval for the true mean number of births expressed as a proportion of the mean is $2 \frac{z_w}{\sqrt{n}} y$ where z_w is the appropriate value from the table of the normal distribution for a $w\%$ confidence interval, y is the coefficient of variation and n is the number of years. Increasing the number of observations from 20 to 50, if y is unaffected,reduces the confidence interval by one third. However, this reduction may be offset if y increases as the number of years over which the births are averaged increases. It was for the purpose of testing this, that we calculated the mean and coefficient of variation y of births averaged over 20, 30 and 50 year time intervals. As mentioned previously the regression equations for these time intervals are shown in tables 1,2 and 3. Graphs III and IV show that for both the 17th and 18th centuries the lowest variability is indeed obtained from the twenty year time interval. Although the difference between the various estimates in the 17th century is not very great, variability is significantly higher in the 18th century using the 50 year time interval rather than the 20 year interval. In fact at a village or parish size of 200 births per year in the 18th century the coefficient of variation, y, is increased by 45% by averaging over 50 rather than 20 years. At 100 births per year the increase in variability is reduced to 33% which would just exactly offset the reduced size of the confidence interval due to the number of years of observations alone.

The explanation for the increased variability of mean births when estimated over the longer time interval is most likely found in changes in the underlying situation over time. For example, if the average number of births in a village or parish increases within the period, the estimate of variability is immediately increased. Of course the longer the time period the more likely this is to occur. Apart from this direct effect on variability this is the real reason why averaging over a larger time period is likely to be counter-productive. Relationships between economic and demographic statistics will not be found if they are obscured by averaging.

We are now in a position to approach the main question of this paper.

What is the approximate mean number of births after which there is relatively little change in the coefficient of variation, so that a researcher could be safe in the knowledge that increased size would not give an improved estimate. By this we mean that if the coefficient of variation, y, is constant, the estimate could not be improved in the sense that the confidence interval for the true mean number of births would remain unchanged in percentage terms, (that is as a percentage of the true mean number of births). This follows from the expression for the percentage confidence interval $2\frac{z_w}{\sqrt{n}}y$ derived in footnote (14).

From graphs III and IV it can be seen that the coefficient of variation initially decreases very sharply. As the number of births per year increases the gain in reduced variability becomes less. A reasonable, although somewhat arbitrary choice of the critical number of births is where the slope of the regression line is one half (15). That is, at this critical value, an increase in the mean births by 10

---

(14) The $w\%$ confidence interval for the true mean number of births on the assumption that the probability of birth remains constant is

$$X_{v,t} - z_w\sqrt{vp_1q_1} < vp < X_{v,t} + z_w\sqrt{vp_1q_1}$$

where $X_{v,t}$ is the number of births in a parish at time t

p_1 is the estimated probability of birth

v is the size of the village or parish

z_w is the appropriate value from the normal distribution or the t distribution for a $w\%$ confidence interval.

The $w\%$ confidence interval for vp if $X_{v,t}$ is averaged over n years to obtain x is

$$x - z_w s_x < vp < x + z_w s_x$$

where $$plim\ (S_x) = \frac{\sqrt{vpq}}{\sqrt{n}}$$

Thus the $w\%$ confidence interval as a percentage of the mean is

$$2\ .\ \frac{z_w}{\sqrt{n}}\ \frac{\sqrt{vpq}}{vp} = 2\frac{z_w y}{\sqrt{n}}$$

(15) The critical value is calculated by solving $x = \sqrt[3]{b^2}$. The regression is of the form:

$$y = a + bx^{-\frac{1}{2}}$$

$$\frac{dy}{dx} = -b\ .\ \frac{1}{2}\ .\ x^{-\frac{3}{2}}$$

$$\frac{dy}{dx} = -\frac{1}{2}\ \text{when}\ x^{\frac{3}{2}} = b$$

reduces the coefficient of variation by 5. As an example take the 18th century with data averaged over 20 years. The estimated critical number of births is 20.37 with a coefficient of variation of 23. At an average birth level of 10 per year, the corresponding coefficient of variation was 31 with the slope of the regression equation at −1.45. With mean births at 30, the coefficient of variation and the slope have reduced to 19 (a drop of 4) and −0.28 respectively. The coefficient of variation increased by 35% if mean births fell from 20.37 to 10 per year but only decreased by 17% if the mean level of births was increased from 21.6 to 30 per year. Between mean births 30 and 40 the coefficient of variation fell by only 2 units. That is, past the critical level the decrease in the coefficient of variation is not substantial. Of course we could have chosen the critical level at say 100 births since at this and higher levels of births the data could almost be described as horizontal with no reduction in variability, (see straight line estimates in Table 3). This however would not suit our purpose since the penalty from studying smaller size villages is not very great. Table 5 shows the critical values for both Belgium and France.

The regularity in the results between the 17th and 18th centuries and between the different time intervals is quite striking. The mean number of births at which the critical value occurs, lies between 19.30 and 22.58. This indicates that as long as the population results in a level of births above 20 per year, most of the advantage of population size in reducing the coefficient of variation will have been achieved.

To check that these results were not specific to Belgium data we obtained information on 13 French villages and three urban parishes of Meulan, France. As in Belgium, births were more variable in 17th century France than in the 18th century. Furthermore at each level of average births, the coefficient of variation calculated over twenty year intervals was less than that obtained using the thirty or fifty year interval (16). The results are shown in table 6. Urban and rural data are analyzed together since from observation there seems to be little difference between them and the overall number of samples is small. It is also of great interest that the average number of births at the turning points of each of the fitted curves was almost identical to the Belgium results (see Table 5). This increases our confidence in the wider applicability of these results.

From an examination of these figures for individual villages it seems clear that the position of the regression equation of the coefficients of variation and thus the critical value is related to the average number of births rather than the population size of the village. Because of different age structures, the population size corresponding to a given average level of births can be quite variable. It appears that mean birth levels of 20 per year can be associated with populations roughly between 450 and 750. Thus a researcher should check the average level of births rather than the population size if he wishes to know if the village he is studying is large enough to have achieved most of the reduced variability due to size.

(16) This can be seen by plotting the regression equations given in table 5. Because the number of observations is small (only 3), the regression was not calculated for 50 years 17th century France.

TABLE 5

Critical Number of Births ·

	17th Century						18th Century					
	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.
I. Belgium Births												
Rural	177	21.08	84	21.38	28	19.63	257	19.30	185	20.75	117	19.45
Rural & Urban	240	21.60	124	22.58	43	22.83	322	20.37	231	20.37	148	19.37
II. French Births												
Rural & Urban	29	21.40	13	22.54	–	–	48	19.56	34	19.81	14	17.04

TABLE 6

Description of Birth Results–France
Coefficient of Variation y Against Mean Births x

	$y = a + b/\sqrt{x}$	n	Years	$sd.b$	R^2	$s.e.$	F	$\bar{x}$	Sx	$\bar{y}$	Sy
17th Century											
Rural & Urban	$y = 2.64 + 98.98/\sqrt{x}$	29	20	13.54	0.66	5.02	53	38.57	39.98	22.83	8.51
	$y = 3.38 + 106.99/\sqrt{x}$	13	30	17.31	0.78	4.13	38	48.76	50.99	23.37	8.36
18th Century											
Rural & Urban	$y = 2.46 + 86.54/\sqrt{x}$	48	20	8.29	0.70	4.36	109	30.34	23.76	21.44	7.92
	$y = 3.16 + 88.19/\sqrt{x}$	34	30	8.59	0.77	3.75	105	31.02	24.86	22.35	7.66
	$y = 9.01 + 70.34/\sqrt{x}$	14	50	19.52	0.52	5.02	13	32.35	22.83	23.52	6.96

VI. MARRIAGES – CHARACTERIZATION OF THE DATA

As expected from the theoretical considerations discussed in Section III, the relationship between the coefficient of variation and the mean number of marriages is similar to that found for births. In fact the regressions estimated on the basis of equation [2] of Section III have very similar coefficients although in each case the variability of the marriages is greater. This can be seen by comparing graphs III and IV for births with the corresponding graphs VII and VIII for marriages. The horizontal axis in graphs VII and VIII is expanded over that for births since generally the number of marriages is less than the number of births per thousand of population. In fact 8 marriages per 1000 seems to be an approximate estimate of the crude marriage rate for Belgium at this time, relative to the 35 or 40 per thousand crude birth rate. The theoretical relationship between the coefficient of variation and mean marriages on the basis of a probability of marriage (17) of 0.1 (follows from equation [1] and Section III) is also plotted in graphs VII and VIII. This theretical line is only 22% above the theoretical line for births based on a probability of birth of 0.4 despite the large difference in the binomial probabilities, since as mentioned in Section IV, the expected coefficient of variation is not very sensitive to the crude rate that is assumed.

As in the case of births, the regressions fitted according to equation [2] are an excellent fit (see Table 7). This is illustrated in graphs V and VI which show the observations on the basis of a 30 year time interval for the 17th and 18th centuries respectively. This separation of the 17th and 18th century is again necessary since the 17th century marriages are more variable, again a reflection of the crises of this time. For example, at 20 marriages per year using a 20 year time interval, the estimated coefficient of variation is 29.97 versus 25.30 for the 18th century. The expected figure at 20 births per year assuming a constant probability

(17) The choice of a suitable probability of marriage is somewhat difficult. In theory the group at risk is the number of single people although even this is complicated by the fact that people from one parish may marry in another. Fortunately this is not such a large problem for Belgium as there is a tendency for women to marry in the parish they live in. In fact, it is common to relate the number of marriages to the number of single women rather than men although this can cause complications especially if the frequency of marriage is different for males and females. A first approach to this question is to remove the remarriages and to calculate the total number of first marriages as a percentage of the total single female population aged 15 to 49. Assuming remarriages are around 17% our estimate of the average number of first marriages per 1000 single women in rural Belgium is around 55. This was obtained by averaging over 7 rural villages with first marriage rates ranging from 45 to 79 per 1000. This may be too low as an estimate of the probability of marriage since the older single women may have a very low change of marriage, and perhaps should not be included in the population at risk. To fully correct for this we would need figures on the numbers of single women in each age group. As a rather gross approximation we therefore use a probability of marriage of 0.1 in our calculations. I am indebted to Professor P. Deprez for the information and calculations in this footnote.

TABLE 7

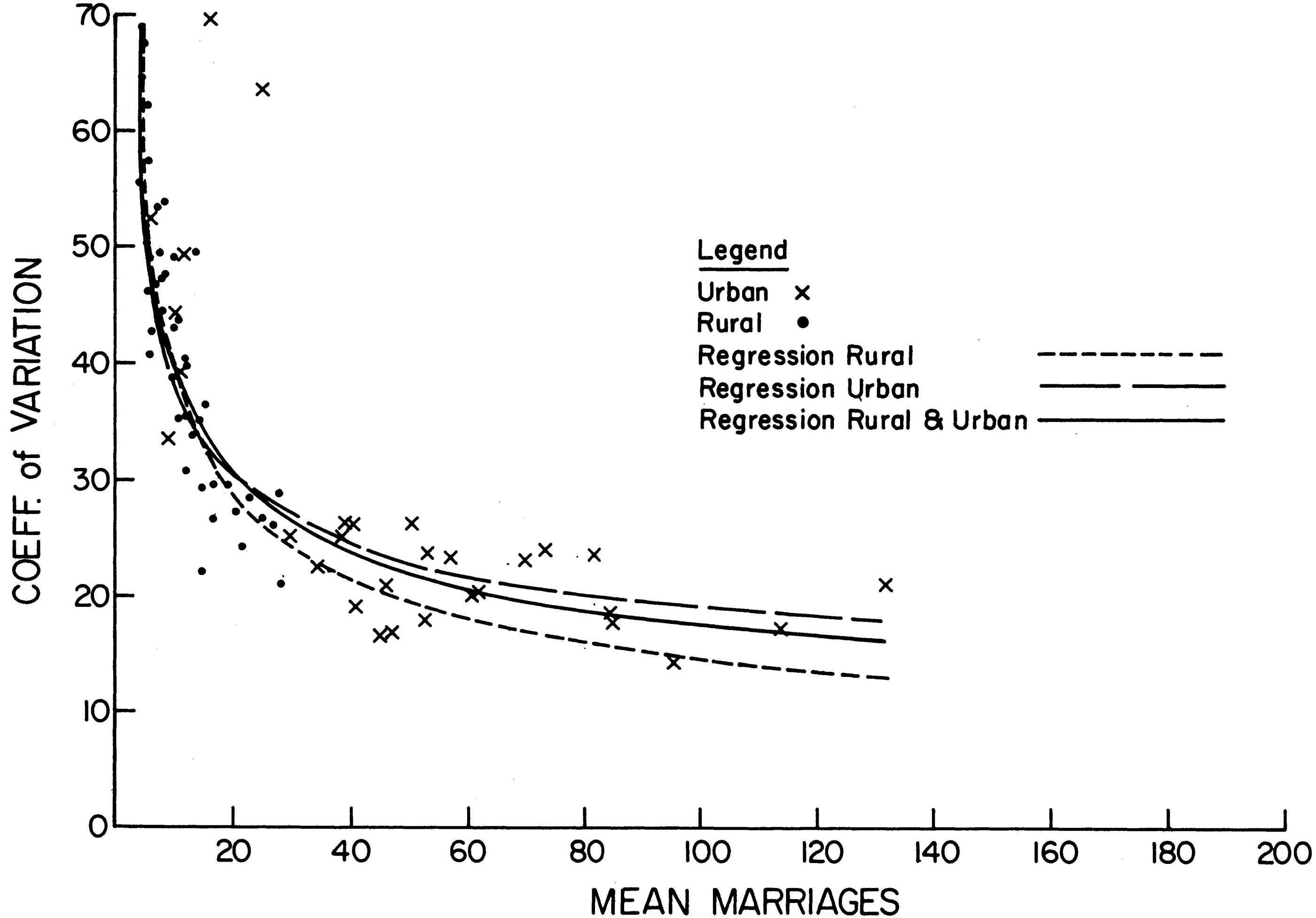
Comparison of Rural and Urban Marriage Results–Belgium
Coefficient of Variation y Against Mean Number of Marriages x

	$y = a + b/\sqrt{x}$	n	Years	$sd.b$	R^2	$s.e.$	F	$\bar{x}$	Sx	$\bar{y}$	Sy
							17th Century				
Rural (a)	$y = 3.35 + 113.63/\sqrt{x}$	110	20	6.61	0.73	8.36	295.91	12.16	7.98	42.10	16.08
	$y = 3.33 + 114.36/\sqrt{x}$	47	30	8.90	0.79	7.11	165.05	11.75	7.82	42.82	15.19
	$y = 7.53 + 108.48/\sqrt{x}$	19	50	7.55	0.94	2.83	206.73	52.26	32.46	26.39	10.03
Urban (b)	$y = 6.75 + 99.3/\sqrt{x}$	46	20	6.85	0.83	4.27	210.07	52.41	31.44	24.11	10.14
	$y = 10.04 + 90.75/\sqrt{x}$	30	30	7.97	0.82	4.01	129.74	51.31	31.64	26.09	9.35
	$y = 7.53 + 108.48/\sqrt{x}$	15	50	7.54	0.94	2.53	206.73	52.26	32.46	26.39	10.03
Rural & Urban (a.b)	$y = 4.67 + 110.03/\sqrt{x}$	156	20	4.35	0.81	7.39	639.42	24.03	25.90	36.80	16.72
	$y = 7.05 + 104.59/\sqrt{x}$	77	30	5.19	0.84	6.16	406.24	27.16	28.21	36.30	15.50
	$y = 9.22 + 97.16/\sqrt{x}$	34	50	4.69	0.93	3.61	429.20	29.76	29.88	35.12	13.50
							18th Century				
Rural	$y = -0.20 + 110.10/\sqrt{x}$	171	20	4.23	0.80	6.64	693	12.79	8.59	36.93	14.94
	$y = -0.69 + 114.88/\sqrt{x}$	124	30	4.75	0.83	6.56	584	12.75	8.53	37.90	15.72
	$y = 1.49 + 113.76/\sqrt{x}$	76	50	5.23	0.86	5.50	473	13.44	8.72	38.47	14.85
Urban (c)	$y = -0.53 + 124.69/\sqrt{x}$	64	20	6.41	0.86	6.93	378	43.10	29.37	25.88	18.32
	$y = -0.42 + 133.99/\sqrt{x}$	48	30	11.40	0.75	9.66	138	43.96	29.91	27.13	19.12
	$y = 3.48 + 118.94/\sqrt{x}$	31	50	10.39	0.82	6.04	131	45.83	29.73	26.52	13.96
Rural & Urban	$y = 0.41 + 111.29/\sqrt{x}$	235	20	3.29	0.83	6.85	1145	21.04	21.65	33.92	16.64
	$y = 0.96 + 113.27/\sqrt{x}$	172	30	4.37	0.80	7.82	673	21.47	22.26	34.90	17.37
	$y = 3.44 + 109.98/\sqrt{x}$	107	50	4.26	0.86	5.75	666	22.82	22.85	35.01	15.52

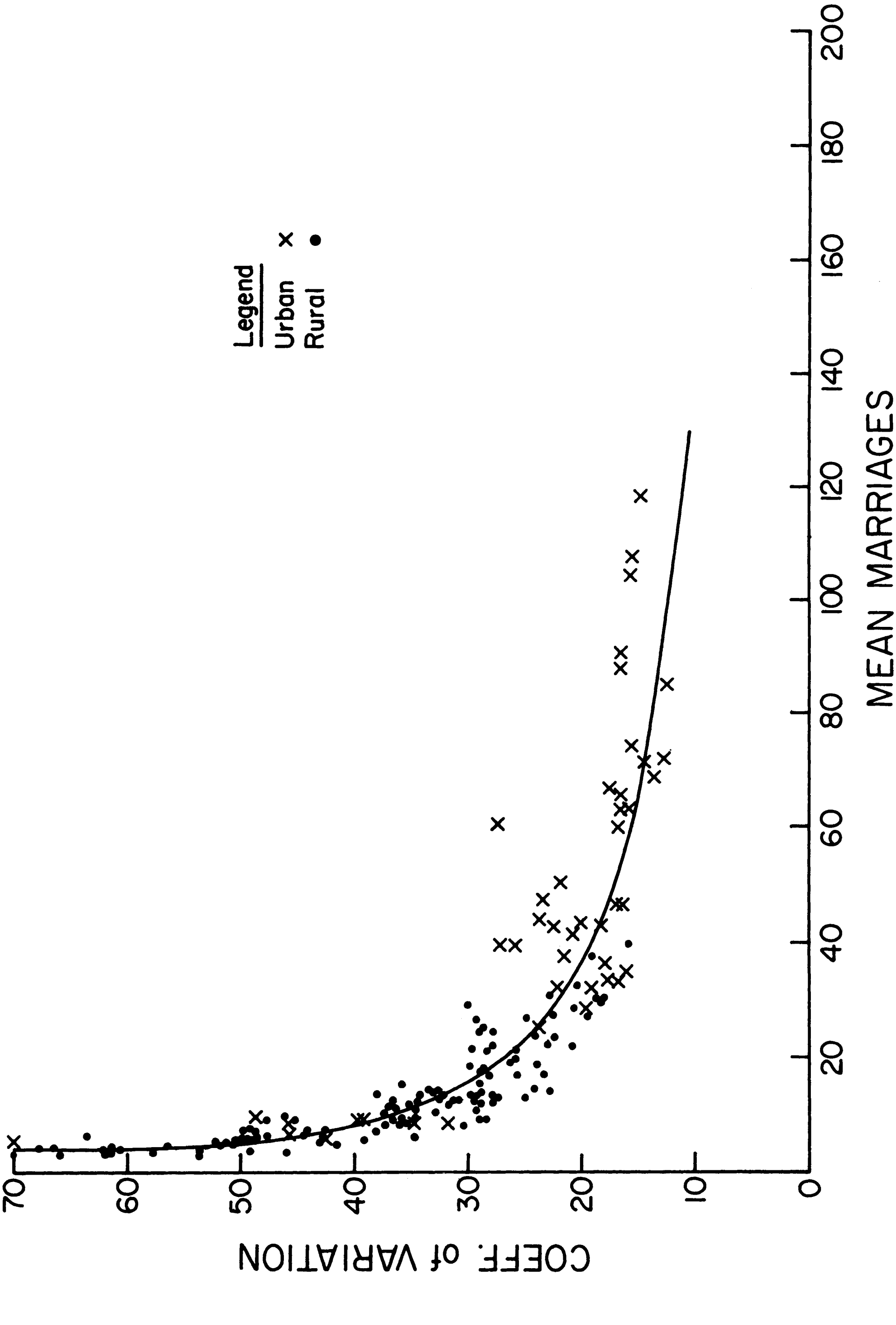
(a) Zeveren 1690-1709 which has mostly a zero yearly marriage rate is excluded from the estimation using a 20 year period.

(b) All the observations of Spanjaards, Ghent, falling in the 17th century groups are excluded because of a sharp rise and then fallin mean number of marriages in this period.

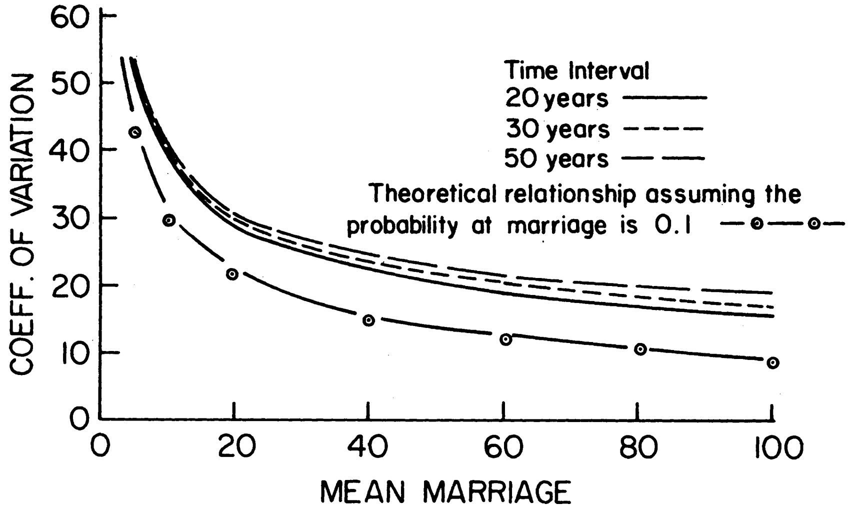
(c) Spanjaards, Ghent, 1696-1745 is excluded from the estimation using a 50 year interval.



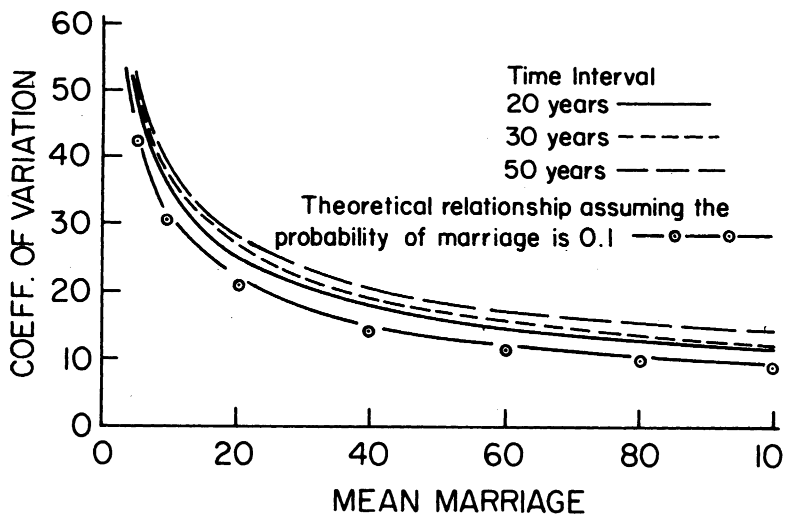
GRAPH V. *17^{th} Century Rural and Urban Marriages, Belgium – 30 Year Time Interval.*



GRAPH VI. *18th Century Rural and Urban Marriages, Belgium – 30 Year Time Interval.*



GRAPH VII. *17th Century Rural and Urban Belgium Marriages*



GRAPH VIII. *18th Century Rural and Urban Belgium Marriages* of marriage of 0.1 is 21.21 indicating that 72% of the value of the coefficient of variation is explained by random forces on the basis of the binomial distribution in the 17th century versus a rather higher 84% in the less crisis prone 18th century. This percentage remains consistently lower for the 17th relative to the 18th century ranging from 79% at 5 marriages per year to 61% at 100 marriages per year, with the corresponding figures at 85% and 82% respectively for the 18th century.

It may also be of interest to see these percentages expressed in terms of the more commonly used statistic, the variance, rather than the coefficient of variation (18). We find that the proportion of variance explained by purely random factors assuming a constant probability of marriage ranges from 62% at 5 marriages per year to 37% at 100 marriages per year in the 17th century with 72% to 67% respectively explained in the 18th century. Thus this proportion decreases as the number of marriages per year and the size of the village increase in the 17th century, with very little change in the 18th century. Of course this is not directly relevant in assessing the cost in terms of variability of studying a small rather than a large parish, since it is the actual variability of the number of marriages which is crucial here (see Section VII below).

Although the actual variability at any given mean number of marriages is higher than that for births, the proportion of the variance that is explained by purely random factors is greater for marriages than for births. This is a consequence of the fact that the probability of the expected coefficient of variation on the basis of the binomial distribution is lower by the 22% mentioned above. To take an example at 20 births per year and a 20 year time interval, the estimated coefficient of variation is 27.38 for the 17th century, 22.51 for the 18th in comparison with 29.27 and 25.30 respectively for marriages. The proportion of the coefficient of variation of the births explained in this case is 0.63 and 0.77 respectively. In terms of the variance, the proportions explained are 0.40 and 0.59 respectively, which are significantly below the corresponding figures for marriages at 0.52 and 0.71 respectively. This would appear to indicate that the variability of marriages is less affected by changes in the underlying probability of marriage than was the case for births.

There is a difference similar to that found for births between the variability of the rural and urban marriage data. For the 17th century the addition of urban observations to the rural observations generally reduces the estimates of the variability of the data. The maximum size of the difference is shown by the regressions in graph V in which is plotted the 17th century data averaged over 30 years. As in the case of births the difference is not as pronounced in the other estimates particularly when a 20 year interval is used. It must be remembered however that perhaps even more than in the case of births the marriages in rural villages are mostly at the low end of the scale whereas the higher levels of mean marriages are mostly

(18) Since the coefficient of variation is the standard deviation divided by the mean, the square of the ratio of two coefficients of variation will be the ratio of the variances.

found in the larger urban parishes. There are few rural villages with mean births above 24 and all of these have means under 45 with almost all below 30.As discussed under Section IV with regard to the birth data there is need for caution in interpreting this seeming difference in variability.

VII. VARIABILITY OF MARRIAGES: EFFECTS OF INCREASING THE SAMPLE SIZE BY EXTENDING OBSERVATIONS OVER TIME OR OVER SPACE.

We also wish to establish for the marriage data whether increasing the number of observations by lengthening the time period can compensate for the small size of the community from which data is collected. As is explained in the discussion of the birth data, the only possible gain by this approach is a reduced standard error for the mean number of marriages and thus a smaller confidence interval for the estimate of the true mean number of marriages. This advantage can however be offset by the increased variability of the observations themselves when they are calculated by averaging over a longer time period. The variability of the marriage data was in fact lower when averaged over 20 rather than 30 or 50 years. The magnitude of the effect depends on the actual data, presumably on the extent of the changes within a 50 year rather a 20 year period. The actual effect is illustrated in graphs VII and VIII and, as can be seen from the graphs and from the formula in footnote 14, is in some cases great enough to make any gain from averaging negligible. However even if there is a significant reduction in the standard error, this may be more than offset by the fact that the averaging may obscur important relationships between various economic and demographic data.

Following the same procedure we used for the birth data, the critical values for the marriages are calculated where the slope of the regression is $-\frac{1}{2}$. For example, for the 18th century using 20 year averages, the estimated critical number of marriages is 23.14 (see Table 8) with the corresponding coefficient of variation at 23.54. It is useful that this particular definition of the critical value also corresponds closely to an elasticity of the coefficient of variation with respect to the mean of about $\frac{1}{2}$ for both births and marriages. An increase in mean marriages of 1% will reduce the coefficient of variation by $\frac{1}{2}$%. Again for the 18th century figures averaged over 20 years, the coefficient of variation is 35.61 with a slope of -1.76 at 10 births per year and only 20.72 with a slope of -0.34 at mean marriages of 30 per year: Thus between 10 and the critical value of 23 marriages per year the coefficient of variation increased by 51%,but it fell by only 12% between mean marriages 23.54 and 30. The gain in increasing the size of the village or parish sampled in terms of variability is not very great past this critical value.

TABLE 8

Critical Number of Marriages

	17th Century						18th Century					
	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.	n	Critical no.
I. Belgium Marriages												
Rural	110	23.46	47	23.56	19	22.75	171	22.97	124	23.63	76	23.48
Rural & Urban	156	22.96	77	22.20	34	21.13	235	23.14	172	23.41	107	22.95
II. French Marriages												
Rural & Urban	25	20.53	10	16.31	–	–	48	22.17	33	22.26	15	21.54

TABLE 9

Description of Marriage Results–France
Coefficient of Variation y Against Mean Marriages x

Rural and Urban Observation Combined

Century	$y = a + b/\sqrt{x}$	n	Years	sd.b	R^2	s.e.	F	$\bar{x}$	Sx	$\bar{y}$	Sy
17	$y = 11.39 + 93.00/\sqrt{x}$	25	20	16.74	0.57	12.18	31	8.97	8.45	50.42	18.25
17	$y = 21.19 + 65.88/\sqrt{x}$	10	30	22.86	0.51	10.14	8.31	8.49	7.39	48.88	13.65
18	$y = 4.30 + 104.37/\sqrt{x}$	48	20	10.65	0.68	11.53	96	6.73	4.83	52.53	20.05
18	$y = 4.46 + 105.05/\sqrt{x}$	33	30	9.49	0.79	8.58	122	6.85	5.03	52.83	18.79
18	$y = 9.33 + 99.98/\sqrt{x}$	15	50	11.34	0.86	7.04	78	8.11	7.29	53.13	17.93

Table 8 shows the critical values for marriages for both Belgium and France. The regularity in the results between the 17th and 18th centuries and the different time intervals is again quite striking. Because of the small number of parishes for which we could get data for France in this period, the critical value is calculated only for the 20 year interval. Even so, the results from France are very similar to those from Belgium. The other interesting observation is of course the similarity of these critical values to the ones found for births. Given that we have found that the expected theoretical curves for marriages and births are very similar this is not so surprising, but it does underscore our comment in the section on births that it is not the size of the population in the parish that is critical in deciding what is small, but the particular variable of interest since the parish size corresponding to a mean level of marriages of twenty per year is rather more than that required for mean level of births of twenty.

VIII. DEATHS

An analysis similar to that for births and marriages was undertaken for deaths. As mentioned in section II we had data from 75 villages as well as parishes of Brugge and Ghent. The means were calculated for the 17th and 18th centuries and for 20, 30 and 50 year intervals. This data when plotted for the 17th century was so variable that it appeared almost at random. For example, using the 20 year average interval, the coefficient of variation ranged from roughly 50 to 130 for the range of mean deaths up to 10 per year, ranged between 30 and 130 for mean deaths between 10 and 20 and still rose as high as 125 in the interval corresponding to 20 to 30 deaths per year. This variability can of course be attributed to the wars and famines of the 17th century referred to in previous sections. Not unexpectedly, the data was rather more uniform for the 18th century. The coefficient of variation formed a thick band roughly 20 percentage points wide. This band appears approximately horizontal after about 20 deaths per year at a height of about 20 to 40. Given the wide variability of the death data there seemed little point in subjecting it to the further detailed analysis accorded the births and marriages unless the data were separated according to crisis and non crisis years.

IX. CONCLUSION

In this paper we have analysed the relationship between the mean and variability of the number of births and marriages in villages and parishes in the 17th and 18th centuries. The results show that although the standard deviation of the births and marriages increases with the size of the village or parish, the coefficient of variation which is the standard deviation as a percentage of the mean decreases with size. In fact, the coefficient of variation decreases at a decreasing rate indicating that although there is an advantage in using large samples because of their reduced variability, this advantage becomes much less after a certain point. When the coefficient of variation becomes constant (which it almost does in the long tails of the observations) we show that the confidence interval for the true mean expressed as a percentage of the mean is also constant, so that there is no further gain from studying larger parishes.

A good description of the data is obtained by fitting a curve derived from the assumption that births and marriages follow a binomial distribution. We have used this curve to estimate a «critical» size after which the gain from studying a village or parish with a larger population is minimal. As we argued in the introduction, models based in highly disaggregated data of an economic, social and demographic nature seems likely to be the most fruitful approach to work in this area.

It should be noted that the «critical» size is determined, not by population size, but by the level of variables of interest such as births or marriages. Since the theoretical estimates of the coefficient of variation obtained by making strict binomial assumptions are very similar (a little higher for marriages than for births) the estimates of the critical numbers of births and marriages are also very close (around 20 and 22 respectively). These levels of births and marriages correspond to different population sizes.

Increasing the apparent size of sample by taking a larger geographic area or increasing the number of years over which the data is collected most often would not improve the situation. With the first approach the population is unlikely to remain homogeneous so that the study will produce only averaged results hiding the underlying relationships. With the second approach, we have shown in this paper that the expected reduction in the standard error of the mean may not materialize since it is partially offset by increased variability in the actual data. We found that the variability of the data for both births and marriages was least when using our shortest time interval, twenty years. This increased variability is a reflection of changes in the underlying situation and in general the longer the time period the greater the risk that basic parameters will change. This indicates that aggregation over time is also liable to obscure the real values of the variables under study and could give very misleading results.

In the past a large number of studies have suffered from being based on too highly aggregated data, both on a geographical basis and over time. In fact, because of this, the massive amounts of results produced in Belgium and France have lost a considerable part of their potential usefulness. We hope that in future more attention will be given to studying more homogeneous but smaller populations over carefully defined time periods.

SUMMARY

This study of the variability of births and marriages is an attempt to examine some of the possible consequences of basing economic and demographic analysis on small populations. Small populations have not been extensively used because of the fear that the variability of the data may be too high for reliable estimates. Using mostly Belgium birth and marriage data from villages and parishes in the 17th and 18th centuries, we show that the coefficient of variation decreases with increased sample size but at a decreasing rate. In fact the advantage of increased sample size in reducing the variability of the data becomes rather insignificant at relatively low numbers of births and marriages per year. Furthermore since the functional form fitted is based on theoretical considerations arising from the binomial distribution, we have reason to believe that similar results will be found with other data. This would seem to indicate that relationships obtained from relatively small villages and parishes may be just as reliable as those from larger populations for the purpose of inferring the existence of the relationships in other populations experiencing similar economic circumstances.

RIASSUNTO

Questo studio della variabilità delle nascite e dei matrimoni si propone di valutare alcune delle possibili conseguenze che scaturiscono dall'analisi economica e demografica quando questa venga condotta su piccoli gruppi. Le ricerche su piccoli gruppi non sono, generalmente, molto utilizzate per il timore che l'elevata variabilità dei dati pregiudichi l'attendibilità delle valutazioni. Ricorrendo soprattutto a dati su nascite e matrimoni di comuni e parrocchie belghe nel XVII e XVIII secolo, si constata che il coefficiente di variazione tende a diminuire quando aumenta l'ampiezza del campione, ma in misura decrescente. Infatti il vantaggio derivante dall'adozione di campioni più grandi al fine della riduzione della variabilità dei dati, diviene presto piuttosto trascurabile, anche in corrispondenza di un numero annuo di nascite e matrimoni ancora relativamente basso. Inoltre, dal momento che la forma di relazione funzionale prescelta si basa su presupposti teorici collegati alla distribuzione binomiale, si è portati a credere che i risultati ottenuti su dati di villaggi e parrocchie relativamente piccoli, possano avere la medesima attendibilità di quelli stimati su popolazioni più ampie, allorché ci si prefigga di inferire l'esistenza delle stesse relazioni in altre popolazioni che sperimentano analoghe condizioni economiche.

RÉSUMÉ

Cette étude de la variabilité des naissances et des mariages est un effort pour examiner certaines des conséquences d'une analyse économique et démographique basée sur des petites populations. Des petites populations n'ont pas été beaucoup utilisées par crainte que la variabilité des données soit trop grande pour des évaluations sûres. Employant principalement des dates de naissances et de mariages des villages et des paroisses belges aux dix-septième et dix-huitième siècles, on montre que le coéfficient de variation diminue quand on augmente le nombre d'observations mais il diminue à un taux décroissant. En effet, l'avantage d'un plus grand échantillon pour réduire la variabilité des données devient assez minime quand le nombre de naissances et de mariages enregistré par an est relativement bas. En outre, puisque la forme fonctionnelle employée est basée sur des considérations théorétiques provenant de la distribution binomiale, on a lieu de croire que des résultats obtenus dans des villages et paroisses relativement petits pourraient être tout aussi sûrs que ceux obtenus dans des populations plus grandes quand on a l'intention d'inférer l'existence de relations dans d'autres populations ayant des conditions économiques pareilles.