The Survey of Consumer Finances:
Sampling and surveying in Cyprus

Alex Karagrigoriou[1]

University of Cyprus

January 27, 2005

1. INTRODUCTION

The University of Cyprus and the Central Bank of Cyprus started in March 1997 a special research project titled *"Portfolios of Cyprus Households"* which is fully funded by the Central Bank of Cyprus and is designed to fulfill the scope of a standard Survey of Consumer Finances, namely to collect information on household wealth from a nationally representative sample of Cyprus households.

The first Cyprus project on the Survey of Consumer Finances (CySCF) took place in 1999 while the second in 2002. The Cyprus research team is currently preparing the launching of the 2005 CySCF.

The conclusions of the 1999 CySCF which is offering a picture of family finances for the period 1998-1999 can be found in Haliassos *et al.* (2001, 2003) while the conclusions of the 2002 CySCF which is offering a picture of family finances for the period 2001-2002 can be found in Antoniou *et al.* (2004). The details for the sampling design of the 1999 and 2002 CySCF can be found in Karagrigoriou and Michael (2001) and Karagrigoriou, Michael and Antoniou (2004).

2. GENERAL CHARACTERISTICS AND SAMPLE DESIGN

The sampling design for the 2002 CySCF is based on the 2001 Census records (Census of Population 2001, 2003) and the 5 counties (Nicosia, Limassol, Larnaca, Paphos, and Famagusta[2]) and 58 statistical areas (11 urban and 47 rural) in which Cyprus is divided according to the Statistical Service of the Republic of Cyprus[3]. Each of these areas has been decided to be considered as a primary sampling unit (psu). This decision was based on the requirement that the primary sampling units should be large enough to ensure on one hand the largest possible non-homogeneity and on the other hand the smallest possible variation.

For the first stage of the design, all primary sampling units were combined in 25 *primary areas or strata* which are:

- **5 Large Standard Metropolitan Statistical Areas (Nicosia Municipality, Nicosia Suburbs, Greater Urban Nicosia, Larnaca and Limassol)**
- **4 Medium Standard Metropolitan Statistical Areas (Greater Urban Larnaca, Greater East Urban Limassol, Greater West Urban Limassol and Paphos) and**
- **16 groups of statistical areas.**

Each of the 9 Standard Metropolitan Statistical Areas (SMSA) is an urban area consisting of one (1) primary sampling unit. All but one (Greater SouthEast and North Urban Paphos) of the 16 groups of statistical areas represent rural areas. Four of these groups (Deftera, Paralimni, Achna, and Pyla) consist of one (1) primary sampling unit while the rest have between 2 and 6 primary sampling units. Note that the 5 large SMSAs represent half of the

---

[2] There is a sixth county, Kyrenia which is not currently under the control of the Cyprus Government.

population (340,000 out of 700,000 people) while the average population size of each of the remaining 20 primary areas is 18,000 people.

The 13 primary units corresponding to the 9 SMSAs and the four (4) single-primary-sampling-unit groups of statistical areas are selected in the sample. One primary unit is selected from each of the remaining 12 groups of statistical areas according to the area-probability technique. The area probability design is a multi-stage design that samples successively smaller geographic areas such as counties, regions, and municipalities. The basic rule of the area probability design is that the sample units at each stage of the design are selected with probability proportional to their population.

During the second phase of the selection the actual sampling units (survey households) were selected. In all 25 primary units at least a single-stage sampling was used. Usually, each psu is divided into a number of clusters according to the geographic location and at least 50% of the clusters are selected. In some cases all clusters are selected. The selection of clusters is based on their area probability. The sampling units are randomly selected from the selected clusters. Note that the cluster is either a quarter for each of the large SMSAs or a small municipality, town, or village in all other cases.

---

[3] There are 59 areas but the first one is not presently controlled by the Cyprus Government.

3. THE MAIN AND WEALTHY SAMPLES

One of the main issues associated with the design of the surveys of consumer finances is the heavily skewed income and wealth overall distribution (Avery, Elliehausen, and Kennickell, 1988 and Kennickell and Woodburn, 1992). Specifically, a relatively large share of wealth is held by a relatively small share of households. To resolve the problem most of such surveys are using a dual-frame design where a representative sample known as the main sample, is supplemented by a special sample of high-income households known as the wealthy sample. The main sample which is based on a standard area-probability multi-stage sampling design is included to ensure adequate representation of broadly distributed characteristics while the wealthy one is included to avoid the under-representation of high-income households in the main sample. The area probability frame is based on geographic information and the sample from that frame is drawn using standard multi-stage area-probability sampling techniques. After the collection of data appropriate sample weights can be used to make the data representative of the population as a whole.

For the CySCF, the wealthy sample consists of a sample from the administrative records maintained by the Electricity Authority of Cyprus (EAC) that contain data coded from household electricity consumption. To obtain good measures of highly concentrated assets one should draw a sample of households with either high net worth or some other unusual characteristic. Although extensive data are collected on income as a product of the administration of federal tax returns in Cyprus, the reliability of such data is questionable. As a result it was decided to fill the wealthy sample with households having high consumption of electricity due to the fact that the EAC list is the only available list that is both accurate in terms of measurements and complete in terms of household population coverage.

Note that the high consumption of electricity has been used in various Cyprus surveys as a criterion of wealth. In order to ensure the highest possible accuracy of the results for the CySCF, the EAC files were preferred to any other Authority's files. Note that although the high consumption of electricity is not regarded as the best indicator of wealth, the files of Electricity Authorities around the world are considered as providing the most accurate and reliable household information. It is reminded that in the United States, the Federal Reserve Board is using for this purpose, the highly reliable tax files of the Internal Revenue Service (IRS).

Data for the 2002 CySCF were obtained between the months of March 2002 and June 2003 by the interviewers of the University of Cyprus. Thus, the survey might be thought of as offering a picture of family finances for the period 2001-2002[4]. A total of 897 interviews were completed, 521 for the main sample and 376 for the wealthy. Note that the same questionnaire was used to interview respondents in both the main and the wealthy samples. The field interviewers contacted by phone the selected households to arrange for a meeting for the interview which was conducted in person and averaged about one and a-half hour. Within each survey household, every effort was made for the questionnaire to be completed by the economically dominant or primary economic unit of the household.

---

[4] Income data for 2001 and asset data for 2002 have been collected.

4. THE HOUSEHOLD LIST

The list of households used for the selection of the samples for the CySCF is the 2001 list of customers of the Electricity Authority of Cyprus (EAC) that consists of approximately 295,000 customers/households. Due to the fact that the list consists of all summerhouses and secondary residencies, all residencies consuming on the average at most 100KW bimonthly, were removed from the list. Approximately 12% of the households (33000 households) fall into this category, one-third of which (10000 households) had zero electricity consumption. The final household population (working list) from which the sample is selected is 260,000 households. Note that the number of households (excluding summer houses and multiple residences) according to the 2001 Census is 230,000. It is expected that the extra 10% of households included in the working list are either summer houses or secondary residences which are used so often that the electricity consumption exceeds the 100KW bimonthly limit. Note that no such cases where encountered in the samples selected.

For the determination of the sample sizes of the two samples required, we assume that the *coverage rate* will be around 99% while the *response rate* will be quite low and approximately equal to 55% for the main sample and below 20% for the wealthy (it is assumed to be around 10%).

Based on the above assumptions regarding the response and the coverage rates and in order to ensure the completion of 696 questionnaires for the main sample, a total sample of 1490 households were selected from the 260,000 households. For the wealthy sample and in order to ensure the completion of 515 questionnaires, a sample of 5517 wealthy households was included in the sample. The wealthy sample consists of approximately 10% of the 63,000 households with the largest average-bimonthly electricity consumption among the 260,000 households in EAC's list. The main and wealthy samples by county appear in Table1.

5. RESPONSE RATES

Regarding the number of questionnaires completed for both samples, one should mention that the overall *response rate* is equal to 75% for the main sample and 70% for the wealthy sample. The *response rate* represents the percentage of completed questionnaires out of the total number of *required* questionnaires. Note that such response rates are comparable to the corresponding ones for the United States SCF (Fries et. al, 1998) and are considered usual for such long surveys. The county response rates range from 65% to 100% with the exception of Paphos county. The lowest response rates were reported in the Paphos county with 30% for the main sample and 50% for the wealthy sample. It is worth noting that the highest response rate in both the main and the wealthy samples were reported in the county of Famagusta (100%). It is important to point out that Famagusta was the county with the lowest response rates during the 1999 CySCF. The county of Larnaca comes second with a response rate equal to 90% for the main and 80% for the wealthy sample. Note that the failure to achieve the required target in each county is not due to the main urban areas but rather due to the rural areas within each county with the exception of Paphos where both urban and rural areas failed to achieve the desired target. In most of these areas the corresponding list of households in the sample was exhausted before the target was achieved.

Note that in the urban area of Paphos as well as in a few rural areas with very small target sample sizes the interviewers were allowed to replace the selected household by an adjacent one in an attempt to improve the extremely low response rate of those areas. It should be pointed out that such replacements occurred (and were allowed) in the main sample only. The general characteristics of the two samples by county are presented in Tables 2 and 3.

8

6. GEOGRAPHIC AND INCOME WEIGHTS

After the data entry into the computer, the database was edited and checked for inconsistencies. A number of primarily minor inconsistencies were reported and after verification the proper corrections were made.

The sampling weights are extremely important in all kinds of surveys because they are used to adjust for various variable distributions. Due to the under-representation in some areas (see Tables 2 and 3) the appropriate geographic weights have been calculated and implemented into the data bank. The weighting procedure effectively scales up the weights of responding households to compensate for similar households (of the same geographic area) that did not respond.

Furthermore, the two samples had to be combined to one representative sample. Due to the fact that the wealthy sample is a sample selected from the list of high-income households the key issue to be investigated is the representative ness of the combined sample in terms of the income distribution. For this purpose the survey's data bank (CySCF Bank) was compared to the Statistical Service's data bank for the 1996/7 Family Budget Survey (FBS Bank). The comparison results of the household income distribution between the FBS and the CySCF data banks are reported separately for the main sample of CySCF, the wealthy sample of CySCF, and the combined CySCF sample. From the visual inspection one concludes that the wealthier families are better represented in the combined sample than the less wealthy families (data not shown). The appropriate income weights for the minimization of the influence of extreme wealth cases on the estimation of net worth have been calculated and

incorporated into the CySCF data bank. Note that the FBS data have been adjusted to Jan. 1, 2002 figures using the CPI index provided by the Central Bank of Cyprus in order to be comparable to the CySCF data.

It should be noted that the reliability of FBS Bank has been confirmed before any comparison between the household income data of CySCF and FBS data banks was made. In particular, FSB Bank was compared with the Cyprus Internal Revenue Service's data bank for 1996 (IRS Bank). The comparison results of the (individual) income distribution between FBS and IRS data banks show that the two income distributions appear to be identical (see Karagrigoriou and Michael, 2001).

After the implementation of the geographic and income weights the weighted main sample size has increased to 686 (98.5% of the target of 696) and the weighted wealthy sample size to 511 (99.2% of the target of 515). For a comparison of the 1999 and 2002 CySCF sample sizes by county refer to Table 4.

7. THE IMPUTATION TECHNIQUE

In any survey, there are several potential sources of error, including inadequate survey responses, nonresponses to the entire survey or to particular questions in the survey, and errors due to sampling. The questionnaire for the Cyprus project was carefully designed and found to be in an ideal correspondence with the questionnaire used for the United States SCF. Furthermore, the questionnaire was tested in a short term pilot study in the early months of 2002. Note also that all interviewers went through a careful training and had the opportunity to test the questionnaire several times before they went out to the field. As it is well known, the estimation of potential sampling errors is not straightforward due to the complexity of the design of the SCF. Finally the nonresponse errors, which are usually due to the fact that respondents are not comfortable in revealing certain information or because they do not know the information being asked, have been handled in the final stage of the analysis when an imputation technique was implemented. The steps of the imputation algorithm are described below:

Step 1: All variables to be imputed are identified.

Step 2: All observations of each variable to be imputed are divided into imputation classes according to a limited number (2-4) of significant predictors.

The variables used as significant predictors are: Age, Education, Employment Status and Number of Family Earners.

Step 3: Identify the number of missing and non-missing observations within each imputation class.

Step 4: Calculate the ratio Missing / Non-Missing = X. For each missing observation select at random without replacement X non-missing observations. Replace the missing observation by the median of the selected observations

or

Step 4 (alternative): For each missing observation select at random one non-missing observation. Replace the missing observation by the non-missing one.

Problematic cases are treated separately (see footnotes on imputation methodology table). The variables selected for the imputation are the ones which satisfy both of the following criteria:

> - **The nonresponse rate of the variable is higher than 5% and**
> - **The variable is necessary for the analysis of the data base.**

There are approximately 50 variables which have been imputed. Among these variables are most gross income variables of Section M on "Employment".

The imputed income distribution by sample type and by sample type and income class for 1999 and 2002 CySCF are given in Figures 1 and 2. Indications of the implications of the wealthy sample used in the survey are evident.

12

7. DISCUSSION

The sampling design is not only the most important but also the most delicate part of any statistical analysis. The sampling design of the Cyprus survey of consumer finances although thoroughly checked and prepared could be tested for possible improvements. In fact some aspects need further attention especially since the 2005 CySCF is about to be launched. These aspects are raised and briefly discussed in this section.

The EAC list of households which is used for the selection of the households that constitute the main and wealthy samples although acceptable has found to have significant inaccuracies. The household main information, namely the mailing address and the phone numbers of households is found in several instances to be incomplete and occasionally misleading. As a result a clarification process is often required which in turn, postpones the completion of the questionnaires and consequently puts to unexpected delays the completion of the survey.

It is clear that alternative methods should be considered. For the main sample the alternative suggestions include the telephone company list of households which is expected to be accurate at least in terms of the telephone numbers of the households. The difficulty associated with this particular list lies on the fact that the Cyprus Telecommunications Authority does not keep separate lists of households and businesses. Another difficulty is associated with the fact that such lists cannot be used for the selection of the wealthy sample. Keeping in mind that the accuracy of the files of the Federal Tax Office is questionable it is necessary to continue to depend on the EAC list for the selection of high-income households for the wealthy sample. Alternatives are not available at present.

The experience from the first and second Cyprus project of SCF showed that the households are more hesitant than expected in providing delicate financial information (Karagrigoriou and Michael, 2001). After the completion of the second CySCF, it has been realised that the actual response rates are lower than the estimated ones used for the selection of the list of households for both samples. In particular for the 2002 CySCF, the overall response rate for the main sample usually assumed to be in the range of 50%, was slightly under 40%. The corresponding estimated and actual overall rates for the wealthy sample were 10% and 7%. As a result, a larger number of households are needed to be pre-selected in order to ensure the completion of the required number of questionnaires for both samples. It should be noted that the CySCF does not keep details of the households that refuse to participate in the survey. This is a drawback of the CySCF because the actual number of refusals is not known and as a result the response rates reported above are inaccurate. The rates reported have been calculated assuming that the household lists have been exhausted and the households that did not completed the questionnaire, refused to participate in the survey.

A final note is reserved for the completion of incomplete data. The imputation technique used in the CySCF is a relatively plain technique that could not only be advanced but also extended to the entire set of variables in the CySCF database. Advanced techniques, such as multiple imputations (Rubin, 1987; Kalton, 1983; Little and Rubin, 1987) seem to be necessary to ensure the accuracy of the derived results. It should be though clearly pointed out that the CySCF does not suffer from heavy non-response rates. In fact almost 30% of the variables in 1999 CySCF had no missing values and another 63% had a non-response rate lower than 5% (6.5% had a non-response rate higher than 5%). Even delicate household characteristics such as the total family gross income had an (un-weighted) non-response rate of 12.8% in 1999 CySCF which reduced to (8%) in 2002 CySCF.

REFERENCES

1996/7 Family Budget Survey (1999), *Dept. of Statistics and Research, Ministry of Finance, Printing Office of the Republic of Cyprus.*

Antoniou, A., Argyridou, C., Haliassos, M., Karagrigoriou, A., Kyriacou, G., Michael, M. C., Papagheorgiou, M. and Syrichas, G. (2004). Assets and Debts of Cyprus Households: Changes between the 1999 and 2002 Cyprus Surveys of Consumer Finances, *http://www.ucy.ac.cy/~alex*.

Avery, R. B., Elliehausen, G., and Kennickell, A. B. (1988). Measuring wealth with survey data: An evaluation of the 1983 survey of consumer finances, *Review of Income and Wealth*, pp. 339-369.

Census of Population 2001, Vol. I, General Demographic Characteristics (2003), Statistics Service, *Ministry of Finance, Printing Office of the Republic of Cyprus.*

Fries, G., Starr-McCluer, M, and Sunden, A. E. (1998). The Measurement of Household Wealth using Survey Data: An Overview of the Survey of Consumer Finances, *Presentation at the 44th Conf. of the Amer. Council of Consumer Interests*, Washington D.C.

Guiso, Luigi, Michael Haliassos, and Tullio Jappelli (Eds.), (2001b). *Household Portfolios*, Cambridge, MA: MIT Press.

Haliassos, M. Hassapis, C., Karagrigoriou, A., Kyriacou, G., Michael, M. C., Syrichas, G. (2003). Debts of Cyprus Households: Lessons form the first Cyprus Survey of Consumer Finances, Working Paper 03-03, *Hermes Center of Excellence on Computational Finance and Economics*, University of Cyprus.

Haliassos, M. Hassapis, C., Karagrigoriou, A., Kyriacou, G., Michael, M. C., Syrichas, G. (2001). Assets of Cyprus Households: Lessons form the first Cyprus Survey of

Consumer Finances, Working Paper 01-22, *Hermes Center of Excellence on Computational Finance and Economics*, University of Cyprus.

Kalton, G. (1983). *Compensating for missing survey data*, Inst. for Social Research, The University of Michigan.

Karagrigoriou, A. and Michael, M. C. (2001). The Sample Design of the 1999 Cyprus Project on the Survey of Consumer Finances, *TR\05\2002, Department of Mathematics and Statistics*, University of Cyprus.

Karagrigoriou, A., Michael, M. C. and Antoniou, G. (2004). The Sample Design of the 2002 Cyprus Project on the Survey of Consumer Finances, mimeo, *http://www.ucy.ac.cy/~alex.*

Kennickell, A. and Woodburn, R. L. (1992). Estimation of household net worth using model-based and design-based weights: Evidence from the 1989 Survey of Consumer Finances, *mimeo*.

Kish, L. (1995). *Survey Sampling*. Wiley Classics Library, Wiley, New York.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*, Wiley, New York.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New

Table 1
The Main and Wealthy Sample Sizes

County	Main Sample	Wealthy Sample
	# of households	# of households
Nicosia	555	2311
Limassol	430	902
Larnaca	245	1564
Paphos	165	476
Famagusta	95	264
Total	1490	5517

Table 2
Response Rate – Main Sample

County	Main Sample (Target)	Main Sample (collected)	Response rate
	# of households	# of households	%
Nicosia	275	217	80%
Limassol	201	142	71%
Larnaca	113	103	91%
Paphos	69	21	30%
Famagusta	36	38	100%
Total	696	521	75%

Table 3
Response Rate – Wealthy Sample

County	Wealthy Sample (Target)	Wealthy Sample (collected)	Response rate
	# of households	# of households	%
Nicosia	229	154	68%
Limassol	155	100	65%
Larnaca	89	71	80%
Paphos	47	24	50%
Famagusta	26	27	100%
Total	547	376	69%

Table 4
Final sample sizes with Geographic and Income Weights

County	MAIN SAMPLE			WEALTHY SAMPLE		
	1999	2002		1999	2002	
Nicosia	242	264		309	196	
Paphos	68	68		56	43	
Famagusta	41	37		30	28	
Larnaca	100	108		111	84	
Limassol	203	209		212	160	
Total	654	686		718	511	

FIGURE 1

IMPUTED INCOME DISTRIBUTION BY SAMPLE TYPE - 2002 CySCF
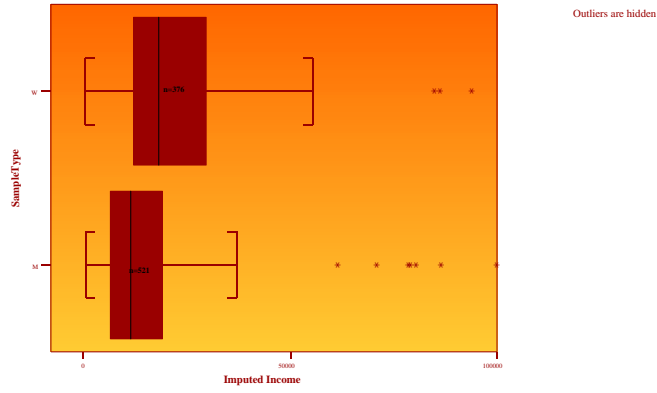
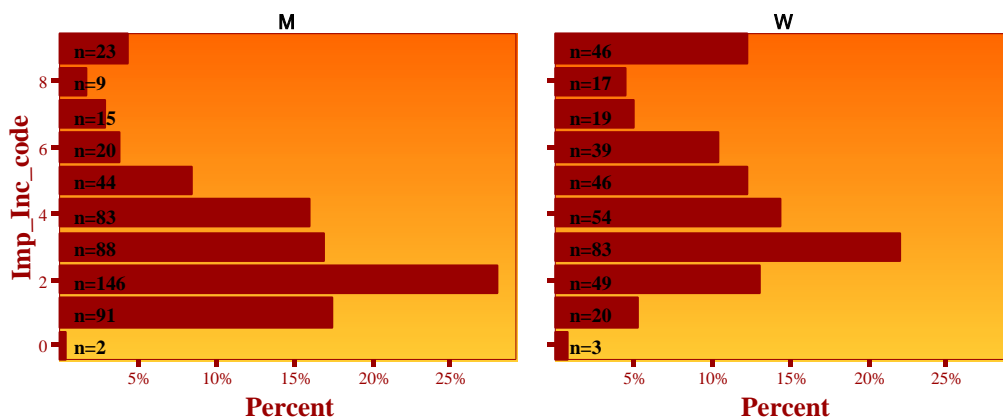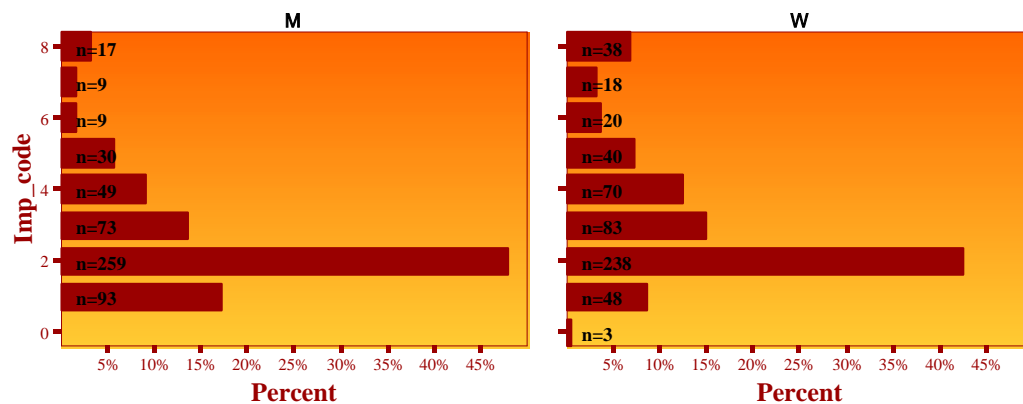IMPUTED INCOME DISTRIBUTION BY SAMPLE TYPE- 1999 CySCF
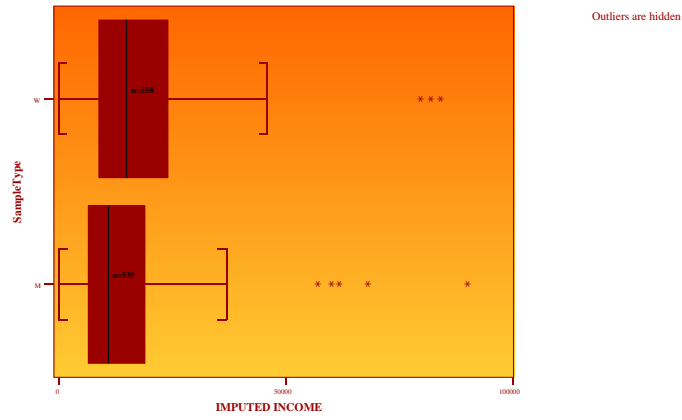
m=main sample

w=wealthy sample

FIGURE 2

INCOME DISTRIBUTION BY SAMPLE TYPE - 2002 CySCF

INCOME DISTRIBUTION BY SAMPLE TYPE - 1999 CySCF

M=Main sample

W=Wealthy sample