Öneri, C.2, S.9.
Ocak, 1998, ss.69-76.

INVESTIGATING HOUSEHOLD INCOME CONCENTRATION AND DISTRIBUTION IN TURKEY

Dr. Harald SCHMIDBAUER

M.Ü. İ.İ.B.F. Almanca İşletme ve Enformatik Bölümleri, Doçent

Abstract

The household income distribution (i.e. how many households had which income) gives important information for social research. We investigate the concentration and distribution of the income earned in 1994 by households in 19 Turkish province centers. The empirical basis is a recent publication of DİE (Devlet İstatistik Enstitüsü). An analysis of the rank correlation between average income and income concentration reveals interesting features ofTurkey.

I-INTRODUCTION

Turkey is among the countries with a high income difference between different social strata (see, for example, ŞEN [11]), and hence a high income concentration, with the consequence that "income distribution has been one of the subjects to be discussed in Turkey" (DİE [2]). A high relative concentration of household income means that a small percentage of households receive a large share of total income. There are many different ways of measuring the extent of inequality of an income distribution (see, for example, FERSCHL [4] or POLASEK [8]). The most widespread seem to be the Lorenz curve and the Gini coefficient. These are also used in the present paper. The empirical basis of the present study is a recent publication of DİE ([2]) concerning the distribution of disposable household incomes in 19 province centers in 1994. Empirical income distributions (in the form of a frequency distribution with grouped data or a histogram) are not given explicitly in [2]. It is possible, however, to derive them using the data in [2]. Our aim here is to investigate properties of household income concentration and distribution and to investigate the relation between several statistical measures (such as Gini coefficient, mean and median income) in order to see different aspects of the income distribution. This also leads to interesting discoveries concerning the 19 province centers. In some cases we give corresponding figures of the household income distribution in Germany. Our aim is not to discuss the notion of disposable income, or to discuss problems of practical statistics. Questions arising from inductive statistics are also beyond the scope of this study.

This paper is organized as follows. The Lorenz curve and Gini coefficient are shortly explained in the next section. We also give properties (which are later used for analyzing the data) which are well-known but not included in most standard textbooks about descriptive statistics or exploratory data analysis. Some aspects of dealing with empirical and theoretical income distributions are discussed in section 3. Results of the empirical study are presented in section 4, which is followed by a section about suggestions for further research. Some conclusions are drawn in the final section.

II- CONCENTRATION — A BRIEF REVIEW

We are concerned here with measuring the inequality in the household income distribution. An appropriate statistical tool for measuring the deviation of the income distribution from an equal distribution (i.e., all the households receive the same income) is the concept of relative concentration. The Lorenz curve is an instrument for visualizing concentration.

Suppose we are given n observed incomes $x_1, \ldots, x_n$, and these are ordered:

$$x_1 \le x_2 \le \ldots \le x_n$$

Then the Lorenz curve $p \mapsto L(p)$, $0 \le p \le 1$, generated by these incomes is the polygon with vertices

$$(0,0) = (p_0,v_0), (p_1,v_1), (p_2,v_2), \ldots, (p_n,v_n) = (1,1)$$

where

$$p_i = i / n$$

(cumulated fractions of households),

$$v_i = \frac{x_1 + \ldots + x_i}{x_1 + \ldots + x_n}$$

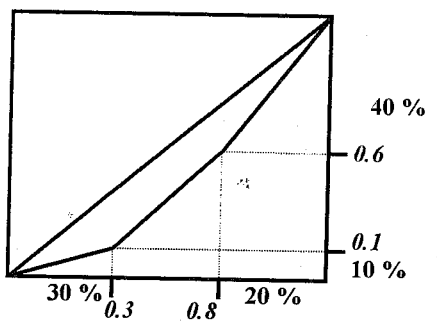


Figure 1: Example of a Lorenz curve

for $i=1,\ldots,n$. Due to the ordering of the x_i, a Lorenz curve is always convex. The Gini coefficient of concentration is the ratio of the area between the 45° line and the Lorenz curve to the area between the 45° line and the abscissa (which is 1/2). If all x_i are equal (equal distribution of income), the Lorenz curve is the 45°line, and the Gini coefficient of concentration becomes 0. The other extreme case is $x_1 = \ldots = x_{n-1} = 0$, $x_n>0$ (one households receives the entire income and all the others nothing). In this case the Gini coefficient is maximal and almost equal to 1 if n is large. An example of a Lorenz curve is given in figure 1. Here, the richest 20% (poorest 30%) of households receive 40% (10%) of the total income. The empirical part of this paper is based on income data concerning 19 province centers in Turkey given in DİE [2]. It is therefore worthwhile to look at the form in which these data are given, in particular because they are not given as grouped income distribution, as is usually the case (see, e.g., RINNE [9]). The data in [2] are, for each province center, organized as shown in table 1.

j	1	2	3	...	18	19	20
p_j	0.05	0.10	0.15	...	0.90	0.95	1.00
s_j	s_1	s_2	s_3	...	s_{18}	s_{19}	s_{20}
s_j/S	s_1/S	s_2/S	s_3/S	...	s_{18}/S	s_{19}/S	s_{20}/S
a_j	a_1	a_2	a_3	...	a_{18}	a_{19}	a_{20}
B_j	b_1	b_2	b_3	...	b_{18}	b_{19}	b_{20}

Table 1: Organization of data in DİE [2]

In table 1, s_j is the sum of all incomes in 1994 in the jth 5%-quantile of households (which are ordered according to income). Thus, s_1 is the sum of all incomes earned by the 5% poorest households and s_{20} is the sum of all incomes earned by the 5% richest households. The s_j/S are shares; $S=s_1+\ldots+s_{20}$. The particular income values of the jth quantile are in the interval $[a_j,b_j]$, that is, a_j (b_j) is the lowest (highest) observed income in the jth group.

The points (p_j,v_j) with $v_j=(s_1+\ldots+s_j)/S$ are the vertices of the Lorenz curve, which is then obtained by linear interpolation. In short, data in the form of table 1 lend themselves very well to drawing a Lorenz curve, but extracting the underlying income frequency distribution requires some effort (see section 3 below).

The Lorenz curve was defined above on the basis of observed income values. It should therefore be more appropriately be called the empirical Lorenz curve. It is also possible to define a Lorenz curve when we assume that the x_i are random variables with distribution function F. The resulting concept of a theoretical Lorenz curve is very useful because properties may be derived which have an immediate interpretation also for the empirical Lorenz curve.

Suppose we are given n random variables with common distribution function F. For large n the total sum of all random variables will then be around

$$n \cdot \int_0^\infty x dF(x) = n\mu,$$

where μ is the expected value of the distribution. (This is the law of large numbers, see FELLER [3].) Let $0 < p < 1$. When the n random variables are arranged in order, the sum of the $p \cdot 100\%$ smallest will be around

$$n \cdot \int_0^{F^{-1}(p)} x dF(x) = n \cdot \int_0^p F^{-1}(x)dx,$$

The Lorenz curve may therefore be defined as

$$p \mapsto L(p) = \mu^{-1} \int_0^p F^{-1}(x)dx \qquad (1)$$

(see GASTWIRTH [5]). When the derivative of this function is equal to 1, we have

$$\frac{d}{dp}L(p) = \mu^{-1}F^{-1}(p) = 1 \text{ or } F(\mu) = p .$$

This leads to

Property 1: Suppose the Lorenz curve has slope 1 in the point p'. Then p' is the fraction of the population receiving less than the average income μ.

It may also be shown that if $L'(p')=1$ then

$$p' - L(p') = \mu^{-1}\int_{\mu}^{\infty}(x-\mu)dF(x) \qquad (2)$$

(see [5]). The difference p'-$L(p')$ is the maximum distance between the 45° line and the Lorenz curve. The right side of (2) is the share of the total income sum which is above the mean μ. Remember that income is equally distributed if all incomes are $=\mu$. Therefore, (2) may be formulated verbally as

Property 2: Suppose the Lorenz curve has slope 1 in the point p'. Then p'-$L(p')$ is the share of the total income which would have to be redistributed in order to obtain an equal distribution of income, i.e. a concentration of zero.

Using the general definition (1), it is also possible to derive error bounds for the Gini coefficient when grouped data are used for its calculation (see[5]).

III- EMPIRICAL AND THEORETICAL INCOME DISTRIBUTIONS

Income data in the form of table 1 contain information about the quantiles of the income distribution in the last two lines. For example, b_{10} is the median income $x_{0.5}$, i.e. 50% of the households are below $x_{0.5}$, and 50% are above. Information about the quantiles can be used in order to gain the necessary input data for a histogram. For example, consider the lowest income class [0,50] (million TL / year 1994). What is the relative frequency of households in this interval? If $50 < b_1$ this frequency is below 5%, and it may be approximated by assuming an equal distribution in $[a_1,b_1]$. Otherwise, we must find j such that $50 \in [a_j,b_j]$, the relative frequency will be greater than $(j-1)\cdot 5\%$, and the value may again be found approximately as described before. In an analogous way the frequencies of all income classes may be estimated.

This method leads to an empirical income distribution, but not to a theoretical model (a probability density). An empirical income distribution will often be only the first step in studying income distribution. A parametric model for the income distribution should be considered. An advantage of using a class of models is that they are derived by postulates concerning the mechanism of income distribution, for example assumptions about the income elasticity. Classical models are the Pareto and the lognormal distribution. On overview is given by DAGUM [1]. The Lorenz curve as a basis for measuring inequality on a parametric basis is also the topic of a recent paper by SARABIA [10]. A different approach is undertaken by HOLM [7] where the income distribution is derived from grouped data on the basis of well-understood principles, namely by maximizing the entropy.

IV- SOME RESULTS FOR 19 PROVINCE CENTERS IN TURKEY

IV.1-Concentration, mean and median income

The (arithmetic) mean and median income as well as the Gini coefficient of household income concentration in 19 province centers are given in table 2.

For further analysis and interpretation, the province centers are also ranked according to these attributes. The city with the highest mean as well as median income is Istanbul, the poorest cities are Diyarbakır and Gaziantep.

The following observations can be made from the data in table 2:

- The median is in all cases lower than the arithmetic mean, which is typical of a left-steep distribution. The ratio between the richest and the poorest place is around 3:1 when the mean income is considered. When we base the comparison on the median income this ratio is only around 2:1. The explanation for this is that the median is not sensitive to very high incomes, in other words, the "tail" of the income distribution is not important for the median.

Place	mean income (million TL)	rank	median income (million TL)	Rank	Gini coefficient	rank
Bursa	162.0	9	111.0	8	0.416	9
Kocaeli	169.1	10	127.1	13	0.411	8
İstanbul	341.3	19	167.6	19	0.576	18
Denizli	157.7	8	112.2	10	0.421	12
İzmir	170.4	11	121.6	11	0.404	7
Adana	220.1	17	103.5	5	0.581	19
Antalya	222.7	18	139.2	16	0.471	16
İçel	143.3	5	106.6	7	0.403	6
Ankara	200.1	15	153.5	18	0.387	5
Eskişehir	126.7	3	95.1	4	0.384	4
Kayseri	172.4	12	94.7	3	0.507	17
Konya	150.3	7	104.3	6	0.429	13
Trabzon	205.8	16	139.6	17	0.441	14
Samsun	177.2	13	124.6	12	0.418	10
Zonguldak	146.4	6	128.8	14	0.318	1
Erzurum	181.9	14	131.0	15	0.442	15
Malatya	138.9	4	111.3	9	0.351	3
Diyarbakır	107.1	2	70.0	1	0.420	11
Gaziantep	102.0	1	81.0	2	0.337	2

Table 2: Income in 19 province centers in 1994.

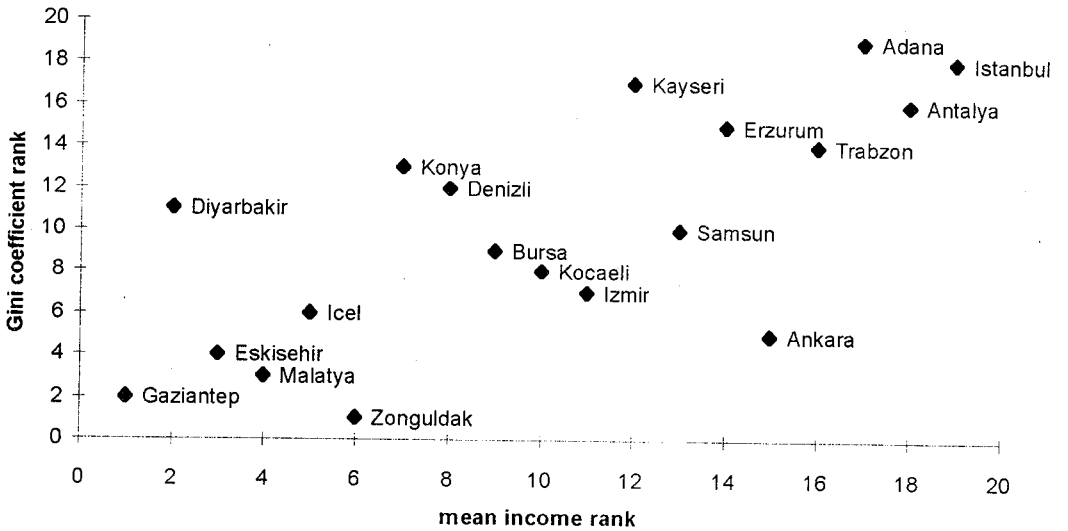


Figure 2: Mean income rank vs. Gini coefficient rank

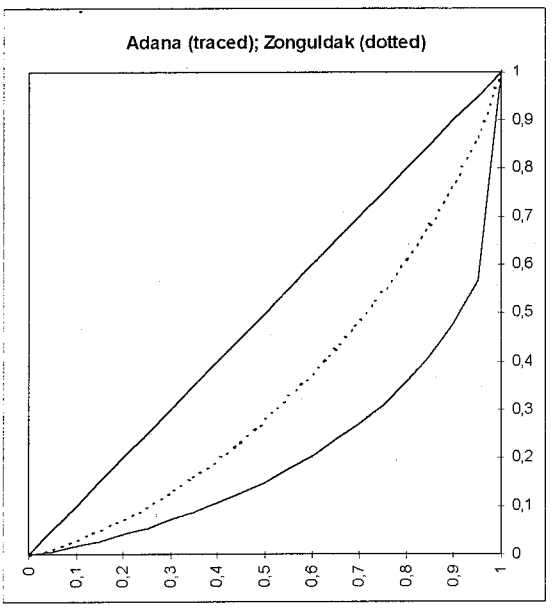
• The rankings according to mean income and according to median income are generally closely associated — but there are three exceptions: For Adana as well as for Kayseri the median income rank is much lower than the mean income rank. In other words: The income level in these cities drops dramatically when we leave away a few relatively high incomes, i.e. when we do not consider the tails of the income distributions explicitly. The general income level is low in Adana as well as in Kayseri; it gets high because of few very high

incomes. This situation is also reflected in a very high income concentration in Adana and Kayseri.

The circumstances are vice versa in the case of Zonguldak whose median income rank is much higher than the mean income rank. Once again this observation is reflected in the concentration — Zonguldak's Gini coefficient is the lowest of all 19 cities.

• There is also a close association between the rankings according to mean income and according to the Gini coefficient. This is displayed in figure 2. Spearman's rank correlation coefficient is +0.71 which indicates a high positive correlation. There are two remarkable exceptions: Ankara and Diyarbakır. Ankara has a high mean income but a low income concentration which suggests that there are very few high-income households but the general income level is high. In other words:

Ankara does not seem to attract extremely rich people on the large scale, or does not provide opportunities for getting extremely rich. On the other hand, Diyarbakır has a very low mean income and a high concentration at the same time. (This is opposed to Gaziantep which also has a low mean income, but with low concentration.) Obviously Diyarbakır is not only a place with low mean income but the income is also very unequally shared by the households. — When we leave out the "exceptions" Ankara and Diyarbakır, the rank correlation coefficient between mean income ranking and Gini coefficient ranking grows to +0.86.



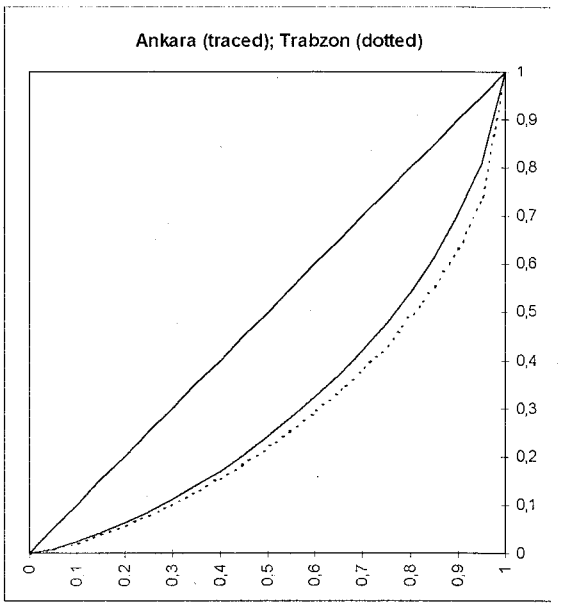


Figure 3: Lorenz curves

• The rank correlation coefficient between *median* income ranking and Gini coefficient ranking is positive but very small: +0.14. This is plausible since, as mentioned before, the median is not sensitive to the tail of the income distribution, but the Gini coefficient is, by notion and definition of concentration.

In a study of the household income and expenditure situation in Germany in 1993 (see HERTEL [6]) the Gini coefficient of concentration of disposable household income was found to be 0.332, the relative difference between mean and median income is 16.9%. The latter is much greater for 18 Turkish cities (up to over 100% for Istanbul and Adana) except for Zonguldak, where it is 13.7%. However, care must be taken when we compare these figures. This will become clear in section 4.4.

VI.2-Lorenz curves

Lorenz curves of concentration for the household income distribution for the extreme cases of Adana (highest concentration among the 19 cities) and Zonguldak (lowest) are shown in figure 3. The Lorenz curve of Istanbul is very similar to that of Adana. The Lorenz curves of the other 17 cities are between the two cases.

Adana's Lorenz curve has slope 1 close to the point (0.8,0.36). This means (see properties 1 and 2

above): The income of ca. 80% of the households is below the Adana average, and 44% of the total income would have to be redistributed in order to achieve an equal distribution of income among the households. (The figures for Istanbul are almost the same.) This shows that the mean income is a rather optimistic figure for a majority of households. The corresponding point in Zonguldak's Lorenz curve is approximately (0.6,0.37), so that only 60% of the households are below Zonguldak's average, and 23% of total income would have to be redistributed.

It is interesting to compare the Lorenz curves of Ankara and Trabzon. The mean income in both cities is nearly the same, but the concentration is very different. Although income concentration in Ankara is relatively low, nearly 70% of the households' income is below average. The corresponding value of Trabzon is slightly higher.

In the case of a *symmetric* income distribution, 50% of the households would be below average.

Income (million TL)	Istanbul		Ankara		Zonguldak		Gaziantep	
	f	$\hat{f}_i$	f	$\hat{f}_i$	f	$\hat{f}_i$	f	$\hat{f}_i$
0 - 50	0.038	0.049	0.046	0.009	0.069	0.006	0.199	0.060
50 - 100	0.192	0.163	0.223	0.168	0.296	0.279	0.436	0.622
100 - 150	0.214	0.171	0.216	0.291	0.246	0.427	0.198	0.269
150 - 200	0.149	0.140	0.182	0.231	0.182	0.201	0.083	0.042
200 - 250	0.099	0.108	0.095	0.140	0.100	0.063	0.037	0.005
250 - 300	0.075	0.081	0.062	0.076	0.045	0.017	0.010	0.002
300 - 350	0.051	0.061	0.048	0.040	0.015	0.005	0.007	0.000
350 - 400	0.034	0.047	0.039	0.021	0.006	0.001	0.006	0.000
400 - 450	0.019	0.036	0.033	0.011	0.005	0.000	0.005	0.000
450 - 500	0.013	0.028	0.009	0.006	0.004	0.000	0.004	0.000
500 - 1000	0.067	0.097	0.017	0.007	0.032	0.000	0.015	0.000
over 1000	0.050	0.020	0.030	0.000	0.000	0.000	0.000	0.000

Table 3: Relative frequencies of income groups, observed and using the lognormal distribution

VI.3- Income distributions

So far we have mainly been concerned with the phenomenon of concentration. The method described in section 3 may be used to find the distribution (relative frequencies of income groups). The income groups are more or less arbitrary, but steps of 50 million TL (income for the whole year 1994) seemed to be a good compromise. The relative frequencies are given in table 3 in the f_i columns. These relative frequencies are displayed in the histograms in figure 4. All distributions are left-steep and have only one peak. With the exception of Istanbul, the income concentration of the distributions is relatively small. This is in accordance with virtually empty high income groups. This is not the case for Istanbul.

Relative frequencies for the income groups were also estimated by assuming that the income distribution is log-normal. These frequencies are given in the $\hat{f}_i$ columns. The log-normal assumption obviously only holds for Istanbul to some extent, but cannot be used for the other cases considered. This is opposed to results concerning the income distribution in Germany (RINNE [9]).

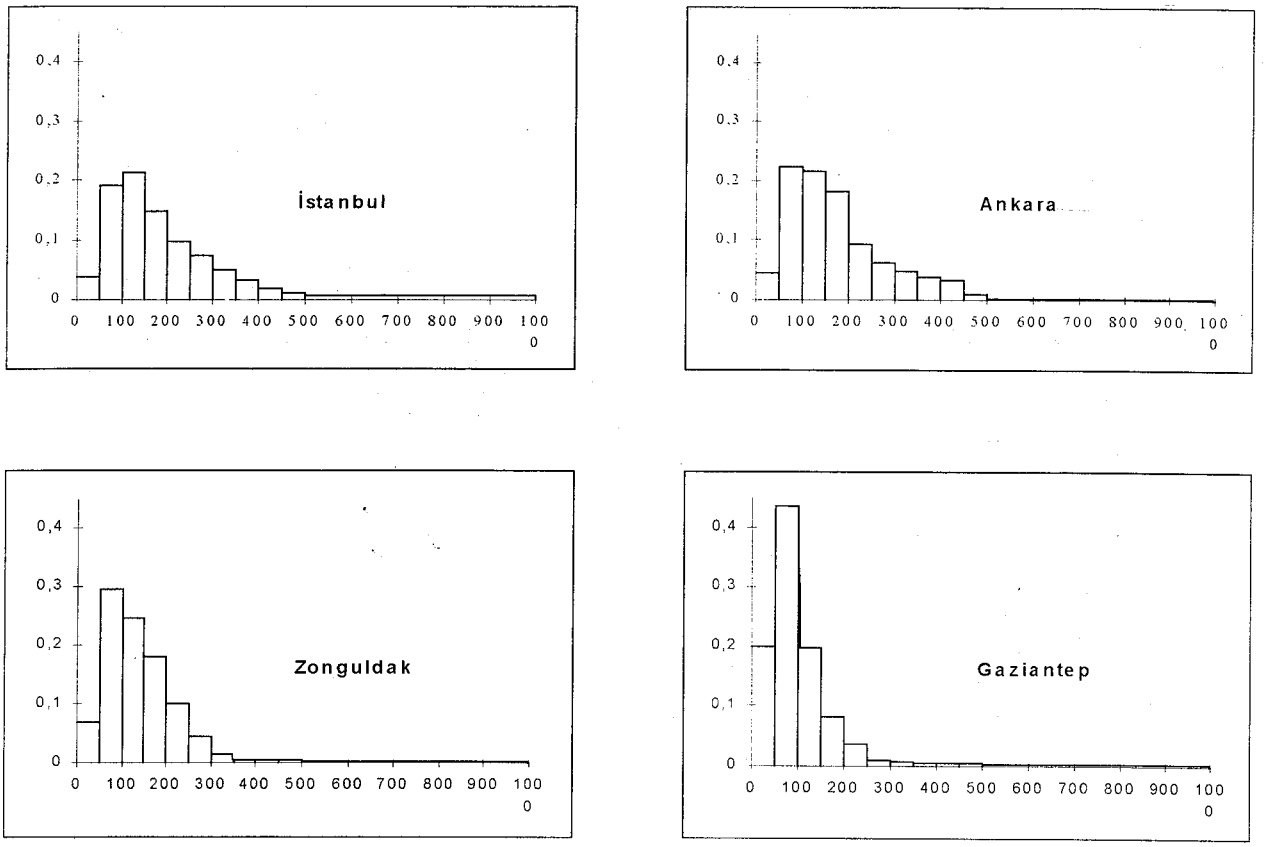


Figure 4: Histograms of income distributions

VI.a-The influence on concentration of the highest incomes

Information about the highest income group is liable to error, mainly because there are only very few cases. It has therefore been suggested to cut off the upper tail of the income distribution and to investigate the concentration of the resulting truncated distribution. The point of truncation was an income of 35000 DM per month in the beforementioned household income study ([6]) in Germany. A similar truncation is not possible for the 19 Turkish cities because income distributions with reliable tail values would be needed. As a surrogate, however, we can compute average incomes and Gini coefficients when the 1% households with the highest income is left out of consideration. Some results are given in table 4. The truncation hardly affects the median income (it will lie in the former 49% quantile) which is therefore not given here.

Obviously this truncation procedure reduces the concentration as well as the mean income, and the higher the concentration the higher the reduction. The conclusion is once again that the mean income (from all households) is a very optimistic figure.

	All households		highest 1% left out		
	mean income (million TL)	Gini coeeficient	Mean income (million TL)	Gini coeeficient	cut-off (million TL)
İstanbul	341.3	0.576	244.9	0.435	2542
Adana	220.1	0.581	158.1	0.444	1694
Ankara	200.1	0.387	188.8	0.359	942
Zonguldak	146.4	0.318	141.1	0.308	570
Diyarbakır	107.1	0.420	99.1	0.384	593

Table 4: Income when the 1% households with highest income are left out

The cut-off values are also given in table 4. These are about 4 (Zonguldak) to 15 (Istanbul and Adana) times the median income. The corresponding value in Germany ([6]; 35000 DM) is about 8.5 times the median. This shows that we should rather compare the Gini coefficients *after* truncation with Germany's value (0.332).

V- SUGGESTIONS FOR FURTHER RESEARCH

We have seen that care should be taken when Gini coefficients of household income concentration from different epochs or countries are compared. The truncation of the underlying income distribution can influence the amount of concentration substantially. However, in order to be in a position to study the truncation influence, the distribution itself should be given, whenever possible. Information about the Lorenz curve should be complemented by information about the underlying income distribution. Knowledge of income distributions before and after the deduction of taxes is also a prerequisite for the discussion of the impact of the tax system on the income distribution. (In DİE [2] the income after the deduction of direct taxes is given.)

Structural properties of the households and their members (e.g., expressed through demographic variables such as age and sex, or social status) will also strongly influence the income distribution. A way of finding an overall measure which is comparable internationally might be gained by using a standardized population for the countries or epochs which are to be compared. Attention was focused here on household income. Results concerning per capita income distributions will be different because household sizes are different in different regions.

Finally, very much research remains to be done in the field of income distributions itself. What creates the income distributions whose effects can be observed? Parametric models might give insight into the mechanism that governs the income distribution in the regions of Turkey.

VI- CONCLUSIONS

It was shown that there is a high positive correlation between mean income rank and Gini coefficient rank of the 19 province centers in Turkey. There are two remarkable exceptions: Ankara, with a high mean income and a low concentration, and Diyarbakır, with a low mean income and a high concentration.

The mean income was found to show the income conditions of a majority of households in a too favourable light. This is due to the high income concentration and to the extreme left-steepness of the income distributions.

For example, about 80% of Istanbul's households receive an income below the average.

The very high income concentration is substantially reduced when the highest incomes are cut off. At the same time the mean income drops substantially, which also shows that the mean income is a rather optimistic figure for many households. The cut off-procedure may also lead to figures which are suitable for international comparisons. This was shown on the example of Germany.

The present study is almost entirely in the framework of descriptive statistics. To investigate the fit of parametric models or base statistical inference on a suitable basis is a practical task which remains to be done.

REFERENCES

[1] DAGUM, CAMILO: Income distribution models. In *Encyclopedia of Statistical Sciences* (Eds.: sc KOTZ/JOHNSON), Vol. 4. Wiley, 1983.

[2] DİE (DEVLET İSTATISTIK ENSTÜTÜSÜ): *1994 Hanehalkı Gelir Dağılımı Anketi Sonuçları*, T.C. Başbakanlık Devlet İstatistik Enstitüsü, 1997.

[3] FELLER, WILLIAM: it *An Introduction to Probability Theory and Its Applications*, Vol. II. Wiley, New York, 1971.

[4] FERSCHL, FRANZ: *Deskriptive Statistik*, 3. Auflage. Physica, 1985.

[5] GASTWIRTH, JOSEPH L.: The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics* 54 (1972), 306 - 316.

[6] HERTEL, JÜRGEN: Einnahmen und Ausgaben der privaten Haushalte 1993. Ergebnis der Einkommens- und Verbrauchsstichprobe. *Wirtschaft und Statistik* (1997), 45 - 58.

[7] HOLM, JUHANI: Maximum entropy Lorenz curves. *Journal of Econometrics* 59 (1997), 377 - 389.

[8] POLASEK, WOLFGANG: *EDA — Explorative Datenanalyse. Einführung in die deskriptive Statistik*, 2. Auf/lage. Springer, 1994.

[9] RINNE, HORST: *Wirtschafts- und Bevölkerungsstatistik*, 2. Auflage. Oldenbourg, 1996.

[10] SARABIA, JOS'E-MAR'IA: A hierarchy of Lorenz curves based on the generalized Tukey's lambda distribution. *Econometric Reviews/* 16 (1997), 305 - 320.

[11] ŞEN, FARUK: *Türkei*, 4. Auflage. Beck, 1996.