THE PERSONAL DISTRIBUTION OF INCOME IN THE UNITED STATES

BY
FREDERICK ROBERTSON MACAULAY
NATIONAL BUREAU OF ECONOMIC RESEARCH, INC.

THE PERSONAL DISTRIBUTION OF INCOME IN THE UNITED STATES

BY
FREDERICK ROBERTSON MACAULAY
NATIONAL BUREAU OF ECONOMIC RESEARCH, INC.

Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in the Faculty of Political Science, Columbia University

NEW YORK
HARCOURT, BRACE AND COMPANY
1922

Printed in the U. S. A.

PREFACE

In the year 1922 the National Bureau of Economic Research, Inc., published in two volumes the result of an investigation into "Income in the United States." Part III of Volume II of that work consisted of the present study. The author acknowledges with thanks the courtesy of the National Bureau of Economic Research in permitting him to have this reprint made from the original plates.[1]

[1] This fact explains the pagination.

TABLE OF CONTENTS

CHAPTER 27

THE PROBLEM

What is the frequency distribution of annual income among personal income recipients in the United States? Before we can give an intelligent answer to this question, we must formulate it more definitely by indicating certain connotations which logic or expediency leads us to attach to some of its terms.

By *income* it seems desirable to mean actual money income, plus the estimated money value of the more important of those items of commodity or service income on which a money value is ordinarily placed. Two of the most important items which are thus included are the annual rental values of owned homes and the value of farm produce consumed by farmers' families.

In line with the ordinary convention, we have excluded from our definition of *income*, that income, whether monetary or non-monetary, which a wife receives from her husband or a child from its parents.[1] Not only is such exclusion practically expedient but it is also theoretically defensible and that quite apart from the fact that a money value is not ordinarily placed on the services of wife or child, wages of housekeepers to the contrary notwithstanding.

The frequency distribution resulting from the exclusion of such quasi incomes will be less heterogeneous and more significant and interpretable than the distribution which would result from inclusion. For the relation of the incomes of wives and children to the economic struggle is derived and secondary, while that of most other incomes is direct and primary. Now, though the distribution of income among persons is not synonymous with distribution among the factors of production, the two problems are very closely related. An individual's income may be thought of as made up of wages, rent, interest and dividends, profits, and gifts or allowances. If we omit this last type of income, the problem of factorial distribution proposes an investigation of how and why the individual received what remains. Even if gifts and allowances admitted of any such systematic and reasoned explanation as may be given of rent, wages, etc., the explanation would be of a totally different kind. Hence, for the purposes of this investigation, it seems undesirable to classify as *income*, the receipts,

[1] That is, while such income has, of course, been counted in the first instance as income of the husband or parent it has not been re-counted as income of the wife or child.

whether monetary or non-monetary, of those persons receiving merely allowances or gifts.[1]

Similar considerations have led us to think of an *income recipient* as an individual rather than a family. Just as it is the husband and not the wife, the parent and not the child, so it is the individual and not the family who, as an income receiver, comes into direct economic relationship with the machinery of distribution.

The chief argument in favor of family rather than individual treatment of the frequency distribution is based upon the idea that, though income accrues to the individual and not the family, the family is a more significant unit of economic need than the individual. But this is a different approach to the question and has, of course, no intimate relation to the problem of factorial distribution. Moreover, we must remember that if we are going to improve appreciably upon the individual, even as a need unit, we cannot stop with actual biological families with their great variation in size and constitution, but must introduce the concept of the theoretical family—father, mother and three children, for example. This last concept is, in its raw form, quite unusable. The population is not made up of such theoretical families. We may discuss what a family of five *ought* to get to maintain a decent standard of living, but we cannot divide the actual population into families of five and discuss what these non-existent hypothetical families actually do get. There remains the alternative of expressing actual families in terms of some *need* unit such as the "ammain." [2] While this last procedure would probably yield an extremely interesting distribution based upon *need* units, it is impractical to attempt any such solution with the data available.[3]

Though a distribution of income among actual biological families would appear to be somewhat less enlightening and interpretable than a distribution by individuals or by ammains, it would have its own peculiar interest and we would have attempted its construction had the data been adequate for such a purpose. Most of the data bearing on income distribution are in the individual form; wages distributions, for example, are

[1] Of course if the wife or child has "independent" income, that income is no longer of the nature of a gift or allowance even though it may arise from property originally deeded by the husband or father. It is now explainable in terms of rent, interest, etc.

If *income* be defined as above, the term *personal income recipient* will correspond closely to the census expression *person gainfully employed.* Perhaps the most important difference is that we do not and the Census does include as separate income recipients, farm laborers working on the home farm.

[2] *Ammain* is a word coined by W. I. King and E. Sydenstricker and defined by them, for any given class of people, as "a gross demand for articles of consumption having a total money value equal to that demanded by the average male in that class at the age when his total requirements for expense of maintenance reach a maximum." *Measurement of Relative Economic Status of Families.* Quarterly Publications of the American Statistical Association, Sept., 1921, p. 852.

[3] It is of course quite possible to estimate the *average* per ammain income, as has been done by Mr. King; the total income of the people can be divided by the estimated number of ammains in the population. See pages 233 and 234.

almost without exception in that form. Now to estimate the frequency distribution of income among families from data which, in the first place, are in the individual form and, in the second place, are extremely inadequate for estimating even the distribution among individuals, could only increase the degree of uncertainty in our results.

A few words explaining the reason for introducing the next chapter at this point are not out of place here. The data upon which an estimate of even the *individual* distribution of income in the United States must be based impress one as being in such shape that it is impossible to arrive at more than the roughest sort of approximation by any mere direct adding process. Some more ingenious plan would seem almost necessary. For example, would it not be possible to formulate a general mathematical "law" for the distribution of incomes which law might then be used for "adjusting" the tentative and hypothetical results obtained from piecing together the existing scanty and inadequate material?

The possibility and desirability of mathematically describing the frequency distribution of income would seem intimately tied up with the case for mathematically describing error distributions and statistical distributions in general. The fact that, in our problem, the "law" would be largely derived from the same data as those which were to be "adjusted" need not greatly disturb us. The procedure of adjusting observations in the light of a mathematical expression derived from the same observations is not novel. A number of attempts, one of which has become world-famous, have been made to demonstrate that the distribution of income follows a definite mathematical law. However, the next chapter will show why we fear that no rational and useful mathematical law will soon be formulated.

CHAPTER 28

PARETO'S LAW AND THE GENERAL PROBLEM OF MATHEMATICALLY DESCRIBING THE FREQUENCY DISTRIBUTION OF INCOME

The problem of formulating a mathematical expression which shall describe the frequency distribution of income in all places and at all times, not only closely, but also elegantly, and if possible rationally as opposed to empirically, has had great attractions for the mathematical economist and statistician. The most famous of all attempts at the solution of this fascinating problem are those which have been made by Vilfredo Pareto. Professor Pareto has been intensely interested in this subject for many years and the discussion of it runs through nearly all of his published work. The almost inevitable result is that "Pareto's Law" appears in a number of slightly different forms and Professor Pareto's feelings concerning the "law" run all the way from treating it as inevitable and immutable to speaking of it as "merely empirical."

In its best known, most famous, and most dogmatic form, Pareto's Law runs about as follows:

1. In all countries and at all times the distribution of income is such that the upper (income-tax) ranges of the income frequency distribution curve may be described as follows: If the logarithms of income sizes be charted on a horizontal scale and the logarithms of the numbers of persons having an income of a particular size or over be charted on a vertical scale, then the resulting observational points will lie approximately along a straight line. In other words, if

x = income size and
y = number of persons having that income or larger
then $\log y = \log b + m \log x$
or $y = bx^m$.[1]

2. In all countries and at all recent times the *slope* of this straight line fitted to the cumulative distribution, that is, the constant m in the equation $y = bx^m$, will be approximately 1.5.[2]

3. The rigidity and universality of the two preceding conclusions strongly

[1] If the cumulative distribution (cumulating from the higher towards the lower incomes as Pareto does) on a double log scale could be exactly described by the equation $y = bx^m$, the non-cumulative distribution could be described by the equation $Y = -mbx^{m-1}$.

[2] Strictly, *minus* 1.5, though Pareto neglects the sign.

suggest that the shape of the income frequency distribution curve on a double log scale is, for all countries and at all times, inevitably the same not only in the upper (income-tax) range but throughout its entire length.

4. If then the nature of the whole income frequency distribution is unchanging and unchangeable there is, of course, no possibility of economic welfare being increased through any change in the proportion of the total income going to the relatively poor. Economic welfare can be increased only through increased production. In other words, Pareto's Law in this extreme form constitutes a modern substitute for the Wages Fund Doctrine.

This is the most dogmatic form in which the "law" appears. In his later work Professor Pareto drew further and further away from the confidence of his first position. He had early stated that the straight line did not seem adequate to describe distributions from all times and places and had proposed more complicated equations.[1] He has held more strongly to the significance of the similarity of slopes but he has wavered in his faith that the lower income portions of the curve (below the income-tax minimum) were necessarily similar for all countries and all times. He has given up the suggestion that existing distributions are inevitable though still speaking of the law as true within certain definite ranges. To translate from his *Manuel* (p. 391): "Some persons would deduce from it a general law as to the only way in which the inequality of incomes can be diminished. But such a conclusion far transcends anything that can be derived from the premises. Empirical laws, like those with which we are here concerned, have little or no value outside the limits for which they were found experimentally to be true." Indeed Professor Pareto has himself drawn attention to so many difficulties inherent in the crude dogmatic form of the law that this chapter must not be taken as primarily a criticism of his work but rather as a note on the general problem of mathematically describing the frequency distribution of incomes.

Almost as soon as he had formulated his law Professor Pareto recognized the impossibility of extrapolating the straight line formula into the lower income ranges (outside of the income-tax data which he had been using). The straight line formula involves the absurdity of an infinite number of individuals having approximately zero incomes. Professor Pareto felt that this zero mode with an infinite ordinate was absurd. He believed that the curve must have a definite mode at an income size well above zero [2] and with a finite number of income recipients in the modal group.

[1] The inadequacy of these more complicated equations is discussed later. See pp. 348, 363 and 364.

[2] This is, of course, not absolutely necessary. It depends upon our definitions of *income* and *income recipient*. If we include the negligible money receipts of young children living at home we might possibly have a mode close to zero. There are few children who do not really earn a few pennies each year. Compare Chart 31A page 416.

Having come to the conclusion that the income frequency distribution curve must inevitably have a definite mode well above zero income and tail off in both directions from that mode, Professor Pareto was led to think of the possibilities of the simplest of all frequency curves, the normal curve of error. However, after examination and consideration, he felt strongly that the normal curve of error could not possibly be used. He became convinced that the normal curve was not the law of the data for the good and sufficient reason that the part of the data curve given by income-tax returns is of a radically different shape from any part of a normal curve.[1]

Professor Pareto finds a further argument against using the normal curve in the irrationality of such a curve outside the range of the data. The mode of the complete frequency curve for income distribution is at least as low as the minimum taxable income. Income-tax data prove this. However, a normal curve is symmetrical. Hence, if a normal curve could describe the upper ranges of the income curve as given by income-tax data then in the lower ranges it would cut the y axis and pass into the second quadrant, in other words show a large number of negative incomes.

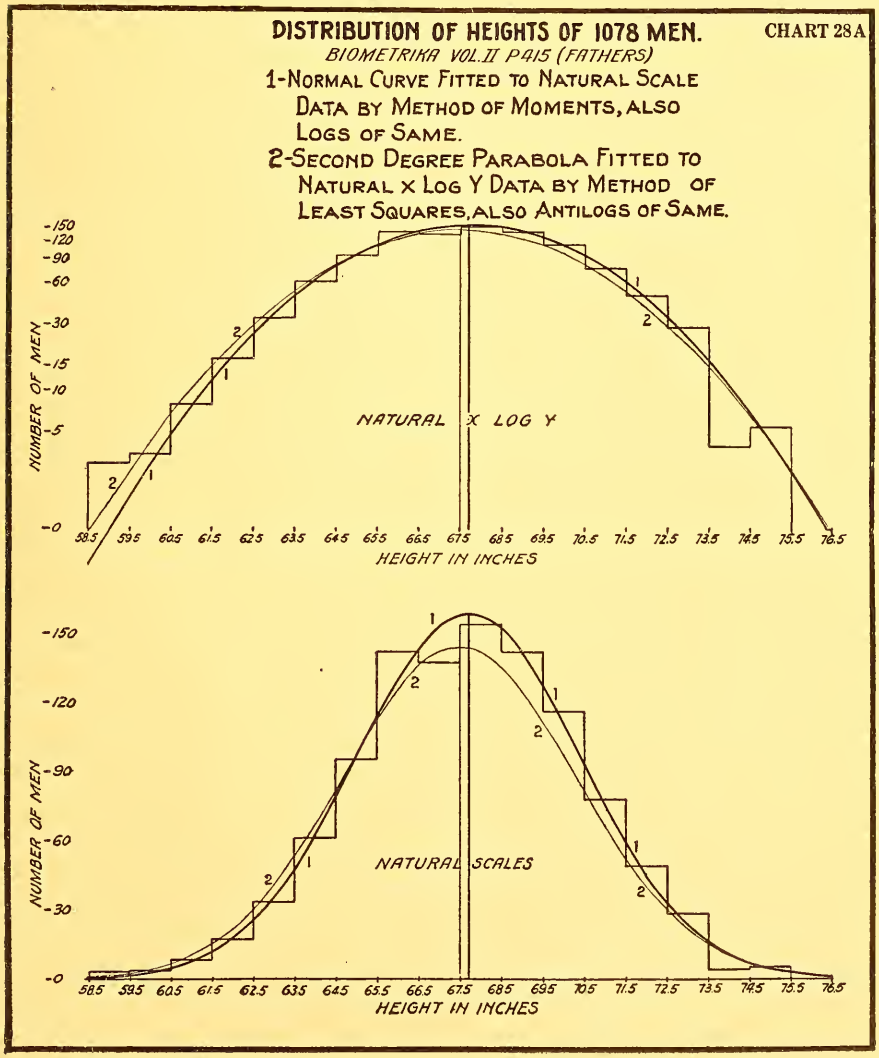
Now, aside from the fact that this whole argument is unnecessary if the data themselves cannot be described even approximately by a normal curve, Professor Pareto's discussion reveals a curious change in his middle term. If he had said that a symmetrical curve on a natural scale with a mode at least as low as the income-tax minimum would show *unbelievably large* negative incomes we could follow him but when he states that not only can there be no zero incomes but that there can be no incomes below "the minimum of existence" we realize that he has unconsciously changed the meaning of his middle term. Having examined a mass of income-tax data, all of which were concerned with *net money* income and from these data having formulated a law, he now apparently without realizing it, changes the meaning of the word income from *net money income* to *money value of commodities consumed*, and assumes that those who receive a *money* income less than a certain minimum must inevitably die of starvation.

[1] Though Pareto seems to have thoroughly understood this fact, his discussion is not altogether satisfactory. He states that the data for the higher incomes show a larger number of such incomes than the normal curve would indicate. This is hardly adequate. To have stated that the upper and lower ranges showed too many incomes *as compared with the middle range* would have been better. An easy way to realize clearly the impossibility of describing income-tax data by a normal curve is to plot a portion of the non-cumulative data on a *natural x log y* basis. When so charted the data present a concave shaped curve. However, if the data were describable by any part of a normal curve of error, they would show a convex appearance, or in the limiting case a straight line, as the equation of the normal curve of error $\left(y_x = y_o e^{\frac{-x^2}{2\sigma^2}}\right)$ becomes, on a natural x log y scale, $\log_e y_x = \log_e y_o - \frac{x^2}{2\sigma^2}$ or a second degree parabola whose axis is perpendicular to the x axis of coördinates.

The reader must note that the limiting straight line case mentioned above is on a *natural x* log *y* scale and not (as the Pareto straight line) on a *log x* log *y* scale. (Note concluded page 347.)

Children receive in general negligible *money* incomes. Many other persons in the community are in the same position. A business man may "lose money" in a given year, in other words he may have a negative money income. There seems no essential absurdity in assuming that a large number of persons receive money incomes much less than necessary to

(Note 1 page 346 concluded.)
Chart 28A showing curves fitted to observations on the heights of men illustrates the appearance of the normal curve on a natural scale and on a natural x log y scale. That chart also illustrates another fact of importance in this discussion, namely, that fitting to a different function of the variable gives a different fit.



support existence. When in 1915 Australia took a census of the incomes of all persons "possessed of property, or in receipt of income," over 14 per cent of the returns showed incomes "deficit and nil." [1]

Professor Pareto's realization of the impossibility of describing income distributions by means of normal curves led him to the curious conclusion that such distributions were somehow unique and could not be explained upon any "chance" hypothesis. "The shape of the curve which is furnished us by statistics, does not correspond at all to the curve of errors, *that is to say* [2] to the form which the curve would have if the acquisition and conservation of wealth depended only on chance." [3] Moreover, while Professor Pareto's further suggestion of possible heterogeneity in the data corresponds we believe to the facts, his reason for making such a suggestion, namely that the data cannot be adequately described by a normal curve, is irrelevant.[4] "Chance" data distributions are no longer thought of as necessarily in any way similar to the normal curve. Even error distributions commonly depart widely from the normal curve. The best known system of mathematical frequency curves, that of Karl Pearson, is intended to describe homogeneous material and is based upon a probability foundation, yet the normal curve is only one of the many and diverse forms yielded by his fundamental equation $\frac{d \log y}{dx} = \frac{x + a}{b_o + b_1 x + b_2 x^2}$.[5]

While Pareto's Law in its straight line form was at least an interesting suggestion, his efforts to amend the law have not been fruitful. His attempts to substitute $\log_e N = \log_e A - \alpha \log_e (x + a)$ or even $\log_e N = \log_e A - \alpha \log_e (x + a) - \beta x$ for the simpler $\log N = \log A - \alpha \log x$ have not materially advanced the subject.[6] The more complicated curves have the same fundamental drawbacks as the simpler one. Among other peculiarities they involve the same absurdity of an infinite number of persons in the modal interval and none below the mode. Along with the doubling of the number of constants, there comes of course the possibility of improving the fit within the range of the data. Such improvement is, however, purely artificial and empirical and without special significance, as can be easily appreciated by noticing the mathematical characteristics of the equation.

A number of other statisticians have at various times fitted different types of frequency curves to distributions of income, wages, rents, wealth,

[1] Compare Table 29A.
[2] My italics.
[3] *Manuel*, p. 385. See also *Cours*, pp. 416 and 417.
[4] Vid. *Cours*, pp. 416 and 417.
[5] Professor A. W. Flux in a review of Pareto's *Cours d'Economie Politique* (*Economic Journal*, March, 1897) drew attention to the inadequacy of Pareto's conception of what were and what were not "chance" data.
[6] Cf. *Cours*, vol. II, p. 305, note.

or allied data.[1] However, no one has advanced such claims for a "law" of *income*[2] distribution as were at one time made by Professor Pareto. When considering the possibility of helpfully describing the distribution of income by any simple mathematical expression, one inevitably begins by examining "Pareto's Law." It is so outstanding. Let us therefore examine Pareto's Law.

> 1. Do income distributions, when plotted on a double log scale, approximate straight lines closely enough to give such approximation much significance?

Before attempting to answer this question it is of course necessary to decide how we shall obtain the *straight line* with which comparisons are to be made.

Professor Pareto fitted straight lines directly by the method of least squares to the *cumulative* distribution plotted on a double log scale. The disadvantage of this procedure is that, though one may obtain the straight line which best fits the *cumulative* distribution, such a straight line may be anything but an admirable fit to the *non-cumulative* figures. For example, if a straight line be fitted by the method of least squares to Prussian returns for 1886 (as given by Professor Pareto) the total number of income recipients within the range of the data is, according to the fitted straight line, only 5,399,000 while the actual number of returns was 5,557,000, notwithstanding the fact that Prussia, 1886, is a sample which runs much more nearly straight than is usual. How bad the discrepancy may be where the data do not even approximate a straight line is seen in Professor Pareto's Oldenburg material. There the least-squares straight line fitted to the cumulative distribution on a double log scale gives 91,222 persons having incomes over 300 marks per annum while the data give only 54,309.

[1] Among others, Karl Pearson, F. Y. Edgeworth, Henry L. Moore, A. L. Bowley, Lucien March, J. C. Kapteyn, C. Bresciani, C. Gini, F. Savorgnan.

[2] Professor H. L. Moore, in his *Laws of Wages*, is concerned primarily with *wages* not *income*.

Professor J. C. Kapteyn has presented a pretty but somewhat hypothetical argument suggesting that the skewness in the income frequency curve should be such that plotting on a log x basis would eliminate it.

"In several cases we feel at once that the effect of the causes of deviation cannot be independent of the dimension of the quantities observed. In such cases we may conclude at once that the frequency curve will be a skew one. To take a single example:

"Suppose 1000 men to begin trading, each with the same capital; in order to see how their wealth will be distributed after the lapse of 10 years, consider first what will be their condition at some earlier epoch, say at the end of the fifth year.

"We may admit that a certain trader A will then only possess a capital of £100, while another may possess £100,000.

"Now if a certain cause of gain or loss comes to operate, what will happen?

"For instance: Let the price of an article in which both A and B have invested their capital, rise or fall. Then it will be evident that if the gain or loss of A be £10, that of B will not be £10, but £10,000; that is to say, the effect of this cause will not be independent of the capital, but proportional to it."

J. C. Kapteyn, *Skew Frequency Curves in Biology and Statistics*, p. 13.

The reason for this peculiarity of the fit to the cumulative distribution becomes clear when we remember that the least-squares straight line may easily deviate widely from the first datum point while a straight line giving the same number of income recipients as the data must necessarily pass *through* the first datum point.[1]

A straight line fitted in such a manner that the total number of persons and total amount of income correspond to the data for these items gives what seems a much more intelligible fit. Charts 28B to 28G show cumulative United States frequency distributions from the income-tax returns for the years 1914 to 1919 on a double log scale (Professor Pareto's suggestion). Two straight lines are fitted to each distribution—one a solid least-squares line fitted to the cumulative data points and the other a dotted line so fitted that the total number of persons and total amount of income correspond to the data figures. While the least-squares line may appear much the better fit to these cumulative data, a mere glance at Tables 28B to 28G will reveal the fact that such a line is, to say the least, a less interpretable fit to the non-cumulative distribution.[2] It is, of course, evident that neither line is in any year a sufficiently good fit to the actual non-cumulative distribution to have much significance. No mathematics is necessary to demonstrate this.[3]

[1] e. g. in the case of Prussia, 1886, the first datum point is $x =$ "over 300M" and $y = 54{,}309$ persons.

[2] Professor Warren M. Persons discussed the fit of the *least-squares* straight line to Professor Pareto's Prussian data for 1892 and 1902 in the *Quarterly Journal of Economics*, May, 1909, and demonstrated the badness of fit of that line to those data.

[3] The income returned for the years 1914 and 1915 was estimated from the *number* of returns. *Income* is not given in the reports for those years.

In fitting straight lines to the data of Tables 28B to 28G the lowest income interval (in which married persons making a joint return are exempt) has always been omitted. To have included in our calculations these lowest intervals would have increased still further the badness of the fit in the other intervals.

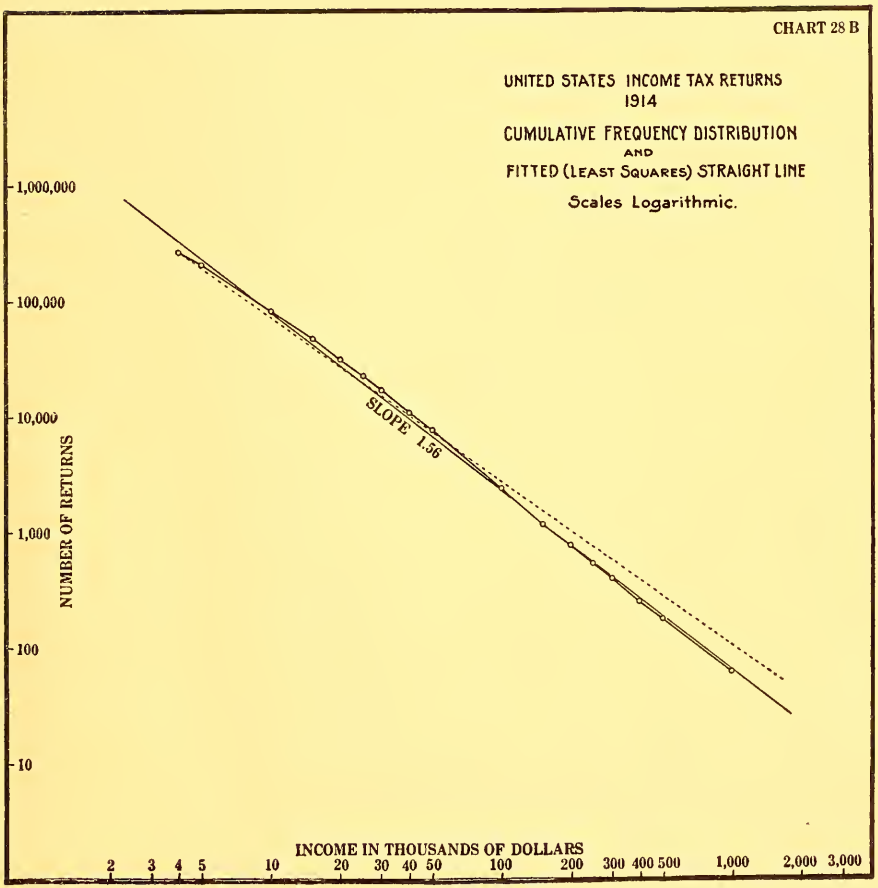


CHART 28 B
UNITED STATES INCOME TAX RETURNS
1914
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic.
1,000,000
100,000
10,000
1,000
100
10
NUMBER OF RETURNS
SLOPE 1.56
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000

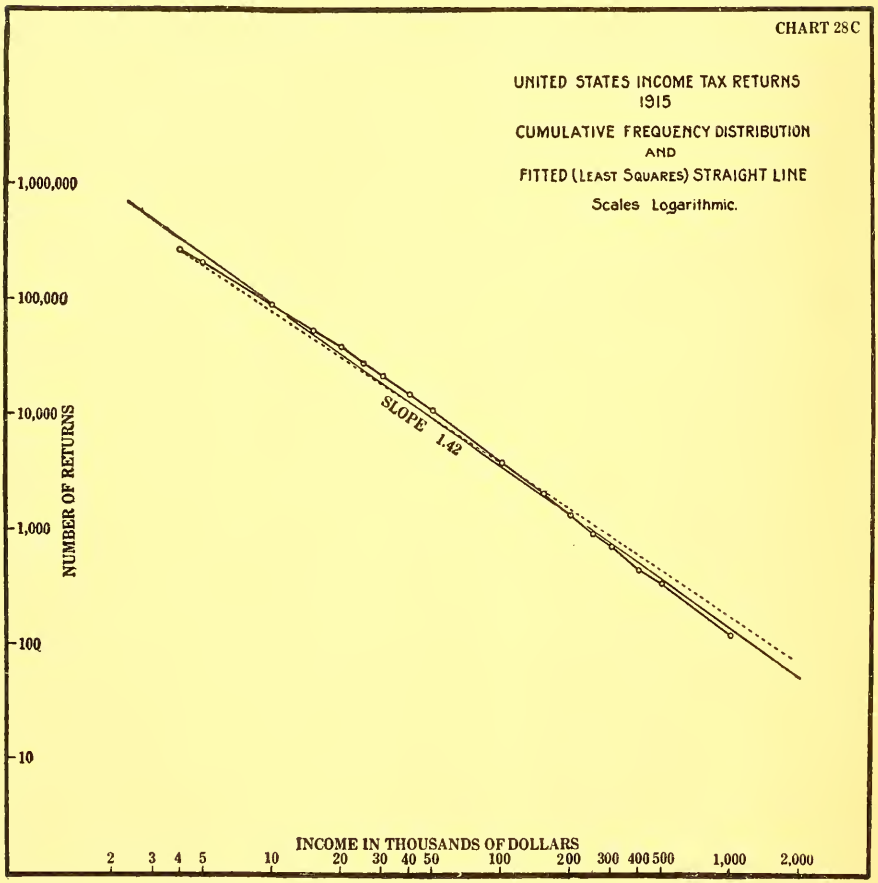


CHART 28C
UNITED STATES INCOME TAX RETURNS
1915
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic.
1,000,000
100,000
10,000
1,000
100
10
NUMBER OF RETURNS
SLOPE 1.42
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000

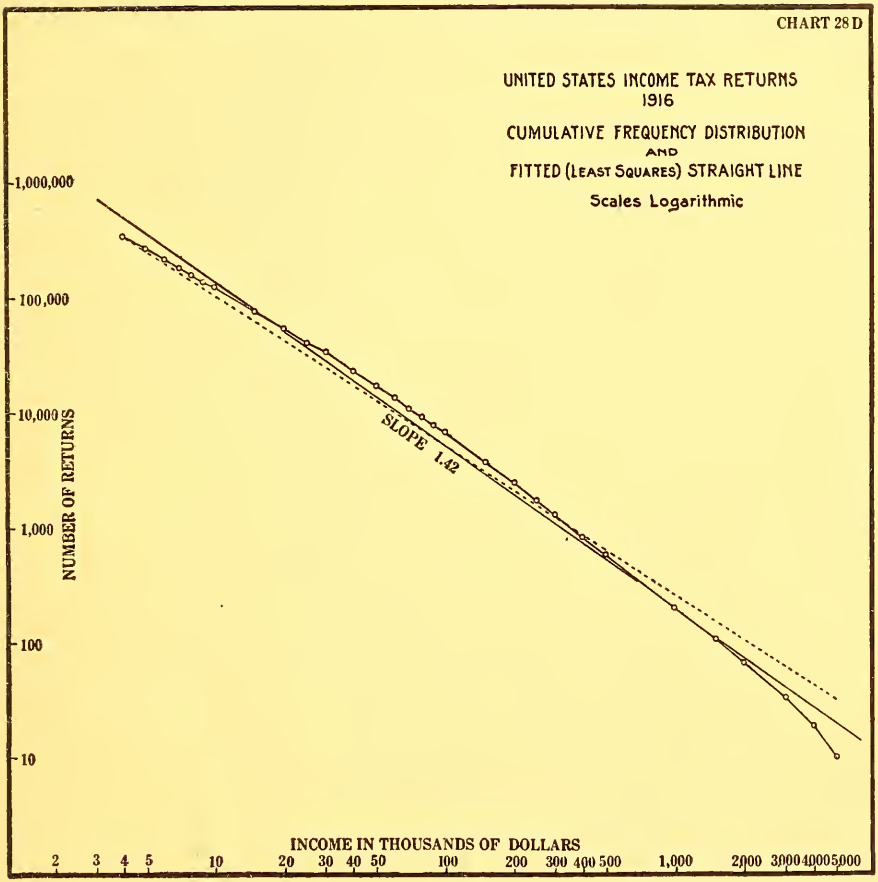


CHART 28 D
UNITED STATES INCOME TAX RETURNS
1916
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic
1,000,000
100,000
10,000
1,000
100
10
NUMBER OF RETURNS
SLOPE 1.42
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2000 3000 4000 5000

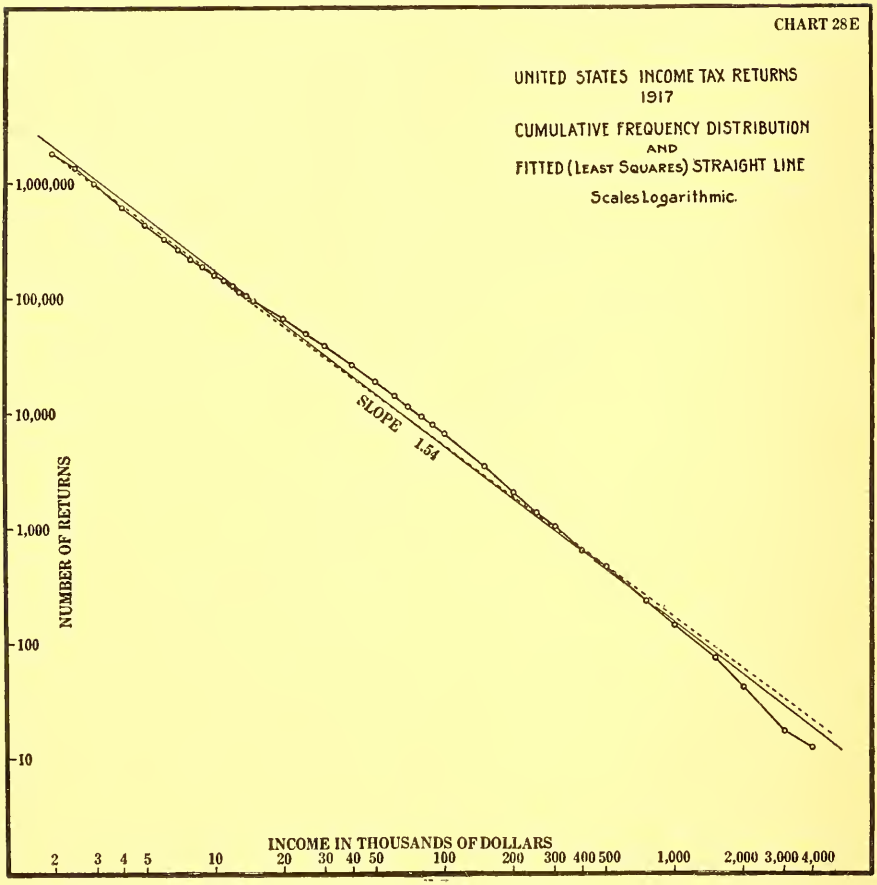


CHART 28E
UNITED STATES INCOME TAX RETURNS
1917
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic.
SLOPE 1.54
NUMBER OF RETURNS
1,000,000
100,000
10,000
1,000
100
10
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000 4,000

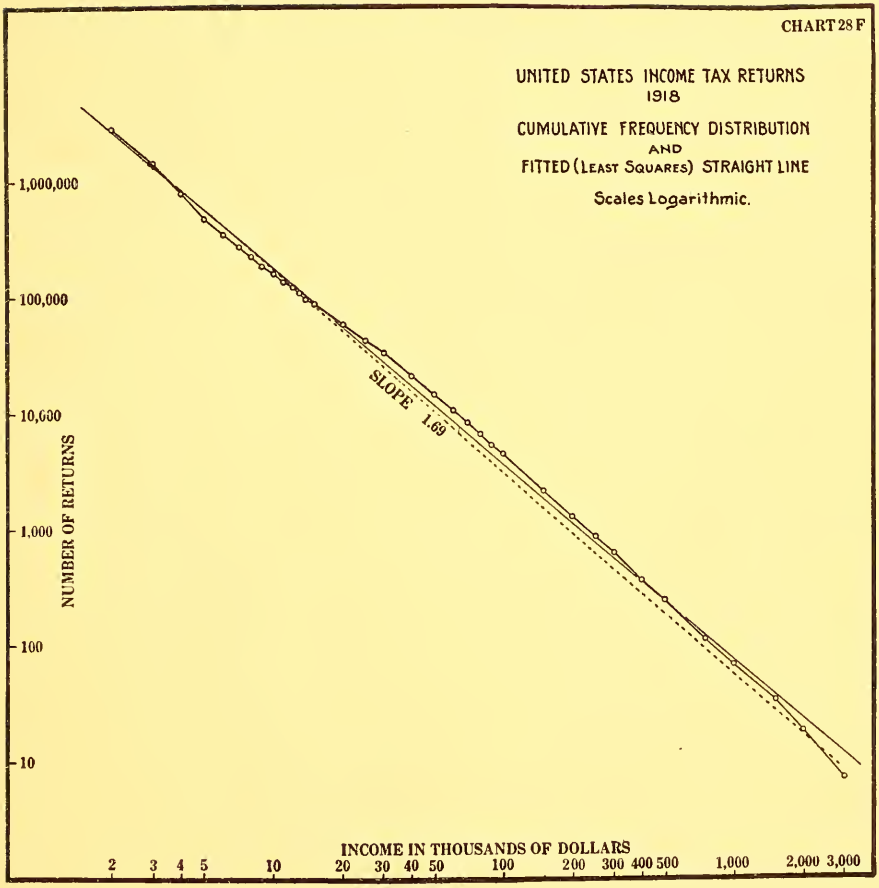


CHART 28 F
UNITED STATES INCOME TAX RETURNS
1918
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic.
1,000,000
100,000
10,000
1,000
100
10
NUMBER OF RETURNS
SLOPE 1.69
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000
INCOME IN THOUSANDS OF DOLLARS

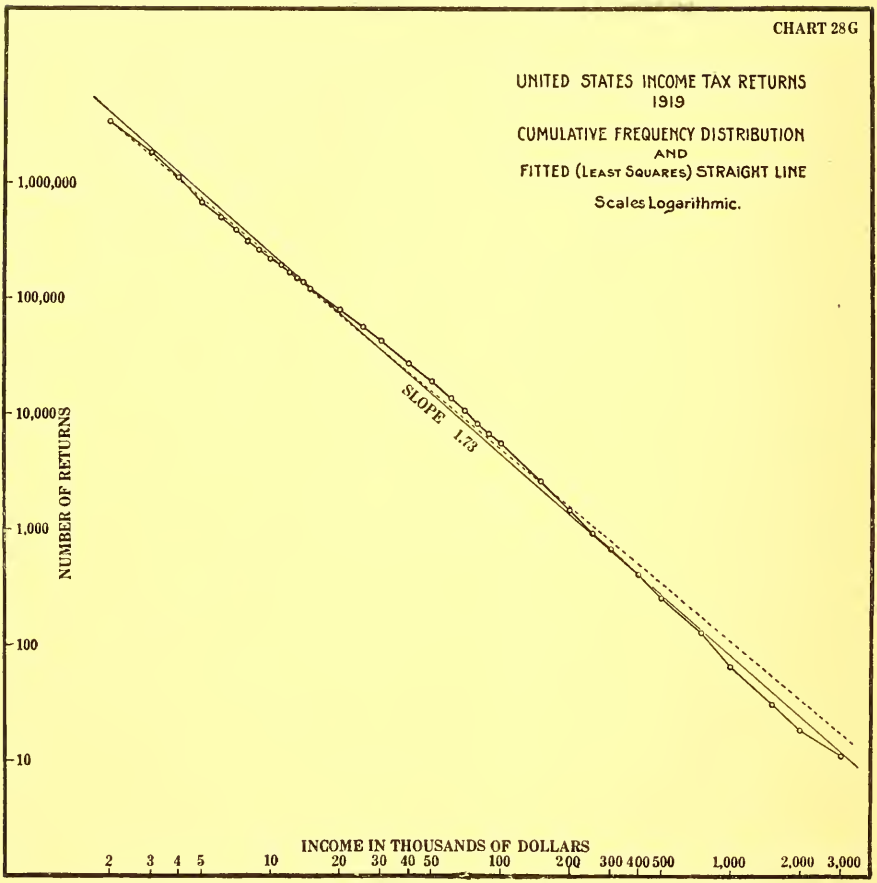


CHART 28G
UNITED STATES INCOME TAX RETURNS
1919
CUMULATIVE FREQUENCY DISTRIBUTION
AND
FITTED (LEAST SQUARES) STRAIGHT LINE
Scales Logarithmic.
SLOPE 1.73
NUMBER OF RETURNS
1,000,000
100,000
10,000
1,000
100
10
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000

TABLE 28B

UNITED STATES INCOME-TAX RETURNS, 1914

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 3,000–$ 4,000	(82,754)				
4,000– 5,000	66,525	101,241	84,683	65.7	78.6
5,000– 10,000	127,448	160,545	115,347	79.4	110.5
10,000– 15,000	34,141	38,630	32,716	88.4	104.4
15,000– 20,000	15,790	15,853	14,102	99.6	112.0
20,000– 25,000	8,672	8,230	7,589	105.4	114.3
25,000– 30,000	5,483	4,879	4,631	112.4	118.4
30,000– 40,000	6,008	5,380	5,267	111.7	114.1
40,000– 50,000	3,185	2,793	2,835	114.0	112.3
50,000– 100,000	5,161	4,430	4,756	116.5	108.5
100,000– 150,000	1,189	1,065.5	1,241	111.6	95.8
150,000– 200,000	406	437.3	535	92.8	75.9
200,000– 250,000	233	227.1	288.1	102.6	80.9
250,000– 300,000	130	134.6	175.5	96.6	74.1
300,000– 400,000	147	148.46	199.9	99.0	73.5
400,000– 500,000	69	77.06	107.6	89.5	64.1
500,000–1,000,000	114	122.20	180.4	93.3	63.2
1,000,000 and over	60	62.78	107.5	95.6	55.8
Total (over $4,000)	274,761	344,256.00	274,761.0		

TABLE 28C

UNITED STATES INCOME-TAX RETURNS, 1915

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 3,000–$ 4,000	(69,045)				
4,000– 5,000	58,949	92,064	68,540	64.0	86.0
5,000– 10,000	120,402	154,507	119,634	77.9	100.6
10,000– 15,000	34,102	40,358	33,013	84.5	103.3
15,000– 20,000	16,475	17,406	14,724	94.7	111.9
20,000– 25,000	9,707	9,372	8,124	103.6	119.5
25,000– 30,000	6,196	5,716	5,050	108.4	122.7
30,000– 40,000	7,005	6,508	5,875	107.6	119.2
40,000– 50,000	4,100	3,503	3,241	117.0	126.5
50,000– 100,000	6,847	5,880	5,653	116.4	121.1
100,000– 150,000	1,793	1,536	1,560	116.7	114.9
150,000– 200,000	724	662.5	695.4	109.3	104.1
200,000– 250,000	386	356.6	383.8	108.2	100.6
250,000– 300,000	216	217.5	238.6	99.3	90.5
300,000– 400,000	254	247.7	277.6	102.5	91.5
400,000– 500,000	122	133.3	153.2	91.5	79.6
500,000–1,000,000	209	223.8	267.1	93.4	78.2
1,000,000 and over	120	133.6	177.3	89.8	67.7
Total (over $4,000)	267,607	338,825.0	267,607.0		

TABLE 28D

UNITED STATES INCOME-TAX RETURNS, 1916

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 3,000-$ 4,000	(85,122)				
4,000- 5,000	72,027	139,096	86,588	51.8	83.2
5,000- 6,000	52,029	84,759	54,221	61.4	96.0
6,000- 7,000	36,470	56,533	36,899	64.5	98.8
7,000- 8,000	26,444	39,846	26,516	66.4	99.7
8,000- 9,000	19,959	29,292	19,801	68.1	100.8
9,000- 10,000	15,651	22,529	15,445	69.5	101.3
10,000- 15,000	45,309	60,668	42,879	74.7	105.7
15,000- 20,000	22,618	26,120	19,311	86.6	117.1
20,000- 25,000	12,953	14,044	10,726	92.2	120.8
25,000- 30,000	8,055	8,558	6,705	94.1	120.1
30,000- 40,000	10,068	9,731	7,854	103.5	128.2
40,000- 50,000	5,611	5,232	4,362	107.2	128.6
50,000- 60,000	3,621	3,189	2,730	113.5	132.6
60,000- 70,000	2,548	2,126	1,857	119.8	137.2
70,000- 80,000	1,787	1,499	1,334.8	119.2	133.9
80,000- 90,000	1,422	1,102	996.8	129.0	142.7
90,000- 100,000	1,074	847	777.5	126.8	138.1
100,000- 150,000	2,900	2,282.1	2,158.4	127.1	134.4
150,000- 200,000	1,284	982.6	972.1	130.7	132.1
200,000- 250,000	726	528.2	539.9	137.4	134.5
250,000- 300,000	427	321.9	337.6	132.6	126.5
300,000- 400,000	469	366.1	395.3	128.1	118.6
400,000- 500,000	245	196.8	219.6	124.5	111.6
500,000-1,000,000	376	329.6	387.4	114.1	97.1
1,000,000-1,500,000	97	85.83	108.7	113.0	89.2
1,500,000-2,000,000	42	36.96	48.88	113.6	85.9
2,000,000-3,000,000	34	31.98	44.19	106.3	76.9
3,000,000-4,000,000	14	13.77	19.91	101.7	70.3
4,000,000-5,000,000	9	7.40	11.05	121.6	81.4
5,000,000 and over	10	19.76	32.87	50.6	30.4
Total (over $4,000)	344,279	510,374.00	344,279.00		

TABLE 28E

UNITED STATES INCOME-TAX RETURNS, 1917

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 1,000–$ 2,000	(1,640,758)				
2,000– 2,500	480,486	618,069	517,512	77.7	92.8
2,500– 3,000	358,221	367,835	284,620	97.4	125.9
3,000– 4,000	374,958	407,366	376,117	92.0	99.7
4,000– 5,000	185,805	212,569	184,854	87.4	100.5
5,000– 6,000	105,988	126,507	111,097	83.8	95.4
6,000– 7,000	64,010	82,746	73,355	77.4	87.3
7,000– 8,000	44,363	57,357	51,285	77.3	86.5
8,000– 9,000	31,769	41,556	37,362	76.4	85.0
9,000– 10,000	24,536	31,551	28,551	77.8	85.9
10,000– 11,000	19,221	24,097	21,900	79.8	87.8
11,000– 12,000	15,035	19,412	17,747	77.5	84.7
12,000– 13,000	12,328	15,707	14,440	78.5	85.4
13,000– 14,000	10,427	12,751	11,761	81.8	88.7
14,000– 15,000	8,789	10,709	9,909	82.1	88.7
15,000– 20,000	29,896	34,161	31,891	87.5	93.7
20,000– 25,000	16,806	17,825	16,876	94.3	99.6
25,000– 30,000	10,571	10,609	10,159	99.6	104.1
30,000– 40,000	12,733	11,749	11,385	108.4	111.8
40,000– 50,000	7,087	6,130	6,021	115.6	117.7
50,000– 60,000	4,541	3,649	3,622	124.4	125.4
60,000– 70,000	2,954	2,387	2,391	123.8	123.5
70,000– 80,000	2,222	1,653.5	1,672	134.4	132.9
80,000– 90,000	1,539	1,198.5	1,217.9	128.4	126.4
90,000– 100,000	1,183	910.0	930.8	130.0	127.1
100,000– 150,000	3,302	2,384.4	2,469.5	138.5	133.7
150,000– 200,000	1,302	985.2	1,039.6	132.2	125.2
200,000– 250,000	703	514.1	550.5	136.7	127.7
250,000– 300,000	342	305.9	330.8	111.8	103.4
300,000– 400,000	380	338.9	371.2	112.1	102.4
400,000– 500,000	179	176.8	196.3	101.2	91.2
500,000– 750,000	225	199.96	225.56	112.5	99.8
750,000–1,000,000	90	82.61	94.97	108.9	94.8
1,000,000–1,500,000	67	68.77	80.51	97.4	83.2
1,500,000–2,000,000	33	28.42	33.90	116.1	97.3
2,000,000–3,000,000	24	23.65	28.71	101.5	83.6
3,000,000–4,000,000	5	9.77	12.10	51.2	41.3
4,000,000–5,000,000	8	5.10	6.40	156.9	125.0
5,000,000 and over	4	12.42	16.25	32.2	24.6
Total (over $2,000)	1,832,132	2,123,640.00	1,832,132.00		

TABLE 28F

UNITED STATES INCOME-TAX RETURNS, 1918

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 1,000-$ 2,000	(1,516,938)				
2,000- 3,000	1,496,878	1,375,372	1,470,366	108.8	101.8
3,000- 4,000	610,095	537,892	566,044	113.4	107.8
4,000- 5,000	322,241	269,674	280,477	119.5	114.9
5,000- 6,000	126,554	155,513	160,366	81.4	78.9
6,000- 7,000	79,152	99,102	101,389	79.9	78.1
7,000- 8,000	51,381	67,184	68,258	76.5	75.3
8,000- 9,000	35,117	47,740	48,266	73.6	72.8
9,000- 10,000	27,152	35,628	35,795	76.2	75.9
10,000- 11,000	20,414	26,793	26,832	76.2	76.1
11,000- 12,000	16,371	21,283	21,231	76.9	77.1
12,000- 13,000	13,202	16,999	16,873	77.7	78.2
13,000- 14,000	10,882	13,638	13,515	79.8	80.5
14,000- 15,000	9,123	11,328	11,165	80.5	81.7
15,000- 20,000	30,227	35,214	34,486	85.8	87.7
20,000- 25,000	16,350	17,654	17,097	92.6	95.6
25,000- 30,000	10,206	10,181	9,762	100.2	104.5
30,000- 40,000	11,887	10,886	10,336	109.2	115.0
40,000- 50,000	6,449	5,458	5,121	118.2	125.9
50,000- 60,000	3,720	3,147	2,928	118.2	127.0
60,000- 70,000	2,441	2,006	1,852	121.7	131.8
70,000- 80,000	1,691	1,359.5	1,246	124.4	135.7
80,000- 90,000	1,210	966.2	881.4	125.2	137.3
90,000- 100,000	934	721.0	653.7	129.5	142.9
100,000- 150,000	2,358	1,822.3	1,636.3	129.4	144.1
150,000- 200,000	866	712.7	629.8	121.5	137.5
200,000- 250,000	401	357.3	312.1	112.2	128.5
250,000- 300,000	247	206.0	178.3	119.9	138.5
300,000- 400,000	260	220.3	188.7	118.0	137.8
400,000- 500,000	122	110.5	93.55	110.4	130.4
500,000- 750,000	132	119.28	99.70	110.7	132.4
750,000-1,000,000	46	46.66	38.36	98.6	119.9
1,000,000-1,500,000	33	36.88	29.88	89.5	110.4
1,500,000-2,000,000	16	14.42	11.50	111.0	139.1
2,000,000-3,000,000	11	11.40	8.96	96.5	122.8
3,000,000-4,000,000	4	4.46	3.44	89.7	116.3
4,000,000-5,000,000	2	2.24	1.71	89.3	117.0
5,000,000 and over	1	4.86	3.60	20.6	27.8
Total (over $2,000)	2,908,176	2,769,408.00	2,908,176.00		

TABLE 28G

UNITED STATES INCOME-TAX RETURNS, 1919

Income class	A U. S. income-tax returns	B Least-squares straight line	C Straight line giving correct total returns and income	Per cent A is of B	Per cent A is of C
$ 1,000–$ 2,000	(1,924,872)				
2,000– 3,000	1,569,741	1,984,285	1,673,688	79.1	93.8
3,000– 4,000	742,334	764,739	660,950	97.1	112.3
4,000– 5,000	438,154	379,330	333,645	115.5	131.3
5,000– 6,000	167,005	216,921	193,470	77.0	86.3
6,000– 7,000	109,674	137,278	123,953	79.9	88.5
7,000– 8,000	73,719	92,511	84,273	79.7	87.5
8,000– 9,000	50,486	65,403	60,066	77.2	84.1
9,000– 10,000	37,967	48,583	44,980	78.1	84.4
10,000– 11,000	28,499	36,386	33,887	78.3	84.1
11,000– 12,000	22,841	28,796	27,027	79.3	84.5
12,000– 13,000	18,423	22,921	21,600	80.4	85.3
13,000– 14,000	15,248	18,329	17,395	83.2	87.7
14,000– 15,000	12,841	15,181	14,459	84.6	88.8
15,000– 20,000	42,028	46,868	45,162	89.7	93.1
20,000– 25,000	22,605	23,249	22,797	97.2	99.2
25,000– 30,000	13,769	13,294	13,228	103.6	104.1
30,000– 40,000	15,410	14,084	14,219	109.4	108.4
40,000– 50,000	8,298	6,986	7,178	118.8	115.6
50,000– 60,000	5,213	3,994	4,162	130.5	125.3
60,000– 70,000	3,196	2,528	2,665	126.4	119.9
70,000– 80,000	2,237	1,704	1,813	131.3	123.4
80,000– 90,000	1,561	1,205	1,292	129.5	120.8
90,000– 100,000	1,113	894	968.3	124.5	114.9
100,000– 150,000	2,983	2,240	2,461.5	133.2	121.2
150,000– 200,000	1,092	863.2	971.6	126.5	112.4
200,000– 250,000	522	428.1	490.4	121.9	106.4
250,000– 300,000	250	245.0	284.4	102.0	87.9
300,000– 400,000	285	259.2	306.0	110.0	93.1
400,000– 500,000	140	128.6	154.4	108.9	90.7
500,000– 750,000	129	137.32	168.2	93.9	76.7
750,000–1,000,000	60	52.89	66.4	113.4	90.4
1,000,000–1,500,000	34	41.25	52.95	82.4	64.2
1,500,000–2,000,000	13	15.89	20.90	81.8	62.2
2,000,000–3,000,000	7	12.40	16.68	56.5	42.0
3,000,000 and over	11	12.15	17.27	90.5	63.7
Total (over $2,000)	3,407,888	3,929,905.00	3,407,888.00		

Why do the least-squares straight lines appear graphically such good fits to the cumulative distributions (for at least the later years) when a merely arithmetic analysis shows even this fit to the cumulative data to be so illusory? *Because the percentage range in the number of persons is so extremely wide.* The deviations of the cumulative data on a double log scale from the least-squares straight line are minute *when compared with the percentage changes in the data from the smallest to the largest incomes.* But this is not helpful. The fact that there are 100,000 times as many persons having incomes over $2,000 per annum as there are persons having incomes over $5,000,000 per annum, does not make a theoretical reading for a particular income interval of twenty or thirty per cent over or under the data reading an unimportant deviation. Charting data on a double log scale may thus become a fertile source of error unless accompanied by careful interpretation.[1] This fact has long been recognized by engineers and others who have had much experience with similar problems in curve fitting.

Another matter of some importance must be noted here. The deviations of the data from the straight lines might be much less than they are and yet constitute extremely bad fits. *The data points (even on a non-cumulative basis) do not flutter erratically from side to side of the fitted lines; they run smoothly, passing through the fitted line at small angles in the way that one curve cuts another.* Now, in curve fitting, such a condition always strongly suggests that the particular mathematical curve used is not in any sense the "law" of the data.

2. Are the slopes of the straight lines fitted to income data from different times and places similar in any significant degree?

[1] The dangers of fitting curves with such a combination as a cumulative distribution and a double log scale, without further analysis, is well illustrated by the results Professor Pareto obtained for Oldenburg. To the Oldenburg data he fitted the rather complicated equation $\log N = \log A - \alpha \log (x + a) - \beta x$ and obtained the following results. (The value Pareto gives for β, namely .0000631, does not check with his calculated figures given below. β = .0000274 is evidently what he intended.)

Income in marks (over)	N	Logarithms of N		Δ
		Observed	Calculated	
300	54,309	4.7349	4.7349	
600	24,043	4.3810	4.4368	—.0558
900	16,660	4.2217	4.2304	—.0086
1,500	9,631	3.9837	3.9409	+.0428
3,000	3,502	3.5443	3.5008	+.0435
6,000	994	2.9974	2.9997	—.0023
9,000	445	2.6484	2.6671	—.0187
15,300	140	2.1461	2.1838	—.0377
30,000	25	1.3979	1.3364	+.0615

(From *Cours d'Economie Politique*, vol. II, p. 307.)

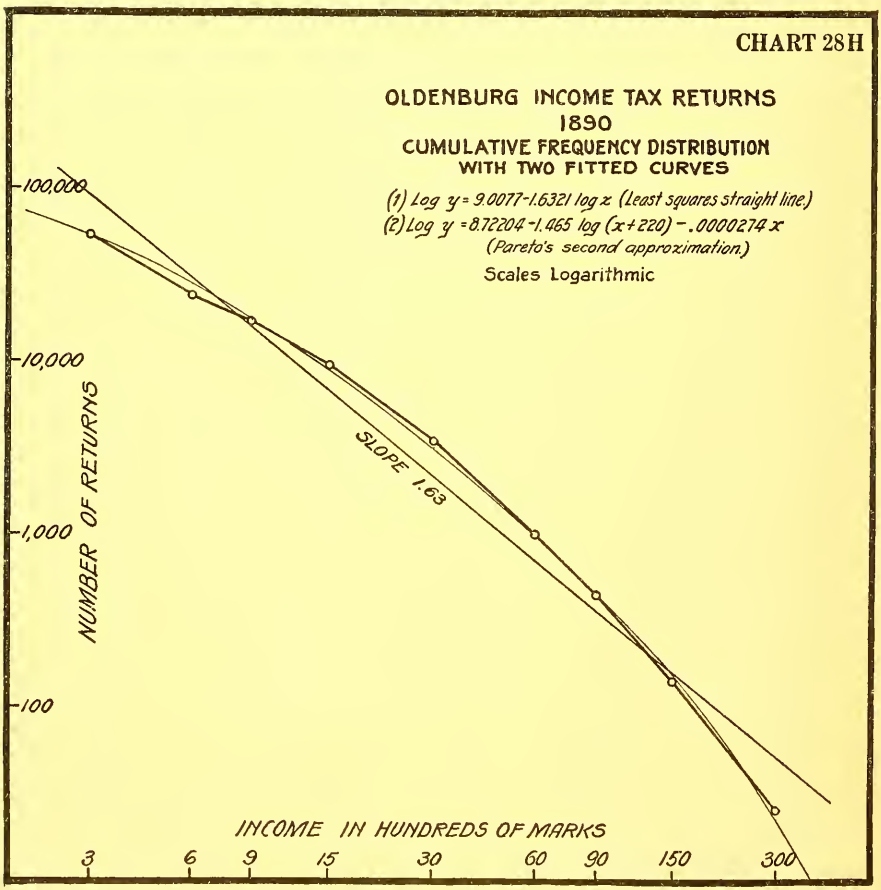
The above table may give the reader a vague idea that the fit is rather good. However, from the above table the following table may be directly derived:

(Note concluded page 364.)

If income distributions charted on a double log scale not only cannot be approximately represented by straight lines, but also differ radically
(Note 1 page 363 concluded.)

Income in marks	Number of persons		Per cent actual are of computed
	Actual	Computed	
300– 600	30,266	26,969	112.2
600– 900	7,383	10,342	71.4
900– 1,500	7,029	8,270	85.0
1,500– 3,000	6,129	5,560	110.2
3,000– 6,000	2,508	2,169	115.6
6,000– 9,000	549	534	102.8
9,000–15,300	305	312	97.8
15,300–30,000	115	131	87.8
Over 30,000	25	22	113.6
Total	54,309	54,309	100.0

The fit no longer impresses one as quite so good. See Chart 28H below.



in shape, it is of course not of great importance whether the straight lines fitted to such data from different times and places have or have not approximately constant slopes. For example, a comparison of Chart 28C showing the cumulative distribution of United States income-tax returns for 1915 on a double log scale and Chart 28F showing similar data for 1918, makes it plain that, even were the slopes of the fitted straight lines for the two years identical, the data curves would still be so different as to make the similarity of slope of the fitted lines of almost no significance.[1]

In considering slopes, let us examine further both the data and the fitted lines for these two years 1915 and 1918. Tables 28I and 28J give some numerical illustrations of the differences between the distributions for the two years. Table 28I gives the number of returns in each income interval each year and the percentages that the 1918 figures are of the 1915 figures.

TABLE 28I

COMPARISON OF UNITED STATES INCOME-TAX RETURNS FOR 1915 AND 1918

Income class	Number of returns		Ratio of 1918 to 1915
	1915	1918	
$ 4,000[a]–$ 5,000	58,949	322,241	5.4664
5,000– 10,000	120,402	319,356	2.6524
10,000– 15,000	34,102	69,992	2.0524
15,000– 20,000	16,475	30,227	1.8347
20,000– 25,000	9,707	16,350	1.6844
25,000– 30,000	6,196	10,206	1.6472
30,000– 40,000	7,005	11,887	1.6969
40,000– 50,000	4,100	6,449	1.5729
50,000– 100,000	6,847	9,996	1.4599
100,000– 150,000	1,793	2,358	1.3151
150,000– 200,000	724	866	1.1961
200,000– 250,000	386	401	1.0389
250,000– 300,000	216	247	1.1435
300,000– 400,000	254	260	1.0236
400,000– 500,000	122	122	1.0000
500,000–1,000,000	209	178	.8517
1,000,000 and over	120	67	.5583

[a] The $3,000–$4,000 class is not included, as in 1915 married persons in that class were exempted while in 1918 they were not.

The change as we pass from the $4,000–$5,000 interval, where the 1918 figures are nearly five-and-a-half times the 1915 figures, to the intervals above $500,000, where the 1918 figures are actually less than the 1915 figures, illustrates the great and fundamental difference between the slopes of the two distributions. However, such a comparison of unadjusted

[1] Compare also the deviations from the fitted lines as given in Tables 28C and 28F.

money intervals, while it throws into relief the differences in slope of the two distributions, is by no means as enlightening for purposes of exhibiting their other essential dissimilarities as a comparison of the two sets of data after they have been adjusted for changes in average (per capita) income and changes in population. Table 28J gives some comparisons between the data for the two years and between the fitted lines for the two years on such an adjusted basis. Two intervals, one in the relatively low income range and the other in the high income range, are used to illustrate the essentially different character of the distributions for the two years.

TABLE 28J

COMPARISONS OF UNITED STATES INCOME-TAX RETURNS FOR THE YEARS 1915 AND 1918 ADJUSTED FOR CHANGES IN AVERAGE (PER CAPITA) INCOME AND CHANGES IN POPULATION

ACTUAL INCOME-TAX DATA

Income intervals	Number of returns (1)	(2)	Fraction of population (3)	(4)	Ratio of Column (4) to Column (3)
	1915	1918	1915	1918	
Between 12 and 13 times average income	21,190	31,197	.00021099	.00029945	1.4193
Between 1,200 and 1,300 times average income	43.85	20.37	.0000004366	.0000001955	.4478
Over 12 times average income	248,600	271,452	.00247536	.00260561	1.0526
	Amount in dollars		Per cent of total income		
	1915	1918	1915	1918	
Over 12 times average income	$4,283,010,735	$5,312,832,516	11.9%	8.7%	.7311

LEAST-SQUARES STRAIGHT LINES

Income intervals	Number of returns (1)	(2)	Fraction of population (3)	(4)	Ratio of Column (4) to Column (3)
	1915	1918	1915	1918	
Between 12 and 13 times average income	32,886	41,730	.00032745	.00040056	1.2233
Between 1,200 and 1,300 times average income	47.63	17.10	.0000004743	.0000001641	.3460

STRAIGHT LINES FITTED TO GIVE THE SAME TOTAL NUMBER OF RETURNS AND THE SAME TOTAL INCOME AS THE INCOME-TAX DATA

Income intervals	Number of returns (1)	(2)	Fraction of population (3)	(4)	Ratio of Column (4) to Column (3)
	1915	1918	1915	1918	
Between 12 and 13 times average income	24,510	42,460	.00024405	.00040756	1.6700
Between 1,200 and 1 300 times average income	54.73	14.15	.0000005450	.0000001358	.2492

NOTES TO TABLE 28J
"Average Income" Intervals

	1915	1918
Average income	$ 358	$ 586
12 times average income	4,296	7,032
13 " " "	4,654	7,618
1,200 " " "	429,600	703,200
1,300 " " "	465,400	761,800

Equations of Fitted Straight Lines on a Cumulative Double Log Basis

	Least-squares lines	Lines giving correct total number of returns and total income
1914	$y = 11.153322 - 1.559256\,x$	$y = 10.557242 - 1.420936\,x$
1915	$y = 10.643299 - 1.419579\,x$	$y = 10.202382 - 1.325598\,x$
1916	$y = 10.839435 - 1.424638\,x$	$y = 10.212702 - 1.298088\,x$
1917	$y = 11.410606 - 1.539996\,x$	$y = 11.170980 - 1.486817\,x$
1918	$y = 12.033697 - 1.693823\,x$	$y = 12.202452 - 1.738497\,x$
1919	$y = 12.320963 - 1.734802\,x$	$y = 12.036155 - 1.667258\,x$

Table 28J needs little discussion. In the section treating actual income-tax data we notice that while the adjusted number of returns in the lower income interval [1] *increased* 41.93 per cent from 1915 to 1918, the adjusted number of returns in the upper income interval [2] *decreased* 55.22 per cent. Moreover, while the adjusted total *number* of returns above the "12-times-average-income" point increased 5.26 per cent, the adjusted *amount of income* reported in these returns decreased 26.89 per cent.

Such figures suggest a rather radical change in the distribution of income during this short three-year period. Similar conclusions may be drawn from the figures for the two pairs of fitted lines, though we must of course remember that these lines describe only very inadequately the actual data. The lines so fitted as to give each year the same total number of returns and total amount of income as the data for that year yield sensational results. While the adjusted number of returns in the lower income-interval *increased* 67 per cent, the adjusted number of returns in the upper income-interval *decreased* 75.08 per cent.

Finally, it has been suggested that changes in the characteristics of the tax-income-distribution in the United States from 1915 to 1918 may be accounted for as the results of the increase in the surtax rates with 1917. We do not believe any large part of these changes can be so accounted for. Notwithstanding the fact that the country entered the European war during the interval, the difference between the 1915 distribution and the 1918 distribution in the United States, extreme as it is, cannot be said to be unreasonably or unbelievably great. Even the changes in the slope of the least-squares line are not phenomenal. Pareto's Prussian figures contain fluctuations in slope from —1.60 to —1.89 while the slope of the least-squares straight line fitted to his Basle data is only —1.25. The

[1] Between 12 and 13 times the average income (per capita) each year.

[2] Between 1,200 and 1,300 times the average income (per capita) each year.

slopes of the least-squares straight lines fitted to the American data are —1.42 for 1915 and —1.69 for 1918.

> 3. If the upper income ranges (or "tails") of income distributions were, when charted on a double log scale, closely similar in shape, would that fact justify the assumption that the lower income ranges were likewise closely similar?

Before attempting to answer the above question, let us summarize the case we have just made against believing the "tails" significantly similar. We can then discuss how much importance such similarity would have did it exist.

We have found upon examination that the approximation to straight lines of the tails of income distributions plotted on double log scales is specious; that the slopes of the fitted straight lines differ sufficiently to produce extreme variations in the relative number of income recipients in the upper as compared with the lower income ranges of the tails; that the upper and lower income ranges of the actual data for different times or places tell a similar story of extreme variation; and that the irregularities in shape of the tails of the actual data, entirely aside from any question of approximating or not approximating straight lines of constant slope, vary greatly from year to year and from country to country, ranging all the way from the irregularities of such distributions as the Oldenburg data, through the American data for 1914, 1915 and 1916 to such an entirely different set of irregularities as those seen in the American data for 1918[1].

At this stage of the discussion the reader may ask whether a general appearance of approximating straight lines on a double log scale, poor as the actual fit may be found to be under analysis, has not some meaning, some significance. The answer to this question must be that, if we were not dealing with a frequency distribution but with a correlation table showing a relationship between *two variables*, an approximation of the regression lines to linearity when charted on a double log scale might easily be the clue to a first approximation to a rational law; but that, on the other hand, approximate linearity in the *tail of a frequency distribution* charted on a double log scale signifies relatively little because it is such a common characteristic of frequency distributions of many and varied types.

The straight line on a double log scale or, in other words, the equation $y = bx^m$, when used to express a relationship between two variables, is, to quote a well-known text on engineering mathematics, "one of the most useful classes of curves in engineering." [2] In deciding what type of equation to use in fitting curves by the method of least squares to data con-

[1] Compare Charts 28H, 28B, 28C, 28D and 28F.
[2] P. Steinmetz, *Engineering Mathematics*, p. 216.

cerning two variables the texts usually mention $y = bx^m$ as "a quite common case." [1] A recent author writes, "simple curves which approximate a large number of empirical data are the parabolic and hyperbolic curves. The equation of such a curve is $y = ax^b$ [$y = bx^m$], parabolic for b positive and hyperbolic for b negative." [2] A widely used text on elementary mathematics speaks of the equation $y = bx^m$ as one of "the three fundamental functions" in practical mathematics.[3] The market for "logarithmic paper" shows what a large number of two-variable relationships may be approximated by this equation. Moreover this equation is often a close first approximation to a rational law. Witness "Boyle's Law." Indeed, sufficient use has not been made of this curve in economic discussions of two-variable problems.

The primary reason why approximation to linearity on a double log scale has no such significance in the case of the *tail of a frequency distribution* as it often has in the case of a two-variable problem is because of the very fact that we are considering the *tail* of the distribution, in other words, a mere fraction of the data. While frequency distributions which can be described throughout their length by a curve of the type $y = bx^m$ are extremely rare, a large percentage of all frequency distributions have *tails* approximating straight lines on a double log scale.[4] It is astonishing how many homogeneous frequency distributions of all kinds may be described with a fair degree of adequacy by means of hyperbolas [5] fitted to the data on a double log scale. Along with this characteristic goes, of course, the possibility of fitting to the tails of such distributions straight lines approximately parallel to the asymptotes of the fitted hyperbola. However we have by no means adequately described an hyperbola when we have stated the fact that one of its asymptotes is (of course) a straight line and that its slope is such and such. Had we even similar information concerning the other asymptote also, we should know little about the hyperbola or the frequency distribution which it would describe on a double log scale. The hyperbola might coincide with its asymptotes and hence have an *angle* at the mode or it might have a very much rounded "top." Such a variation in the shape of the top of the hyperbola [6] would generally correspond to a very great variation in the scatter or "inequality" of the distribution as well as many other characteristics.

[1] D. P. Bartlett, *Method of Least Squares*, p. 33.
[2] J. Lipka, *Graphical and Mechanical Computation*, p. 128.
[3] C. S. Slichter, *Elementary Mathematical Analysis*, preface.
[4] A very large percentage of the remainder have tails approximating straight lines on a natural x log y basis.
[5] N. B. Not a *straight line on the double log scale*, which is a so-called hyperbola on the natural scale, but a true conic section *hyperbola on the double log scale*.
Charts 28K and 28L (Earnings per Hour of 318,946 Male Employees in 1919) illustrate how excellent a fit may often be obtained by means of an hyperbola even though fitted only by selected points. A comparison of the least-squares parabola and the selected-points hyperbola on Chart 28K illustrates also the straight-tail effect.
[6] Compare Karl Pearson's concept of "kurtosis."

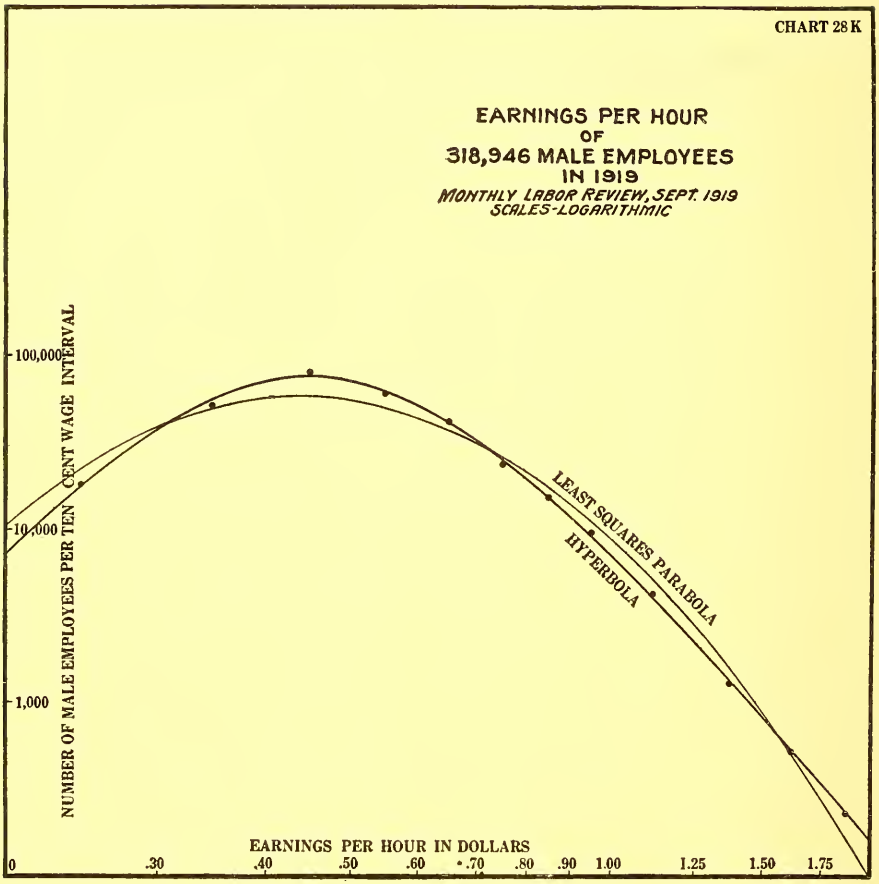
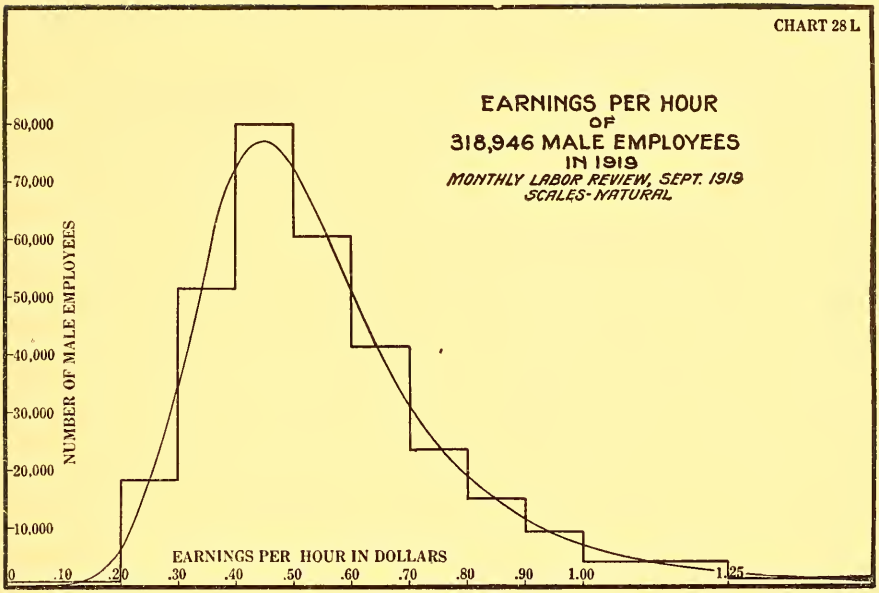
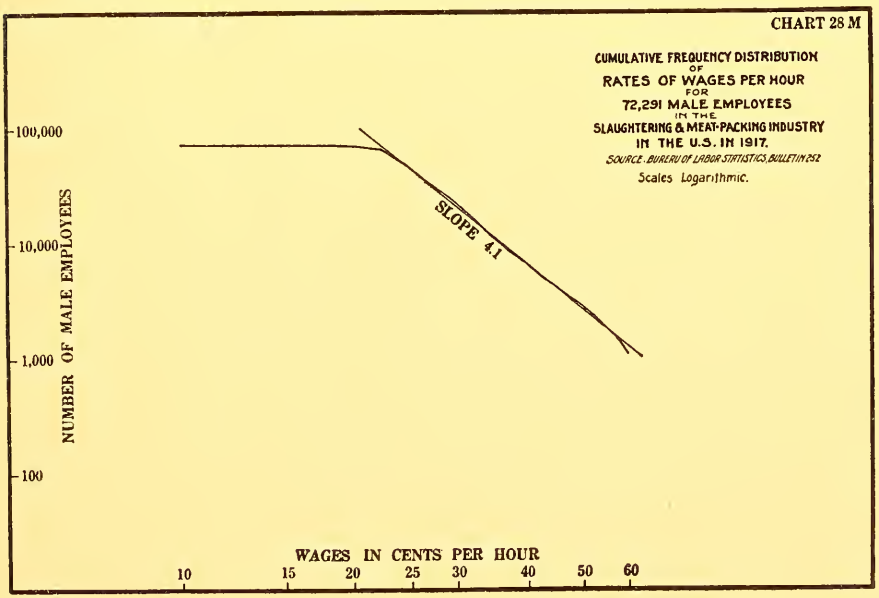
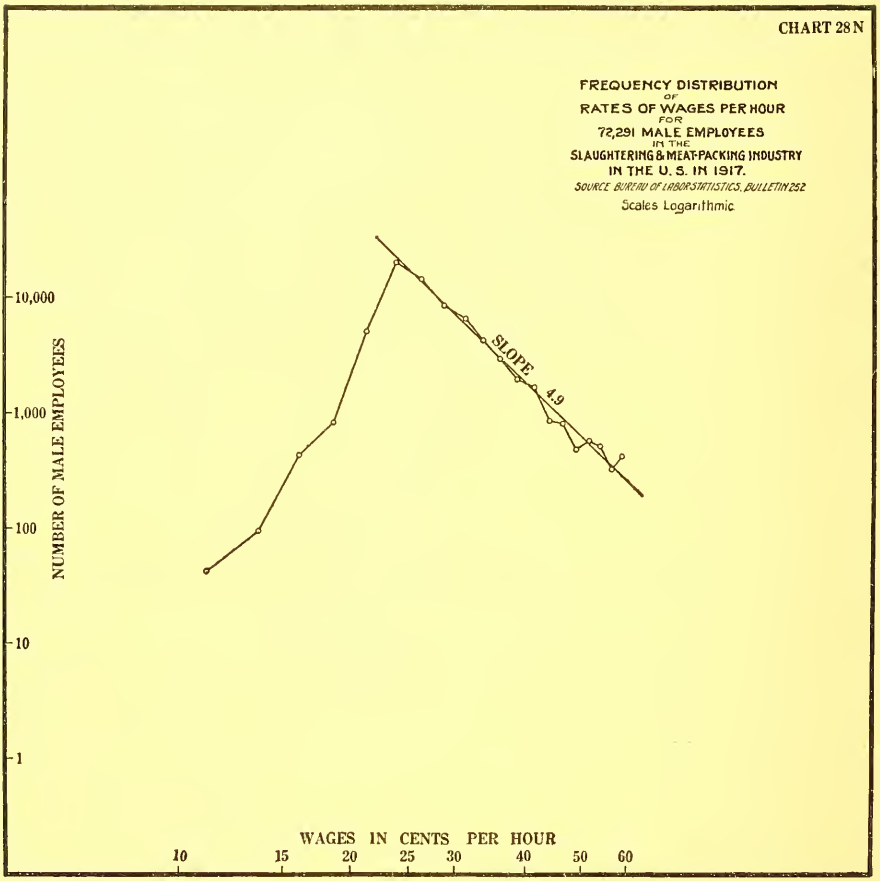


CHART 28 K
EARNINGS PER HOUR
OF
318,946 MALE EMPLOYEES
IN 1919
MONTHLY LABOR REVIEW, SEPT. 1919
SCALES-LOGARITHMIC
NUMBER OF MALE EMPLOYEES PER TEN CENT WAGE INTERVAL
100,000
10,000
1,000
LEAST SQUARES PARABOLA
HYPERBOLA
EARNINGS PER HOUR IN DOLLARS
0
.30
.40
.50
.60
.70
.80
.90
1.00
1.25
1.50
1.75



Rough similarity in the *tails* of two distributions on a double log scale by no means proves even rough similarity in the remainder of the distributions. Charts 28M, 28N, 28O and 28P illustrate both cumulatively



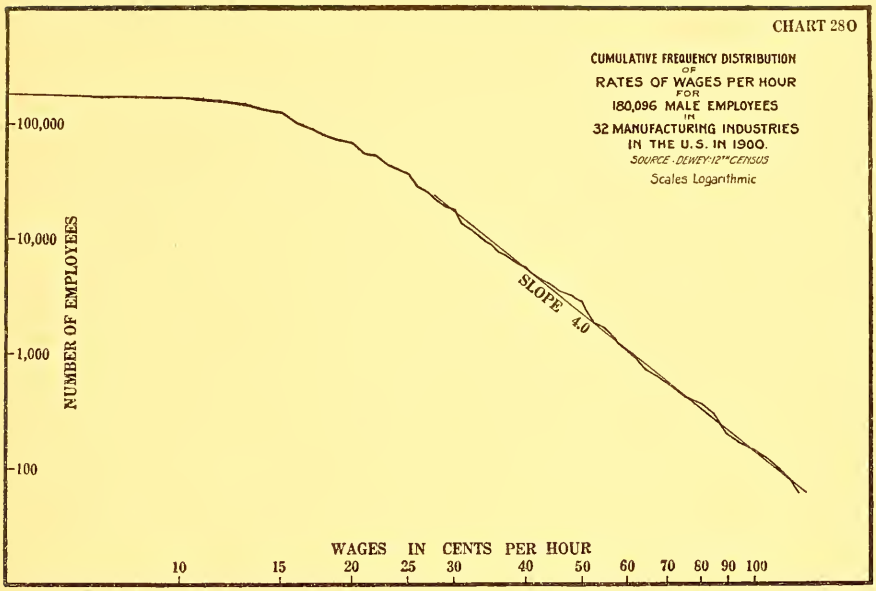


and non-cumulatively on a double log scale two wages distributions whose extreme tails appear roughly to approximate straight lines of about equal slope.[1] Charts 28M and 28N are from data concerning wages per hour of 72,291 male employees in the slaughtering and meat-packing industry in 1917;[2] Charts 28O and 28P are from data concerning wages per hour of 180,096 male employees in 32 manufacturing industries in the United States in 1900.[3] A mere glance at the two non-cumulative distributions will bring home the fact that while they show considerable similarity in the upper income range tails, they are quite dissimilar in the remainder

[1] The illustration shows only "rough similarity" in the extreme tails. However, there seems no good reason for believing that even great similarity in the tails proves similarity in the rest of the distribution. It certainly cannot do so in the case of essentially heterogeneous distributions, such as income distributions.

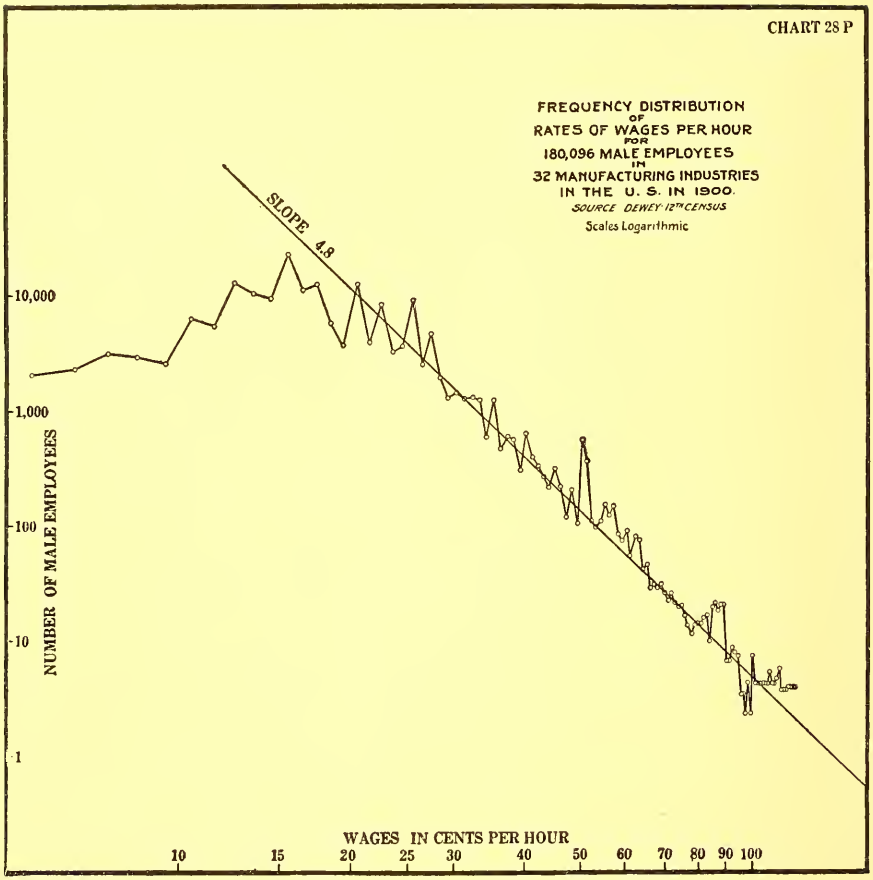
[2] Bureau of Labor Statistics, *Bulletin* No. 252.

[3] Twelfth Census of the United States (1900), *Special Report on Employees and Wages*, Davis R. Dewey.



of the curves. Moreover, in spite of this similarity of tails, the slaughtering and meat-packing distribution has a coefficient of variation of 30.5 while the manufacturing distribution has a coefficient of 47.7. In other words, the relative scatter or "inequality of distribution" is more than one-and-a-half times as great in the manufacturing data as it is in the slaughtering and meat-packing data. Furthermore, no discussion and explanation of greater essential heterogeneity in the one distribution than in the other will offset the fact that the tails are similar but the distributions are different. There seems indeed to be almost no correlation between the slope of the upper-range tail and the degree of scatter in wages distributions. Some distributions showing extremely great scatter have very steep tails, some have not.[1] The frequency curve for the distribution of income in Australia in 1915 is radically different from either the curve for the United States in 1910 constructed by Mr. W. I. King or the curve for the United States in 1918 constructed by the National Bureau of Economic Research.

[1] The tails of wage distributions have in general much greater slopes than those of the upper (i. e., income-tax) range of income distributions. This is an outstanding difference between the two distributions Pareto's conclusions with respect to the convex appearance of the curve for wages are consistent with curves showing *number of dollars* per income-tax interval traceable to wages but not with actual wage distributions showing *number of recipients* per wage interval. Distributions based upon income from effort and distributions based upon income from such sources (mostly profits and income from property) as yield the higher incomes seem to have tails the one as roughly straight as the other. Indeed many wage distributions have tails more closely approximating straight lines than do income-tax data.



Yet all three curves have tails on a double log scale quite as similar as is common with income-tax returns.[1]

From this discussion we may draw the corollary that it is futile to attempt to measure changes in the inequality of distribution of income throughout its range by any function of the mere tail of the income frequency distribution. It seems unnecessary therefore to discuss Pareto's suggestions on this subject.

4. Is it probable that the distribution of income is similar enough from year to year in the same country to make the formulation of any useful general "law" possible?

[1] As will be seen in Chapter 29, there seems reason for believing that the extreme difference between the distribution of incomes obtained by the Australian Census and the estimate made by the National Bureau of Economic Research is due largely to difference in definition of *income* and *income recipient*. However, this does not alter the fact that we have here again two distributions with tails as similar as is usual with income-tax distributions and lower ranges about as different as it is possible to imagine.

Before answering this question we must decide what we should mean by the word *similar*. If income distributions for two years in the same country were such that each distribution included the same individuals and each individual's *income* was twice as large in the second year as it had been in the first year, it would seem reasonable to speak of the distributions as strictly similar. If in a third year (because of a doubling of population due to some hypothetical immigration) the *number* of persons receiving each specified income size was exactly twice what it was in the second year, it would still seem reasonable to speak of the distributions as strictly similar. Tested by any statistical criterion of dispersion which takes account of relative size (such as the coefficient of variation), the dispersion is precisely the same in each of the three years. Moreover the three distributions mentioned above [1] must necessarily have identically the same shape on a double log scale, and furthermore any two distributions which have identically the same shape on a double log scale [2] must necessarily have the same relative dispersion as measured by such indices as the coefficient of variation, interquartile range divided by median, etc. Approximation to identity of shape on a double log scale seems then a useful concept of "similarity." It is the concept implicit in Pareto's work.[3]

Now we have already found considerable evidence that income distributions are not, to a significant degree, similar in shape on a double log scale. The income-tax tails of income distributions for different times and places neither approximate straight lines of constant slope nor approximate one another; they are of distinctly different shapes. Moreover, such tails do not show in respect of their numbers of income recipients and

[1] Or, any distributions whose equations may be reduced to one another by substituting k_1x for x and k_2y for y.

[2] The curve may be thought of as consisting of two parts, which before reduction to logarithms, would be (1) the positive income section and (2) the negative income section with positive signs.

[3] While approximate identity of shape on a natural scale, a natural x and log y scale, or any other similar criterion would constitute a "law," no such approximate identity of shape on such scales has yet been discovered and it seems difficult to advance any very cogent *a priori* reasons for expecting it.

In this connection we must remember that had we the exact figures for the entire frequency curves of the distribution of income in the United States from year to year, if moreover we could imagine definitions of *income* and *income recipient* which would be philosophically satisfactory and statistically usable—and if further we managed year by year to describe our data curves adequately by generalized mathematical frequency curves of more or less complicated variety we should not *necessarily* have arrived at any particularly valuable results. Any series of data may be described to any specified degree of approximation by a power series of the type $y = A + Bx + Cx^2 + Dx^3 + \ldots\ldots$ but such fit is purely empirical and absolutely meaningless except as an illustration of MacLaurin's theorem in the differential calculus. We might be able to describe each year's data rather well by one of Karl Pearson's generalized frequency curves, but if the essential characteristics of the curve—skewness, kurtosis, etc., changed radically from year to year, description of the data by such a curve might well give no clue whatever as to any "law." Not only might the years be different but the fits might be empirical. Professor Edgeworth has well said that "a close fit of a curve to given statistics is not, *per se* and apart from *a priori* reasons, a proof that the curve in question is *the* form proper to the matter in hand. The curve may be adapted to the phenomena merely as the empirically justified system of cycles and epicycles to the planetary movements, not like the ellipse, in favor of which there is the Newtonian demonstration, as well as the Keplerian observations." *Journal of the Royal Statistical Society*, vol. 59, p. 533.

total amounts of income any uniformity of relation to the total number of income recipients and total amount of income in the country, even after adjustments have been made for variations in population and average income.[1] Considerations such as these, reënforce the conclusion which we arrived at from an examination of wage distributions, namely, that there is little necessary relation between the shape of the tail and the shape of the body of a frequency distribution, and have led us to suspect that, even if the tails of income distributions were practically identical in shape, it would be extremely dangerous to conclude therefore that the lower income ranges of the curves were in any way similar.

A most important matter remains to be discussed. What right have we to assume that the heterogeneity necessarily inherent in all income distribution data is not such as inevitably to preclude not only uniformity of shape of the frequency curve from year to year and country to country but also the very possibility of rational mathematical description of any kind unless based upon *parts* rather than the *whole?* What evidence have we as to the extent and nature of heterogeneity in income distribution data?

In the first place we must remember that lower range incomes are predominantly from wages and salaries, while upper range incomes are predominantly from rent, interest, dividends and profits.[2] While 74.67 per cent of the total income reported in the United States in the $1,000–$2,000 income interval in 1918 was traceable to *wages* and *salaries*, only 33.10 per cent of the income in the $10,000–$20,000 interval was from those sources, and only 15.92 per cent of the income in the $100,000–$150,000 interval and 3.27 per cent of the income in the over-$500,000 intervals. On the other hand, while only 1.93 per cent of the total income reported in the $1,000–$2,000 interval in 1918 was traceable to *dividends*, 23.73 per cent was so traceable in the $10,000–$20,000 interval, 43.18 per cent in the $100,000–$150,000 interval, and 59.44 per cent in the over-$500,000 intervals.[3] The difference in constitution of the income at the upper and

[1] Estimated per cent of total income received by highest 5% of income receivers in United States:

1913	33
1914	32
1915	32
1916	34
1917	29
1918	26
1919	24

National Bureau of Economic Research, *Income in the United States*, vol. 1, p. 116.

[2] Compare Professor A. L. Bowley's paper on "The British Super-Tax and the Distribution of Income," *Quarterly Journal of Economics*, February, 1914.

[3] *Statistics of Income 1918*, pp. 10 and 44.

While the reporting of dividends was almost certainly less complete in the lower than in the upper income classes, the difference could not be sufficient to invalidate the general conclusion. Lower range incomes are predominantly wage and salary incomes; upper range incomes are not.

lower ends of the distribution is sufficient to justify the statement that most of the individuals going to make up the lower income range of the frequency curve are wage earners, while the individuals going to make up the upper income range are capitalists and entrepreneurs.[1] What do we know about the shapes of these component distributions? Is the fundamental difference in their relative positions on the income scale their only dissimilarity?

In any particular year the upper income tail of the frequency distribution of income among *capitalists and entrepreneurs* seems not greatly different from the extreme upper income tail of the frequency distribution of income among all classes. This is what we might expect. Not only is the percentage of the total income in the extreme upper income ranges reported as coming from wages and salaries small but much of this so-called wages and salaries income must be merely technical. For example, it is often highly "convenient" to pay "salary" rather than dividends. Furthermore, in so far as the tail of the curve of distribution of income among capitalists and entrepreneurs is not identical with the tail of the general curve, it will show a *smaller* rather than a larger slope, because the percentage of the number of persons in each income interval who are capitalists and entrepreneurs increases as we pass from lower to higher incomes.[2] Now the slopes of the straight lines fitted to the extreme tails of non-cumulative income distributions on a double log scale fluctuate within a range of about 2.4 to 3.0.

The upper range tails of *wages* distributions tell an entirely different story. Aside from surface irregularities often quite evidently traceable to concentration on certain round numbers, the majority of wages distributions have tails which, on a double log scale, are roughly linear.[3] However the *slopes* of straight lines fitted to these tails are much greater than the slopes of corresponding straight lines fitted to income distribution tails.[4] While the slopes of income distribution tails range from about 2.4

[1] Many individuals in the middle income ranges must necessarily be difficult to classify. This does not mean that the concept of heterogeneity is inapplicable. There are countries in which the population is a mixture of Spanish, American Indian, and Negro blood. Now such a population must, for many statistical purposes, be considered extremely heterogeneous even though the percentage of the population which is of *any* pure blood be quite negligible.

[2] In 1917, the only year in which returns are classified according to "principal source of income" (wages and salaries, income from business, income from investment) the difference in slope, in the income range $100,000 to $2,000,000, between the distribution for *all returns* and the distribution for those returns which did not report wages and salaries as their principal source of income was less than .05. The slope in this range of the line fitted to all returns was about 2.64; the business and investment line was about 2.59 and the wages line about 3.21. In 1916, the only year in which returns are classified according to occupations, the distribution of income among *capitalists* shows a slope of only 2.08 while *public service employees* (*civil*) show a slope of 2.70 and *skilled and unskilled laborers* a slope of 2.74.

[3] Attention has already been drawn to the fact that this is a characteristic of many frequency distributions of various kinds.

[4] A further difference between the upper range income distribution among capitalists and entrepreneurs and the upper range of the distribution among all persons seems to be, from the 1916 occupation distributions, that the distribution among all persons shows less of a roll, i. e., is straighter.

to 3.0, the slopes of wages distributions tails commonly range between 4.0 and 6.0. They seldom run below about 4.5; they sometimes run as high as 10.0 and 11.0.

A distribution of wages per hour for 26,183 male employees in iron and steel mills in the United States in 1900 [1] shows a tail with a slope of about 3.35. However, the total of which this is a part, the distribution of wages per hour among 180,096 male employees in 32 manufacturing industries in 1900, shows a tail-slope of about 4.8. The estimated distribution of weekly earnings of 5,470,321 wage earners in the United States in 1905 [2] shows a tail-slope of about 5.0. The distribution of earnings per hour among 318,946 male employees in 29 different industries in the United States in 1919 [3] shows a tail-slope of about 5.86. The distribution of wages per month among 1,939,399 railroad employees in the United States in 1917 [4] shows a tail-slope of about 6.25. The distribution of wages per hour among 43,343 male employees in the foundries and metal working industry of the United States in 1900 [5] shows a tail-slope of about 7.8. The distribution of earnings in a week among 9,633 male employees in the woodworking industry—agricultural implements—in the United States in 1900[6] shows a tail-slope of over 11.0. At the other extreme was the case of the wages-per-hour distribution among 26,183 male employees in American iron and steel mills in 1900 with a slope of 3.35. Both 11.0 and 3.35 are exceptional, but the available data make it clear that wages distributions of either earnings or rates have tail-slopes which are always much greater than the maximum tail-slope of income distributions.

The illustrations in the preceding paragraph are illustrations of the tail-slopes of *wages* distributions among wage earners. However all the evidence points to frequency distributions of *income* among wage earners having tail-slopes only very slightly less steep than the tail-slopes of wages distributions. We have almost no usable data concerning the relation between individual wage distributions and income distributions for the same individuals, but we have a few samples showing the relation between family earnings distributions and family income distributions.[7] Moreover, we can without great risk base certain extremely general conclusions

[1] Twelfth Census of the United States (1900), *Special Report on Employees and Wages*, Davis R. Dewey.
[2] *1905 Census of Manufactures*, Part IV, p. 647.
[3] *Monthly Labor Review*, Sept., 1919.
[4] *Report of the Railroad Wage Commission to the Director General of Railroads*, 1919, p. 96.
[5] Twelfth Census of the United States (1900), *Special Report on Employees and Wages*, Davis R. Dewey.
[6] Twelfth Census of the United States (1900), *Special Report on Employees and Wages*, Davis R. Dewey.
[7] The reader must not confuse the percentage of the income not derived from wages going to *wage-earners* in any particular income class with the percentage of the income not derived from wages going to *all income recipients* in any particular income class. Some of these last recipients are not wage earners at all, they receive no wages. Information concerning the second of these relations *but not the first* is given in the income tax reports.

concerning individual wage-earners' income distributions on these family data. The upper tails of the family-wage distributions *are* the tails of the wage distributions for the individuals who are the heads of the families. This is apparent from an analysis of the samples. Now income from rents and investments belongs almost totally to heads of families. Such income is however so small in amount that it cannot alter appreciably the slope of the tail.[1] While income from other sources than rents and investments (lodgers, garden and poultry, gifts and miscellaneous) may not be so confidently placed to the credit of the head of the family, this item changes its percentage relation to the total income so slowly as to be negligible in its effect upon the tail-slope of the distribution.[2] Notwithstanding the danger of reasoning too assuredly about individuals from these picked family distributions, we seem justified in believing that the tail-slopes of income distributions among individual wage earners are not very different from the tail-slopes of wage distributions among the same individuals.[3]

The upper tail-slopes of income distributions among typical wage earners

[1] For example, in the report on the incomes of 12,096 white families published in the *Monthly Labor Review* for December, 1919, we find the income from rents and investments less than one per cent of the total family income for each of the income intervals.

Income group	Percentage income from rents and investments is of total income
Under $900	.079
$ 900–$1,200	.176
1,200– 1,500	.410
1,500– 1,800	.551
1,800– 2,100	.606
2,100– 2,500	.998
2,500 and over	.778

[2] As a somewhat extreme example, the Bureau of Labor investigation mentioned in the preceding note shows the following relations between total family earnings and total family income (including income from rents and investments, lodgers, garden and poultry, gifts and miscellaneous).

Income group	Percentage that total earnings are of total income
Under $900	96.2
$ 900–$1,200	96.5
1,200– 1,500	96.3
1,500– 1,800	96.0
1,800– 2,100	96.3
2,100– 2,500	95.1
2,500 and over	96.2

[3] Further corroboratory evidence, of some slight importance, that the tail-slopes of wage distributions among wage earners are not very different from the tail-slopes of income distributions among wage earners is yielded by the fact that the tail-slopes of income distributions among families (which are virtually identical with the tail-slopes of both income and wage distributions among the heads of these families) have roughly the same range as the tail-slopes of wage distributions among individuals. The British investigation into the incomes of 7,616 workingmen's families in the United States in 1909 shows a tail-slope of about 3.5. (Report of the British Board of Trade on *Cost of Living in American Towns*, 1911. [Cd. 5609], p. XLIV.) The Bureau of Labor's investigation into the income of 12,096 white families in 1919 shows a tail-slope of about 4.0. Mr. Arthur T. Emery's extremely careful investigation into the incomes of 2,000 Chicago households in 1918 shows a tail-slope of about 4.4. At the other extreme we find that the Bureau of Labor's investigation into the income of 11,156 families in 1903 (*Eighteenth Annual Report of the Commissioner of Labor*, 1903, p. 558) shows a tail-slope of about 10.0, and that Mr. R. C. Chapin's investigation into the income of 391 workingmen's families in New York City (*Standard of Living Among Workingmen's Families in New York City*, p. 44) also shows a slope of about 10.0. The tails of these last two cases are very irregular so that the slope itself is not determinable with much precision.

may then be assumed to have much greater slopes than the upper tail-slopes of income distributions among capitalists and entrepreneurs. It does not seem possible to make any very definite statement concerning the body and lower tail of the capitalist and entrepreneurial distribution—even in so far as that term is a significant one.[1] All the evidence suggests that the mode of what we have termed the capitalist-entrepreneurial distribution is consistently higher than the wage-earners' mode.[2] Its lower income tail undoubtedly reaches out into the negative income range, which the tail of the wage-earners' distribution may, both *a priori* and from evidence, be assumed not to do. It seems a not irrational conclusion then to speak of the capitalist-entrepreneurial distribution as having a lesser tail-slope than the wage-earners' distribution on the *lower* income side as well as on the upper income side,[3] and as a corollary almost certainly a much greater dispersion both actual and relative than the wage-earners' distribution.

Though the above generalizations concerning differences between the wage-earners' income distribution and the capitalist-entrepreneurial income distribution seem sound, they tell but a fraction of the story. Aside from the difficulty of classifying all income recipients in one or the other of these two classes, we are faced with the further fact that investigation suggests that our two component distributions are themselves exceedingly heterogeneous.[4] We have already noted that wage distributions for different occupations and times are extremely dissimilar in shape and we suspect that the same applies to capitalist-entrepreneurial distributions. For example, what little data we possess suggest that the distribution of income among farmers has little in common with other entrepreneurial distributions.

Moreover, the component distributions, into which it would seem necessary to break up the complete income distribution before any rational description would be possible, not only have different shapes and different positions on the income scale (i. e., different modes, arithmetic averages, etc.), but *the relative position with respect to one another on the income scale* of these different component distributions changes from year to year.[5]

[1] In the total income curve there is a broad twilight zone where individuals are often both wage or salary earners and capitalists or even entrepreneurs.

[2] In the 1916 occupation distributions the only occupations showing more returns for the $4,000–$5,000 interval than the $3,000–$4,000 (that is the only occupations showing any suggestion of a mode) are of a capitalistic or entrepreneurial description—bankers; stock-brokers; insurance brokers; other brokers; hotel proprietors and restaurateurs; manufacturers; merchants; storekeepers; jobbers; commission merchants, etc.; mine owners and mine operators; saloon keepers; sportsmen and turfmen.

[3] Of course the very word *slope* is an ambiguous term to use concerning the tail of a curve which enters the second quadrant.

[4] Evidence suggesting definite heterogeneity in the "wage and salary" figures of the income-tax returns is presented in Chapter 30.

[5] This fact is one of the simpler pieces of evidence against the existence of a "law." Of course, even though the income distribution were made up of heterogeneous material, if the

Table 28Q [1] is interesting as showing the changes in the relative positions of the arithmetic averages of different wage distributions in 1909, 1913 and 1918.

TABLE 28Q

CHANGES IN THE RELATIVE POSITIONS OF THE AVERAGE ANNUAL EARNINGS OF EMPLOYEES ENGAGED IN VARIOUS INDUSTRIES

Industry	1909	1913	1918
All Industries	100.0	100.0	100.0
Agriculture	48.2	45.4	54.7
Production of Minerals	95.7	104.4	119.0
Manufacturing:			
Factories	91.2	97.5	103.5
Hand Trades	111.7	103.5	110.8
All Transportation	104.9	105.4	119.3
Railway, Express, Pullman, Switching and Terminal Cos	104.0	108.2	129.3
Street Railway, Electric Light and Power, Telegraph and Telephone Cos	99.5	93.8	81.4
Transportation by Water	123.5	114.1	147.5
Banking	123.0	128.6	135.5
Government	118.1	113.8	83.0
Unclassified Industries	114.4	107.7	97.8

The data are so inadequate that the construction of a similar table for capitalist-entrepreneurial distributions is not feasible. However, there are comparatively good figures for total income of farmers and total number of farmers year by year.[2] The average incomes of farmers, year by year, were the following percentages of the estimated average incomes of all persons gainfully employed in the country.

	Percentages
1910	75.19
1911	69.13
1912	72.41
1913	74.88
1914	76.33
1915	80.45
1916	82.85
1917	104.51
1918	109.68
1919	103.95
1920	63.88

This is a wide range.

Exactly what effects have such internal movements of the component distributions upon the total income frequency distribution curve? This is a difficult question to answer as we have not sufficient data to break

component parts remained constant in shape *and in their relative positions with respect to one another on the income scale*, these relations would of themselves constitute a "law."

[1] Based upon *Income in the United States*, Vol. I, pp. 102 and 103.

[2] See *Income in the United States*, Vol. I, p. 112.

down the total, composite, curve into its component parts with any degree of confidence.[1] However, the movements of wages in recent years would appear to give us a clue to the sort of phenomena we might expect to find if we had complete and adequate data.

The slopes of the upper income tails of wages distributions are great, 4 to 5 or more.[2] Now the wage curve moved up strongly from 1917 to 1918 if we may judge by averages. The average wage of all wage earners in the United States [3] increased 15.6 per cent [4] from 1917 to 1918. During the same period the average income of farmers increased 19.1 per cent [5] and the average income of persons other than wage earners and farmers remained nearly constant. Total amounts of income by sources in millions of dollars were:

	1917	1918	Percentage 1918 was of 1917
Total Wages [a]	$27,795	$32,575	117.20
Total Farmers' Income	8,800	10,500	119.32
All other Income	17,265	17,291	100.15
Total Income	$53,860	$60,366	112.08

[a] Includes pensions, etc., and includes soldiers, sailors, and marines.

Stockholders in corporations saw income from that source actually decline from 1917 to 1918.[6] What happened to American income-tax returns during this time?

[1] The processes by which the income distribution curve published in *Income in the United States*, Vol. I, pp. 132–135 was arrived at were such that to use that material here would practically amount to circular reasoning. The conclusions arrived at here were used in building up that curve.

[2] The slope of the tail of the wage and salary curve in the 1917 income tax returns is only about 3.21 (compare, note 2, p. 377). However we must remember that the individuals there classified are largely of an entirely different type of "wage-earner" from those in the lower groups. In this upper group occur the salaried entrepreneurs, professional men, etc., and those whose "salaries" are really profits or dividends. The evidence points to a rather distinct and significant heterogeneity along this division in the wage and salary distribution. See Chapter 30.

[3] Excluding soldiers, sailors, and marines, and professional classes but including officials and "salaried entrepreneurs."

[4] From $945 per annum in 1917 to $1,092 per annum in 1918.

[5] From $1,370 per annum in 1917 to $1,632 per annum in 1918.

[6] CORPORATION DIVIDENDS, SURPLUS AND EARNINGS

(In millions of dollars)

	Dividends	Surplus	Net earnings
1917	3,995	3,963	7,958
1918	2,568	1,945	4,513

See page 324.

TOTAL AMOUNT OF NET INCOME RETURNED BY SOURCES (RETURNS REPORTING OVER $2,000 PER ANNUM NET INCOME) [a]

(Millions of dollars)

Income class	Wages and salaries		All other sources [b]	
	1917	1918	1917	1918
Over $2,000	$3,648	$6,493	$7,543	$7,198
2,000– 4,000	1,553	3,687	1,799	2,036
4,000– 5,000	301	703	528	736
5,000–10,000	661	849	1,167	1,296
Over 10,000	1,133	1,254	4,049	3,130

[a] Wages income from returns reporting between $1,000 and $2,000 per annum is not available for 1917.

[b] "Other sources" are total *net* income minus wages and salaries, i. e., total *general deductions* have been assumed as deductible from other sources (gross). All things considered, this seems proper here though it may easily be criticised. In connection with changes in the relation between *net* and *gross* income from 1917 to 1918 see Chapter 30, pp. 401 and 402.

While reported income from all other sources than wages and salaries declined 4.6 per cent,[1] reported income from wages and salaries increased 78.0 per cent.[2] Moreover, the great increases in wages and salaries were in the lowest intervals. The wage curve with its steep tail-slope was moving over into the income tax ranges.[3] The effect upon the total curve is very pronounced, as may be seen from Table 28R.

TABLE 28R

AMERICAN INCOME TAX RETURNS IN 1917 AND 1918

Total Number of Returns
(In thousands)

	1917	1918	Percentage 1918 was of 1917
$2,000–$4,000	1,214	2,107	173.56
4,000– 5,000	186	322	173.12
5,000–10,000	271	319	117.71
Over 10,000	162	160	98.77

On a double log scale we see the curve changing its shape radically. While the 1917 curve is comparatively smooth and regular, the 1918 curve develops a distinct "bulge" in the lower ranges.[4]

The preceding discussion has been concerned with equal dollar-income

[1] Had "other sources" been taken gross instead of net, that item would have shown an increase of 5.3 per cent instead of a decrease of 4.6 per cent.

[2] The actual spread is still greater than the figures show. Income from professions, which in 1917 was classed under *wages*, in 1918 and 1919 was classed under *business*.

[3] This seems to be a fact though it is not the whole story. The "intensive drive" of 1919 may easily account for some of the increase. See Chapter 30 for a discussion of the probable extent of this influence.

[4] See *Income in the United States*, Vol. I, Charts 28 and 30.

intervals. However, $2,000 income in 1918 was relatively less than $2,000 income in 1917. The average (per capita) income of the country was $523 in 1917 and $586 in 1918.[1] The adjustment is theoretically crude, but $2,241[2] in 1918 might be considered as in one sense equivalent to $2,000 in 1917. The results of comparisons of the two years upon this basis are given in Table 28S.[3]

TABLE 28S

INCOME RETURNED—BY SOURCES

(Millions of dollars)

1917

Income class	Wages and salaries	Total *net* income	Total *net* income minus wages and salaries	Total *gross* income	Total *gross* income minus wages and salaries
$2,000–$4,000....	$1,553	$3,352	$1,799	$3,713	$2,161
4,000– 5,000....	301	829	528	895	594
5,000–10,000....	661	1,828	1,167	1,951	1,290
Over 10,000....	1,133	5,182	4,049	5,518	4,384
1918					
$2,241–$4,482....	$3,236	$5,359	$2,123	$5,766	$2,530
4,482– 5,602....	498	1,111	613	1,247	749
5,602–11,205....	773	1,960	1,187	2,315	1,542
Over 11,205.....	1,153	4,129	2,976	4,842	3,689
(Multiplied by $\frac{523}{586}$, that is reduced to "1917 dollars")					
$2,241–$4,482....	$2,888	$4,783	$1,895	$5,146	$2,258
4,482– 5,602....	445	992	547	1,113	668
5,602–11,205....	690	1,749	1,059	2,066	1,376
Over 11,205....	1,029	3,685	2,656	4,321	3,292
(Percentages of Total Income of Country)					
1917					
$2,000–$4,000....	2.88	6.22	3.34	6.89	4.01
4,000– 5,000....	.56	1.54	.98	1.66	1.10
5,000–10,000....	1.23	3.39	2.16	3.62	2.39
Over 10,00.....	2.10	9.61	7.51	10.24	8.14
1918					
$2,241–$4,482....	5.30	8.78	3.48	9.45	4.15
4,482– 5,602....	.82	1.82	1.00	2.05	1.23
5,602–11,205....	1.27	3.21	1.94	3.80	2.53
Over 11,205....	1.89	6.77	4.88	7.94	6.05

[1] *Income in the United States*, Vol. I, p. 76.

[2] $\$2,000 \times \frac{586}{523}$.

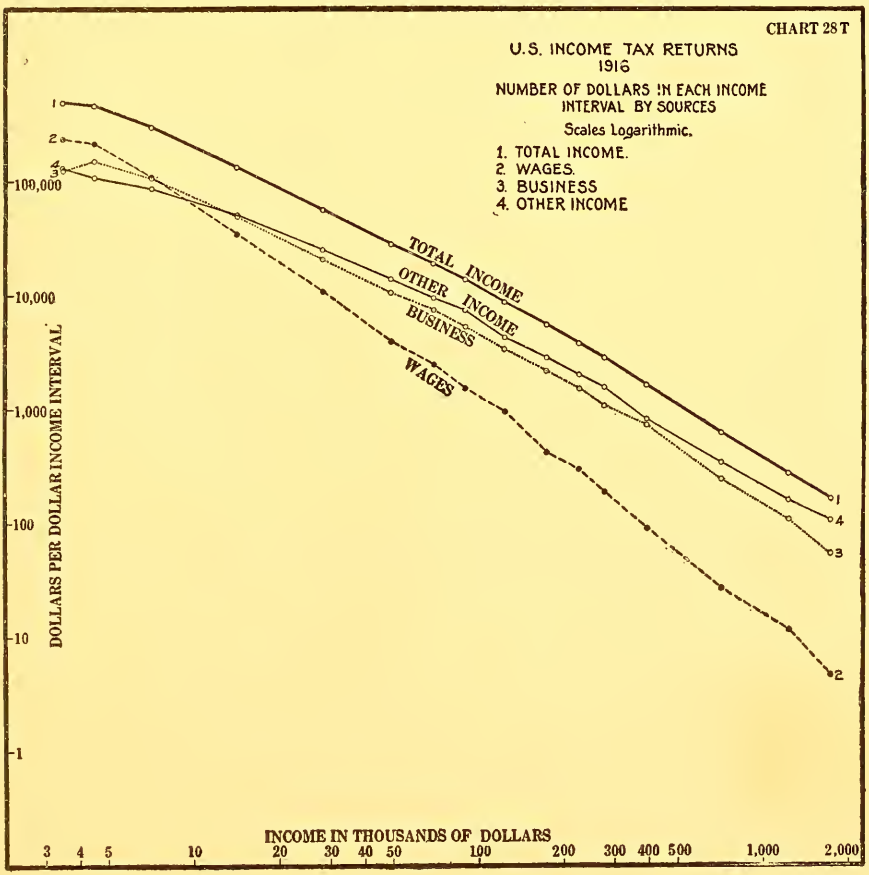
[3] The figures for the amounts of income in the irregular 1918 income intervals of that table ($2,241–$4,482, etc.) were calculated by straight line interpolation on a double log scale applied to the even thousand dollar intervals of the income-tax returns. Though the total income curve does not approximate linearity it may be assumed linear within the small range of one income tax interval without serious error.

(Table 28S concluded.)

NUMBER OF RETURNS
(Thousands)

Income class	1917	Income class	1918	Percentage 1918 was of 1917
$2,000–$4,000.....	1,214	$2,241–$4,482.....	1,758	144.81
4,000– 5,000.....	186	4,482– 5,602.....	220	118.28
5,000–10,000.....	271	5,602–11,205.....	260	95.94
Over 10,000.....	162	Over 11,205.....	136	83.95

It is from this table once again apparent that the wage distribution moved independently up on the income scale and that the effect of this movement was confined to the lowest income intervals. Charts 28T, 28U, 28V, 28W, 28X, 28Y, 28Z, and 28AA which show the number of dollars income per dollar-income interval, by sources, are enlightening as illustrating in still



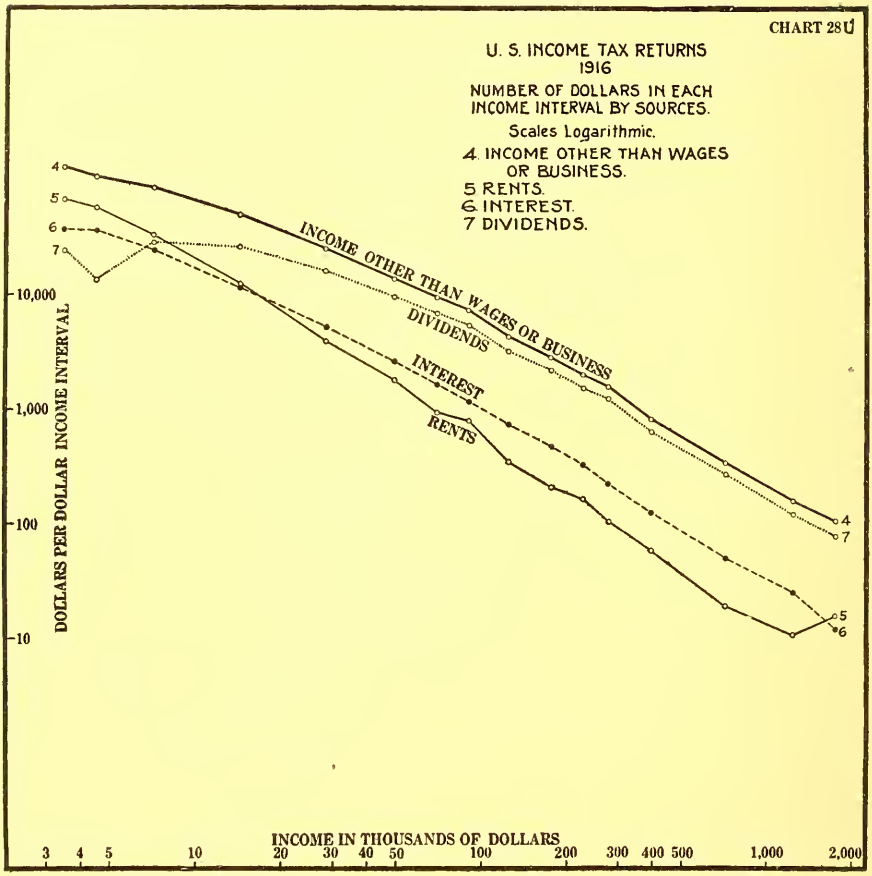


CHART 28U
U. S. INCOME TAX RETURNS
1916
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES.
Scales Logarithmic.
4. INCOME OTHER THAN WAGES OR BUSINESS.
5 RENTS.
6 INTEREST.
7 DIVIDENDS.
INCOME OTHER THAN WAGES OR BUSINESS
DIVIDENDS
INTEREST
RENTS
DOLLARS PER DOLLAR INCOME INTERVAL
10,000
1,000
100
10
INCOME IN THOUSANDS OF DOLLARS
3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000

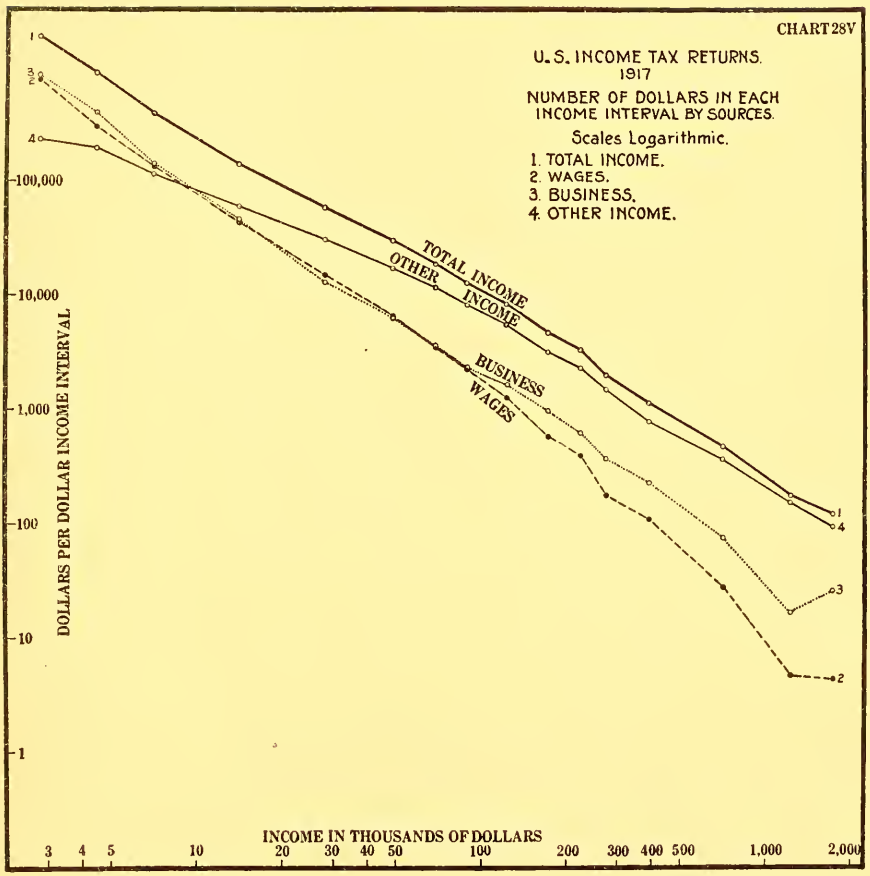


CHART 28V
U. S. INCOME TAX RETURNS.
1917
NUMBER OF DOLLARS IN EACH INCOME INTERVAL BY SOURCES.
Scales Logarithmic.
1. TOTAL INCOME.
2. WAGES.
3. BUSINESS.
4. OTHER INCOME.
TOTAL INCOME
OTHER INCOME
BUSINESS
WAGES
DOLLARS PER DOLLAR INCOME INTERVAL
100,000
10,000
1,000
100
10
1
INCOME IN THOUSANDS OF DOLLARS
3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000

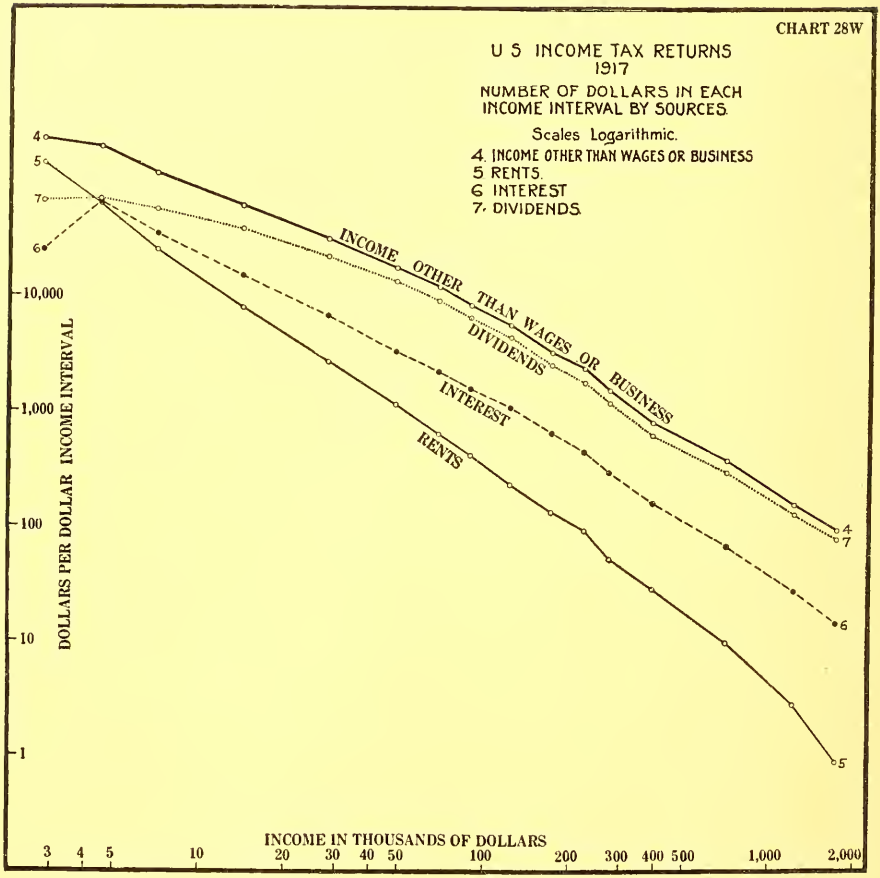


CHART 28W
U S INCOME TAX RETURNS
1917
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES
Scales Logarithmic.
4. INCOME OTHER THAN WAGES OR BUSINESS
5 RENTS.
6 INTEREST
7. DIVIDENDS.
INCOME OTHER THAN WAGES OR BUSINESS
DIVIDENDS
INTEREST
RENTS
DOLLARS PER DOLLAR INCOME INTERVAL
10,000
1,000
100
10
1
INCOME IN THOUSANDS OF DOLLARS
3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000

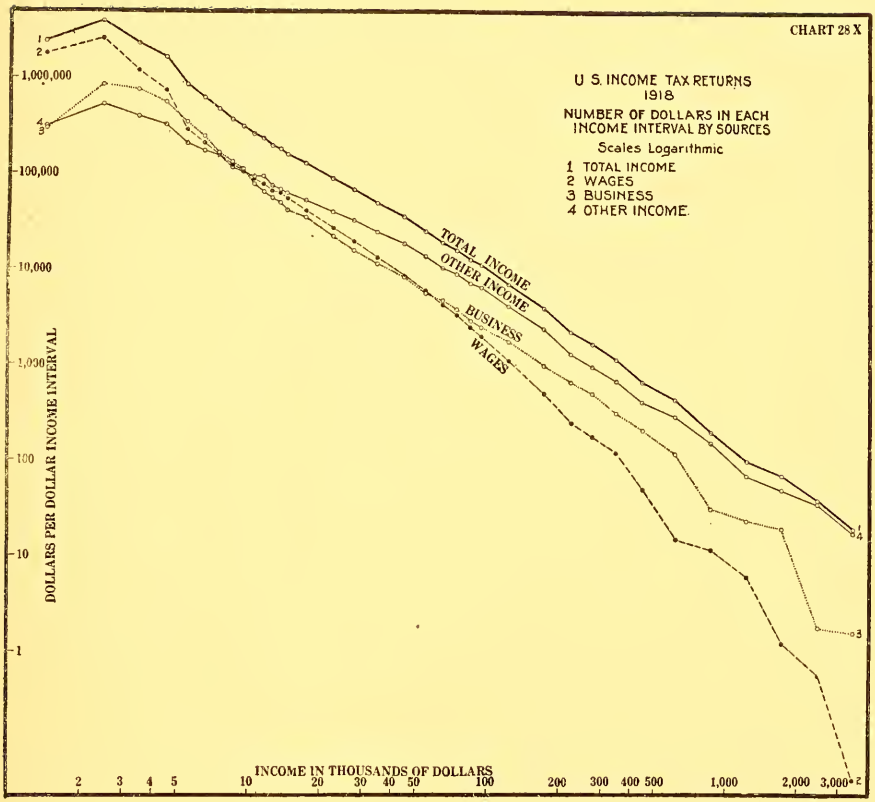


CHART 28 X
U. S. INCOME TAX RETURNS
1918
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES
Scales Logarithmic
1 TOTAL INCOME
2 WAGES
3 BUSINESS
4 OTHER INCOME.
TOTAL INCOME
OTHER INCOME
BUSINESS
WAGES
DOLLARS PER DOLLAR INCOME INTERVAL
1,000,000
100,000
10,000
1,000
100
10
1
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000

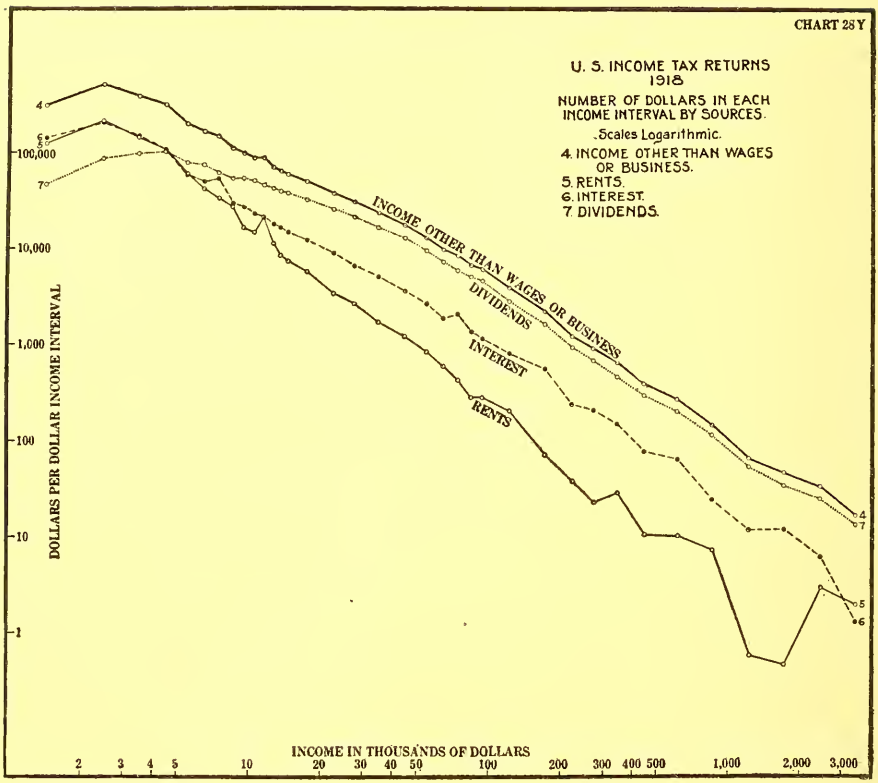


CHART 28 Y
U. S. INCOME TAX RETURNS
1918
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES.
.Scales Logarithmic.
4. INCOME OTHER THAN WAGES
OR BUSINESS.
5. RENTS.
6. INTEREST.
7. DIVIDENDS.
DOLLARS PER DOLLAR INCOME INTERVAL
100,000
10,000
1,000
100
10
1
INCOME OTHER THAN WAGES OR BUSINESS
DIVIDENDS
INTEREST
RENTS
4
7
5
6
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000

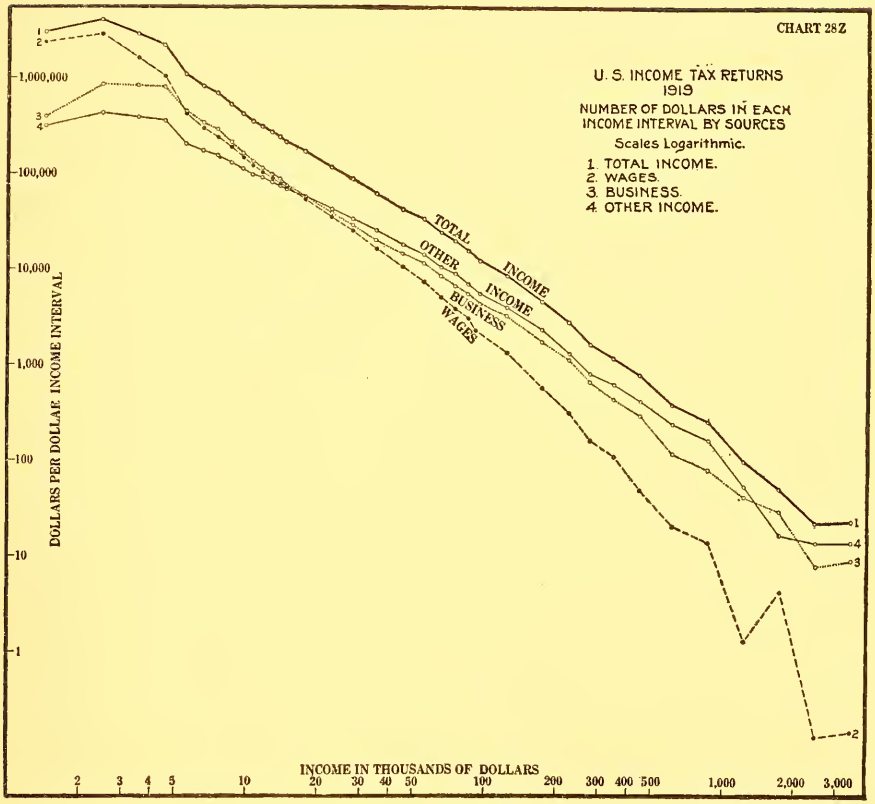


CHART 28Z
U. S. INCOME TAX RETURNS
1918
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES
Scales Logarithmic.
1. TOTAL INCOME.
2. WAGES.
3. BUSINESS.
4. OTHER INCOME.
TOTAL INCOME
OTHER INCOME
BUSINESS
WAGES
DOLLARS PER DOLLAR INCOME INTERVAL
1,000,000
100,000
10,000
1,000
100
10
1
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000

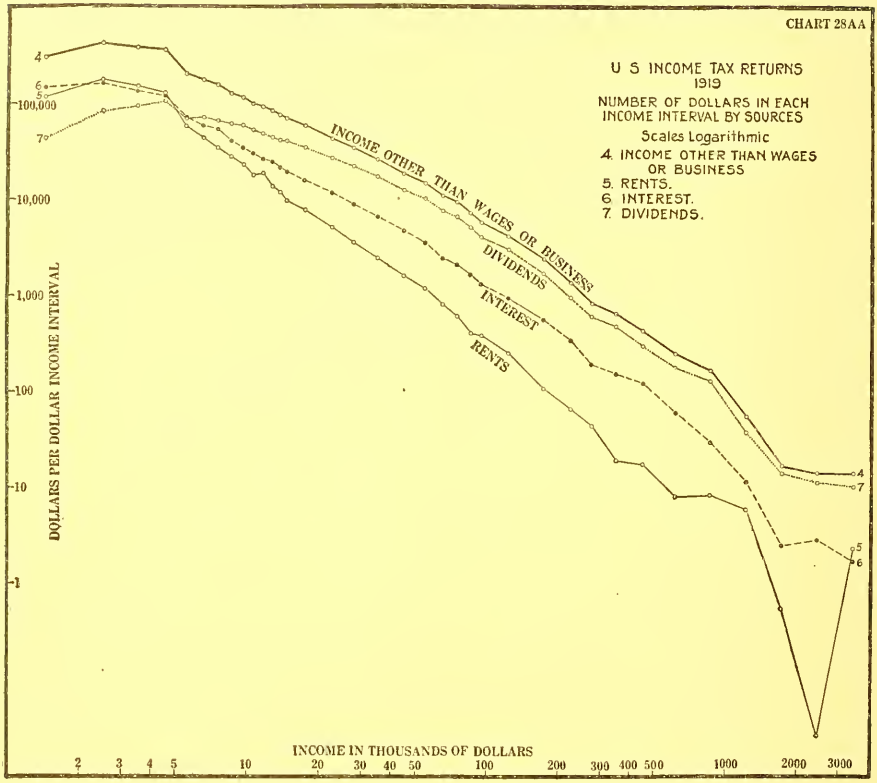


CHART 28AA
U S INCOME TAX RETURNS
1919
NUMBER OF DOLLARS IN EACH
INCOME INTERVAL BY SOURCES
Scales Logarithmic
4. INCOME OTHER THAN WAGES OR BUSINESS
5. RENTS.
6. INTEREST.
7. DIVIDENDS.
INCOME OTHER THAN WAGES OR BUSINESS
DIVIDENDS
INTEREST
RENTS
DOLLARS PER DOLLAR INCOME INTERVAL
100,000
10,000
1,000
100
10
1
INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1000 2000 3000

greater detail the changes in the constitution of the returns from year to year.

Such material and the appearance of the "bulge" on the income-tax curve in the lowest income ranges [1] in the years 1918 and 1919 when wages and salaries were high and average (per capita) incomes also high [2] strongly suggest that the income curve, in so far as it shows any similarity from year to year, changes its general appearance and turns up (on a double log scale) as it approaches those ranges where wages and salaries are of predominant influence.[3] The great *slopes* of wage distributions are on this hypothesis not inconsistent with the smaller *slope* of the general income curve in its higher (income-tax) ranges.[4]

Conclusions:

(1) Pareto's Law is quite inadequate as a mathematical generalization, for the following reasons:
   (a) The tails of the distributions on a double log scale are not, in a significant degree, linear;
   (b) They could be much more nearly linear than they are without that condition being especially significant, as so many distributions of various kinds have tails roughly approaching linearity;
   (c) The straight lines fitted to the tails do not show even approximately constant slopes from year to year or between country and country;
   (d) The tails are not only not straight lines of constant slope but are not of the same shape from year to year or between country and country.

(2) It seems unlikely that any useful mathematical law describing the entire distribution can ever be formulated, because:
   (a) Changes in the shape of the income curve from year to year seem traceable in considerable measure to the evident heterogeneity of the data;
   (b) Because of such heterogeneity it seems useless to attempt to

[1] See Chapter 30 for further discussion of this "bulge" in connection with an examination of how far it may be the result of irregularity in reporting.

[2] Average (per capita) incomes being high means that a definite money income (such as $2,000) takes us relatively further down the income curve than if average incomes were low.

[3] It is difficult to say just where the "bulge" might have appeared in the 1917 distribution if as great efforts had been made to obtain correct returns in that year as were made under the "intensive drive" for 1918 returns. The *wages* line on the 1917 number of dollars income per dollar-income interval chart (Chart 28V) shows signs of turning up somewhere between $4,000 and $5,000 and the *business* line somewhere in the $5,000–$10,000 interval. However neither movement is large nor can their positions be accurately determined on account of the size of the reporting intervals. See also Chapter 30, p. 412.

[4] The "bulge" on the income from wages and salaries curve itself, as seen in the income-tax returns for 1918 and 1919 (see Charts 28X and 28Z), seems the result of heterogeneity in these wage and salary data themselves. This hypothesis is considered in Chapter 30.

describe the whole distribution by any mathematical curve designed to describe homogeneous distributions (as any *simple* mathematical expression must almost necessarily be designed to do);

(c) Furthermore, the existing data are not adequate to break up the income curve into its constituent elements;

(d) If the data were complete and adequate we might still remain in our present position of knowing next to nothing of the nature of any "laws" describing the elements.[1]

(3) Pareto's conclusion that economic welfare can be increased only through increased production is based upon erroneous premises. The income curve is not constant in shape. The internal movements of its elements strongly suggest the possibility of important changes in distribution. The radically different mortality curves for Roman Egypt and modern England,[2] and the decrease in infant mortality in the last fifty years illustrate well what may happen to heterogeneous distributions.

The next four chapters review the data from which any income frequency distribution for the United States must be constructed.

[1] Though all the evidence points to hope of further progress lying in the analysis of the parts rather than in any direct attack upon the unbroken heterogeneous whole.

[2] See *Biometrika*, Vol. I, pp. 261–264.

CHAPTER 29

OFFICIAL INCOME CENSUSES

There has never been a complete income census of the American people. The Federal income-tax data cannot take the place of such a census. Respecting the distribution of income among persons having incomes of less than $1,000 Federal income-tax data give us no information whatsoever. Furthermore, on account of the exemption of married persons, comparatively little use can be made of the $1,000 to $2,000 interval. The number of persons reporting incomes over $2,000 in our best year, 1918, was only 7.3 per cent of the estimated total number of income-recipients in the country. Moreover, not only because of direct evasion and illegal non-reporting, but also because of "legal evasion" and the large amount of tax-exempt income which need not be reported at all, these income-tax data cannot give an approximately correct picture of even that part of the frequency curve which lies above $2,000. The adjustments of the income-tax data necessary to obtain such a picture are extremely large, as we shall presently see.

Only one country in the world has ever taken an official income census which made any pretense of completeness. Under the War Census Act the Commonwealth of Australia took an official income census of incomes received during the year ended June 30, 1915, by everyone, man, woman, or child, who was "possessed of property, or in receipt of income." [1] The results of that census are summarized by G. H. Knibbs, the Commonwealth Statistician, in *The Private Wealth of Australia and its Growth. A Report of the War Census of 1915.* (See Table 29A and Charts 29A, 29B and 29C.)

Now while it would naturally be impossible to construct a complete frequency distribution for American incomes from Australian data,[2] we might perhaps hope to discover some characteristics of income-distribution

[1] While the first clause of the Australian "Wealth and Income Card" stated merely that it was "to be filled in by all persons aged 18 or upwards possessed of property, or holding property on trust, or in receipt of income," etc. (p. 9), "a special instruction was issued that in the case of all persons under the age of 18, possessed of property, or in receipt of income, a return must be furnished by the parent or guardian in respect of such property or income." (p. 10.) The income from such trust funds was not all, but only "in the main," allocated to individual beneficiaries. (p. 22.)

G. H. Knibbs, *The Private Wealth of Australia and its Growth. A Report of the War Census of 1915.*

[2] Aside from the questionableness of such a procedure, the large size of the low income intervals in the Australian distribution and the lack of information concerning the amount of negative income make that distribution a difficult one to work with. A classification by such large intervals tells very little.

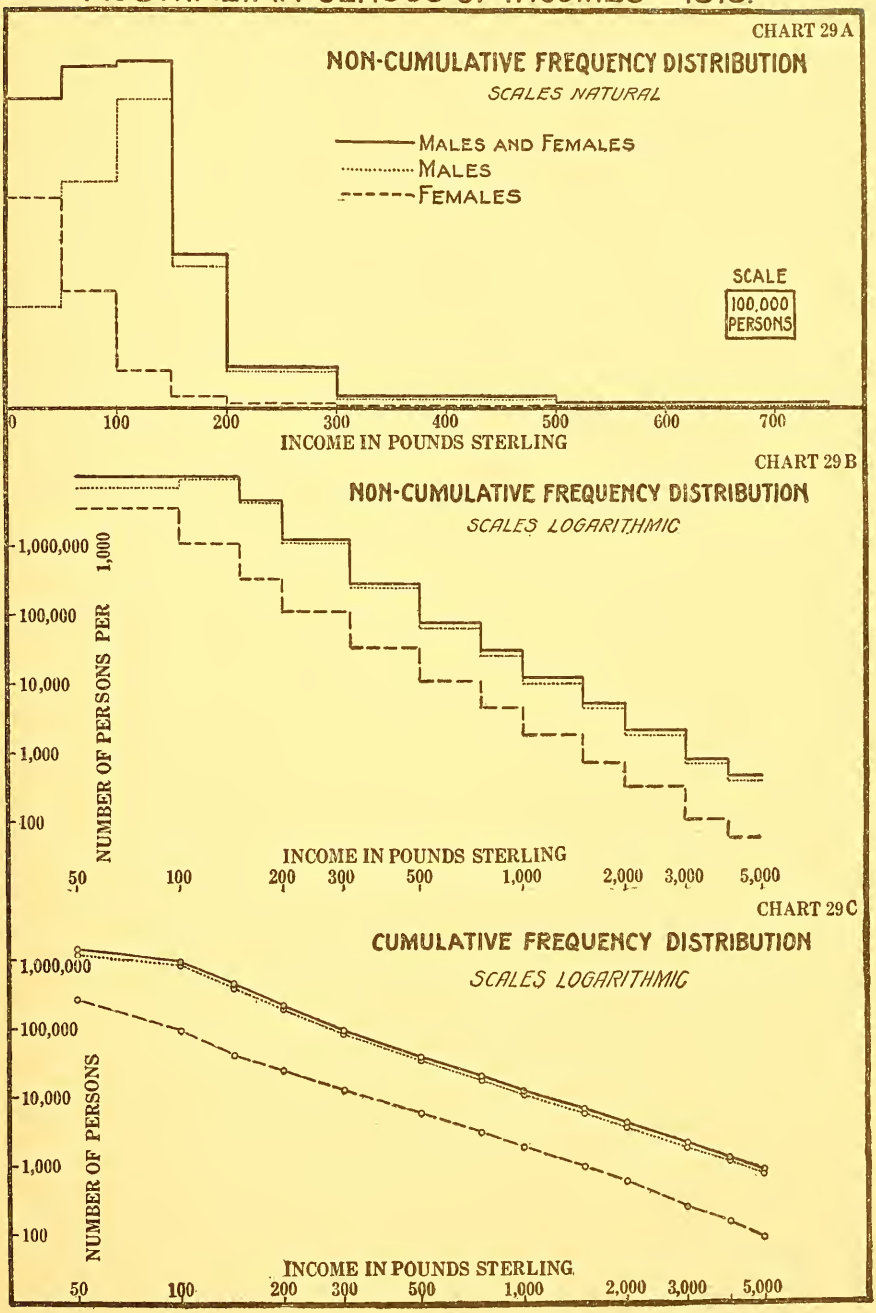
TABLE 29A

AUSTRALIAN WAR CENSUS OF INCOMES
NET INCOME FOR TWELVE MONTHS ENDED JUNE 30, 1915

Income class	Males			Females			Total persons		
	Number	Amount of income (Nearest thousand pounds)	Average income [a] (Pounds)	Number	Amount of income (Nearest thousand pounds)	Average income [a] (Pounds)	Number	Amount of income (Nearest thousand pounds)	Average income [a] (Pounds)
Deficit and nil	66,460			249,476			315,936		
Under £50	145,513	4,163	29	301,592	6,717	22	447,105	10,880	24
£ 50 and under £100	327,835	24,308	74	168,106	11,416	68	495,941	35,725	72
100 " " 150	448,195	55,090	123	52,929	6,250	118	501,124	61,340	122
150 " " 156	46,630	7,093	152	3,651	558	153	50,281	7,651	152
156 " " 200	157,350	27,219	173	12,697	2,211	174	170,047	29,431	173
200 " " 300	106,324	25,191	237	11,001	2,641	240	117,325	27,832	237
300 " " 500	49,108	18,388	374	6,617	2,498	378	55,725	20,887	375
500 " " 750	15,928	9,603	603	2,691	1,633	607	18,619	11,236	603
750 " " 1,000	6,313	5,393	854	1,145	970	847	7,458	6,363	853
1,000 " " 1,500	4,933	5,994	1,215	905	1,089	1,204	5,838	7,083	1,213
1,500 " " 2,000	2,132	3,676	1,724	364	629	1,729	2,496	4,306	1,725
2,000 " " 3,000	1,707	4,149	2,431	317	772	2,434	2,024	4,921	2,431
3,000 " " 4,000	659	2,249	3,412	102	361	3,537	761	2,610	3,429
4,000 " " 5,000	375	1,685	4,494	58	258	4,455	433	1,944	4,489
5,000 " over	746	7,300	9,786	86	656	7,627	832	7,956	9,563
Total	1,380,208	201,503	146	811,737	38,661	48	2,191,945	240,163	110

[a] The above averages are not always consistent with the other figures of the table. They have evidently not been calculated from the *approximate* figures given above for amount of income.

AUSTRALIAN CENSUS OF INCOMES - 1915.



curves in general from this, the only actual census ever taken. A knowledge of such general characteristics might then, quite imaginably, be a little helpful in the problem of describing the American or any other income distribution.

However, when we come to examine the Australian figures, we find that they have certain pronounced peculiarities which would be extremely difficult to read into the American material. For example, the Australian distribution shows a flatness and lack of pronounced mode totally unlike the results we have built up from an analysis of American data. In the Australian distribution there are nearly the same number of persons having incomes between 0 and £50, £50 and £100, and £100 and £150.[1]

What are the causes of this rather startling peculiarity of the Australian frequency curve? [2] In the first place let us suggest a possibly minor but by no means necessarily negligible factor. We know little about the goodness of the Australian reporting in this census. Income is, from its nature, a difficult subject to investigate. When the material is collected by means of schedules to be filled in by the informants, as was the case in the Australian census, the returns may easily be full of errors. The average individual is surprisingly ignorant concerning the amount of his total income. The further fact that the census was taken in order to estimate possibilities of future taxation may well have been a powerful incentive towards great irregularities all along the line, but especially in the lower income groups. Persons whose income brought them distinctly into the upper groups (over £156) were, at the time of the income census, about to make returns under oath for income-tax purposes and would hardly care to show a radical discrepancy between the two returns. On the other hand, many persons, whose true incomes were around £156 and the modal income, might easily have "underestimated" with the idea of evading if possible future taxation based upon a lowering of the exemption limit. The result of such practices would tend to show up graphically in a flattening of the curve in the vicinity of the mode of the distribution and a raising of the numbers in the lowest groups. [3]

However, poor reporting is probably only a secondary element accounting for the peculiarities of the Australian curve. It is most of all the

[1] See Table 29A and Chart 29A.

[2] Notwithstanding the fact that distributions for different times and for different countries probably vary greatly (see Chapter 28), the difference between the Australian curve and the Bureau's American estimate seems too radical to explain upon this basis.

[3] It is difficult to determine the extent of actual non-reporting. The number of males filling out income cards was 2,527,831. All males "possessed of property, or in receipt of income" are supposed to be included in this number. It amounted, however, to only 54.60 per cent of the total male population. Males "possessed of property, or in receipt of income" necessarily constitute a larger percentage of the total male population than do male "breadwinners," yet in the Australian census of 1911 male breadwinners constituted 69.4 per cent of the total male population, and male breadwinners 20 years of age or older 58.9 per cent. Even if we assume that the number of income returns for males under 18 was negligible we still are faced with a discrepancy difficult to account for.

concentration of female returns in the lowest income groups which gives the flat and modeless appearance to the total curve. The Australian frequency distribution among males only, is much more like our estimated American distribution [1] than is the Australian distribution among males and females together. Now the concentration of female returns in the lower income intervals would seem to be the result of a large number of returns made by women and female children receiving petty incomes from property who would be classified, in the Australian Census of Population, as "dependents" and not as "breadwinners." [2]

Of the total female population in 1915, 33.46 per cent made out income cards and 23.18 per cent reported positive incomes (10.28 per cent reported zero or negative incomes). But according to the Australian census of 1911, only 18.6 per cent of the total female population were classified as "breadwinners." Thus the women reporting positive incomes in 1915 constituted a much larger percentage of the total female population than did female "breadwinners" in 1911 of the total female population in that year. The discrepancy seems too great to be accounted for by the increase in the number of women "breadwinners" caused by the war. More than half of the 23.18 per cent of the female population reporting positive incomes in 1915 reported incomes under £50 per annum. Moreover, the average income of this group was only £22 per annum—under the arithmetic average of the interval. This strongly suggests petty incomes from property, and part time occupations such as keeping boarders, lodgers, chickens, etc., rather than any great increase in the number of female "breadwinners." The fact that over 30 per cent of the returns made by females reported zero or negative incomes is further evidence that the large number of extremely small incomes reported was largely the result of the schedule calling for income returns from all persons "possessed of property."

Negative incomes arise in general from business or speculative losses. Bad as may be the condition of any laboring class, its members are seldom faced with negative incomes. It is unlikely that many of the 249,476 females reporting "deficit and nil" were wage-earners. They were in general the owners of small investments which showed losses, such as town lots upon which taxes had been paid.[3]

[1] See *Income in the United States*, Vol. I, pp. 128, 129, 132–135.

[2] All persons are classified as "breadwinners" or as "dependents" by the Australian census. Male "breadwinners" in Australia constituted in 1911, according to the census of that year, 69.4 per cent of the total male population, female "breadwinners" 18.6 per cent of the total female population, and total "breadwinners" 45.0 per cent of the total population. These figures compare with American census figures for 1910 showing males "gainfully employed" to constitute 63.6 per cent of total males, females "gainfully employed" 18.1 per cent of total females, and total "gainfully employed" 41.5 per cent of the total population.

[3] It is worth noting that in the Australian schedule "rates and taxes paid" could be deducted before making an income return. This consideration may be of some importance in explaining the very large number of small, zero, and negative incomes.

While the frequency curve for Australian males is much more like the American distribution than the curve representing both male and female Australian income recipients, even it shows a much greater concentration in the lowest income intervals than does the American distribution. This can probably be accounted for to some extent by a large number of income returns for young male "dependents" "possessed of property."

The essential difference in appearance between the American income-distribution curve which we presented in Volume I and the Australian curve of 1915 is, then, probably traceable to (1) Australian underreporting and (2) Australian inclusion of a large number of "dependents" who received petty incomes from property and who were in no important sense "breadwinners" or "gainfully employed."

What shall we say about the desirability or undesirability of including in an income frequency distribution dependents receiving petty incomes from property? While it is true that their incomes, positive or negative, are in a way as real as any other incomes, we must remember that probably almost all individuals over six years of age not only receive but *earn* some money income during each year. Shall we then include the entire population over six years old in our distribution? As we approach this theoretical limit it is seen that the concept becomes less and less practically or even theoretically interesting. Both practically and theoretically we are interested in the incomes of persons who, though they be minors, have "economically come of age" and have entered into certain definite relations to the machinery of factorial distribution. They are "breadwinners" or "persons gainfully employed," and the concept back of such expressions, though like many economic concepts somewhat of a compromise, seems a good compromise for our purposes.

Defining income recipient as we have, we cannot use the Australian material as an aid to the graduation or adjustment of the American income-distribution curve in its lower ranges. In the upper income ranges, the Australian distribution offers, as we shall see, an interesting illustration of the same double swing (letter S) appearance of the curve seen in some of the more recent American data.[1]

[1] When charted on a double log scale.

CHAPTER 30

AMERICAN INCOME TAX RETURNS

At the beginning of the preceding chapter attention was drawn to some reasons why income-tax returns cannot take the place of an adequate income census. Nevertheless tax returns are in many respects the most important single source of information we have for estimating the frequency distribution of incomes. Were there neither tax returns nor income censuses for any country, it is difficult to see how we could make even an interesting guess as to the distribution of income in the upper ranges.

American income-tax data go back to 1913. We have now at our disposal returns for the seven years, 1913 to 1919, inclusive.[1] However, the amount of information given in the official reports for the earlier years 1913, 1914 and 1915 is not great. Little is shown beyond the number of returns classified by large income intervals and the same returns classified by districts. The 1916 tax report is the most voluminous and in one respect the most adequate report which has yet appeared.[2] It contains a set of tables which we are sorry to miss in the later reports, showing the frequency distribution of incomes by separate occupations. Other features of this report which have been retained in later years are tables showing both number of returns and amount of net income for each income class for the country as a whole, and the same by States; tables showing the sources of the income returned in each income interval, that is the amount from wages, business, property; distribution tables arranged by sex and conjugal condition; amounts of tax collected from each income class, etc.

Changes in the Federal Income Tax Law during the period have not been such as greatly to affect any conclusions which we have drawn from the data. From the standpoint of this investigation, probably the most important changes in the law relate to *general deductions, professions,* and *minimum taxable income.*

In the 1916 returns all deductions were classified as *general deductions.*

[1] The *Annual Reports of the Commissioner of Internal Revenue* are the sources for American income-tax data for the years 1913 to 1915. Since 1915 the data have appeared annually as a separate Treasury Department publication entitled *Statistics of Income.*

[2] A peculiarity of the 1916 data is that the returns are tabulated as family rather than individual returns. "The net incomes reported on separate returns made by husband and wife in 1916 are combined and included as one return in the figures for the several classes." *Statistics of Income,* 1917, p. 22.

In the 1917 returns the types of deductions classified as *general deductions* were greatly reduced; not even *contributions* were included. In 1918 the category was enlarged; *contributions*, for example, were again placed in the *general deductions* class. Now these changes affect greatly the relations between *net* and *total* income from year to year. Reported *net* income was in 1916 only 75.43 per cent of reported *total* income, in 1917 it was 92.67 per cent, in 1918 89.74 per cent, and in 1919 88.51 per cent. As it is the *total* and not the *net* income which in the *Statistics of Income*, is divided up according to source, such fluctuations as the above interfere with comparisons of different years.

While income from *professions* was tabulated separately in 1916, in 1917 it was included in *wages and salaries*, and in 1918 and 1919 in *business*.

In the 1913 to 1916 returns exemptions were $3,000 per annum for an unmarried person, or a married person not living with his wife (or her husband), and $4,000 per annum aggregate exemption for married persons living together.[1] In the 1917 and later returns these minima were reduced to $1,000 and $2,000 respectively. However, the increase in usefulness for our purposes of the 1917 and later returns was even greater than the lowered minima would suggest. Not only was the *minimum taxable income* lowered from $3,000 to $1,000, but this reduction occurred in the face of a rapidly rising general level of incomes. With the rise in incomes, $3,000 in 1918 or 1919 was relatively a much smaller income than $3,000 in 1913. In other words, we might logically expect $3,000 to be relatively further down the income distribution curve in 1918 than in 1916 or 1917.

The accuracy of the reporting is, of course, a matter of great importance for this investigation. Now, while it does not seem possible to measure directly from the data changes in accuracy of reporting during the period, the rapid expansion of the income-tax organization and its increasing attention to the investigation and checking of returns establish the presumption of greater statistical value in the reports for the later years. Offsetting this to an unknown degree is the apparently increasing amount of "legal evasion" in the higher income classes. The reporting for the years 1913, 1914, 1915 and 1916 appears to have been peculiarly bad in the lower income ranges. The distinct improvement in 1917 (compare the 1917 returns with those for earlier years in Tables 28B, 28C, 28D, 28E, and Charts 27 and 28 of Volume I) seems associated with the patriotic enthusiasm engendered by the war. Upon our entry into the war, not only did the Bureau of Internal Revenue make an increased effort to ob-

[1] As the returns for 1913 were for income received for the *ten months* March 1 to December 31, 1913, the actual minima used for reporting purposes were $2,500 and $3,333.33 (i. e., $\frac{10}{12}$ of $3,000 and $4,000 respectively).

tain correct returns but individuals, under the spur of patriotism, seem to have made less effort to evade.[1]

The remainder of this chapter is concerned largely with a discussion of possible irregularities in the *distribution* of non-reporting and understatement in the later years. While the total amount of non-reporting and understatement was almost certainly greater in the returns for 1917 than in those for 1918 and 1919, are we sure that the non-reporting and understatement of these later years are not possibly more irregularly distributed along the frequency curve than was the case in 1917? Is it possible that the improvement in the accuracy of the published returns for 1918, as compared with those for 1917, was so much greater in the income intervals under $5,000 that the resulting change in the shape of the frequency curve may amount to something almost akin to an "over-adjustment"?

Income returns by individuals are made on two types of blanks, a blank to be filled in by persons reporting incomes under $5,000 and another blank to be filled in by persons reporting incomes over that figure. Now, while the returns of incomes under $5,000 and made on "under $5,000" blanks are examined, investigated and audited in the field soon after their receipt, the investigation and audit of the returns for incomes over $5,000 are handled in Washington. If an individual has an actual income of $8,000 but reports $4,600 (on an "under $5,000" blank), as soon as a Field Collector discovers this discrepancy, he passes the matter over to the Revenue Agent in charge of the District for Field Investigation. The return, accompanied by the Agent's report, is forwarded to Washington for final audit. Thus the Field Collectors audit only returns that are (a) made on "under $5,000" blanks and (b) believed, *after investigation*, to be for incomes which are *actually* under $5,000.

While the Field Audit of returns of these incomes is well under way before the preparation of the statistical tables in the *Statistics of Income* and hence appears in that tabulation to an unknown extent, the Washington audit of incomes over $5,000 has hardly begun and hence the amended figures for these higher incomes do not appear in the *Statistics of Income.* It is impossible to say exactly how much of the "bulge" [2] which appears in the $1,000 to $5,000 interval on the double log charts of the 1918 and 1919 tax income distributions is caused by a difference in the accuracy of the published figures for returns of incomes under and over $5,000. However, the Treasury Department states that "the *Statistics of Income*

[1] It must not, of course, be assumed that the increase in the number of returns in 1917 is traceable solely to increased goodness of reporting.

[2] Described in Chapter 28. At many points in the following discussion the reader should refer back to the presentation of the case for heterogeneity in the income-tax data contained in Chapter 28.

are compiled almost entirely from unaudited returns whether they be for 'under $5,000' or 'over $5,000.'" It seems probable therefore that the sudden change in slope of the 1918 curve (on a double log scale) at about $5,000 can be explained only partially by a change in accuracy of the published returns at that point.

Moreover, a considerable amount of evidence, some of which has already been presented in Chapter 28, suggests that the "bulge" on the income curves for the later years corresponds to a reality on the actual income curves. While it may be somewhat over-accented in the published figures for 1918 and 1919, and while the figures for 1917 might have shown more of such a "bulge" [1] had the reporting been better, we must not assume that the published figures for either 1917 or 1918 give a radically incorrect picture of the facts merely because the income curves for the two years are so different. The dogma of the similarity of the income curve from year to year has little evidence to support it.

It is by no means certain that even the apparently definite and sharp angles on the curves in this $4,000 to $6,000 region give an unreal picture. While it is true that we find the same angles on the wages and salaries curve, that curve itself seems heterogeneous. An income distribution curve composed of wage and salary earners (in the ordinary sense of the terms) may well cut an income distribution curve composed of "salaried entrepreneurs," and business and financial experts somewhere in the lower income ranges. The angle on the composite curve may give a decidedly accurate picture of the facts.[2]

Let us see what light the data throw on some of these problems. Table 30A showing the number of returns for the lower income intervals in 1917, 1918, and 1919 and the percentage movements from year to year illustrates the great increase in the number of returns in the under-$5,000 intervals between 1917 and the later years.

Chart No. 28 of Volume I, on which are drawn the frequency distributions for each year from 1916 to 1919 on a double log scale, shows the difference in the appearance of the income curves for the three years. Examining that chart we notice that the 1918 data-points, which in the upper income ranges run nearly as smoothly as the 1917 points, in the $4,000 to $5,000 interval move abruptly upwards and from there on into the lowest income ranges are well above the 1917 points, showing on the chart an irregular, plateau-like effect in these lowest income ranges. No such "plateau" is apparent on the 1917 line. The year 1919 presents in that chart a

[1] While the 1917 curve runs much more smoothly in the $3,000 to $6,000 range than either the 1918 or 1919 curves, it is not without the hint of a bulge beginning at about $4,500. See p. 412.

[2] In constructing the complete income distribution curve for 1918, published in Volume I, the influence of changes in the accuracy of reporting around $5,000 income was probably overestimated.

TABLE 30A

Income intervals	Number of returns			Percentage increases		
	1917	1918	1919	1918 over 1917	1919 over 1918	1919 over 1917
$2,000–$3,000.....	838,707	1,496,878	1,569,741	78.47	4.87	87.16
3,000– 4,000.....	374,958	610,095	742,334	62.71	21.68	97.98
4,000– 5,000.....	185,805	322,241	438,154	73.43	35.97	135.81
5,000– 6,000.....	105,988	126,554	167,005	19.40	31.96	57.57
6,000– 7,000.....	64,010	79,152	109,674	23.66	38.56	71.34
7,000– 8,000.....	44,363	51,381	73,719	15.82	43.48	66.17
8,000– 9,000.....	31,769	35,117	50,486	10.54	43.77	58.92
9,000–10,000.....	24,536	27,152	37,967	10.66	39.83	54.74

similar appearance to 1918 though the absence of small intervals in the range immediately above $5,000 disguises the characteristics of the curve materially.[1]

The change in the contour of the lower range of the tax income frequency curve from 1917 to 1918 and 1919, is, as we have mentioned, associated with a large increase in the relative amount of income from wages and salaries in the lower intervals. Tables 30B and 30C are interesting in this connection.[2]

The 1916 figures in Table 30B are introduced simply because they are computable.[3] However, too much weight must not be attached to them. The 1916 returns are undoubtedly extremely inadequate. The high percentages that year from $3,000 income (the 1916 minimum) up to about $10,000 may possibly be the result of the ease with which *salary* returns (as opposed to *wage, business,* or other returns) are obtainable. The $4,000 to $5,000 interval is the lowest comparable interval for the four years.[4] In that interval the numbers of returns by years were:

1916– 72,027
1917–185,805
1918–322,241
1919–438,154

[1] When chart 28 was drawn for Volume I, only "preliminary" large interval data were available. Final small interval data show a "bulge" very similar to that seen in the 1918 line.

[2] The 1917 official *wages* figures include income from professions. The 1918 and 1919 *wages* figures do not. This makes the increase in the percentages in 1918 still more striking. Income from professions was tabulated separately in 1916, but was included in the *wages* figures for that year in order that 1916 and 1917 might be comparable.

[3] No data are available from which corresponding figures for 1913, 1914 or 1915 might be calculated.

[4] The $3,000–$4,000 interval did not in 1916, include married persons making a joint return.

TABLE 30B

PER CENT THAT INCOME FROM WAGES AND SALARIES IN EACH NET INCOME CLASS WAS OF TOTAL *NET* INCOME IN THAT CLASS

Income class	1916	1917	1918	1919
$ 1,000–$ 2,000			79.45	83.49
2,000– 3,000			69.75	74.53
3,000– 4,000	76.98		55.21	61.86
2,000– 4,000		46.32	(64.42)	(69.45)
4,000– 5,000	66.86	36.30	48.85	52.48
5,000– 10,000	53.31	36.16	39.59	43.24
10,000– 20,000	36.38	32.94	38.60	38.11
20,000– 40,000	24.60	26.82	33.16	33.38
40,000– 60,000	17.23	22.74	27.88	27.57
60,000– 80,000	16.20	19.67	25.36	24.01
80,000– 100,000	13.37	18.51	22.16	22.70
100,000– 150,000	13.34	15.75	18.44	18.75
150,000– 200,000	9.39	12.65	16.16	15.42
200,000– 250,000	9.14	12.30	13.07	13.62
250,000– 300,000	7.87	9.36	12.57	11.92
300,000– 500,000	6.59	10.17	11.27	10.18
500,000–1,000,000	5.21	6.39	5.42	6.80
1,000,000–1,500,000	4.84	2.83	7.54	1.60
1,500,000–2,000,000	3.23	3.76	2.21	10.00
2,000,000 and over	.51	2.39	.85	4.02

The amounts of income from wages and salaries and from other net income in the $4,000–$5,000 interval were year by year in millions of dollars:

	1916	1917	1918	1919
Wages and salaries [a]	216	301	703	1,029
Other net income	107	528	736	931

[a] Income from professions is included in the 1916 and 1917 wages and salaries figures.

The percentage changes in these items from one year to the next were:

	1917/1916	1918/1917	1919/1918
Wages and salaries	139.3	233.7	146.4
Other Net Income	493.0	139.4	126.6

It is plain that the great increase in the $4,000–$5,000 interval [1] in 1917 was in income from other sources than wages and salaries.

Table 30C shows the wage and salary figures compared with *total* income instead of *net* income as in Table 30B. It was, of course, necessary to retain the *net* income intervals as the data are not classified in *total* income

[1] As may be seen from Tables 30B and 30C, the increase from 1916 to 1917 in income from other sources than wages and salaries was greater than the increase in income from wages and salaries not only in the $4,000–$5,000 interval but also in the $5,000–$10,000 interval.

intervals. Though the relations between years are different in this table from what they are in the net income table,[1] the distribution of the percentages in each individual year shows much the same characteristics in both tables.

TABLE 30C

PER CENT THAT INCOME FROM WAGES AND SALARIES IN EACH NET INCOME CLASS WAS OF *TOTAL* INCOME IN THAT CLASS

Income class (Net)	1916	1917	1918	1919
$ 1,000– $2,000.....			74.67	77.25
2,000– 3,000.....			65.42	69.14
3,000– 4,000.....	47.74		51.14	56.71
2,000– 4,000.....		41.82	(60.15)	(64.12)
4,000– 5,000.....	45.96	33.60	44.82	47.12
5,000– 10,000.....	36.38	33.87	33.55	36.60
10,000– 20,000.....	25.76	30.89	33.10	32.70
20,000– 40,000.....	18.81	25.20	28.76	28.36
40,000– 60,000.....	13.75	21.23	23.79	23.39
60,000– 80,000.....	12.76	18.56	21.51	20.33
80,000– 100,000.....	10.74	17.61	19.00	19.25
100,000– 150,000.....	11.06	15.05	15.92	15.40
150,000– 200,000.....	7.68	12.01	13.10	12.41
200,000– 250,000.....	7.83	11.75	11.22	11.26
250,000– 300,000.....	6.64	8.71	10.73	9.80
300,000– 500,000.....	5.50	9.59	9.62	8.19
500,000–1,000,000.....	4.35	5.88	4.37	5.38
1,000,000–1,500,000.....	4.12	2.62	6.29	1.34
1,500,000–2,000,000.....	2.82	3.54	1.81	8.54
2,000,000 and over	.47	2.18	.63	.32

The percentages in Tables 30B and 30C show each year a sudden increase (as we approach the lower income intervals) somewhere in the $4,000 to $5,000 or the $5,000 to $10,000 interval. At *exactly* what point each year do these sudden increases seem to occur? Charts 30D, 30E and 30F present the material in a slightly different form. They illustrate the relationship between the average income from wages and salaries in each net income interval and the average total income in the same net income interval for the years 1917, 1918 and 1919 on a double log scale. The 1918 and 1919 charts immediately suggest the improbability of being able to describe the data by a single simple mathematical expression. To the 1918 data-points have been applied two distinct mathematical curves, which fit the data remarkably well and intersect at about $6,700 total income. The curve fitted to the upper income ranges is a parabola, while that fitted to the lower income ranges is an hyperbola, one of whose asymptotes is the 45° line which divides the chart into a "possible" and an "im-

[1] Some reasons for the changes in relation of *net* to *total* income from year to year are mentioned on pages 401 and 402.

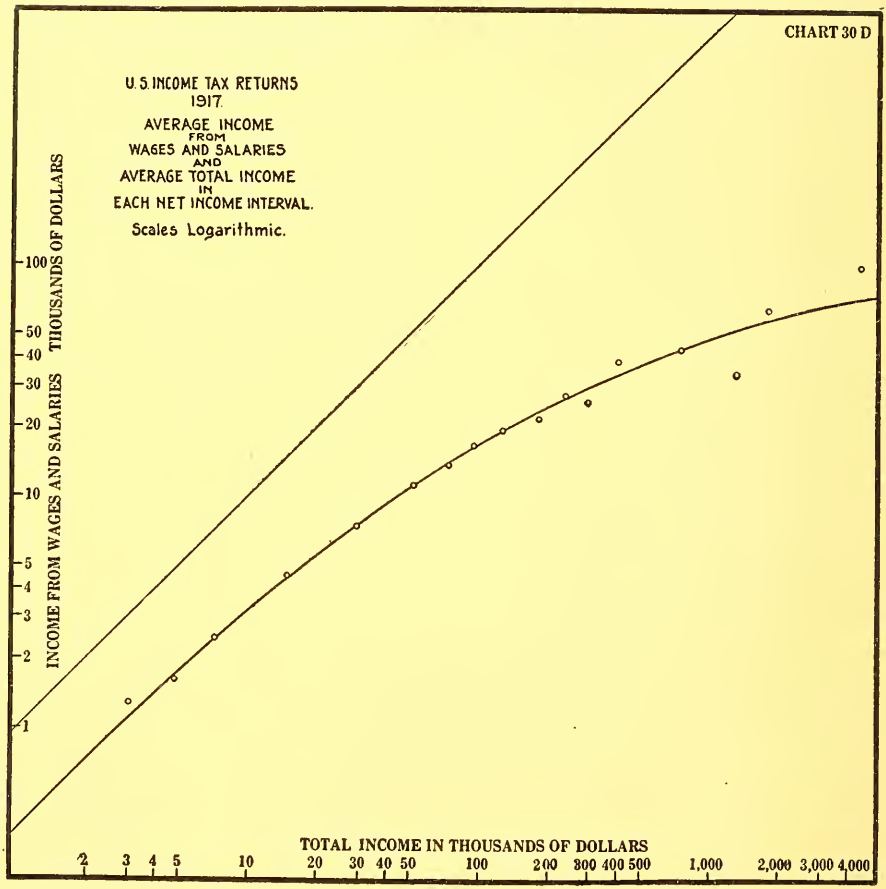


CHART 30 D
U.S. INCOME TAX RETURNS
1917.
AVERAGE INCOME
FROM
WAGES AND SALARIES
AND
AVERAGE TOTAL INCOME
IN
EACH NET INCOME INTERVAL.
Scales Logarithmic.
INCOME FROM WAGES AND SALARIES THOUSANDS OF DOLLARS
100
50
40
30
20
10
5
4
3
2
1
TOTAL INCOME IN THOUSANDS OF DOLLARS
2
3
4
5
10
20
30
40
50
100
200
300
400
500
1,000
2,000
3,000
4,000

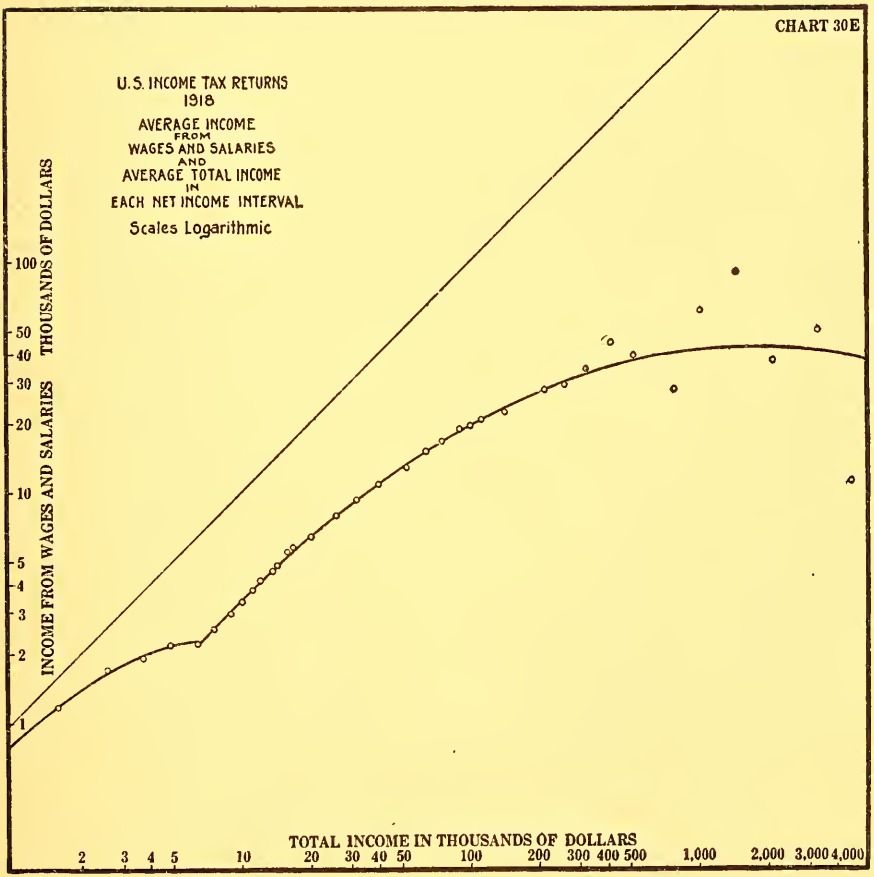


CHART 30E
U.S. INCOME TAX RETURNS
1918
AVERAGE INCOME
FROM
WAGES AND SALARIES
AND
AVERAGE TOTAL INCOME
IN
EACH NET INCOME INTERVAL
Scales Logarithmic
INCOME FROM WAGES AND SALARIES THOUSANDS OF DOLLARS
100
50
40
30
20
10
5
4
3
2
1
TOTAL INCOME IN THOUSANDS OF DOLLARS
2
3
4
5
10
20
30
40
50
100
200
300
400
500
1,000
2,000
3,000
4,000

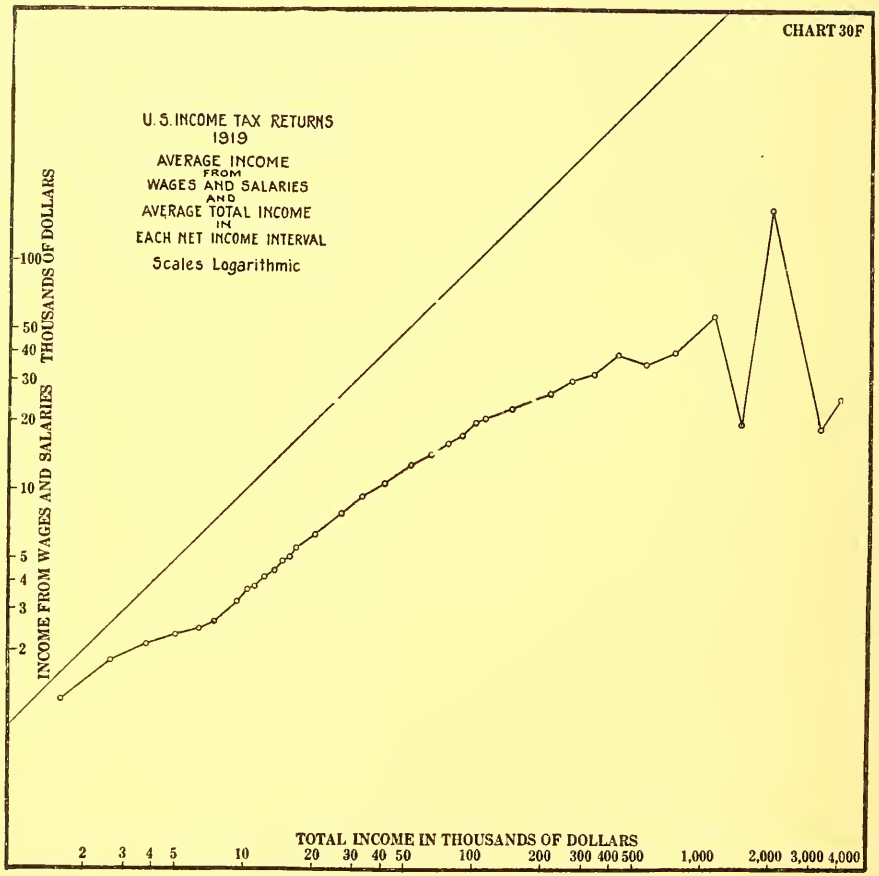


CHART 30F
U.S. INCOME TAX RETURNS
1919
AVERAGE INCOME
FROM
WAGES AND SALARIES
AND
AVERAGE TOTAL INCOME
IN
EACH NET INCOME INTERVAL
Scales Logarithmic
INCOME FROM WAGES AND SALARIES THOUSANDS OF DOLLARS
100
50
40
30
20
10
5
4
3
2
TOTAL INCOME IN THOUSANDS OF DOLLARS
2 3 4 5 10 20 30 40 50 100 200 300 400 500 1,000 2,000 3,000 4,000

possible" area. The equations of the two (1918) curves on a double log scale are (I) $y + 3.92945 - 2.744\,x + .22\,x^2 = 0$ (parabola)

(II) $y^2 - 3.981909\,y - .867246\,xy + 3.981909\,x - .132754\,x^2 - .060262 = 0$ (hyperbola)

As it is difficult to estimate accurately by eye the goodness of fit of a curve to data when charted on a log scale, Table 30E is introduced:

TABLE 30E

WAGES AND INCOME IN THE 1918 INCOME TAX RETURNS

Net income intervals (1918)	Average total income	Average income from wages and salaries		Percentages that data are of mathematical curves
		Data	Mathematical curves	
$ 1,000–$ 2,000...	$ 1,566	$ 1,169	$ 1,178	99.2
2,000– 3,000...	2,583	1,690	1,652	102.3
3,000– 4,000...	3,710	1,897	1,955	97.0
4,000– 5,000...	4,866	2,181	2,117	103.0
5,000– 6,000...	6,388	2,192	2,216	98.9
6,000– 7,000...	7,620	2,537	2,555	99.3
7,000– 8,000...	8,952	2,963	3,012	98.4
8,000– 9,000...	10,148	3,341	3,407	98.1
9,000– 10,000...	11,214	3,747	3,760	99.7
10,000– 11,000...	12,207	4,171	4,078	102.3
11,000– 12,000...	13,707	4,555	4,542	100.3
12,000– 13,000...	14,263	4,806	4,709	102.1
13,000– 14,000...	15,922	5,529	5,204	106.2
14,000– 15,000...	16,778	5,801	5,455	106.3
15,000– 20,000...	20,167	6,375	6,400	99.6
20,000– 25,000...	25,859	7,891	7,860	100.4
25,000– 30,000...	31,704	9,196	9,211	99.8
30,000– 40,000...	39,644	10,711	10,872	98.5
40,000– 50,000...	52,319	12,639	13,192	95.8
50,000– 60,000...	64,327	14,963	15,066	99.3
60,000– 70,000...	74,848	16,576	16,539	100.2
70,000– 80,000...	90,437	18,764	18,459	101.7
80,000– 90,000...	98,379	19,273	19,351	99.6
90,000– 100,000...	111,515	20,447	20,682	98.9
100,000– 150,000...	139,520	22,212	23,163	95.9
150,000– 200,000...	211,959	27,758	27,829	99.7
200,000– 250,000...	259,487	29,107	30,068	96.8
250,000– 300,000...	317,578	34,076	32,226	105.7
300,000– 400,000...	409,756	44,393	34,786	127.6
400,000– 500,000...	514,882	38,967	36,847	105.8
500,000– 750,000...	765,905	27,582	39,765	69.4
750,000–1,000,000...	1,013,846	61,183	41,229	148.4
1,000,000–1,500,000...	1,426,182	89,710	42,199	212.6
1,500,000–2,000,000...	2,084,715	37,118	42,199	88.0
2,000,000–3,000,000...	3,263,673	50,178	40,729	123.2
3,000,000–4,000,000...	4,515,732	11,013	38,753	28.4

The data of table 30E move rather erratically in the intervals above $300,000 per annum income. This is natural in view of the small number

of cases in these upper intervals. There were only 627 returns reporting net incomes of over $300,000 per annum; this is less than one seventieth of one per cent. of the total number of returns. In the 28 intervals under $300,000 per annum 14 of the percentages show the data within one and one half per cent. of the mathematical values.

These mathematical curves have not been introduced as being in any sense the "law" of the data but merely to emphasize how smoothly the data curves run and yet how unmistakable a sensation they give us of two parts, one above about $6,700 total income and one below that figure.[1] It would, of course, be quite impossible to get any sort of approximation to the lower range data by producing the parabola fitted to the upper income ranges. How impossible may be seen from Table 30EE.

TABLE 30EE

WAGES AND INCOME IN THE 1918 INCOME TAX RETURNS

Net income intervals (1918)	Average total income	Average income from wages and salaries			Percentages that data are of	
		Data	Hyperbola	Parabola	Hyperbola	Parabola
$4,000–$5,000.....	$4,866	$2,181	$2,117	$1,574	103.0	138.6
3,000– 4,000.....	3,710	1,897	1,955	1,152	97.0	164.7
2,000– 3,000.....	2,583	1,690	1,652	745	102.3	226.8
1,000– 2,000.....	1,566	1,169	1,178	391	99.2	299.0

The 1919 data show the same two-curve appearance as the 1918 data. This may be clearly seen from chart 30F.[2] The intersection of the two curves would be at about $7,100 instead of $6,700 as on the 1918 chart. Is there any sign of such a change from one curve to another on the 1917 data? There seems to be. Chart 30D shows the 1917 data with a parabola fitted to the observations above the first interval. This curve and Table 30D give us a strong impression that the first interval cannot be described by any simple curve which describes the remainder of the data. The same two-curve characteristics as the 1918 and 1919 data are strongly suggested.

The equation of the 1917 parabola on a double log scale is $y + 1.8417 - 1.8346\,x + .124\,x^2 = 0$. The poorness of the fit to the first interval and the comparative goodness of the fit to the remainder of the data as high as $250,000 per annum may be seen from Table 30D. If the data were numerous enough to permit us fitting two curves they would probably intersect at about $4,500.

[1] An alteration in the size of the intervals in which the data are quoted by the Income Tax Bureau would of course change the data curve to some extent. However, taking the intervals as they come and fitting the curves to them we get the unmistakable impression of great regularity. It seemed scarcely worth while to fit the curves to areas rather than points.

[2] The story told by Chart 30F is so plain it seemed hardly necessary to fit another set of curves.

TABLE 30D

WAGES AND INCOME IN THE 1917 INCOME TAX RETURNS

Net income intervals (1917)	Average total income	Average income from wages and salaries		Percentages that data are of mathematical curve
		Data	Mathematical curve	
$ 2,000–$ 4,000...	$ 3,059	$1,280	$1,101	116.3
4,000– 5,000...	4,818	1,619	1,688	95.9
5,000– 10,000...	7,210	2,442	2,422	100.8
10,000– 20,000...	14,623	4,517	4,374	103.3
20,000– 40,000...	29,236	7,368	7,411	99.4
40,000– 60,000...	51,940	11,024	11,038	99.9
60,000– 80,000...	72,811	13,516	13,699	98.7
80,000– 100,000...	93,742	16,510	15,992	103.2
100,000– 150,000...	126,979	19,108	19,081	100.1
150,000– 200,000...	181,156	21,758	23,147	94.0
200,000– 250,000...	233,880	27,501	26,388	104.2
250,000– 300,000...	293,905	25,587	29,478	86.8
300,000– 500,000...	398,517	38,204	33,877	112.8
500,000–1,000,000...	740,769	43,558	43,632	99.8
1,000,000–1,500,000...	1,294,619	33,973	52,845	64.3
1,500,000–2,000,000...	1,812,388	64,201	58,358	110.0
2,000,000 and over....	4,551,718	99,132	71,945	137.8

Both the regularity of the data curves and the positions of the intersections of the mathematical curves [1] might suggest that heterogeneity of the wages and salaries data was the primary cause of the irregularity in the total income curve. The position of the points of intersection of the mathematical curves might seem inconsistent with a sudden change in accuracy of reporting at exactly $5,000.

However this argument does not appear so conclusive when we examine the actual amount of wages in each income interval. The constitution of the reported income each year may be seen rather plainly in Charts 28T, 28U, 28V, 28W, 28X, 28Y, 28Z, and 28AA.[2] These charts show the number of dollars per dollar income interval reported in each income interval by sources for the years 1916 to 1919.[3] They not only illustrate the fact that the constitution of the income curve changes radically as we move from small to large incomes but also picture the salient characteristics of these changes; each source curve, being charted on a double log scale, may be

[1] Particularly the 1919 intersection which is above the $5,000 to $6,000 *net* income interval.

[2] See pages 385 to 392.

[3] The five lines representing wages, business, rents, interest, and dividends were found to interweave to such an extent when drawn on one chart that two charts were drawn for each year, one representing wages and business and the other incomes from property.

Wages includes "salaries, wages and commissions" and in 1916 and 1917 "professions and vocations."

Business includes "business," "partnerships, personal service corporations, estates, and trusts," and "profits from sales of real estate, stocks, bonds, etc.," and in 1918 and 1919 "professions."

Rents includes royalties.

Interest includes unclassified investment income.

seen at a glance in its entirety. We see from Charts 28X and 28Z that, though the ratio of the income from wages and salaries to total income may, when charted, show an angle above $5,000, the entire "bulge" on the wages and salaries curve itself occurs in the under-$5,000 intervals both in 1918 and 1919. Moreover, while "wages and salaries" is the largest item in these lowest income intervals, and hence is the controlling factor in determining the peculiar shape of the total curve in this region, it is not the only item showing irregularities and "bulges." Some of these movements are extremely difficult to explain. Why should a "bulge" appear on the lower income ranges of the "rent" curve in 1918 and by 1919 become pronounced? [1] The appearance of a bulge on the *wage* curves in 1918 and 1919 seems quite explicable on the basis of heterogeneity within the wage and salary data themselves but one feels a shade less confidence in any explanation of why that curve moved in this peculiar manner if the explanation does not seem also clearly applicable to the rents curve which moved in an apparently similar manner.

[1] A mere increase in rents will not, of course, account for this *unevenness* in their distribution.

CHAPTER 31

INCOME DISTRIBUTIONS FROM OTHER SOURCES THAN INCOME TAX RETURNS

Concerning the frequency distribution of incomes over $3,000 or $4,000 per annum we have almost no information aside from the income tax returns. Existing wage distributions and non-tax income distributions almost never reach higher than $2,500 or $3,000 per annum.

Even in the lower income ranges (under say $2,500 or $3,000) most of the existing non-tax income distributions are of little use in our problem. In the first place there are less than half a dozen distributions of this sort which are not such small samples as to prevent us feeling much confidence in their representative nature.[1] An even more serious defect of every such distribution known to us, with one exception[2] is that the purpose for which the data have been collected almost inevitably makes them extremely ill-adapted to our use. For example, one of the largest recent samples is prefaced by almost a page of introduction explaining what types of recipients were purposely excluded.[3] This is rather typical. To base upon such distributions any wide generalizations with respect to the income curve for the country as a whole or even for the localities from which such data were collected would be unwarranted.

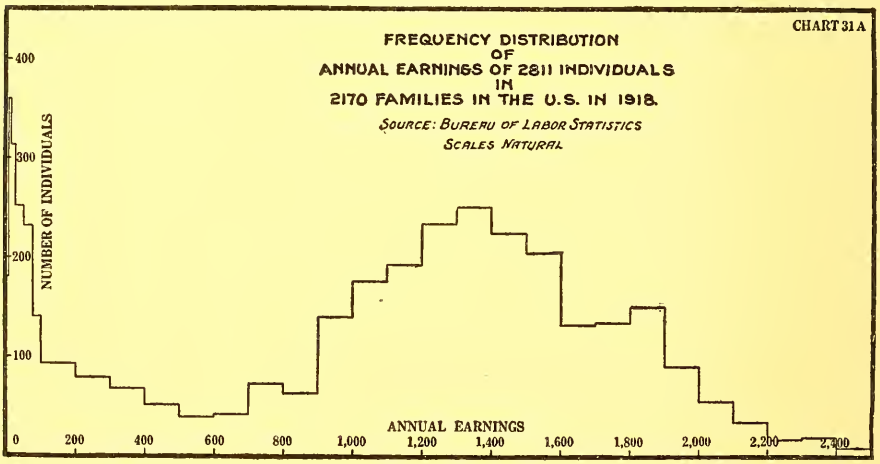
Furthermore, almost without exception these studies in *income* distribution are on a *family* basis. While it is sometimes possible to make a

[1] For example, Chapin's well-known investigation into the distribution of incomes includes only 391 workingmen's families, and the best distribution of farmers' incomes includes only 401 farmers from a single state.

[2] Arthur T. Emery's distribution of income among 1960 Chicago households.

[3] "In studying the sources of income and the importance of each source with relation to the total income of a family the following limitations to the type of family schedules should be kept in mind. No families were scheduled in which there were children who lived as boarders, that is, paid a certain sum per week or per month for board and spent the remainder of their earnings or salary as they saw fit. No families were scheduled which kept any boarders. The number of lodgers to be kept by a family was limited to three at any one time. No families were scheduled in which the total earnings of the family did not equal 75 per cent. or more of the total income. It will be seen that these limitations excluded a large number of families and this materially affects the percentage of families having earnings from children and income from lodgers, and also results in showing a larger percentage of the total income as coming from the earnings of the husband than would be the case if the type of families named had not been excluded from the study. It also reduces the actual amount per family earned by children and received from boarders or lodgers that would be shown in case a cross section of a community including all the types mentioned were used. The object in making the exclusions named was to secure families dependent for support, as largely as possible, upon the earnings of the husband. Of course, it was impracticable to secure a sufficient number of families in which the only source of income was the earnings of the husband, but in following the course named the percentage of families having an income from other sources has been very largely reduced." "Cost of Living in the United States—Family Incomes," *Monthly Labor Review*, Dec., 1919, p. 30.

rough estimate of the individual incomes from the family data, such estimates are not what are needed for our purposes. They can show nothing but the distribution of income among the individuals constituting these families and these families are almost inevitably so chosen as to make the individuals composing them not representative of income recipients at large. Analysis of the distribution of earnings among the individual members of such families discloses an heterogeneity so extreme as to result in a pronouncedly duomodal distribution curve. The fathers' incomes have one mode while the children's incomes have another. Chart 31A showing a natural scale frequency distribution of earnings among 2811 individuals in 2170 families in 1918 [1] exhibits this duomodal appearance in a striking manner. The "families" had been so chosen as to exclude both young



married couples having no children and unmarried but independent wage earners. Investigations planned to bring out the economic characteristics of such "typical families," while they may be extremely valuable for the purposes for which they were undertaken, are necessarily of but little use in the construction of a frequency distribution of all individual incomes in the community. Moreover, even if we were attempting to construct a family and not an individual distribution these data would not generally be particularly helpful for, in addition to the exclusions just mentioned, further narrow and rigid restrictions are usually, and for the purposes in view quite properly, imposed upon the definition of the "typical family."

[1] This is a sample from the 12,096 white families referred to in note 3, page 415. The detailed figures of this sample were tabulated for us by the Bureau of Labor Statistics. They cover 15 cities chosen as representative of the whole list. Each one of the 15 cities shows the duomodal appearance referred to in the text.

As incidentally remarked above, there is one non-tax income frequency distribution to which many of the above criticisms do not apply. It is the distribution of income among 1960 Chicago "households" in 1918 from an investigation made by Mr. Arthur T. Emery for the Chicago *Daily News*.[1] Instead of attempting to describe a "typical family" Mr. Emery attempted to discover the "household" income of each person whose name came at the top of a page in the Chicago city directory. Mr. Emery encountered many difficulties in attempting to follow out this scheme and has himself pointed out sources of error.[2] Notwithstanding the inevitable difficulties, Mr. Emery seems to have made a real effort to obtain a scientific sample. While his distribution shows unmistakable irregularities, it is in many respects for our purposes the most interesting and suggestive recent non-tax income distribution available.

Finally, it seems impossible to obtain from these distributions any but extremely general conclusions concerning the relation between income from effort and income from property. The data have almost always[3] been so chosen as to eliminate any families obtaining an appreciable fraction of their income from property. While they may give us some clues as to the shape of the upper range tail of the wage-earners' income distribution curve[4] they can tell us little about even the upper tail of the *general* income curve and almost nothing about the lower income tail of either the wage-earners' or the general income curve.

[1] While the Bureau is not at liberty to publish this material we were permitted to make what use we could of it in constructing our income curve for the country.

[2] In a letter to the Bureau he writes, "There was, however, one important source of error in this method—the poorer and middle class residents were willing to talk, and with the carefully trained approach of the investigator, the upper class was also won over, but we found in the wealthy districts that the butler and 'not at home' caused a large amount of travel on the part of the investigator," and often a final failure to obtain any report.

[3] These remarks do not apply to the distribution of income among the 401 farmers or Mr. Emery's distribution. However, the Bureau has no figures, in the case of Mr. Emery's distribution, for income from property.

[4] Compare pages 378, 379, 380.

CHAPTER 32

WAGE DISTRIBUTIONS

There is in all an immense amount of American wage data. On the other hand, as an investigator gets into his subject, he begins to realize that the material is more remarkable for its fragmentary nature than for its amount —great as that may be. For no recent year can he obtain wage distributions for more than about 8 per cent. of those gainfully employed. Of course, if these 8 per cent. were scattered over the different types of employment and localities in any truly random fashion, and if their wages were uniformly reported, much might be done with the material. As things are, however, whole occupations as important as agricultural labor and trade are almost unrepresented. Moreover, as we are interested in the amount of wages actually received during the year, it is rather discouraging to find that this is the one type of distribution which practically never occurs. Distributions of amounts actually earned in a month are almost as rare. There are a few distributions of amounts actually earned in a week or fortnight, but the great majority of wage distributions are distributions of wage *rates*—figures by the *hour* being the commonest—or of hypothetical earnings, generally known as full-time earnings per week.

Now it is in general impossible to construct a wage distribution for earnings from a distribution of rates. Earnings depend, of course, not only on rates but also on hours worked. However, we seldom know anything about the distribution of hours worked and almost never do we know anything about the relation between *rates* and hours worked. Chart 32A illustrates how violent may be the difference in shape of the *earnings* and *rates* curves for the same individuals.[1] The *earnings* distribution in this particular case shows not only a much greater scatter than the *rates* distribution but is of an entirely different shape, as may be seen from Chart 32B where the data are drawn on a double log scale. Chart 32C shows the distribution of hours worked in a week for the same individuals. Now, though the slaughtering and meat packing industry may be an extreme example, what evidence we have suggests that distributions of *rates* and of *earnings* are rarely in close agreement. Moreover the relation of the one distribution to the other changes as we pass from industry to industry.[2]

[1] 43,063 Male Employees in the Slaughtering and Meat Packing Industry in 1917. Bureau of Labor Statistics, *Bulletin 252*. For purposes of comparison the two distributions are so placed that the frequency curves show the same arithmetic means and areas.

[2] Resulting largely, of course, from the varying types of distributions of hours-worked-in-

CHART 32 A

FREQUENCY DISTRIBUTIONS OF RATES OF WAGES PER HOUR AND EARNINGS PER WEEK FOR 43,063 MALE EMPLOYEES IN THE SLAUGHTERING AND MEAT PACKING INDUSTRY IN THE U. S. IN 1917.

SOURCE: BUREAU OF LABOR STATISTICS. BULLETIN 252

SCALES NATURAL

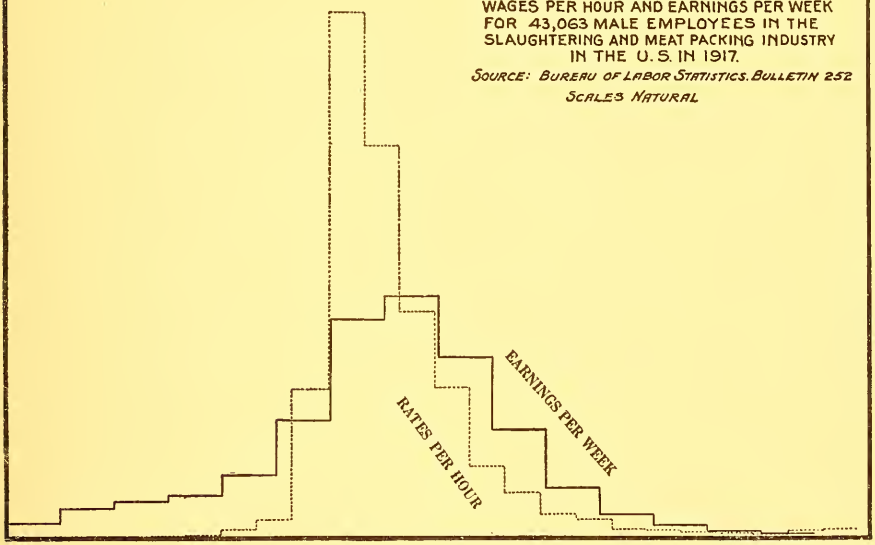
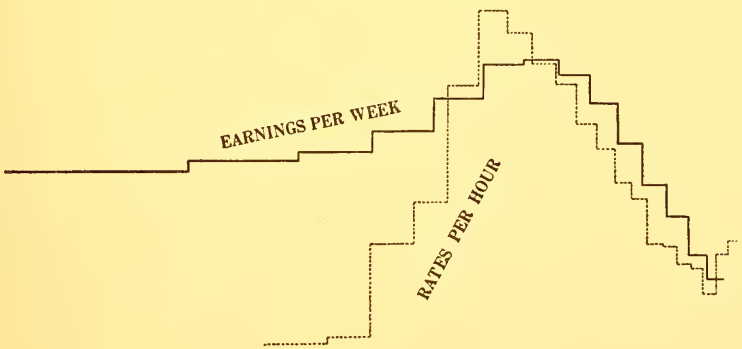


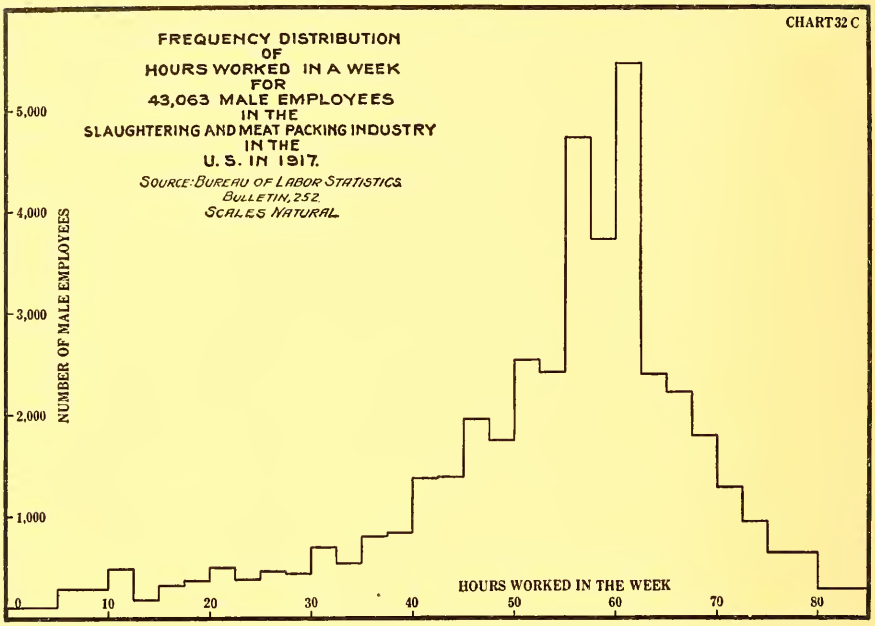
CHART 32 B

FREQUENCY DISTRIBUTIONS OF RATES OF WAGES PER HOUR AND EARNINGS PER WEEK FOR 43,063 MALE EMPLOYEES IN THE SLAUGHTERING AND MEAT PACKING INDUSTRY IN THE U. S. IN 1917.

SOURCE: BUREAU OF LABOR STATISTICS BULLETIN 252

SCALES LOGARITHMIC.





The same difficulty as we find in any attempt to estimate the distribution of earnings per week from the distribution of rates per hour seems inherent in any attempt to estimate the distribution of earnings in a year from the distribution of earnings in a week. The unknown distribution of weeks worked in the year must seriously affect our results.[1]

Estimating the frequency distribution of wages earned in a year for an industry from the frequency distribution of wages earned in another year in the same industry, if we had such data, would involve us in a similar difficulty. Even though we knew the total number of individuals gainfully employed and their total wage bill each year and also the frequency distribution of earnings for one of the years, estimating the frequency distribution for the other year would be hazardous. While some *rates* distributions for the same industry in the same locality show symptoms of not changing in shape very radically from year to year,[2] this does not seem

the-week (month or year) in different industries. Illustrations of lack of uniformity in the relation between rates and earnings of the same persons for the same period but in different industries were worked up from Professor Davis R. Dewey's Special Report on Employees and Wages for the 12th Census.

[1] We have no distributions of amounts earned in a week and in a year for the same industry, with which to illustrate this point directly.

[2] For example, the distribution curve for wages per week among Massachusetts factory workers shows a moderate degree of similarity of shape from year to year.

Professor H. L. Moore (*Political Science Quarterly*, vol. XXII, pp. 61–73) discussed the fluctuation from 1890 to 1900 in the variability of wage rates in a total made up of thirty

a sufficient reason for assuming the same of *earnings* distributions. The shape of the distribution representing hours or days worked in the year may be expected to change greatly from year to year with alternations of prosperity and depression.[1]

What little evidence we possess suggests that wage distributions [2] for individuals of the same sex in the same industry at the same date, but in different *localities*, though generally more dissimilar in shape than distributions for the same industry in the same place but at different *dates*, are less unlike one another than distributions for different *industries* though in the same place and at the same time. The variation in shape of such distributions for different industries is often extreme.[3]

selected manufacturing industries. These distributions (for 1890 and 1900) illustrate both the similarity and the difference in *rates* distributions between the two years.

[1] For example, what little information we have points to the "scatter" of the days-worked-in-a-year distribution being much greater in a year of depression than in a year of prosperity. The extreme variations in shape of the *income* distributions for the same 1240 individuals in the years 1914 to 1919 as seen in the *Statistics of Income*, 1919, page 30, are interesting in this connection.

[2] Whether earnings or rates.

[3] Examples of this are numerous. Charts 32D and 32E show the distribution of wages per week among Massachusetts males working in (a) the boot and shoe industry and (b) the paper and wood pulp industry. For purposes of comparison the two distributions are so placed on the natural scale chart that the frequency curves show the same arithmetic means and areas. The double log chart is based directly upon the natural scale chart. It was necessary to break up the "over $35" interval before calculating the arithmetic means.

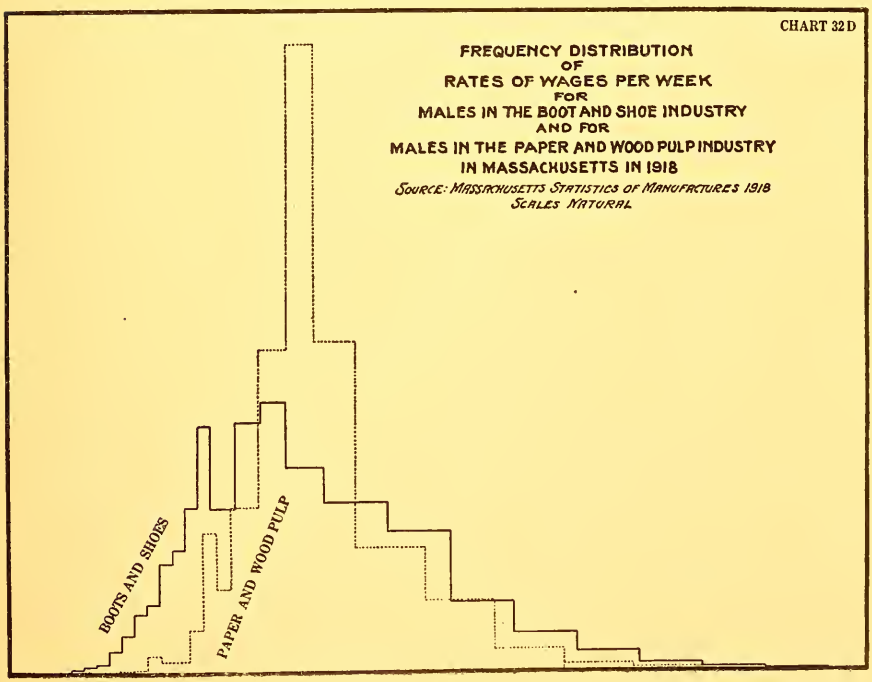
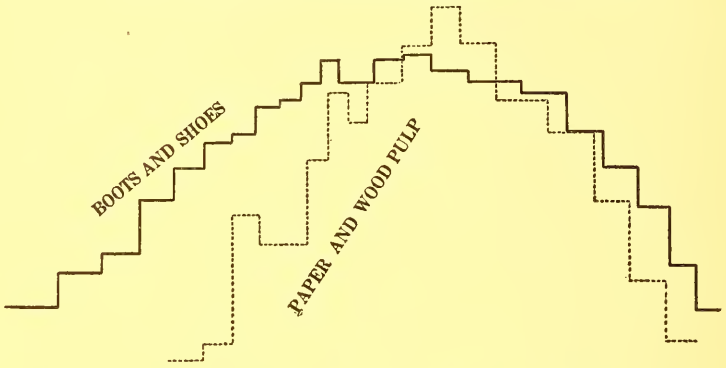


CHART 32 E

FREQUENCY DISTRIBUTION OF RATES OF WAGES PER WEEK FOR MALES IN THE BOOT AND SHOE INDUSTRY AND FOR MALES IN THE PAPER AND WOOD PULP INDUSTRY IN MASSACHUSETTS IN 1918.

Source: Massachusetts Statistics of Manufactures 1918.
Scales Logarithmic



In conclusion, the order of importance of the variables as affecting the shape of the distribution curve seems to be—industry, place, time.

We have but little basis for estimating total income from earnings. In the preceding chapter on Income Distributions from other Sources than Income Tax Returns attention was drawn to the difficulty of arriving at any reliable statement of relationship between earnings and income from such distributions because of the way in which the data were selected. It is even less possible to discover the nature of any such relationship from the income-tax material. Though there is no such apparent "selection" in the income-tax data as in the case of non-tax income distributions, the material is not arranged to answer our particular question.

The non-statistical reader examining Charts 30D, 30E and 30F, on which are plotted average total income and average income from wages in each income interval, might think that it would be quite simple to estimate the probable average total income of persons having any specified wage. However there is a profound statistical fallacy involved in the use of this material for any such purpose. As given in the official tables, income is the independent variable, wages the dependent. This condition cannot be reversed without retabulation of the original returns. The statistical student recognizes the problem as one involving the impossibility of deriving one regression line from the other when neither the nature of the

equation representing the regression line [1] nor the degree of relationship (correlation in the broad, non-linear sense) is known. Even if we knew that the average *net income* of those persons reporting in 1918 in the $5,000 to $6,000 net income class was $5,474 and the average *wage* obtained by these persons was $2,192, we would be quite unwarranted in concluding that the average *income* of persons receiving $2,192 per annum *wages* was $5,474. If no wage earner received income from any other source than wages we still would have a condition where the average *income* in the income class would be greater than the average *wage*. Total wages would be necessarily less than total income, because in the income class are included not only wage earners but capitalists and entrepreneurs. But both total wages and total income are divided by the same number to get an average—namely total number of persons in that income class.

This suggests a technical criticism of the material contained in the *Statistics of Income*. All data concerning the relation between two variables are always there published in such a manner as to give information concerning only one of the regression lines and no information whatever concerning the "scatter." If such data were published in the form of "correlation tables" the increase in usefulness for statistical analysis would be very great. Such "correlation tables" keep closer to the original data than the usual type of statistical tables. Freer use of them is much to be desired, particularly in cases where it is difficult to anticipate all the problems for whose solutions investigators will go to the tabulated materials.

[1] The difficulty of the problem is, if possible, increased in this particular case because of the fact that the regression is radically non-linear.

CHAPTER 33

THE CONSTRUCTION OF A FREQUENCY CURVE FOR ALL INCOME RECIPIENTS

The direct and only adequate method of discovering what is the frequency distribution of income in the United States would be to define very carefully the terms *income* and *income recipient* and then have a carefully planned census taken by expert enumerators upon the basis of these definitions. The returns brought in by the enumerators should moreover be sworn to by the persons making them and heavy penalties attached to the making of false or inaccurate returns. A less satisfactory method but one which would probably give excellent results would be to have a large number of truly random samples taken by such a census. The results of either procedure could then be adjusted in the light of other statistical information concerning the National Income and also in the light of theoretical conclusions derived from the data themselves.

Constructing an income frequency distribution for all income recipients in the United States from the existing data, a few of whose peculiarities have been noted in the preceding chapters, necessarily involves an extremely large amount of pure guessing. It is only because of the practical value of even the roughest kind of an estimate that any statistician would think of attacking the problem. The method followed in the actual construction of the income frequency distribution has been outlined in volume I.[1] This method contains one assumption after another that is open to question. Moreover we feel in many cases quite unable to estimate the probable errors involved in these assumptions. Their only excuse is their necessity. What is the amount of under-reporting for income tax and how is it distributed? What is the effect upon the returns of "legal evasion?" To what extent is the "bulge" on the income-tax returns in the region under about $5,000 in 1918 the result of the "intensive drive?" What is the relation between wages and total income by wage intervals? What is the relation between wage rates and earnings in any particular industry? Etc., etc. These are all questions which must be answered over and over again and yet they are questions the answers to which must be, in many instances, almost pure guesses. And, to repeat, the margin of possible error is often large.

In view of the sparsity and inadequacy of the data, our first approach to the problem was an attempt to discover, if possible, some general mathematical law for the distribution of income. Were we to get any very defin-

[1] *Income in the United States*, Vol. I, pp. 122–139.

ite and reliable clues as to the mathematical nature of the frequency distribution of income from small sample income distributions and from wages distributions, etc., such clues might of course be invaluable in checking the results obtained from piecing together existing wage distributions, income distributions, and other scattered information. We would be in the position of the astronomer who is able to "adjust" the results of his observations in the light of some known mathematical law. It soon became clear, however, that it is quite impossible to discover any essential peculiarities of the income frequency distribution. The available material is not only insufficient for purposes of such generalizations, but moreover the distribution from year to year is so dissimilar, that any generalization of this nature is too vague to be of any practical value.

The method finally used for the construction of the income curve has therefore, we are sorry to say, practically all the weaknesses of the data from which it has been constructed. The occupations of the country were tabulated and to each occupation was assigned those wage and income distributions which seemed applicable with the least strain. We had then a series of income and wage distributions which nominally covered nearly all the income recipients in the United States, though for some occupations the inadequacy of the wage and income samples was little short of absurd. The wage distributions were converted into income distributions on the assumption that the smaller the wage the larger is its percentage of total income. Beyond this simple assumption the particular functional relationships used for many industries were almost pure guess work. Moreover, not only was there the danger of error in moving from wage distribution to income distribution and the danger of error resulting from estimating a wage distribution for a particular industry in a particular locality from a similar though not identical industry in a different locality, but also there was the danger of error resulting from estimating a wage distribution for one year from a wage distribution for another.

The final results are probably not quite so bad as they might have been had we not had a number of collateral estimates with which roughly to check up and otherwise adjust the first results of our estimates. For example, such independent information as Mr. King's estimate of the total income of the country and Mr. Knauth's estimate of the total amount of income from dividends were pieces of information with which the results of the frequency curve calculations were made to agree.

Some hypothetical reasoning is inevitable in such a statistical study as the present one. The investigator must not lose heart. Sir Thomas Browne in his rolling periods sagely remarks that "what song the Syrens sang, or what name Achilles assumed when he hid himself among women, though puzzling questions, are not beyond all conjecture!"

VITA

Frederick Robertson Macaulay was born in Montreal, Canada, August 12, 1882. He attended McGill University, 1899–1902; Colorado College, 1906–1907; the University of Arizona, 1907–1908; the University of Colorado, 1908–1911. From the University of Colorado he received three degrees, B. A. 1909, M. A. 1910, LL. B. 1911.

He attended Columbia University for three years, 1912–1915. During that time he studied under Professors Edwin R. A. Seligman, Benjamin M. Anderson, Jr., Robert E. Chaddock, John B. Clark, William A. Dunning, Frank A. Fetter, Franklin H. Giddings, Wesley C. Mitchell, Henry L. Moore, Henry R. Mussey, Karl F. Th. Rathgen, James H. Robinson, Joseph Schumpeter, Henry R. Seager, James T. Shotwell, Vladimir G. Simkhovitch. He attended the seminars of Professors Seligman, Seager, and Schumpeter.

He taught miscellaneous economic subjects for one year (1915–1916) in the University of Washington, Seattle, Washington. He then taught Economic Theory and Statistics for three years (1916–1917, 1917–1918, and 1919–1920) in the University of California, Berkeley, California. During the year 1918–1919 he was California District Statistician for the Emergency Fleet Corporation. Since May, 1920, he has been on the research staff of the National Bureau of Economic Research, New York City.