A COMPARATIVE ANALYSIS OF THE ROBUSTNESS OF TWO METHODOLOGIES OF MORTALITY ESTIMATION TO AGE ERRORS

HODA RESHAD[*]

INTRODUCTION

The completeness of recording of deaths may be estimated using two procedures. The first was introduced by Brass(Brass, 1975) and the second procedure elaborated by preston and others (preston,Coale,Trussel and Hill,1980). Both procedures assume the stability of the age distributions and equal underregistration of deaths. They make use of the reported age distributions of the population and deaths.

Age misstatement,wether for the deceased or the population is a common feature in reported age and death distributions.The purpose of this paper is to compare the effect of age errors on the derived measure of mortality using the two afore mentioned procedures.

The effect of age errors on the estimates of mortality using Brass procedure has been discussed in details(Rashad,1978). A brief exposition of the effect of certain types of age errors on procedure and a comparison between the two procedures was presented (Preston et al,1980). The focus of their comparison was guided by the view point expressed as:"the predictable nature of the effect of particular defects in the data on the sequences is an additional diagnostic tool for detecting and sometimes for

[*] Lecturer,Institute of Statistical Studies & Research,Cairo University.

correcting the defects". They came to the conclusion that preston procedure may reveal defects in reported data while Brass procedure is of a limited diagnostic tool.

In our study we approach this comparison from a different point of view since as pointed out in (Preston et al,1980) several types of errors are basically indistinguishable, and also since in actual population the combinations of different types of error would possibly result in irregular sequence of $(\hat{N}(a+)/N(a+)$ - ratios on which Preston procedure are based; the correction of reported data my prove to be a complicated process[1]. On the other hand, the fact that both procedures use different formulations may result in different reactions to the same types of age errors and the predictable nature of this reaction would suggest the use of a specific procedure when certain types of age errors prevail. Thus our concern lies with a comparison of the extent and direction of errors introduced in in the estimates of the crude death rate using the two procedures, when no correction of data is attempted.

This study is divided into three parts. The first presents the types of age distortions used in our analysis. It draws heavily on three sources: a general model for age errors developed in (Rashad,1978), and the innacuracies considered in(Preston et al,1980) as well as an empircal type of age error clearly witnessed in Egyptian data (Committee on Population and Demography,1982) The second part deals in brief with the two procedures of mortality estimation and finally the results of subjecting a stable age and death distributions to the different models of age errors are presented and conclusions drawn.

1) Types of Age Distortions.

The two main reasons for age misreport in developing countries are ignorance of age and/or bias associated with this age. The persons aged X may be divided into two classes; the first includes everyone who knows his age correctly while the second include those ignorant of their age. The first class may be subdivided to a_{1x} and a_{2_x} where a_{1x} includes those knowing their age and not bia sed in their reporting of this age, and a_{2x} those knowing their age and biased in their report. Similarly,the second class is divided to a_{3x} and a_{4x}; where a_{3x} includes those not knowing their age and not biased, and finally a_{4x} includes those not knowing their age but biased.

The model for age reporting may be given by:

$Y_x = x$ in group a_{1x}

$Y_x = x + BI_x$ in group a_{2x}

$Y_x = x + er_x$ in group a_{3x}

$Y_x = x + er_x + BI_{x+er_x}$ in group a_{4x}

where, Y_x: reported age when the true age is x.

BI_x: bias associated with age x.

er_x: random error associated with age x.

in other words, if a person knows his age and is not biased against this age or towards a neighbouring age he will state his age correctly. If he is biased the reported age depends on the kind of bias prevailing. If he does not know his age but is not biased, he will attempt to state his age correctly, the deviation between the reported and actual age is simply a random error. Finally if a person does not know his age but is biased against or towards a certain age-which is usually in the neighbourhood of his actual age-he will either avoid or report this age as his actual age.

Instead of dealing with exact age x, we will consider single years age group x, where x denotes the age between x-½ and x+½.

For a full description of this model, the distribution of a_{jx}, er_x and BI_x has to be specified. We will consider two situations, the first when the same kind of error affects both the age and death distributions and the second when differrent errors affect the age and death distributions.

In general, the younger the age the closer the incident of birth and the more likely the age is known; it is expected that the probability of being in group a_{1x} and a_{2x} is a decreasing function of age.

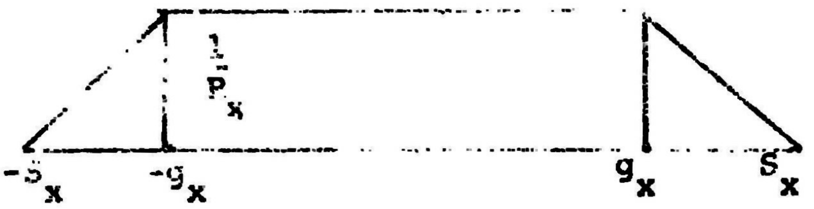
In the first case we will assume the following arbitrary values:

$$P(a_{1x} + a_{2x}) = 70\% \qquad x < 5$$
$$= 50\% \qquad x > 5$$
$$P(a_{2x}) = 30\% \; P(a_{1x}+a_{2x})$$
$$P(a_{4x}) = 40\% \; P(a_{3x}+a_{4x})$$

For the second case, the responsibility of reporting the age of the deceased lies with another person, thus it is more likely that the proportion who do not know their age is higher in the death distribution than in the population age distribution.

The probability of being in group $(a_{1x}+a_{2x})$ for ages higher than 5 was reduced to 30%, keeping other proportions constant.

The distribution of the random error was chosen in the form:



$$f(er_x) = \frac{(er_x + S_x)}{R_x(S_x - g_x)} \qquad -S_x < er_x < -g_x$$

$$= \frac{1}{R_x} \qquad -g_x < er_x < g_x$$

$$= \frac{(S_x - er_x)}{R_x(S_x - g_x)} \qquad g_x < er_x < S_x$$

Where $R_x = S_x + g_x$. $E(er_x) = 0$ and $Var(er_x) = \frac{g_x^2 + S_x^2}{6}$

The parameters of the distribution (g_x,S_x)were set arbitrary, given that they satisfy that for persons not knowing their, age, there is an upper limit for the S_x imposed by several factors, such as : appearance, social status, type of job ...etc. Also the older the person, the higher is the upper limit of his deviation (the higher the values of S_x). Note that although $S_x \geq S_z$ $x \geq z$, $x-S_x \geq z-S_x$ and $x + S_x \geq z+S_x$.

For the first case, the values of the different parameters(g_x,S_x) are illustrated in the following graph. Note that for age 0-1 a slightly different type of error was considered:

$$f(er_x) = \frac{1}{R_x} \qquad -g_x < er_x < g_x$$

$$= \frac{(S_x - er_x)}{R_x(S_x - g_x)} \qquad g_x < er_x < S_x$$

where

$$R_x = \frac{S_x + 3g_x}{2}$$

For the second case, the value of the parameter S_x is increased by 2 for all ages greater or equal 1. Thus implying a bigger random error. For the distribution of the bias error, we note that though there are several types of bias prevailing in developing countries such as digit preference, concentration of women in the middle of the reproduction period, overstatement of age for old people...etc. These biases are basically the same. They show attraction to some ages and avoidance of others.

Two types of bias are studied in this model. The first generally described as digit preference, shows itself as heaping on digits terminating with:0,5,8,2,6 and 4 which of course imply shunning from ages terminating with 3,7,1 and 9. The second bias that characterizes most developing countries is a general movement on the age scale; we will consider the movement from ages 11-19 to ages 20-29 (This movement is clear in female age distributions for African societies) and the movement from ages 51-59 to ages 60-69.

In dealing with digit preference, if x is a preferred end digit, persons whose age end with x state it correctly, unless they are affected by another error. Persons whose age ends with a digit different from x, states their age correctly or ending with another digit according to the following probabilities.

Movement out of age 1 and 9 are stronger than movement out of 3 and 7 as they are close to one of the most preferred end digits.

If f(x/y) denotes the probability of moving from an age ending in y to the closest age ending in x, the different probabilities may be given as:

$f(0/1)=.55$	$f(1/1)=.20$	$f(2/1)=.25$	$f(2/3)=.25$
$f(3/3)=.35$	$f(4/3)=.25$	$f(5/3)=.15$	$f(6/7)=.25$
$f(7/7)=.35$	$f(8/7)=.25$	$f(5/7)=.15$	$f(8/9)=.25$
$f(9/9)=,20$	$f(0/9)=.55$		

In dealing with the general movement on the age scale, a person aged between 11-19 or 51-59 affected by this bias will move up the age scale from 1 to 10 years uniformally, thus:

$$P(BI_x=y)=\frac{1}{9} \qquad 1 < y < 10 \text{ and } 11 < x < 19 \text{ or } 51 < x < 59$$

Where $P(BI_x=y)$ denotes the probability a person aged x will add y years to his age.

Finally, aperson aged 11-19 or 51-59 is subjected to either of the previous biases(digit preference,movement up the age scale) with equal prbability. Thus,a random number decides first which type of bias a person is subjected to and another number reflects the value of this bias. the same kind of bias was considered for the first and second case. In applying the afore mentioned model,the reported age is a stochastic variable. It depends on a random number drawn from any of the distributions specified earlier, according to the group discussed. For example,for a person in group a_{3x}, a random number generated from the distributic of er_x, specifies the value of er_x,required to calculate the reported age as: the actual age + er_x.

Methods for directly generating random numbers from particular distributions are not usually available, but they exist for generating random numbers from uniform distributions. This number may be transformed to the requirec sequence using the relation: F(er)=RN, where F(er) is the cumulative function of the error distribution and RN is th random number drawn from a uniform distribution (0,1). Onc an expression for the inverse cumulative distribution function is available, the error may be easily calculated.

Starting with single years age groups, stable population and death distribution and using atotal population of 100,000 with the assigned probabilities for being in different groups. The numbers in each single year age group is obtained. Each individual in each group is subjected to the appropriate error. The simulated number in each single year age group is summed over all four groups and the reported distribution obtained in single years. The equivalent five years age groups are readily calculated.

Applying the previous procedure twice-First, when the same error affects the age and death distribution.Then when different error affects the age and death distribution - on a stable distribution corresponding to model west, males, with rate r=15% and mortality level 6. We get the following simulated age and death distributions presented in table(1).

Graphs (2) and (3) represent the actual and simulated age distributions in single years and five years age groups respectively. Graph(4) represents the average percentage female age distribution of 30 sets of census or survey data of various African countries and the stable model fitted to this average. This data are extracted from a study of the United Nations on age error in African data. (United Nations,1975). The similarity between the characteristics of age mis-statements in African countries and in the simulated data is apparent.

Preston et al(1980) considered several types of age distortions such as an increasing overstatement of age at death among older persons, starting at age 55. An increasing overstatement of age of persons at older ages beginning at age 55 as well as an increasing overstatement of age among older persons and older decedents, beginning at age 55. They also considered a severe misreporting of age of persons and of deaths of a kind that would duplicate the innacurate age distributions of the female population of India as reported in the census of 1911.

Table(1)

Stable and Simulated AgeDistributions, When the Age Error Model is Applied.

age group	Stable Data		Simulation(1) Same Error Affecting Pop. & Death Dist.		Simulation Different Err Affecting Pop. Death Dist.	
	Pop. Dist.	Death Dist.	Pop. Dist.	Death Dist	Pop. Dist	Death
0-	16161.0	1867.2	16653.0	1877.2	16460.8	1861
5-	13085.4	114.4	12413.4	100.4	12559.1	114
10-	11690.6	73.2	9888.6	67.2	9936.2	77
15.	10451.7	90.1	9310.7	80.1	9327.2	75
20-	9205.1	113.4	11760.1	134.4	11599.5	127.
25-	8001.4	110.3	8489.4	108.3	8608.8	115.
30-	6892.5	110.1	7166.5	108.1	7068.8	105.
35-	5863.6	110.3	5510.6	113.3	5717.9	109.
40-	4903.6	112.2	5009.7	119.2	4924.0	124.
45-	4019.3	108.0	3806.3	108.0	3904.4	104.
50-	3203.0	110.3	2745.0	98.3	2720.2	94.
55-	2453.7	104.9	2163.7	91.9	2146.8	92.
60-	1771.6	104.7	2370.6	126.7	2374.7	107.
65-	1166.9	93.5	1301.9	81.5	1239.9	93.
70-	675.6	75.9	758.6	68.9	725.6	76.
75-	320.3	51.9	374.3	46.9	386.3	54.
80-	134.4	34.7	277.4	54.7	299.8	51.

Table(2) presents the distorted population and deaths age distributions resulting from subjecting the stable distribution-corresponding to model west, males, with growth rate r=15% and mortality level 6-to an overstatement of age by five years for an increasing fraction for age intervals from 55-59 to 80+.The fractions used for the age distributions of the population are: 2, 4, 8, 16, 32 and 40 percent respectively. The fractions used for the death distributions are: 3, 6, 12, 24, 48 and 60 percent respectively.

Table(2)

Simulated age distributions, when stable data are affeoted by overstatement of age starting at age 55.

Age group	Stable data		Simulated data	
	Pop. dist.	death dist.	Pop. dist.	death dist.
55-	2453.7	104.9	2404.626	101.753
60-	1771.6	104.7	1749.81	101.565
65-	1166.412	93.5	1144.412	88.562
70-	675.6	75.9	660.856	68.904
75-	320.3	51.9	325.9	45.204
8o+	134.4	34.7	236.896	61.688

The committee on population and Demography (1982), in attempting to estimate the recent trends in fertility and mortality in Egypt, was faced by a certain type of age mis-statement for decedents. In comparing the cumulative age distributions of the Egyptian population and deaths with a corresponding appropriate stable distribution, they found that the Egyptian popuiation is fairly closely fitted by the stable population except for deficits at ages 15 to 35 in the recorded population . On the other hand, an ever widening gap between the cumulative distribution of recorded deaths and the corresponding stable distribution, starting from 10, suggests that deaths whose age are unknown are assigned to ages over 70. the proportion whos age is unknown in Egypt seems to constitute a large fraction of the population and it is probable that ages of the dead ar known even less frequently than ages of the living. the commi estimated that the ratio of the reported population over 70 t the actual population over 70 ranges from 2 to 15 for the period 1937 to 1976. In attempting to duplicate Egyptian age error, we will assume that the reported number of deathes age over70 constitute on average 1.7 of the actual number. This excess of deaths causes a uniform proportionate reduction in number of deaths in ages 10-70.

In comparing the two procedures for mortality estimation we considered the following types of age distortions

1) Error only affects the population age distributions.

i- model of age error resulting in simulated population ag distribution in table(1). (simulation1).

ii- exxageration of age at death considered in Preston at al (1980). Simulated death distribution in table(2).

iii- exxageration of age at death and misallocation of deaths whose ages are unknown to age group(70+).(combination of error in II.ii and Egyptian type of age error).

III) Error affecting both the age and death distributions

i- model of age error resulting in simulated age and death distributions table(1). (simulation1)

ii- model of age error resulting in simulated age and death distributions in table(2). (simulation 2).

iii- exxageration of age of persons and at death considered in Preston et al (1980). (table(2)).

II - Procedures for Mortality Estimation.

Using basic relations in a stable population, Brass showed that:

$$\frac{N(y)}{N(y+)} = r + CDR \cdot \frac{D(y+)}{N(y+)} \qquad (1)$$

Where: N(y) denotes the proportion of persons at exact age y and N(y+) the proportion over age y and D(y+) the proportion of deaths over age y.

Given the usual five years age groupings, the estimates of D(y+) and N(y+) are straight forward. N(y) is usually estimated as 1/10 $({}_5N_{y-5}+{}_5N_y)$; where ${}_5N_y$ is the proportion of population in age group y to y+5.

In certain situations, it was found that the previous procedure for estimating N(y) may introduce bias in the CDR value. An alternative formula for estimating CDR was suggested (Rashad , 1978). This formula takes the form:

$$\frac{{}_5N_y}{5.p^*(y)} = r + CDR \cdot \frac{D^*(y)}{p^*(y)}$$

where $p^*(y)$ and $D^*(y)$ denote the average of the cumulative distributions over ages y and y+5.

Equations(1) and (2) will be denoted as Formula (A) an Formula (B) respectively.

The second procedure for estimating the crude death rat uses the following relation,which holds for stable populatio

$$\hat{n}(a) = \sum_{y=a}^{\omega} d(y).\ \text{Exp.}(r(y-a)),\ \text{where}$$

$\hat{n}(a)$ is the number of population at exact age a and d(y) is the number of death. At exact age (y). Note that Preston's procedure requires knowledge of the growth rate while in Bras procedure the growth rate is a by product.

If deaths are not completely recorded, a comparison of $\hat{n}$ with the reported population would provide us with an estimat of the completeness of recording. In practical situations, a comparison of the entire population estimated from the number of deaths to the entire enumerated population is not the best option. The use of $\frac{\hat{n}(a+)}{n(a)}$ (where $\hat{n}(a+) = \sum_{a}^{\omega} {}_x n_a$), for sequent values of a from 5 or 10 to 65 or 70 is preferred.

$\hat{n}(a+)$ is calculated using the following steps:

-$\hat{n}(80)=d(80+).(1+r.e(80))$where e(80) is the expectation o life at age 80 in the model life table embodied in the stable population.

-$\hat{n}(80+) = \hat{n}(80).e(80).(1-\frac{2r.e(80))}{3}$

-$\hat{n}(a) = \hat{n}(a+5).\exp(5r) + {}_5d_a\ \exp(2.5r)$, for a = 75,70,...,5;

where ${}_5d_a$ is the number of registered deaths from age 5 to a+5

$$-{}_5\hat{n}_a = 2.5(\acute{n}(a) + \hat{n}(a+5)).$$

$$-\acute{n}(a+) = \sum_{x=a}^{75} {}_5n_x + \hat{n}(80+)$$

III) Summary and Conclusions.

In comparing the effect of age errors on the estimates of underregistration using both Preston and Brass Procedure, we used the mean of $\hat{n}(a+)/n(a+)$ for ages 10-65(12 ratios) and 5-70. The same ages were used in applying Brass procedure.The Slope of the line relating $N(y)/N(y+)$ and $D(y+)/N(y+)$ $({}_5N_y/5P^*(y)$ and $D^*(y)/P^*(y)$ in formula (B))was estimated using the group average method: $(Y_2-Y_1)/(X_2-X_1)$; where the ratios used are divided into 2 groups and Y_1,Y_2 are the mean of $N(y)/N(y+)$ $({}_5N_y/5p^*(y)$ for the first and second group.

Table (3) presents a summary of the underregistration estimated by both procedures. Note that the actual underregistration is 1 in all cases considered and the deviation from 1 is due to the effect of age errors.

It is clear that age errors distort the ratios at old ages and the younger the ages used in the analysis, the smaller the error introduced in the estimates. Thus age group 10-65 always provide better estimates than age group 5-70.

The same observation explains why formula (A) always seem to provide better estimates than formula (B). This is simply because $p^*(65)$ and $D^*(65)$ are the means of the ratios at ages 65 and 70. Thus it may be more acceptable to compare preston's estimates and those obtained from formula (A) using age group 5-70 with the estimates of formula (B) using age group 10-65 .

Table (3)

Extent of Underregistration Estimated Using Both Preston and Brass Procedure When Different Models of Age Errors Prevail.

Type of Error		Preston	Brass	
			Formula(A)	Formula(B
Errors in Pop.dist. alone	age groups used			
I.i	10-65	.953	.980	.964
	5-70	.947	.945	.940
I.ii	10-65	1.000	1.017	1.036
	5-70	.995	1.018	1.043
Errors in death dist.alone				
II.i	10-65	1.069	.988	1.002
	5-70	1.074	.983	1.002
II.ii	10-65	1.059	1.010	1.037
	5-70	1.074	1.029	1.076
II.iii	10-65	1.355	1.300	1.540
	5-70	1.257 m	1.650	1.780
Errors in both age & death dist.				
III.i	10-65	.988	.982	.971
	5-70	.990	.959	.975
III.ii	10-65	1.017	1.034	1.062
	5-70	1.016	1.045	1.082
III.iii	10-65	1.055	1.032	1.065
	5-70	1.063	1.057	1.112

We will not adopt this suggestion here and we will limit our comparison to the performance of formula (A) and preston's procedure using age group 10-65.

Two Type of age errors affecting the population age distributions are considered here. The first, affects the whole age distribution and the second-exxageration of age-is only restricted to old.The effect of the second type of age errors is negligible on both preston and Brass procedure (formula (A). This result has been expected, considering that the cumulative age distributions at young ages up to age 55 remains unaffected with the second type of age error. Brass procedure provides better estimate of registration when the first type of age errors prevail. The first model of age errors represents a fair description of actual data as illustrated in graph(3).

When age errors only affect the death distribution, error II.i & II.ii, it is obvious that Brass procedure is much more robust than preston procedure. But when Egyptian type of age errors prevail (II.iii), the estimates of both procedures are totally unacceptable preston procedure providing better estimates. Note that preston procedure is totally dependent on the death distribution to estimate n(a+) and it involves cumulation of error. This explains why the estimates of completeness of registration using preston procedure is sensitive to errors in death data.

When age errors affect both the age and death distributios - error III.i, III.ii - it is reassuring that the extent of deviations using both preston and Brass procedure is modest , Preston procedure performs better than Brass procedure in this case. When exxageration of age -III.iii- is the source of error, both procedure are more strongly affected than when the general of age errors prevail(II.iii), The estimates of both procedures are totally unacceptable with Preston procedure providing better estimates. Note that Preston procedure is totally dependent on the death distribution to estimate n(a+) and it involves cumulation of error. This explains why the estimates of completeness of registration using Preston procedure is sensitive to errors in death data.

When age errors affect both the age and death distributions-error III.i,III.ii- it is reassuring that the extent of deviations using both Preston and Brass procedure is modest. Preston procedure performs better than Brass procedure in this case. When exxageration of age - III.iii-is the source of error, both procedure are more strongly affected than when the general model of age errors prevail.

Thus, we conclude that when both the population and death distributions are affected with error and when the error follows the general model described in section II; both procedures are robust to age errors with a preference for Preston' procedure(assuming that the growth rate is known precisely). When exxageration of age seems to be the main cause of errors, Brass procedure is preferred. If the exxageration starts from an early age(whic covers the inclusion of deaths of unknown ages to older ages) ;

both procedures provide unacceptable estimates underregistration.

Notes

1- When age errors follow the pattern described by III.i and III.ii the values of $\hat{n}(a+)/n(a+)$ presented in table(4) do not follow a clear pattern.

Table(4) : values of $\hat{n}(a+)/n(a+)$.

Age	5	10	15	20	25	30	35	40	45	50	55	60	65
Pattern of age error													
III.i	1.020	1.013	.983	.955	1.003	1.018	1.032	1.019	1.032	1.022	.963	.871	.954
III.ii	1.025	1.020	.966	.966	1.012	1.032	1.045	1.045	1.058	1.066	1.012	.927	1.029

REFERENCES

BRASS,W. (1975) Methods for estimating fertility and mortality from limited and defective data. Laboratories for population Statistics. An occasional publication, The university of North Carolina at hapel Hill.

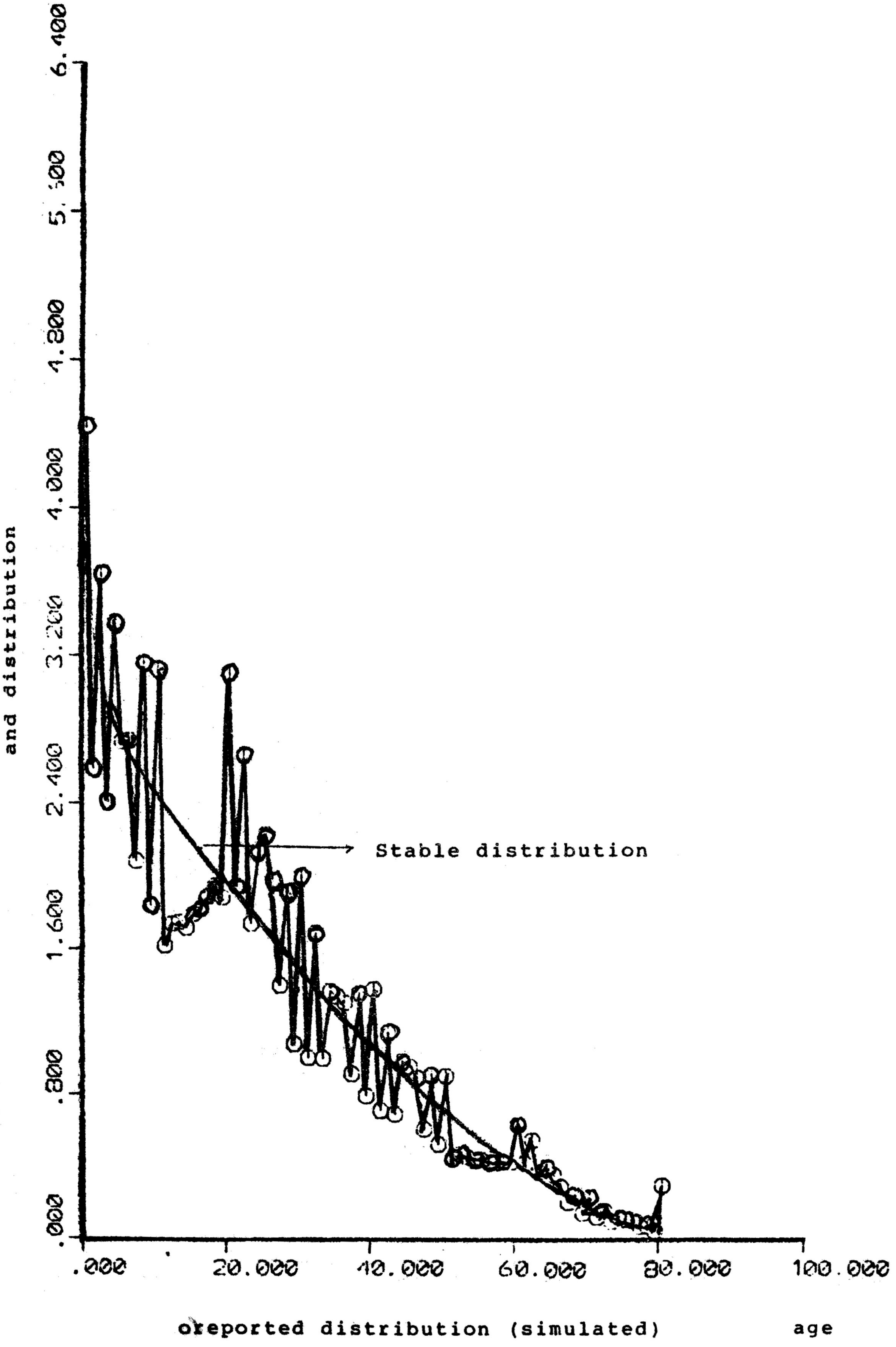
COMMITTEE on POPULATION and DEMOGRAPHY. (1982) The estimation of recent trends in fertility and mortality in Egypt. report no. 9, National Academy Press.

PRESTON,S., A.J.COALE,TRUSSELL and M.WEINSTEIN (1980) Estimating the completeness of reporting of adult deaths in populations that are approximately stable. Population Index 46(2).

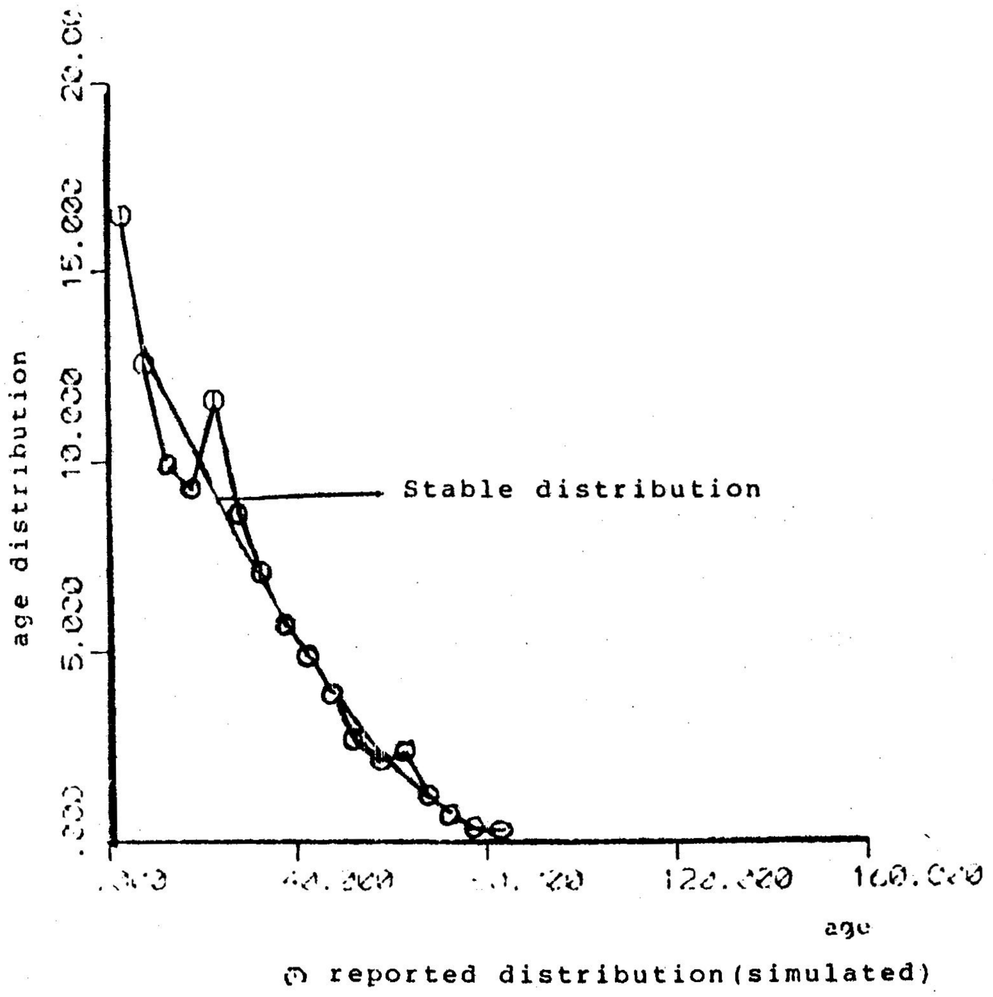
RASHAD,H. (1978) The estimation of adult mortality form defective registration data. Unpublished Ph.d thesis,University of London.

UNITED NATIONS (1975) Techniques of evaluation of basic demographic data. African Population Studies Series,No.2 Addis Ababa.

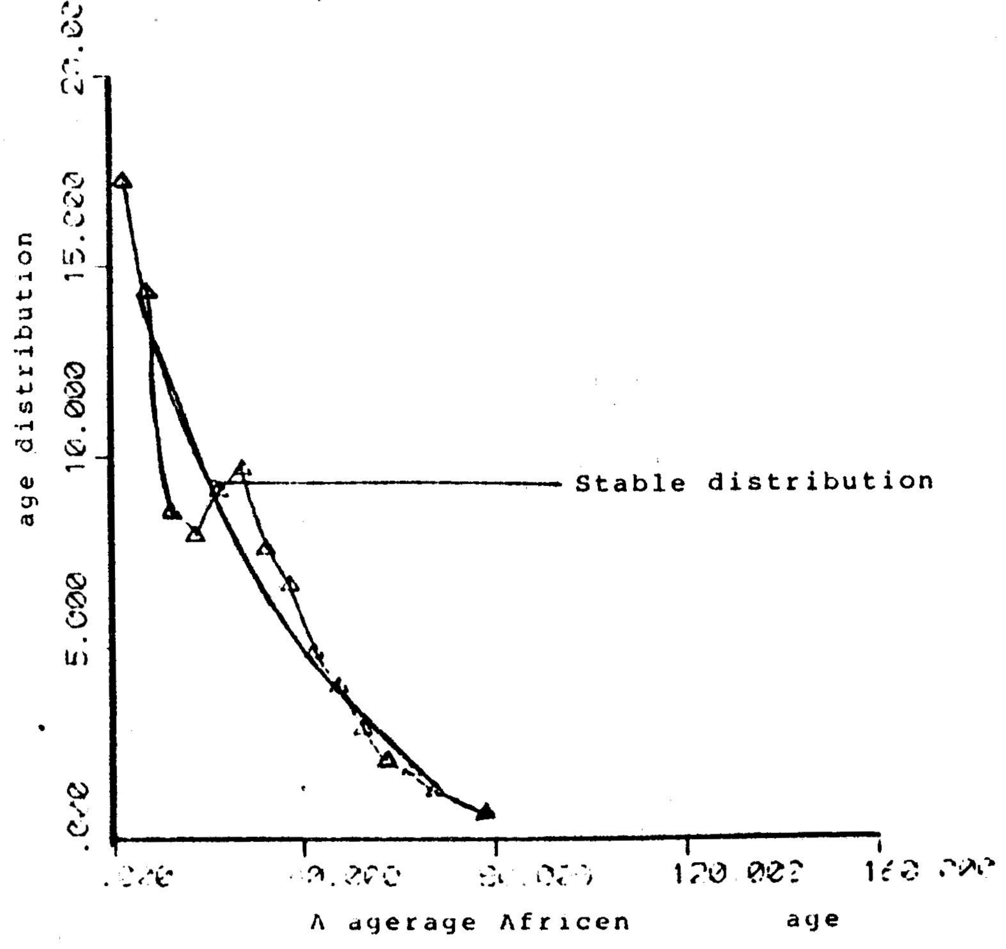
Graph(2) The actual and sumulated age distribution single years.



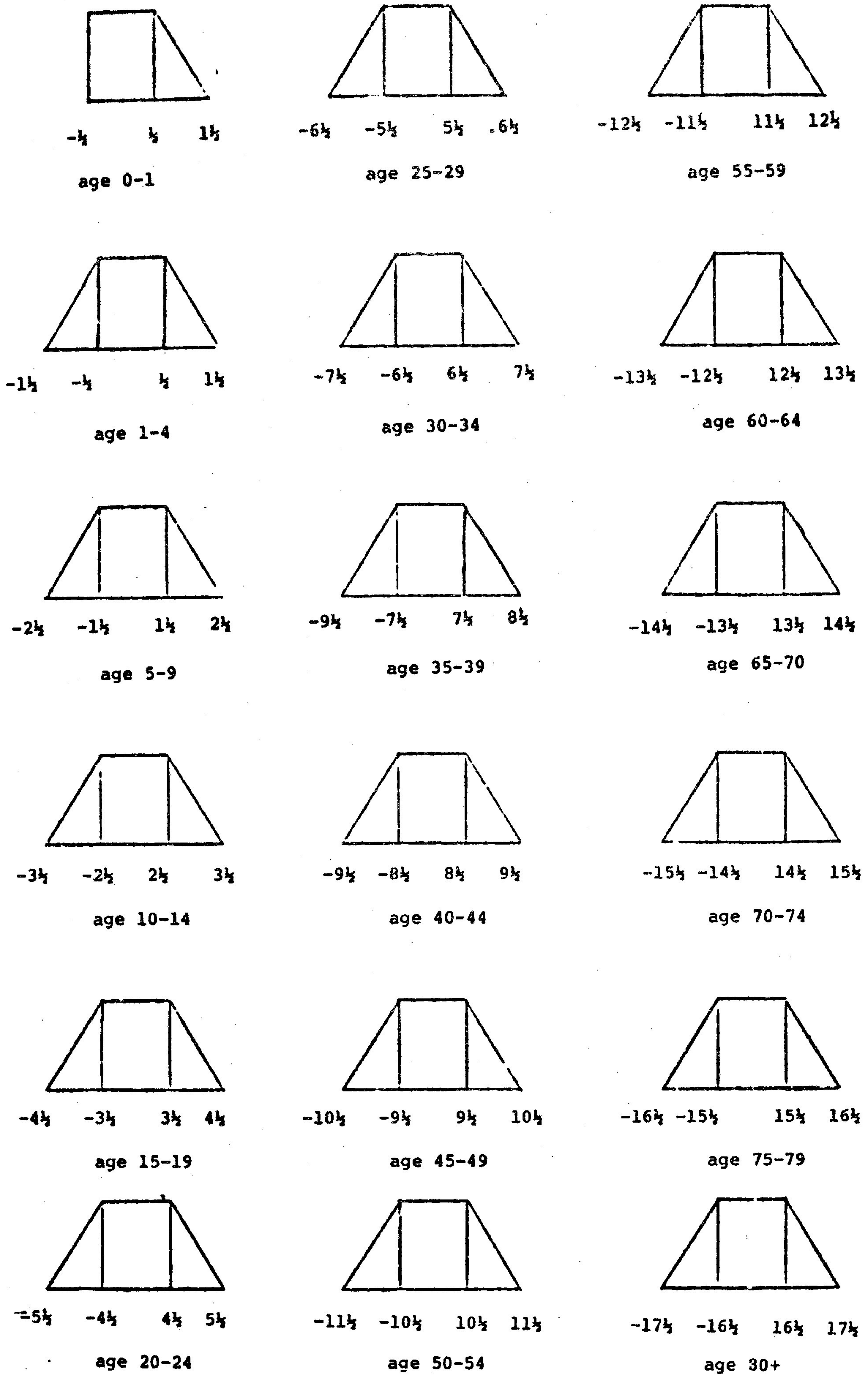
Graph(3) The actual and simulated age distribution 5 years ages group.



Graph(4) The averge.African and fitted stable and distribution



Graph (6.1) Distribution of er_x for different x



age 0-1 age 25-29 age 55-59

age 1-4 age 30-34 age 60-64

age 5-9 age 35-39 age 65-70

age 10-14 age 40-44 age 70-74

age 15-19 age 45-49 age 75-79

age 20-24 age 50-54 age 30+