Percentile-Age Curves for Marathon Times

Gilbert W. Bassett Jr.
January 26, 2004

Abstract: Percentile-age curves are presented for marathon times. The curves show the distribution of finishing times as a function of a runner's age. Separate estimates are presented for Women and Men. Unlike previous studies based on age-group records, our estimates utilize all finishing times for the Chicago Marathon in 2000, 2001 and 2002—more than 87,000 observations. The results show how performance changes as a function of age and place in the overall distribution.

Draft: Comments Welcome

University of Illinois at Chicago
Department of Finance
601 South Morgan (MC 168)
Chicago, Illinois 60607-7121
312.996.5777
gib@uic.edu

1. Introduction

This report presents percentile-age charts for marathon times. Separate charts are presented for Women and Men. The charts are based on all the finishing times in the Chicago Marathon for 2000, 2001, and 2002. While the charts are based on the Chicago Marathon, the estimates for relative changes in performance can be applied to other races.

The charts show how performance changes with age. They allow comparison of an individual's performance with others of the same age. They can be used to forecast performance based on both age and percentile in the overall distribution.

Marathon times as a function of age have been studied by Fair(1994) and recently by Sterken (2003). In addition, the World Masters Athletes (WMA), formerly the World Association of Veteran Athletes (WAVA), publishes age-group dependent correction factors. The correction factors are used to grade results in track and field events. All of the previous estimates have been based on age-group records.

Unlike previous studies, our estimates are based on the entire distribution of finishing times of all runners, best to worst. We can therefore assess changes in performance as a function of age for different classes of runners.

Section 2 summarizes the data and presents summary statistics. We also discuss modifications of the data to try and eliminate selection bias due to prize money for the top finishers. Section 3 presents the percentile-age curves. They are based on regression quantile smoothing estimates that account for year effects. Both absolute time and relative change as a function of age are presented. We also indicate how our estimates compare with previous estimates. Corroborating evidence for the results is the close match between our estimates in the fastest tail of the distribution and the previous estimates based on age-group records. Since fastest and other times change at different rates, our estimates can be used to predict times for non elite runners.

2. Summary Statistics

The Number and Age-Distribution of Runners. The number of finishers in the Chicago marathon for 2000, 2001 and 2002 are given by:

Year	Women	Men	Year Total
2000	11,073	16,819	27,892
2001	11,364	17,264	28,628
2002	12,982	18,128	31,110
Gender Total	35,419	52,211	

The age-distribution of runners was similar for each of the three years; see Figures 1W and 1M for Women and Men. They show that the bulk of the runners are 20 to 40 with Women tending to be slightly younger. The Women mode-range is 25 to about 30 whereas for Men it is about 30 to 40.

Empirical Distributions. Empirical percentiles 1, 50, and 90 are shown in Figures 2W and 2M for each year. (The x percentile is quantile(x/100) where the quantile function is the

inverse of the empirical distribution function). For example 1% of the times were faster than percentile 1 (and 99% of the times were slower). The Figures show that times were similar over the three years. The slightly larger variability at percentiles 1 and 90 is due the expected greater variability in the tails of the distribution. The slight increase in variability evident at ages greater than 45 is due to the smaller number of older runners.

Figure 3W and 3M show percentiles for the combined data over the three years. The "min" curve indicates the fastest time at each age. The percentiles show a similar pattern for Women and Men with times declining steadily after about age 40. Closer inspection shows that the very fastest times peak earlier and decline faster.

Discussion. Assessing the effect of age on performance would, ideally, utilize a large panel of data consisting of times for the same individuals at ages, a, a+1, a+2, and so on.[1] It would be the basis for "growth" curves similar to those for children's height and weight as a function of age; see, for example, Cole(1988), and Koenker (2003). We would estimate F(z|a),the distribution of times at age a, so that, F(z|a+k)-F(z|a), would indicate how the performance distribution changed between ages a and a+k.

While we have a large data set, it is different runners at different ages[2]. We can nevertheless estimate how performance depends on age if our sample is representative of F(z|a), that is, what would be observed in a panel consisting of the same runners at different ages.

To estimate F we assume times are independent with distribution F(.|a). There is however one important exception. It is likely that, relative to what we would expect to see in a random sample based on F, the very fastest times are over-represented at younger ages and underrepresented at older ages. We have too many elite younger runners and too few older runners who used to be elite. The source of the bias is the prize money earned by the fastest runners. It encourages the best runners in the world (who are in their 20s and 30s) to enter the race. Equally important, those runners do not participate when they are older and no longer likely to earn prize money. The prize money means our sample includes elite runners from all over the world, but not formerly elite runners who are in their 40s and 50s and no longer likely to earn prize money.

To diminish and hopefully eliminate this bias, our estimates are based on a modified sample in which the very fastest times are excluded. That is, the estimation data set deletes the overall (no matter what the age) fastest 20 Men and 15 Women times for each year. The idea is that each year there are at most 20 Men and 15 Women in the race who would not participate except for the prize money. Deleting these times leaves a sample that hopefully is representative of the distribution of F.

It should be noted that estimates based on the original, uncorrected, sample would make the fastest percentiles (mistakenly) appear to decrease too quickly. Such a tendency can be seen in Figure 3. For both Women and Men the median time stays roughly constant until the early 40's and then declines. The fastest percentiles, in contrast, peak earlier and decline more rapidly. At least part of this is due to the sample bias discussed

---

[1] Technically, (since we ignore, not only those who never participate, but also those who start but do not finish Marathons) the reference population for the analysis would be individuals (respectively, Women, Men) who finish Marathons at ages 20 to 60.

[2] Actually, some unidentified runners participate in two or more of the races.

above. But part could also be an intrinsic tendency for fast times to decline more quickly even after controlling for bias. By eliminating the top 20 Men and 15 Women from the data each year the bias due to the overrepresentation of elite runners at younger ages is, hopefully, eliminated.

3. Percentile-Age Curves

Let the finish-time of a runner be denoted by T(A,y) where A is age, and y is a an indicator for the year of the race, representing, e.g., race-day conditions(the notation for Women and Men is suppressed for simplicity). The distribution function of T(A,y) is denoted by F(z|A,y) and the associated quantile function is Q(τ|A,y). The quantile function is assumed to be a smooth function of age, a fourth degree polynomial given by

$$Q(\tau \mid A, y) = a(\tau) + a_{2000}(\tau)\,year2000 + a_{2001}(\tau)\,year2001 + \beta_1(\tau)A + \beta_2(\tau)A^2 + \beta_3(\tau)A^3 + \beta_4(\tau)A^4.$$

The model is estimated using regression quantiles at selected τ values; see Koenker and Bassett(1978). The estimates are essentially smoothed versions of the empirical percentiles that account for the differing age proportions and the slight year-to-year differences in times.[3] The estimates of the quantile coefficients are presented in Tables 1W and 1M.

The percentile-age times for Women and Men are depicted in Figures 3W and 3M. The predicted times are based on the regression quantile estimates reported in Table 1W and 1M. The estimated times in the Figure are for 2002; a simple correction factor for finishing times for the other years is indicated below[4].

Year	Women	Men
2000	add 4:09	add 1:20
2001	add :06	add 0:49
2002	add 0	add 0

Figure 3W and 3M can be used by a runner to find their place in the overall age distribution of times. They can also be used to forecast how performance will change with age as a function of one's position in the age distribution. For example, a time of 3:30 for a woman, age 30 puts her in the top 5% of her age group. The forecasted for such a runner at age 40 is 3:35 and 3:47 at age 50.

The percentile-age curves expressed in relative terms, that is, as a fraction of a base age, are presented in Figures 4W and 4M. This form of the estimates was used in the previous studies and can be used to forecast performance at other Marathons. The relative value expresses performance relative to performance at age 35. While the curves are all similar, the most rapid declines are seen to occur at the fastest percentiles of the distribution. At age 50 the time for a Man in percentile 1 is 1.10 his time at age 35, whereas the time for a Man at the median increases by only a factor of 1.05.[5]

---

[3] For discussion of the difference between the regression quantile estimates and related estimates based on the empirical quantiles see Bassett et al. (2002).

[4] Note that the model includes quantile-year effects, though, for simplicity, we report only a single year-correction factor (for all quantiles).

[5] Note that the relative and absolute measures can differ. The same ten minute increase at percentile 1 and percentile 50 corresponds to a faster *rate* at the faster percentile. The same relative increase of 1.1 at percentiles 1 and 50 means a smaller decrease in seconds for the faster time.

Comparison to other Estimates. Figure 5W and M present a rough comparison to other estimates. The figures show relative performance at ages 50 and 60 compared to age 40. Recall that previous studies were based on age-group records and hence would be expected to be most similar to our estimates for the top percentiles. This is indeed the case. The Figures show that the previous estimates are most similar to our percentiles for the fastest times, and all tend to be higher than the estimates in the middle of the distribution, especially for the bulk of the data between ages 40 and 50. This means that a correction factor based on the previous studies or on our fastest percentiles would overestimate the decline in relative times if applied to average or poor runners in the age distribution.

References

Gilbert Bassett, Mo-Yin S. Tam and Keith Knight (2002). Quantile Models and Estimators for Data Analysis, *Metrika*, v,55, 2002, 17-26

Cole, T. J. (1988). Fitting Smoothed Centile Curves to Reference Data, Journal of the Royal Statistical Society, Series A, 151, Part 3, pp 385-418, with discussion.

Fair, Ray C. (1994). How Fast Do Old Men Slow Down? *Review of Economics and Statistics,* 76(1), 103-118.

Koenker, Roger (2003). Quantile Regression of Longitudinal Data, manuscript.

Koenker, Roger and Gilbert Bassett (1978). Regression Quantiles, Econometrica, v.46, p.33-50.

Long-Distance Running Association (2001). http://www.usaldr.org Mundle, P., Dietderich, S., Wood, A., Henry, D. and Wallace, G. (1989). Masters Age Records. Eugene, OR:National Masters.

Sterken, Elmer(2003). From the Cradle to the Grave: How Fast Can We Run*? Journal of Sports Sciences*, 21,479-491.

Figure 1W
Proportion
Women

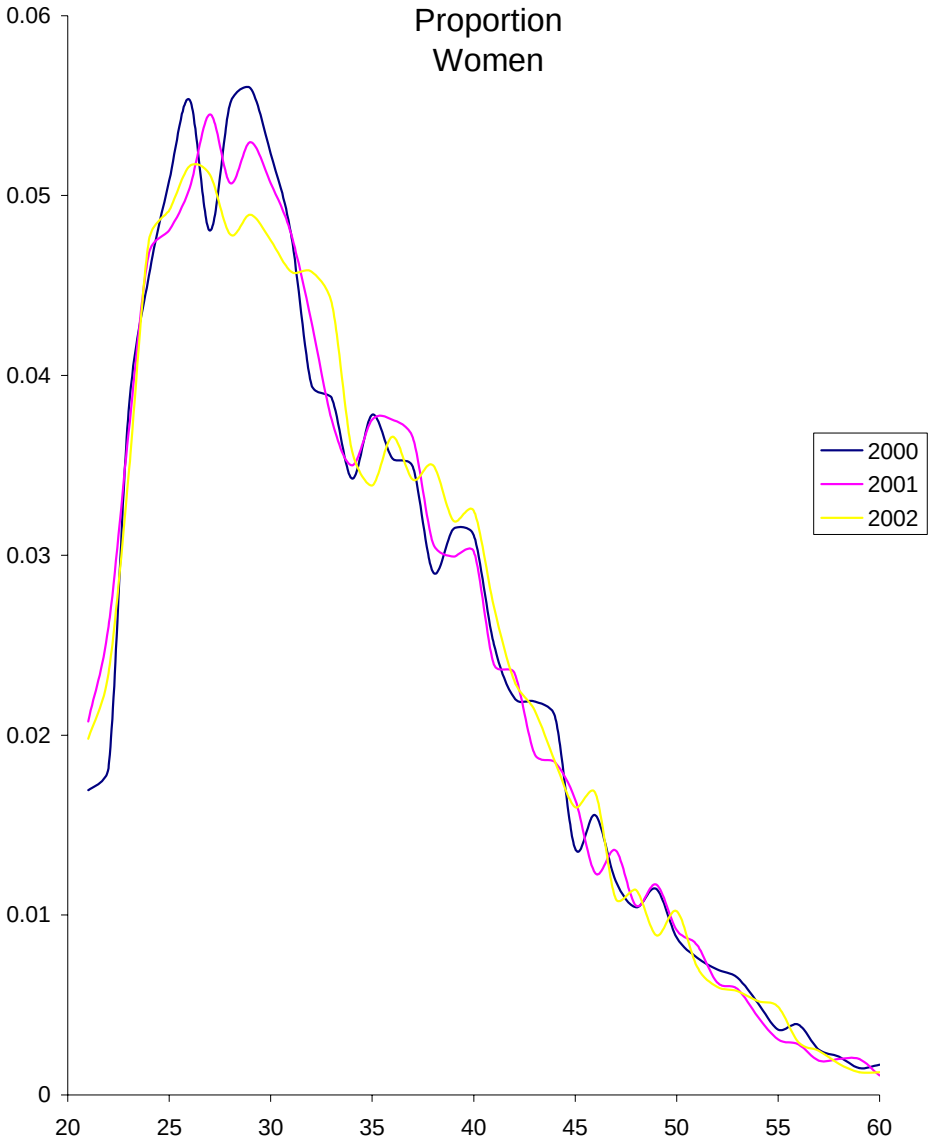

Figure 1M
Proportion
Men

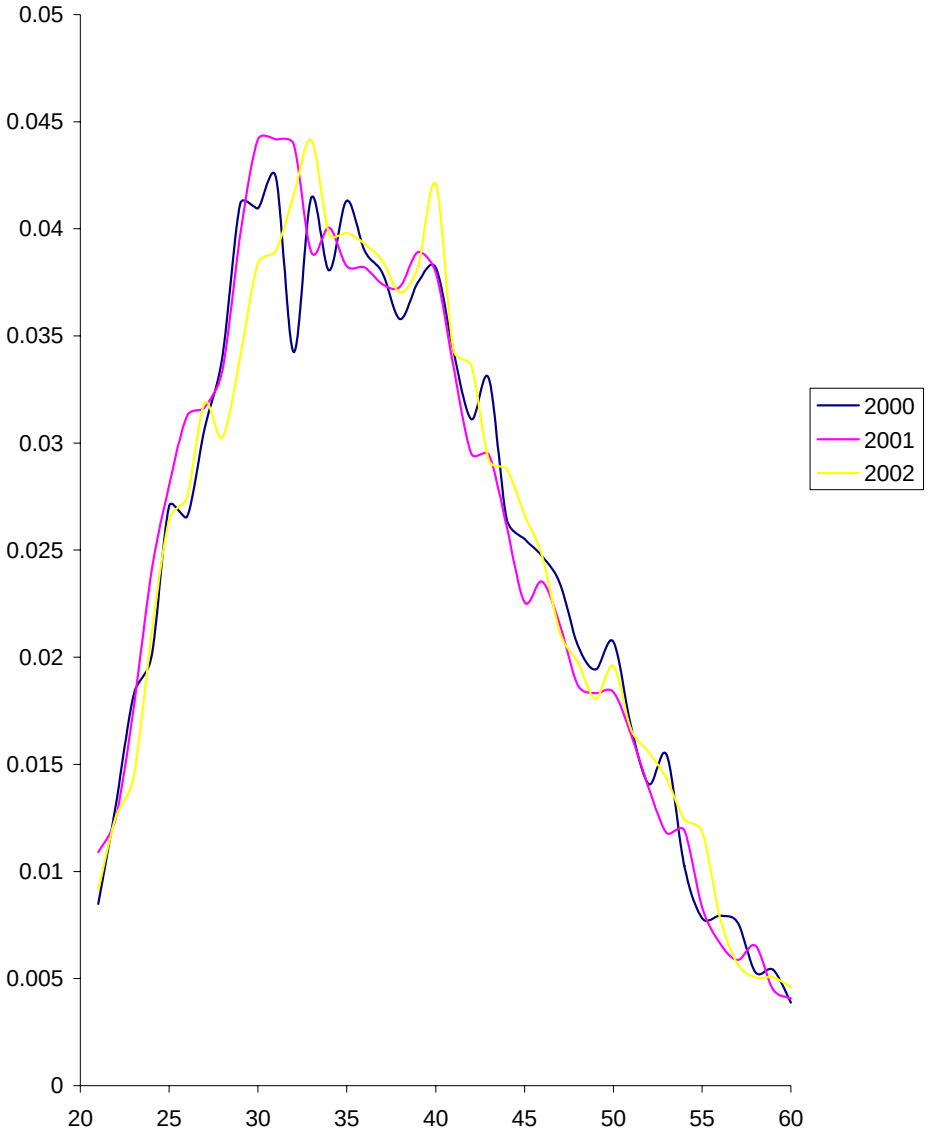

Figure 2W
Percentiles 1, 50, 90
Years 2000, 2001,2002
Women

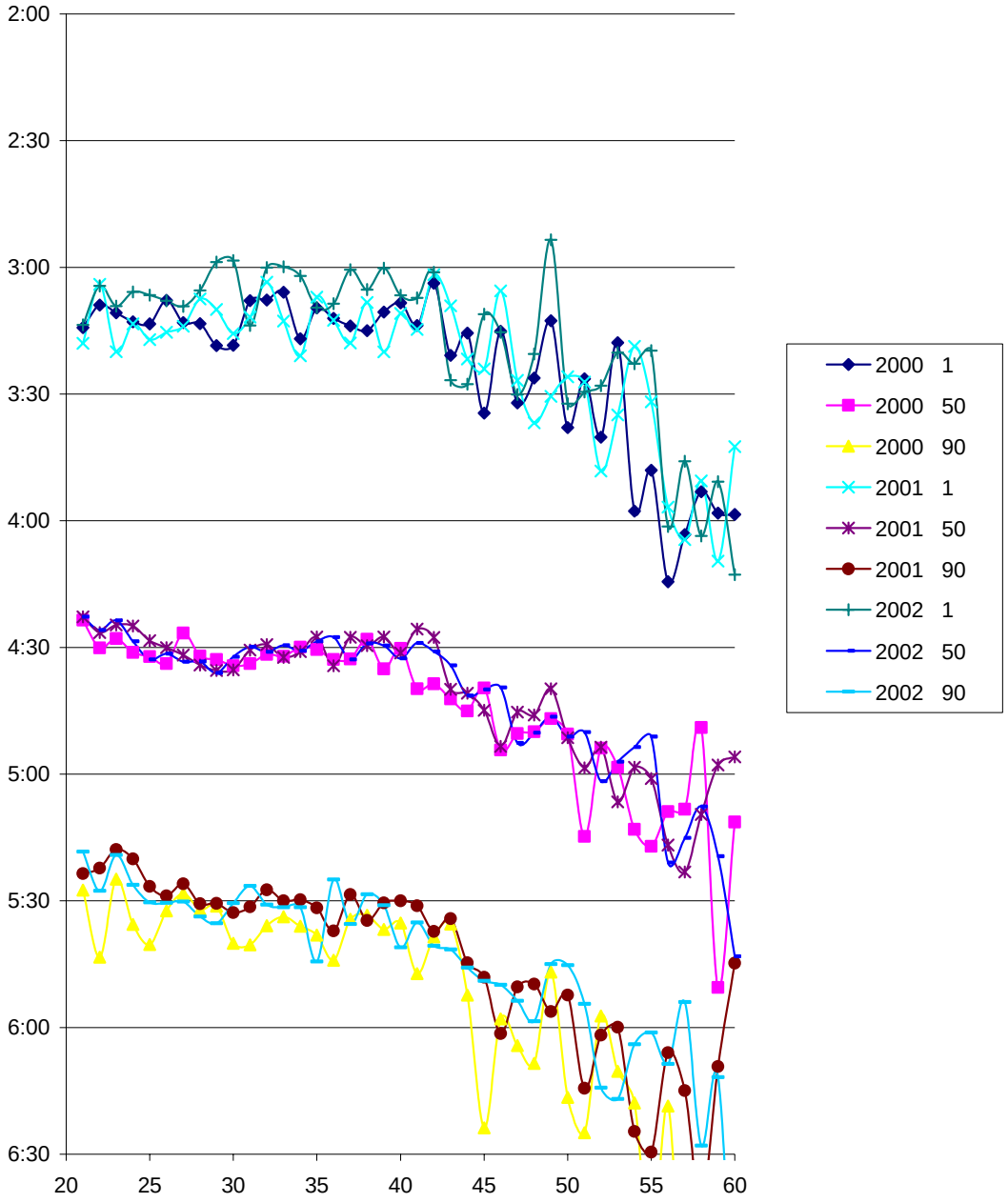

Figure 2M
Percentiles 1, 50, 90
Years 2000, 2001,2002
Men

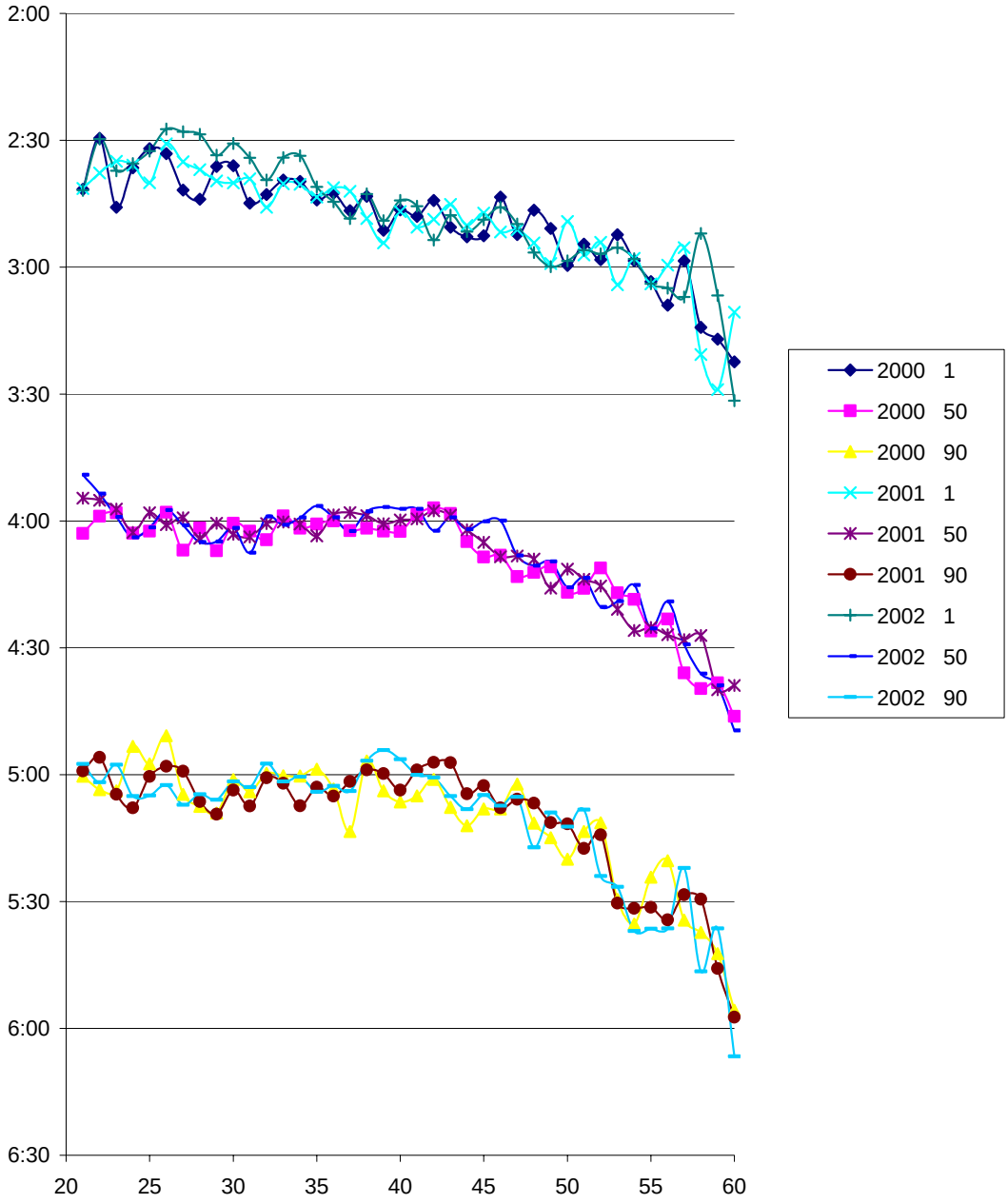

Figure 3W
Empirical Percentiles:
2000-2002
Women

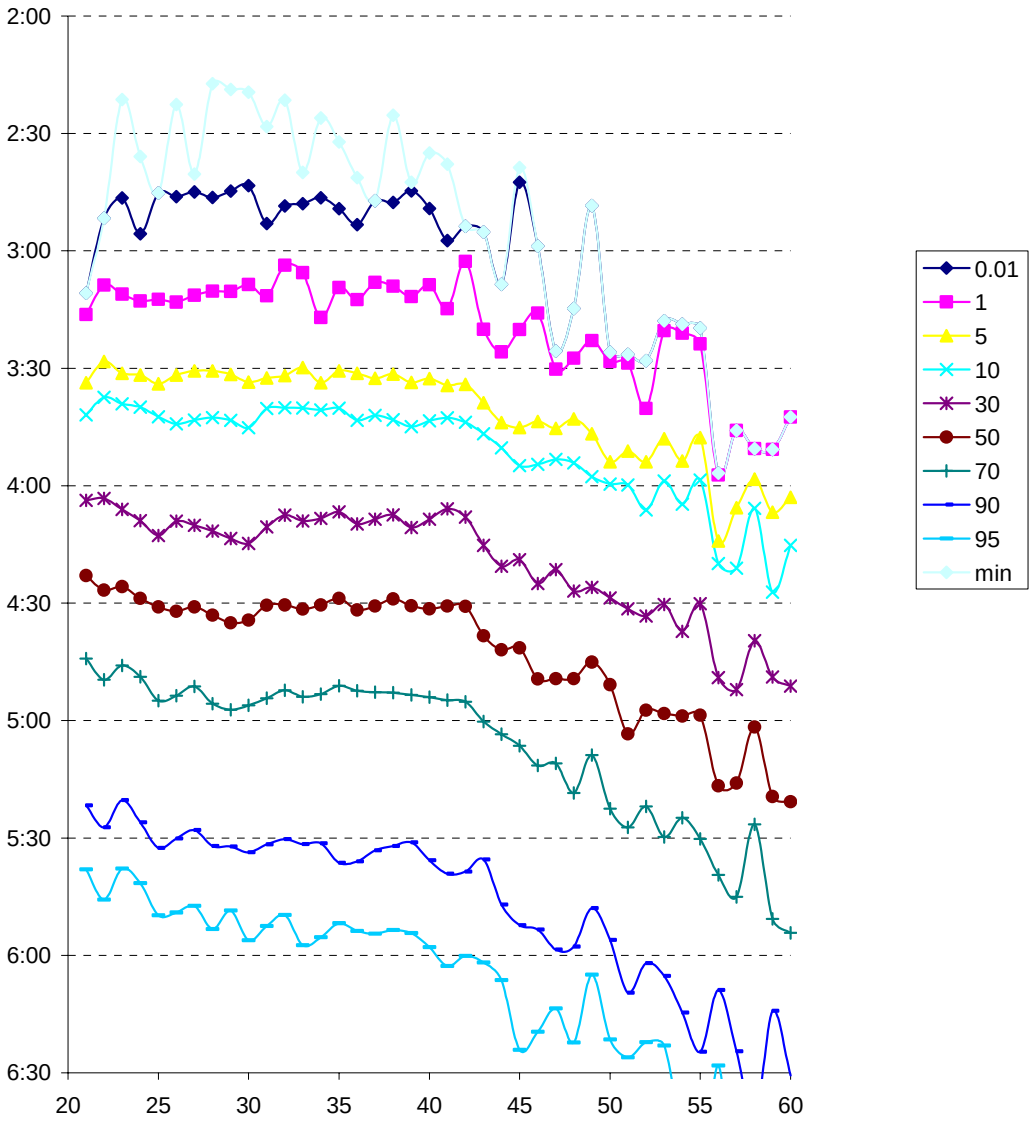

Figure 3M
Empirical Percentiles:
2000-2002
Men

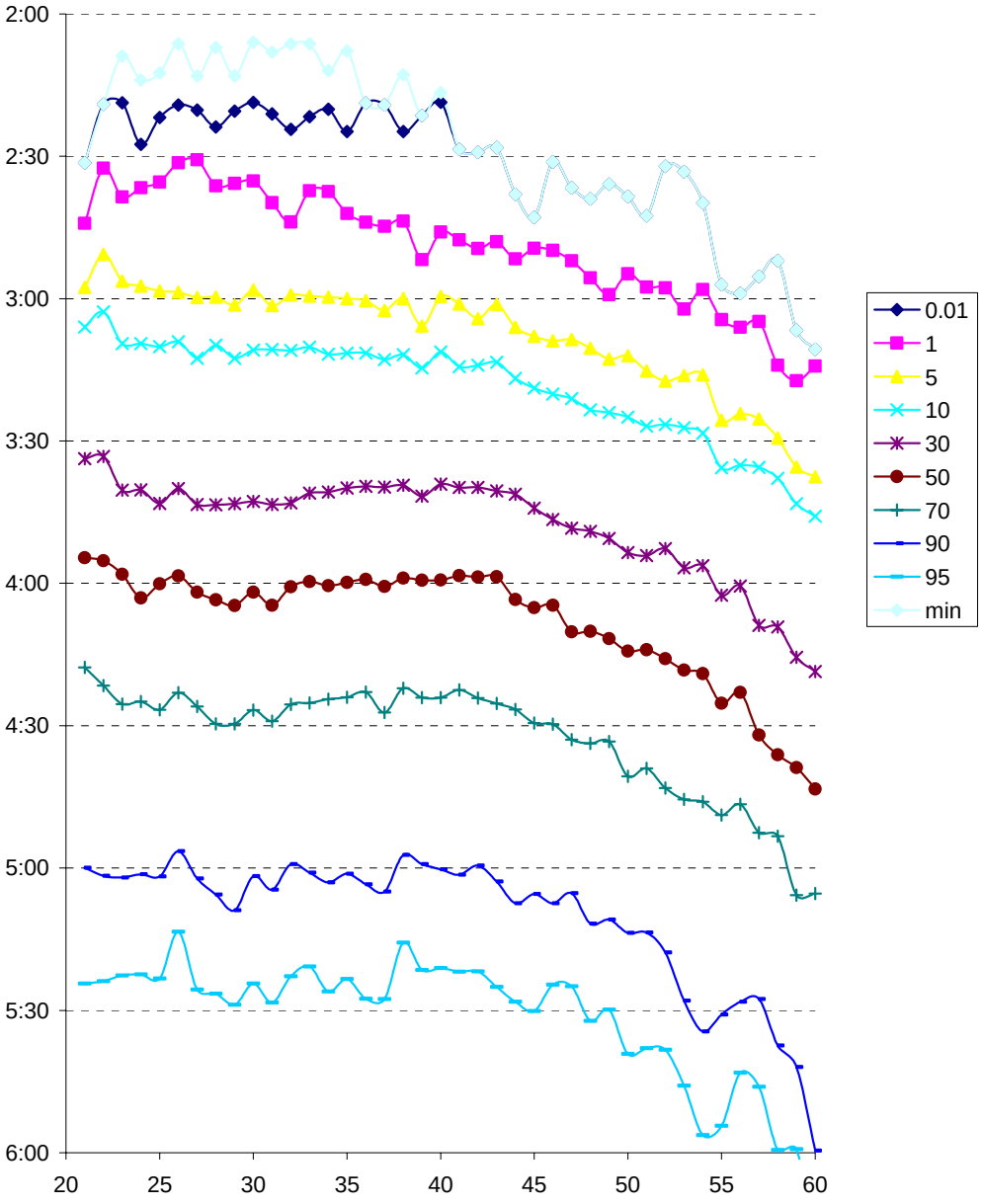

Figure 4W
Percentile-Age Times
Women

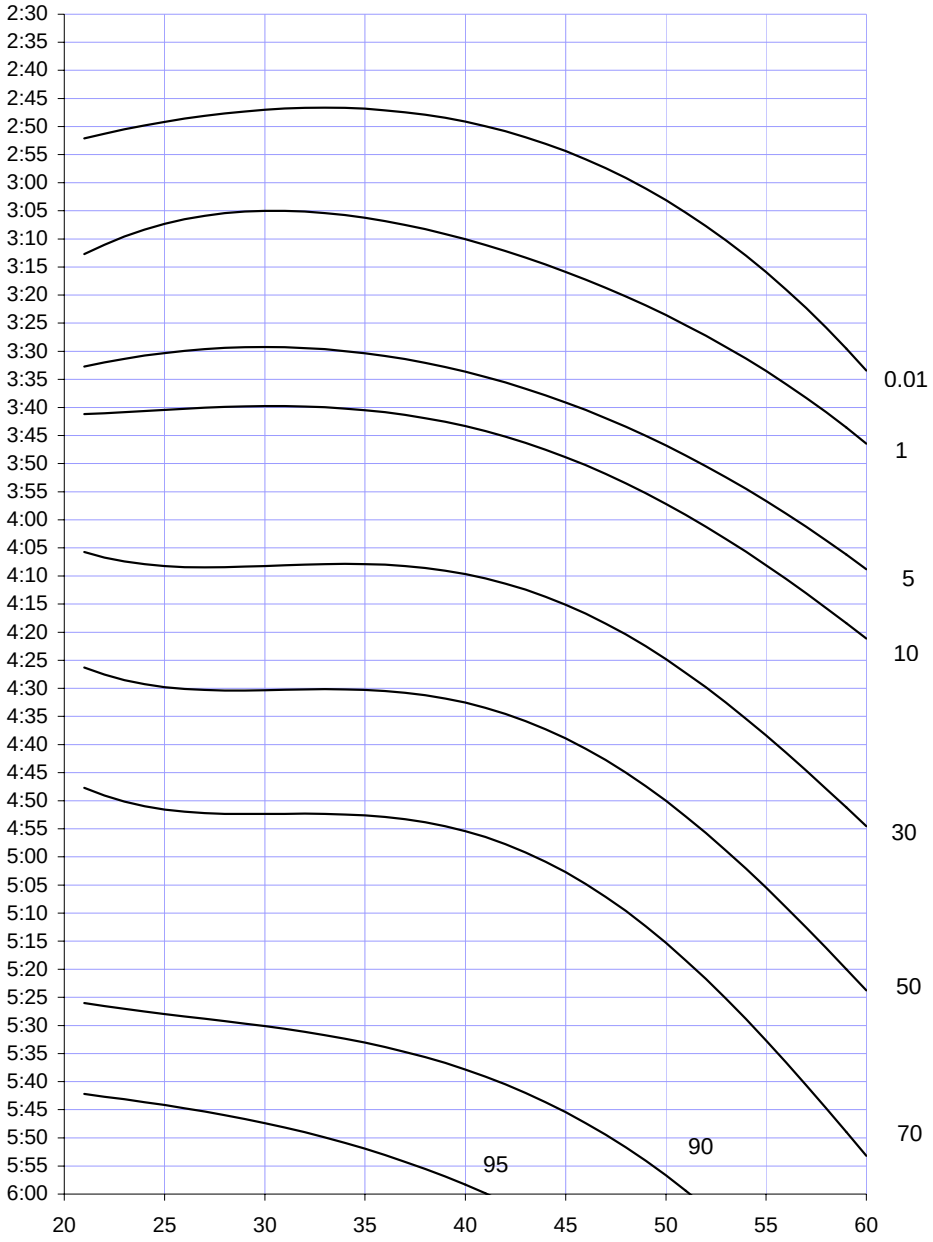

Figure 4M
Percentile-Age Times
Men

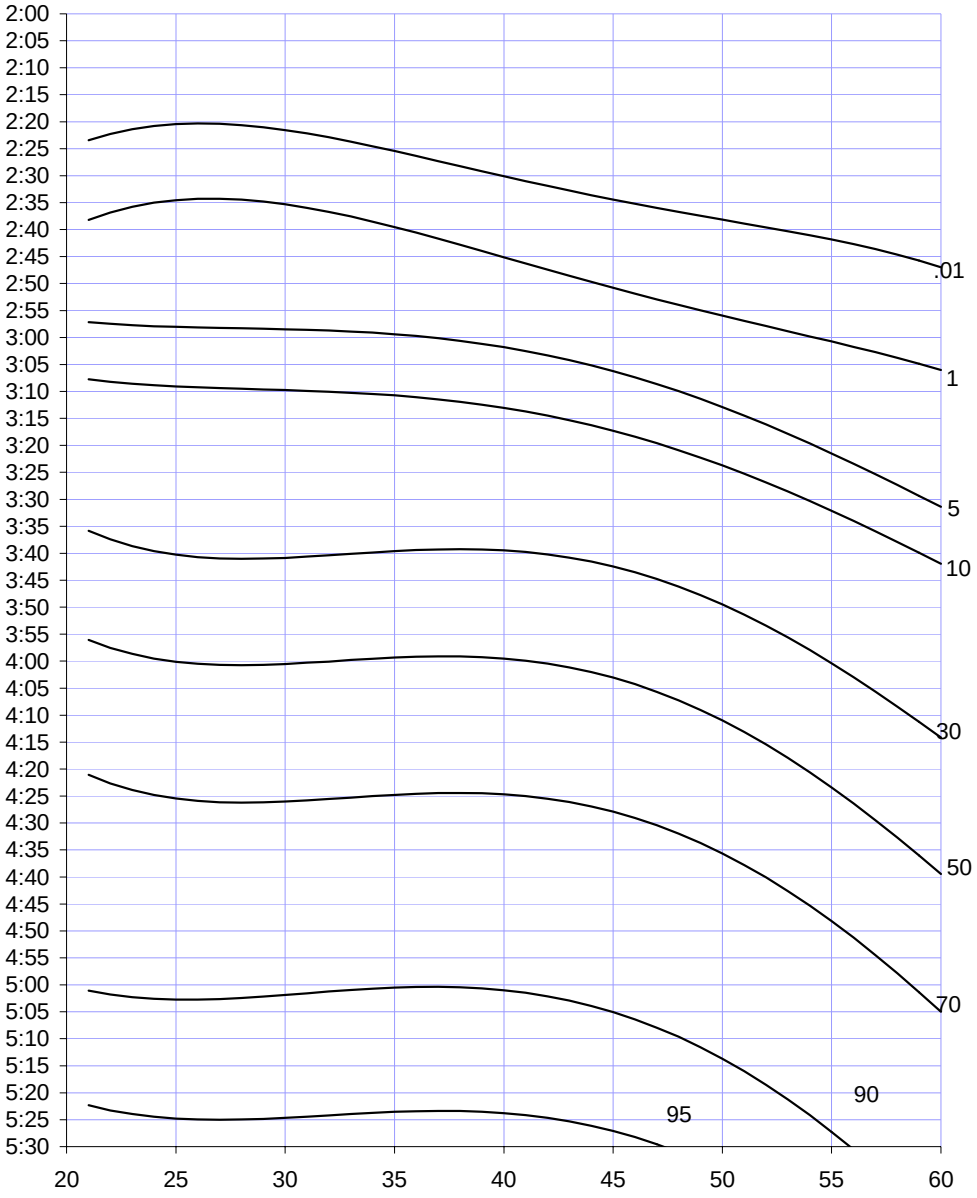

Figure 5W
Time(Age)/Time(35)
Selected Percentiles
Women

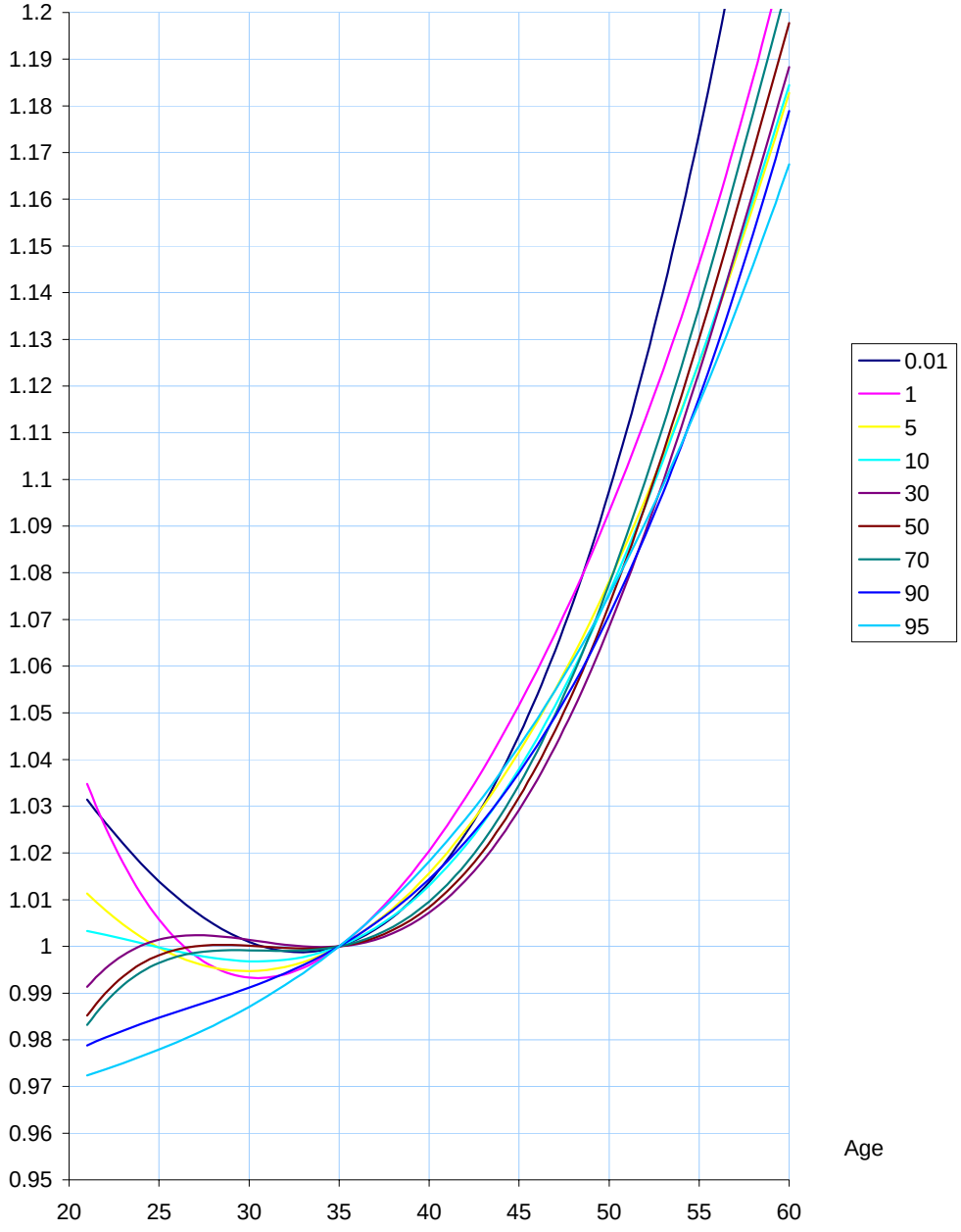

Figure 5M
Time(Age)/Time(35)
Selected Percentiles
Men

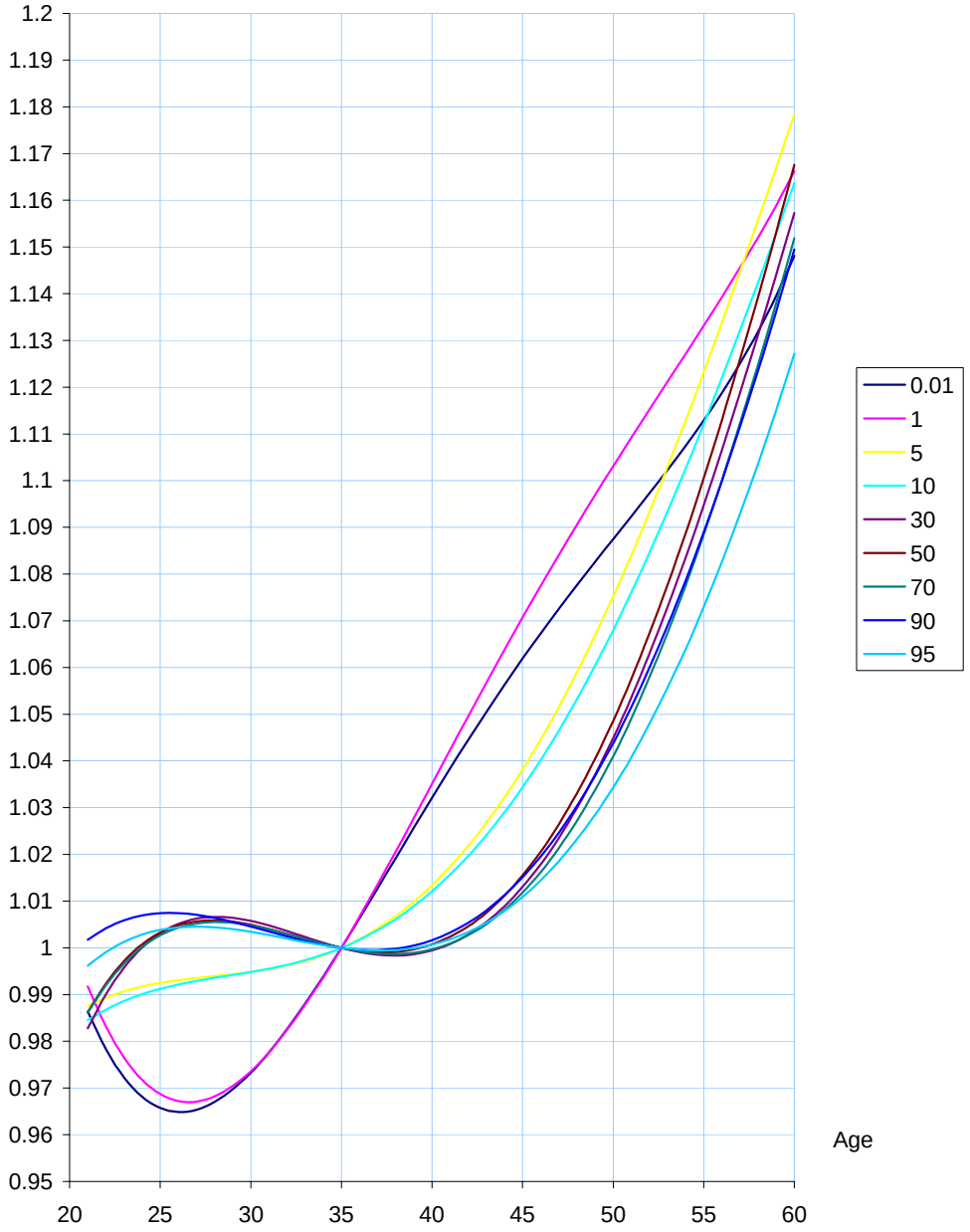

Figure 6W

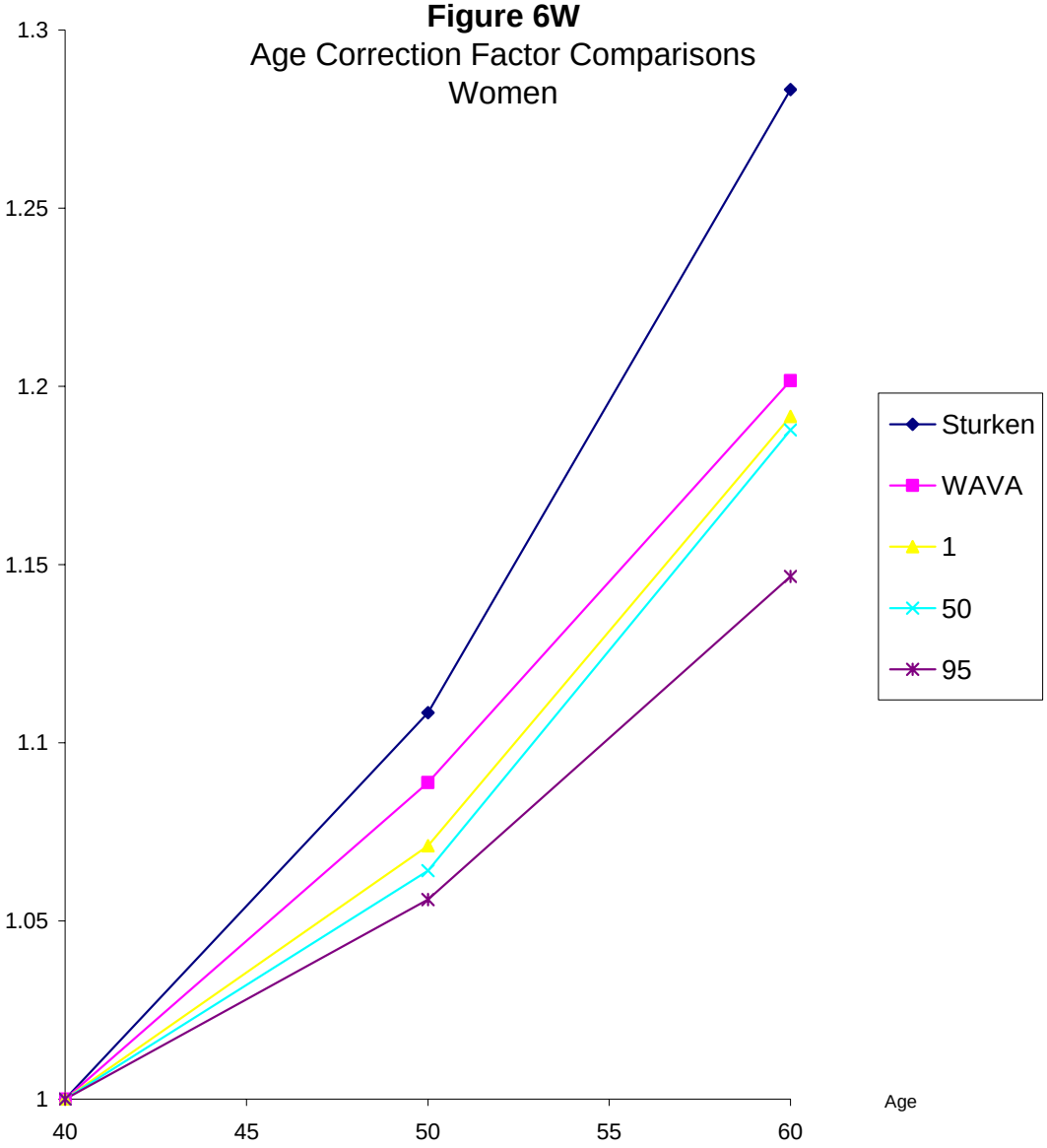
Age Correction Factor Comparisons
Women
1.3
1.25
1.2
1.15
1.1
1.05
1
40
45
50
55
60
Age
Sturken
WAVA
1
50
95

Figure 6M
Age Correction Factor Comparisons
Men

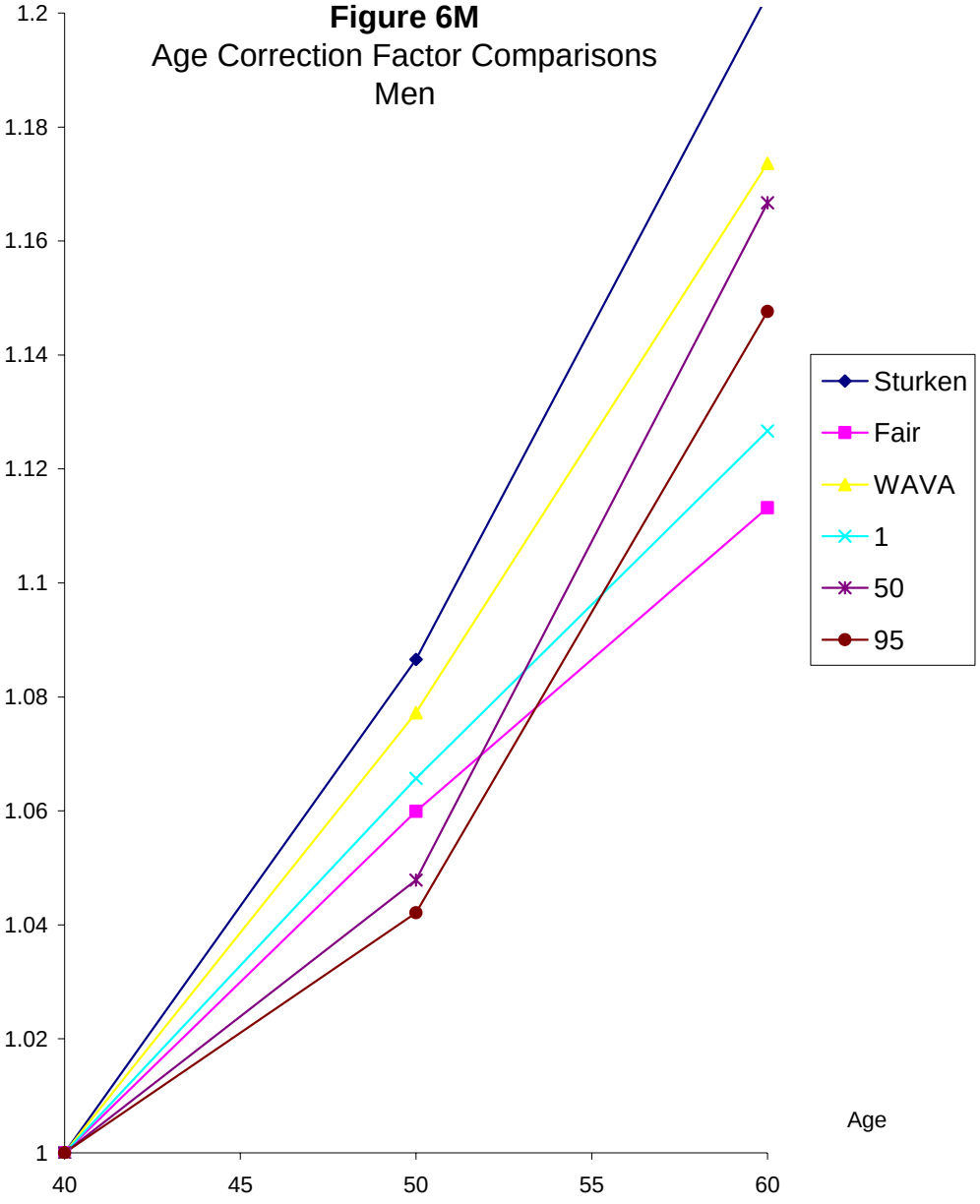

Table 1W
Women Estimates

Percentile		age	age^2	age^3	age^4	y2000	y2001	_cons
0.1	Coeff	-92.8	0.573	0.008	0.0002	269	457	11909
	t	-0.28	0.04	0.04	0.18	2.62	4.52	3.95
1	Coeff	-720	22.87	-0.315	0.0018	381	345	19166
	t	-1.92	1.61	-1.38	1.38	3.97	3.63	5.42
5	Coeff	-166	3.041	-0.008	0.0001	177	108	14981
	t	-1.01	0.49	-0.08	0.1	4.11	2.54	9.69
10	Coeff	244.6	-12.2	0.231	-0.0013	166	98	11623
	t	1.53	-1.98	2.29	-2.16	4.19	2.48	7.81
30	Coeff	1053	-41.86	0.688	-0.0038	207	82	5446
	t	5.7	-5.88	5.9	-5.49	4.62	1.86	3.15
50	Coeff	1249	-49.11	0.804	-0.0044	122	12	4822
	t	7.58	-7.71	7.65	-7.05	3.24	0.32	3.14
70	Coeff	1346	-52.91	0.866	-0.0047	176	-16	5222
	t	5.89	-5.96	5.87	-5.36	3.63	-0.33	2.47
90	Coeff	272.6	-10	0.152	-0.0005	287	-126	16936
	t	0.89	-0.86	0.81	-0.47	4.11	-1.82	5.82
95	Coeff	68.28	-2.577	0.053	-0.0001	659	-176	19763
	t	0.15	-0.15	0.19	-0.07	6.7	-1.8	4.57

n=35,374

Table 1M

Men Estimates

Percentile		age	age^2	age^3	age^4	y2000	y2001	_cons
0.1	Coeff	-1052	39	-0.60	0.0034	268	212	18368
	t	-3.69	3.78	-3.78	3.89	2.99	2.31	6.51
1	Coeff	-1038	37	-0.54	0.0029	153	188	19439
	t	-6.13	5.94	-5.56	5.3	2.87	3.54	11.78
5	Coeff	397	-17	0.29	-0.0016	64	55	7252
	t	3.23	-3.67	4.06	-4.03	1.75	1.51	6.04
10	Coeff	421	-17	0.29	-0.0016	65	66	7570
	t	3.36	-3.68	3.96	-3.83	1.86	1.90	6.21
30	Coeff	1318	-50	0.78	-0.0041	115	80	771
	t	9.9	-10.04	9.9	-9.19	3.28	2.30	0.60
50	Coeff	1251	-47	0.73	-0.0038	94	36	2640
	t	9.35	-9.45	9.21	-8.32	2.85	1.11	2.04
70	Coeff	1256	-47	0.71	-0.0036	15	21	3965
	t	7.73	-7.65	7.28	-6.37	0.40	0.55	2.54
90	Coeff	887	-34	0.54	-0.0027	8	33	10183
	t	3.11	-3.24	3.17	-2.74	0.13	0.51	3.70
95	Coeff	846	-31	0.47	-0.0022	15	-43	11528
	t	2.56	-2.54	2.36	-1.9	0.20	-0.58	3.63

n=52,149