PIRLS-test.R

Carl Schmertmann

Mon Aug 12 16:51:17 2019

```r
library(tidyverse)
```

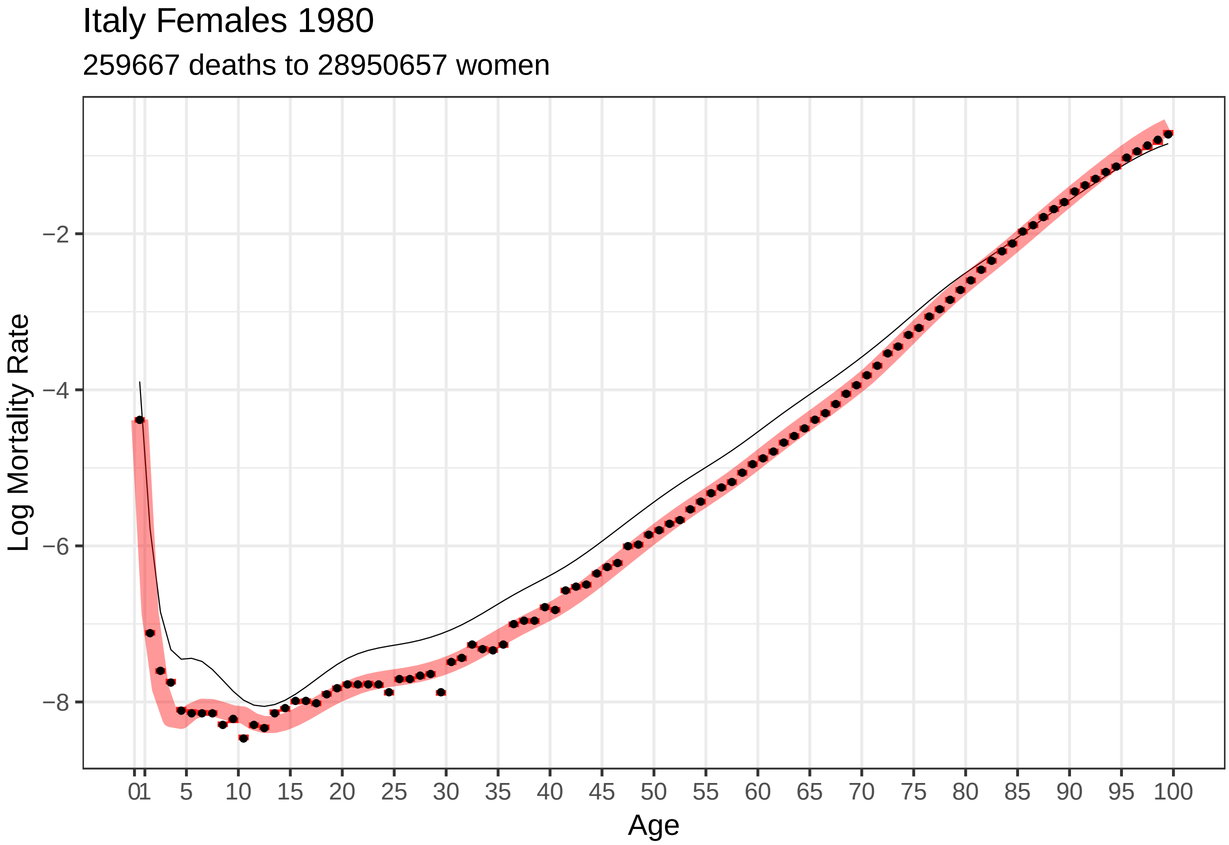
```
## Warning: package 'tidyverse' was built under R version 3.5.3

## -- Attaching packages -------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.5.2

## Warning: package 'readr' was built under R version 3.5.2

## Warning: package 'purrr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## -- Conflicts ----------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
rm(list=ls())

#============== DATA ===========================
## Italy 1980 Female data from HMD (true e0 from HMD is 77.42)

ITA = read.csv(file='ITA-Female-1980.csv')

# standard schedule = smoothed CAN females 1959 log rates at 0,1,...99
std =c(-3.8933, -5.7776, -6.8474, -7.3298, -7.4519, -7.4408, -7.4807,
       -7.5845, -7.7219, -7.8628, -7.9771, -8.041, -8.0568, -8.0329,
       -7.9779, -7.9004, -7.8088, -7.7101, -7.6113, -7.5195, -7.4415,
       -7.3823, -7.3393, -7.308, -7.2837, -7.2619, -7.238, -7.2082,
       -7.1711, -7.1264, -7.0735, -7.0118, -6.9414, -6.8648, -6.7849,
       -6.7047, -6.6272, -6.5544, -6.4845, -6.4147, -6.3423, -6.2644,
       -6.1791, -6.0872, -5.9904, -5.8903, -5.7887, -5.6869, -5.586,
       -5.4866, -5.3895, -5.2953, -5.205, -5.1186, -5.0347, -4.9513,
       -4.8664, -4.778, -4.6847, -4.5877, -4.4887, -4.3895, -4.2918,
       -4.1969, -4.1041, -4.0122, -3.9199, -3.8261, -3.7297, -3.6303,
       -3.5279, -3.4221, -3.3129, -3.2004, -3.0861, -2.9716, -2.8589,
       -2.7497, -2.6457, -2.5482, -2.4556, -2.3659, -2.2771, -2.187,
       -2.0942, -1.9991, -1.9027, -1.8062, -1.7105, -1.6164, -1.5242,
       -1.434, -1.3458, -1.2596, -1.1758, -1.0958, -1.0212, -0.9535,
       -0.8944, -0.8455)



#####################################################
```

```r
# note that this sources TOPALS_fit.R (the grouped version)
# rather than TOPALS_fit function.R (the single-year version)

source('TOPALS_fit.R')

## plotting function
show_fit = function(fit, std, true_schedule, fit_color='red') {

  df_grouped = data.frame(
    L = fit$L,
    U = fit$U,
    N = fit$N,
    D = fit$D
  ) %>%
    mutate(logmx_obs = log(D/N))


  df_single  = data.frame(
    age=seq(std)-0.5,
    std = std,
    logmx_true = true_schedule,
    logmx_fit  = myfit$logm
  )

  this_plot =
    ggplot(data = df_single, aes(x=age,y=logmx_true)) +
    geom_line(aes(x=age,y=std), color='black', lwd=0.2) +
    geom_line(aes(x=age,y=logmx_fit), color=fit_color, lwd=3, alpha=.40) +
    geom_segment(data=df_grouped,aes(x=L,xend=U,
                                     y=logmx_obs,
                                     yend=logmx_obs),
                 color=fit_color,lwd=1, alpha=.90) +
    geom_point(size=0.80) +
    labs(x='Age',y='Log Mortality Rate',
         title='Italy Females 1980',
         subtitle = paste(sum(D),'deaths to',round(sum(N)),'women')) +
    scale_x_continuous(breaks=c(0,1,seq(5,100,5)),minor_breaks = NULL) +
    theme_bw()

  print(this_plot)
} # show_fit

# trapez approx of life expectancy from a logmx schedule over ages 0..99
e0 = function(logmx) {
  mx = exp(logmx)
  px = exp(-mx)
  lx = c(1,cumprod(px))
  return( sum(head(lx,-1) + tail(lx,-1)) / 2)
}

#=======================================
# FULL DATASET WITH 1-YEAR GROUPS
#=======================================
## full dataset: 0,1,...99
N = ITA$N[1:100]
D = ITA$D[1:100]

L = 0:99
U = 1:100

myfit = TOPALS_fit(D=D, N=N, std=std,
                   group_lower_age = L,
                   group_upper_age = U,
                   details=TRUE)
```

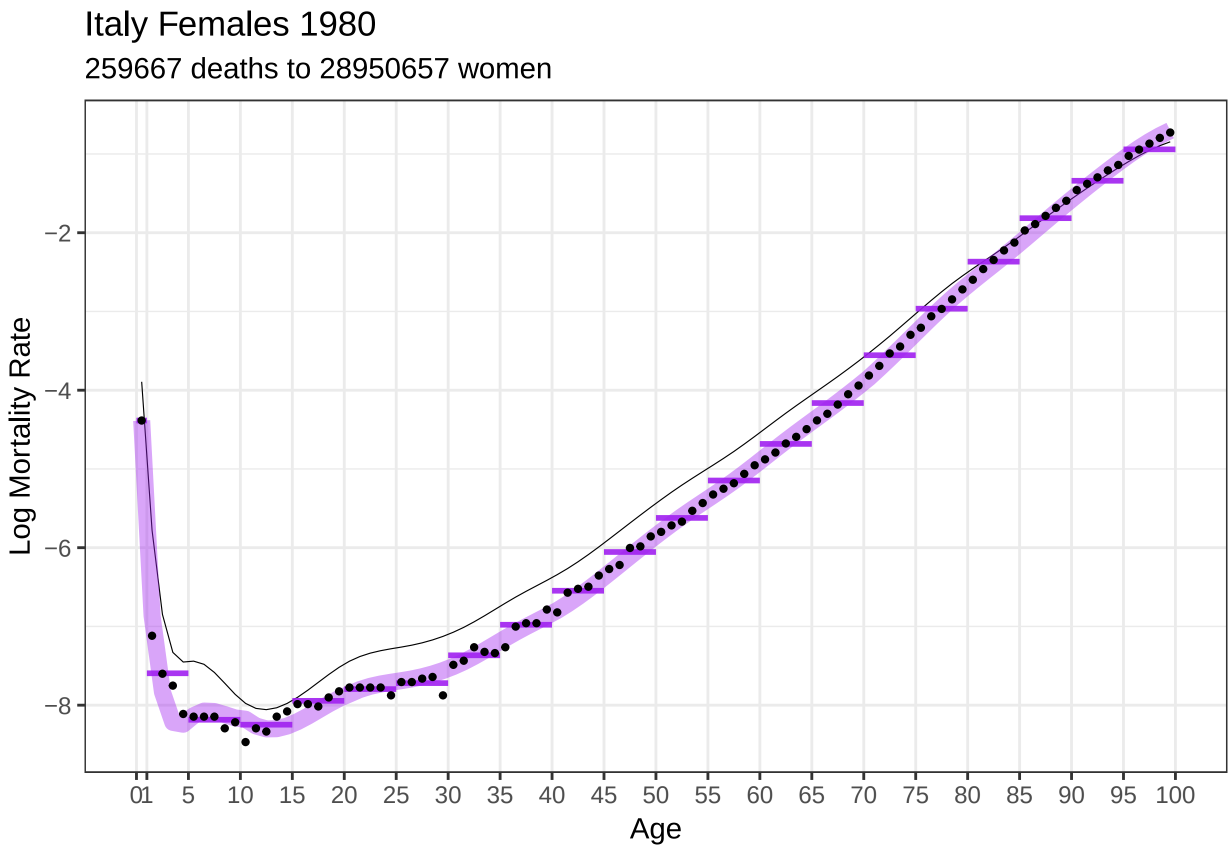
```
## Loading required package: splines
```

```r
show_fit( myfit, std, true_schedule = ITA$logmx[1:100],
          fit_color = 'red')
```

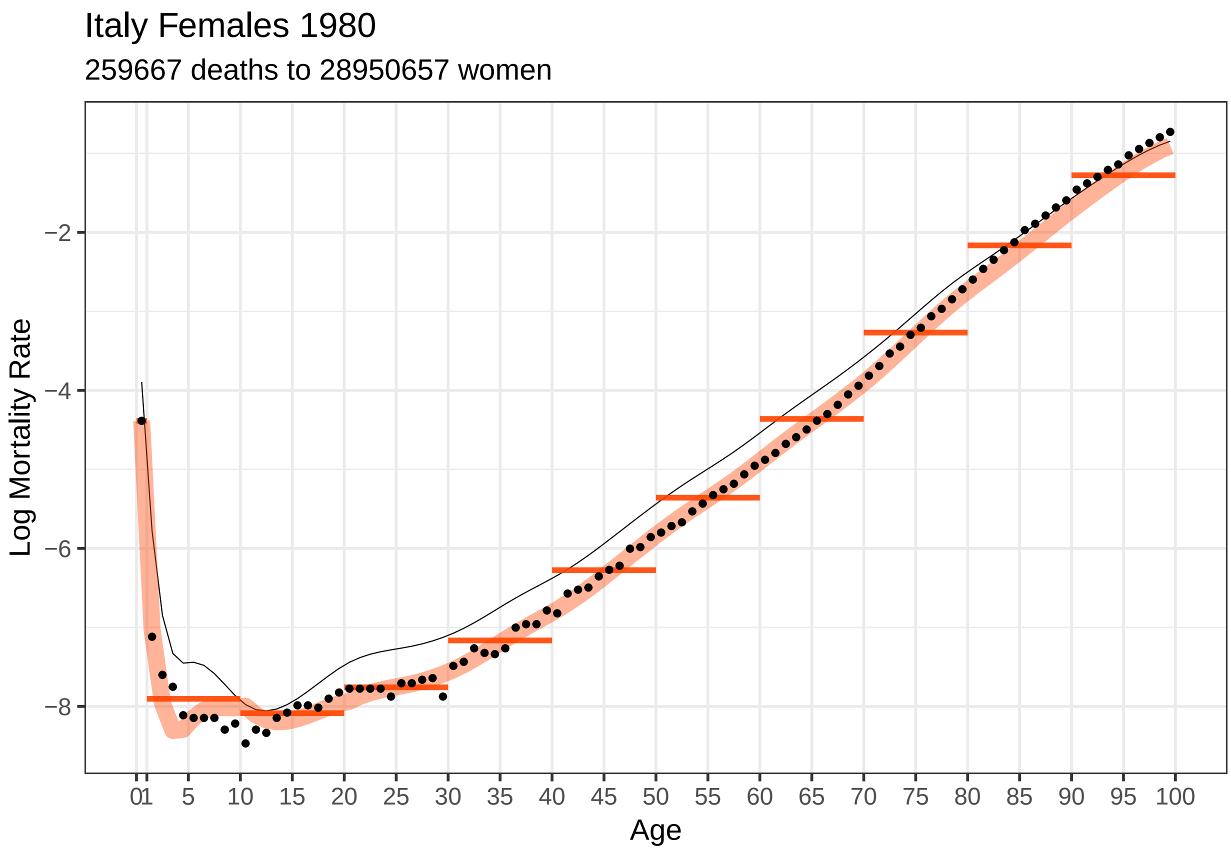


```r
#======================================
# FULL DATASET WITH 5-YEAR GROUPS
#======================================

L = c(0,1,seq(5,95,5))
U = c( tail(L,-1), 100)

N = sapply(seq(L), function(g) {
             tmp = sum(filter(ITA, L[g] <= age, age < U[g])$N)
```

```
           } )

D = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$D)
} )

myfit = TOPALS_fit(D=D, N=N, std=std,
                   group_lower_age = L,
                   group_upper_age = U,
                   details=TRUE)

show_fit( myfit, std, true_schedule = ITA$logmx[1:100],
          fit_color = 'purple')
```



```
#=====================================
# FULL DATASET WITH 10-YEAR GROUPS
#=====================================

L = c(0,1,seq(10,90,10))
U = c( tail(L,-1), 100)

N = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$N)
} )

D = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$D)
} )

myfit = TOPALS_fit(D=D, N=N, std=std,
                   group_lower_age = L,
                   group_upper_age = U,
                   details=TRUE)

show_fit( myfit, std, true_schedule = ITA$logmx[1:100],
          fit_color = 'orangered')
```

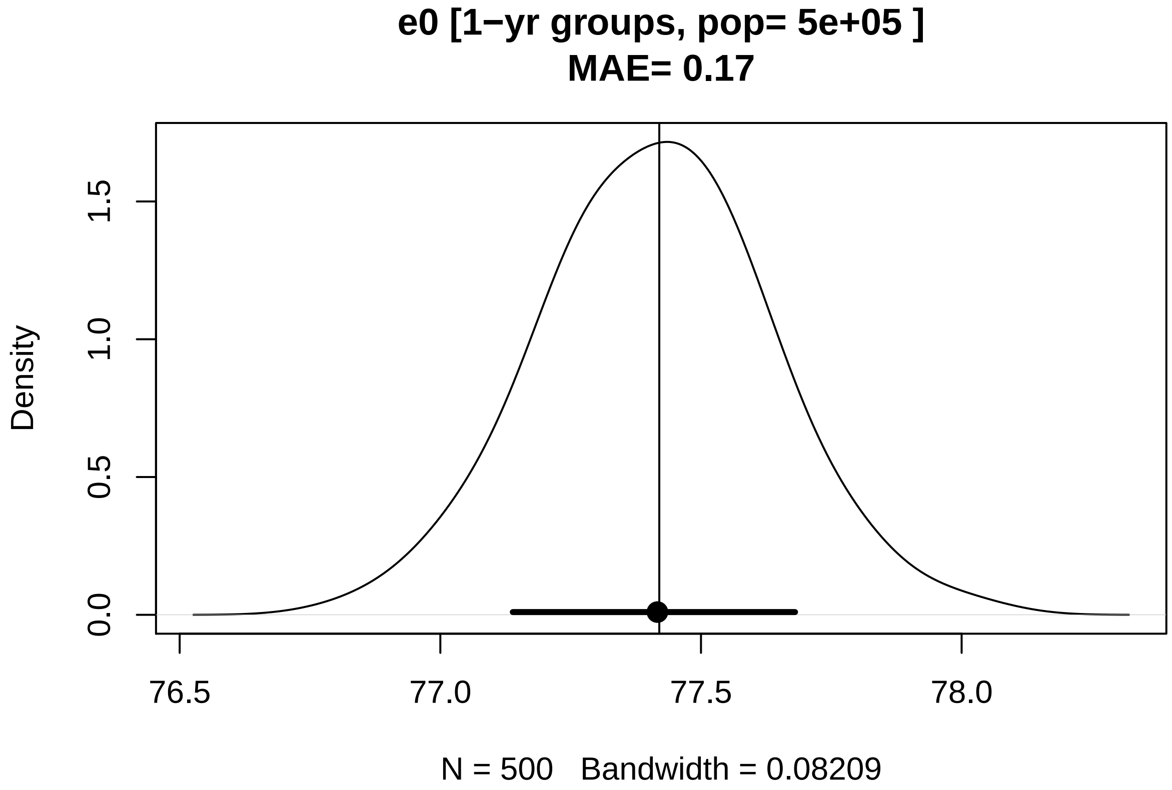


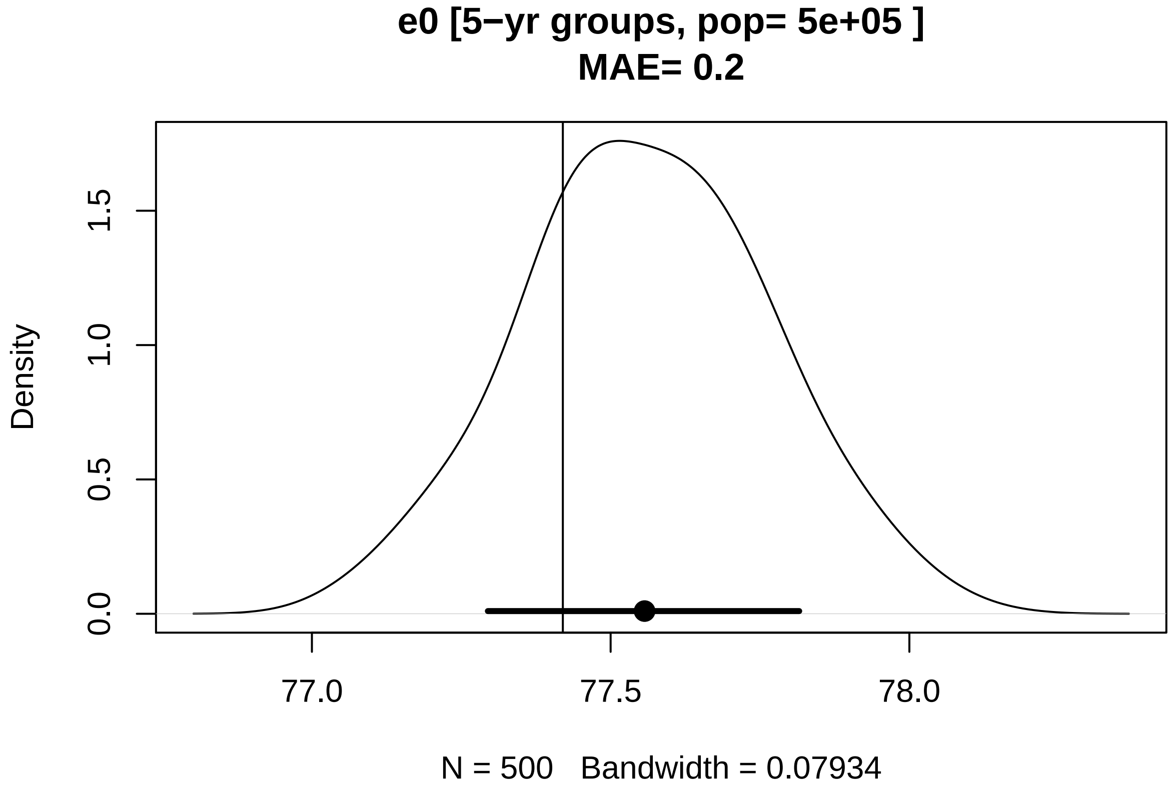
```
#  SMALL POPULATION SIMULATIONS
```

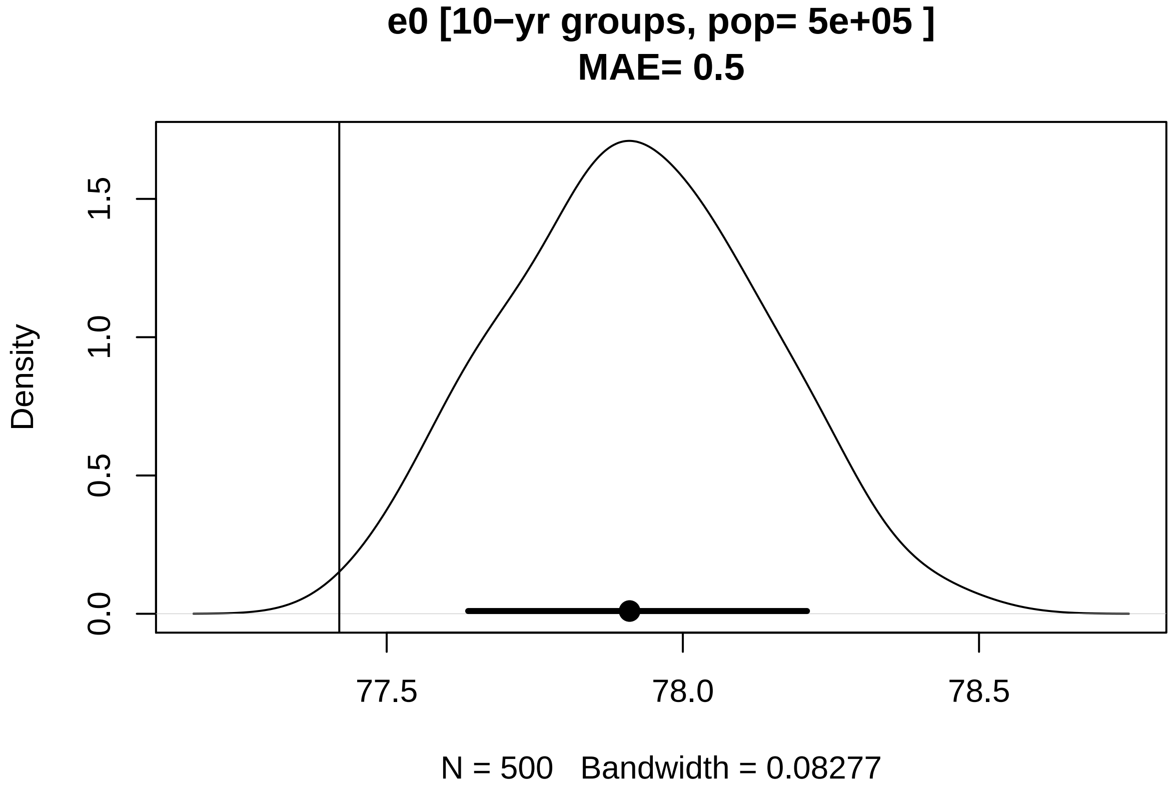
```
for (target_pop in c(500000, 100000, 10000, 1000)) {

    #=======================================
    # SMALL POPULATION DATASETS WITH 1-YEAR GROUPS
    #=======================================

    L = 0:99
    U = 1:100

    bigN = sapply(seq(L), function(g) {
      tmp = sum(filter(ITA, L[g] <= age, age < U[g])$N)
    } )
```

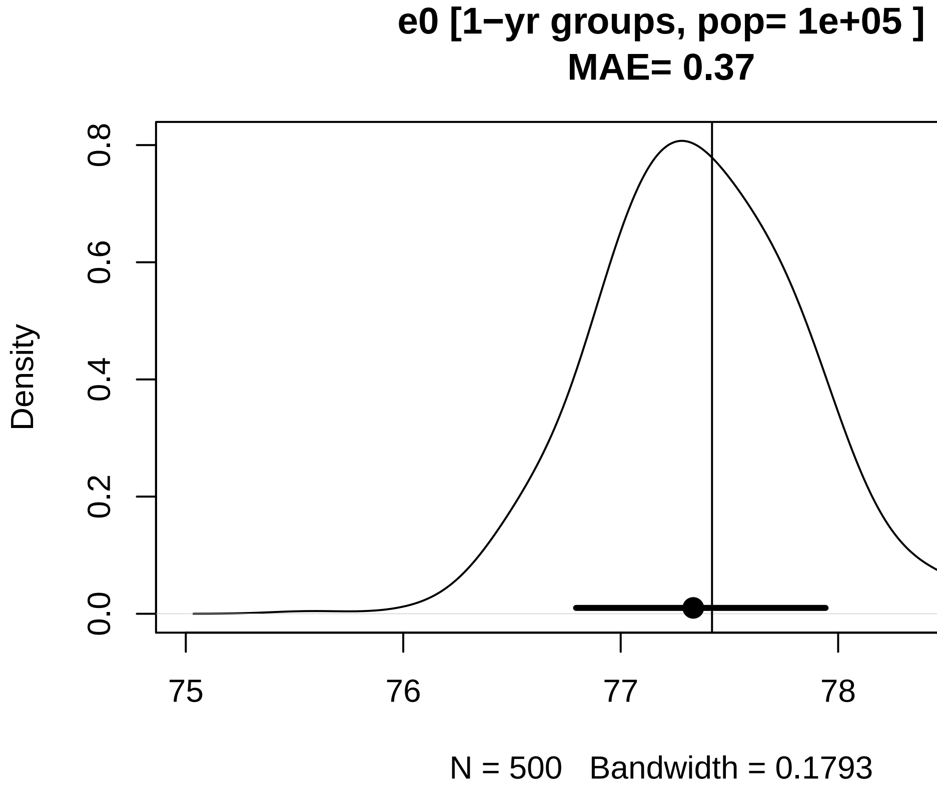
```
bigD = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$D)
} )

nsim = 500
e    = rep(NA,nsim)

N          = bigN * target_pop/sum(bigN)

for (i in 1:nsim) {
  D = rpois(length(N), N * bigD/bigN)

  myfit = TOPALS_fit(D=D, N=N, std=std,
                     group_lower_age = L,
                     group_upper_age = U,
                     details=TRUE)

  e[i] = e0(myfit$logm)
}

MAE = round(mean( abs(e-77.42)),2)

plot( density(e, adj=1.5), main=paste('e0 [1-yr groups, pop=',target_pop,']\nMAE=',MAE))

qq = quantile(e, probs=c(.10,.50,.90))
abline(v=77.42)
points( qq['50%'], .01, pch=16, cex=1.5)
segments( x0=qq['10%'], y0=0.01, x1=qq['90%'], y1=0.01, lwd=3)


#=======================================
# SMALL POPULATION DATASETS WITH 5-YEAR GROUPS
#=======================================

L = c(0,1,seq(5,95,5))
U = c( tail(L,-1), 100)

bigN = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$N)
} )

bigD = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$D)
} )

nsim = 500
e    = rep(NA,nsim)

N          = bigN * target_pop/sum(bigN)

for (i in 1:nsim) {
  D = rpois(length(N), N * bigD/bigN)
  myfit = TOPALS_fit(D=D, N=N, std=std,
                     group_lower_age = L,
                     group_upper_age = U,
                     details=TRUE)

  e[i] = e0(myfit$logm)
}

MAE = round(mean( abs(e-77.42)),2)

plot( density(e, adj=1.5), main=paste('e0 [5-yr groups, pop=',target_pop,']\nMAE=',MAE))
qq = quantile(e, probs=c(.10,.50,.90))
abline(v=77.42)
points( qq['50%'], .01, pch=16, cex=1.5)
segments( x0=qq['10%'], y0=0.01, x1=qq['90%'], y1=0.01, lwd=3)


#=======================================
# SMALL POPULATION DATASETS WITH 10-YEAR GROUPS
#=======================================

L = c(0,1,seq(10,90,10))
U = c( tail(L,-1), 100)

bigN = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$N)
} )

bigD = sapply(seq(L), function(g) {
  tmp = sum(filter(ITA, L[g] <= age, age < U[g])$D)
} )

nsim = 500
e    = rep(NA,nsim)

N          = bigN * target_pop/sum(bigN)

for (i in 1:nsim) {
  D = rpois(length(N), N * bigD/bigN)

  myfit = TOPALS_fit(D=D, N=N, std=std,
                     group_lower_age = L,
                     group_upper_age = U,
                     details=TRUE)

  e[i] = e0(myfit$logm)
}

MAE = round(mean( abs(e-77.42)),2)

plot( density(e, adj=1.5), main=paste('e0 [10-yr groups, pop=',target_pop,']\nMAE=',MAE))
qq = quantile(e, probs=c(.10,.50,.90))
abline(v=77.42)
```

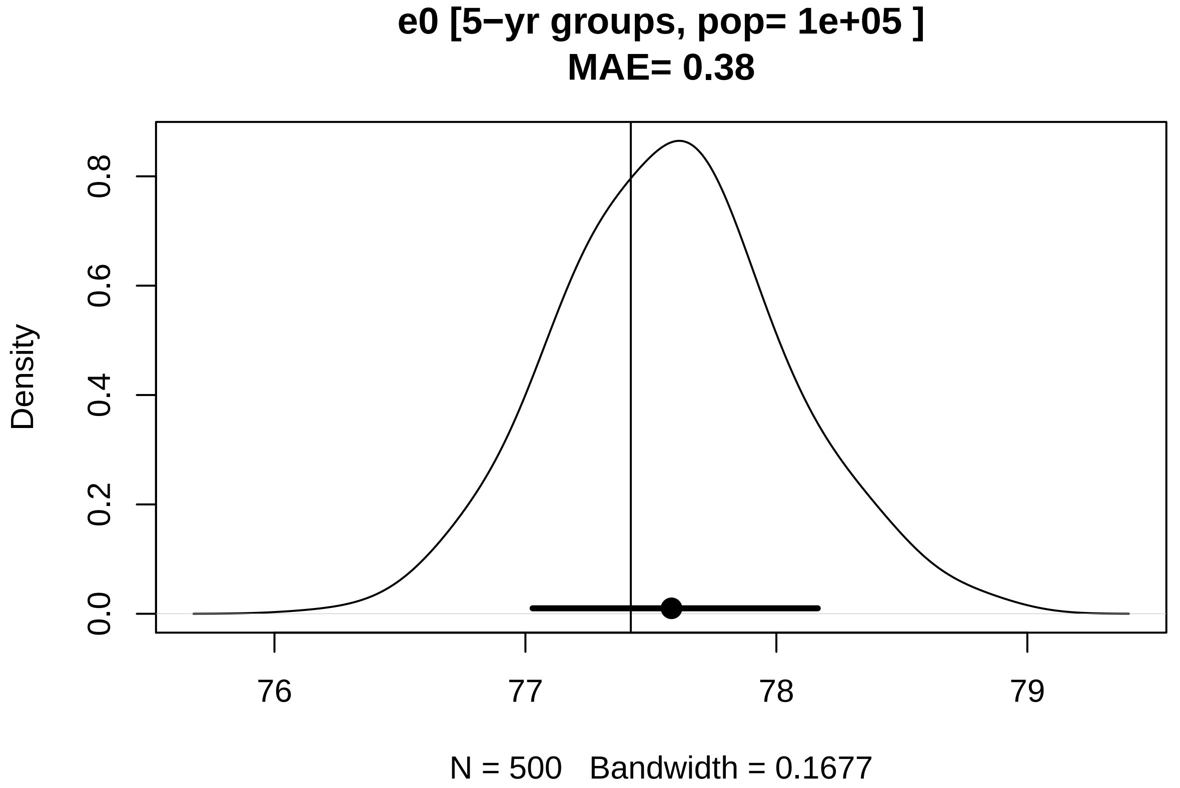
```
    points( qq['50%'], .01, pch=16, cex=1.5)
    segments( x0=qq['10%'], y0=0.01, x1=qq['90%'], y1=0.01, lwd=3)


} # for target_pop
```

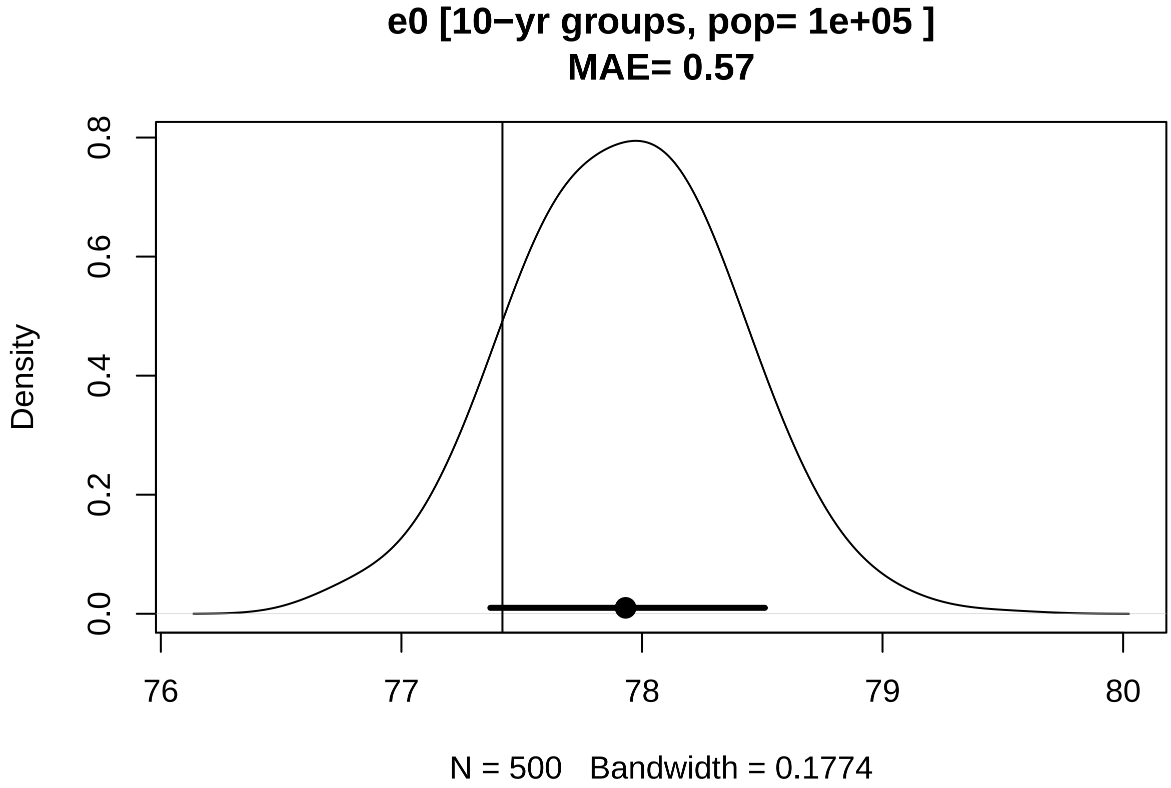


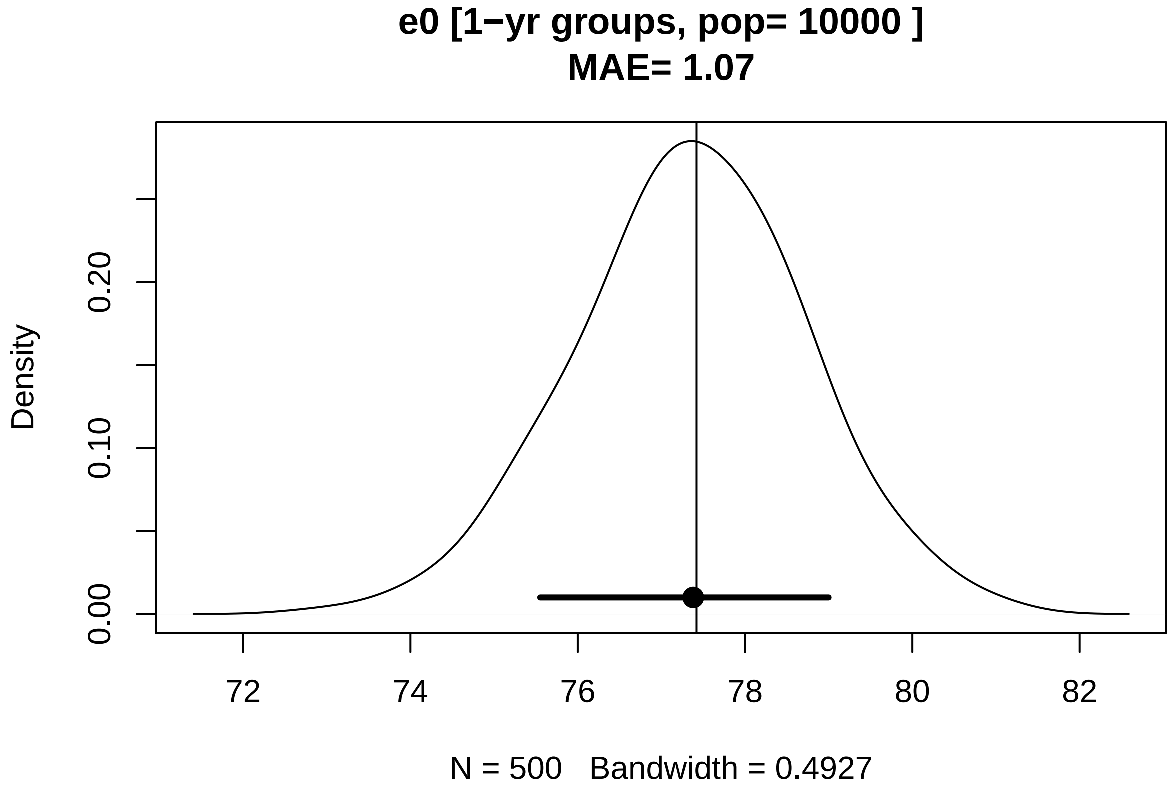

e0 [5−yr groups, pop= 5e+05]
MAE= 0.2
Density
0.0
0.5
1.0
1.5
77.0
77.5
78.0
N = 500   Bandwidth = 0.07934

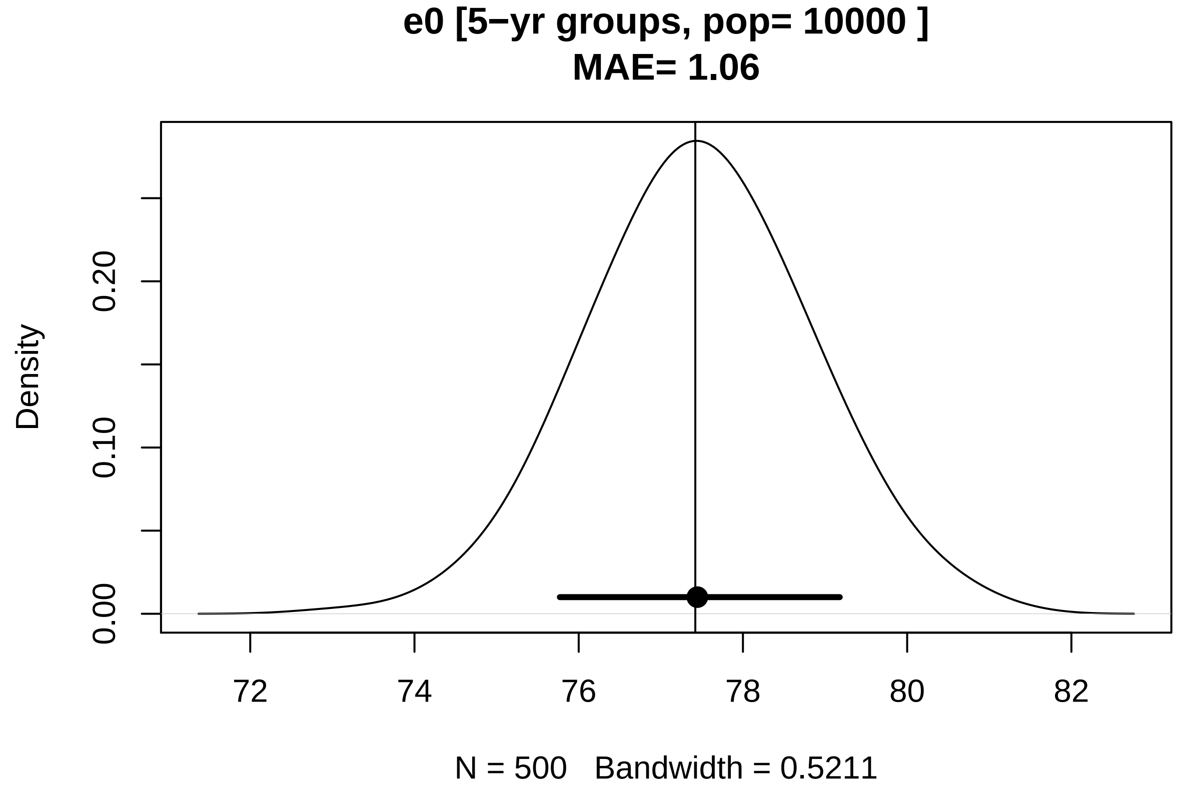

e0 [10−yr groups, pop= 5e+05]
MAE= 0.5
Density
0.0
0.5
1.0
1.5
77.5
78.0
78.5
N = 500 Bandwidth = 0.08277

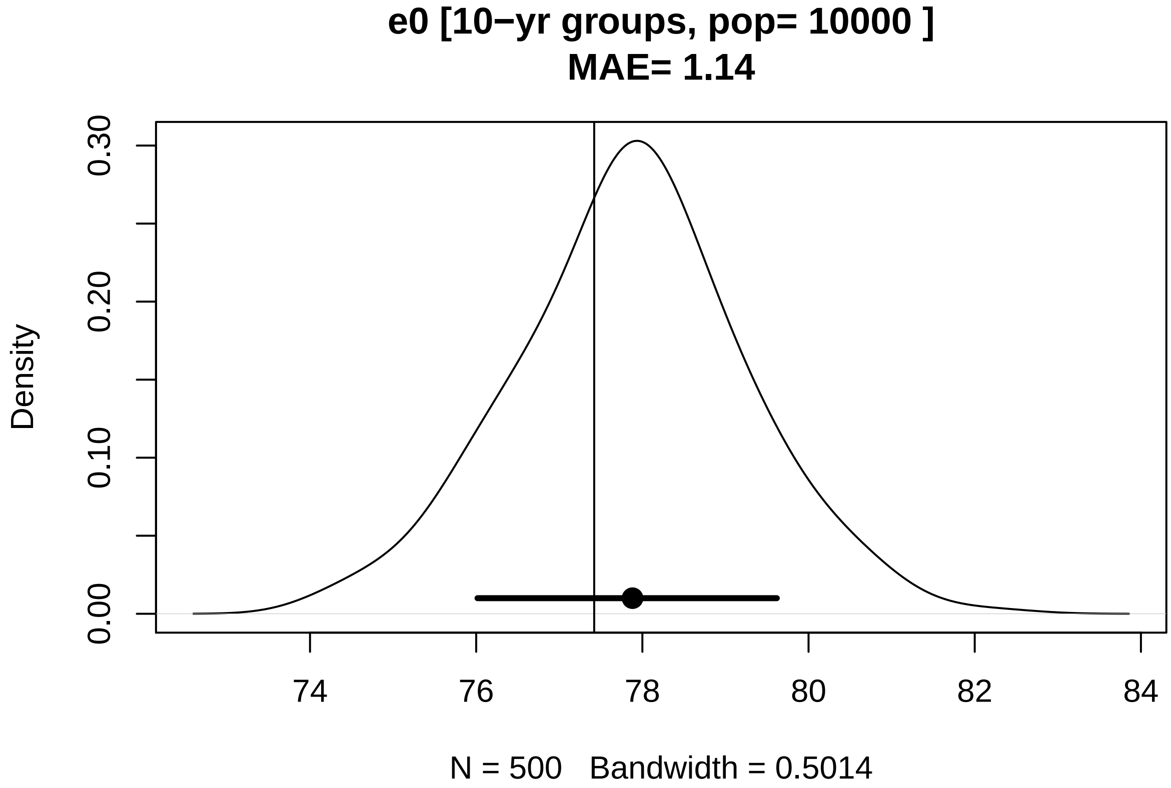

e0 [1−yr groups, pop= 1e+05]
MAE= 0.37
Density
0.0
0.2
0.4
0.6
0.8
75
76
77
78
79
N = 500   Bandwidth = 0.1793

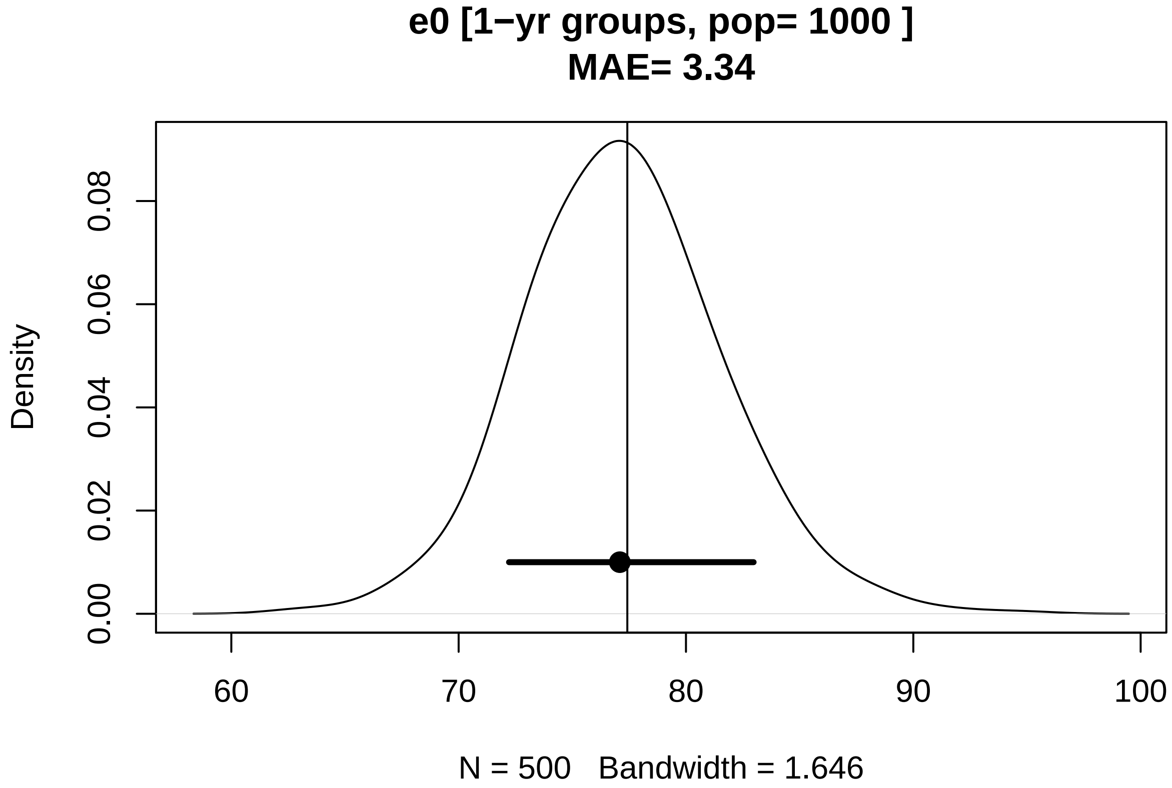

e0 [5−yr groups, pop= 1e+05]
MAE= 0.38
Density
0.0
0.2
0.4
0.6
0.8
76
77
78
79
N = 500   Bandwidth = 0.1677

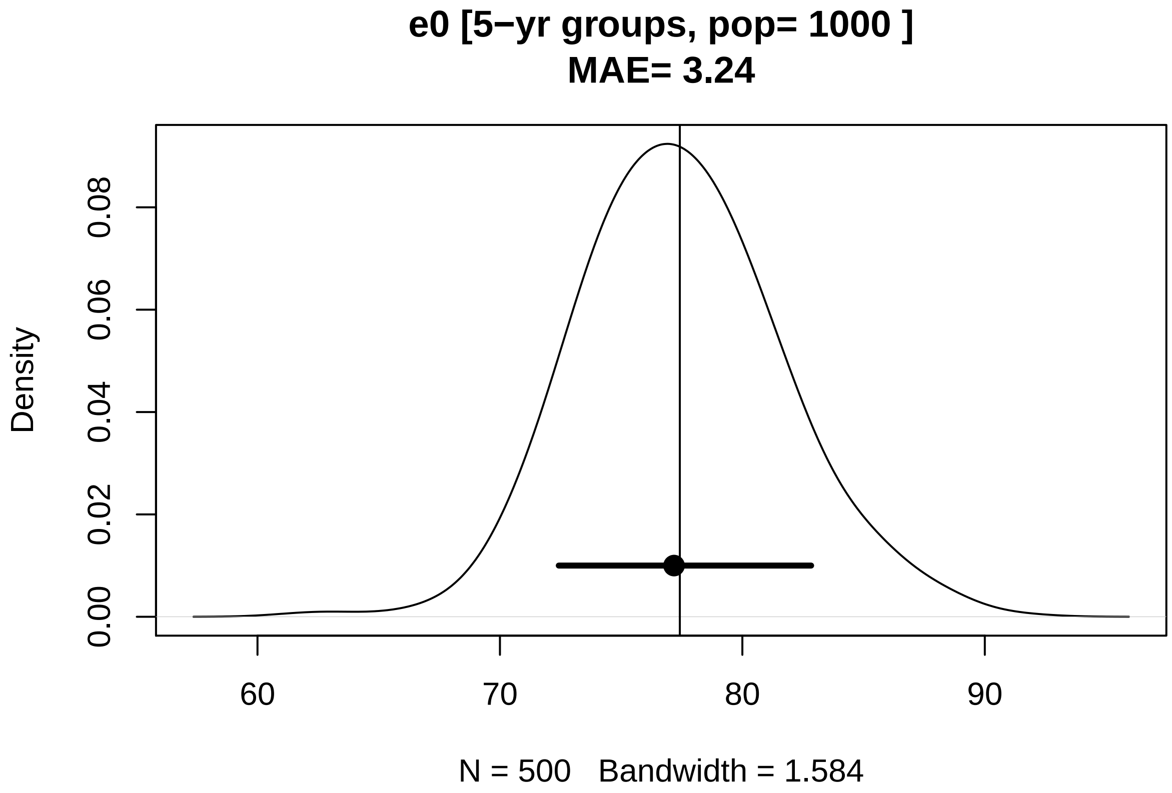

e0 [10−yr groups, pop= 1e+05]
MAE= 0.57
Density
0.0
0.2
0.4
0.6
0.8
76
77
78
79
80
N = 500   Bandwidth = 0.1774

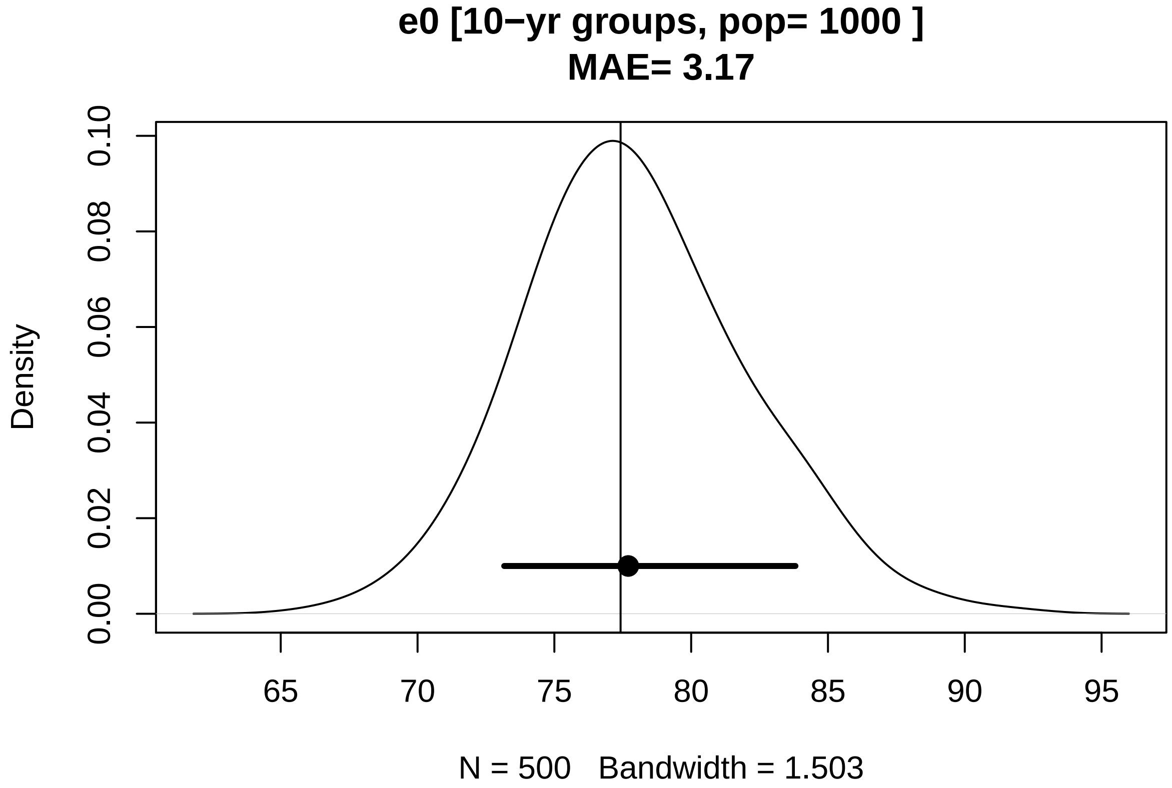

e0 [1−yr groups, pop= 10000]
MAE= 1.07
Density
0.00
0.10
0.20
72
74
76
78
80
82
N = 500   Bandwidth = 0.4927


e0 [5−yr groups, pop= 10000]
MAE= 1.06
Density
0.00
0.10
0.20
72
74
76
78
80
82
N = 500   Bandwidth = 0.5211


e0 [10−yr groups, pop= 10000]
MAE= 1.14
Density
0.00
0.10
0.20
0.30
74
76
78
80
82
84
N = 500   Bandwidth = 0.5014


e0 [1−yr groups, pop= 1000]
MAE= 3.34
Density
0.00
0.02
0.04
0.06
0.08
60
70
80
90
100
N = 500 Bandwidth = 1.646


e0 [5−yr groups, pop= 1000]
MAE= 3.24
Density
0.00
0.02
0.04
0.06
0.08
60
70
80
90
N = 500 Bandwidth = 1.584


e0 [10−yr groups, pop= 1000]
MAE= 3.17
Density
0.00
0.02
0.04
0.06
0.08
0.10
65
70
75
80
85
90
95
N = 500   Bandwidth = 1.503