Title: A statistical approach to distinguish telomere elongation from error in longitudinal datasets

Authors: Mirre J.P. Simons[1*], Gert Stulp[2] & Shinichi Nakagawa[3]

Affiliations:

[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

[2]Department of Sociology, University of Groningen, Groningen, 9712 TG – 31, The Netherlands

[3]Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand

[*]Author for correspondence: Mirre J.P. Simons, mirresimons@gmail.com, Tel: +441142220123, Fax: +441142220002

Type of paper: 'Methods'

Abstract

Telomere length and the rate of telomere attrition vary between individuals and have been interpreted as the rate at which individuals have aged. The biology of telomeres dictates shortening with age, although telomere elongation with age has repeatedly been observed within a minority of individuals in several populations. These findings have been attributed to error, rather than actual telomere elongation, restricting our understanding of its possible biological significance. Here we present a method to distinguish between error and telomere elongation in longitudinal datasets, which is easy to apply and has few assumptions. Using simulations, we show that the method has considerable statistical power (> 80%) to detect even a small proportion (6.7%) of TL increases in the population, within a relatively small sample (N = 200), while maintaining the standard level of Type I error rate ($\alpha \leq 0.05$).

Keywords: telomere length, statistics, telomere shortening, within individual, aging, human

Telomeres are DNA sequence repeats at the end of chromosomes. These repeats shorten at each cell replication or by damage, and critical telomere lengths lead to cellular senescence, apoptosis and/or genome instability (Riethman 2008). These properties of telomeres suggest direct involvement in aging mechanisms, but telomere length (TL) may also be an indicator of the progression of aging within individuals and/or differences in aging between individuals (Mather et al. 2011; Riethman 2008). Indeed, short TL is associated with higher mortality risk in humans (Boonekamp et al. 2013) and other free-living animals (e.g. Barrett et al. 2013; Bize et al. 2009; Heidinger et al. 2012; Salomons et al. 2009). Yet, comparative analyses do not support that shorter telomeres dictate shorter lifespans between species (Gorbunova and Seluanov 2009). The rate at which telomeres shorten is also variable between individuals (e.g. Aviv et al. 2009; Nordfjäll et al. 2009) and higher rates of telomere attrition are associated with increased risk of mortality (Epel et al. 2009).

The biological properties of telomeres dictate shortening rather than lengthening in tissues in which telomeres are not actively elongated (Gorbunova and Seluanov 2009). Yet in the majority of studies TL increases are apparent within a small group of individuals. These elongations are often attributed to error (e.g. Aviv et al. 2009; Beaulieu et al. 2011; Bize et al. 2009; Chen et al. 2011; Ehrlenbach et al. 2009; Epel et al. 2009; Foote et al. 2011; Nordfjäll et al. 2009; Salomons et al. 2009; Shalev et al. 2012; Steenstrup et al. 2013) which is composed of *both* measurement error of TL and other unknown causes of within-individual variability (e.g. variation in TL of the tissue sampled). An alternative explanation is that telomeres do elongate in some individuals. To our knowledge, no statistical approach exists to distinguish telomere elongation from error within longitudinal studies. Here we present a method, which is

59 easy to apply and has few assumptions. Using simulations, we show that this method
60 has considerable statistical power (> 80%), while it retains the standard level of Type
61 I error rate ($\alpha \leq 0.05$).
62
63 Our method first requires estimating variance due to measurement errors (error
64 variance) in two distinct ways related to two different assumptions: 1) TL increases
65 and/or decreases and 2) telomeres do not elongate. Under the first assumption, error
66 variance can be estimated in two steps. First, we estimate the residual variance for
67 each individual using an ordinary (least square) linear regression:
68 $$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i , \quad (1)$$
69 $$\varepsilon_i \sim N(0,\ \sigma_\varepsilon^2) , \quad (2)$$
70 where t_i is the ith time point at which TL, y_i, is measured (i = 1, 2,…, n; n is the
71 number of TL measurements and n > 2), β_0 is the intercept (TL at t = 0), β_1 is the
72 slope (regression coefficient for t), and ε_i is the ith residual value. Residuals are
73 normally distributed (N) with a variance of σ_ε^2. If $\sigma_{\varepsilon j}^2$ represents the jth individual's
74 residual variance (j = 1, 2,…, N; N is the number of individuals in a study), then, an
75 overall error variance estimate of TL ($\bar{\sigma}_\varepsilon^2$) can be obtained by taking an average of
76 $\sigma_{\varepsilon j}^2$:
77 $$\bar{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{j=1}^{N} \sigma_{\varepsilon j}^2 . \quad (3)$$
78 Perhaps, more practically, Equation 3 can be re-written using the residual sum of
79 squares:
80 $$\bar{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{n_j - 2}\sum_{i=1}^{n} \varepsilon_{ij}^2 , \quad (4)$$

where n_j is the number of TL measurements n for jth individual and ε_{ij}^2 is the squared residual value for the ith time point for the jth individual (cf. Crawley 2005).

Under the second assumption (i.e. no telomere elongation), the measurement error variance ($\sigma_\varepsilon'^2$) can be obtained by:

$$\sigma_\varepsilon'^2 = \frac{1}{2(m-1)} \sum_{k=1}^{m} D_k^2 , \qquad (5)$$

where D_k^2 is the difference in TL between the initial and last measurements in the kth individuals that showed an increase in TL (k = 1, 2,…, m; m is the number of individuals whose TL elongated). When observed TL increases are not due to error, but consistent telomere elongation is present in the population, the largest increases of TL are between the first and the last measurement in time. Therefore to increase sensitivity of detecting telomere elongation we define telomere increases as the TL at the last measurement minus the TL at the first measurement per individual as in Equation 5. Note that the same equations can be used to ask the question whether telomere increases occur at any point in time in the population. A mathematical derivation of Equation 5 is given in the Appendix.

When the estimated error variance $\sigma_\varepsilon'^2$ (Equation 5) is larger than the error variance $\bar{\sigma}_\varepsilon^2$, when TL *is* allowed to increase or decrease (Equation 4), the hypothesis that telomeres show no elongation in the sample can be rejected. Statistically, such a comparison can be achieved using a variance ratio test between $\sigma_\varepsilon'^2$ and $\bar{\sigma}_\varepsilon^2$. The ratio of these two estimated error variances should follow an F distribution, which can be written as:

104 $$\frac{\sigma'^2_\varepsilon}{\bar{\sigma}^2_\varepsilon} \sim F(m-1,\ N-1) \qquad (6)$$

105 where the *F* distribution is defined by two degrees of freedom (DF): 1) the numerator
106 DF is the number of observed TL increases minus 1 (*m*–1), and 2) the denominator
107 DF is the number of individuals in a study minus 1 (*N*–1) (Crawley 2007).
108
109 When telomere elongation is statistically detected within a population, the
110 identification of individuals within the population that are likely to show true telomere
111 elongation (i.e. not the resultant of measurement errors) can be identified using the
112 upper confidence limit (UCL) of $\bar{\sigma}^2_\varepsilon$ (Crawley 2007). The UCL of the 95%
113 confidence interval (note that the 95% here is rather arbitrary and can be changed
114 depending on the level of certainty required) can be written as:

115 $$97.5\%\ \text{UCL} = \frac{(N-1)\bar{\sigma}^2_\varepsilon}{\chi^2_{N-1(0.975)}} \qquad (7)$$

116 where $\chi^2_{N-1(0.975)}$ is the value at $p = 0.975$ of the χ^2 distribution defined by DF = *N*–1.
117 This UCL of $\bar{\sigma}^2_\varepsilon$ can be used to determine the normal distribution of the UCL of the
118 underlying measurement error distribution. Subsequently individual telomere
119 increases (note that the increases should be divided by 2 as in Equation 5, because the
120 TL increases are a result from the addition of two equal error distributions) that are at
121 the boundary of this normal distribution (with e.g. 95% confidence) can be looked up
122 with, for example, the function 'qnorm' (Wichura 1988) from R (R Development
123 Core Team 2011). These specific individuals can be selected for follow-up studies, to
124 examine biological and environmental correlates (see also the worked example
125 provided with the manuscript).
126

To investigate the statistical power of the approach proposed here, simulations were conducted in R (code is available upon request). Individual based data of 3 time points per individual were generated. Individuals were set to lose an average TL of 3 per time, which varied among individuals with a given standard deviation (labeled slope SD). At each time point TL was subject to error (labeled error SD). Simulations were run for different combinations of sample size (range 50-500), error and slope SDs (both range 1-5) and each simulation was run 1000 times. The resulting statistical power was calculated as the fraction of times the null hypothesis was rejected when it was actually false (Fig 1), in other words, if the method detected telomere elongation when true telomere elongation was present in the simulated data. As expected, power increased with lower error, larger sample size and higher incidence of telomere elongation in the sample (i.e. higher slope SD). The average proportion of individuals showing a 'real' positive slope in these simulations was 0.13%, 6.7%, 16%, 23% and 27% for slope SD 1, 2, 3, 4, 5 respectively. Note that 0.13% might not be a biologically relevant proportion of individuals that show true telomere elongation, yet in the continuum presented in figure 1. It does give an impression of sensitivity and reliance on outliers of the method presented, and for this reason we included it in our power simulations. The statistical approach presented here is thus able to detect telomere elongation of only a small proportion ($\geq$ 6.7%) of a relatively small sample (under 500 individuals) with considerable power. In addition, the chance of rejecting the hypothesis when it is actually true (Type I error) was simulated using a slope SD of 0 and without an average decrease in TL for a range of sample sizes (50-500) and error SD (of 3 and 4). Type I error rates were equal to the expected α, 5% (4.8% of 10,000 simulations) and were independent of sample size and error SD. Note that if there is an average decrease of TL across the population, type I error rates will be

much lower given that the decline of TL over time reduces the amount of increases due to error.

The formal test to distinguish true telomere elongation from error, described here, forms an incentive to measure individuals at least *three* times longitudinally. The detection of significant elongation of TL within a population will likely spur research into the mechanisms regulating telomere elongation and into specific properties or circumstances of the individuals that show true telomere elongation.

Online supplement

As a supplement we have added a worked example on simulated data in R code with the relevant explanations embedded in the code. We also included the results of a run of this script and its results as a supplement, thereby demonstrating all the necessary steps required in the analyses and also providing the necessary tools for researches to employ our method with ease.

Figures

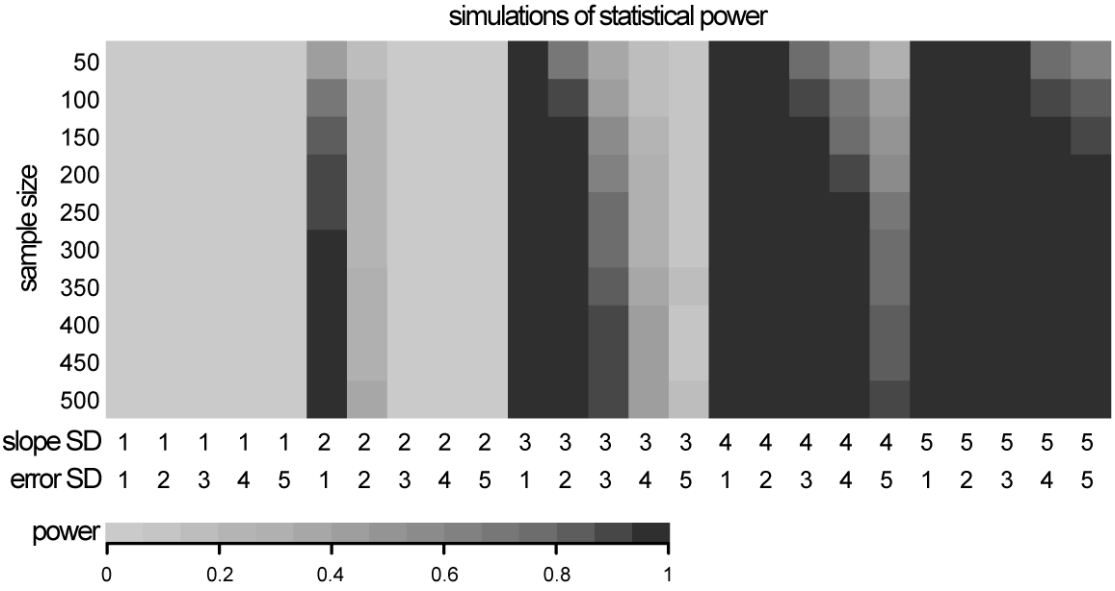

Fig 1 Result of the statistical power simulations. The statistical power (indicated by the grayscale, darker means higher power, the fraction of times the null hypothesis is rejected when it is actually false) is dependent on the sample size on the *y*-axis, and the error standard deviation (error SD) and slope standard deviation (slope SD), both depicted on the *x*-axis.

Acknowledgements

MJPS is supported by the Natural Environment Research Council (J024597/1) (United Kingdom). SN is supported by the Rutherford Discovery Fellowship (New Zealand). GS is supported by a grant by The Netherlands Organisation for Scientific Research (452-10-012), granted to M. Mills

References

Aviv A, Chen W, Gardner JP, Kimura M, Brimacombe M, Cao X, Srinivasan SR, Berenson GS (2009) Leukocyte telomere dynamics: longitudinal findings among young adults in the Bogalusa Heart Study. Am J Epidemiol 169:323–329.

Barrett ELB, Burke TA, Hammers M, Komdeur J, Richardson DS (2013) Telomere length and dynamics predict mortality in a wild longitudinal study. Molecular Ecology 22:249–259.

Beaulieu M, Reichert S, Le Maho Y, Ancel A, Criscuolo F (2011) Oxidative status and telomere length in a long-lived bird facing a costly reproductive event. Funct Ecol 25:577–585.

Bize P, Criscuolo F, Metcalfe NB, Nasir L, Monaghan P (2009) Telomere dynamics rather than age predict life expectancy in the wild. Proc R Soc B 276:1679–1683. doi:

Boonekamp JJ, Simons MJP, Hemerik L, Verhulst S (2013) Telomere length behaves as biomarker of somatic redundancy rather than biological age. Aging Cell

Chen W, Kimura M, Kim S, Cao X, Srinivasan SR, Berenson GS, Kark JD, Aviv A (2011) Longitudinal versus Cross-sectional Evaluations of Leukocyte Telomere Length Dynamics: Age-Dependent Telomere Shortening is the Rule. J Gerontol A Biol Sci Med Sci 66A:312–319.

Crawley MJ (2005) Statistics: an introduction using R. Wiley, Chichester.

Crawley MJ (2007) The R book. Wiley, Chichester,

Ehrlenbach S, Willeit P, Kiechl S, Willeit J, Reindl M, Schanda K, Kronenberg F, Brandstatter A (2009) Influences on the reduction of relative telomere length over 10 years in the population-based Bruneck Study: introduction of a well-controlled high-throughput assay. Int J Epidemiol 38:1725–1734.

Epel ES, Merkin SS, Cawthon R, Blackburn EH, Adler NE, Pletcher MJ, Seeman TE (2009) The rate of leukocyte telomere shortening predicts mortality from cardiovascular disease in elderly men. Aging (Albany NY) 1:81–88.

Foote CG, Gault EA, Nasir L, Monaghan P (2011) Telomere dynamics in relation to

209 early growth conditions in the wild in the lesser black-backed gull. J Zool
210 283:203–209.

211 Gorbunova V, Seluanov A (2009) Coevolution of telomerase activity and body mass
212 in mammals: From mice to beavers. Mech Ageing Dev 130:3–9.

213 Heidinger BJ, Blount JD, Boner W, Griffiths K, Metcalfe NB, Monaghan P (2012)
214 Telomere length in early life predicts lifespan. P Natl Acad Sci U S A 109:1743–
215 1748.

216 Mather KA, Jorm AF, Parslow RA, Christensen H (2011) Is telomere length a
217 biomarker of aging? A review. J Gerontol A Biol Sci Med Sci 66:202–213.

218 Nordfjäll K, Svenson U, Norrback K-F, Adolfsson R, Lenner P, Roos G (2009) The
219 Individual Blood Cell Telomere Attrition Rate Is Telomere Length Dependent.
220 PLoS Genet 5:e1000375.

221 Riethman H (2008) Human telomere structure and biology. Annu Rev Genomics Hum
222 Genet 9:1–19.

223 Salomons HM, Mulder GA, van de Zande L, Haussmann MF, Linskens MHK (2009)
224 Telomere shortening and survival in free-living corvids. Proc R Soc B 276:3157–
225 3165.

226 Shalev I, Moffitt TE, Sugden K, Williams B, Houts RM, Danese A, Mill J, Arseneault
227 L, Caspi A (2012) Exposure to violence during childhood is associated with
228 telomere erosion from 5 to 10 years of age: a longitudinal study. Mol Psychiatry
229 18:576-581.

230 Steenstrup T, Hjelmborg JB, Kark JD, et al. (2013) The telomere lengthening
231 conundrum—artifact or biology? Nucleic Acids Research 14:e131.

232 R Development Core Team (2011) A Language and Environment for Statistical
233 Computing. R Foundation for Statistical Computing Vienna Austria

234 Wichura MJ (1988) Algorithm AS 241: The percentage points of the normal
235 distribution. J R Stat Soc Ser C Appl Stat 37:477–484.

236

Appendix

The derivation of Equation 5

A two-level regression which model telomere length (TL) can be expressed as:

$$y_{ij} = \beta_0 + \gamma_j + (\beta_1 + \varphi_j)t_{ij} + \varepsilon_{ij}, \quad \text{(A1)}$$

$$\begin{pmatrix} \gamma_j \\ \varphi_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\gamma^2 & \rho\sigma_\gamma\sigma_\varphi \\ \rho\sigma_\gamma\sigma_\varphi & \sigma_\varphi^2 \end{pmatrix} \right), \quad \text{(A2)}$$

$$\varepsilon_{ij} \sim N(0,\ \sigma_\varepsilon^2) \quad \text{(A3)}$$

where t_{ij} is the ith time point at which TL, y_{ij} is measured for the jth individual (i = 1, 2,…, n; n is the number of TL measurements and $n > 2$; j = 1, 2,…, N; N is the number of individuals in a study), β_0 is the grand intercept (TL at $t = 0$), β_0 is the grand slope (regression coefficient for t), γ_j is the deviation from β_0 for the jth individual, φ_j is the deviation from β_0 for the jth individual, γ_j and φ_j has a multivariate normal distribution with the variance-covariance structure specified in A2, and ε_{ij} is the ith residual value and residuals are normally distributed with a variance of σ_ε^2.

When we consider A1 at the time points 1 and n (i.e. i = 1 and $i = n$), TL can be written as:

$$y_{1j} = \beta_0 + \gamma_j + (\beta_1 + \varphi_j)t_{1j} + \varepsilon_{1j}, \quad \text{(A4)}$$

$$y_{nj} = \beta_0 + \gamma_j + (\beta_1 + \varphi_j)t_{nj} + \varepsilon_{nj}. \quad \text{(A5)}$$

When we have two measurements in time, 1 and m (the final time point) of telomeres the difference in telomere length is described by:

$$y_{nj} - y_{1j} = \beta_1(t_{nj} - t_{1j}) + \varphi_j(t_{nj} - t_{1j}) + \varepsilon_{nj} - \varepsilon_{1j}. \quad \text{(A6)}$$

By setting $d_j = y_{nj} - y_{1j}$, the variance of d_j can be expressed as:

$$Var(d_j) = (t_{nj} - t_{1j})^2 \sigma_\varphi^2 + 2\sigma_\varepsilon^2 . \quad \text{(A7)}$$

Note that the constant $\beta_1(t_{nj} - t_{1j})$ disappears. Using the definition of variance and further rearranging;

$$\frac{1}{(N-1)} \sum_{j=1}^{N} (d_j - \bar{d})^2 = (t_{nj} - t_{1j})^2 \sigma_\varphi^2 + 2\sigma_\varepsilon^2 , \quad \text{(A8)}$$

$$\frac{1}{2(N-1)} \sum_{j=1}^{N} d_j{}^2 = \sigma_\varepsilon^2 + \frac{(t_{nj} - t_{1j})^2 \sigma_\varphi^2}{2} + \frac{\bar{d}^2}{2} , \quad \text{(A9)}$$

where $\bar{d}$ is the mean value of *dj*. As $\bar{d} = \beta_1(t_{nj} - t_{1j})$ and setting $(t_{nj} - t_{1j}) = u$;

$$\frac{1}{2(N-1)} \sum_{j=1}^{N} d_j^2 = \sigma_\varepsilon^2 + \frac{u^2}{2}(\sigma_\varphi^2 + \beta_1^2) . \quad \text{(A10)}$$

When we assume that TL does not increase or decrease, i.e. $(\sigma_\varphi^2 + \beta_1^2)$=0, A10 reduces to:

$$\sigma_\varepsilon^2 = \frac{1}{2(N-1)} \sum_{j=1}^{N} d_j^2 . \quad \text{(A11)}$$

If we estimate σ_ε^2 in A11 only from individuals that show an increase of TL, or $d_j > 0$ (set such d_j as D_j), we have Equation 5 from the main text;

$$\sigma_\varepsilon'^2 = \frac{1}{2(m-1)} \sum_{k=1}^{m} D_k^2 , \quad \text{(A12)}$$

where D_k^2 is the difference in TL between the initial and last measurements in the *k*th individuals that showed an increase in TL (*k* = 1, 2,…, *m*; *m* is the number of individuals whose TL elongated). Note that we assume $\sigma_\varepsilon'^2$ is also normally distributed as with σ_ε^2 (A3). Due the symmetric nature of the normal distribution, $\sigma_\varepsilon'^2$ can be correctly estimated from restricted data, D_j under our assumption.

```
> #A WORKED EXAMPLE, supplement to A statistical approach to distinguish telomere elongation from error in longitudinal datasets
>
> #Read in data using comma seperated format. Each line is an individual with columns for the different timepoints of measurement. Note that in this example we assume that the
time between all measurements is equal, yet different times can be implemented using a seperate independent variable coding for time for each individual in the individual
regressions below. Simulated data are the result of a sample size of 300, average TL start of 100, with average decrease of TL of 3 per time and a SD of the slope of TL of 3,
and error SD of 2. Please refer to the simulation section of main manuscript for additional details. For help please email corresponding author, Mirre Simons at
mirresimons@gmail.com
>
> data=read.csv("simulateddata.csv")
>
> #Estimating error using individual regressions, Equation 4 in main manuscript
>
> matrixresi<-matrix(0,dim(data)[1],1)  # a matrix to put the residuals of the individual regressions in in
> x=1:dim(data)[2] #number of columns is the number of timepoints
> j=1
> while(j<(dim(data)[1]+1)) #loop the individual regressions for the amount of individuals in the dataset
+ {
+ fit<-lm(t(data[j,])~x)  #individual linear regression
+ matrixresi[j,]<-sum((residuals(fit))^2)/(length(x)-2) #residual sum of squares
+ j<-j+1
+ }
> sigma1<-mean(c(matrixresi)) #average residual sum of squares across the individuals
>
>
> #Estimating error under the assumption that TL cannot increase over time, Equation 5 in main manuscript
>
> #first we calculate the increases TL increases over time (between first(1) and last(3) timepoint)
> deltaTL=data[,3]-data[,1]
>
> #Next we determine which individuals increase in TL between the first and last timepoint
> indexTLincreases=which(deltaTL>0)
>
> #We create a new variable including only the data of individual increases
> TLincreases=deltaTL[indexTLincreases]
> #sigma2
> sigma2=0.5*sum(TLincreases^2)/(length(TLincreases))
>
>
> #compare both estimates of error variance (sigma1 and sigma2), Equation 6 in main manuscript
> vratio=sigma2/sigma1
> pvalue<-pf(vratio,length(TLincreases)-1,dim(data)[1]-1,lower.tail=F)
> print(pvalue)
[1] 1.020768e-09
>
>
> #designating a set of individuals who show TL increases with a set confidence interval, e.g. 97.5%, equation 7.
> upperlimit<-((dim(data)[1]-1)*sigma1)/qchisq(0.025,(dim(data)[1]-1)) #Note, change 0.025 in the qchisq to change the confidence at which individual increases are determined.
This upperlimit is the upper confidence of the variance determined by the individual regressions. This variance is the variance of the upper confidence limit of the underlying
normally distributed error function. Using this we can look up individual TL increases that are at the boundary (with 95% confidence) of this normal distribution (with standard
deviation equal to the upper confidence of sqrt(sigma1)
> upperTL=qnorm(0.05,0,sd=sqrt(upperlimit),lower.tail=F)
> outsideconfindex=which(0.5*TLincreases>upperTL)  #because TL increases are a result from the addition of two equal error distributions divide by 2 (i.e. *0.5).
> print(indexTLincreases[outsideconfindex]) #initial row of individual in the dataset that shows TL increase beyond the set confidence interval
[1]  31  93 138 193 210 231 270 276
>
>
>
>
>
>
```