B.2 Estimation of lifetime distributions and Markov models

1. We call the random variable of a random dinosaur lifetime X, and the observations $x_1, \ldots, x_{22}$. We do the estimates two different ways. In black we do the estimates purely on curtate lifespans, in red we include estimates for the fractional part of lifespan, which we assume to be uniform on $[0,1]$ and independent of the integer part.

   (a) The life expectancy is $\mathbb{E}[X]$, which we estimate by $\frac{1}{22}\sum x_i = 14.6$ years. The correction is $\frac{1}{2}$, so we estimate the total life expectancy to be 15.1 years.

   (b) The observed curtate lifespans have mean $\bar{x} = 14.6$ years and SD

   $$\hat{\sigma} = \sqrt{\frac{1}{n-1}\left(\sum (x_i - \bar{x})^2\right)} = 6.28 \text{ years.}$$

   Using Student's T distribution (with 21 degrees of freedom) to estimate the distribution of $(\bar{x} - \mathbb{E}[X])/\hat{\sigma}$, we estimate a 95% confidence interval for $\mathbb{E}[X]$ to be

   $$\bar{x} \pm t_{21}(.975) \cdot \frac{\hat{\sigma}}{\sqrt{n}} = 14.6 \pm 2.8 \text{ years} = (11.8, 17.4) \text{ years.}$$

   where $t_{21}(.975) = 2.08$ is the quantile function of a Student T random variable with 21 degrees of freedom; that is, $t_d(\alpha)$ satisfies $P\{T \le t_d(\alpha)\} = \alpha$ when T is a Student T random variable with d degrees of freedom.

   We impute a true average observed lifespan of $\bar{x}^* = 15.1$ years; if we assume in addition that the fractional part is independent of the integer part, the variance of the fractional part is $1/12$, which must be added to the variance of the curtate lifespan, yielding $\hat{\sigma}^* = \sqrt{\hat{\sigma}^2 + \frac{1}{12}}$, which differs from $\hat{\sigma}$ only in the fourth decimal place. This results in a confidence interval of $(12.3, 17.9)$ years.

   (c) The formula for life expectancy (ignoring the correction for partial years) is

   $$e_0 = \sum_{x=1}^{\infty} \mathbb{P}\{\text{reach age } x\} = \sum_{x=1}^{\infty} (1-q_0)(1-q_1)\cdots(1-q_{x-1}).$$

   Since $1 - \hat{q}_x^d = (\ell_x - d_x)/\ell_x = \ell_{x+1}/\ell_x$, and $\ell_0 = n$, the size of the starting population, we see that

   $$e_0 = \frac{1}{n}\sum_{x=1}^{\infty} \ell_x = \frac{1}{n}\sum_{x=1}^{\infty} \#\{x_i \ge x\} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

   (d) Counting only the curtate lifespans, we observe in age range 0–4 there are 20 who survive the full 5 years, one who survives 2 years, and one who survives 4 years, yielding a total of 106 lived years. Continuing through all the age ranges:

age	$\tilde{\ell}_x$	d_x	$\hat{\mu}_x$	$\hat{q}_x^c = 1 - e^{-\hat{\mu}_x}$
0–4	106	2	.019	.019
5–9	93	3	.032	.032
10–14	74	5	.068	.065
15–19	36	8	.22	.20
20–24	9	3	.33	.28
25–29	3	1	.33	.28

   The life expectancy of this life table is given by

   $$e_0 = \int_0^{\omega} \exp\left\{-\int_0^x \mu_s ds\right\} dx,$$

   where μ_s is the step function of estimated mortality rates, and ω is the maximum age, which we may take to be 29 or 30, or ∞ if we choose to model mortality as constant forever after

age 25. If you have constantant mortality μ_i over time intervals $[t_{i-1}, t_i)$, taking $t_0 = 0$, this comes out to

$$e_0 = \frac{1}{\mu_1}\left(1 - e^{-t_1\mu_1}\right) + \frac{1}{\mu_2}e^{-t_1\mu_1}\left(1 - e^{-(t_2-t_1)\mu_2}\right) + \cdots + \frac{1}{\mu_k}e^{-t_1\mu_1 - \cdots - (t_{k-1}-t_{k-1})\mu_{k-1}}\left(1 - e^{-(t_k - t_{k-1})\mu_k}\right).$$

Plugging in the above estimates of μ, we get $e_0 = 14.5$ years (regardless of which endpoint we choose, since the difference between an endpoint of 30 and an endpoint of ∞ is only 0.04 years in life expectancy).

The difference from the result of part (b) comes from the constraint that we have forced upon the data, that mortality rates are constant over periods of 5 years. That is, we have not directly averaged the data, but found the average that is closest to fitting this assumption.

We may estimate life expectancy as

$$\hat{e}_0 = \frac{1}{2} + \sum_{x=1}^{\infty}(1 - \hat{q}_0^c)\cdots(1 - \hat{q}_{x-1}^c) = 14.5 \text{ years}.$$

A better estimate comes from including half of the integer year in which the individual died in the count of years at risk, we replace the estimate of total years lived by $\tilde{\ell}_x^* = \tilde{\ell}_x + d_x/2$. We end up with an alternative estimate $\hat{e}_0^* = 14.8$ years.

age	$\tilde{\ell}_x$	d_x	$\hat{\mu}_x$	$\hat{q}_x^c = 1 - e^{-\hat{\mu}_x}$
0–4	107	2	.019	.019
5–9	94.5	3	.032	.031
10–14	76.5	5	.065	.063
15–19	40	8	.20	.18
20–24	10.5	3	.29	.25
25–29	3.5	1	.29	.25

(e) The lifetime is reported at 25 years. That is to say, we know the lifetime X is between 25 and 26 years. Thus $Z := X - 25$ has the distribution of an exponential random variable conditioned on $X < 1$. If the rate is μ, this yields

$$\mathbb{E}[Z \mid Z < 1] = \frac{\int_0^1 \mu e^{-\mu z} z\, dz}{\int_0^1 \mu e^{-\mu z} dz} = \frac{e^\mu - 1 - \mu}{\mu(e^\mu - 1)} \approx \frac{1}{2} - \frac{\mu}{12}.$$

When $\mu = 0.333$ this evaluates to 0.472 (and the linear approximation is accurate to 5 decimal places), giving us an average time of death of 25.472 years. One lesson is that the approximation that adds 1/2 year to each curtate lifetime is fairly close, even when the mortality rates are quite high.

(f) The log likelihood is

$$\ell(B, \theta) = \sum_i \left[\log h(x_i) - H(x_i)\right] = n \log B + \sum_i \left[\theta x_i - \frac{B}{\theta}\left(e^{\theta x_i} - 1\right)\right].$$

The maximum of ℓ could come either at an interior point or at a boundary point. The likelihood goes to 0 as B goes to 0, so the maximum on the boundary is attained along $\theta = 0$, and at an interior point in B.

The partial derivatives are

$$\frac{\partial \ell}{\partial B} = \frac{n}{B} - \sum_i \left[\frac{1}{\theta}\left(e^{\theta x_i} - 1\right)\right] = -\frac{n}{B} + \frac{n}{\theta}\left(Q(\theta) - 1\right)$$

$$\frac{\partial \ell}{\partial \theta} = -n\|x\|_1 + nB\frac{\theta Q'(\theta) - Q(\theta) + 1}{\theta^2}.$$

For any given choice of $\hat\theta$, the maximum likelihood estimate of B must be $\hat B = \hat\theta/(Q(\hat\theta)-1)$. Define

$$\ell_*(\theta) := \ell(\theta/(Q(\theta)-1),\theta) = -\log\frac{Q(\theta)-1}{\theta} + \|x\|_1\theta - 1.$$

The maximum-likelihood estimate is $(\hat\theta/(Q(\hat\theta)-1),\hat\theta)$, where $\hat\theta$ maximises ℓ_*. We note that

$$\ell_*'(\theta) = -\frac{Q'(\theta)}{Q(\theta)-1} + \frac{1}{\theta} + \|x\|_1,$$
$$\ell_*''(\theta) = \frac{Q'(\theta)^2 - Q''(\theta)Q(\theta) + Q''(\theta)}{(Q(\theta)-1)^2} - \theta^{-2}$$

Critical points for ℓ_* will thus be solutions to

$$\frac{Q'(\theta)}{Q(\theta)-1} - \frac{1}{\theta} = \|x\|_1 \tag{B.5}$$

The optional question asked why there should be a unique maximum which is also a critical point. Define $\|x\|_k = (x_1^k + \cdots + x_n^k)^{1/k}$. Then

$$\lim_{\theta\to\infty}\frac{Q'(\theta)}{Q(\theta)-1} = \max x_i \geq \|x\|_1,$$

and by L'Hôpital's rule,

$$\begin{aligned}\lim_{\theta\downarrow 0}\frac{Q'(\theta)}{Q(\theta)-1} - \frac{1}{\theta} &= \lim_{\theta\downarrow 0}\frac{\theta Q'(\theta) - Q(\theta) + 1}{\theta(Q(\theta)-1)}\\ &= \lim_{\theta\downarrow 0}\frac{\theta Q''(\theta)}{Q(\theta) - 1 + \theta Q'(\theta)}\\ &= \frac{Q''(0)}{2Q'(0)}\\ &= \frac{\|x\|_2^2}{2\|x\|_1}.\end{aligned}$$

If this is below $\|x\|_1$ (meaning that $2\|x\|_1^2 > \|x\|_2^2$) then there must be a solution to (B.5), which will be a solution to $\ell_*'(\hat\theta) = 0$. On the other hand, if $2\|x\|_1^2 \leq \|x\|_2^2$ then ℓ_*' starts below 0 and ends below 0. In this case, the maximum may be expected to lie at the boundary point $(0, 1/\|x\|_1)$; the proof that the likelihood does not come back up is not easy, though, and will not be required or given here.

In the former case, the first critical point along the curve $B = \theta/(Q(\theta)-1)$ will be a local maximum, unless it is a saddle point; we do need to check the second derivative to be sure.

Solving these equations numerically, we find the solution $(\hat\theta, \hat B) = (0.145, .0130)$. The second derivative at this point, in particular, may be computed by $Q(\hat\theta) = 12.2$, $Q'(\hat\theta) = 241$, $Q''(\hat\theta) = 5152$, implying that $\ell^*(\hat\theta) = -35.8$. (In fact, the second derivative is always negative, which implies that this is the only critical point. This is not easy to prove, but it can be checked numerically.)

Note: We compute $Q'(\theta) = n^{-1}\sum x_i e^{\theta x_i}$, and $Q''(\theta) = n^{-1}\sum x_i^2 e^{\theta x_i}$.

We have

$$\mathcal{I}(\theta, B) = \begin{pmatrix} B(2\theta^{-3}q_0(\theta) - \theta^{-2}q_1(\theta) + \theta^{-1}q_2(\theta)) & \theta^{-1}q_1(\theta) - \theta^{-2}q_0(\theta) \\ \theta^{-1}q_1(\theta) - \theta^{-2}q_0(\theta) & B^{-2}\end{pmatrix}$$

As usual, we estimate the information at the true parameter by substituting the estimated parameters, yielding

$$\mathcal{I}(\hat\theta, \hat B) \approx \begin{pmatrix} 408 & 1068 \\ 1068 & 5917\end{pmatrix}$$

It follows that $(\hat{\theta} - \theta, \hat{B} - B)$ is approximately normal with covariance matrix

$$n^{-1}\mathcal{I}^{-1} \approx \begin{pmatrix} 2.1 \times 10^{-4} & -3.9 \times 10^{-5} \\ -3.9 \times 10^{-5} & 1.5 \times 10^{-5} \end{pmatrix}.$$

Thus a 95% confidence interval for B is about $0.0130 \pm 1.96 \cdot \sqrt{1.5 \times 10^{-5}} = (.0054, .0206)$.
The optional question asks for a 95% confidence interval for $h(20) = Be^{20\theta}$. We treat B and θ as normal random variables, using the asymptotic theory. Note that

$$\mathbb{E}[Q(\theta)] = \mathbb{E}\left[e^{\theta X}\right],$$

where X is chosen from a Gompertz distribution with parameters (B, θ). We define

$$q_k(\theta) := \mathbb{E}[Q^{(k)}(\theta)] = e^{B/\theta}\theta^{1-k}B^{-1}\int_{B/\theta}^{\infty}(\log u + \log\theta/B)^k u e^{-u} du.$$

While $q_0(\theta) = 1 + \theta/B$, the others have no simple closed-form solutions. Still, there is no difficulty in computing them. We get $q_0(\hat{\theta}) = 12.2$, $q_1(\hat{\theta}) = 239$, and $q_2(\hat{\theta}) = 5020$.
We can represent $(\hat{\theta} - \theta, \hat{B} - B) = (\alpha Z_1 + \beta Z_2, \gamma Z_1 + \delta Z_2)$, where (Z_1, Z_2) are independent standard normal random variables and

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = (n\mathcal{I})^{-1/2} = \begin{pmatrix} 1.4 \times 10^{-2} & -2.2 \times 10^{-3} \\ -2.2 \times 10^{-5} & 3.2 \times 10^{-3} \end{pmatrix}.$$

(Note: This is the square root of the matrix $(n\mathcal{I})^{-1}$. How do we turn this into an estimate for $h(20)$? We use a Taylor series to approximate

$$\begin{aligned} h(20;\theta,B) - h(20;\hat{\theta},\hat{B}) &\approx \frac{\partial h(20)}{\partial\theta}(\theta - \hat{\theta}) + \frac{\partial h(20)}{\partial B}(B - \hat{B}) \\ &= 20Be^{20\theta}(\theta - \hat{\theta}) + e^{20\theta}(B - \hat{B}) \\ &\approx 4.7(\theta - \hat{\theta}) + 18(B - \hat{B}) \\ &\approx 0.028Z_1 + .047Z_2, \end{aligned}$$

which is normal with mean 0 and variance $\sigma^2 = .0030$. Thus, we estimate a 95% confidence interval to be

$$h(20;\hat{\theta},\hat{B}) \pm 1.96\sigma = .23 \pm .11.$$

(g) The survival function for a Gompertz population is

$$\exp\left\{-\frac{B}{\theta}\left(e^{\theta x} - 1\right)\right\},$$

so the life expectancy is

$$e_0 = \int_0^{\infty} \exp\left\{-\frac{B}{\theta}\left(e^{\theta x} - 1\right)\right\} dx.$$

Substituting $u = \frac{B}{\theta}e^{\theta x}$ yields

$$e_0 = \frac{e^{B/\theta}}{\theta}\int_{B/\theta}^{\infty} e^{-u}\frac{du}{u} = \frac{e^{B/\theta}}{\theta}\,\mathrm{E}_1(B/\theta).$$

Plugging in the estimates yields $\hat{e}_0 = 14.5$ years.

2.

A. sarcophagus survival bootstrap

Fitting Gompertz model

We know that the MLE $(\hat{B}, \hat{\theta})$ satisfies

$$\frac{Q'(\hat{\theta})}{Q(\hat{\theta}) - 1} - \frac{1}{\hat{\theta}} - \bar{x} = 0$$

and

$$\hat{B} := \frac{\hat{\theta}}{(Q(\hat{\theta}) - 1)},$$

where $Q(\theta) := \frac{1}{n} \sum e^{\theta x_i}$ and $\bar{x} := \frac{1}{n} \sum x_i$.

```
########### A. sarcophagus bootstrap ############
require(stats)
ages=c(2,4,6,8,9,11,12,13,14,14,15,15,16,17,17,18,19,19,20,21,23,28)
numSkelets=length(ages)
# MLE
Q=function(theta,A){
  mean(exp(theta*A))
}
Qprime=function(theta,A){
  mean(A*exp(theta*A))
}
scoreFunction=function(theta,A){
  Qprime(theta,A)/(Q(theta,A)-1)-1/theta-mean(A)
}

MLE.theta=uniroot(function(theta) scoreFunction(theta,ages+.5),c(1e-9,10))$root
MLE.B=MLE.theta/(Q(MLE.theta,ages+.5)-1)
cat(c(MLE.theta,MLE.B))
```

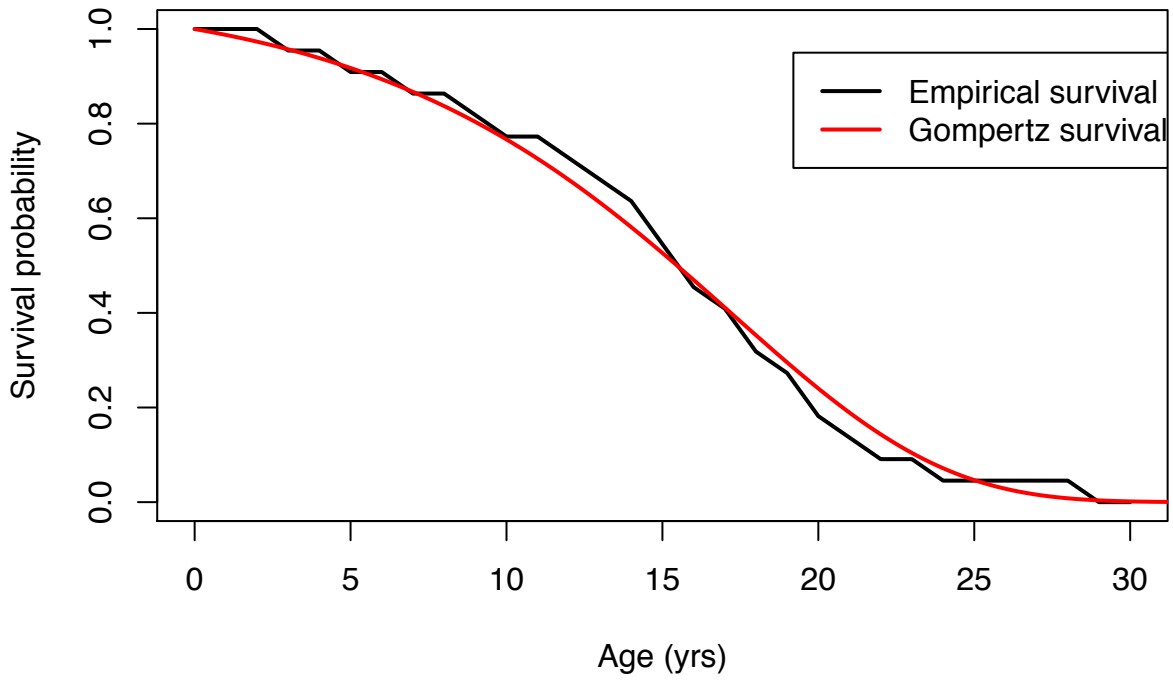
```
## 0.147417 0.01163253
```

Plot empirical survival and compare to Gompertz

```
### Plot empirical survival with slopes for
### integer periods; add Gompertz fit with parameter (B,\theta)
survivalplot=function(A,B=NA,theta=NA){
  allSurvive=survivalplotCalc(A)
  plot(allSurvive$ages,allSurvive$Surv,type='l',lwd=2,xlab='Age (yrs)', ylab='Survival probability')
  if(!is.na(B)&!is.na(theta)){
    maxAge=max(allSurvive$ages)+2
    gompAges=seq(0,maxAge,.1) # Smooth curve
    gompSurv=exp(-B/theta*(exp(theta*gompAges)-1))
    lines(gompAges,gompSurv,col=2,lwd=2)
    legend(.6*max(maxAge),.95,c('Empirical survival','Gompertz survival'),lwd=2,col=1:2)
  }
}
```

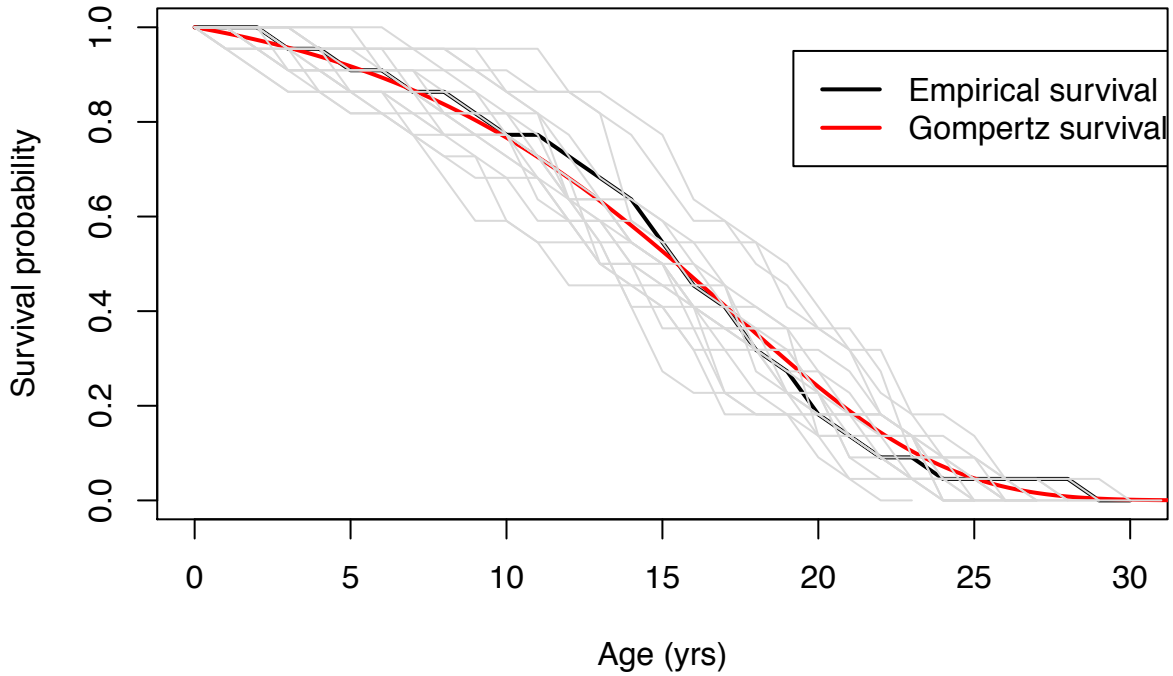
```
survivalplotCalc=function(A){
   N=length(A)
  ageTable=N-cumsum(table(A))
  ageX=as.integer(names(ageTable))
  surviveY=as.vector(ageTable)/N #Survival to age X
  allAges=seq(0,max(ageX)+2)
  allSurvive=rep(1,max(ageX)+3) # Initialise survival to 1
  allSurvive[ageX+2]=surviveY # Survival at given ages is assigned
  allSurvive=cummin(allSurvive) # Other ages extended as constant
  list(ages=allAges,Surv=allSurvive)
}

survivalplot(ages,MLE.B,MLE.theta)
```



Parametric bootstrap

```
### Function to simulate n samples from Gompertz
### Inverse cumulative hazard
gompSample=function(B,theta,n){
  floor(log(1+theta/B*rexp(n))/theta)
  # We take only the integer part, to make it comparable to our observations
}
### Plot 20 bootstrap samples
survivalplot(ages,MLE.B,MLE.theta)
for (i in 1:20){
  gS=gompSample(MLE.B,MLE.theta,numSkelets)
  spC=survivalplotCalc(gS)
  lines(spC$ages,spC$Surv,col=gray(.85))
}
```



We now simulate 10000 bootstrap samples, and for each one we estimate $\hat{B}$ and $\hat{\theta}$. We use this to compute bootstrap confidence intervals for the parameters.

```
alpha=.95  # level of confidence intervals
n.samples=10000
allB=rep(0,n.samples)
alltheta=rep(0,n.samples)
for (i in 1:n.samples){
  gS=gompSample(MLE.B,MLE.theta,numSkelets)
  alltheta[i]=uniroot(function(theta) scoreFunction(theta,gS+.5),c(.01,10))$root
allB[i]=alltheta[i]/(Q(alltheta[i],ages+.5)-1)
}
```

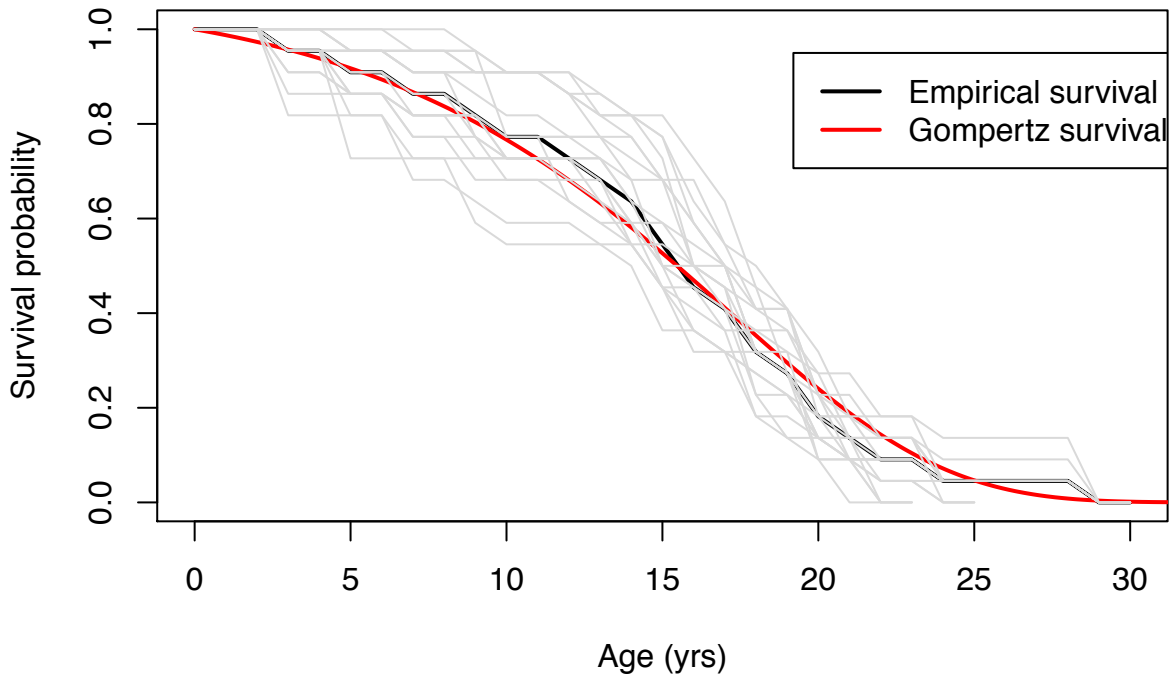
We now extract the appropriate quantiles of the samples.

```
## 95 % confidence interval for theta (  0.097 , 0.241 )
```

```
## 95 % confidence interval for B (  0.002 , 0.023 )
```

Nonarametric bootstrap

```
### Plot 20 bootstrap samples
survivalplot(ages,MLE.B,MLE.theta)
for (i in 1:20){
  gS=sample(ages,numSkelets,replace=TRUE)
  spC=survivalplotCalc(gS)
  lines(spC$ages,spC$Surv,col=gray(.85))
}
```



We now simulate 10000 bootstrap samples, and for each one we estimate $\hat{B}$ and $\hat{\theta}$. We use this to compute bootstrap confidence intervals for the parameters.

```
alpha=.95  # level of confidence intervals
n.samples=10000
allB=rep(0,n.samples)
alltheta=rep(0,n.samples)
for (i in 1:n.samples){
  gS=sample(ages,numSkelets,replace=TRUE)
  alltheta[i]=uniroot(function(theta) scoreFunction(theta,gS+.5),c(.01,10))$root
allB[i]=alltheta[i]/(Q(alltheta[i],ages+.5)-1)
}
```

We now extract the appropriate quantiles of the samples.

```
## 95 % confidence interval for theta (  0.099 , 0.287 )
```

```
## 95 % confidence interval for B (  0.001 , 0.023 )
```

3. (a) i. Just write the quantities as integrals and sums and interchange the order of integration and summation:

$$E_x^c = \sum_{i=1}^{n} \int_{a_i}^{b_i} dt = \int_K^{K+N+1} \sum_{i=1}^{n} 1_{\{a_i \le t \le b_i\}} dt = \int_K^{K+N+1} P_{x,t} dt. \tag{B.6}$$

ii. Under the assumption of piecewise linearity we calculate

$$E_x^c = \sum_{k=K}^{K+N} \int_0^1 (r P_{x,k} + (1-r) P_{x,k+1}) dr = \sum_{k=K}^{K+N} \frac{P_{x,k} + P_{x,k+1}}{2}. \tag{B.7}$$

The assumption of piecewise linear $P_{x,t}$ cannot hold exactly since $P_{x,t} \in \mathbb{N}$, but for large n this is negligible.

(b) Denote by $E^c_{[t,t+1]}$ the total time exposed to risk between ages t and $t+1$ for $t \in \mathbb{R}_+$. Then, for the first three, clearly,

$$d_x^{(1)} / E^c_{[x,x+1]} \quad \text{estimates} \quad \text{if } \mu_{x+\frac{1}{2}} \quad \mu_t \text{ constant for } t \in [x, x+1);$$

A death contributing to $d_x^{(4)}$ can be due to somebody aged anywhere in $(x-1, x+1)$, so there is overlap between the one-year age groups. We can define

$$P_{x,t}^{(2)} = \#\text{ lives at risk at time } t \text{ with } x\text{th birthday in calendar year } \lfloor t \rfloor, \tag{B.8}$$

but also ought to adjust

$$E_x^{c,4} = \int_K^{K+N+1} P_{x,t}^{(4)} dt \approx \sum_{k=K}^{K+N} \frac{P_{x,k} + P_{x+\mathbf{1},k+1}}{2} \tag{B.9}$$

since it is more natural to assume that the cohort of lives $P_{x,k}$ with xth birthday in calendar year k changes linearly to $P_{x+1,k+1}$ since this will count the same people (being age k at the start of year x and age $k+1$ at the start of year $x+1$).

4. (a) The times between arrivals in a Poisson process with intensity λ are independent exponential with parameter λ. Thus, the number of cumulative sums that are $\le t$ is precisely the number of arrivals on the interval $[0, t]$, which has a Poisson distribution with parameter λt.

(b) By definition, the interval $(a(x), b(x))$ is a $(1-\alpha)100\%$ confidence interval for the parameter μ if and only if $P_\mu\{x : \mu \notin [a(x), b(x)]\} = \alpha$. For simplicity, let's say we're looking for a symmetric confidence interval, so with $P\{x : \mu < a(x)\} = P\{x : \mu > b(x))\} = \alpha/2$.
The key point is that the lower limit of the confidence interval depends on the upper tail of the distribution, and vice versa. Thus, if we want to find $a(x)$, we need to find λ small enough that the probability in the upper tail, of X as big as the x we observed, is below $\alpha/2$. That is, we need to solve the equation

$$P_a(X \ge x) = \alpha/2.$$

Of course, we could do that directly, numerically, though it is slightly awkward that we are searching for a fixed quantile of a distribution whose parameter is a, rather than a fixed distribution.
Letting $T_1, \ldots, T_x$ be i.i.d. exponential random variables with parameter 1, and Z_{2x} a random variable with χ^2 distribution with $2x$ degrees of freedom, we know that

$$P_\lambda(X \ge x) = P(T_1 + \cdots + T_x \le \lambda) = P(Z_{2x} \le 2\lambda).$$

Thus, $a(x)$ must satisfy $P(Z_{2x} \le 2a) = \alpha/2$, meaning that $a = \frac{1}{2} c_{\alpha/2}(2x)$. Similarly, we calculate $b(x)$ from the bound $P_b(X \le x) = \alpha/2$, which is equivalent to

$$P_b(X \ge x+1) = 1 - \frac{\alpha}{2}.$$

# events observed	lower	upper
0	*	3.69
1	0.025	5.57
2	0.242	7.22
3	0.619	8.77
4	1.09	10.2
5	1.62	11.7
6	2.20	13.1
7	2.81	14.4
8	3.45	15.8

Table B.1: 95% confidence intervals for Poisson distribution.

age	# deaths	# at risk	lower	upper
0–4	2	22	.0022	.066
5–9	3	20	.0062	.088
10–14	5	17	.019	.137
15–19	8	12	.058	.263
20–24	3	4	.031	.438
25–29	1	1	.0051	1.11

Table B.2: 95% confidence intervals for mortality rates in dinosaur data.

(c) Each individual has probability $p = 1 - e^{-\lambda t}$ of dying during this time, and the events are independent. Since there are n individuals at risk, k may be seen as a sample from a binomial distribution with parameters (n, p). If np is moderately small, and n is moderately large, then this distribution is approximately Poisson with parameter $n(1 - e^{-\lambda t})$, which for moderately small values of λt is approximately $nt\lambda$. A confidence interval for λ will thus be $1/nt$ times the corresponding confidence interval for the Poisson parameter. A slightly better approximation, will be

$$\left(\log \left[1 - \frac{1}{2n} c_{\alpha/2}(2k) \right], \log \left[1 - \frac{1}{2n} c_{1-\alpha/2}(2k+2) \right] \right),$$

which avoids the approximation of e^{-x} by $1 - x$. This will still be only approximate, since it depends on replacing the binomial by Poisson distribution, and so will not be good when n is small. However, when n is large but np small — which we expect to be the case when n is large and k is not very large — the Poisson approximation beats the normal approximation. In particular, when $k = 0$, the normal approximation yields a confidence interval which extends into the negative, which makes little sense.

(d) In table B.1 we see the (exact) confidence intervals for a Poisson random variable. In fact, since for 0 observed events there is no lower bound except 0, it makes sense to use a one-sided confidence interval, with upper bound $c_{.95}(2)/2 = 3.00$ instead of $c_{.975}(2)/2 = 3.69$. (Of course, we always have the option of using asymmetric confidence intervals.)

In Table B.2 we take the corresponding row of the previous table (for the number of deaths in that period), and divide it by 5 times the number at risk. The normal confidence intervals are based on the Standard Error computation

$$SE(\hat{\mu}) = \sqrt{\mu/\tilde{\ell}} \approx \sqrt{\hat{\mu}/\tilde{\ell}},$$

where $\tilde{\ell}$ is the total time at risk. A symmetric 95% confidence interval is then $\hat{\mu} \pm 1.96SE$.

age	$\tilde{\ell}_x$	d_x	$\hat{\mu}_x$	Standard Error	Confidence Interval
0–4	106	2	.019	.013	(-0.007,.045)
5–9	93	3	.032	.018	(-0.004,.068)
10–14	74	5	.068	.029	(-.010,.126)
15–19	36	8	.22	.07	(.08,.36)
20–24	9	3	.33	.16	(.01,.65)
25–29	3	1	.33	.29	(-.25,.91)

Table B.3: 95% confidence intervals based on the normal approximation.

5. We compute z_x for each age class as follows:

Age	Exposed to risk E_x	Observed deaths d_x	Expected deaths $E_x q_s^x$	z_x
20–24	35000	35	34	0.180
25–29	33000	30	29	0.178
30–34	30000	31	35	-0.692
35–39	30000	45	52	-0.959
40–44	31000	84	81	0.379
45–49	28000	138	129	0.813
50–54	25000	229	213	1.14
55–59	23000	360	345	0.814
60–64	20000	522	500	0.996

Table B.4: Computing z_x

(a) The X^2 statistic is of the form

$$X^2 = \sum z_x^2, \text{ where } z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s (1 - q_x^s)}}. \tag{B.10}$$

Substituting in the values from Table B.4 we get $X^2 = 5.21$. Since this corresponds to χ^2 with 9 degrees of freedom, we get a p-value of 0.82.

(b) The cumulative deviations test statistic is

$$Z = \frac{\sum d_x - E_x q_x^s}{\sqrt{\sum E_x q_x^s (1 - q_x^s)}},$$

which should have approximately $N(0, 1)$ distribution under the null hypothesis. We compute $Z = 1.53$, yielding a p-value (for the 2-sided Z test) of 0.13. This lower p-value makes sense, since there are only two negative deviations, but it is still far too high to be considered any significant evidence that the data did not come from the standard table.

(c) Signs test: We observe 7 positives out of 9 tries. Under the null hypothesis these 7 should be like $P = \text{Binom}(9, \frac{1}{2})$. The p-value is

$$P\{0, 1, 2, 7, 8, 9\} = \sum_{0,1,2,7,8,9} \left(\frac{1}{2}\right)^9 \binom{9}{k} = 0.18.$$

While there is no strong evidence that the population differs from the standard population, it looks as though there may under-estimation of mortality using the standard life table. This would be a bad thing for an insurance company, which would then set its premiums too low. Graduation, possibly using $q_x^0 = b + a q_x^s$ may help.