Exposed to Risk and Census methods

CHRIS SUTTON, NOVEMBER 2024

5 models in first 6 weeks

- Kaplan Meier estimator
- Cox's P-H model
- multi-state Markov process
- Binomial type models
- Poisson model

typical estimator calculations

$$\frac{\text{number of transitions}}{\text{total waiting time}}$$

$$\frac{\text{decrement count}}{\text{exposed to risk}}$$

This week we are concerned with issues around the calculation of the denominator especially where we have incomplete data

Topic outline

1. Central and Initial Exposed to Risk
2. Homogeneity
3. Principle of correspondence
4. Census approximations
5. Definitions of age

Central and Initial E-to-R

definitions

E_x^c the Central exposed to risk

- the observed waiting time
- used in multi-state & Poisson models

E_x the Initial exposed to risk

- approx $E_x \approx E_x^c + ½d_x$
- for the actuarial estimate in Binomial type models

comparison

Central exposed-to-risk = observed waiting time, is a very intuitive measure

Initial exposed-to-risk requires an adjustment to what actually observed for lives who die so its interpretation more complicated

- unless we can use the naïve binomial with N lives observed for whole year

Central exposed-to-risk extends unchanged to multi-decrement and multi-state models in way that Initial exposed-to-risk cannot

Where the Central exposed-to-risk needs adjustments from available life assurance data, it is hard to justify a 2nd set of adjustments needed for Initial exposed-to-risk

Initial exposed-to-risk historically important for actuaries from time when binomia-type models formed the basis of most life tables

- Today multi-state (or Poisson) models are more attractive in many situations

our focus

in most actuarial investigations the multi-state or Poisson models will be usable and the additional limitations of initial exposed-to-risk and binomial models are not needed

for the remainder of this topic we will focus on central exposed-to-risk E_x^c

Homogeneity

a valid assumption?

Our models have carried the assumption we can observe **identical** lives

- or at least ones with the same mortality characteristics, so that we can assume they follow the same distribution T_x
- in practice this will never be entirely true

homogeneity = the quality of all being the same or of the same kind

Hence we sub-divide populations by characteristics known to affect mortality in attempt to reduce heterogeneity

common sub-divisions

Age

Type of policy

Smoker / Non-Smoker

Male / Female

Level of underwriting

Duration policy in force

how much subdivision?

life assurance companies can only sub-divide where the data has been collected [statement of the obvious]

- usual source is proposal form
- marketing reasons to keep these short

more sub-divisions result in smaller populations making use of statistical methods more difficult

- balance required between the desire for homogeneity and need for large enough populations

other potential sub-divisions

sales channel	policy size
occupation	known impairments
postcode	marital status

Principle of correspondence

correspondence

our q_x and μ_x estimators use deaths and exposed-to-risk data

- these 2 data sets need to be consistent [should be obvious]

however in life assurance they often come from 2 different sources

- deaths from claims data
- exposed-to-risk from premiums collected data
- need care to ensure that these two use the same definition of age x

the **principle of correspondence**

> A life alive at time *t* should be included in the exposure at age x at time *t* if and only if, were that life to die immediately, they would be counted in the deaths data d_x at age x.

Census approximations

exact calculation

record all dates of birth

record dates of entry into observation

record dates of exit from observation

calculate central exposed to risk E_x^c

if we add data on cause of exit from observation we have d_x as well

however

often the exact calculation is not possible because either:

- precise dates of entry or exit are not recorded
- the age definition does correspond to [x, x+1]

In these cases, what are known as **census approximations** are necessary

CMI

Continuous
Mortality Investigation

Institute and Faculty of Actuaries

https://www.actuaries.org.uk/learn-and-develop/continuous-mortality-investigation/about-cmi

$P_{x,t}$

death data is often in the form

d_x = total number deaths age x last birthday in the calendar years K, K+1, … K+N

- so N+1 calendar years of data for deaths between ages x and x+1

CMI does not have access to precise entry & exit from observation data, instead it receives **census** data

$P_{x,t}$ = number of lives under observation, aged x last birthday at time t

where t in this (CMI) case is 1st January in calendar years K, K+1, … K+N

- so N+1 calendar years of total number policies in-force on 1st January

E^c_x

for any t (i.e. not just 1st January census)

$$E^c_x = \int_K^{K+N+1} P_{x,t}\, dt$$

our problem then reduces to estimating this integral when we have $P_{x,t}$ at only a few calendar dates (e.g. 1st January's)

CMI then uses the **trapezium approximation** assumes $P_{x,t}$ is linear between census dates

$$E^c_x \approx \sum_{t=K}^{K+N} ½(P_{x,t} + P_{x,t+1})$$

with census data in years K,..,K+N+1

(although easily adaptable to different intervals)

Definitions of age

different definitions

earlier we used "age last birthday" in d_x which gives year of age [x, x+1]

other age definitions are possible:

$d_x^{(2)}$	Number of deaths at age x <u>nearest birthday</u>
$d_x^{(3)}$	Number of deaths at age x <u>next birthday</u>

these different years of age are called the **rate interval**

resulting estimates

Definition of x	Rate interval	q estimates	μ estimates
Age last birthday	$[x, x + 1]$	q_x	$\mu_{x+½}$
Age nearest birthday	$[x – ½ , x + ½]$	$q_{x-½}$	μ_x
Age next birthday	$[x – 1, x]$	q_{x-1}	$\mu_{x-½}$

$\hat{q}$ estimates q at the start of the rate interval

$\hat{\mu}$ estimates μ at the mid-point of the rate interval

census data correspondence

with different age definitions we need to check that the principle of correspondence is satisfied

- census data {P} is consistent with death data {d} if and only if any of the lives counted in P were to die *on the census date* itself then they would be included in {d}

so $P^{(2)}_{x,t}$ should be used for 'age nearest' data with rate interval $[x - ½ , x + ½]$, the number of lives under observation age x nearest birthday at time t

- (where e.g. t is 1st January in calendar years K, K+1, …, K+N+1

and $P^{(3)}_{x,t}$ should be used for 'age next' data with rate interval $[x–1, x]$, the number of lives under observation age x next birthday at time t

- (where e.g. t is 1st January in calendar years K, K+1, …, K+N+1

death data has priority

if we find death and census date with different age definitions we must adjust the census data not the death data because as mortality rates are usually small each piece of death data carries more information and should be preserved intact.

example if we have age nearest birthday death data, $d_x^{(2)}$ but age last birthday census data then we can use

$P'_{x,t} = ½(P_{x-1,t} + P_{x,t})$ as an approximation of $P_{x,t}^{(2)}$