Graduation and Statistical tests of mortality experience

CHRIS SUTTON, NOVEMBER 2023

Over the next two weeks we will cover

1. Comparing mortality experience with standard tables
2. Graduation and the reasons for it
3. Desirable features of graduated rates
4. Statistical tests of mortality experience
5. Methods of graduation
6. Statistical tests of a graduation

Introduction to the topic

from observations

The models we have considered (e.g. multi-state, Poisson, Binomial) give mortality for a single year of age [x, x+1]

- for practical work we generally need a life table covering a large age range

If our observations yield data for all ages $x_1, x_2, \ldots, x_n$ then we can obtain:

- death and exposed-to-risk data
- crude estimates of the rate of mortality or force of mortality
- asymptotic distributions for D_x q_x or μ_x

model data, outputs and distributions

Poisson or multi-state	Binomial type
Deaths d_x	Deaths d_x
Exposed to risk E_x^c	Initial Exposed to risk $E_x \approx E_x^c + ½d_x$
Crude estimate of hazard rate $\mu_{x+½}$	Crude estimate of mortality rate q_x
Asymptotically $D_x \sim \text{Normal}\ (E_x^c\mu_{x+½}\ ,\ E_x^c\mu_{x+½}\ )$	Approximately $D_x \sim \text{Binomial}\ (E_x\ ,\ q_x)$
or $\tilde{\mu}_{x+½} \sim \text{Normal}\ (\mu_{x+½}\ ,\ \mu_{x+½}\ /\ E_x^c\ )$	With the further approximation $D_x \sim \text{Normal}\ [\ E_xq_x\ ,\ E_xq_x(1\text{-}q_x)\ ]$ or $\tilde{q}_x \sim \text{Normal}\left[q_x\ ,\ \frac{q_x(1-q_x)}{E_x}\right]$

comparison with standard tables

standard tables

a life assurance company will want to compare its own mortality experience with:

1. past experience
   - e.g. to assess premiums

2. published life tables also known as "standard tables"
   - e.g. if considering those tables for financial reporting purposes

National Life Tables

- from census data every 10 years
- e.g. English Life Tables

Tables from Life Office data

- calculated by CMI in the UK
- e.g. "92 series"

checking

Life assurance actuaries use standard tables a lot, so it is important to check the company's own experience is consistent with the tables used

if q_x^s and $\mu_{x+½}^s$ are from standard tables

and $\hat{q}_x$ and $\hat{\mu}_{x+½}$ are from models with data taken from experience

we want to know whether the two are consistent, or more formally

> Our hypothesis is that standard table values $\{q_x^s\}$ and $\{\mu_{x+½}^s\}$ are the "true" parameters of our model [multi-state / Poisson / Binomial] at each age x

we test the hypothesis using the [Normal] distribution assumptions for D_x already given, comparing death data d_x with expected deaths given the standard table

Graduation

what? $\mathring{q}$ and $\mathring{\mu}$

our model outputs $\{\hat{q}_x\}$ and $\{\hat{\mu}_{x+½}\}$ will proceed roughly not smoothly with x

- because each value is estimated independently and therefore will contain its own sampling errors

we would prefer q_x and $\mu_{x+½}$ that are smooth functions of age

hence we **graduate** or smooth crude estimates $\hat{q}_x$ and $\hat{\mu}_{x+½}$ to produce graduated estimates $\mathring{q}_x$ and $\mathring{\mu}_{x+½}$

- There are various methods used to do this which we will cover next week

the reasons for graduation

why?

1

there is intuitive appeal in a smooth function (although this is effectively an assumption)

2

smoothing can use data in adjacent ages to improve estimates

3

it may be a way of reducing sampling error at individual ages

4

hard to justify jumps when using q or μ in practical work

Note however, graduation cannot remove:

1. bias in the data
2. mistakes

desirable features

1. smoothness
2. adherence to data or "goodness of fit"
3. suitability for intended practical work

features 1. and 2. are usually in conflict. The art of graduation lies in finding a satisfactory compromise

this depends on the nature of the work. In life assurance we must not underestimate mortality; for annuities we must not overestimate it

Statistical tests of mortality experience

testing smoothness

testing smoothness

the usual test for smoothness in mathematics is differentiability, but that is no good in this situation

in practice smoothness is usually easily obtained in 3 of the 4 graduation methods we will meet later

where we do wish to test, the most usual criteria is that the <u>third difference</u> of graduated quantities $\{\mathring{q}_x\}$ or $\{\mathring{\mu}_{x+½}\}$ should:

a) be small compared to the quantities themselves

b) progress regularly

test set-up

2 things we might test

comparing mortality experience to a standard table

testing adherence to data (or goodness of fit) of a graduation

set up

Hypothesis

(if comparing experience to a standard table)

<u>Two State or Poisson</u>

$D_x \sim Normal(E^c_x \mu^s_{x+½} , E^c_x \mu^s_{x+½})$

<u>Binomial</u>

$D_x \sim Binomial(E_x , q^s_x)$

Hypothesis

(if testing adherence to data of a graduation)

<u>Two State or Poisson</u>

$D_x \sim Normal(E^c_x \mathring{\mu}_{x+½} , E^c_x \mathring{\mu}_{x+½})$

<u>Binomial</u>

$D_x \sim Binomial(E_x , \mathring{q}_x)$

deviation and standardised deviation

In multi-state and Poisson models:

deviation = actual deaths – expected deaths

$$= d_x - E^c_x \mu^s_{x+½} \quad \text{or} \quad d_x - E^c_x \mathring{\mu}_{x+½}$$

standardised deviation is

$$z_x = \frac{d_x - E^c_x \mu^s_{x+½}}{\sqrt{(E^c_x \mu^s_{x+½})}} \quad \text{or} \quad z_x = \frac{d_x - E^c_x \mathring{\mu}_{x+½}}{\sqrt{(E^c_x \mathring{\mu}_{x+½})}}$$

deviation and standardised deviation 2

In Binomial type models:

deviation = actual deaths – expected deaths

$= d_x - E_x q_x^s \quad \text{or} \quad d_x - E_x \mathring{q}_x$

standardised deviation is

$$z_x = \frac{d_x - E_x q_x^s}{\sqrt{(E_x q_x^s)}} \quad \text{or} \quad z_x = \frac{d_x - E_x \mathring{q}_x}{\sqrt{(E_x \mathring{q}_x)}}$$

large samples

if sufficient number of lives at all ages we can apply the *Central Limit Theorem* for both hypotheses and all 3 models giving

z_x ~ Normal (0,1) $x = x_1, x_2, \ldots, x_n$

and the z_x's are mutually independent

χ^2 test

χ^2 test

let

$$X = \sum_{\text{all ages}} z_x^2 \qquad \text{is the } \chi^2 \text{ statistic}$$

if we are comparing experience with a standard table, X can be assumed to have a χ^2 distribution on m degrees of freedom (m being the number of age groups $x=x_1, x_2, \ldots, x_m$)

Large X suggests excessive deviations from the standard table so we test X against the upper 5% point of χ^2_m and the hypothesis fails if $X > \chi^2_{m:0.95}$

χ^2 continued

if we are testing adherence to data of a graduation, X can be assumed to have a χ^2 distribution on $<m$ degrees of freedom

how many fewer than m d-of-f depends on the method of graduation used

we will return to this question next week

The χ^2 test fails to detect some defects

1. a few large deviations offset by many smaller deviations
2. cases where the test is satisfied even though the data does not satisfy the assumptions that underpin χ^2
3. small biases may remain left undetected
4. “runs” or “clumps” where the data set as a whole satisfies the test but there are groups of ages with bias
5. use of squared deviations here z_x^2 means we learn nothing about the direction of individual biases

standardised deviations test

standard deviations test i

Use this to check 1. a few large deviations offset by many smaller deviations

If z_x's are m independent samples from Normal(0,1) this tests for that normality

consider some intervals e.g. (-∞,-2) (-2,-1) (-1,0) (0,1) (1,2) (2,∞)

count number of observed z_x in each interval and compare with what would be expected under Normal(0,1)

standard deviations test ii

Interval	(-∞,-2)	(-2,-1)	(-1,0)	(0,1)	(1,2)	(2, ∞)
Expected number	0.02m	0.14m	0.34m	0.34m	0.14m	0.02m
Actual z_x count	☐	☐	☐	☐	☐	☐

we can then set up a new χ^2 statistic

$$X = \sum_{\text{intervals}} \frac{(\text{actual} - \text{expected})^2}{\text{expected}}$$

X in this example should be χ^2 on 5 degrees of freedom (6 intervals so 5 d-of-f)

alternative

an alternative standard deviations test uses the fact that if they are indeed from Normal(0,1) then ½ the z_x's should be in the interval (-⅔, ⅔)

the number of z_x's outside the range (-⅔, ⅔) should be Binomial(m, 0.5)

We reject the hypothesis if the number outside the range (that is $|z_x| > ⅔$) is in the upper 5% tail of Binomial(m, 0.5)

If we have subdivided our z_x's more and have large enough m, we could similarly check

- 1 in 20 of $|z_x| > 1.96$

  1 in 100 of $|z_x| > 2.57$

  In general we will be suspicious of any z_x's >2 or <-2

signs test

signs test

Let P = number of z_x's which are positive

under our hypothesis, P ~ Binomial(m, ½)

We can test this by finding k* the smallest k for which $\sum_{j=0}^{k} \binom{m}{j} ½^m \geq 0.025$

our (two-tailed) test is satisfied at the 5% level if

$$k^* \leq P \leq m - k^*$$

if m is large we can use the approximation P ~ Normal(½m, ¼m)

cumulative deviations test

cumulative deviations test (i)

Tests for overall bias or a long run of deviations of the same sign

for [graduated data] we have the hypothesis: $d_x \sim \text{Normal}(E^c_x \mathring{\mu}_{x+½} , E^c_x \mathring{\mu}_{x+½})$

then the deviation has the approximate distribution

$$d_x - E^c_x \mathring{\mu}_{x+½} \sim \text{Normal}(0, E^c_x \mathring{\mu}_{x+½})$$

and accumulated deviation over the whole range has the distribution

$$\sum_{\text{all ages}} [d_x - E^c_x \mathring{\mu}_{x+½}] \sim \text{Normal}(0, \sum_{\text{all ages}} E^c_x \mathring{\mu}_{x+½})$$

cumulative deviations test (ii)

we can standardise this

$$\frac{\sum_{\text{all ages}} [d_x - E^c_x \mathring{\mu}_{x+½}]}{\sqrt{\sum_{\text{all ages}} (E^c_x \mathring{\mu}_{x+½})}} \sim \text{Normal}(0, 1)$$

which can be tested in the usual way (with a two-tailed test usually)

this can be applied just to parts of the age range with financial significance rather than all ages if desired

This test will not work if the method of graduation used is designed to produce a cumulative deviation of 0

grouping of signs / Steven's test

grouping of signs test (i)

Detects "clumping" a group of deviations of the same sign

Let G = number of groups of positive z_x's

of m deviations if we say n_1 are positive and n_2 are negative ($n_1+n_2=m$)

under our hypothesis, the n_1 and n_2 deviations are arranged in random order

we calculate the probability of at least G groups of positive deviations given n_1 and n_2 deviations and test this at the 5% level

There are $\binom{n_2+1}{t}$ ways to arrange t ($\leq$G) positive groups amongst n_2 negative signs

grouping of signs test (ii)

and there are $\binom{n_1-1}{t-1}$ ways to arrange n_1 positive signs into t groups

finally there are $\binom{m}{n_1}$ ways to arrange n_1 positive and n_2 negative signs

so the probability of exactly t positive groups is

$$\frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}}$$

grouping of signs test (iii)

Then we seek k* which is the smallest k such that

$$\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} \geq 0.05$$

and the test fails if $G < k^*$

if m large (generally if m>20) we can use the approximation

$$G \sim \text{Normal}\left(\frac{n_1(n_2+1)}{n_1+n_2}, \frac{(n_1 n_2)^2}{(n_1+n_2)^3}\right)$$

serial correlations test

serial correlations test

the last test we consider

Our hypothesis is that if the two sequences z_1 z_2… z_{m-1} and z_2 z_3 … z_m (both of length m-1) are uncorrelated, then so too should be the two sequences (of length m-2) z_1 z_2 … z_{m-2} and z_3 z_4 … z_m

- these are known as ‘lagged sequences’

The test will:

- calculate r_j the correlation coefficient of the j^{th} lagged sequence
- then r_j ~ Normal(0,1/m) under a hypothesis of uncorrelated lagged sequences
- we can test $r_j\sqrt{m}$ against Normal(0,1) where high values indicate a tendency for deviations of the same sign to cluster

the correlation coefficient r_j

$$r_j = \frac{\sum_{i=1}^{M-j} (z_i - \bar{z}^{(1)})(z_{i+j} - \bar{z}^{(2)})}{\sqrt{\left[\sum_{i=1}^{M-j} (z_i - \bar{z}^{(1)})^2 \sum_{i=1}^{M-j} (z_{i+j} - \bar{z}^{(2)})^2 \right]}}$$

where $\bar{z}^{(1)} = \frac{1}{M-j} \sum_{i=1}^{M-j} z_i$ $\qquad$ $\bar{z}^{(2)} = \frac{1}{M-j} \sum_{i=1}^{M-j} z_{i+j}$

if m large r_j simplifies to

if m large $\quad \bar{z}^{(1)} \approx \bar{z}^{(2)} \approx \bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$

and then r_j simplifies to

$$r_j = \frac{\sum_{i=1}^{m-j} (z_i - \bar{z})(z_{i+j} - \bar{z})}{\frac{m-j}{m} \sum_{i=1}^{m} (z_i - \bar{z})^2}$$

statistical tests we have covered

χ^2 test	standardised deviations test	signs test
cumulative deviations test	grouping of signs test	serial correlations test

Graduation: methods and tests

CHRIS SUTTON

NOVEMBER 2023

last week we covered

- comparing experience with standard tables
- introduction to graduation
- desirable features of a graduation
- statistical tests of mortality experience and goodness of fit

this week we will cover

- 4 methods of graduation
- comparison of the methods
- duplicate policies

Methods of graduation

4 methods

graphical methods

by some parametric formula

with reference a standard table

using spline functions

Graphical Graduation methods

graphical method

Historically this is the way that graduation was performed by actuaries before the use of computers

draw a curve through a plot of the crude $\hat{q}_x$ or $\hat{\mu}_x$

- quick
- visual
- with obvious limitations
- actually very hard to do well in practice

graphical method

easier to complete if the plot is on a log-scale

- will be closer to a straight line fit (per Gompertz)

accuracy can be improved by plotting some confidence intervals around the crude estimates e.g. $\hat{q}_x \pm 2\sqrt{d_x} / E_x$ or $\hat{\mu}_x \pm 2\sqrt{d_x} / E_x^c$ and then check the graduated curve stays within the interval about 95% of the time

No longer used in practice

Graduation by parametric formula

Formulae

Parametric approaches seek to fit a certain formula for the force of mortality by finding maximum likelihood estimates for formula parameters.

Recall established formulae for μ_x from week 2

Gompertz	$\mu_x = Bc^x$
Makeham	$\mu_x = A + Bc^x$

Gompertz-type terms have been found to be successful in modelling middle-age and older human life.

UK life assurance industry approach

Recent standard tables for UK life assurance usage have adopted the parametric form:

$$\mu_x = \text{polynomial}① + \exp(\text{polynomial}②)$$

we can see Gompertz and Makeham as special cases of this form

we will consider the parametric curve-fitting technique used in the two most recent CMI standard tables: the “92 series” and the “00 series”

92 series building blocks

- Poisson model
- 4 years of deaths data
- central exposed to risk
- male / female
- different classes of life assurance

formula used

form

$\mathring{\mu}_x = f(\alpha_1, \alpha_2, \ldots \alpha_r, \alpha_{r+1}, \ldots \alpha_{r+s}, x)$

and

polynomial① $= \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \ldots + \alpha_r x^{r-1}$

polynomial② $= \alpha_{r+1} + \alpha_{r+2} x + \alpha_{r+3} x^2 + \ldots + \alpha_{r+s} x^{s-1}$

Poisson likelihood

Poisson likelihood = $(\mu_{x+½})^{d_x} \exp(-\,\mu_{x+½}E^c_x)$ x constants

$= f(\alpha_1, \alpha_2, \ldots \alpha_{r+s}, x+½\,)^{d_x} \exp(-\,f(\alpha_1, \alpha_2, \ldots \alpha_{r+s}, x+½\,)E^c_x)$ x constants

$$\text{Total likelihood} = \prod_{\text{all ages}} f(\alpha_1, \alpha_2, \ldots \alpha_{r+s}, x+½\,)^{d_x} \exp(-\,f(\alpha_1, \alpha_2, \ldots \alpha_{r+s}, x+½\,)E^c_x)$$

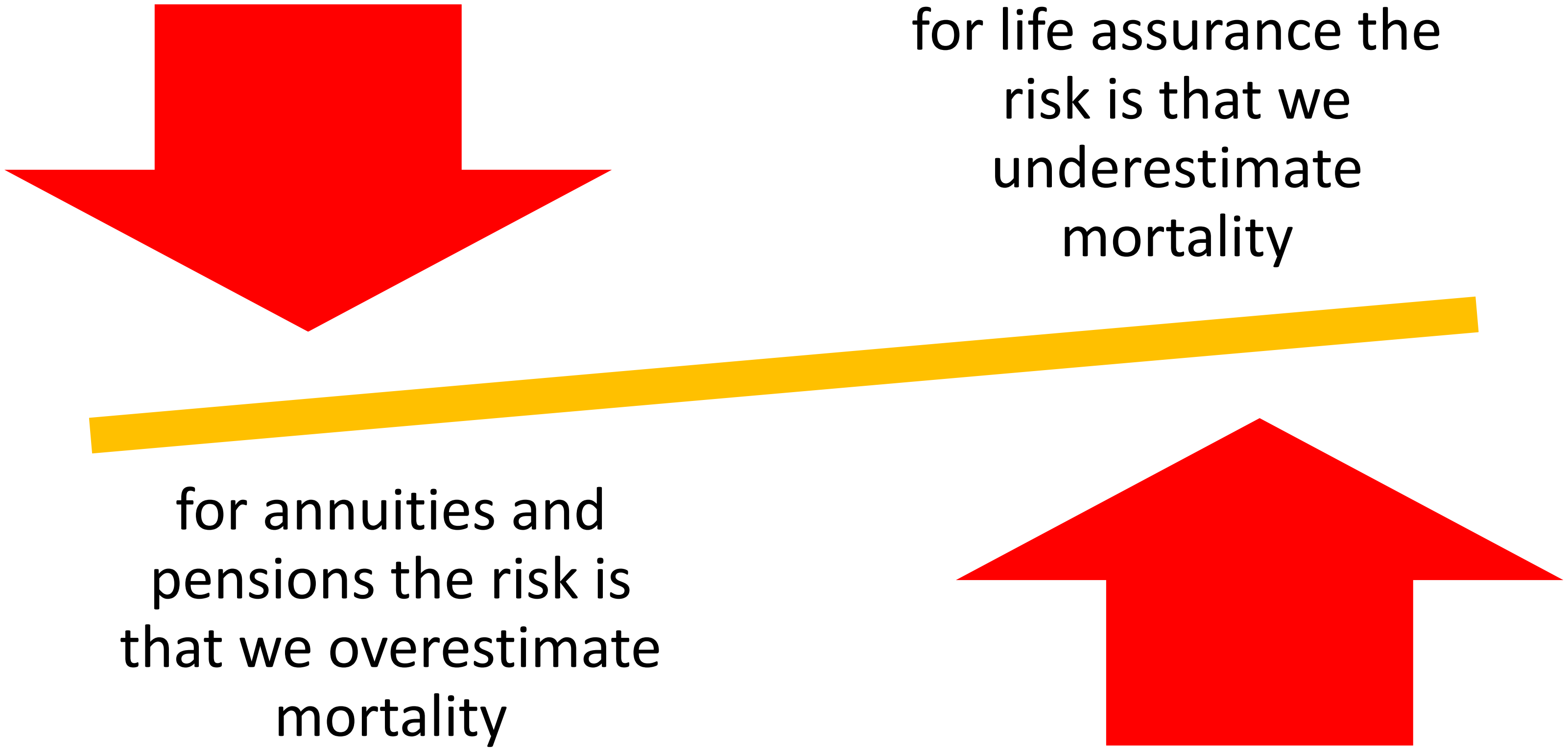
This likelihood is maximised by numerical methods to obtain MLEs for $\alpha_1, \alpha_2, \ldots \alpha_{r+s}$ and hence $\mathring{\mu}_x$

sense checks

with all graduations we should perform some simple sense-checks on the results:

- male mortality > female mortality
- mortality for those with life assurance < general population
- mortality lower for those who recently took out life assurance

financial risks



mortality trends

Graduated mortality tables will be used in practice to estimate **future mortality**

however the table will be based on an investigation of *past mortality*

Mortality trends are important:

- rates have generally been falling over time
- this gives a margin of safety for life assurance business but means that projections of future improvements are needed for annuity and pensions business

We will consider Mortality Projections in more detail in week 11

tests

In practice the CMI and others doing graduation will try fitting different formulae

- Gompertz $\alpha_1 \exp(\alpha_2 x)$
- Makeham $\alpha_1 + \alpha_2 \exp(\alpha_3 x)$
- $\alpha_1 + \alpha_2 \exp(\alpha_3 x + \alpha_3 x^2)$
- …

Then use the statistical tests of ‘goodness-of-fit’ discussed last week to evaluate the different formulae

Graduation by reference to a standard table

the appeal of standard tables

Published life tables (e.g. 92 series, English Life Tables and others)

- are based on a well-defined class of lives
- although never a fully homogeneous group of course

If our mortality experience is from a similarly constituted group, but with fewer lives, the characteristics of mortality in a standard table can be a useful basis for graduation

the approach

if q_x^s or μ_x^s are from a standard table

and $\mathring{q}_x$ or $\mathring{\mu}_x$ are the graduated rates we seek to produce

then we look for some <u>simple</u> function f() such that

$$\mathring{q}_x = f(q_x^s) \qquad \text{or} \qquad \mathring{\mu}_x = f(\mu_x^s)$$

examples:

$$\mathring{q}_x = a + bq_x^s$$

$$\mathring{\mu}_x = \mu_x^s + c$$

$$\mathring{\mu}_x = \mu_{x+d}^s$$

a, b, c, d some constants

where to begin

in looking for suitable function f() we can begin with a simple graph

- plot of $\tilde{q}_x$ against q_x^s might show a linear relationship in q
- plot of $-\log(1-\tilde{q}_x)$ against $-\log(1-q_x^s)$ might indicate a linear relationship in μ
- concentrate on where $\tilde{q}$ has the most data, not at the extreme ages
- once a type of relationship has been detected, best fitting parameters for f() need to be found. There are a number of ways to do this, 2 common ones are:
  - maximum likelihood
  - least squares

fitting parameters

1. <u>maximum likelihood</u>

let $q_x = f(\alpha_1, \alpha_2, .. \alpha_n, q^s_x)$

or $\mu_x = f(\alpha_1, \alpha_2, .. \alpha_n, \mu^s_x)$

where

$\alpha_1, \alpha_2, .. \alpha_n,$ are unknown parameters

2. <u>least squares</u>

parameters are those which minimise

$$\sum_{\text{all ages}} w_x (\hat{q}_x - \mathring{q}_x)^2$$ or the equivalent in μ

where $\{w_x\}$ are weights

- the inverse of the estimated variance of q or μ are good initial candidates for weights

comments

we would normally try a number of candidate functions f(q) or f(μ) and compare them using the statistical tests for 'goodness-of-fit' we introduced last week

as with parametric graduation, we would again want to:

- perform sense checks (male/female; insured/population; recent medical/not)
- consider the direction of financial risks (life assurance / annuity)

Comparison of methods

which method?

parametric method for larger data sets

standard table method for smaller data sets

parametric formula graduation

- this method is a natural extension of the simple probabilistic models for single years of age
- it is straightforward to extend the statistical theory of estimation to several parameters
- computer programs exist to complete the necessary optimisations
- with a small number of parameters, the graduation is sure to be smooth
- a good method for comparing a number of different experiences (different sales channels, time periods, life offices etc) is to fit the same type of parametric formula to each and then the differences between the parameter values fitted will give insight into the differences between the various experiences

parametric (continued)

- this method is a good way to produce standard tables from large amounts of data
- it can be difficult to find one curve that fits all ages well as different features are important at different ages:
  - infant mortality
  - the accident hump
  - exponential mortality in older ages
- the existence of heterogeneity in all data sets increases the difficulty further
- need to be careful with any extrapolations which will be prone to large inaccuracy
  - because most studies are heavy in data in middle ages, many tables and studies are effectively an extrapolation of graduation at very young and very old ages

standard table graduation

- a good way of adapting small data sets for use in practical actuarial work
- the CMI themselves have found that later year's life assurance company data can be fitted as linear functions of the '92 series
- as we begin with smooth standard tables, as long as the f() is simple, the graduation will be smooth too
- the choice of standard table is clearly very important with this method
  - an inappropriate choice of standard table for graduation could lead to the wrong shape for the whole exercise

Graduation using spline functions

splines

An alternative to a single parametric formula

Split the age range of the graduation into different parts

- Fit a polynomial of a specified degree to each part of the range
- The different pieces join at **knots**
- Selecting the knots is not easy
- Certain conditions are imposed on the **spline** functions and on their derivatives to ensure continuity at the knots

Polynomials of degree 3 are common – so called **cubic splines**

The English Life Tables which cover the whole population have used this

Statistical tests of mortality experience

tests

The statistical tests described last week can be used to test a graduation

χ^2 test
- standard deviations
- signs
- cumulative deviations *
- grouping of signs
- serial correlations

* cannot use the *cumulative deviations* test if the method of graduation gives zero cumulative deviation by design

χ^2 test

we need to modify the χ^2 test when testing a graduation:

recall that when testing goodness-of-fit with a standard table, the chi-squared statistic $\sum z_x^2$ had a χ^2 distribution on m degrees of freedom

- m = the number of years of age we are comparing

when comparing actual experience with a standard table, the actual deaths and expected deaths come from different data

however when comparing actual experience with a graduation, the $\{\mathring{q}\}$ or $\{\mathring{\mu}\}$ use the data from actual deaths

- so we have to reduce the number of degrees of freedom in the χ^2 test

how many fewer d-of-f ?

graphical

- very difficult to know how many degrees-of-freedom is right
- suggest lose 2 or 3 d-of-f for every 10 ages fitted corresponding to height / slope / (and maybe) curvature in that range

parametric formula

- lose one degree-of-freedom for each parameter fitted

how many fewer d-of-f ?

standard table

- lose one degree-of-freedom for each parameter fitted
- plus other number of d-of-f due to the constraints imposed by choosing the standard table
- best approach is not to be prescriptive but calculate the $\chi 2$ statistic and allow plenty of margin when comparing with the critical value at m minus no. parameters

splines

- will vary considerably depending on the nature of the polynomials fitted
- certainly one d-of-f for each parameter fitted and one for each knot
- more if placement of the knots a result of examining the data

Duplicate policies

the issue

CMI observes policies not lives, so

E_x = number of policy years (not person years)

d_x = number of policies becoming a claim by death (not number of deaths)

because some people have multiple policies, claim observations are not entirely independent [more than 1 claim might result from 1 death]

this is the problem of **duplicate policies**

framework

assume we observe N lives age [x, x+1]

- no censoring
- no new entrants
- so lives are statistically independent

a proportion π_i of the lives own i policies (i=1,2,3,…)

- so π_2 own 2 policies etc

so total number of policies observed is $\sum_{\text{all } i} i\pi_i N$

in practice, in a real investigation, we will observe the total policies but will not know the $\{\pi_i\}$

each life still have probability q_x of death within the year

distributions

if total number of claims is C we <u>cannot</u> assume C is Binomial because of the non-independence of claims

instead, let D_i = number of deaths amongst the $\pi_i N$ lives with i policies

C_i = number of claims amongst the same lives

then $D_i \sim \text{Binomial}(q_x, \pi_i N)$ as deaths <u>are</u> assumed independent

and expected number of claims

$$E[C] = E[\sum_i C_i] = E[\sum_i iD_i] = \sum_i iE[D_i] = \sum_i i\pi_i N q_x$$

variance

$\text{Var}[C] = \text{Var}[\sum C_i] = \text{Var}[\sum iD_i]$

$= \sum i^2 \text{Var}[D_i]$ because of the independence of deaths

$= \sum i^2 \pi_i N q_x(1-q_x)$

if policies were independent [which they are not] then the variance of C would be $\sum i \pi_i N q_x(1-q_x)$

therefore duplicate policies increase the variance of the number of claims

r_x

Duplicate policies increase Var(C) in the ratio r which varies by age (because the π_i proportions will be different at each age) giving a set of r_x values

$$r = \frac{\sum i^2 \pi_i}{\sum i \pi_i}$$

CMI have estimated some r_x values by tracing duplicate policies in-force in special investigations of a sub-set of their data