# Mandatory group assignment in *Pension Systems* and *Demography and Mortality*

University of Copenhagen, Autumn 2021

24 September – 03 October

## Introduction

This is the compulsory group assignment for the courses *Pension Systems* and *Demography and Mortality*. The solution is to be made in groups of 2-4 persons, and must be submitted as a .pdf-file uploaded to Absalon. You should also submit the R code (either as an .R-file or .txt-file) used to answer the problems.

The assignment is made available on Absalon Friday, September 24. The deadline for uploading the solution is Sunday, October 3 at 23:59. Only one person from each group should upload the solution. The names of all group members must be clearly stated on the front page. If the assignment is not approved, you will be notified on October 6, and an improved version must be handed in no later than October 10 at 23:59.

Regarding style, references to syllabus and lecture notes should be concise, and plots should all bear readable and informative axis labels, legends and titles. The solution must be written in English. Remember to write in clear text what you do and the conclusions you draw along the way – though they may seem "obvious" from a plot or a table.

Clarifying questions should be asked on Absalon or during the exercise class on September 29.

## Part 1: Decomposing differences in life expectancy

Methods of decomposition to foster a better understanding of historical evolutions are at the heart of demography. In Part 1, we will look at a method for decomposing life expectancy differences between two populations in terms of age group contributions.

Let $\mu_i(x)$ denote the force of mortality at age $x$ in population $i = 1, 2$, and assume that $0 < \mu_2(x) < \mu_1(x)$ for all $x$. For populations $i = 1, 2$ define

- The survival function: $S_i(x) = \exp(-\int_0^x \mu_i(u)\,\mathrm{d}u)$;
- The conditional survival function: $S_i(y|x) = S_i(y)/S_i(x)$ for $y \geq x$;
- The remaining life expectancy: $e_i(x) = \int_x^\infty S_i(z|x)\,\mathrm{d}z = \int_0^\infty S_i(x+u|x)\,\mathrm{d}u$;
- The remaining life expectancy before age $y$: $e_i(x, y) = \int_x^y S_i(z|x)\,\mathrm{d}z$.

Preston et al. (p. 64) describes the two main decompositions: a continuous and a discrete approach. Both approaches are linked to the quantity

$$A(x,y) = \int_x^y (\mu_1(u) - \mu_2(u)) \exp\left(\int_x^u (\mu_1(v) - \mu_2(v))\,\mathrm{d}v\right) S_1(u) e_1(u)\,\mathrm{d}u, \tag{1}$$

for two ages $x, y \in \mathbb{R}_+$ satisfying $x < y$.

**Question 1.1:** Show Pollard's continuous decomposition $e_2(0) - e_1(0) = A(0, \infty)$.

Arriaga's discrete decomposition is given in Preston et al. (p. 64, formula 3.11). It states that the effect of a difference in mortality rates between ages $x$ and $y$ on the life expectancy at birth can be expressed as

$$_{y-x}\Delta_x = \frac{l_x^1}{l_0^1}\left(\frac{_{y-x}L_x^2}{l_x^2} - \frac{_{y-x}L_x^1}{l_x^1}\right) + \frac{T_y^2}{l_0^1}\left(\frac{l_x^1}{l_x^2} - \frac{l_y^1}{l_y^2}\right), \tag{2}$$

with notation being consistent with Preston et al.[1] The first term in the right-hand side of (2)

$$D(x,y) = \frac{l_x^1}{l_0^1}\left(\frac{_{y-x}L_x^2}{l_x^2} - \frac{_{y-x}L_x^1}{l_x^1}\right), \tag{3}$$

expresses the direct effect of changing mortality between age $x$ and $y$, while the second term in the right-hand side of (2) can be expressed as an indirect effect and an interaction effect

$$I(x,y) = \frac{T_y^1}{l_0^1}\left(\frac{l_x^1}{l_x^2}\frac{l_y^2}{l_y^1} - 1\right), \tag{4}$$

$$X(x,y) = \left(\frac{T_y^2 - T_y^1 \frac{l_y^2}{l_y^1}}{l_0^1}\right)\left(\frac{l_x^1}{l_x^2} - \frac{l_y^1}{l_y^2}\right). \tag{5}$$

Note that $_{y-x}\Delta_x = D(x,y) + I(x,y) + X(x,y)$. We want to express these effects in terms of our usual notation.

**Question 1.2:** Rewrite (3)–(5) in terms of $S_1, S_2, e_1$, and $e_2$ to get expressions for the direct (D), the indirect (I), and the interaction (X) effect.

**Question 1.3:** Show that the sum of Arriaga's direct and indirect effects correspond to (1), i.e. show $D(x,y) + I(x,y) = A(x,y)$ for any $x$ and $y$ with $x < y$.

**Question 1.4:** Let $0 = x_0 < x_1 < \cdots < x_n = \infty$ and show that

$$\sum_{i=1}^{n} [D(x_{i-1}, x_i) + I(x_{i-1}, x_i) + X(x_{i-1}, x_i)] = e_2(0) - e_1(0). \tag{6}$$

[1] All quantities in (2) can be found in Preston et al. p. 69.

How does Arriaga's decomposition relate to that of Pollard? Specifically, why is Arriaga's approach described as being discrete?

**Question 1.5:** Justify the approximation

$$e_2(0) - e_1(0) \approx \int_0^\infty (\mu_1(x) - \mu_2(x))S_1(x)e_1(x)\,\mathrm{d}x, \tag{7}$$

by considering either the limit of $\sum_{i=1}^n [D(x_{i-1}, x_i) + I(x_{i-1}, x_i)]$ as the age intervals become finer and finer, or via Taylor expansion of the exponential term in $A(0, \infty)$.

**Question 1.6:** Describe in words the effects of mortality change on life expectancy via $D$, $I$, and $X$. That is, describe the effects captured by $D$, $I$, and $X$, respectively.

Like the rest of the Western world, Denmark has experienced huge gains in life expectancy at birth over the past two centuries. In the following, we are interested in studying the contribution to $e(0)$ from mortality improvements in the age groups 0, 1–10, 11–20, …, 81–90, 90+ during the periods $1835-1900$, $1900-1950$, and $1950-2020$. First, let us get a better grasp on the historical evolution of life expectancy.

**Question 1.7:** Compute and visualize the period life expectancy at birth for both sexes over the period $1835 - 2020$.

**Question 1.8:** Compute and visualize, using e.g. a bar plot, the average annual rate of mortality improvement for both sexes and the age groups and periods stated above.

**Question 1.9:** What does this preliminary analysis tell us about the evolution of $e(0)$?

For a more thorough analysis, we will apply Arriaga's decomposition method.

**Question 1.10:** Implement a function in R that computes (2) for given ages $x$ and $y$ and mortality profiles $\mu_1$ and $\mu_2$. (Hint: Verify your implementation using Equation (6).)

**Question 1.11:** Decompose the gains in life expectancy at birth for females and males using Arriaga's decomposition (2) for the age groups and time periods stated above. Visualize the results using e.g. a stacked bar plot. Based on this plot, expand on the preliminary analysis you made in Question 1.9 describing the evolution of $e(0)$ in Denmark. In particular, you should comment on the additional insights you have gained from applying Arriaga's decomposition.

# Part 2: The Cairns-Blake-Dowd model

Consider the mortality model of Cairns et al. (2006)[2], commonly referred to as the CBD-model. As usual, we assume that $\mu(x,t)$ is constant on age-period cells $[x, x+1) \times [t, t+1)$ for integer ages $x$ and calendar years $t$. Suppose that we have death counts $D(x,t)$ and exposures $E(x,t)$ available for ages $x \in \{x_{\min}, \dots, x_{\max}\}$ and $t \in \{t_{\min}, \dots, t_{\max}\}$ available. Denote by $q(x,t) = 1 - \exp(-\mu(x,t))$ the probability that an individual aged $x$ at time $t$ dies before $t+1$. The CBD-model is then

$$\operatorname{logit} q(x,t) := \log\left(\frac{q(x,t)}{1-q(x,t)}\right) = \alpha_t + \beta_t(x - \bar{x}), \tag{8}$$

where $\bar{x} = \sum_{x=x_{\min}}^{x_{\max}} \frac{x}{x_{\max}-x_{\min}+1}$ denotes the mean of the age-span considered and $\alpha_t$ and $\beta_t$ are time-varying parameters to be estimated from data.

In the original work, (8) is fitted by an OLS-regression on the logit transformed (empirical) death probabilities. Instead, we could use the Poisson assumption

$$D(x,t)|E(x,t) \overset{\text{indep.}}{\sim} \text{Poisson}(E(x,t)\mu(x,t)), \tag{9}$$

to estimate the model's parameters.

**Question 2.1:** What are the pros and cons of using (9) to estimate the model compared to the original OLS approach?

**Question 2.2:** Using data for England and Wales (HMD-code: *GBRCENW*), estimate the CBD-model under the Poisson assumption (9) for both sexes, ages 20–100, and calendar years 1970–2018. Describe how you fitted the model, make a plot of the resulting parameters, and investigate the model's fit to data. Comment on the output.

For the estimation, you can find inspiration in the paper by Brouhns et. al. (2002)[3] (in the Week 2 folder on Absalon) for how the Poisson Lee-Carter model is fitted using the Newton-Raphson method. You are not required to use this method though.

**Question 2.3:** As in the original work (Cairns et al. (2006), p. 6), we use a two-dimensional random walk with drift to forecast the model. Estimate the model for each sex and state the parameters. In your opinion, is the use of a random walk structure justifiable?

**Question 2.4:** Specify the distribution of $(\alpha_{t_{\max}+h}, \beta_{t_{\max}+h})|(\alpha_{t_{\max}}, \beta_{t_{\max}})$ for some $h \in \mathbb{N}_+$ and give a two-sided pointwise 95%-CI. Plot the estimated parameters along with the median forecast and the 95%-CI for $h = 100$. Comment on the output.

Denote by $e_p^s(x,t)$ the period remaining life expectancy for sex $s \in \{\text{f, m}\}$ (denoting female and male, respectively) at age $x$ in year $t$. You may assume a maximum attainable age of $\omega = 100$ in the following.

[2] Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, 73(4):687–718.

[3] Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance Mathematics and Economics*, 31(3):373–393.

**Question 2.5:** Simulate 10,000 period life expectancy trajectories using the CBD model. Plot the resulting forecasting distribution of $e_p^f(60, t)$ and $e_p^m(60, t)$ for $t = \{2025, 2050, 2100\}$, i.e. for a short-, medium- and long horizon, using e.g. a density plot. Comment on the output.

**Question 2.6:** Repeat Question 2.5 for the Lee-Carter model calibrated to the same data as you used to estimate the CBD-model in Question 2.2.

**Question 2.7:** Discuss how the CBD model differs from the Lee-Carter model. In particular, discuss the structural differences between the two models, the differences in their ability to fit historical data, and the differences between their forecasts.