PARTE IV

MÉTODOS AVANÇADOS DE ANÁLISE DEMOGRÁFICA

CAPÍTULO 17

MÉTODOS MATEMÁTICOS NA ANÁLISE DE POPULAÇÃO

Até aqui a discussão dos conceitos e métodos se fez com um uso mínimo de recursos matemáticos, estatísticos e informáticos. Muitos destes conceitos podem ser entendidos, pelo menos na sua essência, sem precisar de tais recursos. Entretanto, na medida em que se avança na análise quantitativa dos fenômenos demográficos, se faz necessário lançar mão de um acervo de instrumentos um pouco mais sofisticados. Muitos artigos publicados em revistas profissionais hoje em dia partem do suposto de que os seus leitores são familiarizados com esses instrumentos, de forma que pode tornar-se difícil acompanhar a exposição sem dominá-los. Trata-se particularmente dos seguintes recursos:

1. A língua de programação "R";
2. Introdução a alguns outros pacotes para a análise demográfica;
3. O cálculo diferencial e integral;
4. Princípios da álgebra matricial; e
5. Conceitos básicos de estatística, particularmente a estatística bayesiana.

Uma introdução aos métodos de interpolação, que exige um pouco mais espaço, será oferecida separadamente no próximo capítulo. Todos estes conceitos e técnicas foram desenvolvidos e possuem aplicações muito além da demografia. Portanto, a discussão mais detalhada de cada um

deles não cabe num livro como este. As seções que seguem se limitam a alguns princípios básicos e alguns exemplos da relevância dos métodos para a análise demográfica. Além disso serão providenciadas algumas referências a livros e documentos que tratam da matéria de uma forma mais completa do que será possível dentro das limitações do presente texto.

17.1 A LÍNGUA DE PROGRAMAÇÃO "R"

Até agora, vários procedimentos de cálculo demográfico foram ilustrados com o uso de EXCEL. De fato, EXCEL é um recurso muito útil e, em muitos casos, amplamente suficiente. Entretanto, desde o início da década de 2010 a tendência que se observa na comunidade acadêmica nos países mais desenvolvidos – e cada vez mais nos países em desenvolvimento também – é a substituição de EXCEL por "R", a não ser para as tarefas mais básicas. As vantagens de "R" são as seguintes:

1. Diferentemente de EXCEL, que é um programa comercial da Microsoft, pelo qual o usuário em princípio precisa pagar cada vez que instala uma versão mais atualizada, o "R" (e o seu complemento, RStudio) está disponível gratuitamente no sítio da CRAN (de "Comprehensive R Archive Network"), para ser descarregado cada vez que o usuário sinta a necessidade de atualizá-lo.

2. O EXCEL só funciona no sistema operacional Windows. Muitos profissionais da área de ciências exatas, estatística e informática hoje em dia preferem outros sistemas operacionais como UNIX ou Ubuntu que não dão suporte para EXCEL mas sim para "R".

3. O "R" é um programa de fonte aberta que pode ser estendida pelas contribuições dos próprios usuários. Efetivamente, muitas extensões (funções, bibliotecas de procedimentos) têm sido contribuídas por usuários e estão disponíveis para a comunidade em geral. Em muitos casos, estes procedimentos inclusive superam a qualidade de programas comerciais disponíveis para a execução de tarefas semelhantes.

4. Especificamente, existem bibliotecas de procedimentos para a produção de gráficos que superam a qualidade dos gráficos que podem ser produzidos em EXCEL.

5. As facilidades para a leitura de diferentes tipos de arquivos (ficheiros) de dados em "R" são muito mais abrangentes do que em EXCEL.

6. Pela sintaxe compacta de "R", programas escritos em "R" geralmente são mais curtos e eficientes do que programas semelhantes feitos em EXCEL ou Visual BASIC, principalmente se incorporam funções ou procedimentos já existentes (estatísticos ou outros) desenvolvidos por outros usuários.

7. O "R" tem vários recursos para a leitura de arquivos (ficheiros) externos, numa variedade de formatos.

8. Mesmo se não tivesse as vantagens anteriores, o fato de que cada vez mais as análises estatísticas de dados demográficos são apresentadas em "R" obriga o professional da área a entender essa linguagem.

Por outro lado, a desvantagem de "R" para iniciantes é de ser menos intuitivo, com comandos mais complexos, do que EXCEL de modo que exige um investimento inicial de tempo maior para aprendê-lo.

O R pode ser usado interativamente, por meio de comandos que recebem uma resposta imediata, ou por meio de funções, que são conjuntos de expressões que levam a um resultado, dependendo dos valores iniciais escolhidos para o cálculo, e que podem ser executadas no seu conjunto quantas vezes for necessário. Por exemplo, ao entrar em R, o sistema mostra o símbolo ">", como sinal de que espera um comando. Tal comando poderia ser algo tão simples como a adição de dois números. Ao digitar

```
> 2.66+3.14
```
(17.1)

o sistema responde

```
[1] 5.8
```
(17.2)

Também se pode definir uma variável e passar um valor a ela, usando a seguinte sintaxe:

```
> Idade <- 56
```
(17.3)

A partir deste ponto, o nome "Idade" pode ser usado em outras expressões e estas serão calculadas com o valor de 56 anos para esta variável. Uma característica menos usual de "R" é que aceita como valores não só números simples, mas também vetores, ou seja, sequências de números. O comando usado para tal propósito é o seguinte:

```
> Idade <- c(0,1,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80)
```
(17.4)

As posições dentro do vetor podem ser identificadas por meio de índices em parênteses quadradas. Por exemplo, Idade[4] se refere ao quarto número na sequência, que é 10. Para saber o comprimento de um vetor, usa-se o comando length:

```
> length(Idade)
```
(17.5)

ao qual o sistema responde

```
[1] 18
```
(17.6)

O mesmo resultado de (17.4) também pode ser obtido mais sinteticamente da seguinte forma:

```
> Idade <- -1:16·5; Idade[1] <- 0; Idade[2] <- 1
```
(17.7)

Aqui já se pode perceber a tendência de programas em "R" de gerar expressões muito compactas mas ao mesmo tempo um pouco críticas.

Para ter uma noção de como "R" pode ser usado para programar problemas comuns de análise demográfica, se desenvolve a seguir uma aplicação para calcular a esperança de vida ao nascer a partir de uma série de Taxas Específicas de Mortalidade, usando como exemplo a sequência da Tabela 8.4, de São Tomé & Príncipe. As probabilidades de morte ${}_nq_x$ serão estimadas usando a fórmula de Greville (8.26). Cabe alertar que o exemplo é puramente ilustrativo, pois em realidade o uso da fórmula de Greville não é aconselhável para as idades de 0 e 1-4 anos, mas numa aplicação real tais detalhes poderiam ser resolvidos posteriormente. O que interessa aqui é o procedimento.

O primeiro passo é definir um vetor com o comprimento dos intervalos etários (nx), compatível com o vetor Idade:

```
> nx <- rep(5,times=18); nx[1] <- 1; nx[2] <- 4; nx[18] <- 15
```
(17.8)

Uma maneira mais simples para conseguir o mesmo resultado é

```
> nx <- diff(Idade); nx[18] <- 15
```
(17.9)

Aqui o comprimento do intervalo aberto final foi definido arbitrariamente como 15. O próximo passo consiste em passar os valores das Taxas Específicas de Mortalidade da Tabela 8.4 para um vetor Mx:

```
> Mx <- c(0.0231,0.0031,0.000707,0.001353,0.001571,0.00338,0.002632,0.003274,
0.006269, 0.006473,0.007764,0.014355,0.013627,0.032319,0.031169,0.054313,0.072
998,0.170058)
```
(17.10)

Não existe limite para o número de caracteres por linha. Para melhorar a legibilidade, pode-se introduzir alguns espaços.

A fórmula de Greville depende de uma constante c, cujo valor pode variar entre 1,08 e 1,10. Aqui se supõe que c = 1,09:

```
> c <- 1.09
```
(17.11)

Agora vem o passo essencial, a conversão de Mx para qx, usando a fórmula de Greville:

```
> qx <- Mx / (1/nx + Mx * (0.5 + (nx/12)*(Mx – log(c))))
```
(17.12)

Nota-se que a fórmula é aplicada a todas as suas 18 componentes, sem precisar de uma instrução explícita para tal. Também nota-se que "R" usa o símbolo log para o logaritmo natural (ln), ou seja o logaritmo de base *e = 2,71828*.....

A partir do vetor qx se constrói o vetor de sobrevivência, usando o recurso de "looping" com uma variável índice i:

```
> lx <- rep(100000, times=18); for (i in 2:18) lx[i] <- lx[i-1] * (1-qx[i-1])
```
(17.13)

Nota-se que, antes de preencher o vetor com os seus devidos valores, primeiro é preciso criá-lo, o qual se faz na primeira parte da expressão, onde se define o vetor inicialmente como uma sequência de 18 vezes o número 100.000.

O número de óbitos em cada intervalo de idade se constrói de forma análoga:

```
> dx <- rep(l[18], times=18); for (i in 1:17) dx[i] <- lx[i]-lx[i+1]
```
(17.14)

ou, usando o comando diff,

```
> dx <- -diff(lx); dx[18] <- lx[18]
```
(17.15)

O vetor de anos-pessoa se cria por uma simples divisão do vetor dx com o vetor Mx. Nota-se que o "R" distingue entre lx e Lx como dois vetores diferentes.

```
> Lx <- dx / Mx
```
(17.16)

O vetor acumulado Tx se constrói de forma semelhante com dx:

```
> Tx <- rep(Lx(18), times=18); for (i in 17:1) Tx(i) <- Tx(i+1)+ Lx(i)
```
(17.17)

Finalmente, a esperança de vida ao nascer é dada por:

```
> e0 <- Tx[1] / lx[1]
```
(17.18)

Agora, ao digitar "e0" o espaço de comando, o sistema responde

```
[1]  65.347
```
(17.19)

Se o cálculo da esperança de vida ao nascer for uma tarefa recorrente, vale a pena juntar todos os comandos numa função, da seguinte forma:

```
greville <- function(Mx,c) {
   # Função ilustrativa para calcular uma tábua de vida usando o método de Greville
   (ver 8.18) a partir de nMx, com uma constante de ajuste c
   ult <- length(Mx)
   Idade <- -1:(ult-2)*5; Idade[1] <- 0; Idade[2] <- 1
   nx <- diff(Idade); nx[ult] <- 15
   qx <- Mx / (1/nx + Mx*(.5 + (nx/12)*(Mx-log(c[1]))))
   lx <- rep(100000, times=ult)
   for (i in 2:ult) lx[i] <- lx[i-1]*(1-qx[i-1])
   dx <- -diff(lx); dx[ult] <- lx[ult]
   Lx <- dx / Mx; Tx <- rep(Lx[ult], times=ult)
   for (i in (ult-1):1) Tx[i] <- Tx[i+1]+Lx[i]; qx[18] <- 1; ex <- Tx/lx
   result <- cbind(Idade, Mx, qx, lx, dx, Lx, Tx, ex)
   format(result, scientific=FALSE)
   result
}
```
(17.20)

Ao digitar o comando

```
> greville(M,1.09)
```
(17.21)

a resposta que aparece agora é

	Idade	Mx	qx	lx	dx	Lx	Tx	ex
[1,]	0	0.023100	0.022838983	100000.00	2283.8983	98870.06	6534700.1	65.347001
[2,]	1	0.003100	0.012327801	97716.10	1204.6246	388588.59	6435830.0	65.862533
[3,]	5	0.000707	0.003529206	96511.48	340.6089	481766.51	6047241.4	62.658262
[4,]	10	0.001353	0.006743802	96170.87	648.5572	479347.56	5565474.9	57.870694
[5,]	15	0.001571	0.007826429	95522.31	747.5986	475874.33	5086127.3	53.245438
[6,]	20	0.003380	0.016768086	94774.71	1589.1905	470174.71	4610253.0	48.644336
[7,]	25	0.002632	0.013079926	93185.52	1218.8597	463092.61	4140078.3	44.428343
[8,]	30	0.003274	0.016246212	91966.66	1494.1098	456356.09	3676985.7	39.981724
[9,]	35	0.006269	0.030893070	90472.55	2794.9749	445840.62	3220629.6	35.597864
[10,]	40	0.006473	0.031883318	87677.58	2795.4521	431863.45	2774789.0	31.647646
[11,]	45	0.007764	0.038128290	84882.13	3236.4103	416848.31	2342925.5	27.602107
[12,]	50	0.014355	0.069432382	81645.72	5668.8565	394904.67	1926077.2	23.590671
[13,]	55	0.013627	0.066021787	75976.86	5016.1280	368102.15	1531172.5	20.153144
[14,]	60	0.032319	0.150017946	70960.73	10645.3830	329384.67	1163070.4	16.390339
[15,]	65	0.031169	0.145059738	60315.35	8749.3285	280706.10	833685.7	13.822116
[16,]	70	0.054313	0.239860917	51566.02	12368.6726	227729.51	552979.6	10.723722
[17,]	75	0.072998	0.309185001	39197.35	12119.2316	166021.42	325250.1	8.297759
[18,]	80	0.170058	1.000000000	27078.11	27078.1148	159228.70	159228.7	5.880347

Dispor dos cálculos no formato de uma função permite substituir diferentes valores, para verificar a sensibilidade do procedimento a variações nos parâmetros. Por exemplo, substituindo diferentes valores de *c*, se verifica que as diferenças nos resultados são mínimas. Evidentemente, o mesmo poderia ser feito por meio de uma planilha em EXCEL, mas é impressionante como o "R" permite dar todas as instruções necessárias em só 15 linhas, incluindo o cabeçalho de identificação.

A função, tal como foi redigida em (17.20), assume que o primeiro intervalo etário seja o de 0 anos, o segundo 1-4 e que todos os outros tenham um comprimento de 5 anos. Mas ela também introduz algumas generalizações e extensões em comparação com as fórmulas anteriores. A primeira linha contém um comentário (identificado pelo símbolo "#"), para descrever a finalidade da função. Ao introduzir a variável top na segunda linha, a função se torna independente do formato de 18 idades usado anteriormente. A programa lê o vetor *M* e estabelece o número de faixas etárias necessárias para a análise. Em vez de mostrar só uma esperança de vida específica, o resultado da função é uma tábua de vida inteira, com todas as suas componentes. A sintaxe de "R" facilita este tipo de procedimento, já que aceita matrizes, vetores e números da mesma forma, como resultados de um procedimento de cálculo. Por exemplo, o comando *result <- cbind(Idade, Mx, qx, lx, dx, Lx, Tx, ex)* dá a instrução para considerar os vetores *Idade, Mx, qx, lx, dx, Lx, Tx* e *ex* como colunas de uma única matriz chamada *result*. O último comando da função pede para mostrar esta matriz por inteiro, sem necessidade de identificar o número de filas e colunas. Daria para melhorar o procedimento ainda mais, incluindo instruções para lidar com as primeiras duas e as últimas duas faixas etárias onde a fórmula de Greville reconhecidamente não funciona muito bem, mas tais detalhes não serão abordados aqui.

Para maiores detalhes sobre a língua de programação já existe um grande número de manuais e orientações, alguns inclusive em português, como Ribeiro (2015). O número de manuais e orientações existentes em inglês é demasiado grande para listar aqui. Muitos estão disponíveis na internet, como a introdução à análise demográfica de James H. Jones, da Universidade de Stanford, que pode ser encontrada em https://jhj1.people.stanford.edu/sites/g/files/sbiybj3091/f/file/jones-rintro050508.pdf. Também existem vários livros que aplicam "R" a diferentes tipos de análise demográfica e técnicas afins (por exemplo, Beyersman e Allignol, 2011; Moore, 2016; Willekens, 2014). Estes últimos, devido aos temas tratados, geralmente se destinam a usuários mais avançados. O sítio web da CRAN (https://cran.r-project.org/web/packages/available_packages_by_name.html) contém mais de 10.000 pacotes e procedimentos de aplicativos em "R" contribuídos por usuários. Os seguintes são alguns dos aplicativos mais relevantes do ponto de vista demográfico, com os nomes dos seus criadores:

bayesLife	Faz projeções probabilísticas da esperança de vida para todos os países do mundo, usando um modelo Baysiano hierárquico (Hana Ševčiková, Adrian Raftery, Jennifer Chunn).
bayesPop	Produz projeções de população para todos os países do mundo usando várias componentes, tais como a Taxa de Fecundidade Total e esperança de vida (Hana Sevčiková, Adrian Raftery, Thomas Buettner).

bayesTFR	Faz projeções probabilísticas da Taxa de Fecundidade Total para todos os países do mundo, usando um modelo Baysiano hierárquico (Hana Ševčiková, Leontine Alkema, Adrian Raftery, Bailey Fosdick, Patrick Gerland).
Biograph	Calcula taxas de transição a partir das transições e exposições, com gráficos e indicadores de ciclo de vida. O pacote estrutura os dados para a modelação estatística e demográfica por múltiplos estados das histórias de vida (Frans Willekens).
childhoodmortality	Calcula taxas de mortalidade na infância (neonatal, pós-neonatal, infantil, de crianças e abaixo de 5) usando microdados dos DHS. O pacote foi desenvolvido segundo a metodologia descrita em DHS *Guide to Statistics*. Especificamente, se usa uma abordagem de tábuas de vida baseadas em coortes sintéticas, combinando as probabilidades de morte para segmentos etários com experiências reais de coorte para obter a mortalidade de faixas etárias mais convencionais. Os erros padrão para as estimativas de mortalidade são computados usando o método "jackknife" de replicação repetida descrita no Apêndice *Estimates of Sampling Errors* dos Informes finais das pesquisas DHS (Casey Breen).
DBKGrad	Graduação não paramétrica de taxas de mortalidade usando um estimador fixo ou adaptivo do tipo "beta kernel" (Angelo Mazza, Antonio Punzo).
DDM	Um conjunto de métodos aplicados a dois censos para estimar o grau de cobertura do registro de óbitos de uma população. Os métodos incluem o Método Generalizado de Balanço de Crescimento (GGB), o Método Sintético de Gerações Extintas (SEG) e um híbrido dos dois, GGB-SEG. Cada método oferece uma estimação automática, mas os usuários também podem especificar parâmetros exatos ou usar uma interface gráfica para adivinhar parâmetros do modo tradicional, se assim se deseja (Tim Riffe, Everton Lima e Bernardo Queiroz).
demogR	Constrói e analisa modelos matriciais de população em "R" (James Holland Jones).
demography	Calcula funções para a análise demográfica, inclusive tábuas de vida, o método Lee-Carter, faz análise funcional de taxas de mortalidade, taxas de fecundidade, números de migração líquida, e prognósticos estocásticos de população (Rob J. Hyndman, com contribuições de Heather Booth, Leonie Tickle e John Maindonald).
europop	Contém estimativas das populações de todas as cidades europeias com 10.000 habitantes ou mais no período de 1500-1800, baseado nos dados adaptados de *European Urbanization, 1500-1800* (1984) (Matthew Lincoln, Jan De Vries).

GENLIB	Análise de dados genealógicos, incluindo estatísticas descritivas (por exemplo, coeficientes de parentesco e consanguinidade) e simulações de omissão de genes (Louis Houde, Jean-François Lefebvre, Valéry Roy-Lagace, Sébastien Lemieux, Michael J. Fromberger, Marie-Helène Roy-Gagnon).
Giza	Fornece uma maneira simples para criar pirâmides etárias múltiplas numa única janela de gráficos, aproveitando o potencial do pacote "lattice". É uma maneira conveniente de visualisar dados longitudinais agrupados (isto é, estruturados por idade e educação) (Erich Striessnig).
HPbayes	Fornece todas as funções necessárias para estimar os 8 parâmetros do modelo de mortalidade de Heligman-Pollard usando um procedimento Bayesiano com IMIS e converte os parâmetros em probabilidades específicas por idade e a tábua de vida correspondente (David J. Sharrow).
Ilc	Ajusta uma classe de modelos de mortalidade do tipo Lee-Carter usando um algoritmo iterativo (Zoltan Butt, Steven Haberman, Han Lin Shang).
LexisPlotR	Funções para desenhar diagramas de Lexis para propósitos demográficos (Philipp Ottolinger, Marieke Smilde-Becker).
lifecontingies	Classes e métodos que permitem o manejo de tábuas de vida e tábuas atuariais, incluindo tábuas de múltiplos decrementos. Além disso, o programa contém funções para cálculos de matemática demográfica, financeira e atuarial sobre contingências de tábuas de vida para efeitos de seguros (Giorgio A. Spedicato, Reinhold Kainhofer, Kevin J. Owens, Christophe Dutang, Ernesto Schirmacher e Gian Paolo Clemente).
LifeTables	Sistema de tábuas de vida modelo de dois parâmetros baseado na Human Mortality Data Base (David J. Sharrow e Hana Ševčiková).
Maples	Método geral para estimar perfis etários com base nos dados típicos obtidos em inquéritos demográficos, com o objetivo de obter padrões etários suavizados, conjuntamente com os riscos relativos de covariados fixos e variáveis no tempo (Roberto Impicciatore).
migest	Métodos indiretos para estimar fluxos migratórios bilaterais com dados parciais ou deficientes. Os métodos apresentados poderiam ser relevantes para outras situações de dados categóricos com dados não migratórios, onde, por exemplo, os totais marginais são conhecidos mas há informação incompleta sobre os dados bilaterais (Guy J. Abel).
migration.indices	Fornece vários índices, como a Taxa Bruta de Migração, diferentes tipos de índices de Gini ou o Coeficiente de Variação, entre outros, para medir a (des) igualdade da migração (Lajos Bálint e Gergely Daróczi).

mortAAR	Analisa dados arqueológicos de mortalidade. Aceita dados demográficos em vários formatos e mostra o resultado numa tábua de vida convencional, além de gráficos de diferentes índices (percentagem de óbitos, sobrevivência, probabilidade de morte, esperança de vida, percentagem da população (Nils Mueller-Scheessel, Martin Hinz, Clemens Schmid, Christoph Rinne, Daniel Knitter, Wolfgang Hamer, Dirk Seidensticker, Franziska Faupel, Carolin Tietze e Nicole Grunert).
MortalityLaws	Ajusta as 'leis' de mortalidade mais comuns e constrói tábuas de vida completas e abreviadas a partir de vários índices de entrada. Também fornece uma função elegante para descarregar dados da Human Mortality Database <http://www.mortality.org> (Marius D. Pascariu e Vladimir Canudas-Romo).
MortalitySmooth	Ajusta contagens com distribuição de Poisson em uma ou duas dimensões usando P-splines especificamente configurados para dados de mortalidade. É possível incorporar variação adicional de Poisson e projetar os resultados. O programa facilita a colheita de dados de mortalidade e a sua seleção por país, sexo, idade e anos (Carlo G. Camarda).
MortCast	Estima taxas específicas de mortalidade e as projeta por meio dos métodos de Kannisto, Lee-Carter e outros relacionados (Hana Ševčiková, Nan Li e Patrick Gerland).
MSDem	Executa projeções de população por múltiplos estados, tendo a idade, sexo e nível de educação como estratos potenciais e a região e área de residência como suas unidades geográficas potenciais (Marcus Wurzer, Samir K. C. e Markus Speringer).
POPdemog	Faz gráficos de fenômenos demográficos para uma ou várias populações a partir de dados coalescentes de simulação. Atualmente o programa funciona com os programas de simulação 'ms', 'msHot', 'MaCS', 'msprime', 'SCRM' e 'Cosi2'. O programa não verifica se os dados simulados são corretos, mas supõe que os dados de entrada tenham sido validados pelos próprios programas pelos quais foram gerados (Ying Zhou).
pyramid	Produz pirâmides de população baseadas em (1) data.frame ou (2) vetores. O primeiro é chamado pyramid() e o segundo pyramids() (Minato Nakazawa).
ROMIplot	Possibilita a representação gráfica de mapas de superfície de Lexis (mapas de "calor") que mostram taxas de melhoria da mortalidade. Os dados brutos que serão representados podem ser lidos da Human Mortality Database, usando um programa originalmente escrito por Tim Riffe (Roland Rau, Tim Riffe).
smoothAPC	Suavização/graduação de taxas de mortalidade como soma de quatro componentes: uma função bivariada suave da idade e o tempo, funções univariadas suaves para modelar o efeito de coorte, funções univariadas suaves

	para modelar o efeito de período e erros aleatórios (Alexander Dokumentov e Rob J. Hyndman).
StMoMo	Implementa a família de modelos de mortalidade de idade-período-coorte generalizados. Esta família de modelos incorpora muitos modelos propostos na literatura demográfica e atuarial, incluindo os modelos de Lee-Carter e Cairns-Blake-Dowd (2006). O programa inclui as funções para ajustar modelos de mortalidade com recursos de avaliação da qualidade do ajuste, projeções e simulações (Andres Villegas, Pietro Millossovich e Vladimir Kaishev).
SUMMER	Fornece métodos para estimar, projetar e graficar taxas de mortalidade de menores de 5 anos localizados no espaço e no tempo, da forma como foi descrita por Mercer et al. (2015) (Bryan D. Martin e Zehang R. Li).
vitality	Fornece rotinas de ajuste para quatro versões da família Vitality de modelos de mortalidade (Gregor Passolt, James J. Anderson, Ting Li, David H. Salinger e David J. Sharrow).

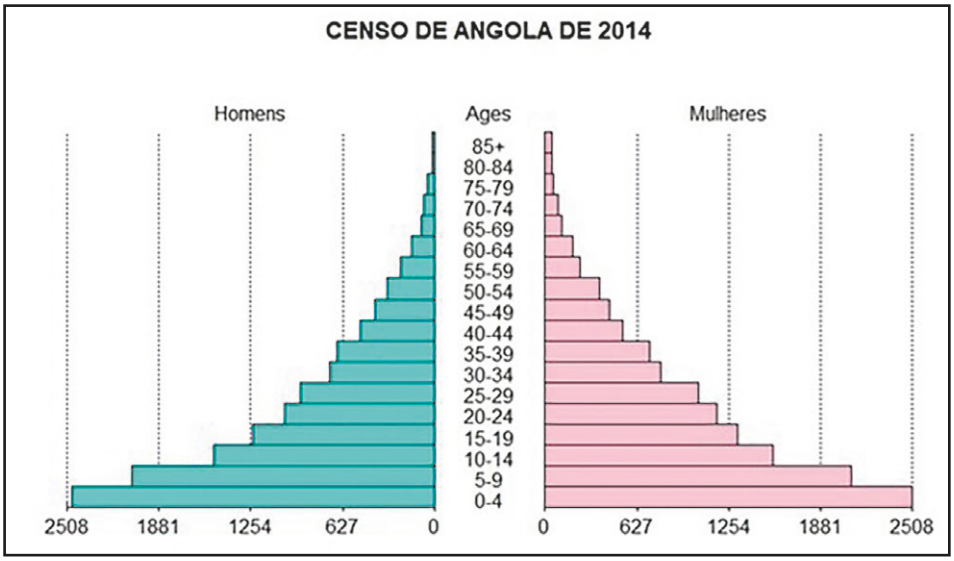
Estes pacotes podem ser carregados via internet usando o comando *install.packages*. Como exemplo, se instalou o pacote "pyramid':

```
> install.packages("pyramid")
> library(pyramid)
```
(17.22)

Estes comandos disponibilizam as instruções de "pyramids" como se fossem comandos do próprio R. Por exemplo, as seguintes instruções resultam no desenho do Gráfico 17.1, com os dados da população de Angola em 2014 que já foram apresentados em EXCEL no Gráfico 6.1:

```
> ages <- c('0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39','40-44','45-49','50-54','55-59','60-64','65-69','70-74','75-79','80-84','85+')
> males <- c(2474,2063,1507,1245,1027,914,720,663,509,412,323,234,162,89,73,49,18,18)
> females <- c(2508,2089,1557,1311,1173,1044,785,712,534,437,372,235,186,113,89,57,44,44)
> data <- data.frame(males,females,ages)
> pyramid(Llab="Homens",Rlab="Mulheres",data,main="CENSO DE ANGOLA DE 2014")
```
(17.23)

Gráfico 17.1: A mesma pirâmide do Gráfico 6.1, refeita com a ajuda do pacote "pyramid" em "R"



Fonte: Recenseamento Geral da População e Habitação: Resultados Definitivos (2016).

Embora este exemplo seja simples, muitos outros destes programas são especializados e exigem um conhecimento adicional sobre os métodos específicos sendo implementados. Uma outra coletânea, de programas, mais básicos, pode ser encontrada no sítio web de Tim Riffe (https://sites.google.com/site/timriffepersonal/). Estes programas também podem ser instalados no editor de R, mas o procedimento neste caso é um pouco diferente (veja o sítio web para detalhes). Alguns também podem ser carregados no formato da fonte em "R". Um destes programas[1] é o seguinte, que implementa o (primeiro) método de Keyfitz para a construção de tábuas de vida discutido no Capítulo 9:

```
axKeyfitz <- function(Mx,axsmooth=TRUE) {
  # Iterative ax-dx process described on pages 44-45 of Preston et al., Demography:
  Measuring and Modelling Population Processes. Blackwell Publishing, 2001
   N <- length(Mx); n <- rep(5,N); n[1] <- 1; n[2] <- 4
   if (axsmooth) {
        ages <- cumsum(n) - n; span <- ifelse(N>30,.15,.4); Mx <- log(Mx)
        Mx[2:N] <- predict(loess(Mx ~ ages, span = span,
              control=loess.control(surface="interpolate")),
              newdata=ages[2:N])
        Mx <- exp(Mx)
   }
```

[1] O programa foi ligeiramente modificado para os propósitos deste livro. O programa original tem a amplitude dos intervalos etários como argumento num vetor n. Aqui o vetor n é definido sempre como n = c(1,4,5,5,5,5,5,5,5,.....). Isso evita que o usuário entre com vetores de intervalos etários desiguais, já que a fórmula de Keyfitz só se aplica a intervalos de tamanhos iguais. Seria possível; adaptar o programa para que aceite intervalos etários de tamanhos arbitrários, mas isso exigiria modificações mais profundas.

```
  axit <- .5*n
  axit[1] <- .07+1.7*Mx[1]
  for (i in 1:7) {
        qx <- (n*Mx) / (1 + (n - axit)*Mx)
        qx[length(Mx)] <- 1; px <- 1 - qx; lx <- 1
        for (i in 2:length(Mx)) lx[i] <- lx[i-1] * px[i-1]
        dx <- diff(lx)
        for (i in 2:(length(Mx)-1)) {
          axit[i] <- (-(n[i-1]/24)*dx[i-1] + (n[i]/2)*dx[i] + (n[i+1]/24)*dx[i+1]) / dx[i]
        }
        axit[N-1] <- axit[N-2] - (axit[N-3]-axit[N-2])*1.5
        axit[N] <- axit[N-1] - (axit[N-2]-axit[N-1])*1.5
  }
  axit[1] <- .07+1.7*Mx[1]
  return(axit)
}
```

(17.24)

Esta função consiste de duas partes. Na primeira, a série Mx é suavizada (opcionalmente) usando uma regressão não paramétrica local do tipo LOESS (parecido com o LOWESS discutido na seção 16.3.7 do Capítulo 16). Os detalhes deste procedimento fogem ao alcance deste texto. A segunda parte da função constitui a aplicação propriamente dita do método iterativo descrito na Figura 8.1. O comando "*for (I in 1:7)"* significa que se usam 7 iterações, o que normalmente é mais do que suficiente. Usando Mx como antes (17.10), obtém-se os seguintes resultados suavizados e não suavizados para a sequência ${}_{n}a_{x}$:

```
> axKeyfitz(Mx,axsmooth=TRUE)

[1] 0.109270 2.020886 2.057144 2.606796 2.688693 2.576620 2.555669 2.636451
2.610847 2.583760 2.599265 2.632146 2.617196
[14] 2.602525 2.565465 2.557393 2.545286 2.527124

> axKeyfitz(Mx,axsmooth=FALSE)

[1] 0.109270 1.980610 2.307408 2.630805 2.761938 2.561711 2.483884 2.719651
2.596851 2.533065 2.684666 2.565612 2.706287
[14] 2.572742 2.541460 2.557873 2.582494 2.619424
```

(17.25)

Carl Schmertmann mantém um sítio web (http://schmert.net/BayesBrass/) chamado "Bayes + Brass" que, entre outros documentos, contém os procedimentos em "R" usados para fazer as estimativas probabilísticas de fecundidade para pequenas áreas desenvolvidas por Assunção et al. (2005) e Schmertmann et al. (2013) (ver também na seção 17.5 deste capítulo). Vale mencionar também o pacote fertestR (https://github.com/josehcms/fertestr), da autoria de Everton Lima, José Monteiro da Silva, Patrick Gerland e Helena C Castanheira, que implementa diferentes métodos indiretos para a estimação da fecundidade. Outro tipo de biblioteca de recursos disponível na web são os procedimentos para baixar e analisar bases de dados no domínio público, usando "R". Um conjunto de procedimentos deste tipo vem sendo mantido por *Anthony Joseph Damico em* https://github.com/ajdamico/asdfree/tree/master/docs. *Entre as bases de dados que podem ser acessadas por meio dos protocolos fornecidos neste sítio de web estão os Censos Demográficos e o Censo Escolar (2007) do Brasil, as PNADs, Pesquisa de Orçamentos Familiares, Pesquisa Mensal de Emprego, Pesquisa Nacional de Saúde, PISA, os DHS e o World Values Survey.*

Para usuários de EXCEL que querem continuar usando EXCEL como plataforma de análise, mas que também querem ter acesso aos recursos oferecidos por "R", uma solução pode ser o RExcel que permite justamente fazer isso. O RExcel pode ser baixado gratuitamente da internet no sítio http://sunsite.univie.ac.at/rcom/. Para maiores informações sobre o uso de RExcel, se refere aos autores do pacote (Baier, Neuwirth e De Meo, 2011; Heiberger e Neuwirth, 2009).

17.2 INTRODUÇÃO A ALGUNS OUTROS PACOTES PARA A ANÁLISE DEMOGRÁFICA

Além dos programas em "R" mencionados na seção anterior, existem muitos programas baseados em outras plataformas de programação, particularmente o EXCEL, que têm sido desenvolvidos ao longo das últimas décadas. Como as próprias plataformas de programação estão em constante evolução, alguns dos programas que foram desenvolvidos em DOS na década de 80 já estão superados, mas outros têm acompanhado a tecnologia e continuam sendo usados na atualidade. Uma discussão exaustiva de todos os programas e pacotes atualmente em circulação tomaria muito espaço, de modo que esta seção se limitaria a alguns dos mais usados dentro de cada categoria de aplicações.

Programas de Manejo de Dados

Os censos e em menor medida os inquéritos de população criam bases de dados muito grandes que precisam ser editadas e posteriormente analisadas e divulgadas para os usuários, sem ferir o princípio de confidencialidade estatística. Para a edição da informação o US Bureau of the Census tem criado um conjunto de programas que compartilha com outros institutos nacionais de estatística. Ao longo dos anos esses diferentes programas foram consolidados num programa unificado de edição de dados que se chama CSPro (Census and Survey Processing System). Muitas entidades produtoras de estatísticas hoje em dia usam o CSPro para organizar as suas bases de dados e para detectar e corrigir erros de forma automática, por meio da formulação de regras de edição. Embora o programa seja pouco aplicado fora do âmbito das instituições produtoras de estatísticas oficiais, ele é do domínio público e em princípio qualquer um pode descarregá-lo e usá-lo. No momento da publicação deste livro a versão mais recente era a versão 7.2.1.

A situação no Brasil é excepcional pela facilidade de acesso dos usuários aos microdados dos censos (amostra), PNADs e outros inquéritos que são disponibilizados pelo IBGE em formato ASCII, além de alguns formatos comerciais. Na maioria dos países o acesso é mais restrito e se usam programas de disseminação da informação que, pelo seu desenho, limitam o tipo de informação que pode ser gerada. O próprio CSPro pode ser usado para a criação de tabulações com informação censitária e de inquéritos, mas também se usam bases de dados mais especializadas. O programa REDATAM (REcuperação de DAdos para Áreas pequenas por Microcomputador), atualmente na sua versão 7, foi criado pelo CELADE no fim da década de 80, quando a capacidade de armazenamento de dados em microcomputadores era muito menor do que hoje em dia, com o objetivo de criar bases de dados altamente comprimidos e protegidos contra o vazamento de informações sigilosas para a análise de áreas pequenas. Para garantir a confidencialidade, a base de dados do REDATAM (que precisa ser preparada previamente) tem uma estrutura que impossibilita a identificação de indivíduos, mesmo quando se usa a base de dados inteira do censo. Inicialmente foi aplicado só na América Latina, mas hoje em dia também existem muitas aplicações na África (inclusive Cabo Verde e Moçambique) e Ásia onde o sistema é comumente conhecido como IMIS (Integrated Multi-sectoral Information Systems).

Os dados podem vir de qualquer combinação de censos, pesquisas ou outras fontes. É possível definir, a partir de uma base de dados, qualquer área geográfica de interesse (como blocos de uma cidade) ou combinações dessas áreas; criar novas variáveis e recodificar variáveis existentes; obter vários tipos de tabulações rapidamente e exportar saídas para outros softwares, como de mapeamento digital. A versão 7 faz uso extensivo do formato XML, que permite uma interconexão com outros softwares, bem como uma nova arquitetura de armazenamento de dados, que acelera a execução (processa um milhão de registros por segundo) e uma melhoria na apresentação de tabelas definidas pelo usuário. A versão 7 também incorpora novos comandos para tabulações (GINI e NTIL); o processamento para Análise (MultiTally), que permite obter várias estatísticas de uma variável não categorizada (por exemplo: renda, superfície de exploração) num único processo: Casos, Soma, Máximo, Mínimo, Média; e o contrário de uma destas estatísticas para diversas variáveis.

Até a versão 6, o suporte de variáveis alfanuméricas em REDATAM estava disponível apenas para a obtenção de tabulados. Agora, pode-se trabalhar com estas variáveis da mesma forma que as outras (INTEGER, REAL e BOOL), como em filtros, por exemplo. Um exemplo de sua utilização serão as causas de morte de forma direta. A nova versão também define uma gramática ou sintaxe de programação capaz detectar e visualizar eventuais erros de uso. A atualização ainda permite que existam aplicativos nos diversos idiomas falados na América Latina, como o quéchua, o creole e o guarani, além de línguas de outras regiões como árabe ou chinês, ou ainda, aplicativos feitos sob demanda para um determinado país. Muitos países usam o REDATAM como planilha de cálculo com que os usuários podem criar as suas tabelas "on-line", o que evita a necessidade de disponibilizar a própria base de dados. Entretanto, estas aplicações "on-line" geralmente não possuem a mesma flexibilidade do programa original.

O REDATAM possui algumas extensões, na forma de programas acessórios que executam tarefas específicas, especialmente o mapeamento de dados geográficos. O principal se chama ZONPLAN, que pode ser usado para mapear indicadores demográficos em nível de regiões de planejamento, províncias, municípios, áreas de enumeração etc. A Universidade de Waterloo,

no Canadá, desenvolveu três Sistemas de Informação Geográfica para serem usados conjuntamente com REDATAM, chamados AccessPlan, EduPlan e TourPlan, para as áreas de saúde, educação e turismo.

A base de dados DevInfo do UNICEF, antes conhecida como ChildInfo, e a sua versão CensusInfo para a divulgação de dados censitários muitas vezes é apresentada como uma alternativa para o REDATAM/IMIS, mas em realidade se trata de programas com finalidades distintas. Enquanto o REDATAM/IMIS é um programa de análise, que permite ao usuário criar as suas próprias tabelas e construir os seus próprios indicadores, DevInfo é um programa de divulgação de indicadores predefinidos. No DevInfo/CensusInfo os indicadores são pré-calculados de modo que o usuário não tem acesso direto aos microdados. Os recursos de exibição de informação em DevInfo são mais sofisticados do que em REDATAM/IMIS, mas o usuário não tem nenhum controle sobre a definição da informação.

O programa NACAOB (NAscimentos, CAsamentos e ÓBitos) é um software de manejo de informação com características diferentes, na medida em que foi desenvolvido especificamente para aplicações na área de demografia histórica, particularmente para apoiar a reconstituição de famílias. Outra particularidade é que se trata de uma base de dados cumulativa que é continuamente expandida pelos pesquisadores participantes. O software vem sendo desenvolvido desde a década de 90 pelo Grupo de Pesquisa CNPq – Demografia e História.

A ideia inicial surgiu a partir de um projeto sobre os comportamentos demográficos e familiares de uma comunidade no noroeste de Portugal. Assim a arquitetura lógica da versão original e sua estrutura de banco de dados teve como referência os registros paroquiais portugueses. Portanto não incorporava, por exemplo, a população escrava, segmento significativo da população brasileira, não apenas por conta de aspectos demográficos, como também a questões ligadas às hierarquias sociais. Esses elementos foram acrescentados posteriormente. Nos anos 2000, o "Sistema de Reconstituição de Paróquias" (SRP), houve uma mudança na metodologia em que as fichas de família eram inseridas num banco de dados. As versões sucessivas do NACAOB também têm procurado incorporar as mudanças tecnológicas. Atualmente a arquitetura lógica está orientada para o trabalho em redes colaborativas, com uma versão visual e multiusuária que foi introduzida em 2009. A lógica do trabalho isolado de cada pesquisador está sendo, gradativamente, substituída pela alimentação simultânea das bases de dados, por distintos pesquisadores, embora as bases permaneçam organizadas por paróquia/freguesia.

A estrutura do banco de dados é relativamente simples, com quatro tabelas principais que descrevem os eventos (batizados, casamentos e óbitos) e os indivíduos (que permite a inserção de todos os indivíduos arrolados em cada evento). Além dessas, há tabelas de apoio, para ajudar os pesquisadores a incluir dados relativos a cada evento, de forma padronizada, por exemplo, naturalidade, ocupação, título ou patente, cor ou causa de morte. As tabelas de apoio são permanentemente atualizadas com dados novos. Entre os recursos da versão atual está o campo "nome" que permite que o pesquisador efetue o registro na grafia original e a sua equivalente, padronizada (por exemplo Joseph, Joze, José que acaba sendo padronizado para José), o qual facilita o posterior cruzamento com outros dados. O programa também dispõe de tabelas auxiliares fixas, que padronizam informações sobre o papel que cada indivíduo desempenha no evento. Os dados inseridos por cada pesquisador são disponibilizados através de extrações periódicas ou por demanda, em formato de planilha EXCEL. Ainda não há a possibilidade de efetuar extrações diretas do banco

de dados, que está hospedado num provedor comercial e, por isso requer que o administrador disponibilize as extrações de dados atualizadas para os respectivos pesquisadores. As atualizações são feitas periodicamente, ou sob a demanda dos interessados.

Scott (2018: Tabela 3) apresenta o estágio do banco de dados NACAOB em março de 2018, quando havia 24 freguesias cadastradas, que conformavam um conjunto de mais de 98.000 assentos de batizado, 12.000 assentos de casamento e 54.000 registros de óbito. Neste universo, a base contava com mais de 770.000 registros de indivíduos.

É preciso mencionar que existem outros programas para o manejo de bancos de dados em demografia histórica. Talvez o mais conhecido seja o IDS (de "Intermediate Data Structure"), que foi desenvolvido para facilitar o manejo de dados longitudinais do tipo usado na análise de biografias individuais. Na data da publicação deste livro o programa estava na sua Versão 4.0. Para maiores detalhes sobre o programa e o tipo de análises que permite, ver Alter e Mandemakers (2014).

Construção de Tábuas de Vida

Como se viu no Capítulo 9, o algoritmo básico de construção de tábuas de vida é relativamente simples e pode ser programado em EXCEL sem maiores dificuldades, embora também exista o modulo LIFTB de MORTPAK para este fim. O que justifica o uso de software especializado para esse fim são os recursos acessórios que podem apoiar a construção de tábuas de vida, tais como algoritmos mais sofisticados para a conversão de ${}_nM_x$ em ${}_nq_x$, mecanismos de graduação e projeção (incluindo procedimentos do tipo Lee-Carter a ser discutido na seção 21.4.1), inclusão de recursos estatísticos para quantificar a variabilidade das estimativas ou o uso de modelos para fechar as tábuas a partir de uma determinada idade. Outro motivo pode ser a extensão do conceito de tábuas de vida para modalidades mais complexas como as tábuas de múltiplo decremento ou múltiplos estados que serão discutidas no Capítulo 19. Vários dos pacotes que foram mencionados na seção anterior ou que serão mencionados nesta seção contêm algum recurso para apoiar a construção de tábuas de vida. Entretanto, existem alguns pacotes que foram desenhados especificamente para esta finalidade. Um é o programa SURVIVAL (atualmente na sua versão 6.0), do Population Studies Center da Universidade de Michigan (http://www.psc.isr.umich.edu/dis/data/demosoft.html), que contém vários recursos estatísticos avançados e ajuda a projetar a mortalidade para o futuro. Um programa semelhante, chamado DeRaS, foi desenvolvido pela Universidade Carolina de Praga e está disponível em http://deras.natur.cuni.cz/en/. Especificamente para o cálculo de tábuas de vida de múltiplos estados, o Population Studies Center da Universidade de Michigan desenvolveu um software separado, disponível no mesmo endereço mencionado acima.

Programas de Projeção

Da mesma forma como acontece na construção de tábuas de vida, o algoritmo básico das projeções demográficas baseadas na metodologia de coortes-componentes é relativamente fácil de construir em EXCEL (ver Capítulo 21), embora as planilhas possam ser grandes e a preparação dos dados possa ser trabalhosa. Algumas projeções mais especializadas, como a metodologia de

Relação de Coortes (ver seção 21.7.4), têm um grau de complexidade que eventualmente justifica o uso de um programa específico, mesmo que a construção de uma planilha em EXCEL para implementá-las não esteja além das possibilidades do usuário médio.

As vantagens principais do uso de programas pré-montados são as seguintes:

1. Um processo muito trabalhoso que precisa ser executado antes da projeção propriamente dita é a conciliação censitária (ver seção 16.5 do Capítulo 16), para determinar a população de base. A maioria dos programas de projeção não contém recursos para facilitar esse procedimento, mas caso houvesse seria uma vantagem importante.

2. Muitos destes programas já vêm com certos modelos internos de mortalidade, fecundidade e eventualmente migração (ver Capítulo 20), o que poupa muito trabalho na hora de especificar um conjunto de tábuas de vida ou esquemas de fecundidade: em vez de ter que especificar cada tábua ou cada modelo, basta especificar os parâmetros básicos dos modelos.

3. Alguns programas contêm recursos para ajudar na projeção das tendências da mortalidade, fecundidade e migrações, seja pelo uso de modelos matemáticos simples ou pela projeção dos determinantes.

4. Alguns dos programas mais novos contêm providências para modelar a influência do AIDS (SIDA), o que pode ser importante no contexto africano.

5. Se a projeção for realizada não só em nível nacional, mas também por unidades nacionais, as eventuais facilidades que o programa oferece para fazer isso de forma consistente podem ser uma consideração importante para a sua adoção.

6. Alguns programas de projeção da população contêm recursos para derivar projeções funcionais, da força de trabalho, número de domicílios (agregados familiares), população escolar, incidência de deficiências ou doenças etc.

Por outro lado, o grande perigo do uso de pacotes em geral é que podem gerar resultados absurdos ou não justificados quando alimentados com dados que não satisfazem as condições necessárias para a aplicação dos métodos. Este perigo está menos presente nos programas de projeção demográfica do que nos programas de análise que serão discutidos abaixo, mas é importante sempre estar vigilante e verificar a plausibilidade dos resultados.

Nos anos 80 do século passado foram desenvolvidos muitos programas de projeção demográfica, mas a maioria (PROJ3S, PDPM/PC, PEOPLE, PRODEM, FIVFIV, o programa do East-West Center para projeções por ordem de nascimento) caiu em desuso porque não evoluíram em termos de sistemas operacionais (migração de DOS para Windows) ou porque não desenvolveram os recursos acessórios mencionados acima para aumentar a sua utilidade. Os programas descritos abaixo ainda continuam em uso.

No contexto africano, o recurso mais usado é o DEMPROJ, que faz parte do pacote SPECTRUM, distribuído gratuitamente pela USAID. Por exemplo, várias projeções da população

da Guiné-Bissau foram realizadas usando esse programa (Antunes, 2010; INE, 2013). O programa faz projeções de coortes-componentes para um máximo de 50 anos de um país ou uma região, eventualmente divididos por área urbana e rural. O acessório EASYPROJ fornece os dados necessários para preparar projeções a partir das estimativas de parâmetros produzidas pela Divisão de População das Nações Unidas. Além do DEMPROJ, SPECTRUM também contém um programa chamado AIM, para fazer projeções da incidência de AIDS (SIDA) e avaliar os seus impactos sociais e econômicos. O módulo PROJCT de MORTPAK também faz projeções, mas é menos usado do que DEMPROJ.

O programa RUP (Rural-Urban Projections), do US Bureau of the Census, que originalmente foi escrito para apenas duas áreas geográficas (rural e urbana), está disponível para ilimitado número de unidades usando a extensão RUPAGG. O programa contém providências para garantir que todas essas projeções subnacionais sejam consistentes. Outra extensão, chamada RUPCombine, permite a localização exata de choques demográficos como resultado de desastres naturais ou conflitos armados. Para fazer projeções subnacionais, o US Bureau of the Census também disponibiliza o SPToolkit, um programa que funciona conjuntamente com RUP e com PAS (ver abaixo). O RUP é o componente central de um pacote mais amplo, que reúne todos os diferentes elementos da análise e que se chama Demographic Analysis and Population Projection System (DAPPS). Ainda é preciso mencionar dois outros programas. O POPGROUP, da Universidade de Manchester na Inglaterra, é baseado em EXCEL, produz projeções de população, domicílios (agregados familiares) e força de trabalho, e tem a particularidade de permitir a projeção de grupos sociais, além da projeção de áreas geográficas.

Tanto DemProj como RUP usam a metodologia convencional de coortes-componentes. Mas como será discutido no Capítulo 21, esta metodologia pode ser refinada com o uso de álgebra matricial, o qual dá origem à metodologia multirregional ou de múltiplos estados, onde cada indivíduo é caracterizado não só pela sua idade e sexo, mas também por um "estado" como o lugar de residência, escolaridade ou estado civil. Como os cálculos neste caso são mais complexos, o uso de software especializado faz mais sentido. Além de alguns programas mais antigos que atualmente são pouco usados, três programas que continuam mais ou menos vigentes são PDE, do IIASA na Áustria, POPSTAR2, da Universidade de Queensland na Austrália (Wilson e Cooper, 2013) e MULTIPOLES, desenvolvido por Kupiszewski e Kupiszewska (2011), na Polônia. A limitação de todos esses programas é que eles foram desenvolvidos para aplicações mais ou menos específicas, o que pode dificultar a sua aplicação em outros contextos. Por exemplo, MULTIPOLES foi desenvolvido para o contexto da EU, com um nível supranacional, um nível nacional e um nível subnacional. POPSTAR2 foi desenvolvido para o contexto geográfico do Estado de Queensland, que foi o cliente que encomendou o programa. O programa faz parte de um conjunto de módulos de projeção que foram produzidos nesse projeto (POPULATES, HOUSEPRO, POPACTS, POPSAS, PROBREG, POPCORN, PROBPOP e POPART). O mesmo autor também produziu um programa de projeção chamado RePPP (Regional Population Projection Program) em EXCEL baseado numa metodologia multi biregional que tem a vantagem de ser mais econômica em termos das suas demandas de dados. Finalmente deve ser mencionado o pacote MSDem em "R", que faz parte da lista de programas em "R" da seção anterior. Para uma discussão mais completa do software disponível nesta área, ver Willekens e Putter (2014).

Programas de Análise

Os dois pacotes principais com versões atualizadas que continuam sendo usadas para fins de análise demográfica, principalmente para a aplicação de métodos indiretos (ver Capítulo 23), são MORTPAK, desenvolvido pela Divisão da População das Nações Unidas, e PASEX, do US Bureau of the Census. Os pacotes PANDEM e PREVIO, do CELADE, e os programas de Correção Consistente e de Análise Fecundidade, que foram desenvolvidos nos anos 80 pelo East-West Center, são mais antigos e só existem em versões para DOS. O programa PRODEMOG (versão 3.0), desenvolvido em Visual BASIC, por Luciano Petrioli da Universidade de Siena e que contém uma grande variedade de modelos demográficos, foi publicado pela última vez em 2000, mas é pouco usado. Tanto MORTPAK como PASEX são versões atualizadas para Windows de programas que originalmente foram desenvolvidos em DOS. Ambos estão organizados como coleções de procedimentos que não interagem entre eles. Os programas fornecem um formato comum para a entrada de dados e a organização dos resultados, mas no demais os procedimentos são independentes.

Como o nome sugere, o pacote MORTPAK foi desenvolvido originalmente para a análise da mortalidade, mas ao longo do tempo foram incorporados alguns procedimentos na área da fecundidade e projeção. A versão 4.3 de MORTPAK que periodicamente é atualizada está disponível em: https://www.un.org/en/development/desa/population/publications/mortality/mortpak.asp. Contém os seguintes procedimentos:

BENHR	Implementa o método de Bennett-Horiuchi para estimar a cobertura do sistema de registro de mortalidade (ver seção 23.4.2 do Capítulo 23).
BESTFT	Identifica a tábua de vida modelo do conjunto de Princeton ou das Nações Unidas (ver seção 20.2.3 do Capítulo 20) que melhor se ajusta a uma série observada ${}_nm_x$ ou ${}_nq_x$.
COMPAR	Parecido com o anterior, mas em vez de escolher o ajuste melhor, o COMPAR imprime todos os índices de semelhança com todas as tábuas de vida modelo do sistema. Além de ${}_nm_x$ e ${}_nq_x$, também se pode usar ℓ_x.
CEBCS / QFIVE	Estimam a mortalidade infantil e de crianças com diferentes métodos indiretos do tipo que serão explicados na seção 23.3 do Capítulo 23. O CEBCS se baseia no tempo desde a primeira união, enquanto o QFIVE classifica as mulheres pela idade. A primeira variante não é discutida em detalhe neste livro. A segunda variante pode ser aplicada com as tábuas de vida modelo de Princeton ou das Nações Unidas (versão de Palloni-Heligman). Para maiores detalhes sobre as tábuas de vida modelo, ver seção 20.2.3 do Capítulo 20.
CENCT	Aplica a metodologia descrita na seção 23.4.4 do Capítulo 23, baseada na equação generalizada de balance, para estimar a cobertura de um censo em relação a outro, usando os modelos de mortalidade de Princeton ou de Nações Unidas.

COMBIN	Combina uma estimativa da mortalidade infantojuvenil com uma estimativa da mortalidade adulta (e_{20} ou ${}_{n}q_{15}$) para construir uma tábua de vida inteira.
CORMOR	Mostra as correspondências entre diferentes tábuas de vida modelo. O usuário escolhe um parâmetro da tábua de vida e o sexo e recebe os valores correspondentes de ${}_{n}m_{x}$, ${}_{n}q_{x}$, ℓ_{x} ou e_{x} para todas as idades e todos os modelos de Princeton e Nações Unidas.
FERTCB	Estima TEFs com base no número total de filhos nascidos vivos classificados pela idade das mães em um ou dois momentos do tempo.
FERTPF	Aplica o método P/F explicado na seção 23.2.2 do Capítulo 23 em um ou dois momentos do tempo.
ICM	Estima ${}_{1}q_{1}$, ${}_{1}q_{2}$, ${}_{1}q_{3}$ e ${}_{1}q_{4}$ a partir de ${}_{1}q_{0}$, ${}_{4}q_{1}$ e ${}_{5}q_{5}$.
LIFTB	Constrói uma tábua de vida a partir de uma série ${}_{n}m_{x}$, ${}_{n}q_{x}$ ou ℓ_{x}.
MATCH	Encontra uma tábua de vida que corresponde a um determinado nível de mortalidade, especificado em termos de ${}_{n}m_{x}$, ${}_{n}q_{x}$, ℓ_{x} ou e_{x} e um modelo de mortalidade que pode ser o de Princeton, Nações Unidas ou um padrão fornecido pelo próprio usuário.
ORPHAN	Estima a mortalidade adulta a partir de informação acerca da sobrevivência da mãe ou do pai do respondente. A lógica do método é parecida com a estimação da mortalidade na infância (seção 23.3 do Capítulo 23), mas os detalhes não são abordados neste livro.
PRESTO	Estima a mortalidade e fecundidade e ajusta a distribuição etária da população com base nas distribuições etárias observadas em dois censos sucessivos e o suposto de um modelo de mortalidade de Princeton, Nações Unidas ou fornecido pelo usuário. O método de Preston subjacente a este procedimento não é discutido neste livro.
PROJCT	Faz projeções anuais de uma população inicial especificada por sexo e grupos quinquenais de idade e hipóteses sobre a evolução futura da mortalidade, fecundidade e migração.
STABLE	Calcula uma distribuição etária estável (ver Capítulo 22) com base numa série de valores ${}_{n}m_{x}$ ou ${}_{n}q_{x}$ e uma taxa intrínseca de crescimento.
TIMESER	Parecido com MATCH, mas organizado no formato de uma série temporal.
UNABR	Gradua uma série quinquenal de valores ${}_{n}q_{x}$ para obter uma série por idades simples.

WIDOW Estima a mortalidade adulta a partir de informação acerca da sobrevivência do(a) primeiro(a) esposo(a). A lógica do método é parecida com a estimação da mortalidade na infância (seção 23.3 do Capítulo 23), mas os detalhes não são abordados neste livro.

O pacote PASEX, do US Census Bureau, é uma coleção de planilhas de cálculo em EXCEL com os seguintes procedimentos:

AGEINT Interpola linear ou exponencialmente entre duas distribuições etárias.

AGESEX Calcula vários dos índices introduzidos no Capítulo 16 para medir a qualidade da declaração de idade e sexo.

AGESMTH Gradua a distribuição etária da população usando diferentes métodos, alguns dos quais foram introduzidos no Capítulo 16.

BASEPOP e BPSTRNG Prepara a distribuição por sexo e idade de uma população para uma projeção, aplicando um processo de suavização.

GRPOP-YB Faz um gráfico da estrutura etária em duas ou três datas, por ano de nascimento de cada coorte.

MOVEPOP Desloca a distribuição etária da população para uma data diferente.

OPAG Abre o intervalo final de uma distribuição etária para que o último grupo seja 80+ anos.

PYRAMID Faz uma pirâmide etária por sexo, com números absolutos e relativos, seguindo os procedimentos explicados na seção 6.1 do Capítulo 6.

SINGAGE Calcula os índices de Whipple, Myers e Bachi com base numa distribuição etária por idades simples, como foi descrito no Capítulo 16.

ADJASFR Ajusta um padrão de TEFs para reproduzir um determinado número de nascimentos.

ARFE-2 e ARFE-3 Usam uma técnica desenvolvida por Arriaga discutida na seção 23.3.2.5 do Capítulo 23 para estimar taxas de fecundidade com base no número médio de filhos tidos das mulheres e um padrão de fecundidade por idade.

ASFRPATT Fornece as TEFs típicas referentes a uma determinada TFT.

CBR-GFR Calcula a TBN e a TFG com base na TFT.

CBR-TFR Estima a TBN e TFT com base na TFG.

PFRATIO Aplica o método P/F de Brass para estimar a fecundidade (ver seção 23.2.2 do Capítulo 23).

RELEFERT Aplica o método de Rele (não discutido neste livro) para estimar a TBR para um ou dois períodos quinquenais anteriores ao censo.

REL-GMPZ Faz a análise de Gompertz relacional que é explicada na seção 23.2.2 do Capítulo 23.

REVCBR Calcula taxas de natalidade em dois ou três períodos quinquenais antes do censo, com base na estrutura etária do censo.

TFR-GFR Estima a TFT e a TFG com base na TBN.

TFRLGST e E0LGST Ajustam uma função logística a dois ou mais valores da TFT ou e_0, respectivamente, e dados os valores assintóticos, faz interpolações e extrapolações.

TFRSINE Faz o mesmo com a função seno.

CSRMIG Estima a migração intercensitária líquida entre duas áreas.

ADJMX Ajusta um padrão de TEMs para reproduzir um determinado número de óbitos.

BTHSRV Estima taxas de mortalidade infantil com base no número de crianças nascidas durante o ano anterior ao censo e o número ainda vivo no momento do censo.

GRBAL Aplica o método de Balanço de Crescimento de Brass para a estimação da mortalidade de maiores de 5 anos, conforme a discussão na seção 23.3 do Capítulo 23.

INTPLTM e INTPLTF Interpolam tábuas de vida masculinas e femininas, respectivamente, entre duas tábuas de vida dadas.

LOGITQX e LOGITLX Suavizam as funções de uma tábua de vida usando os logitos de q_x e ℓ_x (ver seção 20.2.2.2 do Capítulo 20).

LTMXQXAD Constrói uma tábua de vida a partir de TEMs ou das probabilidades de morte entre duas idades específicas.

LTNTH, LTSTH e LTWST Selecionam a tábua de vida modelo de Princeton Norte, Sul ou Oeste que reproduzirá uma dada TBM para uma estrutura etária dada.

LTPOPDTH Constrói e suaviza uma tábua de vida para um ou ambos os sexos, com base em dados de população e óbitos.

PREBEN	Estima o nível de mortalidade para idades maiores de 5 anos durante o período intercensitário pelo método de Preston e Bennett (1983), que não é discutido neste livro.
PRECOA	Implementa a técnica de Preston et al. (1980) para estimar o sub-registro de óbitos, conforme a discussão no Capítulo 23.
URBINDEX	Calcula vários índices de urbanização e distribuição da população.
CTBL32	Aplica o método do ajuste biproporcional iterativo explicado na seção 21.7.3 do Capítulo 21.
FITLGSTC	Ajusta uma função logística a 3 (ou um múltiplo de 3) valores observados equidistantes de qualquer índice, sem requerer valores assintóticos.
LOGISTIC	Ajusta uma função logística a 2 ou mais valores observados de qualquer índice, dadas as assíntotas inferior e superior.
SP	Constrói uma população estável com base em tábuas de vida por sexo e taxas específicas de fecundidade por idade ou então uma taxa intrínseca de crescimento.

Finalmente, devem ser mencionadas as planilhas em EXCEL que foram desenvolvidas por Moultrie et al. (2013), como parte do projeto que ensina a aplicação de procedimentos atualizados de metodologia indireta de estimação e que estão disponíveis no sítio web da IUSSP (http://demographicestimation.iussp.org/). Embora não se trate de um "pacote" no sentido formal, a coleção de planilhas fornece um apoio valioso aos usuários que queiram aplicar os métodos desenvolvidos ou reproduzidos na guia.

Programas de Microssimulação

O Capítulo 13 se referiu brevemente aos modelos de microssimulação e suas implementações em software, para fins de estudo da influência de processos demográficos sobre a estrutura familiar. Por meio deles é possível obter projeções dos processos demográficos em nível agregado pela soma dos comportamentos individuais, em vez de projetar as tendências em nível macro, como nas metodologias de projeção demográfica mais convencionais (ver Capítulo 21). Por esses métodos define-se um conjunto de funções demográficas e submete-se cada indivíduo da população a essas funções individualmente por meio de sorteios aleatórios que definem se e quando cada indivíduo experimenta cada evento demográfico. Isso é vantajoso em situações onde existem tantas combinações possíveis para caracterizar os comportamentos individuais que a projeção de todas as categorias resultantes seria muito complexa. Por exemplo, a fecundidade das mulheres depende da sua idade, estado civil, nível de educação, atividade econômica, história migratória e outros fatores mais. Levar todos esses fatores em conta numa projeção convencional é pouco factível, mas a microssimulação pode oferecer uma alternativa mais viável. Outra vantagem destes métodos é

que eles fornecem um critério intrínseco da variabilidade dos resultados. Para mais detalhes, ver Bijak et al. (2018).

O programa mais antigo de microssimulação é o SOCSIM (de "Social Simulation") que foi desenvolvido por Hammel, Wachter e Laslett (1978) no contexto dos estudos históricos sobre os domicílios na Inglaterra pré-industrial, aos quais já se fez referência nos Capítulos 13 e 15. Mais em particular, foi feito para estudar os efeitos da ordem de nascimento dos filhos na estrutura dos domicílios. O programa CAMSIM (de "Cambridge Simulation") foi desenvolvido por Smith (1987) e tem muitas semelhanças com SOCSIM, mas também algumas diferenças fundamentais na sua arquitetura (ver Zhao, 2006, para uma comparação sistemática).

O programa LIPRO (de "LIfestyle PROjections"), que já foi mencionado no Capítulo 13, usa a metodologia de múltiplos estados em que as unidades projetadas se referem a domicílios (agregados familiares) e os "estados" do modelo se referem às suas estruturas. A metodologia subjacente é descrita num livro de autoria de Van Imhoff e Keilman (1991) que pode ser baixado gratuitamente do sítio web https://www.nidi.knaw.nl/en/research/al/270101. A versão original de 1988 foi feita em DOS, mas a versão 4.0, de 1999, foi implementada em Windows, com uma interface em EXCEL. O programa tem sido amplamente usado na Europa, mas como exige dados bastante detalhados sobre a dinâmica dos domicílios não existem aplicações para América Latina ou África. Uma alternativa para o LIPRO é o programa PROFAMY, que é mais fácil de aplicar na medida em que exige menos dados. Contrariamente a LIPRO, PROFAMY se baseia em informação demográfica padrão e não exige muitos dados sobre as transformações que podem acontecer dentro dos domicílios. A versão mais recente do programa no momento da publicação deste livro é a versão 2.1.

Outros programas do mesmo tipo incluem o DYNAMOD, da Universidade de Canberra, o MODGEN, de Statistics Canada, e o DEMOFAM, também do Canadá. O MOSART-H, do Instituto Nacional de Estatística da Noruega, segue uma lógica parecida com o LIPRO. No Brasil há o programa SADEPREV (Simulador Atuarial-Demográfico de Regimes Próprios de Previdência Social) que foi desenhado para a projeção de beneficiários titulares e pensionistas em planos previdenciários de funcionários públicos brasileiros por meio de uma metodologia de microssimulação (Corrêa, 2014).

Outros

Acima já se mencionou o pacote SPECTRUM, desenvolvido pelo Futures Group e Research Triangle Institute e distribuído pela USAID. Este software é a consolidação de vários programas anteriores que, além do DEMPROJ e AIM, que já foram mencionados acima, incluem o FAMPLAN, BenefitCosts e RAPID. O objetivo principal destes programas é o apoio ao planejamento familiar. O programa RAPID é um recurso de publicidade para visualizar os impactos de diferentes cenários de crescimento demográfico. Mais recentemente, se desenvolveram pacotes para modelar o dividendo demográfico (ver Capítulo 14). O programa DemDiv (Moreland et al., 2014) tem sido amplamente usado para este propósito nos países africanos.

Finalmente, é útil saber que o Departamento de Demografia da Universidade de Berkeley mantém uma página web (https://applieddemogtoolbox.github.io/Toolbox/) que reúne um grande

número de ferramentas de programação, incluindo programas em "R", outros tipos de programas e bases de dados do domínio público, que estão à disposição dos pesquisadores da área.

17.3 CÁLCULO DIFERENCIAL E INTEGRAL

Evidentemente este não é um lugar adequado para entrar em detalhes sobre o cálculo diferencial e integral, um assunto que tipicamente exige vários semestres de cursos de matemática para dominá-lo completamente. Entretanto, os conceitos que precisam ser compreendidos para acompanhar a aplicação do cálculo diferencial e integral na demografia são muito mais limitados do que aqueles que se aplicam na física ou na engenharia. As definições da *derivada* (valor diferenciado) e *integral* de uma função podem ser entendidas intuitivamente e com relativa facilidade a partir de um Gráfico como 17.2. A curva representa uma função matemática que aqui será chamada *f(x)*. No ponto *x=2,5*, se desenhou a reta tangente da função *f(x)* que naquele ponto assume o mesmo valor da função *f(x)*, com um coeficiente de inclinação que depende da rapidez com que *f(x)* aumenta naquele ponto. Sem entrar nos detalhes de como se determina este coeficiente de inclinação (assunto coberto em qualquer curso formal de cálculo diferencial e integral), basta dizer aqui que o coeficiente de inclinação da reta tangente é chamada a *derivada* da função *f(x)* no ponto *x=2,5.* O processo de determinar as derivadas de *f(x)* em todos os pontos *x* se chama a *diferenciação da função f(x)* e o seu resultado é uma função *derivada* que se nota como *f'(x)* ou *df/dx*. No caso do Gráfico 17.2, se pode ver que *f'(x)* é positiva até aproximadamente *x=2,96* e depois se torna negativa.

Dependendo da complexidade da função *f(x)*, a determinação de *f'(x)* ou *df/dx* pode ser mais ou menos difícil, mas no caso das funções mais comuns as regras são simples:

$$\frac{d}{dx}\, A\, x^n + B\, x^{n-1} + C\, x^{n-2} + \ldots + F\, x + G = n\, A\, x^{n-1} + (n-1)\, B\, x^{n-2} + \ldots + F$$

$$\frac{d}{dx}\, e^{A\, x} = A\, e^{A\, x}$$

$$\frac{d}{dx} \ln(A\, x) = \frac{1}{x}$$

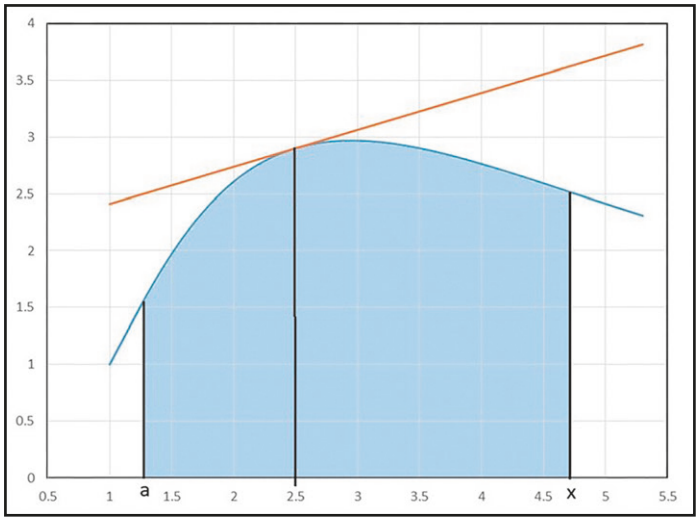
$$\frac{d}{dx}\, f(x) g(x) = f(x)\, g'(x) + f'(x)\, g(x) \qquad \text{(17.26.a-d)}$$

onde *e = 2,71828.....* e *ln* se refere ao logaritmo natural com esta mesma base[2]. Finalmente é importante conhecer a chamada *regra da cadeia* que determina que

[2] A razão de ser do número *e = 2,71828....* é justamente que o uso deste número em funções exponenciais e como base do logaritmo leva aos resultados muito simples mostrados em (17.26.a) e (17.26.b). Também é possível diferenciar funções exponenciais ou logaritmos com outras bases, mas o resultado envolve um fator *ln(b)* onde *b* é o número de base escolhido.

$$\frac{d}{dx} f\big(g(x)\big) = f'\big(g(x)\big)\, g'(x) \tag{17.27}$$

Gráfico 17.2: Representação esquemática dos conceitos de derivada (inclinação) no ponto 2,5 e integral (superfície embaixo da curva)



Fonte: Elaboração própria.

O outro elemento destacado no Gráfico 17.2 é a superfície entre a curva e o eixo horizontal de *a* até *x*. Esta quantidade é notada da seguinte forma que é conhecida como a *integral* de *f(t)* entre *a* e *x*:

$$Superfície\ contida\ entre\ a\ e\ x\ e\ entre\ a\ curva\ e\ o\ eixo\ horizontal = \int_a^x f(t)\, dt \tag{17.28}$$

Pode ser demonstrado que integração e diferenciação são processos opostos, ou seja, se *f(x)* for uma função mais ou menos suave, sem descontinuidades ("saltos"), o seguinte pode ser verificado:

$$\frac{d}{dx} \int_a^x f(t)\, dt = f(x) \tag{17.29}$$

ou seja, integrando a função *f(x)* e depois diferenciando o resultado, volta-se para a mesma função *f(x)*.

Com esta introdução superbreve ao cálculo diferencial e integral (que provavelmente precisa ser complementada com outras leituras), agora dá para aplicar os conceitos expostos à análise demográfica. Para poder articular as relações entre as diferentes funções usadas na demografia, seria desejável que elas fossem definidas como funções contínuas, cujo valor é conhecido em qualquer ponto. Mas na prática isso é difícil. Para conhecer a Taxa Específica de Fecundidade de mulheres com uma idade exata de 24 anos e 140 dias ou entre 24 anos e 140 dias e 24 anos e 141 dias e todos os outros pontos da curva de fecundidade seria necessário compilar uma quantidade

proibitiva de informação. Além disso os resultados seriam baseados em pouquíssimos casos, de modo que a variação aleatória seria enorme. Por isso, os dados normalmente são agregados em intervalos mais manejáveis de 1 ou 5 anos. Mas isso traz o inconveniente de que a intensidade dos fenômenos não é exatamente a mesma ao longo do intervalo. Buscam-se, então, funções contínuas definidas a partir de dados discretos. Antes de entrar em mais detalhe sobre as maneiras como isso pode ser feito, vale a pena explicitar como algumas das relações investigadas nos capítulos anteriores podem ser reformuladas no formato de funções contínuas. Como exemplo concreto, se usarão as funções da tábua de vida introduzidas no Capítulo 9.

O equivalente contínuo da função de sobrevivência ℓ_x é a função contínua $\ell(x)$ que é definida para qualquer valor não negativo de x e não, como no caso de ℓ_x, só para $x=0, 1, 2, 3, 4, \ldots$ ou $x=0, 1, 5, 10, 15, \ldots$ Enquanto a raiz ℓ_0 da variante discreta normalmente é definida como sendo igual a 100.000 ou 10.000, o valor de $\ell(0)$ geralmente é considerado como igual a 1. A derivada negativa de $\ell(x)$ indica a rapidez com que o número de sobreviventes diminui. Esta quantidade - $\ell'(x)$ também é chamada a *função instantânea de óbitos*, já que para valores pequenos Δx a quantidade - $\ell'(x)\,\Delta x$ é o equivalente contínuo da função discreta ${}_{\Delta x}d_x$ da tábua de vida.

No Capítulo 9 foi introduzida uma outra função contínua, a saber, a *força da mortalidade ou taxa instantânea de mortalidade* na idade x, $\mu(x)$. Observou-se que a forma mais fácil de entender esta quantidade é como a TEM na idade x para um intervalo de idade muito pequeno, tão pequeno que a distribuição da população ao longo do intervalo (que influi em ${}_nM_x$) se torna irrelevante. Sendo uma TEM, $\mu(x)$ pode ser entendida como a razão entre a função instantânea de óbitos e a versão instantânea de ${}_nL_x$, que é $\ell(x)$, de modo que $\mu(x)$ pode ser escrita como

$$\mu(x) = -\frac{\ell'(x)}{\ell(x)} = -\frac{d}{dx}\ln(\ell(x)) \tag{17.30}$$

onde foram usadas as relações (17.26.c) e (17.27). Invertendo a fórmula, isso significa que

$$\ell(x) = \ell(0)\exp\left(-\int_0^x \mu(t)\,dt\right) \tag{17.31}$$

onde $\ell(0)$, como se observou acima, é uma constante de escala geralmente definida como sendo igual a 1.

Com base em (17.31) é fácil de ver que

$${}_nd_x = \ell_x - \ell_{x+n} = \int_x^{x+n} \ell(t)\;\mu(t)dt \tag{17.32}$$

A partir de (17.31) também dá para ver que, para um intervalo muito curto $(x,x+\Delta x)$

$${}_{\Delta x}q_x = 1 - \frac{\ell(x+\Delta x)}{\ell(x)} = 1 - \exp\left(-\int_x^{x+\Delta x}\mu(t)\,dt\right) \approx \int_x^{x+\Delta x}\mu(t)\,dt \approx \Delta x\,\mu(x) \tag{17.33}$$

o qual significa que μ*(x)* é a versão contínua de ${}_nq_x$, tanto como é de ${}_nM_x$, ou seja, no limite, para intervalos muito curtos, ${}_nq_x$ e ${}_nm_x$ ambas são iguais a μ*(x)*.

No caso em que μ*(x)* é constante no intervalo *(x,x+n)*, também se pode derivar a partir de (17.30) que

$$ {}_{\mathrm{n}}q_x = 1 - \frac{\ell(x+\mathrm{n})}{\ell(x)} = 1 - \exp\left(-\int_x^{x+\mathrm{n}} \mu(t)\,dt\right) = 1 - \exp\left(-n\ {}_nM_x\right) \qquad (17.34) $$

o qual é a fórmula para ${}_nM_x$ constante (9.22) que já foi derivada por outros meios no Capítulo 9.

A função ${}_nL_x$, como já foi sugerido no Gráfico 9.5, se obtém pela integração de *ℓ(x)*, ou seja

$$ {}_nL_x = \int_x^{x+n} \ell(\mathrm{t})\,dt = \int_x^{x+n} \exp\left(-\int_0^t \mu(s)\,ds\right) dt \qquad (17.35) $$

Substituindo *x+n* por ∞, a fórmula (17.35) dá como resultado a função T_x.

A esperança de vida contínua é dada por:

$$ e_x = \int_x^{\infty} \ell(\mathrm{t})\,dt \,/\, \ell_x = \int_x^{\infty} \exp\left(-\int_0^t \mu(s)\,ds\right) dt / \exp\left(-\int_0^x \mu(t)\ dt\right) \qquad (17.36) $$

Numa formulação contínua evidentemente não há necessidade para os fatores de separação ${}_na_x$, mas as mesmas podem ser calculadas da seguinte forma:

$$ {}_na_x = -\int_x^{x+n} t\,\ell'(t)\,dt \,/\, (\ell_x - \ell_{x+n}) \qquad (17.37) $$

17.3.1 Interação com o exemplo da fórmula de De Moivre

Para ilustrar melhor o que as relações acima significam na prática, pode-se usar a relação estipulada por De Moivre, que foi introduzida na seção 9.2 do Capítulo 9:

$$ \ell(x) = 1 - x/\omega \qquad (17.38) $$

o que implica que

$$ -\ell'(x) = 1/\omega \qquad (17.39) $$

e

$$ \mu(x) = -\frac{\ell'(x)}{\ell(x)} = \frac{\frac{1}{\omega}}{1-\frac{x}{\omega}} = \frac{1}{\omega - x} \qquad (17.40) $$

e também

$$ {}_nL_x = \int_x^{x+n} \ell(t)\, dt = \int_x^{x+n} 1 - \frac{t}{\omega}\, dt = n - \frac{(x+n)^2 - x^2}{2\,\omega} = n\,(1 - \frac{2x+n}{2\omega}) \quad (17.41) $$

A esperança de vida na idade *x* seria dada por

$$ e_x = {}_{\omega-x}L_x \,/\, \ell_x = (\omega - x)\left(1 - \frac{2x + \omega - x}{2\omega}\right) / \,(1 - \frac{x}{\omega}) = \frac{1}{2}\,(\omega - x) \quad (17.42) $$

resultado que intuitivamente faz sentido porque indica que a esperança de vida na idade *x* é igual à metade do número de anos que ainda faltam até ω, que marca a extinção total da coorte.

17.3.2 Derivação da fórmula para ${}_na_x$ no método iterativo de Keyfitz

Como ilustração do uso de cálculo diferencial e integral, (17.37) agora pode ser usada para derivar o resultado (9.27) do Capítulo 9. Supõe-se que *-ℓ'(x)* pode ser representada por um polinômio de segundo grau $(A{\cdot}x^2 + B{\cdot}x + C)$ nos intervalos *(-n,0), (0,n)* e *(n,2n).* Integrando o polinômio sobre os três intervalos em questão, obtém-se as seguintes expressões:

$$ \frac{1}{3}\,A\,n^3 - \frac{1}{2}\,B\,n^2 + C\,n = {}_nd_{-n} $$

$$ \frac{1}{3}\,A\,n^3 + \frac{1}{2}\,B\,n^2 + C\,n = {}_nd_0 $$

$$ \frac{7}{3}\,A\,n^3 + \frac{3}{2}\,B\,n^2 + C\,n = {}_nd_n \quad (17.43.a\text{-}c) $$

Com estas expressões é possível determinar *A, B* e *C*:

$$ A = (\tfrac{1}{2}\,{}_nd_{-n} - {}_nd_0 + \tfrac{1}{2}\,{}_nd_n) \,/\, n^3 $$

$$ B = (-\,{}_nd_{-n} + {}_nd_0) \,/\, n^2 $$

$$ C = (\tfrac{1}{3}\,{}_nd_{-n} + \tfrac{5}{6}\,{}_nd_0 - \tfrac{1}{6}\,{}_nd_n) \,/\, n \quad (17.44.a\text{-}c) $$

Aplicando (17.37) ao intervalo *(0,n)* aparece a seguinte expressão para ${}_na_x$:

$$ {}_na_x = \left(\frac{1}{4}\,A\,n^4 + \frac{1}{3}\,B\,n^3 + \frac{1}{2}\,C\,n^2\right) / \,{}_nd_0 = \left(-\frac{1}{24}\,{}_nd_{-n} + \frac{1}{2}\,{}_nd_0 + \frac{1}{24}\,{}_nd_n\right)\, n \,/\, {}_nd_0 \quad (17.45) $$

ou seja, a equação (9.27) onde, para facilitar a derivação, mas sem perda de generalidade, x foi escolhido igual a 0.

17.3.3 Uma fórmula alternativa para ${}_nq_x$

A mesma lógica seguida acima pode ser usada para derivar uma fórmula alternativa para ${}_nq_x$ que considera a estrutura etária subjacente da população, da mesma forma que (8.28). Mas em vez de usar uma interpolação linear, este procedimento usa uma interpolação com um polinômio de segundo grau, do mesmo tipo usado em (17.43.a-c).

O primeiro passo consiste em multiplicar as três TEMs pelas suas populações subjacentes, para obter números de óbitos. A estes óbitos se aplica o mesmo procedimento descrito acima para obter os parâmetros *A, B* e *C* do polinômio de interpolação. Também se deriva um polinômio de interpolação $(P{\cdot}x^2 + Q{\cdot}x + R)$ para interpolar a população nas três faixas etárias. Agora ${}_nq_x$ pode ser escrito da seguinte forma:

$$ {}_nq_x = 1 - \exp\left(-\int_0^n \frac{Ax^2 + B\,x + C}{Px^2 + Q\,x + R}\,dx\right) \tag{17.46} $$

Embora (17.46) possa ser calculada de forma analítica, a avaliação da integral é trabalhosa, mesmo nos dias de hoje, com a ajuda de EXCEL e outros recursos. Mas a sua avaliação numérica em "R" é bastante fácil, como mostra o programa abaixo.

```
Mtoq <- function(mlow,mmid,mhigh,plow,pmid,phigh,n) {
    # Calcula nqx a partir de nMx usando o suposto de que os óbitos e a população
    têm um perfil etário representado por um polinômio de segundo grau
    dlow <- mlow*plow; dmid <- mmid*pmid; dhigh <- mhigh*phigh
    a = (dlow/2 - dmid + dhigh/2) / n^3
    b = (-dlow + dmid) / n^2
    c = (dlow/3 + 5*dmid/6 - dhigh/6) / n
    p = (plow/2 - pmid + phigh/2) / n^3
    q = (-plow + pmid) / n^2
    r = (plow/3 + 5*pmid/6 - phigh/6) / n
    integrand <- function(x) {(a*x^2+b*x+c)/(p*x^2+q*x+r)}
    result <- integrate(integrand,0,n)
    nqx <- 1 - exp(-result[[1]])
    nqx
}
```
(17.47)

Seria possível melhorar este programa, acrescentando alguns recursos adicionais, como os seguintes:

1. Da mesma forma como no programa axKeyfitz (17.24), seria aconselhável suavizar as curvas de ${}_nM_x$ e da população antes de aplicar os cálculos mostrados acima.
2. Em vez de aplicar o procedimento a três faixas etárias de cada vez, poder-se-ia logo aplicá-lo aos vetores que descrever ${}_nM_x$ e a população subjacente.
3. Seria desejável acrescentar um teste para verificar se $A{\cdot}x^2 + B{\cdot}x + C$ e $P{\cdot}x^2 + Q{\cdot}x + R$ não assumem valores negativos dentro do intervalo *(0,n),* porque isso poderia prejudicar o realismo do modelo.

A Tabela 17.1 mostra os resultados de diferentes formas para calcular ${}_nq_x$ e os compara com o procedimento (17.47). As diferenças na parte inferior da tabela referem-se à raiz média dos desvios quadráticos relativos. Como se pode perceber, o segundo método de Keyfitz (9.28) se aproxima mais do procedimento (17.47), enquanto os métodos de Reed-Merrell e Greville são um pouco mais próximos do que as fórmulas simples (9.21) e (9.22).

Tabela 17.1: Valores de ${}_5q_x$ masculinos para o Brasil, 2012-2014

	(9.21)	(9.22)	(9.25)	(9.26)	(9.28)	(17.47)
0	0.015921	0.015921	0.015931	0.015930		
5	0.001309	0.001309	0.001309	0.001309	0.001344	0.001342
10	0.001968	0.001968	0.001968	0.001968	0.001959	0.001960
15	0.010020	0.010019	0.010023	0.010023	0.010022	0.010022
20	0.013207	0.013207	0.013214	0.013213	0.013206	0.013207
25	0.012590	0.012590	0.012596	0.012596	0.012590	0.012590
30	0.013335	0.013335	0.013342	0.013342	0.013345	0.013344
35	0.015627	0.015627	0.015636	0.015635	0.015662	0.015662
40	0.019983	0.019983	0.019999	0.019997	0.020033	0.020033
45	0.027453	0.027451	0.027481	0.027478	0.027531	0.027529
50	0.038624	0.038619	0.038679	0.038673	0.038815	0.038804
55	0.056266	0.056251	0.056377	0.056365	0.056650	0.056639
60	0.079988	0.079943	0.080199	0.080173	0.080676	0.080640
65	0.116672	0.116532	0.117074	0.117021	0.117955	0.117901
70	0.176243	0.175745	0.176975	0.176860	0.177928	0.176538
75	0.253290	0.251751	0.254265	0.254046		
Diferença	0.002118	0.002187	0.001947	0.001962	0.000576	0.000000

Fonte: Cálculos baseados na Tabela 8.3.

17.3.4 A entropia da tábua de vida

Outro uso de cálculo diferencial e integral é para introduzir o conceito de *entropia* da tábua de vida que quantifica a sensibilidade da esperança de vida a mudanças no nível global da mortalidade por idade $\mu(x)$ (Goldman, 1986; Hakkert, 1987; Hill, 1993). Suponha que o nível de mortalidade em todas as idades seja multiplicado por um fator r, ou seja, $\mu_r(x) = r\,\mu(x)$. Então a função $\ell_r(x)$ correspondente a este novo nível de mortalidade seria

$$\ell_r(x) = \exp\left(-\int_0^x r\,\mu(t)\,dt\right) = \ell^r(x) \tag{17.48}$$

ou seja, multiplicar toda a função por um fator r equivale a elevar a função $\ell(x)$ ao expoente r. Usando (17.36) tem-se que

$$e_r(0) = \int_0^\infty \ell_r(x)dx = \int_0^\infty \ell^r(x)dx \tag{17.49}$$

Agora, diferenciando $e_r(0)$ no ponto $r=1$ e dividindo por $e_r(0)$, se obtém

$$H = -\frac{1}{e_r(0)}\frac{d}{dx}\,e_r(0)|_{r=1} = -\int_0^\infty \ln(\ell(x))\;\ell(x)\;dx \;/\; \int_0^\infty \ell(x)dx \tag{17.50}$$

O valor de H depende da forma da curva $\ell(x)$ e inclusive pode ser considerado uma medida da distribuição da mortalidade por idades (Anson, 2002). Se $\ell(x)$ for um retângulo, retratando uma situação em que (quase) ninguém morre antes da idade a e depois (quase) todos morrem por volta desta idade, a multiplicação de $\mu(x)$ por r tem praticamente nenhum efeito sobre a esperança de vida e H acaba sendo quase 0. Por outro lado, se $\ell(x)$ for uma função exponencial negativa ($\ell(x) = e^{-\alpha r}$), H é igual a 1, ou seja, um aumento de 1% no nível global de mortalidade ($\mu(x)$) diminuirá a esperança de vida ao nascer em 1%. Na maioria das tábuas de vida H se encontra entre estes extremos, geralmente mais próximo a 0 do que a 1.

17.3.5 Derivadas parciais

Para os propósitos deste livro, as derivadas se referem a apenas uma variável de cada vez. Entretanto, em certas aplicações as quantidades podem ser diferenciadas em relação a mais de uma variável. Por exemplo, a variável $\ell(x,t)$ depende tanto da idade x da pessoa como do tempo t e em certas aplicações tanto uma como outra derivada é relevante. Como este livro não entra nesse tipo de análises, não vale a pena fazer maiores considerações a respeito. Só cabe apontar que, em situações onde a derivada pode referir-se a mais de uma variável, se usa o símbolo de *diferenciação parcial*, com ∂ redondo em vez de d reto. Portanto, em situações onde apenas a idade é relevante se escreve

$$\frac{d}{dx}\,\ell(x)$$

mas quando existem duas ou mais variáveis relevantes se escreve

$$\frac{\partial}{\partial x}\,\ell(x,t) \quad \text{ou} \quad \frac{\partial}{\partial t}\,\ell(x,t)$$

17.4 PRINCÍPIOS DA ÁLGEBRA MATRICIAL

Da mesma forma como no caso do cálculo diferencial e integral, seria irrealista esperar que um livro como este possa fornecer uma introdução completa aos princípios da álgebra matricial, um assunto que normalmente exige um curso universitário de pelo menos um semestre para dominar. Para uma introdução mais completa, o leitor é referido a diversos livros sobre o tema, como Shokranian (2009). Mas para efeitos do uso da álgebra matricial na demografia basta introduzir apenas algumas noções básicas. Como o nome sugere, a álgebra matricial trata da forma como se pode manipular matrizes para fazer cálculos. Matrizes são conjuntos de números organizados de forma retangular, por exemplo *n* filas por *m* colunas. Quando *n=1* ou *m=1*, a matriz também é chamada um *vetor de fila* ou um *vetor de coluna*. Vetores são números generalizados para mais de uma dimensão. Por exemplo, o vetor *(1990, 2015, 25)* poderia representar o ano de nascimento, o ano corrente e idade atual de uma pessoa. Quando um vetor de tamanho *m* é transformado linearmente num vetor de tamanho *n*, é preciso especificar *n·m* fatores de multiplicação que configuram a *matriz de transformação.*

Matrizes podem ser somadas (desde que tenham a mesma dimensão) e multiplicadas (desde que o número de colunas da primeira matriz seja igual ao número de filas da segunda). A definição da multiplicação de duas matrizes é a seguinte:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} =$$

$$\begin{pmatrix} a_{11}\cdot b_{11} + a_{12}\cdot b_{21} + a_{13}\cdot b_{31} & a_{11}\cdot b_{12} + a_{12}\cdot b_{22} + a_{13}\cdot b_{32} \\ a_{21}\cdot b_{11} + a_{22}\cdot b_{21} + a_{23}\cdot b_{31} & a_{21}\cdot b_{12} + a_{22}\cdot b_{22} + a_{23}\cdot b_{32} \end{pmatrix} \tag{17.51}$$

As matrizes quadradas têm algumas características especiais. Elas podem ser consideradas transformações lineares do espaço n-dimensional em si mesmo. Uma característica importante desta transformação é se ela preserva o mesmo número de dimensões ou se ela as reduz, por exemplo transformando um espaço 3-dimensional num plano 2-dimensional. As transformações que preservam todas as dimensões podem ser invertidas (tem como voltar), mas se o resultado possui menos dimensões do que o espaço original isso não é possível. Para saber qual é o caso, calcula-se o *determinante* da matriz. No caso de uma matriz de 2 por 2, a fórmula é muito simples:

$$det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}\,a_{22} - a_{21}\,a_{12} \tag{17.52}$$

No caso de uma matriz de 3 por 3, a fórmula é

$$det\begin{pmatrix} a_{11} & \cdots & a_{13} \\ \vdots & \ddots & \vdots \\ a_{31} & \cdots & a_{33} \end{pmatrix} = a_{11}\, a_{22}\, a_{33} + a_{12}\, a_{23}\, a_{31} + a_{13}\, a_{21}\, a_{32} - a_{31}\, a_{22}\, a_{13} - a_{32}\, a_{23}\, a_{12} - a_{33}\, a_{21}\, a_{12} \quad (17.53)$$

Para dimensões mais altas aplicam-se fórmulas recursivas para reduzir o número de dimensões, que estão além do escopo deste livro. Entretanto, geralmente não há necessidade de calcular determinantes manualmente porque a função MATRIZ.DETERM() de EXCEL o faz automaticamente. Para esse fim basta indicar a área da planilha (por exemplo, MATRIZ.DETERM(D5:G8)) onde se encontra a matriz. A área precisa ser quadrada; senão EXCEL acusa um erro. Em "R" a função que calcula o determinante é simplesmente det().

Nas fórmulas (17.51), (17.52) e (17.53), as matrizes foram escritas por extenso, com todos os seus elementos. Entretanto, em muitas fórmulas elas são representadas apenas por um símbolo, geralmente uma maiúscula em negrito, para deixar claro que se trata de uma matriz ou um vetor, não de um número escalar. As matrizes quadradas com determinante não igual a 0 podem ser *invertidas*, ou seja, para uma matriz quadrada $\boldsymbol{A}$ com $det(\boldsymbol{A}) \neq 0$ é possível encontrar uma matriz $\boldsymbol{A^{-1}}$, de tal forma que

$$A\, A^{-1} = A^{-1}\, A = I \quad (17.54)$$

onde $\boldsymbol{I}$ é a matriz de identidade, uma matriz quadrada com as mesmas dimensões de $\boldsymbol{A}$ que consiste de 1 em todas as posições da diagonal (da esquerda superior até a direita inferior) e 0 em todas as posições fora da diagonal. Não se entrará aqui na questão como $\boldsymbol{A^{-1}}$ pode ser manualmente calculada a partir de $\boldsymbol{A}$.

Uma das aplicações mais básicas da álgebra matricial é a solução de um sistema de n equações lineares com n variáveis desconhecidas $x_1, \ldots, x_n$:

$$\begin{aligned} a_{11} \cdot x_1 + a_{12} \cdot x_2 + \ldots\ldots + a_{1n} \cdot x_n &= y_1 \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 + \ldots\ldots + a_{2n} \cdot x_n &= y_2 \\ \ldots\ldots\ldots & \\ a_{n1} \cdot x_1 + a_{n2} \cdot x_2 + \ldots\ldots + a_{nn} \cdot x_n &= y_n \end{aligned} \quad (17.55)$$

Em notação matricial, esse sistema pode ser escrito simplesmente como

$$\boldsymbol{A\, x = y} \quad (17.56)$$

onde $\boldsymbol{A}$ simboliza a matriz de n por n de coeficientes $a_{11}, \ldots, a_{nn}$, $\boldsymbol{x}$ é um vetor de coluna com os valores desconhecidos de $x_1, \ldots, x_n$ e $\boldsymbol{y}$ é um vetor de coluna com os valores conhecidos de $y_1, \ldots, y_n$. A solução desse sistema pode ser escrita simplesmente como

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\,\boldsymbol{y} \tag{17.57}$$

Em "R", o comando correspondente é

$$> \; x <\text{- } solve(A, y) \tag{17.58}$$

O EXCEL também permite fazer cálculos matriciais, com as funções MMULT() para a multiplicação e MMINVERSE() para a inversão de matrizes. Quando essas funções são chamadas, inicialmente aparece apenas a primeira célula do resultado. Para ver o resto, é preciso apertar F2 e depois a combinação Shift+Ctrl+Enter. Com estas regras básicas de cálculo, é possível manipular matrizes de uma forma muito parecida com números comuns e obter resultados em formato de matrizes.

As matrizes têm muitos usos potenciais na demografia (Caswell, 2001), alguns dos quais serão discutidos nos próximos capítulos. A utilidade principal está na sua aplicação às projeções de população que será abordada no Capítulo 21.

17.5 CONCEITOS BÁSICOS DE ESTATÍSTICA

Como já se mencionou no Capítulo 4, a demografia tradicionalmente sempre foi vista como uma ciência de grandes números, onde o investigador tipicamente tem acesso ao universo estudado inteiro, por meio do censo ou do registro civil. Entretanto, cada vez há mais aplicações em que técnicas estatísticas e probabilísticas têm um papel a cumprir. Por exemplo, muitos parâmetros demográficos hoje em dia são estimados por meio de inquéritos amostrais. Por outro lado, as estimativas para pequenas áreas, mesmo que se baseiem no universo inteiro, precisam lidar com números reduzidos de observações que exigem uma interpretação probabilística. Técnicas estatísticas como a regressão múltipla são amplamente aplicadas para analisar relações entre diferentes variáveis demográficas. Como no caso do cálculo diferencial e integral que foi abordado nas seções anteriores, a introdução sistemática das técnicas estatísticas e probabilísticas exige um tratamento aprofundado que vai muito além das possibilidades deste livro. Entretanto, como certas técnicas estatísticas, tais como a estimação bayesiana de parâmetros demográficos, estão se tornando mais comuns na literatura da área, é preciso pelo menos dedicar algumas palavras a estas técnicas, para situá-las minimamente.

A seção 17.3 tratou da conversão de várias das funções da tábua de vida para funções contínuas. Várias destas funções podem ser entendidas como distribuições probabilísticas. Por exemplo, a função contínua $\ell(x)$ que foi introduzida na seção 17.3 pode ser entendida como 1 menos a função cumulativa de probabilidade $F_X(x) = P(X \leq x)$ que define a distribuição de probabilidade para uma

variável aleatória X que descreve a idade de morte individual. A densidade probabilística $f_X(x)$ da idade de morte x, como sempre, é a derivada de $F_X(x)$, ou seja, $f_X(x) = - \ell'(x)$. Usando o conceito de força da mortalidade, a densidade de probabilidade da idade de morte pode ser escrita como

$$f_X(x) = -\ell'(x) = \ell(0)\mu(x)\exp\left(-\int_0^x \mu(t)\,dt\right) \tag{17.59}$$

No caso onde $\ell(x)$ descreve uma função de De Moivre, a função de probabilidade acumulada para a variável aleatória $T(x)$ seria

$$F_{T(x)}(t) = 1 - \frac{\omega - x - t}{\omega - x} = \frac{\omega - x - (\omega - x - t)}{\omega - x} = \frac{t}{\omega - x} \tag{17.60}$$

e a função de densidade simples

$$f_{T(x)}(t) = \frac{d}{dt}F_{T(x)}(x) = \frac{d}{dt}\left(\frac{t}{\omega - x}\right) = \frac{1}{\omega - x} \tag{17.61}$$

A função contínua $T(x)$ (que não deve ser confundida com a variável discreta T_x) simboliza a variável aleatória $X - x$, o tempo restante de vida de um indivíduo a partir da idade x. Quando for interpretada desta forma, a sua distribuição de probabilidade pode ser escrita como

$$F_{T(x)}(t) = P(T(x) \leq t) = {}_tq_x \tag{17.62}$$

e a função de densidade simples como

$$f_{T(x)}(t) = {}_tp_x\mu(x+t) \tag{17.63}$$

A mesma função também pode ser calculada como variável discreta (número de anos inteiros vividos) e quando calculada desta forma geralmente é identificada com o símbolo $K(x)$. A sua distribuição de probabilidade é a seguinte:

$$P[K(x) = k] = {}_kp_x - {}_{k+1}p_x = \frac{\ell_{x+k}}{\ell_x} - \frac{\ell_{x+k+1}}{\ell_x} = \frac{{}_1d_x}{\ell_x} \tag{17.64}$$

Esta é a probabilidade de que um indivíduo com idade (x) sobreviva t anos e morra no instante de tempo seguinte, entre t e $t+dt$. A esperança de vida de uma população a partir da idade x é a expectativa matemática da distribuição de probabilidade $f_X(x) = - \ell'(x)$: $e_0 = E\{x\, f_X(x)\} = E\{- x\, \ell'(x)\}$.

Existem duas interpretações distintas sobre o conceito de probabilidade: a *objetivista* ou *frequentista* e a *subjetivista* ou *bayesiana*. Na primeira concepção, a probabilidade de um evento é definida como a sua frequência relativa caso a experiência subjacente seja repetida muitas

vezes. Por exemplo, ao selecionar uma carta de um baralho aleatoriamente, a priori não há como saber que naipe sairá, mas se a experiência for repetida muitas vezes e supondo que o baralho seja regular, a tendência é que cada naipe será selecionado 25% das vezes. O problema desta concepção é que muitas experiências da vida real não podem ser repetidas, pelo menos não nas mesmas circunstâncias, o que torna a noção de uma frequência relativa dentro de uma sequência de experiências idênticas um pouco hipotética. Os textos de Chiang (1968, 1984) são a referência padrão para a teoria da probabilidade referente à tábua de vida e alguns outros indicadores demográficos, dentro da perspectiva objetivista. Entre outras coisas, esses textos calculam as variâncias e covariâncias de muitas medidas demográficas quando são construídas a partir de amostras, tema que está além dos propósitos deste livro.

A interpretação bayesiana de probabilidade parte da ideia subjetiva do grau de certeza que o observador possui sobre o resultado esperado. Sem qualquer informação anterior, um observador da seleção de uma carta de um baralho deve considerar que todos os naipes são igualmente prováveis, portanto 25% cada um. Mas se o resultado for "copas" cinco vezes seguidas, o observador pode desconfiar de que talvez o baralho não seja regular (com 13 cartas de cada naipe) e em função disso pode adaptar a sua avaliação subjetiva das probabilidades para a próxima seleção. Portanto, na concepção bayesiana a probabilidade de um evento é uma distribuição probabilística que depende da experiência acumulada do observador.

Tomando um exemplo mais relevante desde o ponto de vista demográfico, um observador que precisa estimar a esperança de vida de uma certa população a partir da sobrevivência de diferentes indivíduos, provavelmente tomaria em conta qual é a esperança de vida conhecida de outras populações parecidas. Se o primeiro indivíduo sobrevive até os 98 anos, seria racional supor que a população subjacente tenha uma esperança de vida relativamente alta, mas provavelmente não 98 anos pois nenhuma população conhecida tem uma esperança de vida tão elevada. Uma aposta razoável poderia ser, por exemplo, 80 anos. Se o segundo indivíduo sobrevive até os 85 anos, isso confirma a ideia de uma esperança de vida relativamente alta e o observador poderia ajustar a sua distribuição subjetiva de expectativas para uma média de 81 anos. Por outro lado, se o segundo indivíduo morre aos 50 anos, isso sugere que a primeira observação tenha sido excepcional e que seria melhor ajustar a curva para baixo. Continuando desta forma, as expectativas do observador vão se modificando em função de novos dados.

Atualmente há uma tendência ao aumento de análises de dados demográficos baseadas em algum tipo de metodologia bayesiana. Alguns dos exemplos são os seguintes:

1. Desde 2012 a metodologia de projeção demográfica seguida pela Divisão de População das Nações Unidas se baseia na análise bayesiana da mortalidade e fecundidade que incorpora tanto dados sobre a história de cada país como dados sobre as tendências observadas em países mais adiantados na sua transição demográfica (Raftery et al., 2012). Em vez de valores determinísticos, os resultados deste tipo de projeções são distribuições probabilísticas que descrevem as probabilidades de que certos parâmetros demográficos terão um ou outro valor no futuro. Wisniowski et al. (2015) propõem uma metodologia alternativa, também baseada em estatística bayesiana.

2. Freire (2001), Freire, Gonzaga e Gomes (2019) e Muniz (2018) adaptaram métodos existentes de projeção demográfica de pequenas áreas para uma abordagem probabilística usando métodos bayesianos empíricos.

3. Neves e Migon (2004) propuseram um procedimento bayesiano para a graduação de tábuas de vida, usando uma função matemática (a de Makeham, descrita em (20.3)) com parâmetros estimados com base em critérios bayesianos. Olivieri e Patacco (2011) propõem uma metodologia bayesiana alternativa para a graduação de tábuas de vida.

4. Outras aplicações foram feitas por Assunção, Potter e Cavenaghi (2002), Assunção et al. (2005), Potter et al. (2010), Schmertmann et al. (2013) e Schmertmann e Gonzaga (2018).

Dentro da abordagem bayesiana ainda existem duas vertentes distintas que são conhecidas como o método bayesiano *clássico*, *apriorístico* ou *pleno* e o método bayesiano *empírico*. Como o nome sugere, o método apriorístico parte de uma distribuição inicial do parâmetro a estimar (geralmente notado como θ) que é considerado "razoável" por razões apriorísticas. No exemplo anterior, da estimação da esperança de vida de uma população, a distribuição inicial de $\theta = e_0$ poderia ser uma distribuição uniforme no intervalo de 40 a 90 anos, considerando que a priori não há nenhuma informação que privilegie um valor sobre outro, exceto que esperanças de vida de menos de 40 ou mais de 90 anos são muito improváveis. Esta distribuição apriorística depois vai sofrer modificações em função da recepção de novos dados (para um exemplo, ver Schmertmann e Gonzaga, 2018).

Na abordagem bayesiana empírica, por outro lado, os próprios dados são usados para definir a distribuição de θ. Geralmente, as fórmulas resultantes desta metodologia exibem uma estrutura que pode ser descrita como *contração* ("shrinkage", em inglês), conforme a descrição dada por Marshall (1991). Assunção et al. (2005) ilustram isso para o caso da estimação da fecundidade e Freire, Gonzaga e Gomes (2019) para a projeção de pequenas áreas. Supõe-se que um conjunto de parâmetros θ_i (por exemplo, TFTs municipais) seja distribuído segundo uma distribuição normal com valor esperado μ e variância σ^2. Além disso, supõe-se que os parâmetros θ_i são estimados imperfeitamente por estimadores que, além do seu componente sistemático θ_i, contêm erros aleatórios u_i que se distribuem normalmente com um valor esperado 0 e variância ω_i^2:

$$\widehat{\vartheta_i} = \vartheta_i + u_i \; com \; \vartheta_i \sim N(\mu, \sigma^2) \; e \; u_i \sim M(0, \omega_i^2) \tag{17.65}$$

Então, pode ser demonstrado que o melhor estimador (não enviesado e de variância mínima) para θ_i não é $\widehat{\vartheta_i}$ mas

$$\widetilde{\vartheta_i} = \widehat{\vartheta_i} + \frac{\omega_i^2}{\omega_i^2 + \sigma^2} (\mu - \widehat{\vartheta_i}) \tag{17.66}$$

A interpretação desta fórmula é a seguinte: Se ω_i^2 for pequeno em comparação com σ^2, pode-se confiar bastante em , mas se ω_i^2 for grande, não se pode confiar muito neste estimador do

parâmetro e é mais seguro escolher um valor mais próximo de μ. A priori não se sabe quais são os valores de ω_i^2 e de σ^2, pois estes valores precisam ser estimados a partir dos dados, o que modifica (17.66). Mas o princípio de contração ("shrinkage") ilustrado por (17.66) é bastante típico de muitos estimadores bayesianos empíricos, inclusive dos estimadores da fecundidade municipal derivados por Assunção et al. (2005).