DEMOGRAPHIC ANALYSIS: LECTURE NOTES

19 September 2019

Joseph George Caldwell, PhD (Statistics)
1432 N Camino Mateo, Tucson, AZ 85745-3311 USA
Tel. (001)(520)222-3446, E-mail jcaldwell9@yahoo.com
Website http://www.foundationwebsite.org

Contents

1. OVERVIEW

THIS PRESENTATION IS A SURVEY OF THE BASIC CONCEPTS OF DEMOGRAPHIC ANALYSIS. IT IDENTIFIES AND DESCRIBES THE MAJOR ASPECTS AND TECHNIQUES OF DEMOGRAPHIC ANALYSIS, AND ILLUSTRATES SOME OF THE TECHNIQUES USING DATA AND COMPUTER PROGRAMS THAT ARE AVAILABLE FREE FROM THE INTERNET.

THE COURSE DOES NOT INCLUDE DESCRIPTIVE DEMOGRAPHIC MATERIAL, SUCH AS SUMMARIES OF THE CURRENT WORLD POPULATION, POPULATION TRENDS AND PROSPECTS, OR DEMOGRAPHIC ASPECTS OF SUBSTANTIVE FIELDS SUCH AS ECONOMICS, EDUCATION AND HEALTH.

THE COURSE MATERIAL IS DIVIDED INTO TWO PARTS, THE FIRST PART, DEALING WITH OLDER (“CLASSICAL”) TECHNIQUES THAT CAN BE IMPLEMENTED USING MAINLY ARITHMETIC, BASIC ALGEBRA, AND MATRIX ARITHMETIC; AND THE SECOND PART, DEALING WITH MODERN TECHNIQUES THAT UTILIZE TECHNIQUES OF MATRIX ALGEBRA, CALCULUS AND STATISTICS.

A LIST OF THE SPECIFIC TOPICS COVERED IS LISTED IN THE COURSE SYLLABUS (INCLUDED AT END OF THESE NOTES).

THE PRIMARY TEXT FOR THE COURSE IS *THE METHODS AND MATERIALS OF DEMOGRAPHY* 2ND EDITION BY JOSEPH S. SIEGEL AND DAVID A. SWANSON, EDS. (ELSEVIER ACADEMIC PRESS, 2004). FOR THE TOPIC OF FORECASTING OF DEMOGRAPHIC COMPONENTS, SUCH AS MORTALITY, THE TEXT, *DEMOGRAPHIC FORECASTING* BY FEDERICO GIROSI AND GARY KING (PRINCETON UNIVERSITY PRESS, 2008) IS USED. BOTH OF THESE TEXTS ARE AVAILABLE FREE FROM THE INTERNET (WEBSITE ADDRESS GIVEN IN THE REFERENCES). THERE ARE MANY OTHER EXCELLENT TEXTS IN DEMOGRAPHIC ANALYSIS. THESE TWO WERE SELECTED AS PRIMARY TEXTS FOR THE COURSE BECAUSE THEY ARE COMPREHENSIVE AND DETAILED, AND INCLUDE BOTH CLASSICAL AND MODERN TECHNIQUES.

THE ORDER OF COVERAGE OF THE COURSE TOPICS IS SPECIFIED IN THE SYLLABUS. THE ORDER OF TOPICS AND THE LEVEL OF DETAIL OF THE COVERAGE OF THE TOPICS DIFFERS FROM THAT OF THE PRECEDING TEXTS. THE MAJOR REASONS FOR THE DIFFERENCE ARE TWO. FIRST, THE COURSE EMPHASIZES THE USE OF COMPUTER SOFTWARE TO PERFORM POPULATION-BASED FORECASTING.

SECOND, THE COURSE IS A *SURVEY* COURSE, WITH THE OBJECTIVE OF IDENTIFYING MAJOR TOPICS IN DEMOGRAPHIC ANALYSIS AND DESCRIBING THEIR NATURE, BUT NOT PROVIDING DETAILED DESCRIPTION OF THEIR MATHEMATICAL DERIVATION OR APPLICATION. AN ANALOGY OF THE APPROACH WOULD BE A DESCRIPTION OF A MULTIPLE REGRESSION MODEL, SHOWING THE MODEL SPECIFICATION (FORMULA) AND ASSUMPTIONS AND IDENTIFYING THE APPROACH TO ESTIMATION (E.G., LEAST-SQUARES, MAXIMUM LIKELIHOOD), WITHOUT DESCRIBING FORMULAS OR ALGORITHMS FOR IMPLEMENTING THE ESTIMATION PROCEDURES, BUT PROVIDING EXAMPLES OF APPLICATION OF THE TECHNIQUE USING AVAILABLE STATISTICAL COMPUTER PROGRAMS.

FOR THE COMPUTER ANALYSIS, COMPUTER SOFTWARE WILL BE USED THAT IS AVAILABLE FREE FROM THE INTERNET. THE PRIMARY TOOL FOR MAKING POPULATION PROJECTIONS AND POPULATION-BASED FORECASTS WILL BE THE *SPECTRUM* PACKAGE OF DEMOGRAPHIC COMPUTER PROGRAMS, AVAILABLE FROM THE UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT (USAID) *HEALTH POLICY PLUS PROJECT* OR ITS PARTNERS, SUCH AS *AVENIR HEALTH*.

2. DEFINITIONS AND SCOPE OF DEMOGRAPHY AND DEMOGRAPHIC ANALYSIS

DEMOGRAPHY IS THE SCIENTIFIC STUDY OF HUMAN POPULATION, INCLUDING ITS SIZE, GEOGRAPHIC DISTRIBUTION, AND COMPOSITION, AND THE FACTORS THAT AFFECT THEM.

THE WORD DEMOGRAPHY COMES FROM TWO ANCIENT GREEK WORDS, DEMOS, MEANING "PEOPLE," AND GRAPHY, MEANING "WRITING ABOUT OR RECORDING SOMETHING." SO, LITERALLY, DEMOGRAPHY MEANS "WRITING ABOUT PEOPLE.” OVER TIME, HOWEVER, THE TERM HAS BEEN APPLIED TO INCLUDE STUDY OF LIVING POPULATIONS OF ANY SORT, INCLUDING PLANTS AND ANIMALS. IN THIS PRESENTATION, WE ARE CONCERNED ONLY WITH HUMAN POPULATIONS.

MAJOR ASPECTS OF DEMOGRAPHY

THE FIVE MAJOR ASPECTS OF DEMOGRAPHY ARE:

- SIZE
- GEOGRAPHIC DISTRIBUTION
- COMPOSITION (DISTRIBUTION BY FACTORS SUCH AS AGE, SEX, AND RACE, THAT AFFECT POPULATION GROWTH)
- POPULATION DYNAMICS (CHANGE IN SIZE OR COMPOSITION OVER TIME; FACTORS THAT AFFECT GROWTH, SUCH AS BIRTH RATES, DEATH RATES AND MIGRATION RATES)
- SOCIOECONOMIC DETERMINANTS AND CONSEQUENCES OF POPULATION CHANGE

THESE ASPECTS WILL NOW BE DISCUSSED.

SIZE: TOTAL NUMBER OF PEOPLE IN AN AREA OF INTEREST, CLASSIFIED IN MAJOR CATEGORIES, INCLUDING:

- *DE FACTO*: PEOPLE PRESENT IN A GIVEN AREA AT A GIVEN TIME
- *DE JURE*: PEOPLE ASSOCIATED WITH A GIVEN AREA ACCORDING TO SPECIFIED CRITERIA, SUCH AS LEGAL RESIDENCE OR USUAL RESIDENCE
- NOMADS
- MILITARY PERSONNEL
- FOREIGN OFFICIAL PERSONNEL
- LEGAL ALIENS
- ILLEGAL ALIENS
- PERSONS IN INSTITUTIONS
- PERSONS IN HOUSEHOLDS

GEOGRAPHIC DISTRIBUTION:

- AREAS OR REGIONS (E.G., COUNTRIES, PROVINCES, STATES, COUNTIES)
- PLACES (METROPOLITAN AREAS, CITIES, TOWNS, VILLAGES)

COMPOSITION:

- ASCRIBED CHARACTERISTICS: AGE, SEX, RACE, YEAR OF BIRTH, PLACE OF BIRTH

- ACHIEVED OR ACQUIRED CHARACTERISTICS: BASIC SOCIOECONOMIC CHARACTERISTICS, INCLUDING NATIVITY, LANGUAGE, ETHNICITY, ANCESTRY, RELIGION, CITIZENSHIP, MARITAL STATUS, HOUSEHOLD CHARACTERISTICS, LIVING ARRANGEMENTS, EDUCATIONAL LEVEL, SCHOOL ENROLMENT, LABOR-FORCE STATUS, INCOME AND WEALTH
- CHARACTERISTICS ASSOCIATED WITH ANY FIELD RELATED TO DEMOGRAPHY: INSURANCE, HEALTH, DISABILITY, INSTITUTIONAL STATUS, COMMERCE, MARKET RESEARCH, URBAN AND REGIONAL PLANNING, TRANSPORTATION PLANNING, POLITICS, DEFENSE

POPULATION DYNAMICS (CHANGE OVER TIME):

- NUMBERS OF BIRTHS, DEATHS, IMMIGRANTS (“IN-MIGRANTS”), EMIGRANTS (“OUT-MIGRANTS”)
- POPULATION RATES OF CHANGE (GROWTH OR DECLINE)
- CHANGES IN GEOGRAPHIC DISTRIBUTION OR COMPOSITION
- CHANGE IN STATUS (AGE, MARITAL STATUS, SOCIOECONOMIC STATES)
- COMPONENTS OF CHANGE:
    - NARROW SCOPE: BIRTHS, DEATHS, MIGRANTS
    - BROAD SCOPE: ALL VARIABLES AFFECTING THESE (E.G., MARRIAGE, SICKNESS, EMPLOYMENT)

SOCIOECONOMIC DETERMINANTS AND CONSEQUENCES OF POPULATION CHANGE:

- MANY VARIABLES IN MANY FIELDS OF INTEREST (E.G., MARITAL STATUS, EDUCATION, EMPLOYMENT, HEALTH, ENVIRONMENT)

MAJOR CATEGORIES OF DEMOGRAPHY:

- DEMOGRAPHIC ANALYSIS (METHODS AND MATERIALS)
- POPULATION STUDIES (DESCRIPTION OF STATUS AND TRENDS IN A SUBSTANTIVE AREA, SUCH AS POVERTY, HEALTH OR ENVIRONMENT)

THIS PRESENTATION ADDRESSES DEMOGRAPHIC ANALYSIS, NOT ON POPULATION STUDIES.

CLASSICAL VS. MODERN DEMOGRAPHY

UNTIL ABOUT THE MIDDLE OF THE TWENTIETH CENTURY, THE MATHEMATICS INVOLVED IN MOST DEMOGRAPHIC ANALYSIS WAS VERY BASIC, INVOLVING MOSTLY JUST ARITHMETIC AND BASIC ALGEBRA. THE MAIN REASON FOR THIS IS THAT THE SALIENT FEATURE OF POPULATION CHANGE CAN BE REPRESENTED BY A VERY SIMPLE EQUATION, KNOWN AS THE DEMOGRAPHIC EQUATION (OR THE BALANCING EQUATION OR THE DEMOGRAPHIC BALANCING EQUATION), WHICH STATES THAT THE CHANGE IN THE POPULATION OF AN AREA OVER A PERIOD OF TIME IS EQUAL TO THE NUMBER OF BIRTHS MINUS THE NUMBER OF DEATHS PLUS THE NUMBER OF IMMIGRANTS MINUS THE NUMBER OF EMIGRANTS:

$$P_{end} - P_{beg} = \text{BIRTHS} – \text{DEATHS} + \text{IMMIGRANTS} – \text{EMIGRANTS},$$

OR

$$P_{end} = P_{beg} + B – D + I – E,$$

WHERE

P_{end} = POPULATION SIZE (NUMBER OF PERSONS) AT END OF PERIOD

P_{beg} = POPULATION SIZE AT BEGINNING OF PERIOD

B = NUMBER OF BIRTHS DURING PERIOD

D = NUMBER OF DEATHS DURING PERIOD

I = NUMBER OF IMMIGRANTS DURING PERIOD

E = NUMBER OF EMIGRANTS DURING PERIOD.

THE MAGNITUDES OF THESE QUANTITIES DEPEND ON A VARIETY OF FACTORS, SUCH AS AGE AND LOCATION. THIS FORMULA CAN BE USED TO ESTIMATE RATES OF CHANGE BY AGE, SEX AND LOCATION, WHICH MAY BE USED TO PROJECT FUTURE POPULATION LEVELS, DISTRIBUTION AND COMPOSITION CONDITIONAL ON PAST OBSERVED RATES OR ON OTHERWISE-SPECIFIED RATES.

THE ESSENTIAL FUNCTION OF AGENCIES CONCERNED WITH DEMOGRAPHIC DATA WAS TO ASSEMBLE DATA FROM REGISTRATION SYSTEMS, CENSUSES AND SURVEYS, AND ESTIMATE NUMBERS AND RATES ASSOCIATED WITH THE BASIC COMPONENTS. UNTIL THE MID-TWENTIETH CENTURY, THE COMPUTATIONS INVOLVED IN THE PROCESSING OF DEMOGRAPHIC DATA WERE DONE MANUALLY, WITHOUT THE AID OF CALCULATING MACHINES OR ELECTRONIC COMPUTERS. THE ACCOMPLISHMENT OF THIS BASIC FUNCTION, USING THE METHODS OF BASIC ALGEBRA, IS REFERRED TO AS “CLASSICAL,” OR “TRADITIONAL” DEMOGRAPHY.

OVER TIME, THE DEMAND FOR DEMOGRAPHIC DATA INCREASED, AND EFFORT WAS EXPENDED ON THE DEVELOPMENT OF STATISTICAL TECHNIQUES FOR MAKING IMPROVED ESTIMATES OF DEMOGRAPHIC QUANTITIES, WHERE THE TERM “IMPROVED” REFERS TO INCREASED DETAIL, HIGHER PRECISION, AND ASSESSMENT OF ACCURACY (ESTIMATION OF PRECISION, CHARACTERIZATION OF BIAS). THESE NEW DEVELOPMENTS INVOLVED USE OF THE MODERN METHODS OF STATISTICAL ANALYSIS, SUCH AS ESTIMATION AND FORECASTING. THE USE OF THESE STATISTICAL TECHNIQUES INVOLVES THE USE OF CALCULUS AND MATRIX ALGEBRA, AND MANY OF THEM REQUIRE A COMPUTER TO PERFORM THE REQUIRED NUMERICAL CALCULATIONS. THE METHODS OF DEMOGRAPHIC ANALYSIS THAT INVOLVE THE USE OF METHODS BEYOND BASIC ALGEBRA ARE REFERRED TO AS “MODERN” DEMOGRAPHY. IN MOST INSTANCES, APPLICATION OF THESE METHODS INVOLVES THE USE OF COMPUTER SOFTWARE.

SINCE ABOUT 1960, THE METHODS OF DEMOGRAPHIC ANALYSIS HAVE EXPANDED TO INCLUDE METHODS FOR INDIRECT ESTIMATION OF DEMOGRAPHIC QUANTITIES, AND THE USE OF MODERN STATISTICAL METHODS.

IN THIS PRESENTATION, WE SHALL USE THE TERM “BASIC” TO REFER TO THE CLASSICAL OR TRADITIONAL METHODS OF DEMOGRAPHY, THAT MAY BE IMPLEMENTED USING JUST BASIC ALGEBRA AND SIMPLE MATRIX ARITHMETIC, AND “ADVANCED” TO REFER TO THE MODERN METHODS, WHICH REQUIRE CALCULUS, GENERAL MATRIX ALGEBRA, AND INFERENTIAL STATISTICS.

IT IS NOTED THAT THE BASIC MATERIAL COULD EASILY INCLUDE CALCULUS AS A PREREQUISITE. IF THIS WERE DONE, SOME OF THE BASIC DEMOGRAPHIC

CONCEPTS, SUCH AS EXPONENTIAL GROWTH AND THE PROBABILITIES AND EXPECTATIONS ASSOCIATED WITH THE LIFE TABLE, WOULD BE DEFINED IN TERMS OF INTEGRALS AND DERIVATIVES INSTEAD OF SUMS. TO KEEP PART 1 OF THE PRESENTATION ACCESSIBLE TO A LARGER AUDIENCE, A KNOWLEDGE OF CALCULUS IS NOT ASSUMED FOR THAT PART.

3. USES OF DEMOGRAPHIC DATA AND DEMOGRAPHIC ANALYSIS

USES OF DEMOGRAPHIC DATA

IN MANY FIELDS, VARIABLES OF INTEREST ARE RELATED TO POPULATION AND BASIC POPULATION ATTRIBUTES SUCH AS AGE, SEX, AND GEOGRAPHIC DISTRIBUTION. IF THE RELATIONSHIP OF A VARIABLE TO A BASIC POPULATION CHARACTERISTIC IS KNOWN, THEN IT IS POSSIBLE TO ESTIMATE THE VALUE OF THE VARIABLE CONDITIONAL ON THE POPULATION CHARACTERISTICS, AND THE PRECISION OF SUCH A FORECAST MAY BE SUBSTANTIALLY HIGHER THAN IF THESE CHARACTERISTICS ARE NOT TAKEN INTO ACCOUNT. FURTHERMORE, IF A FORECAST OF THE BASIC POPULATION CHARACTERISTICS IS AVAILABLE, THEN IT IS POSSIBLE TO FORECAST THE VALUE OF THE VARIABLE CONDITIONAL ON THE POPULATION FORECAST. FORECASTS THAT DEPEND HEAVILY ON POPULATION ARE CALLED “POPULATION-BASED FORECASTS.”

EXAMPLES OF THIS TYPE OF APPLICATION INCLUDE:

- ESTIMATION OF HEALTH AND WELFARE CASELOADS AND BUDGETS
- ESTIMATION OF SCHOOL ENROLMENTS
- DEMAND FOR INFRASTRUCTURE, SUCH AS HOUSING, OFFICE SPACE, SCHOOLS, STORES, HOSPITALS, WATER, WATER-TREATMENT PLANTS, SEWAGE-TREATMENT PLANTS, ELECTRICITY-GENERATION PLANTS AND ROADS
- ESTIMATION OF DEMAND FOR AGRICULTURAL COMMODITIES AND PRODUCTS
- ESTIMATION OF DEMAND FOR MANUFACTURED PRODUCTS, SUCH AS FOOD, MEDICINES, HOME APPLIANCES, CLOTHING AND AUTOMOBILES

- ESTIMATION OF DEMAND FOR SERVICES, SUCH AS MEDICAL SERVICES, INSURANCE AND BANKING
- DEMAND FOR NATURAL RESOURCES, SUCH AS WATER, PETROLEUM, NATURAL GAS, COAL, MINERAL ORES, FOREST PRODUCTS, SEAFOOD AND ARABLE LAND
- ESTIMATION OF TAX REVENUES
- ESTIMATION OF POLITICAL TRENDS AND POWER

USES OF DEMOGRAPHIC ANALYSIS

IN ORDER TO PROVIDE THE POPULATION ESTIMATES AND FORECASTS REQUIRED TO SUPPORT POPULATION-BASED FORECASTS, RELIABLE ESTIMATES MUST BE AVAILABLE OF POPULATION, DISAGGREGATED BY LOCATION AND COMPOSITION. SOME OF THE REQUIRED DATA ARE AVAILABLE AS DIRECT ESTIMATES FROM VITAL-STATISTICS REGISTRATION SYSTEMS AND CENSUSES, BUT IN MANY INSTANCES, THE BASIC POPULATION DATA MUST BE ESTIMATED. ESTIMATES ARE REQUIRED NOT ONLY FOR NUMERICAL TOTALS (COUNTS BY AGE, SEX AND LOCATION), BUT OF DYNAMIC QUANTITIES SUCH AS BIRTHS, DEATHS AND MIGRATION, WHICH INVOLVE RATES SUCH AS FERTILITY, MORTALITY AND MIGRATION. ESTIMATES ARE REQUIRED FOR INTERCENSAL TIMES AND POST-CENSAL TIMES. ESTIMATES, INCLUDING HYPOTHETICAL POPULATION PROJECTIONS AND PROBABILISTIC FORECASTS, ARE REQUIRED FOR FUTURE TIMES.

THE ESTIMATION PROCESS INVOLVES A VARIETY OF NUMERICAL AND STATISTICAL METHODS, INCLUDING MATRIX ALGEBRA, NONPARAMETRIC STATISTICAL METHODS, PARAMETRIC STATISTICAL METHODS, MULTIVARIATE ANALYSIS AND STATISTICAL FORECASTING. MOST OF THESE METHODS ARE IMPLEMENTED USING COMPUTER SOFTWARE PROGRAM PACKAGES. THESE METHODS AND SOFTWARE ARE REFERRED TO AS THE “TOOLS” OF DEMOGRAPHIC ANALYSIS.

ONCE RELIABLE ESTIMATES ARE AVAILABLE FOR POPULATION TOTALS, COMPOSITION AND DISTRIBUTION (FOR PAST, PRESENT AND FUTURE TIMES), POPULATION-BASED ESTIMATES CAN BE CONSTRUCTED THAT ARE CONDITIONAL ON THESE ESTIMATES. THE CONSTRUCTION OF POPULATION-BASED ESTIMATES INVOLVES A RANGE OF ANALYTICAL TECHNIQUES, INCLUDING INTERPOLATION, EXTRAPOLATION, STANDARDIZED RATES, SYNTHETIC ESTIMATION, SMALL-AREA ESTIMATION, REGRESSION, MULTIVARIATE ANALYSIS AND FORECASTING.

4. DEMOGRAPHIC DATA

DEFINITION OF DEMOGRAPHIC DATA

THE TERM "DEMOGRAPHIC DATA," IN A NARROW SENSE, REFERS TO DATA RELATING TO THE THREE BASIC COMPONENTS OF POPULATION CHANGE, BIRTHS, DEATHS, AND MIGRATION. IN A WIDER SENSE, DEMOGRAPHIC DATA INCLUDES DATA ON VARIABLES THAT HAVE A SIGNIFICANT EFFECT ON, OR A SIGNIFICANT ASSOCIATION WITH, THE THREE BASIC COMPONENTS, SUCH AS AGE, RACE, NUPTIALITY, EDUCATION AND EMPLOYMENT.

THE MAJOR SOURCES OF BASIC DEMOGRAPHIC DATA ARE REGISTRATION RECORDS ("VITAL STATISTICS DATA"), CENSUSES AND SURVEYS. WITHIN NATIONS, THESE DATA ARE ASSEMBLED, PROCESSED AND DOCUMENTED BY NATIONAL STATISTICAL AGENCIES, SUCH AS NATIONAL CENSUS BUREAUS. AT THE INTERNATIONAL LEVEL, DATA ARE ASSEMBLED, COMPILED, AND DISTRIBUTED BY INTERNATIONAL AGENCIES SUCH AS THE UNITED NATIONS, THE WORLD BANK, AND OTHER INTERNATIONAL AND NATIONAL ORGANIZATIONS (SUCH AS THE ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, THE U.S. CENSUS BUREAU, AND THE U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT). THE DATA ARE AVAILABLE IN HARDCOPY IN PUBLICATIONS SUCH AS THE *U.N. DEMOGRAPHIC YEARBOOK*, AND IN ELECTRONIC FORM FROM AGENCY INTERNET WEBSITES. THESE DATA INCLUDE DATA FOR VARIOUS CENSUSES AND SURVEYS, AS WELL AS DERIVED DATA SUCH AS LIFE TABLES AND POPULATION PROJECTIONS.

FOR DEMOGRAPHIC APPLICATIONS, SUCH AS FORECASTING THE VALUES OF VARIABLES ASSOCIATED WITH POPULATION, INTEREST FOCUSES ON ANY VARIABLES THAT MAY BE ASSOCIATED WITH POPULATION LEVELS OR CHANGES. SUCH DATA ARE AVAILABLE FROM A WIDE VARIETY OF SOURCES, SUCH AS GOVERNMENT AGENCIES (CENTRAL BANKS, DEPARTMENTS / MINISTRIES OF

STATISTICS, FINANCE, EDUCATION, HEALTH, AGRICULTURE, HOUSING, TRANSPORTATION AND OTHERS).

SOURCES OF BASIC DEMOGRAPHIC DATA

HERE FOLLOWS A SUMMARY OF BASIC DEMOGRAPHIC DATA FROM MAJOR SOURCES.

UNITED NATIONS

EACH YEAR, THE U.N. PUBLISHES *World Population Prospects*, WHICH PRESENTS A DESCRIPTION OF CURRENT WORLD POPULATION AND PROPULATION PROJECTIONS. THE PUBLICATION IS AVAILABLE AT https://population.un.org/wpp/ . THE AMOUNT OF DETAIL PUBLISHED VARIES EACH YEAR. FOR 2019, *World Population Prospects Highlights* IS AVAILABLE, DESCRIBING THE GLOBAL AND REGIONAL SITUATION. IN SOME YEARS, A LARGE AMOUNT OF COUNTRY-BY-COUNTRY DEMOGRAPHIC DATA ARE AVAILABLE. THE MOST RECENT YEAR IN WHICH A SUBSTANTIAL REVISION AND AMOUNT OF DETAIL WERE PRESENTED WAS 2017: *World Population Prospects, the 2017 Revision*. FOR THE COMPREHENSIVE REVISION OF 2017, THE DOCUMENTS ARE QUITE LARGE, E.G., *VOLUME I, COMPREHENSIVE TABLES*, IS 377 PAGES LONG, AND *VOLUME II, DEMOGRAPHIC PROFILES*, IS 883 PAGES LONG.

THE U.N. PUBLISHES ANNUAL *DEMOGRAPHIC YEARBOOKS*, WHICH MAY BE ACCESS AT WEBSITE https://unstats.un.org/unsd/demographic-social/products/dyb/ .

THE UNITED NATIONS MAINTAINS A GLOBAL DATABASE, AT WEBSITE http://data.un.org/ . THE DATA SETS ARE ORGANIZED BY TOPIC INTO A NUMBER OF "DATAMARTS," ONE OF WHICH, http://data.un.org/ CONTAINS DEMOGRAPHIC DATA. THE DATA MAY BE FILTERED BY COUNTRY AND YEAR, AND DOWNLOADED IN A VARIETY OF FILE FORMATS.

WORLD BANK

THE WORLD BANK MAINTAINS A WEBSITE, *HEALTH, NUTRITION AND DATA PORTAL* (http://datatopics.worldbank.org/health/population) THAT PROVIDES

POPULATION AND OTHER DEMOGRAPHIC ESTIMATES AND PROJECTIONS FROM 1960 TO 2050. THEY ARE DISAGGREGATED BY AGE-GROUP AND SEX AND COVER MORE THAN 200 ECONOMIES. A VERY POWERFUL AND INFORMATIVE GRAPHICAL USER INTERFACE ALLOWS THE USER TO CONSTRUCT A WIDE VARIETY OF GRAPHICS. THE USER INTERFACE IS CALLED *Population Dashboard*. IT IS COMPRISED OF THREE PARTS, CALLED the *Population Dynamics Dashboard*, the *Population Size and Composition Dashboard*, and the *Fertility and Mortality Dashboard*. THE USER MAY SELECT A COUNTRY ON A WORLD MAP, AND SEE RESULTS FOR THE SELECTED COUNTRY.

THE World Bank Open Data WEBSITE (https://data.worldbank.org/,) AND DataBank (https://databank.worldbank.org/source/population-estimates-and-projections) WEBSITES PROVIDES DOWNLOADABLE POPULATION ESTIMATES, PROJECTIONS AND DEMOGRAPHIC DATA IN SEVERAL FILE FORMATS (comma-separated-value (CSV), Extensible Markup Language (XML), AND Microsoft Excel).

U.S. CENSUS BUREAU

THE U.S. CENSUS PROVIDES ACCESS TO A LARGE AMOUNT OF DEMOGRAPHIC DATA ABOUT THE COUNTRY AND STATES. IT ALSO PROVIDES ACCESS TO A SUBSTANTIAL AMOUNT OF DATA FOR COUNTRIES OF THE WORLD.

HERE FOLLOWS A SUMMARY DESCRIPTION OF WHAT IS AVAILABLE FROM THE U.S. CENSUS BUREAU'S *INTERNATIONAL PROGRAMS / INTERNATIONAL DATABASE* WEB PAGE:

Overview

The International Data Base (IDB) was developed by the U.S. Census Bureau to provide access to accurate and timely demographic measures for populations around the world. The database includes a comprehensive set of indicators, as produced by the U.S. Census Bureau since the 1960s. Through sponsorship from various U.S. Government agencies, the IDB is updated on a regular basis to provide information needed for research, program planning, and policy-making decisions, in the U.S. and globally.

Data included in the IDB consist of indicators developed from censuses, surveys, administrative records, and special measures of HIV/AIDS-related mortality. Through evaluation and adjustment of data from these sources, measures of population, mortality, fertility, and net migration are estimated for current and past years and then used as the basis for projections to 2050.

The IDB provides estimates and projections for 228 countries and areas which have populations of 5,000 or more and as recognized by the U.S. Department of State. Population size (by single year of age and sex) and components of change (fertility, mortality, and migration) are provided from an initial or base year through 2050, for each calendar year. This level of detail provides an important foundation for tracking the demographic impacts of HIV/AIDS and related conditions, as well as events of concern that are affecting populations around the globe.

UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT (USAID) DEMOGRAPHIC AND HEALTH SURVEYS (DHS) PROGRAM

HERE FOLLOWS A SUMMARY OF THE USAID DHS PROGRAM:

Since 1984, The Demographic and Health Surveys (DHS) Program has provided technical assistance to more than 400 surveys in over 90 countries, advancing global understanding of health and population trends in developing countries.

The DHS Program has earned a worldwide reputation for collecting and disseminating accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, and nutrition.

The DHS Program is funded by the U.S. Agency for International Development (USAID). Contributions from other donors, as well as funds from participating countries, also support surveys. The project is implemented by ICF.

Since September 2013, ICF has been partnering with seven internationally experienced organizations to expand access to and use of the DHS data:

Avenir Health
Blue Raster
EnCompass
Johns Hopkins Bloomberg School of Public Health/Center for Communication Programs

PATH
Vysnova

THE DHS PROGRAM WEBSITE IS https://dhsprogram.com/. DHS SURVEY DATA MAY BE DOWNLOADED FROM WEBSITE https://dhsprogram.com/data/available-datasets.cfm ; GEOSPATIAL DATA ARE AVAILABLE AT THE *Spatial Data Repository Modeled Surfaces* WEBSITE, https://spatialdata.dhsprogram.com/modeled-surfaces/.

THE PRECEDING DEMOGRAPHIC DATA SOURCES (U.N., WORLD BANK, U.S. CENSUS BUREAU, USAID) PROVIDE (AT NO COST) A SUBSTANTIAL AMOUNT OF DETAILED DEMOGRAPHIC DATA, AND ARE WIDELY USED. THERE ARE MANY ADDITIONAL SOURCES OF DEMOGRAPHIC DATA. THE CENSUS BUREAUS OF EACH COUNTRY MAINTAIN DEMOGRAPHIC DATA AT THE NATIONAL AND REGIONAL LEVELS. SEVERAL DATABASES ARE IDENTIFIED AT THE WEBSITE OF THE International Union for the Scientific Study of Population, https://iussp.org/en/population-databases.

5. THE MATHEMATICS OF BASIC DEMOGRAPHY

MATHEMATICS USED FOR BASIC DEMOGRAPHY

THE MATHEMATICS OF BASIC DEMOGRAPHY IS BASIC (“HIGH SCHOOL”) ALGEBRA, PLUS A KNOWLEDGE OF THE LOGARITHMIC AND EXPONENTIAL FUNCTIONS (WHICH RELATE TO CONTINUOUS GROWTH RATES), AND A BASIC KNOWLEDGE OF MATRIX ALGEBRA (ADDITION, SUBTRACTION, AND MULTIPLICATION, BUT NOT HIGHER-LEVEL CONCEPTS SUCH AS DETERMINANTS, RANK, VECTOR SPACES, EIGENVALUES, EIGENVECTORS, FACTORIZATION OR INVERSES).

SOME BASIC DEFINITIONS AND CONCEPTS FROM BASIC ALGEBRA (RATES, PROPORTION, AND GROWTH)

RATIO: A QUOTIENT OF TWO NUMBERS, a AND b, DENOTED AS a/b, a to b, OR a:b. EXAMPLE: THE RATIO 1 TO 10, OR .1

IF THE NUMBERS a AND b ARE MEASURED IN THE SAME UNITS, THE RATIO IS A *DIMENSIONLESS NUMBER* (I.E., HAS NO UNITS ASSOCIATED WITH IT). IF THE NUMBERS a AND b ARE MEASURED IN DIFFERENT UNITS, THE RATIO IS CALLED A *RATE*, AND THE UNIT OF MEASUREMENT OF THE RATE IS THE UNIT OF THE NUMBERATOR DIVIDED BY (OR "PER") THE UNIT OF THE DENOMINATOR.

NOTE THAT SOME AUTHORS IN DEMOGRAPHY USE THE TERM "RATE" ONLY WHEN THE DENOMINATOR IS A UNIT OF TIME, SUCH AS A YEAR, AND USE THE TERM "RATIO" IF THE UNITS OF THE NUMERATOR ARE THE SAME, OR ARE DIFFERENT BUT THE DENOMINATOR IS NOT A UNIT OF TIME. MOST AUTHORS DO NOT ADHERE TO THIS CONVENTION. IN PHYSICS (AND IN OTHER FIELDS, SUCH AS ECONOMETRICS), THE DENOMINATOR IN A RATE DEFINITION MAY BE ANY TYPE OF UNIT (OR "DIMENSION"). IF IT IS A TIME UNIT, AND ATTENTION IS CALLED TO THAT FACT, THEN THE RATE MAY BE REFERRED TO AS THE "TIME RATE OF CHANGE OF THE NUMERATOR".

IT IS NOTED THAT EVEN WHEN AUTHORS CLAIM TO USE "RATE" ONLY TO APPLY TO TIME RATES, THEY MAY DEPART FROM THIS USAGE FOR SOME STANDARD RATES, SUCH AS THE CRUDE BIRTH RATE, WHICH HAS UNITS (BIRTHS/PERSON-YEARS) NOT RELATIVE TO TIME."

PROPORTION: IF a IS A PART (PORTION) OF b, THE RATIO IS CALLED A PROPORTION. EXAMPLE: IF 100 PEOPLE IN A GROUP OF 1,000 ARE EMPLOYED (AND THE OTHERS ARE NOT EMPLOYED), THEN THE PROPORTION OF PERSONS EMPLOYED IS 100/1000 = .1 (DIMENSIONLESS).

PERCENTAGE: A PROPORTION MULTIPLIED BY 100, WITH UNITS PERCENTAGE POINTS, OR EXPRESSED AS "PER CENT." IN THE PRECEDING EXAMPLE, THE PERCENTAGE OF PEOPLE UNEMPLOYED IS 10 PERCENT.

EXAMPLE: IF A GROUP OF 50 PERSONS CONTAINS 20 FEMALES AND 30 MALES, THEN THE RATIO OF FEMALES TO MALES IS 20 TO 30, OR 20/30, OR 2/3, THE PROPORTION OF FEMALES IS 20/50 = .4, AND THE PERCENTAGE OF FEMALES IS .4 x 100 = 40 PERCENT.

EXAMPLE: IF A POPULATION OF 1,000 PERSONS GROWS IN SIZE TO 1,100 IN FIVE YEARS, THEN THE FIVE-YEAR RATE OF GROWTH IS 100/1,000 = .1 PERSONS PER FIVE YEARS, OR TEN PERCENT INCREASE IN POPULATION PER FIVE YEARS.

GROWTH RATES OF POPULATIONS

LINEAR GROWTH

IF A POPULATION GROWS BY A FIXED AMOUNT PER UNIT OF TIME, THEN THE GROWTH IS SAID TO BE *LINEAR*. IN THE EXAMPLES THAT FOLLOW, WE SHALL ASSUME THAT TIME IS MEASURED IN YEARS.

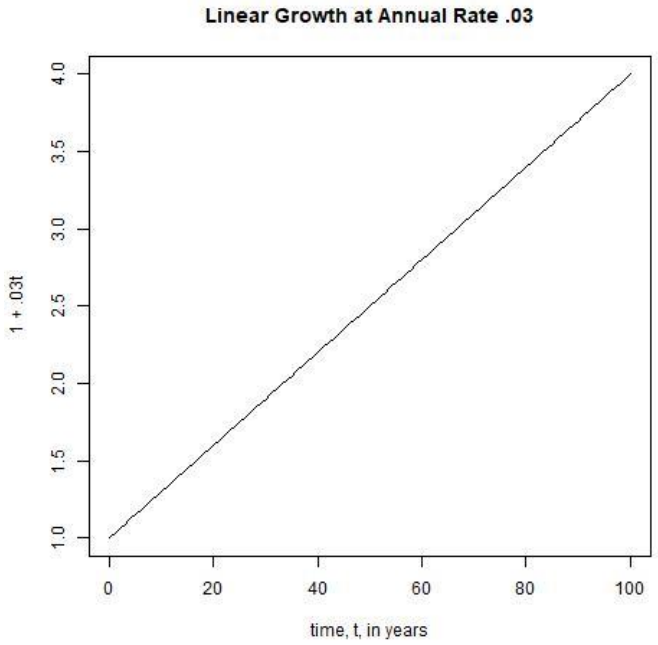
EXAMPLE: SUPPOSE THAT A SCHOOL GRADUATES TEN PEOPLE PER YEAR. THEN THE GROWTH RATE OF THE TOTAL NUMBER OF GRADUATES IS TEN PERSONS PER YEAR. (THIS IS ANALOGOUS TO THE YIELD ON A BOND THAT PAYS A PREMIUM OF, SAY, 5 PERCENT PER YEAR. THE SAME AMOUNT IS EARNED EACH YEAR. THIS IS CALLED *SIMPLE INTEREST*.)

THE FORMULA FOR LINEAR GROWTH IS

$$P_t = P_0(1 + at)$$

WHERE t DENOTES TIME (IN YEARS), P_0 DENOTES THE POPULATION SIZE AT TIME $t = 0$, P_t DENOTES THE POPULATION SIZE AT TIME t, AND a DENOTES THE (LINEAR) RATE OF GROWTH.

FIGURE 1 SHOWS A PLOT OF LINEAR GROWTH AT ANNUAL RATE $a = .03$.



GEOMETRIC GROWTH (OR COMPOUND GROWTH)

IF A POPULATION GROWS, EACH YEAR, BY A FIXED PROPORTION, a, OF ITS SIZE AT THE BEGINNING OF THE YEAR, THEN THE GROWTH IS SAID TO BE *GEOMETRIC*, OR *COMPOUND*, WITH COMPOUNDING PERIOD ONE YEAR.

SUPPOSE THAT THE POPULATION SIZE AT TIME $t = 0$ IS P_0, AND THAT THE GEOMETRIC RATE OF GROWTH IS a. THEN THE POPULATION SIZE AT TIME $t = 1$ IS

$$P_1 = P_0(1 + a).$$

THIS IS THE SAME AMOUNT, AT THE END OF ONE YEAR, AS FOR LINEAR GROWTH. AT THE END OF TWO YEARS, THE POPULATION SIZE IS

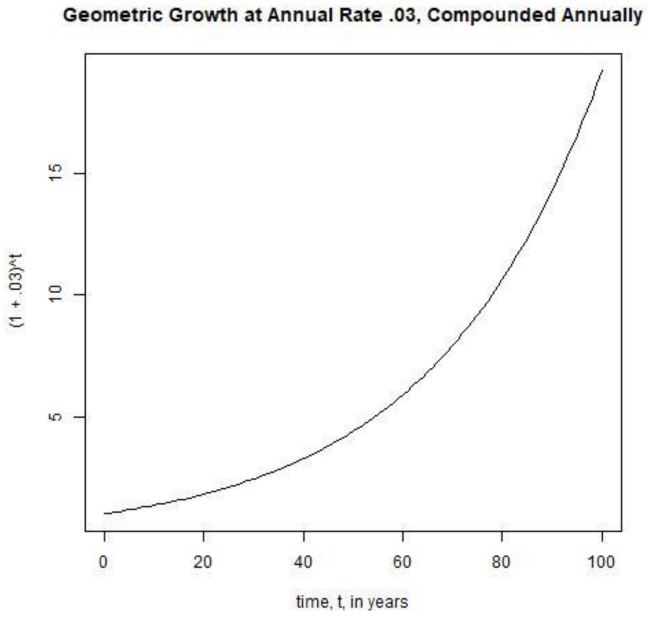
$$P_2 = P_1(1 + a) = P_0 (1 + a) (1 + a) = P_0 (1 + a)^2,$$

AND, BY MATHEMATICAL INDUCTION, THE POPULATION SIZE AT THE END OF t YEARS IS

$$P_t = P_0 (1 + a)^t.$$

THE FORMULA FOR GEOMETRIC GROWTH IS ANALOGOUS TO GROWTH OF AN INTEREST-BEARING ACCOUNT FOR WHICH INTEREST IS COMPOUNDED ANNUALLY.

FIGURE 2 SHOWS A PLOT OF GEOMETRIC GROWTH AT ANNUAL RATE a = .03.



GEOMETRIC, OR COMPOUND, GROWTH, WITH ARBITRARY COMPOUNDING FREQUENCY (OR PERIODIC COMPOUNDING)

IN THE PRECEDING CASE, THE GROWTH (OR INTEREST) WAS COMPOUNDED ANNUALLY. THE FOLLOWING FORMULA SHOWS THE GROWTH IF COMPOUNDING IS DONE AT A FREQUENCY, n, PER YEAR.

LET a/n DENOTE THE SIMPLE GROWTH (INTEREST) RATE TO BE APPLIED AT THE END OF EACH OF THE f PERIODS IN THE YEAR (n IS CALLED THE *COMPOUNDING FREQUENCY*). THEN THE *SIMPLE ANNUAL GROWTH (INTEREST) RATE*, OR THE *NOMINAL ANNUAL GROWTH (INTEREST) RATE, OR "ANNUALIZED" RATE*, IS DEFINED AS THIS AMOUNT TIMES n, OR (a/n)n = a. THE FORMULA FOR THE GROWTH IS:

$$P_t = P_0 (1 + a/n)^{nt}$$

WHERE t, P_t AND P_0 ARE DEFINED AS BEFORE.

EXPONENTIAL GROWTH

IF THE FREQUENCY OF COMPOUNDING IS INCREASED WITHOUT LIMIT, SO THAT THE COMPOUNDING PERIOD SHRINKS TO ZERO LENGTH, THEN THE SITUATION IS REFERRED TO AS *CONTINOUS COMPOUNDING*. IN THIS CASE, THE QUANTITY $(1 + a/n)^n$ IN THE ABOVE FORMULA CONVERGES TO THE NATURAL EXPONENTIAL FUNCTION:

$$\exp(a) = e^a = \lim_{n=\infty}\left(1 + \frac{a}{n}\right)^n ,$$

WHERE e IS THE BASE OF NATURAL LOGARITHMS, AN IRRATIONAL NUMBER APPROXIMATELY EQUAL TO 2.71828.

THE INVERSE FUNCTION OF THE EXPONENTIAL FUNCTION IS THE NATURAL LOGARITHMIC FUNCTION:

$$log_e(e^a) = ln(e^a) = a.$$

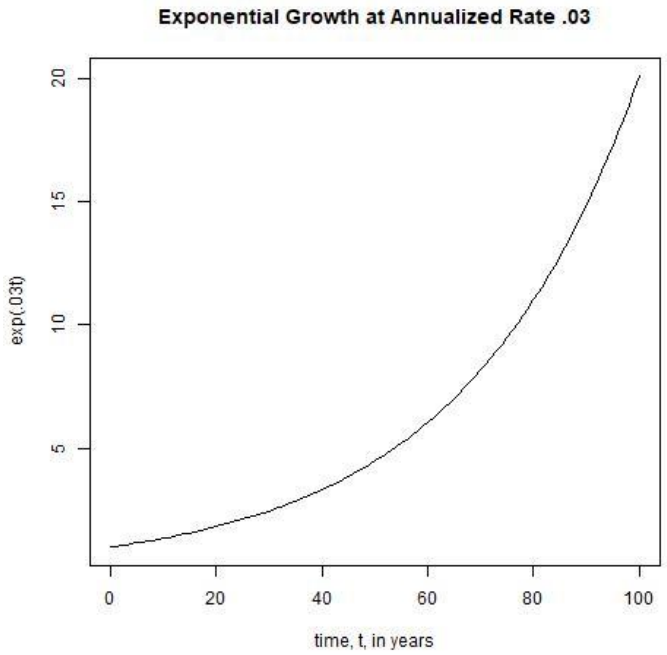
WITH CONTINUOUS COMPOUNDING AT A NOMINAL ANNUAL INTEREST RATE OF a (I.E., AN ANNUALIZED RATE OF a), THE FORMULA FOR THE POPULATION SIZE AT TIME t (STARTING AT A POPULATION SIZE P_0 AT TIME t = 0) IS

$P_t = P_0 e^{at}$.

(THE POPULATION SIZE AT TIME t_2 RELATIVE TO THE SIZE AT t_1 IS

$P_{t2} = P_{t1} e^{a(t2 – t1)}$.)

FIGURE 3 SHOWS A PLOT OF EXPONENTIAL GROWTH AT ANNUALIZED RATE a = .03.



THE (NATURAL, BASE e) LOGARITHM OF THIS QUANTITY IS

$$\log_e(P_t) = \log_e(P_0) + \log_e(e^{at}) = \log_e(P_0) + at.$$

THAT IS, IF THE GROWTH OF THE POPULATION IS EXPONENTIAL (AT ANNUALIZED RATE a), THEN THE GROWTH OF THE LOGARITHM IS LINEAR (AT ANNUAL RATE a).

IT IS NOT NECESSARY TO USE THE BASE e FOR LOGARITHMS IN THE PRECEDING EXPRESSIONS, BUT IT SIMPLIFIES THINGS. FOR EXAMPLE, SUPPOSE THAT WE USED LOGARITHMS TO THE BASE 10 (I.E., "COMMON" LOGARITHMS, INSTEAD OF NATURAL (BASE e) LOGARITHMS), THE LOGARITHM OF THE POPULATION SIZE AT TIME t IS

$$\log_{10}(P_t) = \log_{10}(e^{at}) = at \log_{10}(e).$$

THAT IS, THE LOGARITHM OF THE POPULATION SIZE AT TIME t IS at TIMES A CONSTANT, $\log_{10}(e)$. IF WE USE AN ARBITRARY BASE, b, FOR THE LOGARITHMS, THEN THIS CONSTANT IS $\log_b(e)$. ONLY FOR THE BASE b = e IS THIS CONSTANT EQUAL TO ONE.

THE REASON FOR USING THE NATURAL EXPONENTIAL FUNCTION IN THE EXPRESSION FOR POPULATION GROWTH, AND HENCE THE BASE e FOR THE LOGARITHMS, IS SIMPLICITY. THE SIMPLIFICATION IS VERY SIGNIFICANT. USING BASE e, THE COEFFICIENT OF t IN THE LOGARITHMIC EQUATION IS THE GROWTH RATE, a, WHICH IS AN ESSENTIAL FEATURE OF THE GROWTH PROCESS.

FOR POSITIVE VALUES OF r, THE EXPONENTIAL FUNCTION OF RATE r (e^{rt}) IS MONOTONICALLY INCREASING. FOR NEGATIVE VALUES OF r, IT IS MONOTONICALLY DECREASING. FOR $r = 0$ IT IS A STRAIGHT HORIZONTAL LINE PASSING THROUGH ORDINATE VALUE 1.

FOR POSITIVE r, THE EXPONENTIAL FUNCTION OF RATE r IS A CONCAVE FUNCTION. THIS FACT HAS TWO IMPORTANT RESULTS. FIRST, IF THE CURVE IS EXTRAPOLATED FROM A TIME t_1 USING THE SIMPLE GROWTH RATE BETWEEN TIMES t_0 TO t_1, WHERE $t_0 < t_1$, THE EXTRAPOLATION WILL ALWAYS FALL BELOW THE ACTUAL CURVE. THAT IS, LINEAR GROWTH AND COMPOUND GROWTH (COMPOUNDED A FINITE NUMBER OF TIMES, NOT CONTINUOSLY) ARE ALWAYS LESS IN THE FUTURE THAN EXPONENTIAL GROWTH AT THE SAME NOMINAL RATE.

SECOND, IF A POPULATION IS COMPRISED OF SEVERAL SUBPOPULATIONS OF VARIOUS POSITIVE GROWTH RATES (I.E., THE POPULATION IS HETEROGENEOUS WITH RESPECT TO THE GROWTH RATES OF ITS COMPONENT SUBPOPULATIONS), THE GROWTH OF A (HYPOTHETICAL) POPULATION HAVING THE AVERAGE GROWTH RATE WILL ALWAYS BE LESS THAN THE GROWTH OF THE ACTUAL POPULATION. (THIS IS A RESULT KNOWN AS JENSEN'S INEQUALITY.)

IF A POPULATION SIZE AT TIME t, P_t, IS GIVEN AT TWO SUCCESSIVE TIMES, t_1 AND t_2, THE LINEAR GROWTH RATE IS $a = (P_{t2} – P_{t1})/(t_2 – t_1)$. THE EXPONENTIAL GROWTH RATE IS

$$r = \ln(P_{t2}/ P_{t1})/(t_2 – t_1).$$

THIS IS EASY TO SEE. BY THE DEFINITION OF EXPONENTIAL GROWTH, WE HAVE:

$$P_{t2} = P_{t1}\, e^{r(t2 – t1)}$$

OR

$$e^{r(t2 - t1)} = P_{t2}/ P_{t1}.$$

TAKING LOGARITHMS OF BOTH SIDES YIELDS

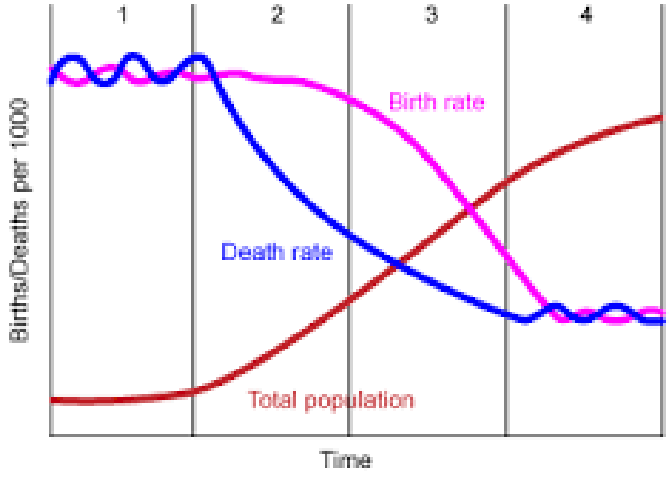
$$r(t_2 - t_1) = \ln(P_{t2}/ P_{t1})$$

OR

$$r = \ln(P_{t2}/ P_{t1})/(t_2 - t_1).$$

AN INTERESTING FEATURE OF THE EXPONENTIAL FUNCTION e^{rt} IS THAT THE SLOPE OF THE FUNCTION INCREASES AS t INCREASES. THE SLOPE IS THE CHANGE IN POPULATION DIVIDED BY THE CHANGE IN TIME, I.E., THE LINEAR RATE OF GROWTH. SO, FOR AN EXPONENTIAL PROCESS, THE LINEAR RATE OF GROWTH BECOMES LARGER AND LARGER AS TIME PASSES (EVEN THOUGH THE EXPONENTIAL RATE OF GROWTH REMAINS CONSTANT).

EXPONENTIAL GROWTH IS “EXPLOSIVE.” IT CANNOT CONTINUE FOR VERY LONG AT A POSITIVE RATE WITHOUT REACHING EXTREMELY LARGE VALUES. FOR THIS REASON, REASONABLE MODELS OF LONG-TERM POPULATION GROWTH MUST ALWAYS ALLOW FOR THE GROWTH RATE TO DROP TO ZERO BEFORE VERY LONG. THIS FACT IS A SALIENT FEATURE OF THE “DEMOGRAPHIC TRANSITION” MODEL OF WORLD POPULATION GROWTH, SHOWN IN FIGURE 4.



THE FIGURE SHOWS THAT WORLD POPULATION REMAINED AT VERY LOW LEVEL FOR A VERY LONG TIME, WITH BIRTHS BALANCING DEATHS. A FEW CENTURIES AGO, MORTALITY DECLINED, FERTILITY REMAINED HIGH, AND POPULATION BEGAN TO INCREASE. BECAUSE OF EARTH'S FINITE SPACE AND RESOURCES, THIS INCREASE – THE SO-CALLED "POPULATION EXPLOSION" – CANNOT CONTINUE FOR VERY LONG (EXPLOSIONS DO NOT LAST FOR A VERY LONG TIME!). AT SOME POINT, BIRTH AND DEATH RATES MUST RETURN TO BE IN BALANCE.

VECTORS AND MATRICES

WHILE MUCH OF BASIC DEMOGRAPHY CAN BE DESCRIBED IN TERMS OF SIMPLE ARITHMETIC AND BASIC ALGEBRA, EVEN THE BASIC FORMULAS ARE A LITTLE COMPLICATED, INVOLVING SUMS OF PRODUCTS. THE COMPUTATIONS INVOLVED ARE OF A KIND THAT CAN BE REPRESENTED VERY COMPACTLY USING THE NOTATION OF VECTORS AND MATRICES, AND SIMPLE VECTOR / MATRIX ARITHMETIC. USE OF THIS NOTATION MAKES THE PRESENTATION MUCH SIMPLER, ENHANCES UNDERSTANDING, AND FACILITATES THE USE OF COMPUTER SOFTWARE FOR DEMOGRAPHIC ANALYSIS. BOOKS ON MATHEMATICAL DEMOGRAPHY ALWAYS USE MATRIX NOTATION, BUT BOOKS ON GENERAL DEMOGRAPHY DO NOT.

BECAUSE OF THE SUBSTANTIAL ADVANTAGES OF MATRIX NOTATION, IT WILL BE USED IN THIS PRESENTATION.

IN THIS PRESENTATION, THE MATERIAL ON MATRIX NOTATION WILL BE PRESENTED IN TWO PARTS, CORRESPONDING TO THE TWO PARTS OF THE PRESENTATION (BASIC AND ADVANCED). THE FIRST PART, PRESENTED HERE, WILL DEFINE VECTORS AND MATRICES AND BASIC ARITHMETIC COMPUTATIONS INVOLVING VECTORS AND MATRICES.

A *COLUMN VECTOR* IS A VERTICAL ARRAY OF A SEQUENCE OF n ELEMENTS

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

A *ROW VECTOR* IS A HORIZONTAL ARRAY OF A SEQUENCE OF n ELEMENTS

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n).$$

IN THIS APPLICATION, THE ELEMENTS ARE SYMBOLS OR VARIABLES OR NUMBERS.

VECTORS ARE INDICATED BY BOLDFACE OR UNDERLINED FONT. A VECTOR CONSISTING OF ONE ELEMENT IS CALLED A *SCALAR*. FOR EXAMPLE, x MAY DENOTE A SCALAR AND **x** AND $\underline{x}$ MAY DENOTE VECTORS.

THE ELEMENT x_i IS CALLED THE i-th COMPONENT OF **x**. THE NUMBER OF COMPONENTS IN **x** IS VARIOUSLY CALLED THE DIMENSION OR SIZE OR LENGTH OF **x**. (THE TERMS “LENGTH” AND “SIZE” HAVE DIFFERENT MEANINGS IN OTHER CONTEXTS, TO BE DEFINED LATER.)

THE *TRANSPOSE* OF A COLUMN VECTOR **x**, DENOTED BY **x**’ or $\mathbf{x}^T$ IS THE ROW VECTOR OF LENGTH n, $\boldsymbol{x}' = \boldsymbol{x}^T = (x_1, x_2, \dots, x_n)$ OR $\boldsymbol{x} = (x_1, x_2, \dots, x_n)' = (x_1, x_2, \dots, x_n)^T$. THE TRANSPOSE OF A ROW VECTOR IS DEFINED SIMILARLY.

THE TERM “VECTOR” MAY REFER TO EITHER A COLUMN VECTOR OR A ROW VECTOR. ABSENT AN EXPLICIT INDICATOR (PRIME OR “T”), AN ARBITRARY VECTOR IS ASSUMED TO BE A COLUMN VECTOR.

A *MATRIX* **X** OF m ROWS AND n COLUMNS (AN “m by n” MATRIX) IS A RECTANGULAR ARRAY OF ELEMENTS:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$$

WE SHALL DENOTE MATRICES IN BOLDFACE FONT. (THIS IS NOT A UNIVERSAL CONVENTION – MOST AUTHORS WRITE VECTORS IN BOLDFACE, BUT SOME AUTHORS WRITE GENERAL MATRICES IN STANDARD FONT.)

IF x_{ij} DENOTES THE ELEMENT IN ROW i AND COLUMN j OF MATRIX **X**, THEN THE MATRIX **X** MAY BE DENOTED AS **X** = $[x_{ij}]$. IF WE WISH TO MAKE THE NUMBER OF ROWS AND COLUMNS EXPLICIT, WE WRITE **X** = $[x_{ij}]_{mxn}$. THE TRANSPOSE **X**’ (OR $\mathbf{X}^T$) IS DEFINED AS THE MATRIX HAVING ELEMENT x_{ji} IN ROW i AND COLUMN j.

(NOTATION AMBIGUITY WARNING: SOMETIMES, THE SUPERSCRIPT T WILL REFER TO EXPONENTIATION, NOT TO TRANSPOSITION.) THE ROWS AND THE COLUMNS OF A MATRIX ARE VECTORS. A MATRIX HAVING JUST ONE ROW OR ONE COLUMN IS A VECTOR (OR, IF JUST ONE ROW AND ONE COLUMN, A SCALAR).

IF A MATRIX HAS THE SAME NUMBER OF ROWS AS COLUMNS, IT IS CALLED *SQUARE*, AND THE NUMBER OF ROWS (OR COLUMNS) IS CALLED THE *SIZE* OR *ORDER* OF THE MATRIX.

BASIC OPERATIONS ON MATRICES ARE THE FOLLOWING. SUPPOSE THAT **A** = $[a_{ij}]$ IS AN m x n MATRIX AND **B** = $[b_{ij}]$ IS A p x q MATRIX.

ADDITION: IF m=p AND n=q, THEN **A** + **B** = $[a_{ij} + b_{ij}]_{mxn}$.
SUBTRACTION: IF m=p AND n=q, THEN **A** - **B** = $[a_{ij} - b_{ij}]_{mxn}$.
SCALAR MULTIPLICATION: IF c IS A SCALAR, THEN c**A** = $[ca_{ij}]$.
MULTIPLICATION: **AB** = $[\sum_{k=1}^{n} a_{ik}b_{kj}]$ where n=p.

THE PRODUCT OF AN n BY m MATRIX **A** AND AN m BY k MATRIX **B** IS THE n BY k MATRIX WHOSE i,j-th ELEMENT (I.E., ENTRY IN ROW i AND COLUMN j) IS THE VECTOR PRODUCT OF THE i-th ROW OF **A** AND THE j-th COLUMN OF **B**. NOTE THAT THE PRODUCT IS DEFINED ONLY IF THE MATRICES ARE *CONFORMABLE*, I.E., THE NUMBER OF COLUMNS OF **A** IS EQUAL TO THE NUMBER OF ROWS OF **B**.

SINCE VECTORS ARE MATRICES, THE PRECEDING DEFINITIONS APPLY TO VECTORS. FOR EXAMPLE, THE PRODUCT OF A SCALAR a AND A VECTOR **x** WHOSE i-th COMPONENT IS x_i IS THE VECTOR WHOSE i-th COMPONENT is ax_i: a **x'** = (ax_1, ax_2,…,ax_n).

IF VECTORS **a** AND **b** ARE OF THE SAME LENGTH n, THE *VECTOR PRODUCT* (OR INNER PRODUCT) IS $\boldsymbol{a}'\boldsymbol{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n$. TWO VECTORS WHOSE INNER PRODUCT IS ZERO ARE SAID TO BE ORTHOGONAL. (THIS DEFINITION REFERS TO GEOMETRIC ORTHOGONALITY; IN STATISTICS, TWO RANDOM VECTORS ARE SAID TO BE ORTHOGONAL IF THEY ARE UNCORRELATED.)

THE *DIAGONAL* (OR MAIN DIAGONAL OR PRINCIPAL DIAGONAL) OF A MATRIX IS THE VECTOR OF ELEMENTS FOR WHICH THE ROW INDEX EQUALS THE COLUMN INDEX.

THE SQUARE MATRIX, **I**, SUCH THAT ALL OF THE DIAGONAL ELEMENTS EQUAL TO ONE AND ZEROS ELSEWHERE IS CALLED THE *IDENTITY* MATRIX, SINCE **IA** = **A** FOR ANY MATRIX **A**. (IF **I** IS AN n x n MATRIX, IT IS DENOTED AS $\mathbf{I}_n$ IF IT IS DESIRED TO INDICATE ITS SIZE.)

6. COMPUTER SOFTWARE FOR BASIC DEMOGRAPHY

SOFTWARE CATEGORIES: GENERAL AND DEMOGRAPHY-SPECIFIC

A WIDE VARIETY OF COMPUTER SOFTWARE IS AVAILABLE TO PERFORM DATA PROCESSING (COLLECTION, ASSEMBLY, RETRIEVAL, ANALYSIS AND DISPLAY) RELATED TO DEMOGRAPY. THERE ARE LITERALLY THOUSANDS OF PROGRAMS AVAILABLE THAT MIGHT BE USED, AND THIS PRESENTATION WILL NOT ATTEMPT TO SUMMARIZE THE FIELD. WE SHALL DISCUSS DEMOGRAPHIC SOFTWARE IN SEVERAL MAJOR FUNCTIONAL CATEGORIES:

- GENERAL-PURPOSE SOFTWARE (NOT SPECIFICALLY ORIENTED TO DEMOGRAPHY)
  - STANDARD “OFFICE” SOFTWARE
  - DATA ENTRY, EDITING, (BASIC) PROCESSING AND DISTRIBUTION (CENSUS AND SURVEY DATA ENTRY AND STORAGE)
  - DATABASE MANAGEMENT SYSTEM
  - GEOGRAPHIC INFORMATION SYSTEM
  - MATRIX OPERATIONS
  - STATISTICAL ANALYSIS (ESTIMATION, HYPOTHESIS TESTING AND FORECASTING)

- SOFTWARE SPECIFICALLY ORIENTED TO DEMOGRAPHIC ANALYSIS
  - DISPLAY OF POPULATION FEATURES
  - POPULATION PROJECTIONS
  - POPULATION-BASED FORECASTS
  - DIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS
  - INDIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS
  - FORECASTING DEMOGRAPHIC PARAMETERS

IN EACH MAJOR CATEGORY, ONE OR MORE SOFTWARE SOURCES (PACKAGES, PRODUCTS) WILL BE IDENTIFIED. MAJOR SOFTWARE PACKAGES WILL BE IDENTIFIED, BUT PRESENTATION OF EXAMPLES WILL BE RESTRICTED TO SOFTWARE PROGRAMS THAT ARE AVAILABLE FREE FROM THE INTERNET. MUCH FREE USEFUL SOFTWARE IS AVAILABLE IN R LIBRARIES, BUT THERE IS ALSO FREE USEFUL SOFTWARE FROM OTHER SOURCES.

ATTENTION WILL CENTER ON SOFTWARE THAT GENERATES “END-USER” OUTPUT, RATHER THAN ON SOFTWARE THAT PERFORMS INTERMEDIATE STEPS OF COMPLEX NUMERICAL OR STATISTICAL PROCEDURES.

THE DISCUSSION AT THIS POINT IN THE PRESENTATION IS SIMPLY TO IDENTIFY A NUMBER OF PROGRAM PACKAGES THAT ARE USEFUL FOR DEMOGRAPHY, AND TO IDENTIFY THOSE THAT WILL BE USED TO PRODUCE EXAMPLES IN THE PRESENTATION.

GENERAL-PURPOSE SOFTWARE (NOT SPECIFICALLY ORIENTED TO DEMOGRAPHY)

STANDARD “OFFICE” SOFTWARE

STANDARD “OFFICE” SOFTWARE INCLUDES PROGRAMS FOR TEXT EDITING, WORD PROCESSING, ELECTRONIC SPREADSHEET AND PRESENTATION GRAPHICS. ALL OF THESE FUNCTIONS ARE INCLUDED IN THE BASIC MICROSOFT *OFFICE* SUITE OF PRODUCTS, WHICH INCLUDES Word (WORD PROCESSOR), Excel (ELECTRONIC SPREADSHEET, GRAPHING) AND PowerPoint (PRESENTATION PROGRAM), OR IN THE Corel WordPerfect Office SUITE OF PRODUCTS, WordPerfect (WORD PROCESSING), Quattro Pro (ELECTRONIC SPREADSHEET AND GRAPHING) AND Presentations (PRESENTATION PROGRAM). OPEN-SOURCE OFFICE SUITES ARE AVAILABLE (Apache OpenOffice, LibreOffice). (Note: Both Apache OpenOffice and The Document Foundation’s LibreOffice are descendants of OpenOffice. LibreOffice is the most actively developed free and open-source office suite, with approximately 50 times the development activity of Apache OpenOffice, the other major descendant of OpenOffice (earlier StarOffice).)

OFFICE SUITE SOFTWARE IS NECESSARY TO VIEW DOCUMENTATION (WORD PROCESSOR), TO ASSEMBLE DATA (TEXT EDITOR AND ELECTRONIC SPREADSHEET) AND TO TRANSFER DATA BETWEEN APPLICATIONS (ELECTRONIC SPREADSHEET).

DATA ENTRY, STORAGE, EDITING, (BASIC) PROCESSING AND DISTRIBUTION (FOR CENSUS AND SURVEY DATA)

THE MAJOR PROGRAMS FOR DATA ENTRY ARE CSPro, AVAILABLE FREE FROM THE U.S. CENSUS BUREAU, AND Epi Info, AVAILABLE FREE FROM THE U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION. CSPro AND Epi Info ARE DESIGNED TO COLLECT AND EDIT CENSUS AND SURVEY DATA AND RUN A NUMBER OF BASIC STATISTICAL PROCEDURES. THEY ARE PRIMARILY DATA MANAGEMENT TOOLS, NOT GENERAL-PURPOSE STATISTICAL ANALYSIS PROGRAM PACKAGES.

CSPro, SHORT FOR THE *CENSUS AND SURVEY PROCESSING* SYSTEM, IS USED WORLDWIDE BY STATISTICAL AGENCIES, INTERNATIONAL ORGANIZATIONS, NONGOVERNMENTAL ORGANIZATIONS (NGOs), CONSULTING FIRMS, COLLEGES AND UNIVERSITIES, HOSPITALS, AND PRIVATE SECTOR GROUPS, IN MORE THAN 160 COUNTRIES. MAJOR INTERNATIONAL HOUSEHOLD SURVEY PROGRAMS, SUCH AS *MULTIPLE INDICATOR CLUSTER SURVEYS* (MICS) AND DEMOGRAPHIC AND HEALTH SURVEYS (DHS) ALSO USE CSPro. THE MAIN PURPOSE OF THIS SOFTWARE FRAMEWORK IS DATA ENTRY, EDITING, TABULATION AND DISSEMINATION. VERSIONS ARE AVAILABLE FOR USE ON SMARTPHONES AND TABLETS THAT USE THE ANDROID OPERATING SYSTEM. COMPLEX STATISTICAL ANALYSIS, SUCH AS ANALYSIS OF COMPLEX SURVEY DATA OR ECONOMETRIC ANALYSIS, WOULD TYPICALLY BE DONE USING A DIFFERENT STATISTICAL PROGRAM PACKAGE, IMPORTING THE DATA FILE CONSTRUCTED BY CSPro.

THE TABULATION APPLICATION OF CSPro CAN CROSS-TABULATE VARIABLES, AND IF APPLICABLE, PRODUCE MAP RESULTS BY GEOGRAPHICAL AREA USING BOTH EXISTING VARIABLES AND NEWLY CREATED VARIABLES. OUTPUT TABLES CAN CONTAIN SELECTED SUMMARY STATISTICS INCLUDING SIMPLE DATA COUNTS, PERCENTAGES, MEANS, MEDIANS, MODES, STANDARD DEVIATIONS, VARIANCES, N-TILES, PROPORTIONS, MINIMUMS, AND MAXIMUMS. TABULATIONS CAN BE MADE BASED ON VALUES FROM THE DATA FILE (AS IT IS) OR BY APPLYING WEIGHTS.

Epi Info IS STATISTICAL SOFTWARE FOR EPIDEMIOLOGY DEVELOPED BY THE U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC). IT IS CURRENTLY AVAILABLE FOR THE MICROSOFT WINDOWS, SAMSUNG ANDROID AND APPLE IOS OPERATING SYSTEMS, ALONG WITH A WEB AND CLOUD VERSION. THE PROGRAM ALLOWS FOR ELECTRONIC SURVEY CREATION, DATA ENTRY, AND ANALYSIS. WITHIN THE ANALYSIS MODULE, ANALYTIC ROUTINES INCLUDE t-TESTS, ANOVA, NONPARAMETRIC STATISTICS, CROSS TABULATIONS AND STRATIFICATION WITH ESTIMATES OF ODDS RATIOS, RISK RATIOS, AND RISK DIFFERENCES, LOGISTIC REGRESSION (CONDITIONAL AND UNCONDITIONAL), SURVIVAL ANALYSIS (KAPLAN-MEIER AND COX PROPORTIONAL HAZARD), AND ANALYSIS OF COMPLEX SURVEY DATA.

AN ANALYSIS CONDUCTED IN 2003 DOCUMENTED OVER 1,000,000 DOWNLOADS OF Epi Info FROM 180 COUNTRIES.

DATA IN Epi Info ARE STORED IN MICROSOFT Access DATABASE FORMAT.

A MAJOR BENEFIT OF Epi Info IS THAT IT INTEGRATES SUPPORT FOR EVERY STEP OF THE SURVEY PROCESS, FROM DEVELOPING THE QUESTIONNAIRE TO DATA ANALYSIS AND CREATING CUSTOM REPORTS. USERS DEVELOP A QUESTIONNAIRE, DEVELOP THE DATA-ENTRY PROCESS, ENTER DATA INTO THE DATABASE (INTO SCREENS WHICH WERE CREATED WHILE DEVELOPING THE QUESTIONNAIRE) AND ANALYZE THE DATA. FOR EPIDEMIOLOGICAL USES SUCH AS OUTBREAK INVESTIGATIONS, BEING ABLE TO RAPIDLY CREATE AN ELECTRONIC DATA ENTRY SCREEN AND THEN DO IMMEDIATE ANALYSIS ON THE COLLECTED DATA CAN SAVE CONSIDERABLE AMOUNTS OF TIME COMPARED TO USING PAPER SURVEYS.

AS SUCH, Epi Info IS ONE OF THE BEST SOFTWARE PACKAGES FOR SURVEY DEVELOPERS AND RESEARCHERS, ESPECIALLY THOSE WHO DO EPIDEMIOLOGICAL RESEARCH/SURVEYS. IT IS NOT DESIGNED TO ANALYZE A DATA SET CREATED USING OTHER SOFTWARE.

DATABASE MANAGEMENT SYSTEM

WHILE POPULATION DATA FROM A PARTICULAR CENSUS OR SURVEY MAY BE STORED IN TABLES OF A STATISTICAL PROGRAM PACKAGE, THE STORAGE OF DATA FOR MANY TIME PERIODS OR REGIONS IS STORED IN A DATABASE

MANAGEMENT SYSTEM. TO ALLOW FOR EASY RETRIEVAL, THE TYPE OF DATABASE USED WOULD TYPICALLY BE A *RELATIONAL* DATABASE MANAGEMENT SYSTEM. ACCORDING TO *DB-Engines*, IN JULY 2019, THE MOST WIDELY USED RELATIONAL DATABASE MANAGEMENT SYSTEMS WERE Oracle, MySQL (free software), Microsoft SQL Server, PostgreSQL (free software), IBM DB2, Microsoft Access, SQLite (free software), and MariaDB (free software).

SOME FREE RELATIONAL DATABASE MANAGEMENT SYSTEMS, SUCH AS MySQL, ARE NOT EASY TO USE. A KEY INGREDIENT IN A USER-FRIENDLY SYSTEM IS THE AVAILABILITY OF AN AUTOMATED QUERY-GENERATION SYSTEM (TO GENERATE SQL CODE). THE MICROSOFT Access SYSTEM (PART OF THE MICROSOFT OFFICE SUITE OF PRODUCTS) IS LOW-COST AND RELATIVELY EASY TO USE (SINCE IT CONTAINS A CAPABILITY FOR AUTOMATED GENERATION OF SQL CODE). RELATIONAL DATABASE SYSTEMS ARE INCLUDED IN OTHER OFFICE-SUITE PACKAGES (Paradox IN WordPerfect Office) AND IN OPEN-SOURCE SOFTWARE (Base in Apache OpenOffice AND LibreOffice).

ALTHOUGH INCLUDED IN OFFICE SUITES, RDBMSs ARE TYPICALLY PRICED SEPARATELY.

GEOGRAPHIC INFORMATION SYSTEM

TO DISPLAY THE GEOGRAPHIC DISTRIBUTION OF POPULATION, GEOGRAPHIC MAPPING SOFTWARE IS USED. A GEOGRAPHIC INFORMATION SYSTEM (GIS) IS A SYSTEM DESIGNED TO ACQIRE, STORE, PROCESS, ANALYZE AND PRESENT SPATIAL OR GEOGRAPHIC DATA. A FULL-FEATURED GIS INCLUDES A CAPABILITY TO PEFORM COMPLEX GEOGRAPHIC TRANSFORMATIONS AND EXECUTE SPATIAL QUERIES. IF ALL THAT IS DESIRED IS THE PRESENTATION OF MAPS, WITH LITTLE PROCESSING OR ANALYSIS, ALL THAT IS REQUIRED IS “MAPPING SOFTWARE.”

THERE ARE A SUBSTANTIAL NUMBER OF GEOGRAPHIC INFORMATION SYSTEMS AVAILABLE, INCLUDING A NUMBER OF FREE ONES. THE ESRI ArcGIS AND MapInfo SYSTEMS ARE POPULAR COMMERCIAL PRODUCTS. A VERY POWERFUL FREE SYSTEM IS THE GRASS (GEOGRAPHIC RESOURCES ANALYSIS SUPPORT SYSTEM) GIS, DEVELOPED MANY YEARS AGO AND STILL SUPPORTED BY THE U.S. ARMY CORPS OF ENGINEERS. HERE FOLLOWS AN EXCERPT FROM THE WIKIPEDIA ARTICLE ON GRASS:

Geographic Resources Analysis Support System (commonly termed GRASS GIS) is a geographic information system (GIS) software suite used for geospatial data management and analysis, image processing, producing graphics and maps, spatial and temporal modeling, and visualizing. It can handle raster, topological vector, image processing, and graphic data.

GRASS GIS contains over 350 modules to render maps and images on monitor and paper; manipulate raster and vector data including vector networks; process multispectral image data; and create, manage, and store spatial data.

It is licensed and released as free and open-source software under the GNU General Public License (GPL). It runs on multiple operating systems, including OS X, Windows and Linux. Users can interface with the software features through a graphical user interface (GUI) or by plugging into GRASS via other software such as QGIS. They can also interface with the modules directly through a bespoke shell that the application launches or by calling individual modules directly from a standard shell. The latest stable release version (LTS) is GRASS GIS 7, which is available since 2015.

The GRASS Development Team is a multinational group consisting of developers at many locations. GRASS is one of the eight initial Software Projects of the Open Source Geospatial Foundation.

THE QGIS OPEN SOURCE GIS IS ONE OF THE MORE POPULAR AND USER-FRIENDLY OPEN SOURCE GIS PACKAGES AVAILABLE. HERE IS A SUMMARY OF THE FUNCTIONALITY OF QGIS FROM THE WIKIPEDIA ARTICLE ON QGIS:

QGIS functions as geographic information system (GIS) software, allowing users to analyze and edit spatial information, in addition to composing and exporting graphical maps. QGIS supports both raster and vector layers; vector data is stored as either point, line, or polygon features. Multiple formats of raster images are supported, and the software can georeference images.

QGIS supports shapefiles, coverages, personal geodatabases, dxf, MapInfo, PostGIS, and other formats. Web services, including Web Map Service and Web Feature Service, are also supported to allow use of data from external sources.

QGIS integrates with other open-source GIS packages, including PostGIS, GRASS GIS, and MapServer. Plugins written in Python or C++ extend QGIS's capabilities. Plugins can geocode using the Google Geocoding API, perform geoprocessing functions similar to those of the standard tools found in ArcGIS, and interface with PostgreSQL/PostGIS, SpatiaLite and MySQL databases.

MATRIX OPERATIONS

DEMOGRAPHIC ANALYSIS INVOLVES MATRIX ALGEBRA. THERE ARE A NUMBER OF MATRIX ALGEBRA / LINEAR ALGEBRA PACKAGES AVAILABLE, INCLUDING VERY GOOD FREE ONES.

ONE OF THE MOST POPULAR COMMERCIAL LINEAR ALGEBRA PACKAGES IS MathWorks' MATLAB PACKAGE.

THE *NUMERICAL RECIPES* BOOK (BY PRESS ET AL.) INCLUDES A FULL RANGE OF ROUTINES FOR MATRIX ALGEBRA.

POPULAR FREE SOFTWARE FOR MATRIX ALGEBRA INCLUDES ALGLIB, ATLAS, Dlib, GNU Scientific Library, LAPACK, Math.NET Numerics, SciPy, EIgen, Armadillo, AND lbrsb.

A FULL RANGE OF MATRIX OPERATIONS IS AVAILABLE IN R.

STATISTICAL ANALYSIS (ESTIMATION, HYPOTHESIS TESTING AND FORECASTING)

THE MAJOR COMMERCIALLY AVAILABLE STATISTICAL PROGRAM PACKAGES ARE SAS, SPSS AND Stata. SOME OF THESE PACKAGES, SUCH AS SAS AND SPSS, CONTAIN MANY SUB-PACKAGES, WHICH ARE SOLD SEPARATELY. THERE IS ALSO MUCH FREE STATISTICAL SOFTWARE AVAILABLE IN R, AND ALSO IN PYTHON.

THE ADVANTAGES OF COMMERCIAL PACKAGES ARE NUMEROUS:

- AS A COMMERCIAL PRODUCT, THE PACKAGE IS WARRANTED TO PEFORM CORRECTLY. IF THE PRODUCT DOES NOT WORK AS ADVERTISED, THE FIRM WILL FIX IT OR REFUND YOUR MONEY. THIS IS

IN STARK CONTRAST TO R SOFTWARE, WHICH CARRIES THE FOLLOWING CAVEAT:

> This document is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2, or (at your option) any later version.
>
> This document is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

- COMMERCIAL STATISTICAL SOFTWARE PACKAGES ARE GENERALLY POWERFUL, WITH MANY FEATURES, AND DESIGNED FOR EASE OF USE (MANY AND REASONABLE DEFAULT SETTINGS).
- TRAINING PROGRAMS ARE AVAILABLE.
- MANY AFTER-MARKET REFERENCE TEXTS ARE AVAILABLE.
- GOOD DOCUMENTATION.
- HIGH QUALITY OF RESULTS (FEW ERRORS, REASONABLE PRESENTATION).
- TECHNICAL SUPPORT FROM THE COMPANY.
- TECHNICAL SUPPORT FROM THE INTERNET.
- IT IS MUCH EASIER FOR AN ORGANIZATION TO CONTROL QUALITY OF OUTPUT IF A SINGLE COMMERCIAL PACKAGE IS USED. IF AN ORGANIZATION’S STAFF IS FREE TO USE AN UNSPECIFIED SELECTION OF UNCERTIFIED R OR PYTHON ROUTINES, THE STATISTICAL ANALYSIS PROCESS IS NOT UNDER REASONABLE CONTROL. CERTIFICATION OF QUALITY WOULD REST ON THE CAPABILITIES OF THE PERSON WHO PERFORMED THE STATISTICAL ANALYSIS, MAKING IT DIFFICULT TO DOCUMENT THE PROCESS OR PRODUCT.
- IT IS EASIER TO ASSESS SKILLS OF POTENTIAL NEW EMPLOYEES OR CONTRACTORS RELATIVE TO PARTICULAR PACKAGES (SUCH AS SAS, SPSS, AND Stata) THAN FOR R OR PYTHON, WHICH INCLUDE A MASSIVE NUMBER OF LIBRARIES AND ROUTINES.
- SOME CLIENTS (SUCH AS BANKS, U.S. NATIONAL INSTITUTES OF HEALTH) REQUIRE CONTRACTORS TO USE HIGH-QUALITY, TESTED, WELL-KNOWN SOFTWARE (SUCH AS SAS).
- AT THE END OF A PROJECT, WHEN DATABASES AND STATISTICAL ANALYSIS COMMAND FILES ARE DELIVERED TO A CLIENT, THE CLIENT

MAY PREFER OR REQUIRE DELIVERY IN A PARTICULAR WELL-KNOWN PACKAGE, SUCH AS SAS, SPSS, OR Stata.

- WHILE THE LEARNING CURVE FOR COMPETENT USE OF ANY PARTICULAR SOFTWARE MAY BE STEEP, ONCE A USER IS FAMILIAR WITH IT, HE IS MORE EFFICIENT USING THAT SOFTWARE, INCLUDING RE-USE OF CODE, THAN TO LEARN A NEW ROUTINE.
- PRIOR TO USING FREE SOFTWARE, SUCH AS R OR PYTHON, IT MUST BE TESTED FOR ACCURACY. THIS IS NOT NECESSARY WITH A MAJOR COMMERCIAL PACKAGE.
- COMPREHENSIVE FACILITY FOR IMPORT AND EXPORT OF DATA.
- COMMUNICATION WITH CONSULTANTS AND SUBCONTRACTORS. SUBCONTRACTORS AND CONSULTANTS INVOLVED IN STATISTICAL ANALYSIS WILL HAVE EXPERTISE IN AT LEAST ONE MAJOR STATISTICAL PACKAGE, AND WILL GENERALLY BE ABLE TO RESPOND TO A STATISTICAL PROGRAMMING PROBLEM FASTER IF IT HAS BEEN PROGRAMMED IN A MAJOR SYSTEM (BY REVIEWING AND TESTING COMMAND-FILE CODE), THAN IF PROGRAMMED IN R OR A SMALL-MARKET-SHARE PRODUCT.

DISADVANTAGES OF MAJOR STATISTICAL SOFTWARE PACKAGES INCLUDE:

- THE COST CAN BE EXTREMELY HIGH. (THE PACKAGES ARE LEASED, NOT SOLD OUTRIGHT.)
- THE COMPLETE SOFTWARE SYSTEM MAY BE LEASED IN THE FORM OF MANY SUB-PACKAGES (E.G., SURVEY, TIME SERIES). IT IS NOT ECONOMICAL TO PURCHASE ALL OF THE COMPONENTS WHEN SOME MAY RARELY OR NEVER BE USED. IF THE NEED FOR A PARTICULAR FUNCTIONALITY ARISES, IT MAY NOT BE PRACTICAL TO ACQUIRE IT.
- WHILE MAJOR PACKAGES INCLUDE A WIDE VARIETY OF PROGRAMS, THEY MAY NOT INCLUDE A PROGRAM FOR A NEW TECHNIQUE, OR A HIGHLY SPECIALIZED ONE. SUCH APPLICATIONS ARE MORE LIKELY TO BE AVAILABLE IN R THAN IN MAJOR COMMERCIAL PACKAGES.
- SOME COMMERCIAL PACKAGES, SUCH AS SAS AND SPSS, ARE “CLOSED,” I.E., THE PACKAGE INCLUDES ONLY ROUTINES ISSUED BY THE VENDOR. OTHER COMMERCIAL PACKAGES, SUCH AS Stata ARE “OPEN,” AND INCLUDE MANY SPECIAL-PURPOSE ROUTINES DEVELOPED BY USERS (BUT VETTED BY THE SOFTWARE-PACKAGE FIRM).

- SINCE THE MAJOR STATISTICAL SOFTWARE PACKAGES INCLUDE A FIXED SET OF ROUTINES AND HAVE LARGE USER BASES, ERRORS ARE NOTICED AND CORRECTED.
- NEW TEXTBOOKS AND REFERENCE TEXTS OFTEN PROVIDE EXAMPLES OF ANALYSIS IN R, MORESO THAN IN A COMMERCIAL PACKAGE SUCH AS Stata. EMPLOYEES MAY BE ABLE TO ACCOMPLISH A ONE-TIME ANALYSIS REQUIREMENT MUCH FASTER AND MORE EFFICIENTLY BY FOLLOWING THE R EXAMPLE IN A REFERENCE TEXT.
- TO AN INCREASING EXTENT, RECENT COLLEGE GRADUATES WITH SOME STATISTICAL TRAINING ARE LIKELY TO HAVE SOME EXPERIENCE IN R.

SOFTWARE SPECIFICALLY ORIENTED TO DEMOGRAPHIC ANALYSIS

SOFTWARE PACKAGES HAVE BEEN DEVELOPED THAT ARE SPECIFICALLY ORIENTED TO DEMOGRAPHIC ANALYSIS. THESE PACKAGES CONTAIN MODULES THAT PERFORM MANY IF NOT ALL OF THE FOLLOWING FUNCTIONS:

- DISPLAY OF POPULATION FEATURES
- POPULATION PROJECTIONS
- POPULATION-BASED FORECASTS
- DIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS
- INDIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS
- FORECASTING DEMOGRAPHIC PARAMETERS

EACH OF THE PRECEDING FUNCTIONS WILL BE QUICKLY SUMMARIZED. BECAUSE THE PRECEDING FUNCTIONS ARE USUALLY PACKAGED TOGETHER, THE DISCUSSION ABOUT SOFTWARE WILL FOCUS ON A FEW PACKAGES IN THEIR ENTIRETY, AND NOT DESCRIBE THE IMPLEMENTATION OF THE FUNCTION WITHIN THE PACKAGE.

DISPLAY OF POPULATION FEATURES

THIS FUNCTION INCLUDES GRAPHIC PRESENTATIONS (GRAPHS, FIGURES, TABLES, CHARTS) SHOWING

- POPULATION BY COMPOSITION (AGE, SEX): POPULATION “AGE PYRAMIDS”

- POPULATION BY REGION (TABLES OR MAPS)
- POPULATION BY VARIOUS DEMOGRAPHIC CATEGORIES (RACE, MARITAL STATUS, FAMILY STATUS)
- TOTAL POPULATION OR SUBPOPULATIONS OVER TIME
- POPULATION PARAMETERS (MORTALITY, FERTILITY) OVER TIME

POPULATION PROJECTIONS (DETERMINISTIC)

THIS FUNCTION IS SIMILAR TO THE PRECEDING ONE, BUT FOR FUTURE TIMES, CORRESPONDING TO ASSUMPTIONS ABOUT MORTALITY, FERTILITY AND MIGRATION.

POPULATION FORECASTS (STATISTICAL; WITH ASSESSMENTS OF PRECISION OR LIKELIHOOD)

THIS FUNCTION IS SIMILAR TO THE PRECEDING ONE, BUT FOR FUTURE TIMES. THE FORECASTS ARE PREDICTIONS OF LIKELY FUTURE POPULATIONS, BASED ON STATISTICAL MODELS.

POPULATION-BASED FORECASTS

THESE ARE FORECASTS OF QUANTITIES, SUCH AS WELFARE BUDGETS AND CASELOADS, THAT ARE SUBSTANTIALLY DEPENDENT ON POPULATION LEVELS, COMPOSITION AND GEOGRAPHIC DISTRIBUTION.

DIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS

THIS FUNCTION RELATES TO ESTIMATION OF DEMOGRAPHIC PARAMETERS SUCH AS MORTALITY, FERTILITY AND MIGRATION RATES FROM VITAL REGISTRATION RECORDS, CENSUSES AND SURVEYS THAT RECORD EVENTS DIRECTLY RELATED TO THE RATES (SUCH AS BIRTHS AND DEATHS).

INDIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS

THIS FUNCTION RELATES TO ESTIMATION OF DEMOGRAPHIC PARAMETERS SUCH AS MORTALITY AND FERTILITY BY INDIRECT MEANS (I.E., NOT BY RECORDING BIRTHS AND DEATHS), SUCH AS A SURVEY OF FAMILY CHARACTERISTICS.

FORECASTING DEMOGRAPHIC PARAMETERS

THIS FUNCTION RELATES TO STATISTICAL FORECASTING OF FUTURE VALUES OF DEMOGRAPHIC PARAMETERS SUCH AS MORTALITY AND FERTILITY.

MAJOR DEMOGRAPHIC ANALYSIS SOFTWARE PACKAGES

THERE ARE A LARGE NUMBER OF COMPUTER SOFTWARE PACKAGES AVAILABLE FOR CONDUCTING DEMOGRAPHIC ANALYSIS AND MAKING POPULATION PROJECTIONS. THE BOOK, *BEYOND SIX BILLION: FORECASTING THE WORLD'S POPULATION*, PUBLISHED IN 2000, LISTS THE MAJOR ONES IN USE AT THAT TIME. SOME OF THEM WERE AVAILABLE TO THE PUBLIC, AND SOME WERE NOT. SOME ARE NO LONGER SUPPORTED.

IN THIS SURVEY COURSE, WE SHALL DESCRIBE AND ILLUSTRATE USE OF A FEW OF THE MORE WIDELY USED FREE SOFTWARE PACKAGES FOR DEMOGRAPHIC ANALYSIS. THE ONES TO BE DISCUSSED ARE:

- *Demographic Analysis & Population Projection System (DAPPS) Software, FROM THE U.S. CENSUS BUREAU*
- *Spectrum, FROM THE U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT HEALTH POLICY PLUS PROJECT OR ITS PARTNERS (Avenir Health AND OTHERS)*
- *MortPak, FROM THE UNITED NATIONS*
- *Tools from the International Union for the Scientific Study of Population (IUSSP)*
- THE R PACKAGE, DEMOGRAPHY
- YourCast, from Gary King

HERE FOLLOWS A BRIEF DESCRIPTION OF THE PRECEDING PACKAGES.

Demographic Analysis & Population Projection System (DAPPS) Software, FROM THE U.S. CENSUS BUREAU

THE CENSUS BUREAU'S DAPPS SOFTWARE PACKAGE IS AVAILABLE FREE FROM THE WEBSITE https://www.census.gov/data/software/dapps.html . HERE FOLLOWS A DESCRIPTION OF THAT PACKAGE, FROM THE WEBSITE:

Demographic Analysis & Population Projection System (DAPPS) Software. DAPPS is a program designed to help users analyze and produce population projections with ease. It accomplishes this through a user-friendly spreadsheet interface for data entry and the projection power of RUP.

In order to create a population projection, DAPPS requires at least three inputs: a base population, by age and sex (usually based on a census or estimate); a mortality structure, by age and sex (usually a life table or deaths, by age and sex); and a fertility pattern, by age of mother (births or age-specific fertility rates). A fourth input, a pattern of net migration (by age and sex of migrant), is optional but recommended.

The data for these components can originate from either a RUP input file or a spreadsheet-based program, such as Microsoft Excel or United Nations MORTPAK.

A DESCRIPTION OF ADDITIONAL SOFTWARE RESOURCES AVAILABLE FROM THE US CENSUS BUREAU IS AS FOLLOWS:

The following software products have been developed by staff of the U.S. Census Bureau. These programs are used by researchers in statistical offices, universities, and private organizations around the world for data collection, data processing, and population analysis work. The programs are available free of charge.

CSPro is a public domain software package used by organizations and individuals for entering, editing, tabulating, and disseminating census and survey data.

Population Analysis System (PAS) Software. PAS contains tools for analyzing age structure, mortality, fertility, migration, population distribution, and urbanization, to generate population projections.

Rural Urban Projection (RUP) Software. RUP is a computer program for projecting age and sex cohorts over time [for two regions, typically rural and urban].

Subnational Projections Toolkit (SPToolkit) Software. The Toolkit supports preparation of subnational population projections using cohort-component and/or non-cohort-component (mathematical extrapolation) methods.

Tool for Assessing Statistical Capacity (TASC). TASC provides a quantitative measure to the capacity of a National Statistical Office to conduct Population & Housing Censuses or household-based surveys.

Spectrum, FROM THE U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT HEALTH POLICY PLUS PROJECT OR ITS PARTNERS (Avenir Health AND OTHERS)

HERE FOLLOWS A DESCRIPTION OF THE SPECTRUM PACKAGE, FROM THE AVENIR WEBSITE, https://www.avenirhealth.org/software-spectrum.php .

SPECTRUM consists of several software models including:

- DemProj: Demography
- FamPlan: Family Planning
- LiST: Lives Saved Tool (Child Survival)
- AIM: AIDS Impact Model
- Goals: Cost and impact of HIV Intervention
- Resource Needs Module: Costs of implementing an HIV/AIDS program
- RAPID: Resources for the Awareness of Population Impacts on Development
- TIME: TB Impact Model and Estimates – Epidemiological and cost-effectiveness analysis of TB control strategies
- Malaria: Impact of malaria interventions
- STI: Estimation of burden and trends in Sexually Transmitted Infections
- NCD: Non-communicable diseases and mental health, substance abuse, and neurological disorders

Most models are available in English, French, and Spanish. Some are also available in Portuguese, Arabic, and Russian

MortPak, FROM THE UNITED NATIONS

THE UNITED NATIONS MAINTAINS A SOFTWARE PACKAGE, *MortPak*, WHICH PERFORMS A VARIETY OF DEMOGRAPHIC ANALYSES. THE PACKAGE CAN BE DOWNLOADED FROM WEBSITE https://www.un.org/en/development/desa/population/publications/mortality/mortpak.asp.

HERE IS AN EXCERPT DESCRIBING THE MORTPAK PACKAGE FROM THE PREFACE OF THE DOCUMENT, *MORTPAK FOR WINDOWS VERSION 4.3* (POSTED AT https://www.un.org/en/development/desa/population/publications/pdf/mortality/mortpak_manual.pdf).

The present volume contains the working manual for MORTPAK for Windows, the United Nations software package for demographic measurement in developing countries. The MORTPAK software packages for demographic measurement have had widespread use throughout research institutions in developing and developed countries since their introduction in 1988. Version 4.0 of MORTPAK included 17 applications in the areas of population projection, life-table and stable-population construction, graduation of mortality data, indirect mortality estimation, indirect fertility estimation, and other indirect procedures for evaluating age distributions and the completeness of censuses. Version 4.3 of MORTPAK enhanced many of the original applications and added 3 more to bring the total to 20 applications. The package incorporates techniques that take advantage of the United Nations model life tables and generalized stable-population equations. The package, as presented here, has been constructed with worksheet-style, full screen data entry which takes advantage of the interactive microcomputer environment and reduces dependence on a manual. The Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat has long conducted demographic estimation and projection activities at the country level, incorporating methodological advances in the construction of model life tables, for example. As a by-product of these activities, this extensive body of computer software has been developed.

MORTPAK has already been well tested and is now widely used for analysis of developing country data and in developing country institutions. The design of the applications in MORTPAK as well as the program MATCH has its origins in the United States Census Bureau package, Computer Programs for Demographic Analysis (Arriaga, Anderson and Heligman, 1976).

Tools from the International Union for the Scientific Study of Population (IUSSP)

THE IUSSP MAINTAINS A WEBSITE, Tools for Demographic Estimation by the International Union for the Scientific Study of Population (IUSSP), AT http://demographicestimation.iussp.org/ , WHICH PROVIDES A SET OF SOFTWARE TOOLS FOR ESTIMATING DEMOGRAPHIC PARAMETERS FROM LIMITED DATA.

HERE FOLLOWS A DESCRIPTION OF THE SITE, FROM THE SITE:

This site represents the major output arising from a joint IUSSP and UNFPA project to produce a single volume containing updated tools for demographic estimation from limited, deficient and defective data.

The material here follows in a direct line of descent from Manual X and subsequent works (for example, the 2002 UN Manual of Adult Mortality Estimation). The principal aspect of this website is a series of (mostly) static webpages describing and documenting the tools and methods of demographic estimation from limited, deficient and defective data. The material is organised thematically first, and then by the kinds of data that may be available. Where appropriate, downloadable spreadsheets are provided that allow users to apply the methods to their own data.

R FOR DEMOGRAPHY

FOR MOST APPLICATIONS, THE SOFTWARE PACKAGES DESCRIBED ABOVE ARE RECOMMENDED FOR USE. THEY ARE OF HIGH QUALITY, FREE, WELL SUPPORTED (MAINTENANCE, DOCUMENTATION AND TRAINING), WIDELY USED, AND COMPREHENSIVE IN CAPABILITY.

THAT SAID, THERE ARE SOME ROUTINES THAT ARE NOT INCLUDED IN SOME OF THE PACKAGE, SUCH AS THE LEE-CARTER METHOD. IN THAT CASE, IT MAY BE

DESIRABLE TO USE ONE OF THE ABOVE PACKAGES FOR GENERAL USE, AND USE A SPECIAL-PURPOSE R ROUTINE AS NEEDED. IN ADDITION, IT MAY BE DESIRED TO UNDERTAKE INTERACTIVE MODELING AND ANALYSIS, WHICH MIGHT BE MORE EFFICIENTLY DONE IN R, GIVEN THE EXPERIENCE OF THE USER.

HERE FOLLOWS A DESCRIPTION OF SOME R RESOURCES IN THE FIELD OF DEMOGRAPHY. NOTE THAT THE R LIBRARY ARCHIVE IS CONSTANTLY GROWING, AND NEW SOFTWARE IS BEING ADDED ALL THE TIME.

THE PENN STATE R USER GROUP POSTS THE FOLLOWING SUMMARY OF R SOFTWARE FOR DEMOGRAPHY, AT https://sites.psu.edu/rpsu/forums/topic/r-for-demography/ .

November 7, 2013 at 1:57 pm

R packages and code of particular interest to demographers.

Yes, demographers can use R! Here are a few resources to get you started…

- German Rodriguez also has a nice intro to R handout on his website http://data.princeton.edu/R/
- Another good introduction to the language is R Basics by Jamie Jones, a demographer at Stanford University
- There is also a demography package by Rob Hyndman that includes functions for common calculations used in demography such as lifetable calculations; Lee-Carter modelling and variants; functional data analysis of mortality rates, fertility rates, net migration numbers; and stochastic population forecasting. We show you how to use this package with data from the Human Mortality Database and from your own files.
- Also worth a look is the USCensus2000 suite of packages that allow for convenient handling of the 2000 US Census spatial and demographic data.
- Anthony Damico has written R code to read and analyze many large US government datasets http://www.asdfree.com . Check out his set of two-minute tutorials as well.
- Look at the Task View for Social Sciences on cran, new packages are always being added

- Eddie Hunsinger's Applied Demography Toolbox has some good R resources and there are a couple nice examples of raking & life tables on Sebatian Daza's site
- Tim Riffe, a PhD student in demography at the Centre for Demographic Studies (CED) and the Autonomous University of Barcelona, has put together a set of packages of interest to demographers.

Comprehensive R Archive Network (CRAN)

THE R PACKAGE, "DEMOGRAPHY" IS DESCRIBED AT https://cran.r-project.org/web/packages/demography/demography.pdf

HERE IS A SUMMARY OF THE PACKAGE:

Package: 'demography', April 22, 2019, Version 1.22
Title: Forecasting Mortality, Fertility, Migration and Population Data
Description: Functions for demographic analysis including life table calculations; Lee-Carter modelling; functional data analysis of mortality rates, fertility rates, net migration numbers; and stochastic population forecasting.
Author: Rob J Hyndman with contributions from Heather Booth, Leonie Tickle and John Maindonald.

YourCast from Gary King

FEDERICO GIROSI AND GARY KING WROTE A BOOK, *DEMOGRAPHIC FORECASTING*, WHICH IS AVAILABLE FROM WEBSITE https://gking.harvard.edu/files/abs/smooth-abs.shtml . SOFTWARE (CALLED *YourCast*) TO IMPLEMENT THE PROCEDURES DESCRIBED IN THE BOOK IS AVAILABLE FROM WEBSITE https://gking.harvard.edu/publications/yourcast .

THIS SOFTWARE WILL BE DISCUSSED FURTHER IN PART 2 OF THIS PRESENTATION.

7. POPULATION STATIC CHARACTERISTICS

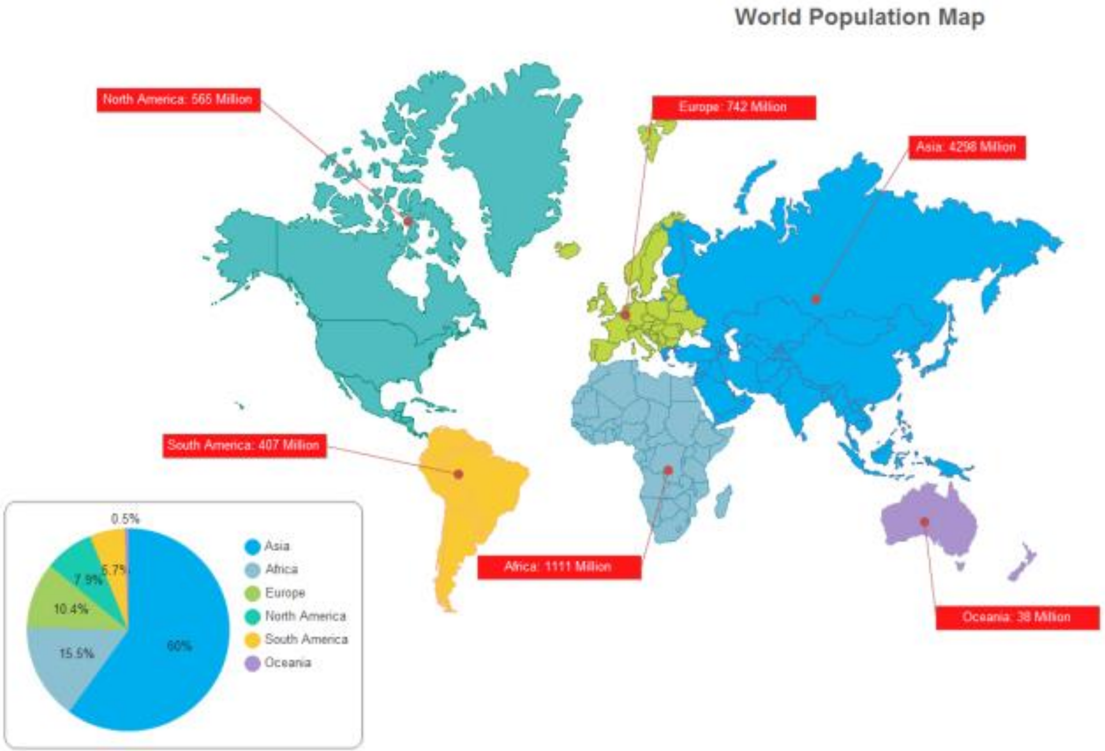
POPULATIONS ARE DYNAMIC ENTITIES. DESCRIPTORS OF POPULATIONS MAY BE DIVIDED INTO TWO CATEGORIES: DESCRIPTORS OF A POPULATION AT A POINT IN

TIME; AND MEASURES OF POPULATION CHANGE OVER TIME. THE FIRST CATEGORY OF DESCRIPTORS ARE STATIC DESCRIPTORS, AND THE SECOND CATEGORY ARE DYNAMIC DESCRIPTORS. THIS SECTION DESCRIBES STATIC CHARACTERISTICS OF POPULATIONS.

STATIC DESCRIPTORS INCLUDE THE FOLLOWING:

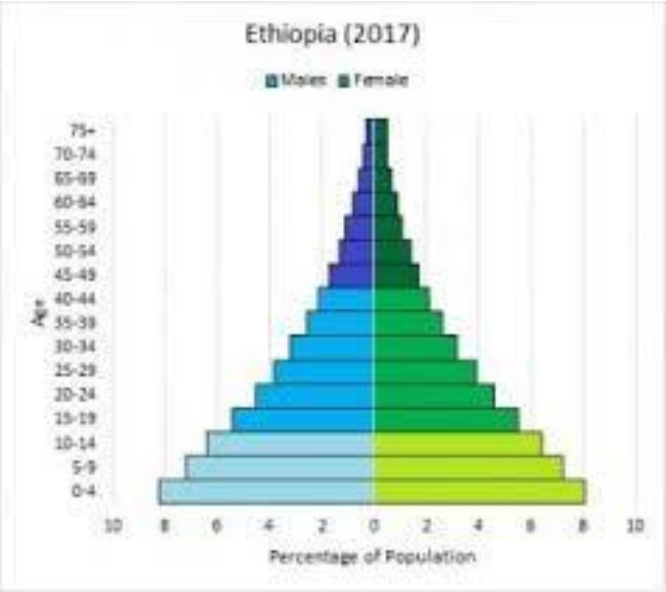
- POPULATION TOTAL
- POPULATION GEOGRAPHIC DISTRIBUTION
- POPULATION COMPOSITION (DISTRIBUTION BY DEMOGRAPHIC CHARACTERISTICS OTHER THAN OR ADDITIONAL TO GEOGRAPHY; AT LEAST BY AGE AND SEX, BUT ALSO BY ANY OTHER CHARACTERISTICS OF INTEREST, SUCH AS RACE, MARITAL STATUS, FAMILY STATUS, ECONOMIC STATUS, OR EDUCATIONAL STATUS)

THE GEOGRAPHIC DISTRIBUTION OF A POPULATION MAY BE DESCRIBED IN A TABLE, INDICATING COUNTS BY REGION OR PLACE, OR ON A MAP, SHOWING DENSITY BY SHADING OR COLOR. EXAMPLES ARE SHOWN IN FIGURE 5.



THE DISTRIBUTION OF A POPULATION BY ANY SINGLE CHARACTERISTIC, SUCH AS AGE, SEX, RACE, OR ECONOMIC STATUS, MAY BE SHOWN IN A VARIETY OF GRAPH TYPES, SUCH AS HISTOGRAMS, PIE CHARTS, OR DENSITY FUNCTIONS. AN EXAMPLES IS THE PIE CHART SHOWN IN 5.

A GRAPH CALLED A POPULATION AGE PYRAMID SHOWS THE DISTRIBUTION OF POPULATION BY AGE AND SEX (TOGETHER) AT A POINT IN TIME, USING AGE CATEGORIES OF THE SAME TIME SPAN. FIGURE 6 PRESENTS AN EXAMPLE OF A POPULATION AGE PYRAMID. FOR A POPULATION WITH A HIGH BIRTH RATE (GREATER THAN THE “REPLACEMENT RATE” OF TWO BABIES PER FEMALE), THE PYRAMID HAS A BROAD BASE; FOR A POPULATION WITH A BIRTH RATE CLOSE TO REPLACEMENT RATE, THE “PYRAMID” IS MORE “OBELISK-SHAPED.”



INTEREST IN POPULATION SIZE (COUNTS) ALONE, AT A SINGLE POINT IN TIME, IS LIMITED. OF GREATER INTEREST ARE RELATIONSHIPS OF POPULATION SIZE TO OTHER VARIABLES, SUCH AS THE FOLLOWING:

- MEASURES OF POPULATION CHANGE OVER TIME (I.E., DYNAMIC MEASURES SUCH AS GROWTH OR DECLINE);
- VARIABLES THAT AFFECT POPULATION SIZE (SUCH AS BIRTH RATES, DEATH RATES, AND MIGRATION RATES);

- VARIABLES THAT ARE AFFECTED BY POPULATION SIZE (SUCH AS WELFARE COSTS AND CASELOADS, DEMAND FOR MANUFACTURED PRODUCTS, AND QUALITY OF LIFE (SUCH AS CROWDING)); AND
- JOINT RELATIONSHIPS AMONG VARIABLES THAT ARE AFFECTED BY POPULATION SIZE.

THE NEXT SECTION WILL DISCUSS THE LAST TWO ITEMS (VARIABLES THAT ARE AFFECTED BY POPULATION SIZE). AFTER THAT, THE FIRST TWO ITEMS (DYNAMIC MEASURES) WILL BE ADDRESSED.

8. POPULATION-BASED ESTIMATES

THE REASON WHY DEMOGRAPHY IS A SUBJECT OF CONSIDERABLE INTEREST IS BECAUSE SO MANY OTHER VARIABLES IMPORTANT TO HUMAN LIFE ARE RELATED TO DEMOGRAPHIC VARIABLES.

MANY EXAMPLES CAN BE PRESENTED OF VARIABLES THAT ARE AFFECTED BY POPULATION. THIS SECTION WILL PRESENT SOME EXAMPLES OF POPULATION-BASED ESTIMATES.

STANDARDIZED (AGE-ADJUSTED) RATES

FREQUENTLY, IT IS DESIRED TO COMPARE TWO POPULATIONS WITH RESPECT TO A PARTICULAR CHARACTERISTIC, SUCH AS A DEATH RATE, OR A MORBIDITY RATE, OR A CRIME RATE, IN A SITUATION IN WHICH THE CHARACTERISTIC DEPENDS ON AGE, AND THE AGE DISTRIBUTION DIFFERS FOR THE TWO POPULATIONS. THE TWO POPULATIONS MIGHT BE, FOR EXAMPLE, TWO DIFFERENT CITIES.

A DIRECT COMPARISON OF THE MEAN RATE FOR THE TWO CITIES IS MISLEADING, BECAUSE IT IS AFFECTED BOTH BY THE AGE-SPECIFIC RATES AND THE AGE DISTRIBUTIONS OF THE TWO CITIES. IN A DIRECT COMPARISON OF THE MEAN FOR THE TWO CITIES, IT IS NOT KNOWN WHETHER AN OBSERVED DIFFERENCE IS DUE TO A DIFFERENCE IN THE AGE-SPECIFIC RATES OR A DIFFERENCE IN THE AGE DISTRIBUTION. IT IS SAID THAT THE AGE IS A CONFOUNDING VARIABLE – SINCE

THE AGE DISTRIBUTION MAY DIFFER FOR THE TWO CITIES, THE MEAN RATE MAY DIFFER, EVEN IF THE AGE-SPECIFIC RATES ARE THE SAME FOR BOTH CITIES.

IN ORDER TO COMPARE THE TWO CITIES WITH RESPECT TO THE CHARACTERISTIC IN A MORE UNDERSTANDABLE WAY, IT IS DESIRED TO TAKE INTO ACCOUNT THE DIFFERENCE IN THEIR AGE DISTRIBUTIONS. TWO WAYS OF DOING THIS ARE *DIRECT STANDARDIZATION* AND *INDIRECT STANDARDIZATION*.

DIRECT STANDARDIZATION MAY BE EMPLOYED WHEN AGE-SPECIFIC RATES OF THE CHARACTERISTIC ARE KNOWN FOR THE TWO CITIES. IN DIRECT STANDARDIZATION, THE TWO SETS OF AGE-SPECIFIC RATES ARE APPLIED (SEPARATELY) TO A “REFERENCE” OR “STANDARD” POPULATION, TO OBTAIN TWO ESTIMATES OF THE MEAN RATE FOR THAT POPULATION (CORRESPONDING TO THE TWO SETS OF AGE-SPECIFIC RATES).

THE TWO ESTIMATES OF DIRECT STANDARDIZATION ARE COMPLETELY HYPOTHETICAL, REFERRING TO FICTITIOUS SITUATIONS IN WHICH THE TWO DIFFERENT SETS OF AGE-SPECIFIC RATES ARE APPLIED TO THE REFERENCE POPULATION.

INDIRECT STANDARDIZATION MAY BE EMPLOYED WHEN AGE-SPECIFIC RATES OF THE CHARACTERISTIC ARE KNOWN FOR A REFERENCE POPULATION (E.G., ALL CITIES IN THE COUNTRY). IN INDIRECT STANDARDIZATION, THESE AGE-SPECIFIC RATES ARE APPLIED TO THE POPULATIONS FOR EACH OF THE TWO CITIES, TO OBTAIN ESTIMATES OF THE MEAN RATE FOR EACH CITY.

UNLIKE THE DIRECT-STANDARDIZED RATES, THE INDIRECT STANDARIZED RATES MAY BE REGARDED AS ESTIMATES OF THE MEAN RATES FOR THE TWO CITIES, UNDER THE ASSUMPTION THAT THE AGE-SPECIFIC RATES FOR THE REFERENCE POPULATION ARE THE SAME AS THOSE FOR EACH CITY.

INDIRECT RATES OF THIS SORT ARE AN EXAMPLE OF WHAT ARE CALLED “SYNTHETIC ESTIMATES.”

THE PROCESS OF STANDARDIZATION IS DESCRIBED IN DETAIL IN THE BOOK, *STATISTICAL METHODS FOR SURVIVAL DATA ANALYSIS*, 2nd ed., BY ELISA T. LEE (WILEY, 1992).

THE FORMULA FOR DIRECT STANDARDIZATION IS AS FOLLOWS:

SUPPOSE THAT THERE ARE k AGE GROUPS, AND THAT THE AGE-SPECIFIC RATES FOR THE CHARACTERISTIC ARE DENOTED AS r_{ij}, WHERE i DENOTES CITY (i = 1, 2) AND j DENOTES AGE GROUP (j = 1,…,k). LET p_j DENOTE THE PROPORTION OF THE POPULATION IN EACH AGE GROUP (j), IN THE REFERENCE POPULATION. THEN THE AGE-ADJUSTED MEAN RATE FOR CITY i IS:

$$R_{direct,i} = \sum_{j=1}^{k} r_{ij}\, p_j.$$

THE FORMULA FOR INDIRECT STANDARDIZATION IS AS FOLLOWS:

SUPPOSE THAT THERE ARE k AGE GROUPS, AND THAT THE AGE-SPECIFIC RATES FOR THE CHARACTERISTIC IN THE REFERENCE POPULATION ARE DENOTED AS s_j, WHERE j DENOTES AGE GROUP (j = 1,…,k). LET n_{ji} DENOTE THE POPULATION IN EACH OF THE j GROUPS, FOR CITY i (i=1,2). LET r DENOTE THE CRUDE RATE OF THE CHARACTERISTIC IN THE REFERENCE POPULATION. LET D_i DENOTE THE OBSERVED TOTAL FOR THE CHARACTERISTIC, IN CITY i. THEN THE AGE-ADJUSTED MEAN RATE FOR CITY i IS:

$$R_{indirect,i} = \frac{D_i}{\sum_{j=1}^{k} n_{ij} s_j} r.$$

RATIO ESTIMATES; SYNTHETIC ESTIMATES; SMALL-AREA ESTIMATION

THE PRECEDING SECTION (ON INDIRECT ESTIMATION) DESCRIBED A VERY BASIC METHOD OF ESTIMATING A RATE FOR A CHARACTERISTIC, WHERE AGE-SPECIFIC RATES ARE KNOWN FOR A REFERENCE POPULATION AND IT IS ASSUMED THAT THOSE SAME RATES APPLY TO ANOTHER POPULATION OF INTEREST. THIS APPROACH IS OFTEN APPLIED WHEN THE AGE-SPECIFIC RATES ARE KNOWN FOR A LARGE POPULATION, SUCH AS A COUNTRY OR STATE, AND IT IS DESIRED TO OBTAIN ESTIMATES FOR A SMALLER REGION WITH THE LARGER REGION. SUCH ESTIMATES ARE CALLED “SMALL-AREA ESTIMATES.”

THERE ARE A NUMBER OF TECHNIQUES FOR SMALL-AREA ESTIMATION. THE PARTICULAR SMALL-AREA ESTIMATION PROCEDURE DESCRIBED ABOVE (INDIRECT STANDARDIZATION) IS CALLED “SYNTHETIC ESTIMATION.”

IN THE PRECEDING EXAMPLE, SYNTHETIC ESTIMATION WAS APPLIED IN THE CASE OF A SINGLE COVARIATE, AGE. IN GENERAL, RATE DATA ARE OFTEN AVAILABLE FOR A NUMBER OF DEMOGRAPHIC CHARACTERISTICS, INCLUDING AGE, SEX, RACE AND GEOGRAPHIC REGION. IF THESE DATA ARE AVAILABLE IN TABULAR FORM, THEN THE PROCEDURE FOR IMPLEMENTING SYNTHETIC ESTIMATION IN THE CASE OF MORE THAN ONE COVARIATE IS A DIRECT EXTENSION OF THE METHOD DESCRIBED ABOVE. IF THE RELATIONSHIPS ARE DESCRIBED IN THE FORM OF REGRESSION-TYPE RELATIONSHIPS, THEN THE PROCEDURE IS MORE COMPLICATED.

IN GENERAL, THE PROCEDURE FOR MAKING SYNTHETIC ESTIMATES IS ANALOGOUS TO THE PROCEDURE FOR MAKING A CAUSAL ESTIMATE. THE ESTIMATE IS OBTAINED BY AVERAGING OVER ALL VARIABLES THAT HAVE A CAUSAL INFLUENCE ON THE VARIABLE OF INTEREST. IN THE CASE OF A SINGLE COVARIATE, THE FORMULA FOR ESTIMATING THE CAUSAL EFFECT OF A VARIABLE x ON ANOTHER, y, IN THE PRESENCE OF A CONFOUNDING VARIABLE, z, IS:

$$P\big(y\big|do(x)\big) = \sum_{z} P(y|x,z)P(z)$$

WHERE P DENOTES THE PROBABILITY FUNCTION, y DENOTES THE CHARACTERISTIC OF INTEREST, x DENOTES THE VARIABLE FOR WHICH THE CAUSAL EFFECT IS DESIRED (CITY, IN THE EXAMPLE GIVEN ABOVE), AND z DENOTES THE CONFOUNDING VARIABLE (A VARIABLE THAT AFFECTS BOTH y AND x; AGE, IN THE EXAMPLE GIVEN ABOVE).

THE SYNTHETIC ESTIMATION PROCEDURE IS IMPLEMENTED IN THE SPECTRUM SOFTWARE TO BE DESCRIBED LATER IN THIS PRESENTATION. THIS METHODOLOGY IS RELEVANT TO A WIDE RANGE OF APPLICATIONS, SUCH AS ESTIMATION OF WELFARE CASELOADS AND BUDGETS, ESTIMATION OF SCHOOL ENROLMENTS, ESTIMATION OF PRISON POPULATIONS, AND MARKETING.

THE METHODOLOGY FOR SMALL-AREA ESTIMATION IS DESCRIBED IN DETAIL IN THE BOOK, *SMALL AREA ESTIMATION* BY J. N. K. RAO (WILEY, 2003). A SHORT COURSE ON THIS SUBJECT IS INCLUDED IN THIS PRESENTATION SERIES. FOR ADDITIONAL INFORMATION ON SMALL-AREA ESTIMATION, REFER TO THE RAO BOOK OR THE SHORT-COURSE LECTURE NOTES.

9. POPULATION DYNAMIC CHARACTERISTICS

POPULATION DYNAMIC CHARACTERISTICS ARE CHARACTERISTICS RELATING TO TIME, SUCH AS NUMBERS OR MEASURES (SUCH AS RATES) RELATING TO BIRTHS, DEATHS, IMMIGRANTS AND EMIGRANTS.

PERIODS AND COHORTS; THE LEXIS DIAGRAM

TO ESTIMATE DYNAMIC CHARACTERISTICS, IT IS NECESSARY TO OBSERVE A POPULATION AT MORE THAN ONE POINT IN TIME. THERE ARE TWO BASIC METHODS FOR OBSERVING POPULATIONS AT MORE THAN ONE POINT IN TIME. THE FIRST METHOD IS TO OBSERVE A (SPECIFIED, WELL-DEFINED) POPULATION OF INDIVIDUALS OF THE SAME AGE AT SUCCESSIVE POINTS IN TIME (I.E., AT SUCCESSIVE AGES). SUCH A POPULATION IS CALLED A COHORT. THE SECOND IS TO OBSERVE A (SPECIFIED, WELL-DEFINED) POPULATION OF INDIVIDUALS OF VARIOUS AGES AT TWO DIFFERENT POINTS IN TIME (SUCH AS AT TWO SUCCESSIVE CENSUSES). THE TWO POINTS IN TIME DEFINE A PERIOD.

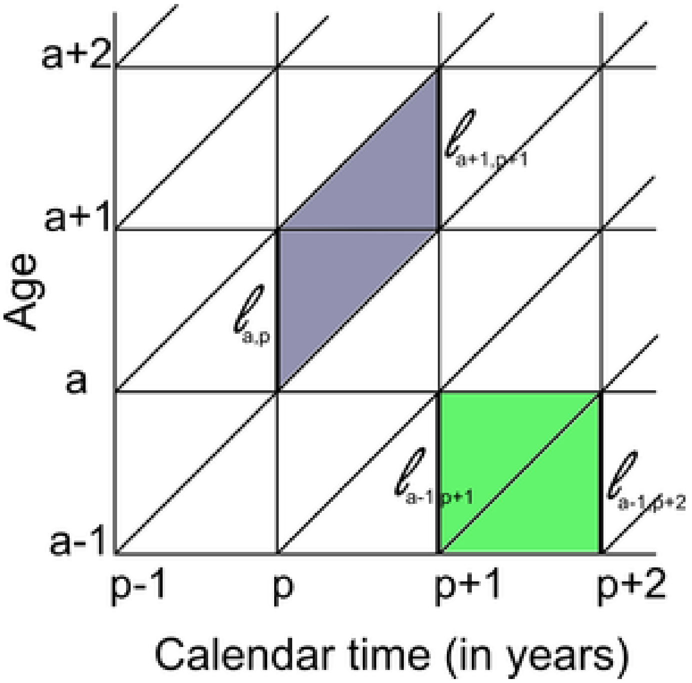
(WE SHALL USE THE TERM “CHARACTERISTIC” GENERALLY TO REFER TO A QUALITATIVE (BUT SPECIFIC) CONCEPT OR ATTRIBUTE (SUCH AS A GROWTH RATE), AND THE TERMS “MEASURE” OR “DESCRIPTOR” TO REFER TO A QUANTITATIVELY DEFINED VARIABLE RELATED TO A CHARACTERISTIC (SUCH AS A LINEAR GROWTH RATE OR AN EXPONENTIAL GROWTH RATE).)

IN WORKING WITH COHORTS, NO MIGRATION IS ALLOWED INTO OR OUT OF THE COHORT. THE CIRCUMSTANCE OF NO MIGRATION APPLIES TO SOME COUNTRIES OVER SOME PERIODS OF TIME. IT ALSO APPLIES TO “CASE CONTROL” STUDIES OF SPECIFIC POPULATIONS OVER TIME, AS IN A CLINICAL TRIAL OF A NEW PHARMACEUTICAL DRUG.

BECAUSE ALL MEMBERS OF A COHORT ARE OF THE SAME AGE, AND BECAUSE NO MIGRATION IS ALLOWED INTO OR OUT OF THE COHORT, A COHORT IS MORE HOMOGENEOUS RELATIVE TO THESE TWO FACTORS THAN A (NON-COHORT) POPULATION DEFINED SOLELY BY A TIME PERIOD, AND SOME POPULATION CONCEPTS ARE EASIER TO MEASURE AND ESTIMATE USING COHORT DATA THAN USING PERIOD DATA. ON THE OTHER HAND, COHORT MEMBERS OF DIFFERENT AGES MAY BE SUBJECT TO DIFFERENT CIRCUMSTANCES (SUCH AS WARS, ECONOMIC RECESSIONS, AND GEOGRAPHIC LOCATIONS), SO THERE ARE CERTAINLY NUMEROUS FACTORS WITH RESPECT TO WHICH THEY ARE NOT HOMOGENEOUS.

NOTE THAT BIRTHS TO COHORT MEMBERS ARE NOT INCLUDED IN THE COHORT. THE COHORT CONSISTS OF ITS ORIGINAL MEMBERS (ALL OF THE SAME AGE), AND DECREASES IN SIZE AS THEY DIE.

FIGURE 7 ILLUSTRATES POPULATIONS DEFINED BY COHORT AND BY PERIOD. THIS GRAPH, DEPICTING A POPULATION IN TERMS OF BOTH AGE AND TIME, IS CALLED A *LEXIS DIAGRAM*. (THE AGE AXIS MAY INCREASE UPWARD OR DOWNWARD, BUT UPWARD IS MORE COMMON.)



THE DYNAMIC CHARACTERISTICS OF INTEREST FOR POPULATIONS RELATE TO INSTANCES OF BIRTHS, DEATHS AND MIGRATION, TO GROWTH, AND TO CHANGES IN DISTRIBUTION AND COMPOSITION OVER TIME.

THE MEASURES USED TO DESCRIBE DYNAMIC CHARACTERISTICS ARE PROBABILITIES AND RATES OF OCCURRENCE OF EVENTS OF INTEREST, SUCH AS BIRTHS AND DEATHS. USUALLY, PROBABILITIES ARE ESTIMATED FROM COHORT DATA AND RATES ARE ESTIMATED FROM PERIOD DATA.

WE SHALL HOW DESCRIBE A NUMBER OF POPULATION MEASURES, RELATIVE TO COHORTS AND TO PERIODS. WE BEGIN WITH PERIODS.

POPULATION DESCRIPTORS FOR PERIODS

UNDERSTANDING OF BASIC DEMOGRAPHIC PHENOMENA SUCH AS BIRTHS, DEATHS, MARRIAGES, AND MIGRATION IS ENHANCED BY RELATING THE OBSERVED NUMBERS OF INSTANCES TO A BASE. OVER A PERIOD OF TIME, THE NUMBER OF OCCURRENCES WILL VARY ACCORDING TO THE SIZE OF THE POPULATION AND THE TIME DURATION (LENGTH) OF THE PERIOD. WHILE THE LENGTH OF THE PERIOD IS FIXED, THE SIZE OF THE POPULATION MAY VARY OVER THE PERIOD.

TO TAKE VARIATIONS IN POPULATION SIZE INTO ACCOUNT, A REASONABLE APPROACH IS TO CONSIDER RATES OF OCCURRENCE PER PERSON PER UNIT OF TIME (SAY, A YEAR). THIS APPROACH IS IMPLEMENTED BY USING “OCCURRENCE/EXPOSURE” RATES. IN THIS CASE, A POPULATION RATE OF OCCURRENCE IS DEFINED AS

> RATE OF OCCURRENCE OF AN EVENT IN A PERIOD = (NUMBER OF OCCURRENCES OF THE EVENT IN THE PERIOD) / (TOTAL TIME DURATION (SUMMED OVER THE POPULATION) OF EXPOSURE TO THE RISK OF OCCURRENCE IN THE PERIOD).

IF PERSON-YEARS IS USED IN THE DENOMINATOR, THEN THE RATE IS CALLED AN ANNUALIZED RATE. IN THIS PRESENTATION, WE SHALL USE ANNUALIZED RATES, UNLESS OTHERWISE SPECIFIED. IN THIS CASE, THE PRECEDING EQUATION READS

> RATE OF OCCURRENCE OF AN EVENT IN A PERIOD = (NUMBER OF OCCURRENCES OF THE EVENT IN THE PERIOD) / (PERSON-YEARS IN THE PERIOD).

A VERY SIGNIFICANT PROBLEM WITH THIS APPROACH IS THAT IN MANY DEMOGRAPHIC APPLICATIONS THE TIME DURATION OF EXPOSURE IS RARELY EVER OBSERVED OR RECORDED, SO THAT THE QUANTITY CANNOT BE CALCULATED FROM DIRECTLY OBSERVED DATA. THE PRECEDING MEASURE IS OF THEORETICAL INTEREST, BUT OF LIMITED PRACTICAL INTEREST.

FOR RELATIVELY SHORT TIME PERIODS, IN WHICH NOT MUCH POPULATION GROWTH OCCURS, THE PRECEDING EXPRESSION MAY BE APPROXIMATED AS

> RATE OF OCCURRENCE OF AN EVENT IN A PERIOD = (NUMBER OF OCCURRENCES OF THE EVENT IN THE PERIOD) / (SIZE OF THE POPULATION AT THE PERIOD MIDPOINT x LENGTH (TIME DURATION) OF THE PERIOD).

SUCH A RATE IS CALLED A “CENTRAL” RATE, SINCE IT IS BASED ON THE POPULATION AT THE CENTER (MIDPOINT) OF THE PERIOD. THIS DEFINITION DOES NOT TAKE INTO ACCOUNT VARIATIONS IN THE POPULATION SIZE OVER THE OBSERVATION PERIOD. IF THE POPULATION IS GROWING AT AN EXPONENTIAL RATE, THE RATE USING THIS SECOND DEFINITION WILL BE HIGHER THAN UNDER THE FIRST DEFINITION.

(THIS PART OF THE PRESENTATION DOES NOT DEAL WITH THE STATISTICAL ESTIMATION OF POPULATION MEASURES FROM SAMPLE DATA. THE RATES DEFINED HERE ARE POPULATION MEASURES THAT MAY IN CONCEPT BE CALCULATED FOR ANY FINITE POPULATION (ALTHOUGH DATA MAY NOT BE AVAILABLE TO DO SO IN A PARTICULAR CASE). HERE, NO STATISTICAL MODEL IS SPECIFIED FOR THE MEASURES; THEY ARE NOT VIEWED AS ESTIMATES OF MODEL PARAMETERS.)

BOOKS ON MATHEMATICAL DEMOGRAPHY GENERALLY USE THE FIRST DEFINITION, WHEREAS BOOKS ON GENERAL DEMOGRAPHY GENERALLY USE THE SECOND DEFINITION. THE MAIN TEXT FOR THIS PRESENTATION (*METHODS AND MATERIALS OF DEMOGRAPHY*) USES THE SECOND DEFINITION, AS SHALL THIS

PRESENTATION. THE CHOICE OF THE SECOND DEFINITION RESULTS IN SIMPLER FORMULAS, WITH NO SIGNIFICANT EFFECT ON DISCUSSION OF CONCEPTS.

SOME OF THE KEY MEASURES OF POPULATION DYNAMIC CHARACTERISTICS WILL NOW BE DISCUSSED (USING, AS DISCUSSED, THE SECOND DEFINITION OF RATE).

PRIOR TO DISCUSSING THESE MEASURES, WE RESTATE THE DEMOGRAPHIC BALANCING EQUATION. IT WAS PRESENTED EARLIER, REFERRING TO A PERIOD OF UNSPECIFIED LENGTH (TIME DURATION), AS

$$P_{end} = P_{beg} + B - D + I - E,$$

WHERE

P_{end} = POPULATION SIZE (NUMBER OF PERSONS) AT END OF PERIOD

P_{beg} = POPULATION SIZE AT BEGINNING OF PERIOD

B = NUMBER OF BIRTHS DURING PERIOD

D = NUMBER OF DEATHS DURING PERIOD

I = NUMBER OF IMMIGRANTS DURING PERIOD

E = NUMBER OF EMIGRANTS DURING PERIOD.

LET P DENOTE THE POPULATION SIZE AT THE MIDPOINT OF THE TIME PERIOD, AND T DENOTE THE LENGTH (TIME DURATION, IN YEARS) OF THE PERIOD.

THEN THE FOLLOWING (CENTRAL) CRUDE RATES ARE DEFINED:

CRUDE BIRTH RATE = CBR = B/(PT)

CRUDE DEATH RATE = CDR = D/(PT)

CRUDE IMMIGRATION RATE = CRIM = I/(PT)

CRUDE EMIGRATION RATE = CREM = E/(PT)

CRUDE GROWTH RATE = CGR = $(P_{end} – P_{beg})/(PT)$

CRUDE RATE OF NATURAL INCREASE = CRIN = CBR – CDR

CRUDE NET MIGRATION RATE = CRNM = CRIM – CREM.

DIVIDING THE BALANCING EQUATION BY PT, WE CAN WRITE IT EQUIVALENTLY AS:

CGR = CBR – CDR + CRIM – CREM

= CRIN + CRNM.

THE FORMULAS OBVIOUSLY SIMPLIFY IF THE PERIOD IS OF LENGTH T = 1 YEAR, SINCE THE DENOMINATOR PT BECOMES SIMPLY P. IN APPLIED DEMOGRAPHY, THE PERIOD LENGTH T = 5 YEARS IS OFTEN USED.

IF T=1, THEN THE CRUDE BIRTH RATES AS DEFINED ABOVE ARE MEAN ANNUALIZED GROWTH RATES. IF T IS NOT EQUAL TO ONE, THIS IS NOT THE CASE. IF THE GROWTH RATE IS CONSTANT DURING THE PERIOD (WHETHER POSITIVE OR NEGATIVE), THEN THE MIDPOINT POPULATION WILL ALWAYS UNDERESTIMATE THE TRUE NUMBER OF PERSON-YEARS. FOR T DIFFERENT FROM ONE, INSTEAD OF APPROXIMATING PERSON-YEARS (IN THE FIRST DEFINITION OF CRUDE RATE GIVEN ABOVE) BY PT, A BETTER APPROXIMATION IS

$PY = (P_{end} – P_{beg})T/(\ln(P_{end}/P_{beg}))$.

WHERE PY DENOTES THE NUMBER OF PERSON YEARS IN THE INTERVAL.

(THIS RESULT CORRESPONDS TO ASSUMING A CONSTANT (EXPONENTIAL) GROWTH RATE, r, DURING THE PERIOD. IN THIS CASE, THE VALUE OF r IS

$r = [\ln(P_{end}/P_{beg})]/T$.

THE VALUE OF THE CRUDE GROWTH RATE IS

$CGR = (P_{end} – P_{beg})/PY.$

IF THIS IS SET EQUAL TO r, WE HAVE

$[\ln(P_{end}/P_{beg})]/T = (P_{end} – P_{beg})/PY$

OR

$PY = (P_{end} – P_{beg})T/ \ln(P_{end}/P_{beg}).)$

THE PRECEDING MEASURES RELATE TO TERMS IN THE BALANCING EQUATION. ADDITIONAL BASIC DEMOGRAPHIC RATES ARE THE FOLLOWING.

THE INFANT MORTALITY RATE IS DEFINED AS:

IMR = (NUMBER OF DEATHS UNDER AGE 1 IN THE PERIOD)/(NUMBER OF LIVE BIRTHS IN THE PERIOD).

THE NET REPRODUCTION RATE IS DEFINED AS:

NRR = (TOTAL NUMBER OF DAUGHTERS BORN TO COHORT MEMBERS) / (INITIAL NUMBER OF WOMEN IN THE COHORT).

(Note: Some authors, who use the term “rate” only for ratios in which the denominator is a measure of time, call the IRR the “net reproduction ratio.”)

THE FIELD OF DEMOGRAPHY INVOLVES MANY FORMULAS. THIS PRESENTATION COVERS A LOT OF MATERIAL IN A SHORT TIME, AND ATTEMPTING TO CITE MANY OF THEM WOULD BE DETRIMENTAL TO THE PRIMARY OBJECTIVE OF CONVEYING AN UNDERSTANDING OF THE BASIC CONCEPTS. FOR THIS PRESENTATION, A BRIEF SURVEY, THERE IS A SUBSTANTIAL ADVANTAGE TO USING SIMPLE APPROXIMATIONS, AND THAT WILL BE DONE. IN ACTUAL PRACTICE, THE MORE DETAILED FORMULAS WOULD TYPICALLY BE USED (THEY ARE AVAILABLE IN BOOKS ON MATHEMATICAL DEMOGRAPHY). THE MORE DETAILED FORMULAS ARE USED IN THE SOFTWARE THAT IS ILLUSTRATED IN THE PRESENTATION.

THE CRUDE RATES DEFINED ABOVE MAY BE CALCULATED FOR AN ENTIRE POPULATION OR FOR SUBPOPULATIONS. THE RATES ARE VERY DEPENDENT ON A VARIETY OF DEMOGRAPHIC CHARACTERISTICS, SUCH AS AGE, SEX, MARITAL STATUS, AND COUNTRY.

THE PRECEDING RATES, CALCULATED FROM A PERIOD (I.E., A CROSS-SECTION OF A POPULATION OVER A CLOSED TIME INTERVAL), ARE USEFUL DESCRIPTORS OF THE OVERALL NATURE OF A POPULATION. THEY ARE EASY TO CALCULATE, AND DATA ARE AVAILABLE FROM MOST COUNTRIES TO CALCULATE THEM. THEY MAY REASONABLY BE USED TO PROJECT GROWTH UNDER THE SAME CONDITIONS AS APPLY TO THE DATA FROM WHICH THEY WERE CALCULATED. OFFSETTING THEIR SIMPLICITY AND AVAILABILITY IS THE FACT THAT THEY ARE NOT VERY USEFUL FOR IMPORTANT APPLICATIONS, SUCH AS INSURANCE, PLANNING AND POLICY ANALYSIS.

ONE PROBLEM WITH RATES CALCULATED FROM CROSS-SECTIONAL DATA IS THAT THEY MAY NOT BE USED TO ESTIMATE PROBABILITIES OF DEATH AT SPECIFIED AGES. THE REASON FOR THIS IS THAT THE POPULATIONS REPRESENTING DIFFERENT AGE GROUPS DO NOT INCLUDE THE SAME INDIVIDUALS. FOR EXAMPLE, THE RATIO OF (POPULATION SIZE OF TWENTY-YEAR-OLDS) / (POPULATION OF NINETEEN-YEAR OLDS) DOES NOT REPRESENT THE PROBABILITY THAT A NINETEEN-YEAR-OLD SURVIVES TO AGE TWENTY. FOR CROSS-SECTIONAL DATA, THE POPUATION OF TWENTY-YEAR-OLDS COULD IN FACT EXCEED THE POPULATION OF NINETEEN-YEAR-OLDS, BECAUSE OF MIGRATION, A LOWER MORTALITY RATE, OR RANDOM FLUCTUATIONS, RESULTING IN A RATIO THAT EXCEEDS ONE (AND HENCE CANNOT REPRESENT A PROBABILITY).

AS ANOTHER EXAMPLE, THE INFANT MORTALITY RATE DOES NOT REPRESENT THE PROBABILITY THAT AN INFANT SURVIVES (AGE ONE), FOR THE SAME REASON: THE NUMERATOR (NUMBER OF INFANT DEATHS) MAY EXCEED THE DENOMINATOR (NUMBER OF BIRTHS) BECAUSE THESE NUMBERS ARE CALCULATED FROM DIFFERENT POPULATIONS.

FOR MANY APPLICATIONS, IT IS MUCH MORE USEFUL TO WORK WITH MEASURES THAT REPRESENT STATISTICAL QUANTITIES, SUCH AS PROBABILITIES AND EXPECTED VALUES (MEANS). THE REASONS FOR THIS ARE SEVERAL. IF A STATISTICAL FRAMEWORK (SAMPLE SPACE, PROBABILITIES, RANDOM VARIABLES)

IS ADOPTED, THE THEORY OF STATISTICS MAY BE APPLIED TO ADDRESS ISSUES OF INTEREST.

WE SHALL NOW DISCUSS POPULATION DESCRIPTORS FOR COHORTS, WHICH INCLUDE A NUMBER OF PROBABILITIES AND EXPECTATIONS.

POPULATION DESCRIPTORS FOR COHORTS

THERE ARE A NUMBER OF POPULATION MEASURES THAT ARE DEFINED RELATIVE TO COHORTS. BECAUSE NO MIGRATION IS ALLOWED INTO OR OUT OF THE COHORT, MEASURES DEFINED FOR DIFFERENT AGE GROUPS ARE BASED ON THE SAME POPULATION (OBSERVED LONGITUDINALLY OVER TIME), AND PROBABILITIES AND EXPECTATIONS OF A NUMBER OF AGE-RELATED CHARACTERISTICS MAY BE ESTIMATED WITH HIGH PRECISION DIRECTLY FROM COHORT DATA.

FOR NON-HUMAN POPULATIONS WITH RELATIVELY SHORT LIFE-TIMES, IT MAY BE FEASIBLE TO OBSERVE A COHORT OVER THE LIFETIMES OF ITS MEMBERS. IN MOST SITUATIONS, IT IS NOT FEASIBLE TO OBSERVE A HUMAN COHORT OVER THE FULL LIFE SPANS OF ITS MEMBERS. WHAT IS USUALLY DONE IS TO SPECIFY A HYPOTHETICAL, OR “SYNTHETIC,” COHORT, AND ESTIMATE ITS CHARACTERISTICS INDIRECTLY, FROM CROSS-SECTIONAL DATA. THE SYNTHETIC COHORT IS THEN USED AS A BASIS FOR ESTIMATING OTHER MEASURES OF INTEREST.

THE CRUDE MEASURES PRESENTED ABOVE HAVE COUNTERPARTS RELATIVE TO COHORTS. IT IS CONVENIENT TO SEPARATE THE DISCUSSION INTO TWO PARTS, THE FIRST DEALING WITH MORTALITY (DEATHS) AND THE SECOND DEALING WITH FERTILITY (BIRTHS).

COHORT-BASED MORTALITY MEASURES

THE PRIMARY USE OF A COHORT IS TO FORM THE BASIS FOR CONSTRUCTING A TABLE SHOWING THE MORTALITY EXPERIENCE OF THE COHORT BY AGE OF ITS MEMBERS, SPECIFICALLY, THE NUMBER OF MEMBERS STILL ALIVE (I.E., THE SURVIVORS) AT VARIOUS AGES. SUCH A TABLE IS CALLED A “MORTALITY TABLE” IN MUCH OF THE WORLD, AND A “LIFE TABLE” IN THE UNITED STATES.

A LIFE TABLE IS USED BY INSURANCE COMPANIES TO DETERMINE PROFITABLE PRICES FOR LIFE-INSURANCE POLICIES. IN THAT APPLICATION, THE LIFE TABLE SHOWS THE NUMBER OF SURVIVORS REMAINING AT EACH YEAR. SUCH A TABLE IS CALLED A "COMPLETE" LIFE TABLE. FOR MOST DEMOGRAPHIC APPLICATIONS, THE TABLE SHOWS SURVIVORS REMAINING AT THE END OF THE FIRST YEAR, THE FIFTH YEAR, AND EVERY FIFTH YEAR AFTER THAT. SUCH A LIFE TABLE IS CALLED AN "ABRIDGED" LIFE TABLE.

A LIFE TABLE MAY REPRESENT THE MORTALITY EXPERIENCE FOR A POPULATION OVER A PERIOD OF TIME, REPRESENTED BY THE VERTICAL BAND IN THE LEXIS DIAGRAM. SUCH A LIFE TABLE IS CALLED A "CURRENT" LIFE TABLE. OR, IT MAY REPRESENT THE MORTALITY EXPERIENCE OF A COHORT, IN WHICH CASE IT IS CALLED A "COHORT" LIFE TABLE. FOR LIFE INSURANCE APPLICATIONS, A CURRENT LIFE TABLE IS DESIRED, FOR A RECENT PERIOD.

FIGURE 8 PRESENTS AN EXAMPLE OF A LIFE TABLE.

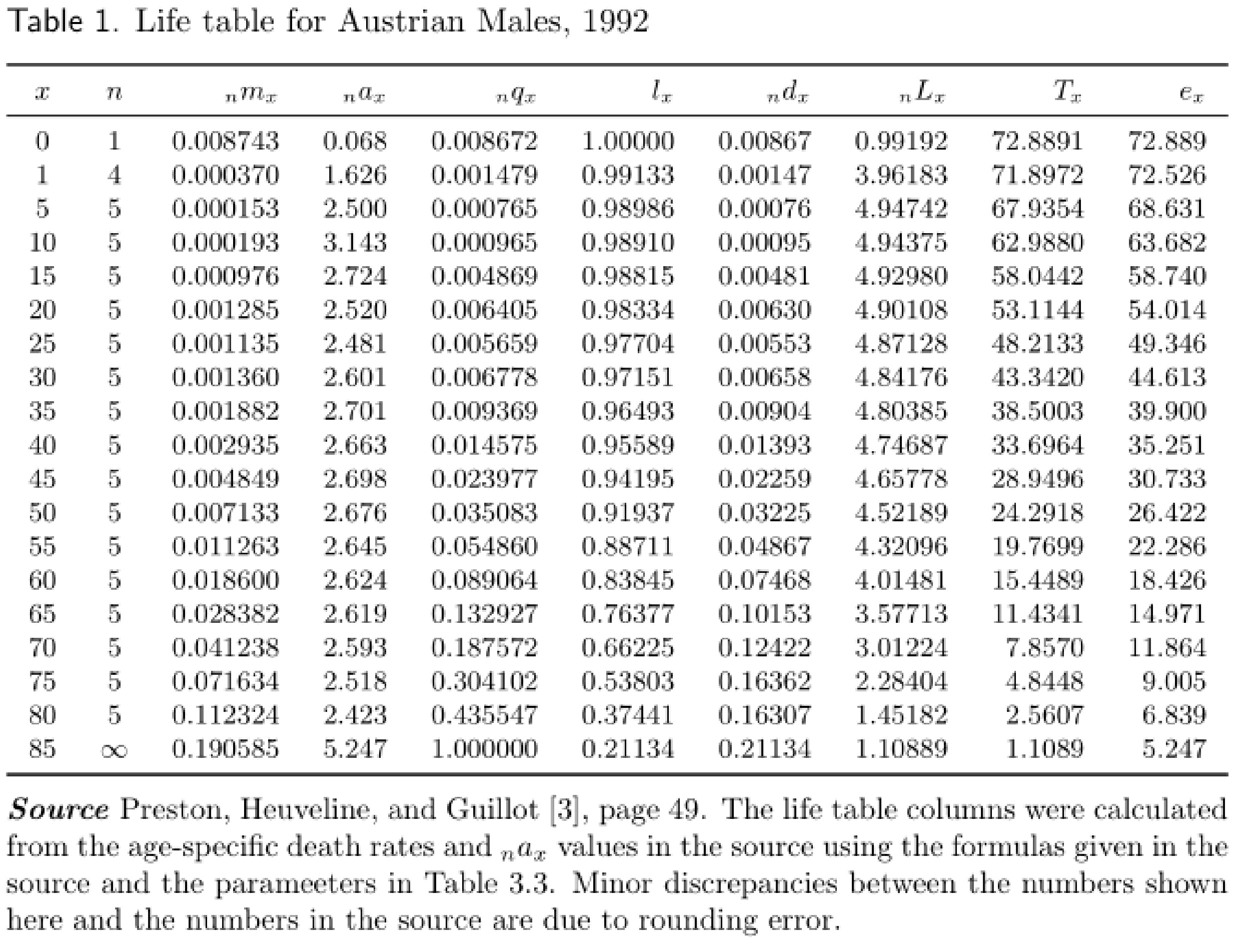


Table 1. Life table for Austrian Males, 1992

x	n	${}_nm_x$	${}_na_x$	${}_nq_x$	l_x	${}_nd_x$	${}_nL_x$	T_x	e_x
0	1	0.008743	0.068	0.008672	1.00000	0.00867	0.99192	72.8891	72.889
1	4	0.000370	1.626	0.001479	0.99133	0.00147	3.96183	71.8972	72.526
5	5	0.000153	2.500	0.000765	0.98986	0.00076	4.94742	67.9354	68.631
10	5	0.000193	3.143	0.000965	0.98910	0.00095	4.94375	62.9880	63.682
15	5	0.000976	2.724	0.004869	0.98815	0.00481	4.92980	58.0442	58.740
20	5	0.001285	2.520	0.006405	0.98334	0.00630	4.90108	53.1144	54.014
25	5	0.001135	2.481	0.005659	0.97704	0.00553	4.87128	48.2133	49.346
30	5	0.001360	2.601	0.006778	0.97151	0.00658	4.84176	43.3420	44.613
35	5	0.001882	2.701	0.009369	0.96493	0.00904	4.80385	38.5003	39.900
40	5	0.002935	2.663	0.014575	0.95589	0.01393	4.74687	33.6964	35.251
45	5	0.004849	2.698	0.023977	0.94195	0.02259	4.65778	28.9496	30.733
50	5	0.007133	2.676	0.035083	0.91937	0.03225	4.52189	24.2918	26.422
55	5	0.011263	2.645	0.054860	0.88711	0.04867	4.32096	19.7699	22.286
60	5	0.018600	2.624	0.089064	0.83845	0.07468	4.01481	15.4489	18.426
65	5	0.028382	2.619	0.132927	0.76377	0.10153	3.57713	11.4341	14.971
70	5	0.041238	2.593	0.187572	0.66225	0.12422	3.01224	7.8570	11.864
75	5	0.071634	2.518	0.304102	0.53803	0.16362	2.28404	4.8448	9.005
80	5	0.112324	2.423	0.435547	0.37441	0.16307	1.45182	2.5607	6.839
85	∞	0.190585	5.247	1.000000	0.21134	0.21134	1.10889	1.1089	5.247

Source Preston, Heuveline, and Guillot [3], page 49. The life table columns were calculated from the age-specific death rates and ${}_na_x$ values in the source using the formulas given in the source and the parameeters in Table 3.3. Minor discrepancies between the numbers shown here and the numbers in the source are due to rounding error.

THE LEFT-MOST COLUMN SPECIFIES THE AGE, AND THE SUCCESSIVE COLUMNS TO THE RIGHT SPECIFY A SELECTION OF THE FOLLOWING DESCRIPTORS. (THE DESCRIPTORS DEFINED HERE ARE NOT INDEPENDENT: MANY OF THE DESCRIPTORS ARE DEFINED IN TERMS OF OTHER DESCRIPTORS.)

Starting age for age group: x

Width of the age interval: n

Cohort survivors (number of survivors at age x): ℓ_x

Cohort members at age zero (initial size, the “radix” of the table): ℓ_0

Cohort deaths (number of persons dying in the interval x to x + n):

$$ {}_n d_x = \ell_x - \ell_{x+n} $$

Probability of dying in the interval from x (inclusive) to n (exclusive):

$$ {}_n q_x = \quad {}_n d_x / \ell_x = 1 - (\ell_{x+n} / \ell_x) $$

Probability of surviving the interval from x (inclusive) to n (exclusive) (i.e., of surviving from age x to age x+n):

$$ 1 - \quad {}_n q_x = 1 - \quad {}_n d_x / \ell_x = (\ell_{x+n} / \ell_x) $$

Cohort survivors at age x+n:

$$ \ell_{x+n} = (1 - \quad {}_n q_x) \ell_x $$

Hazard rate (minus the slope of the logarithm of the number of cohort survivors as a function of age):

$$ \ell_{x+n} = \ell_x e^{-n \quad {}_n h_x} $$

or

$$_n h_x = -\frac{ln\left(\frac{\ell_{x+n}}{\ell_x}\right)}{n} = -\frac{ln(\ell_{x+n}) - ln\ (\ell_x)}{n}$$

(Note: The hazard rate depends on n, since it is the slope from age x to age x+n. In Part 2 of the presentation, we will discuss the hazard function, which is the limit of the hazard rate as n approaches zero.)

Expected time lived for persons dying in the interval from x to x+n:

$$_n a_x$$

Cohort person-years lived (CPYL) between ages of x and x+n:

$$_n L_x = (n)\ell_{x+n} + \quad _n a_x \quad _n d_x$$

(The value of $_n a_x$ is close to n/2. Estimation of $_n a_x$ involves substantial effort, and a data-based statistical estimate is generally not available. In this case, it is approximated by the value n/2, in which case

$$_n L_x = \left(\frac{n}{2}\right)(\ell_x + \ell_{x+n}).$$

Since $_n a_x$ is often approximated by n/2, measures that depend on it, such as the ones that follow, are also based on this approximation.)

Life table death rate:

$$_n m_x = \quad _n d_x / \quad _n L_x$$

Person-years of life remaining:

$$T_x = \sum_{i=0}^{\infty} \quad _n L_{x+in}$$

Expectation of life beyond age x (this function is not necessarily a decreasing function of x):

$$e_x = T_x/\ell_x$$

Expectation of age at time of death (this function is an increasing function of x):

$$x + e_x$$

A LIFE TABLE IS A DETAILED DESCRIPTION OF THE AGE-RELATED MORTALITY CHARACTERISTICS OF A POPULATION THAT EXPERIENCES NO MIGRATION. IT IS USEFUL FOR INSURANCE CALCULATIONS AND FOR MAKING POPULATION PROJECTIONS (AMONG OTHER THINGS).

A LIFE TABLE IS A COMPLICATED, CUMBERSOME NONPARAMETRIC DESCRIPTION OF MORTALITY, SINCE IT SPECIFIES VALUES OF DESCRIPTORS FOR MANY AGES. TO PROMOTE UNDERSTANDING IT IS OFTEN DESIRABLE TO CONSIDER SIMPLER PARAMETRIC DESCRIPTORS OF MORTALITY. EARLY PARAMETRIC DESCRIPTORS OF THE HAZARD FUNCTION INCLUDE THE GOMPERTZ CURVE AND THE MAKEHAM CURVE. THE GOMPERTZ CURVE IS A STRAIGHT LINE DRAWN THROUGH A PLOT OF THE LOGARITHM OF THE HAZARD FUNCTION. THERE ARE BETTER DESCRIPTORS OF MORTALITY THAN THE GOMPERZ AND MAKEHAM CURVES. DISCUSSION OF THEM IS DEFERRED TO PART 2 OF THE PRESENTATION, WHICH CONSIDERS PROBABILITY MODELS RELATING TO MORTALITY.

COHORT-BASED FERTILITY MEASURES

COHORT-BASED MEASURES RELATED TO FERTILITY INCLUDE:

Age-specific fertility rate:

ASFR = (NUMBER OF CHILDREN BORN TO WOMEN OF THE COHORT BETWEEN AGES x AND x+n) / (PERSON-YEARS LIVED BY WOMEN IN THE COHORT BETWEEN AGES x AND x+n)

$$ASFR = \quad {}_n f_x$$

Net reproduction rate:

NRR = (NUMBER OF DAUGHTERS BORN TO WOMEN OF THE COHORT BETWEEN AGES x AND x+n) / (NUMBER OF WOMEN IN THE COHORT BETWEEN AGES x AND x+n)

= (NUMBER OF BABIES BORN TO WOMEN OF THE COHORT BETWEEN AGES x AND x+n) (FRACTION FEMALE AT BIRTH) / (NUMBER OF WOMEN IN THE COHORT BETWEEN AGES x AND x+n)

AGE-BASED FORMULA FOR THE NET REPRODUCTION RATE:

$$NRR = \sum_x \ {}_nf_x \ \ {}_nL_x f_{fab} / \ell_0$$

WHERE ${}_nf_x$ DENOTES THE FERTILITY RATE FOR WOMEN AGED x TO x+n, AND f_{fab} DENOTES THE FRACTION FEMALE AT BIRTH.

GROSS REPRODUCTION RATE CALCULATED FROM AGE-SPECIFIC FERTILITY RATES, IGNORING EFFECTS OF MORTALITY (REPLACE ${}_nL_x/\ell_0$ BY n IN FORMULA FOR NRR)

$$GRR = \sum_x \ {}_nf_x \ \ {}_n(n) f_{fab}$$

TOTAL FERTILITY RATE CALCULATED FROM AGE-SPECIFIC FERTILITY RATES, IGNORING EFFECTS OF MORTALITY:

$$TFR = \sum_x \ {}_nf_x \ \ {}_n(n)$$

10. POPULATION PROJECTIONS AND FORECASTS

DEMOGRAPHIC ANALYSIS HAS A VARIETY OF USES, SUCH AS ASSISTING THE PRICING OF INSURANCE PREMIUMS, DETERMINING THE NUMBER OF REPRESENTATIVES IN A LEGISLATURE, AND ALLOCATING FEDERAL GOVERNMENT FUNDS TO STATES. ONE OF THE MAJOR USES OF DEMOGRAPHIC ANALYSIS IS TO ESTIMATE THE SIZE, COMPOSITION AND DISTRIBUTION OF POPULATION IN THE

FUTURE. THAT INFORMATION IS VERY USEFUL FOR ASSISTING THE ESTIMATION OF POPULATION-RELATED PHENOMENA SUCH AS DEMAND FOR INFRASTRUCTURE, GOVERNMENT SERVICES, WELFARE BUDGETS, AND CONSUMER PRODUCTS AND SERVICES.

AN ESTIMATE OF A FUTURE POPULATION IS BASED ON ASSUMPTIONS ABOUT THE FUTURE DYNAMICS OF THE POPULATION. POPULATION DYNAMICS MAY BE SPECIFIED IN TERMS OF A NUMBER OF PARAMETERS, SUCH AS OVERALL GROWTH RATES OR SPECIFIC VARIABLES SUCH AS MORTALITY RATES, FERTILITY RATES AND MIGRATION RATES.

ESTIMATES OF FUTURE POPULATION ARE CLASSIFIED INTO TWO CATEGORIES:

- *PROJECTIONS*, WHICH ARE ESTIMATES OF FUTURE POPULATION STATUS (NUMBERS, DISTRIBUTION, COMPOSITION) WITH NO ASSESSMENT OF THE STATISTICAL PROPERTIES OF THE ESTIMATES; AND
- *FORECASTS*, WHICH ARE ESTIMATES OF FUTURE POPULATION STATUS TOGETHER WITH SOME SORT OF QUANTITATIVE ASSESSMENT OF THE ACCURACY OF THE ESTIMATE (CLOSENESS TO THE TRUE VALUE), SUCH AS A DESCRIPTION OF THE STATISTICAL PROPERTIES OF THE ESTIMATES (SUCH AS STANDARD ERRORS, CONFIDENCE INTERVALS, LIKELIHOOD FUNCTIONS, OR THE MEAN-SQUARED-ERROR OF PREDICTION).

PROJECTIONS AND FORECASTS ARE CONDITIONED ON VALUES OF VARIABLES OR PARAMETERS THAT AFFECT POPULATION DYNAMICS. THE CONDITIONING VARIABLES MAY BE AS SIMPLE AS THE MOST RECENTLY OBSERVED POPULATION SIZE AND GROWTH RATE, OR THEY MAY BE HIGHLY DETAILED DESCRIPTIONS OF THE CURRENT POPULATION STATUS AND CONSIDERED ESTIMATES OF THE PRESENT AND FUTURE VALUES OF VARIABLES THAT AFFECT POPULATION DYNAMICS, SUCH AS AGE-SPECIFIC MORTALITY RATES, FERTILITY RATES, AND MIGRATION RATES.

FOR FORECASTS, IT IS DESIRED THAT THE ESTIMATE BE IN SOME SENSE CLOSE TO THE VALUE BEING ESTIMATED. TYPICALLY, FORECASTS ARE CONDITIONED ON “BEST ESTIMATES” OF THE VALUES OF THE CONDITIONING VARIABLES OR PARAMETERS. PROJECTIONS MAY BE CONDITIONED ON ANY VALUES OF THE CONDITIONING VARIABLES OR PARAMETERS, INDEPENDENT OF THE LIKELIHOOD

OF THOSE VALUES. FOR EXAMPLE, A PROJECTION MAY BE MADE FOR EXTREME VALUES OF A CERTAIN PARAMETER, TO ILLUSTRATE A LIMITING CASES OR SENSITIVITY.

STATISTICAL FORECASTS ARE BASED ON STATISTICAL MODELS OF POPULATION, VIEWED AS A STOCHASTIC PROCESS. PROJECTIONS MAY BE BASED ON STATISTICAL MODELS, BUT MORE OFTEN THEY ARE BASED ON SIMPLE DETERMINISTIC MODELS. PART 2 OF THIS PRESENTATION DEALS WITH FORECASTS BASED ON STATISTICAL MODELS AND STATISTICAL ANALYSIS.

THIS PART OF THE PRESENTATION DEALS WITH SIMPLE DETERMINISTIC MODELS. THEY COULD BE DESCRIBED AS “EXPECTED VALUE” MODELS AND ANALYSIS. IMPLEMENTATION OF THEM INVOLVES ARITHMETIC AND ALGEBRA, AND SIMPLE MATRIX ARITHMETIC, BUT NO CALCULUS, ADVANCED LINEAR ALGEBRA, OR STATISTICS.

THE POPULATION ESTIMATES CONSIDERED IN THIS PART OF THE PRESENTATION INCLUDE BOTH PROJECTIONS AND FORECASTS. IN THIS PART, HOWEVER, THE FORECASTS WILL BE VERY SIMPLE ONES, SUCH AS PROJECTIONS USING “BEST ESTIMATES” OF POPULATION GROWTH PARAMETERS, WITH LIMITED ASSESSMENT OF FORECAST ACCURACY (SUCH AS GENERATION OF A SET OF PROJECTIONS OVER A RANGE OF PARAMETER VALUES THAT ARE CONSIDERED LIKELY).

POPULATION PROJECTIONS ARE USEFUL FOR A NUMBER OF PURPOSES, INCLUDING:

- TEACHING, TO ASSIST UNDERSTANDING OF POPULATION DYNAMICS;
- CONDUCTING APPROXIMATE SENSITIVITY ANALYSES, SUCH AS CONSTRUCTING A ROUGH ASSESSMENT OF THE EFFECT OF A PARTICULAR VARIABLE ON POPULATION STATUS;
- COMPONENTS OF SIMULATION STUDIES;
- COMPONENTS OF STATISTICAL MODELS (SUCH AS A TRANSITION MATRIX IN A KALMAN FILTER).

POPULATION PROJECTIONS, BY THEMSELVES, ARE NOT VERY USEFUL AS BASES FOR DECISIONS. FOR DECISIONS, IT IS IMPORTANT TO ASSESS THE LIKELIHOOD

OF FUTURE ESTIMATES. TO DO THAT IN A REASONABLE FASHION, IT IS NECESSARY TO HAVE A POPULATION FORECAST.

PART 1 OF THIS PRESENTATION DISCUSSES POPULATION PROJECTIONS, AND PART 2 DISCUSSES POPULATION FORECASTS.

WE SHALL NOW DESCRIBE METHODS FOR MAKING POPULATION PROJECTIONS.

ALTERNATIVE PROJECTION METHODS

PROJECTIONS ARE INVARIABLY BASED ON ASSUMPTIONS ABOUT THE FUTURE DETERMINANTS OF POPULATION GROWTH. THESE ASSUMPTIONS MAY BE AS SIMPLE AS "CURRENT GROWTH RATES CONTINUE" TO ELABORATE SPECIFICATION ABOUT FUTURE MORTALITY RATES, FERTILITY RATES, AND MIGRATION, BY YEAR OR OTHER TIME INTERVAL.

IN ORDER TO MAKE A PROJECTION CONDITIONAL ON ONE OR MORE VARIABLES, IT IS NECESSARY TO HAVE AVAILABLE A MATHEMATICAL MODEL THAT SPECIFIES THE RELATIONSHIP OF POPULATION TO THE VARIABLES AND TO ONE OR MORE CONSTANTS. THE CONDITIONING VARIABLES ARE REFERRED TO USING VARIOUS NAMES, SUCH AS EXPLANATORY VARIABLES, CONTROL VARIABLES OR COVARIATES. THE CONSTANTS ARE REFERRED TO AS MODEL PARAMETERS.

IN THIS PART OF THE PRESENTATION, WE HAVE DISCUSSED SIMPLE GROWTH MODELS, WHICH SPECIFY FUTURE POPULATION AS A FUNCTION OF CURRENT POPULATION AND A GROWTH RATE; AND WE HAVE DISCUSSED THE BALANCING EQUATION, WHICH SPECIFIES GROWTH AS A FUNCTION OF MORTALITY, FERTILITY AND MIGRATION.

TO MAKE PROJECTIONS USING THE GROWTH MODEL, ALL THAT IS REQUIRED IS A VALUE FOR THE (EXPONENTIAL) GROWTH RATE, AND THIS MAY BE ESTIMATED FROM RECENT DATA.

TO MAKE PROJECTIONS FOR THE BALANCING EQUATION MODEL, IT IS NECESSARY TO SPECIFY VALUES FOR BIRTHS, DEATHS, IMMIGRATION AND EMIGRATION FOR FUTURE TIMES. THESE QUANTITIES ARE FUNCTIONS OF AGE, AND THEIR VALUES ARE ESTIMATED USING THE COHORT-BASED (AGE-SPECIFIC) MORTALITY

ESTIMATES FROM A LIFE TABLE, COHORT-BASED (AGE-SPECIFIC) FERTILITY ESTIMATES, AND MIGRATION RATES FROM RECENT HISTORY OR PLANNED MIGRATION POLICY.

BIRTHS, DEATHS AND MIGRATION ARE CALLED THE "COMPONENTS" OF POPULATION, AND A PROJECTION METHOD THAT IS BASED ON PROJECTING POPULATION BASED ON COHORT COMPONENTS IS CALLED A "COHORT-COMPONENT PROJECTION MODEL."

WE SHALL NOW DISCUSS THESE TWO PROJECTION METHODS.

EXPONENTIAL GROWTH POPULATION PROJECTION MODEL

AS MENTIONED EARLIER, EXPONENTIAL GROWTH IS EXPLOSIVE. IT CANNOT CONTINUE FOR VERY LONG WITHOUT REACHING ASTRONOMICALLY HIGH LEVELS. THE USE OF AN EXPONENTIAL GROWTH PROJECTION MODEL MAY BE USED TO FORECAST POPULATION LEVELS A SHORT TIME INTO THE FUTURE, BUT IT IS NOT AT ALL REASONABLE FOR PREDICTING POPULATION LEVELS VERY FAR INTO THE FUTURE. FOR LONGER-TERM PROJECTIONS, IT SHOWS HOW QUICKLY EXTREMELY HIGH POPULATION LEVELS WILL BE REACHED, IF IT CONTINUES.

AS AN EXAMPLE, CONSIDER THE CASE OF MEXICO. THE POPULATION OF MEXICO IN 1970 WAS 50 MILLION, WITH A GROWTH RATE OF 3.5 PERCENT. THE FORMULA FOR EXPONENTIAL GROWTH IS

$$P_t = P_0 \, e^{rt},$$

WHERE P_t DENOTES POPULATION SIZE AT TIME t AND r DENOTES THE GROWTH RATE, IN THIS CASE, $r = .035$. TO PROJECT THE POPULATION SIZE OF MEXICO FROM 1970 TO 2100 USING THIS FORMULA, WE SUBSTITUTE $r = .035$, $P_0 = 50{,}000{,}000$, AND $t = 2100 – 1970 = 130$ INTO THE FORMULA. WE OBTAIN

$$P_{2100} = P_{1970} \, e^{\,.035(130)} = 50{,}000{,}000\ (94.63240831) = 4{,}731{,}620{,}416.$$

THIS NUMBER, ALMOST FIVE BILLION, IS MORE THAN HALF THE PRESENT POPULATION OF EARTH. GIVEN THE AMOUNT OF ARABLE LAND THAT MEXICO POSSESSES, IT IS AN ABSURDLY LARGE NUMBER. THIS SIMPLE EXAMPLE, IF

WORKED OUT IN 1970, WOULD SHOW THE IMPOSSIBILITY OF CONTINUING TO GROW AT AN EXPONENTIAL GROWTH RATE OF 3.5 PERCENT PER YEAR.

THIS RESULT WOULD BEG THE QUESTION OF WHAT THE POPULATION SIZE WOULD BE IF THE GROWTH RATE WERE REDUCED IN FUTURE YEARS. PROJECTIONS COULD BE CONSTRUCTED FOR VARIOUS SCENARIOS ABOUT THE VALUES OF FUTURE GROWTH RATES, BUT THEY WOULD BE OF LIMITED USE. THE PROBLEM WITH THE SIMPLE GROWTH MODEL IS THAT IT DOES NOT DEPEND ON PHYSICALLY MEANINGFUL VARIABLES OR PARAMETERS, SUCH AS BIRTH RATES, DEATH RATES AND MIGRATION RATES. ESTIMATES OF FUTURE POPULATION SIZES BASED ON ASSUMPTIONS ABOUT THESE RATES ARE EASIER TO COMPREHEND THAN ESTIMATES BASED ON OVERALL GROWTH RATES. TO DO THIS, WE NEED A MODEL THAT EMBODIES A RELATIONSHIP OF POPULATION GROWTH TO THESE VARIABLES / PARAMETERS. WE SHALL NOW DISCUSS ONE SUCH MODEL, THE COHORT-COMPONENT PROJECTION MODEL.

NOTE THAT, IN MAKING POPULATION PROJECTIONS OR FORECASTS, THERE IS LITTLE POINT TO CONDITIONING ON VARIABLES THAT CANNOT BE FORECAST, INFLUENCED OR CONTROLLED. THE REASON FOR THIS IS THAT IN MAKING A PROJECTION THAT IS CONDITIONAL ON ONE OR MORE VARIABLES, IT IS NECESSARY TO SPECIFY THE FUTURE VALUES OF THESE VARIABLES.

COHORT-COMPONENT POPULATION PROJECTION MODEL

THE COMPONENTS OF POPULATION GROWTH, DEPICTED IN THE DEMOGRAPHIC BALANCING EQUATION, ARE BIRTHS, DEATHS AND MIGRATION. PREVIOUSLY, IT WAS MENTIONED THAT THERE IS A STRONG RELATIONSHIP OF BIRTH AND DEATH RATES TO AGE. THE COHORT-COMPONENT PROJECTION MODEL PROJECTS POPULATION BY ESTIMATING THE SIZE OF AN AGE COHORT POPULATION ENTERING AND LEAVING A PERIOD OF INTEREST, SUBTRACTING DEATHS OF COHORT MEMBERS, ADDING BIRTHS ASSOCIATED WITH COHORT MEMBERS, ADDING IMMIGRANTS IN THE SAME AGE GROUP AS THE COHORT, AND SUBTRACTING EMIGRANTS FOR THE SAME AGE GROUP AS THE COHORT. THE METHOD USES COHORT-SPECIFIC INFORMATION ABOUT THE COMPONENTS, HENCE THE NAME “COHORT-COMPONENT.”

A POPULATION PROJECTION STARTS FROM A POPULATION SPECIFIED AT A POINT IN TIME. IN TYPICAL APPLICATIONS THE STARTING POINT FOR THE PROJECTION IS THE LATEST TIME FOR WHICH A REASONABLE ESTIMATE OF POPULATION IS AVAILABLE (SUCH AS THE LATEST CENSUS). TO IMPLEMENT THE COHORT-COMPONENT METHOD, IT IS NECESSARY THAT THE POPULATION COUNTS BE DISAGGREGATED BY AGE AND SEX. MORTALITY DATA ARE PROVIDED BY A LIFE TABLE, AND IT IS ASSUMED THAT FERTILITY AND MIGRATION DATA ARE AVAILABLE FOR THE SAME AGE-BY-SEX CATEGORIES AS THE LIFE TABLE.

THE BASIS FOR A COHORT-COMPONENT POPULATION PROJECTION IS THE DEMOGRAPHIC BALANCING EQUATION, WHICH WE DENOTED EARLIER AS:

$$P_{end} = P_{beg} + B – D + I – E,$$

WHERE P_{end} DENOTES THE POPULATION AT THE END OF A TIME PERIOD, P_{beg} DENOTES THE POPULATION AT THE START OF THE TIME PERIOD, B DENOTES BIRTHS DURING THE PERIOD, D DENOTES DEATHS DURING THE PERIOD, I DENOTES IMMIGRANTS DURING THE PERIOD, AND E DENOTES EMIGRANTS DURING THE PERIOD.

FOR THE COHORT-COMPONENT METHOD, THE BALANCING EQUATION IS SPECIFIED FOR EACH AGE-BY-SEX CATEGORY OF THE LIFE TABLE ON WHICH THE PROJECTION IS TO BE BASED. WE SHALL USE x TO DENOTE AGE CATEGORY, AND f AND m TO DENOTE SEX (FEMALE AND MALE).

LET US DENOTE, FOR THE MOMENT, THE START TIME OF THE PERIOD AS t, AND THE END TIME AS t+n.

TO IMPLEMENT THE COHORT-COMPONENT POPULATION PROJECTION METHOD, IT IS NECESSARY TO ESTIMATE THE NUMBER OF BIRTHS, DEATHS, IMMIGRANTS AND EMIGRANTS FOR A TIME PERIOD OF INTEREST.

SINCE BIRTHS COME ONLY FROM FEMALES, THE COHORT-COMPONENT METHOD IS IMPLEMENTED BY PROJECTING THE FEMALE POPULATION, AND ESTIMATING THE PROJECTED MALE POPULATION FROM THAT.

IGNORING MIGRATION FOR THE MOMENT, THE PARAMETERS NEEDED TO MAKE A COHORT-COMPONENT PROJECTION ARE, WHERE x DENOTES AGE AND n DENOTES THE TIME-LENGTH OF THE AGE CATEGORY (BOTH FOR THE COHORT AND THE PERIOD):

- THE PROBABILITY THAT A FEMALE ALIVE AT AGE x DIES BEFORE AGE x + n
- THE EXPECTED NUMBER OF DAUGHTERS AGED 0 TO n AT THE END OF THE PROJECTION STEP PER FEMALE AGED x TO x + n AT THE BEGINNING OF THE PROJECTION STEP

WE BEGIN BY CONSIDERING DEATHS.

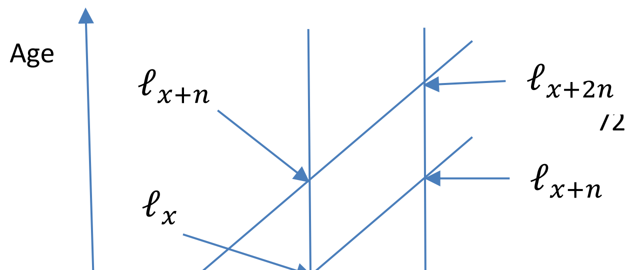
FOR REASONS THAT WILL BECOME CLEAR LATER, IT IS MORE CONVENIENT TO WORK IN TERMS OF SURVIVORS (AND SURVIVAL RATES AND PROBABILITIES) THAN IN TERMS OF DEATHS (AND MORTALITY RATES AND PROBABILITIES).

ESTIMATION OF SURVIVORS FOR A PERIOD

PROJECTION OF THE PROPORTION OF A COHORT THAT SURVIVES A SPECIFIED AGE INTERVAL IS SIMPLE: THE LIFE TABLE SPECIFIES THE PROPORTION OF THE POPULATION REMAINING AT EACH AGE SPECIFIED IN THE TABLE. TO PROJECT THE PROPORTION OF A COHORT THAT SURVIVES A PERIOD OF TIME IS MORE COMPLICATED, HOWEVER, SINCE THE COHORT MEMBERS ARE OF VARYING AGES, BOTH AT THE START AND THE END OF THE PERIOD.

THE WAY THAT A COHORT-COMPONENT PROJECTION OF SURVIVORS IS MADE IS TO ESTIMATE, FOR EACH AGE COHORT, THE NUMBER OF COHORT MEMBERS ENTERING THE PERIOD OF INTEREST AND THE NUMBER OF COHORT MEMBERS SURVIVING UNTIL THE END OF THE PERIOD. THIS ESTIMATION IS DONE FOR EACH AGE CATEGORY, x TO x+n, WHERE x DENOTES AGE AND n DENOTES THE AGE-LENGTH OF THE AGE CATEGORY. THESE NUMBERS ARE ESTIMATED FROM THE DATA IN A SPECIFIED LIFE TABLE. A GRAPHICAL REPRESENTATION OF THE PROCESS IS ILLUSTRATED IN FIGURE 9.

Cohort-Component Population-Projection Method



FOR SIMPLICITY, THE COHORT-COMPONENT PROJECTION METHOD IS IMPLEMENTED BY MAKING PROJECTIONS FOR PERIODS THAT ARE OF THE SAME TIME DURATION AS THE AGE-INTERVAL OF THE COHORT (E.G., 1, 4, OR 5 YEARS), AS ILLUSTRATED IN FIGURE 9.

WE HAVE DENOTED THE INITIAL TIME VALUE OF THE PERIOD AS t. IN THE LEXIS DIAGRAM, BOTH AGE (DENOTED BY x) AND TIME (DENOTED BY t) ARE CONSIDERED TO "MOVE" AT THE SAME RATE (I.E., DIFFER BY JUST A CONSTANT). AT THE BEGINNING OF THE PERIOD, THE YOUNGEST MEMBER OF THE COHORT IS OF AGE x, AND THE OLDEST MEMBER IS OF AGE x+n.

NOTE THAT THE AGES OF THE COHORT MEMBERS AT THE BEGINNING OF THE PERIOD ARE IN THE RANGE x TO x+n, AND THE AGES AT THE END ARE IN THE RANGE x+n TO x+2n. WE REFER TO THIS SEGMENT OF THE POPULATION IN A PERIOD AS AN "AGE GROUP."

THE LIFE TABLE SPECIFIES THE PROBABILITY THAT A PERSON ALIVE AT AGE x DIES BY AGE x + n. TO MAKE A PROJECTION FOR THE POPULATION IN AGE CATEGORY x TO x+n, IT IS DESIRED TO KNOW THE PROBABILITY THAT A PERSON OF AGE x TO x+n WHO IS ALIVE AT THE BEGINNING OF THE PERIOD DIES BEFORE THE END OF THE PERIOD. THE COMPLICATION THAT ARISES IS THAT THE INDIVIDUALS ALIVE AT THE BEGINNING OF THE PERIOD ARE OF VARYING AGES (NOT OF A SPECIFIED AGE, AS IN THE LIFE TABLE). SINCE THE AGE DISTRIBUTION IS TYPICALLY UNKNOWN, THE DESIRED PROBABILITY IS APPROXIMATED, IN THIS CASE BY LINEAR INTERPOLATION.

FROM THIS POINT ON, IT IS CONVENIENT TO WORK IN TERMS OF THE PROBABILITY OF SURVIVAL, RATHER THAN THE PROBABILITY OF DEATH.

THE DESIRED PROBABILITY IS APPROXIMATED BY COMPARING THE NUMBER OF PERSONS IN THE COHORT WHO ARE ALIVE AT THE END OF THE PERIOD TO THE NUMBER OF PERSONS IN THE COHORT WHO ARE ALIVE AT THE BEGINNING OF THE PERIOD. LET US DENOTE THE TIME OF THE BEGINNING OF THE PERIOD AS TIME t, AND THE TIME OF THE END OF THE PERIOD AS TIME t+n.

LET US CONSIDER ALL OF THE COHORT MEMBERS OF AGE x THROUGH x+n WHO ARE ALIVE AT TIME t. IT IS DESIRED TO KNOW HOW MANY OF THIS GROUP ARE ALIVE AT TIME t. FROM FIGURE 9, IT IS SEEN THAT THE NUMBER ALIVE VARIES FROM ℓ_x TO ℓ_{x+n}. INTERPOLATING LINEARLY IN THE LIFE TABLE, THIS IMPLIES THAT THE (APPROXIMATE) NUMBER ALIVE AT TIME t IS

$$(5/2)(\ell_x + \ell_{x+n}).$$

SIMILARLY, CONSIDER THE COHORT MEMBERS OF AGE x+n THROUGH x+2n WHO ARE ALIVE AT TIME t+n. FROM FIGURE 9, IT IS SEEN THAT THE NUMBER ALIVE VARIES FROM ℓ_{x+n} TO ℓ_{x+2n}. INTERPOLATING LINEARLY IN THE LIFE TABLE, THIS IMPLIES THAT THE NUMBER ALIVE AT TIME t+n IS

$$(5/2)(\ell_{x+n} + \ell_{x+2n}).$$

AN ESTIMATE OF THE PROBABILITY THAT A MEMBER OF THE PERIOD WHO IS ALIVE AT THE BEGINNING OF THE PERIOD (TIME t) IS STILL ALIVE AT THE END OF THE PERIOD (TIME t+n) IS GIVEN BY THE RATIO OF THESE TWO QUANTITITES:

$$\frac{(\ell_{x+n} + \ell_{x+2n})}{(\ell_x + \ell_{x+n})}.$$

AN ESTIMATE FOR THE PROBABILITY OF SURVIVAL (FOR AN INDIVIDUAL IN THE AGE-BY-PERIOD CATEGORY) THAT TAKES INTO ACCOUNT THE DISTRIBUTION OF DEATHS WITHIN THE PERIOD IS:

$$\frac{{}_nL_{x+n}}{{}_nL_x}.$$

A SIGNIFICANT SHORTCOMING OF THE SECOND ESTIMATE IS THAT IT DEPENDS ON THE VALUE OF ${}_na_x$, WHICH IS NOT AVAILABLE IN MANY LIFE TABLES (SUCH AS IN DEVELOPING COUNTRIES OR SUB-COUNTRY REGIONS). IF A DATA-BASED ESTIMATE IS NOT AVAILABLE FOR ${}_na_x$, THEN IT IS ESTIMATED AS n/2, IN WHICH CASE THE FORMULA FOR ${}_nL_x$ REDUCES TO

$${}_nL_x = (n)(\ell_{x+n}) + ({}_na_x)({}_nd_x) = (n/2)(\ell_x + \ell_{x+n}),$$

WHICH IS THE FIRST FORMULA PRESENTED.

IN MANY INSTANCES, THE ADDITIONAL PARAMETER ($_na_x$) REQUIRED BY THE MORE COMPLICATED FORMULA PRESENTED ABOVE WOULD NOT BE AVAILABLE. FOR PROJECTION PURPOSES, THIS DOES NOT MATTER. A PROJECTION MAY BE CONDITIONED ON VALUES OF $_nL_x$ OR ON VALUES OF ℓ_{x+n}. A POPULATION PROJECTION IS SIMPLY A CALCULATION, ACCORDING TO A REASONABLE FORMULA, OF FUTURE POPULATON UNDER SPECIFIED CONDITIONS. FOR PROJECTION PURPOSES, IT IS OF NO PRACTICAL SIGNIFICANCE WHETHER THE PROJECTION IS CONDITIONAL ON VALUES OF $_nL_x$ OR ON VALUES OF ℓ_{x+n}. FOR POPULATION *FORECASTS*, WHICH ARE CONSIDERED IN PART 2 OF THE PRESENTATION, THIS COULD BE OF SIGNIFICANCE, BUT, HERE, IT IS NOT.

NOTE THAT, WHILE SMALL DIFFERENCES IN RATES CAN TRANSLATE INTO LARGE DIFFERENCES IN POPULATION PROJECTIONS IN THE DISTANT FUTURE, THE EFFECT OF CHOOSING BETWEEN THE TWO FORMULAS (THAT IS, BETWEEN CONDITIONING ON VALUES OF $_nL_x$ OR ON VALUES OF ℓ_{x+n}) WOULD TYPICALLY PALE IN COMPARISON TO THE EFFECT ASSOCIATED WITH VARIATION IN THE LEVELS OF VALUES OF EITHER PARAMETER.

FOR THE DETERMINISTIC MODELS CONSIDERED IN THIS PART OF THE PRESENTATION, THE EFFECTS OF VARIATIONS IN THE VALUES OF MODEL PARAMETERS ON PROJECTIONS WOULD BE ASSESSED BY A “PERTURBATION ANALYSIS” OR “SENSITIVITY ANALYSIS.” THE VARIATION IN FUTURE PROJECTIONS (REFERRED TO IN TERMS SUCH AS “LOW VARIANT,” “MEDIUM VARIANT,” AND “HIGH VARIANT”) WOULD TYPICALLY BE MUCH LARGER RELATIVE TO CHANGES IN LEVELS OF THE PARAMETER VALUES THAN RELATIVE TO THE CHOICE OF $_nL_x$ OR ℓ_{x+n} AS A CONDITIONING PARAMETER.

IN SUMMARY, FOR PROJECTION PURPOSES, IT DOES NOT MATTER WHICH OF THE TWO SIMILAR PARAMETERS IS CONDITIONED ON. THE MORE COMPLEX FORMULA IS OF THEORETICAL INTEREST, BUT EITHER FORMULA MAY BE USED IN A PROJECTION MODEL OR A SIMULATION MODEL.

LET US DENOTE THIS ESTIMATE OF THE PROBABILITY THAT A PERSON IN AGE GROUP x WHO IS ALIVE AT THE BEGINNING OF THE PERIOD SURVIVES UNTIL THE END OF THE PERIOD AS $_ns_x$ (“s” for “survival probability”).

IN THE FOLLOWING, WE SHALL NOT BE CONSIDERING n AS A VARIABLE – IT SHALL BE WHATEVER THE VALUE IS IN THE LIFE TABLE BEING USED, FOR THE AGE GROUP OF INTEREST. FROM THIS POINT OF VIEW, FOR THIS PRESENTATION THERE IS LITTLE NEED TO CARRY IT ALONG EXPLICITLY IN THE NOTATION, AND, FOR SIMPLICITY, WE SHALL DROP IT. IN THIS CASE, THE NOTATION $_ns_x$ REDUCES TO SIMPLY s_x.

ESTIMATION OF BIRTHS FROM COHORT MEMBERS, FOR A PERIOD

A ROUGH ESTIMATE OF THE NUMBER OF DAUGHTERS IS

$$(n)\quad {}_nf_xf_{fab}$$

WHERE $_nf_x$ DENOTES THE COHORT AGE-SPECIFIC FERTILITY RATE FOR WOMEN AGED x TO x+n, AND f_{fab} DENOTES THE FRACTION FEMALE AT BIRTH.

A FORMULA THAT TAKES INTO ACCOUNT DEATHS OF NEWBORNS AND THE CHANGING FERTILITY OF WOMEN IN THE PERIOD IS:

$$\frac{{}_nL_0}{2\ell_x}\left(\quad {}_nF_0 + \quad {}_nF_{x+1}\frac{{}_nL_{x+n}}{{}_nL_x}\right)f_{fab},$$

WHERE $_nF_x$ DENOTES THE PERIOD AGE-SPECIFIC FERTILTY RATE OF WOMEN AGED FROM x TO x+n OVER THE PERIOD.

DERIVATION OF THE PRECEDING FORMULA IS A LITTLE COMPLICATED, AND A DERIVATION WILL NOT BE PRESENTED HERE. (REFER TO A TEXT ON DEMOGRAPHIC ANALYSIS FOR A DERIVATION.)

WE ESTIMATE THE NUMBERS OF CHILDREN OF BOTH SEXES BY DIVIDING THIS QUANTITY BY f$_{fab}$. LET US DENOTE THIS QUANTITY AS $_nb_x$. THAT IS,

$${}_nb_x = \frac{{}_nL_0}{2\ell_x}\left(\quad {}_nF_0 + \quad {}_nF_{x+1}\frac{{}_nL_{x+n}}{{}_nL_x}\right).$$

FOR THE PURPOSE OF MAKING POPUATION PROJECTIONS, IT DOES NOT MATTER WHICH OF THE TWO PRECEDING FORMULAS IS USED. THE PROJECTION MAY BE CONDITIONED ON EITHER FORMULA (AND THE PARAMETER VALUES THAT DEFINE IT). THE REASON FOR THIS IS THE SAME AS IN THE EARLIER DISCUSSION OF CHOICE OF PARAMETERS AND FORMULAS FOR SPECIFYING MORTALITY.

SPECIFICATION OF THE COHORT-COMPONENT MODEL

WE NOW HAVE ESTIMATES OF THE PARAMETERS NEEDED TO IMPLEMENT THE BIRTH- AND DEATH-RELATED PARTS OF THE COHORT-COMPONENT PROJECTION METHOD.

AS WE DISCUSSED FOR MORTALITY, THERE IS LITTLE NEED IN THIS PRESENTATION TO EXPLICITLY REPRESENT n IN THE FORMULA, AND WE SHALL SIMPLY WRITE b_x IN PLACE OF $_nb_x$.

WE SHALL DENOTE THE STARTING TIME OF THE PROJECTION AS TIME t = 0. DENOTE THE POPULATION IN AGE GROUP x AT TIME t AS $P_t(x)$.

WE HAVE:

$$P_{t+1}(1) = \sum_{j=1}^{k} b_j P_t(j)$$

$$P_{t+1}(x) = s_{x-1} P_t(x-1), \qquad x = 2, \dots, k,$$

WHERE k DENOTES THE NUMBER OF AGE GROUPS.

THE PRECEDING FORMULAS MAY BE REPRESENTED COMPACTLY IN MATRIX FORM AS

$$\boldsymbol{P}_{t+1} = \boldsymbol{A}\boldsymbol{P}_t$$

WHERE

$$\boldsymbol{P}_t = (P_1, P_2, \dots, P_k)'$$

AND

$$\boldsymbol{A} = \begin{pmatrix} b_1 & b_2 & \cdots & b_{k-1} & b_k \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & s_{k-1} & 0 \end{pmatrix}.$$

THE MATRIX **A** IS CALLED A "LESLIE" MATRIX, AFTER THE MAN, P. H. LESLIE, WHO POPULARIZED ITS USE IN 1945.

IN GENERAL, THE POPULATION PROJECTION T STEPS INTO THE FUTURE IS

$$\boldsymbol{P}_{t+1} = \boldsymbol{A}^T \boldsymbol{P}_t.$$

THE MATRIX A IS CALLED THE POPULATION PROJECTION MATRIX. (HERE, T DENOTES EXPONENTIATION, NOT TRANSPOSITION.)

THERE ARE A NUMBER OF SIGNIFICANT ADVANTAGES TO REPRESENTING THE COHORT-COMPONENT PROJECTION METHOD IN MATRIX FORM. THESE INCLUDE

- TRANSPARENCY OF THE MODEL: THE MATRIX FORMULA IS SYMBOLICALLY MUCH SIMPLER, AND HENCE EASIER TO COMPREHEND, THAN A SET OF EQUATIONS
- IMPLEMENTATION OF THE MODEL: IN EQUATION FORM, CALCULATION OF A PROJECTION WOULD REQUIRE PROGRAMMING OF THE EQUATIONS; IN MATRIX FORM, THE PROJECTION IS CALCULATED USING MATRIX MULTIPLICATION (FOR WHICH PROGRAMS ARE READILY AVAILABLE (FROM *NUMERICAL RECIPES*, R, MATLAB, AND MANY OTHER SOURCES))
- MATHEMATICAL PROPERTIES: CERTAIN MATHEMATICAL PROPERTIES CAN BE INFERRED FROM THE MATRIX REPRESENTATION, WHICH ARE NOT AT ALL EVIDENT FROM THE EQUATION REPRESENTATION (THESE WILL BE DISCUSSED LATER)

THE PRECEDING DISCUSSION LIMITED THE MODEL TO THE CASE IN WHICH NO MIGRATION OCCURS. THE CASE IN WHICH MIGRATION MAY BE EXPRESSED AS A

FRACTION OF THE POPULATION IN EACH AGE CATEGORY IS READILY IMPLEMENTED IN THE PRECEDING FORMULATION. ALL THAT IS REQUIRED IS TO REPLACE THE BIRTH PARAMETER, b_x, BY $b_x + nm_x$, WHERE nm_x IS THE NET MIGRATION ASSOCIATED WITH AGE GROUP x.

IN THE PRECEDING DISCUSSION, THE LESLIE MATRIX USED FOR EACH PROJECTION TIME STEP WAS THE SAME. IN GENERAL, IF THE BIRTH RATE EXCEEDS THE DEATH RATE, THE MODEL WILL RESULT IN GROWTH THAT TENDS TO EXPONENTIAL GROWTH. SINCE EXPONENTIAL GROWTH CANNOT CONTINUE FOR LONG, IT IS NOT REALISTIC FOR THE LESLIE MATRIX TO CONTINUE UNCHANGED. IN GENERAL, FUTURE PROJECTIONS WOULD BE REPRESENTED BY THE PRODUCT OF A SERIES OF LESLIE MATRICES:

$$\boldsymbol{P}_{t+1} = \boldsymbol{A}_T \ldots \boldsymbol{A}_2 \boldsymbol{A}_1 \boldsymbol{P}_t.$$

IT CAN BE PROVED THAT, FOR REASONABLE SITUATIONS, IF THE LESLIE MATRIX DOES NOT CHANGE OVER TIME, THEN THE PROCESS CONVERGES TO ONE IN WHICH THE (EXPONENTIAL) GROWTH RATE IS A CONSTANT, AND THE RELATIVE SIZES OF THE AGE CATEGORIES REMAIN UNCHANGED.

A DETAILED BRIEFING ON THE COHORT-COMPONENT POPULATION-PROJECTION IS POSTED AT INTERNET WEBSITE https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1266/2015/03/Wilmoth-March-2-2015.pdf .

MULTISTATE MODELS

THE COHORT-COMPONENT POPULATION-PROJECTION MODEL MAY BE VIEWED AS REPRESENTING WHAT IS CALLED A BIRTH-AND-DEATH PROCESS. FOR EACH TIME STEP OF THE PROJECTION, INDIVIDUALS ARE BORN AND INDIVIDUALS DIE. THIS PROCESS WAS NOT EXPLICIT IN THE PRECEDING DISCUSSION, WHICH DEALT WITH PROBABILITIES OF BIRTHS AND PROBABILITIES OF DEATHS, NOT WITH THE EVENTS OF BIRTH AND DEATH.

IN THE BIRTH-AND-DEATH PROCESS, AN INDIVIDUAL IS CONCEIVED AS BEING IN ONE OF TWO MAIN STATES: LIVING OR DEAD, AND, IN THE CASE OF LIVE INDIVIDUALS, IN ONE OF A NUMBER OF AGE CATEGORIES. THIS MODEL MAY READILY BE EXTENDED TO THE SITUATION IN WHICH THERE ARE ANY NUMBER OF

STATES, REPRESENTING A VARIETY OF HUMAN CONDITIONS. FOR EXAMPLE, THE STATES COULD REPRESENT MARITAL STATUS, CATEGORIES OF WELLNESS, CATEGORIES OF DISEASE, CATEGORIES OF DISABILITY, EDUCATIONAL STATUS OR ECONOMIC STATUS. IN THESE APPLICATIONS, A TRANSTION MATRIX WOULD BE SPECIFIED, SIMILAR TO THE LESLIE MATRIX, WHICH SPECIFIED THE PROBABILITY OF TRANSITIONING, AT EACH TIME STEP, FROM ONE STATE TO A DIFFERENT STATE. IF THE TRANSITION PROBABILITY IS DEPENDENT ON AGE, THEN THE TRANSITION MATRIX WOULD REFLECT THAT.

THE EXPECTED NUMBERS OF PERSONS IN EACH STATE MAY BE PROJECTED IN THE SAME WAY AS WAS DONE (TO PROJECT POPULATION NUMBERS IN AGE CATEGORIES) IN THE CASE OF THE LESLIE MATRIX.

MULTIPLE DECREMENTS

A STANDARD LIFE TABLE IS AN EXAMPLE OF WHAT IS CALLED A “SINGLE DECREMENT” TABLE. THE MORTALITY RATES ARE RATES OF DEATHS FROM ALL CAUSES. IN SOME APPLICATIONS, IT IS OF INTEREST TO CONSIDER MORTALITY RATES ASSOCIATED WITH SPECIFIC CAUSES OF DEATH. A LIFE TABLE THAT INCLUDES REPRESENTATION OF DEATHS FROM MULTIPLE CAUSES IS CALLED A MULTIPLE-DECREMENTS TABLE. THE ALTERNATIVE CAUSES OF DEATH ARE REFERRED TO AS “COMPETING RISKS,” SINCE DEATH IS ASSOCIATED WITH ONLY ONE CAUSE.

THE CONSTRUCTION OF MULTIPLE DECREMENT TABLES IS SIMPLE IN SOME CASES, AND COMPLEX IN OTHERS. THE SIMPLEST CASE IS THE ONE IN WHICH THE “ALL-CAUSE” MORTALITY IS DECOMPOSED INTO A NUMBER OF MUTUALLY EXCLUSIVE AND EXHAUSTIVE CATEGORIES, AND A MORTALITY RATE IS SHOWN FOR EACH ONE OF THEM. THESE MORTALITY RATES INDICATE MORTALITY FROM A CAUSE IN THE PRESENCE OF OTHER CAUSES.

IT MAY BE OF INTEREST TO ESTIMATE MORTALITY FROM A PARTICULAR DISEASE, SUCH AS HIV, IN THE ABSENCE OF OTHER CAUSES. IN OTHER WORDS, WHAT IS DESIRED IS A LIFE TABLE THAT REFLECT DEATHS FROM THAT SINGLE CAUSE, IF ALL OTHER RISKS WERE ELIMINATED. SUCH A LIFE TABLE IS CALLED AN ASSOCIATED SINGLE DECREMENT TABLE. CONSTRUCTION OF THIS TYPE OF TABLE IS COMPLICATED, UNLESS SIMPLIFYING ASSUMPTIONS ARE MADE, SUCH AS

INDEPENDENCE OF COMPETING RISKS, AND CONSTANT SHARES OF DEATHS FROM COMPETING RISKS IN A TIME INTERVAL.

IT MAY BE OF INTEREST TO ESTIMATE LIFE EXPECTANCY UNDER THE ASSUMPTION THAT A PARTICULAR RISK IS REMOVED. A LIFE TABLE CORRESPONDING TO THIS SITUATION IS CALLED A "CAUSE-DELETED" LIFE TABLE. FOR EXAMPLE, FROM DATA FROM 1964, THE EXPECTATION OF LIFE FOR FEMALES WAS 73.78 YEARS. IF HEART DISEASE IS DELETED AS A CAUSE OF MORTALITY, THEN THE LIFE EXPECTANCY BECOMES 90.85 YEARS. IF CANCER IS DELETED, LIFE EXPECTANCY BECOMES 76.34 YEARS. THESE RESULTS SHOW THAT REDUCTIONS IN HEART DISEASE WOULD HAVE A SUBSTANTIALLY GREATER EFFECT ON LIFE EXPECTANCY THAN REDUCTIONS IN CANCER.

BECAUSE THE FORMULAS FOR ESTIMATING MULTIPLE DECREMENTS TABLES ARE COMPLICATED, AND BECAUSE THIS TOPIC IS SOMEWHAT SPECIAL, NO FORMULAS WILL BE PRESENTED ON THIS TOPIC. FOR DESRIPTION AND DISCUSSION OF FORMULAS, REFER TO THE TEXT THE METHODS AND MATERIALS OF DEMOGRAPHY 2nd ed. BY SIEGEL AND SWANSON, DEMOGRAPHY BY PRESTON ET AL., OR TO APPLIED MATHEMATICAL DEMOGRAPHY BY KEYFITZ AND CASWELL.

FROM SIEGEL AND SWANSON: "INCREMENT-DECREMENT TABLES ARE A TYPE OF MULTIPLE DECREMENTS TABLE THA ALLOW FOR BOTH INCREMENTS AND DECREMENTS IN THE INITIAL COHORT, SUCH AS LABOR FORCE ENTRY AND EXIT, SCHOOL ENROLMENT AND WITHDRAWAL, AND MARRIAGE, DIVORCE AND WIDOWHOOD.

COMPUTER SOFTWARE FOR MAKING POPULATION PROJECTIONS

PRINCIPAL SOURCES OF COMPUTER SOFTWARE FOR MAKING POPULATION PROJECTIONS WERE IDENTIFIED IN SECTION 6. THESE ARE:

- *Demographic Analysis & Population Projection System (DAPPS) Software, FROM THE U.S. CENSUS BUREAU*

- *Spectrum, FROM THE U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT HEALTH POLICY PLUS PROJECT OR ITS PARTNERS (Avenir Health AND OTHERS)*

- *MortPak, FROM THE UNITED NATIONS*

- *Tools from the International Union for the Scientific Study of Population (IUSSP)*

- THE R PACKAGE, DEMOGRAPHY

SOME OF THESE SOFTWARE TOOLS WILL NOW BE DEMONSTRATED.

SOURCES OF DEMOGRAPHIC DATA FOR THE COHORT-COMPONENT PROJECTION METHOD

IN ORDER TO IMPLEMENT THE COHORT-COMPONENT PROJECTION METHOD, IT IS NECESSARY TO HAVE AVAILABLE ESTIMATES OF AGE-SPECIFIC MORTALITY RATES, AGE-SPECIFIC FERTILITY RATES, AND AGE-SPECIFIC NET MIGRATION RATES. DATA ABOUT VITAL STATISTICS AND MIGRATION ARE COLLECTED AND PROCESSED BY A NUMBER OF INTERNATIONAL, NATIONAL, AND SUBSTATE AGENCIES, AND ARE AVAILABLE IN ELECTRONIC FORM FROM INTERNET WEBSITES.

AGENCIES THAT PROVIDE DEMOGRAPHIC DATA USED FOR MAKING POPULATION PROJECTIONS INCLUDE:

- UNITED NATIONS (VARIOUS DEPARTMENTS AND AFFILIATES, SUCH AS POPULATION DIVISION, WORLD HEALTH ORGANZATION, UNITED NATIONS POPULATION FUND)
- U.S. CENSUS BUREAU
- U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT DEMOGRAPHIC AND HEALTH SURVEYS
- STATISTICAL AGENCIES OF COUNTRIES AND STATES

11. POPULATION-BASED ESTIMATES

BY THEMSELVES, POPULATION DATA ARE OF LITTLE DIRECT INTEREST TO MOST PEOPLE, EXCEPT DEMOGRAPHERS. THE MOST WIDESPREAD USE FOR

POPULATION DATA IS TO SUPPORT THE ESTIMATION OF QUANTITIES RELATED TO POPULATION, SUCH AS DEMAND FOR PRODUCTS, SERVICES OR INFRASTRUCTURE. ESTIMATES OF QUANTITIES RELATED TO POPULATION ARE REFERRED TO AS POPULATION-BASED ESTIMATES. THE ESTIMATES OF INTEREST MAY BE FOR THE PRESENT TIME, IN WHICH CASE THEY ARE REFERRED TO SIMPLY AS POPULATION-BASED ESTIMATES, OR FUTURE TIME, IN WHICH CASE THEY ARE REFERRED TO MORE SPECIFICALLY AS POPULATION-BASED PROJECTIONS OR FORECASTS.

MATHEMATICAL BASIS FOR POPULATION-BASED ESTIMATES

THIS PRESENTATION FOCUSES ON BASIC PROCEDURES FOR MAKING POPULATION-BASED ESTIMATES. TO THIS END, INTEREST CENTERS ON QUANTITIES OF INTEREST THAT ARE DEPENDENT ON POPULATION. AS DISCUSSED EARLIER, THESE QUANTITIES ARE FROM A LARGE VARIETY OF APPLICATION AREAS, SUCH AS HEALTH, EDUCATION, GOVERNMENT AND BUSINESS.

TO MAKE A POPULATION-BASED ESTIMATE, TWO BASIC THINGS ARE NEEDED:

1. POPULATION DATA, WHICH MAY BE EITHER HISTORICAL DATA (FROM A CENSUS, SURVEY, REGISTRATION SYSTEM, OR OTHER SOURCE) OR FUTURE FIGURES (EITHER PROJECTIONS OR FORECASTS)

2. THE RELATIONSHIP OF A QUANTITY OF INTEREST TO POPULATION CHARACTERISTICS

RELATIVE TO POINT 1, THE PRECEDING SECTION DESCRIBED METHODS FOR MAKING POPULATION PROJECTIONS AND SIMPLE FORECASTS. PART 2 OF THE PRESENTATION DISCUSSES PROCEDURES FOR MAKING STATISTICAL FORECASTS.

RELATIVE TO POINT 2, IT WILL BE ASSUMED THAT THE RELATIONSHIP OF QUANTITIES OF INTEREST TO POPULATION IS SPECIFIED. THE RELATONSHIP COULD BE SPECIFIED BY AN EQUATION THAT SPECIFIES THE EXPECTATION OF A QUANTITY OF INTEREST AS A FUNCTION OF POPULATION CHARACTERISTICS (SUCH AS A REGRESSION-TYPE EQUATION), OR A SET OF TABLES THAT DOES THE SAME.

THE OBJECTIVE HERE IS TO IDENTIFY METHODS FOR COMBINING THE TWO ITEMS LISTED ABOVE TOGETHER TO PRODUCE A POPULATION-BASED FORECAST.

IT IS NOTED THAT, CONCEPTUALLY, THE TWO ITEMS LISTED ABOVE NEED NOT BE SEPARATED – AN ALTERNATIVE APPROACH WOULD BE TO ESTIMATE QUANTITIES OF INTEREST BASED ON A MULTIVARIATE MODEL THAT INCLUDES BOTH THE QUANTITIES OF INTEREST AND POPULATION AS INTERRELATED VARIABLES (ENDOGENOUS VARIABLES IN A SYSTEM). WE SHALL ADOPT THE TWO-STEP APPROACH SPECIFIED ABOVE BECAUSE IT IS MUCH SIMPLER AND IS THE APPROACH COMMONLY TAKEN TO MAKING POPULATION-BASED ESTIMATES. IT IS APPROPRIATE IN ANY SITUATION IN WHICH A VARIABLE OF INTEREST IS AFFECTED BY POPULATION, BUT DOES NOT IN TURN AFFECT POPULATION.

TO KEEP THE PRESENTATION SIMPLE, AND PROMOTE UNDERSTANDING OF BASIC CONCEPTS, WE SHALL DESCRIBE A SINGLE, BASIC PROCEDURE FOR MAKING POPULATION-BASED ESTIMATES, VIZ., THE METHOD OF SYNTHETIC ESTIMATION. THIS METHODOLOGY IS DESCRIBED IN DETAIL, FOR EXAMPLE, IN THE TEXT, *SMALL AREA ESTIMATION* BY J. N. K. RAO (WILEY, 2003). THIS TECHNIQUE IS AN EXAMPLE OF MODEL-BASED ESTIMATION OR INDIRECT ESTIMATION. IT IS ANALOGOUS TO THE PROCEDURE OF POST-STRATIFICATION IN SAMPLE SURVEY.

SUPPOSE THAT A MODEL OF THE RELATIONSHIP OF A VARIABLE OF INTEREST TO POPULATION CHARACTERISTICS (AGE, SEX, RACE, ETC.) IS AVAILABLE FOR A PARTICULAR DOMAIN (WHICH WE SHALL CALL THE “ESTIMATION DOMAIN,” SUCH AS A PARTICULAR COUNTRY FOR A PARTICULAR YEAR. IF THIS MODEL IS USED TO ESTIMATE THE VARIABLE FOR A DIFFERENT DOMAIN, SUCH AS A SUBREGION OR A DIFFERENT YEAR, UNDER THE ASSUMPTION THAT THE RELATIONSHIP OF THE VARIABLE OF INTEREST TO POPULATION CHARACTERISTICS IS THE SAME FOR THE DIFFERENT DOMAIN AS FOR THE ESTIMATION DOMAIN, THE ESTIMATE IS CALLED A “SYNTHETIC ESTIMATE.”

FOR EXAMPLE, SUPPOSE THAT WE ARE INTERESTED IN ESTIMATING NATIONAL SCHOOL ENROLMENT FIVE YEARS IN THE FUTURE, BASED ON A PROJECTION OF POPULATION FIVE YEARS INTO THE FUTURE. SUPPOSE FURTHER THAT WE HAVE AVAILABLE, FROM RECENT HISTORICAL DATA, ESTIMATES OF THE PROPORTION OF POPULATION ENROLLED, BY AGE AND SEX. LET US DENOTE THIS PROPORTION

AS p(a,s), WHERE a DENOTES AGE CATEGORY AND s DENOTES SEX CATEOGORY. THEN, UNDER THE ASSUMPTION THAT THE RELATIONSHIP OF ENROLMENT TO POPULATION IS THE SAME FIVE YEARS INTO THE FUTURE, A SYNTHETIC ESTIMATE OF THE TOTAL ENROLMENT, E, FIVE YEARS IN THE FUTURE, IS

$$E = \sum_{a=1}^{n_a} \sum_{s=1}^{2} p(a,s)N(a,s)$$

WHERE N(a,s) DENOTES THE POPULATION SIZE FOR AGE CATEGORY a AND SEX CATEGORY s.

COMPUTER SOFTWARE FOR POPULATION-BASED ESTIMATES

THE *SPECTRUM* SOFTWARE PACKAGE CALCULATES SYNTHETIC ESTIMATES FOR A RANGE OF ECONOMIC, EDUCATIONAL AND OTHER POPULATION-RELATED VARIABLES.

THE *SPECTRUM* SOFTWARE MAY BE DOWNLOADED FROM WEBSITE https://www.avenirhealth.org/software-spectrum.php .

AT THIS POINT OF THE PRESENTATION, A DEMONSTRATION OF THE SPECTRUM SOFTWARE WILL BE PRESENTED. THE SOFTWARE WILL BE USED TO MAKE A COHORT-COMPONENT POPULATION PROJECTION, AND POPULATION-BASED FORECASTS BASED ON THE PROJECTION.

APPLICATION-SPECIFIC DATA SOURCES FOR POPULATION-BASED ESTIMATES

SOURCES FOR DEMOGRAPHIC DATA (POPULATION, FERTILITY, MORTALITY, MIGRATION) HAVE ALREADY BEEN IDENTIFIED. IN ORDER TO MAKE POPULATION-BASED FORECASTS, DATA ARE REQUIRED THAT DESCRIBE THE RELATIONSHIP OF INCIDENCES AND PREVALENCES OF QUANTITIES OF INTEREST TO POPULATION CHARACTERISTICS (SUCH AS AGE AND SEX). THE INTERNET PROVIDES A RICH SOURCE FOR DATA OF THIS SORT, IN A FULL RANGE OF SUBSTANTIVE FIELDS (INCLUDING ECONOMICS, EDUCATION, HEALTH, ENVIRONMENT, URBAN PLANNING, AGRICULTURE, POLITICS AND INSURANCE).

12. GEOGRAPHIC INFORMATION SYSTEMS (GIS)

THE GEOGRAPHIC DISTRIBUTION OF POPULATION IS WELL ILLUSTRATED ON A MAP ON A COMPUTER SCREEN. COMPUTER SOFTWARE FOR DISPLAYING GEOGRAPHIC DATA ON MAPS IS GENERALLY REFERRED TO AS GEOGRAPHIC INFORMATION SYSTEM (GIS) SOFTWARE. GIS SOFTWARE RANGES IN FUNCTIONALITY FROM DISPLAYING SIMPLE MAPS TO POWERFUL SYSTEMS THAT PERFORM COMPLEX SPATIAL CALCULATIONS.

THERE ARE A LARGE NUMBER OF GIS SYSTEMS AVAILABLE, BOTH COMMERCIAL AND FREE. COMMERCIAL PACKAGES INCLUDE THE ESRI ArcGIS PACKAGE AND THE PITNEY BOWES MapInfo PACKAGE. FREE GIS PACKAGES INCLUDE THE US ARMY CORPS OF ENGINEERS' GRASS SYSTEM, QGIS, OpenStreetMap AND SAGA GIS.

SOME GIS-RELATED SOFTWARE IS AVAILABLE IN R, BUT IT IS NOT A REASONABLE ALTERNATIVE TO THE FREE FULL-FUNCTION GIS PACKAGES SUCH AS GRASS AND QGIS.

IN THIS PRESENTATION, WE WILL ILLUSTRATE PRESENTATION OF POPULATION DATA USING EITHER THE GRASS OR QGIS PACKAGE.

IN ORDER TO USE A GIS EFFICIENTLY, IT IS NECESSARY TO ACCESS GEO-SPATIAL DATA THAT HAVE ALREADY BEEN PREPARED FOR USE IN A GIS, OR "GEOREFERENCED," OR "GEOCODED." THIS INCLUDES DATA SETS THAT CONTAIN GEOGRAPHIC COORDINATES FOR POINTS (SUCH AS CITIES), LINES (SUCH AS ROADS) AND POLYGONS (SUCH AS COUNTRIES). A LARGE AMOUNT OF GEOCODED DATA IS AVAILABLE, FREE, FROM THE INTERNET. FOR STANDARD DEMOGRAPHIC APPLICATIONS, GEOCODED DATA ARE READILY AVAILABLE, AND THERE IS NO NEED TO EXPEND RESOURCES IN GEOCODING DATA FOR BASIC MAP FEATURES.

AT THIS POINT IN THE PRESENTATION, A DEMONSTRATION WILL BE PRESENTED SHOWING THE DISPLAY OF POPULATION DATA ON A MAP.

13. SUMMARY OF PART 2

PART 1 OF THE PRESENTATION DESCRIBED THE BASIC CONCEPTS OF DEMOGRAPHY AND DEMOGRAPHIC ANALYSIS, USING MATHEMATICS THAT INCLUDES ARITHMETIC, BASIC ALGEBRA, MATRIX ARITHMETIC, AND A FEW BASIC CONCEPTS FROM CALCULUS, PROBABILITY AND STATISTICS.

PART 1 INCLUDED MATERIAL ON POPULATION PROJECTION, POPULATION FORECASTING, AND POPULATION-BASED FORECASTS, AND THE USE OF BASIC COMPUTER SOFTWARE TO IMPLEMENT THESE FUNCTIONS.

PART 2 OF THE PRESENTATION DESCRIBES SOME MORE ADVANCED ASPECTS OF DEMOGRAPHIC ANALYSIS, MAKING GREATER USE OF CALCULUS, MATRIX ALGEBRA, AND STATISTICS.

HERE FOLLOWS A SUMMARY OF THE CONTENT OF PART 2.

- THE MATHEMATICS OF ADVANCED DEMOGRAPHIC ANALYSIS
    - VECTOR AND MATRIX ALGEBRA
    - EIGENVALUES, EIGENVECTORS, AND PRINCIPAL COMPONENTS
- THE STATISTICS OF ADVANCED DEMOGRAPHIC ANALYSIS
    - ESTIMATION OF SURVIVAL FUNCTIONS
    - THE LINEAR STATISTICAL MODEL
    - STOCHASTIC PROCESSES
    - STATISTICAL FORECASTING METHODOLOGIES (FORECASTING (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS, VAR MODELS, STRUCTURAL MODELS, KALMAN FILTERING, BAYESIAN ESTIMATION)
    - SMALL-AREA ESTIMATION
- MORE ON POPULATION PROJECTION
    - POPULATION-TRANSITION MATRICES FOR GENERAL LIFE-CYCLE MODELS
    - STABLE POPULATION
- MORE ON POPULATION-BASED FORECASTS
    - SMALL-AREA ESTIMATION
- ESTIMATION OF DEMOGRAPHIC PARAMETERS BY INDIRECT METHODS: NONPARAMETRIC METHODS
- COMPUTER SOFTWARE FOR INDIRECT ESTIMATION
    - INDIRECT METHODS (*TOOLS FOR DEMOGRAPHIC ESTIMATION*)

    - UNITED NATIONS' *MORTPAK*
    - GARY KING'S *YourCast*
    - ROB HYNDMAN'S *DEMOGRAPHY IN R,* CRAN R LIBRARIES)
- ESTIMATION OF DEMOGRAPHIC PARAMETERS BY INDIRECT METHODS: PARAMETRIC METHODS
    - SURVIVAL FUNCTION, FAILURE DENSITY FUNCTION, FORCE OF MORTALITY
    - NONPARAMETRIC ESTIMATION OF THE SURVIVAL FUNCTION
    - PARAMETRIC REPRESENTATIONS OF THE HAZARD FUNCTION
    - SEMIPARAMETRIC MODELS OF THE HAZARD FUNCTION
- STOCHASTIC PROCESSES IN DEMOGRAPHY APPLICATIONS
    - REFERENCE TEXTS
    - LIMITATIONS
    - EFFECTS OF TAKING STOCHASTIC VARIATION INTO ACCOUNT
- FORECASTING OF DEMOGRAPHIC PARAMETERS
    - THE GIROSI-KING METHOD (BAYESIAN ESTIMATION, PRINCIPAL COMPONENTS, MULTIVARIATE CONSTRAINTS)
    - BASED ON THE DEMOGRAPHIC TRANSITION
    - BASED ON A STABLE POPULATION
- COLLECTION OF DEMOGRAPHIC DATA (CSPro, Epi Info)
- SPECIAL TOPICS (OPTIONAL)
- DISCUSSION OF REFERENCES AND OTHER RESOURCES
- COMPLETION OF COURSE EVALUATION FORM

14. THE MATHEMATICS OF ADVANCED DEMOGRAPHY

VECTOR AND MATRIX ALGEBRA

IN PART 1 OF THE PRESENTATION, VECTORS AND MATRICES WERE DEFINED, AND BASIC OPERATIONS INVOLVING THEM, SUCH AS ADDITION, SUBTRACTION AND MULTIPLICATION, WERE DESCRIBED. FOR PART 2, ADDITIONAL ASPECTS OF MATRIX ALGEBRA ARE NEEDED.

THE *INVERSE* OF A SQUARE MATRIX **A** IS A MATRIX **B** = $\mathbf{A}^{-1}$ SUCH THAT **AB** = **I**, IF IT EXISTS, IT IS UNIQUE, AND $\mathbf{A}^{-1}$ **A** = **I**. IF THE INVERSE EXISTS, THE MATRIX IS CALLED *NONSINGULAR* OR *INVERTIBLE*.

THE *RANK*, r, OF A MATRIX IS THE NUMBER OF LINEARLY INDEPENDENT ROWS OR COLUMNS (WHICH ARE EQUAL). FOR A SQUARE MATRIX, IF r=n, THE MATRIX IS SAID TO BE OF FULL RANK, AND IT IS INVERTIBLE.

A *SYMMETRIC MATRIX* IS A SQUARE MATRIX FOR WHICH $x_{ij} = x_{ji}$.

<u>EXAMPLES</u>:

THE MATRIX

$$\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

IS OF RANK 2. ITS INVERSE IS THE SAME MATRIX.

THE MATRIX

$$\begin{matrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 0 & -1 & 1 \end{matrix}$$

IS SINGULAR (NON-INVERTIBLE), SINCE THE THIRD ROW IS THE FIRST MINUS THE SECOND. IT IS OF RANK 2.

THE MATRIX

$$\begin{matrix} 1 & .5 & 0 \\ .5 & 1 & 0 \\ 0 & 0 & 0 \end{matrix}$$

IS SINGULAR, AND OF RANK 2.

IN MATRIX NOTATION, A SYSTEM OF SIMULTANEOUS LINEAR EQUATIONS, FOR EXAMPLE,

$$y_1 = a_{11} x_1 + a_{12} x_2$$
$$y_2 = a_{21} x_1 + a_{22} x_2$$

MAY BE REPRESENTED AS

$$\mathbf{y} = \mathbf{A}\ \mathbf{x}$$

WHERE $\mathbf{y}' = (y_1, y_2)$, $\mathbf{x}' = (x_1, x_2)$ AND $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2)$ WHERE $\mathbf{a}_1' = (a_{11}, a_{21})$ AND $\mathbf{a}_2' = (a_{12}, a_{22})$.

CONSIDER THE SYSTEM OF LINEAR EQUATIONS

$$A\mathbf{x} = \mathbf{b}$$

WHERE A IS AN n x n MATRIX AND **x** AND **b** ARE VECTORS OF LENGTH n. IT IS DESIRED TO DETERMINE NONZERO SOLUTIONS, **x**, TO THIS SYSTEM, WHICH ARE LINEAR COMBINATIONS OF THE COLUMNS OF A. THERE ARE TWO CASES TO CONSIDER, DEPENDING ON WHETHER **b** IS ZERO OR NONZERO. IF **b** IS ZERO, THE SYSTEM IS CALLED *HOMOGENEOUS,* AND IF **b** IS NONZERO THE SYSTEM IS CALLED *NONHOMOGENEOUS*.

IF $\mathbf{b} = \mathbf{0}$, THEN A SOLUTION EXISTS IF AND ONLY IF THERE EXISTS A LINEAR COMBINATION (I.E., A**b**) OF THE COLUMNS OF **A** THAT IS EQUAL TO **0**. THAT IS, THERE IS A NONZERO SOLUTION TO THE SYSTEM IF AND ONLY IF **A** IS NOT OF FULL RANK, I.E., IS SINGULAR.

SUPPOSE THAT **b** IS NONZERO. IT IS DESIRED TO DETERMINE CONDITIONS UNDER WHICH THE SYSTEM HAS SOLUTIONS FOR *ALL* NONZERO VECTORS **b**, NOT JUST FOR SOME VECTORS **b**. IN THIS CASE, THERE IS A SOLUTION IF **A** IS INVERTIBLE, SINCE IN THAT CASE WE HAVE (PREMULTIPLYING BOTH SIDES OF THE SIMULTANEOUS EQUATIONS BY $\mathbf{A}^{-1}$),

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}b$$

OR

$$\mathbf{x} = \mathbf{A}^{-1}b.$$

IF **A** IS NOT INVERTIBLE, THEN THERE IS NO MATRIX $\mathbf{A}^{-1}$ FOR WHICH $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, THAT IS, THERE IS NO MATRIX $\mathbf{A}^{-1}$ FOR WHICH $\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, I.E., NO MATRIX $\mathbf{A}^{-1}$ FOR

WHICH $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. THAT IS, THERE IS NO LINEAR COMBINATION OF THE COLUMNS OF **A** WHICH SATISFY THE SYSTEM OF EQUATIONS.

IN SUMMARY, FOR **b** NOT EQUAL TO ZERO, THE SYSTEM HAS A NONZERO SOLUTION THAT MAY BE EXPRESSED AS A LINEAR COMBINATION OF THE COLUMNS OF **A** IF AND ONLY IF A IS NONSINGULAR. THE SOLUTION IS $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

(HERE FOLLOWS SOME ADDITIONAL INFORMATION ABOUT SOLUTIONS TO LINEAR HOMOGENEOUS SYSTEMS OF EQUATIONS, FROM THE WIKIPEDIA ARTICLE ON THAT TOPIC:

There is a close relationship between the solutions to a linear system and the solutions to the corresponding homogeneous system:

$A x = b$ and $A x = 0$.

Specifically, if p is any specific solution to the linear system Ax = b, then the entire solution set can be described as

$\{ p + v : v$ is any solution to $A x = 0 \}$.

Geometrically, this says that the solution set for Ax = b is a translation of the solution set for Ax = 0. Specifically, the flat for the first system can be obtained by translating the linear subspace for the homogeneous system by the vector p.

This reasoning only applies if the system Ax = b has at least one solution. This occurs if and only if the vector b lies in the image of the linear transformation A. [END OF WIKIPEDIA EXTRACT.]

THE TRACE OF A MATRIX IS THE SUM OF ITS DIAGONAL ELEMENTS: IF **A** IS AN nxn MATRIX, THEN $tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$. IT HOLDS THAT tr(**A** + **C**) = tr(**A**) + tr(**C**), tr(**AC**) = tr(**CA**) AND tr(**A**) = tr(**A**’) (WHERE THE MATRICES ARE ASSUMED CONFORMABLE).

SUPPOSE THAT **A** IS AN nxn SQUARE MATRIX. A NUMBER λ AND AN nx1 NONZERO VECTOR **b** ARE A RIGHT EIGENVALUE AND EIGENVECTOR PAIR OF **A** IF $\mathbf{Ab}=\lambda\mathbf{b}$. THERE ARE UP TO n POSSIBLE EIGENVALUES FOR **A**. THEY MAY BE

COMPLEX NUMBERS, IN WHICH CASE THEY OCCUR IN CONJUGATE PAIRS. DENOTE THE n EIGENVALUES AS λ_i FOR i= 1,…,n. THEN tr(**A**) = $\sum_{i=1}^{n} \lambda_i$.

THE DETERMINANT OF A MATRIX, **A**, DENOTED BY |**A**|, IS DEFINED AS $|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$. THE NOTATION det(**A**) IS ALSO COMMON FOR |**A**|.

IT CAN BE SHOWN THA T THE DETERMINANT OF A MATRIX **A** = [a_{ij}] MAY BE CALCULATED AS THE SUM OF ALL PRODUCTS

$$(-1)^p a_{1i_1}, \ldots, a_{mi_m}$$

CONSISTING OF ONE ELEMENT FROM EACH ROW AND EACH COLUMN, WHERE p IS THE NUMBER OF INVERSIONS REQUIRED TO TRANSFORM THE PERMUTATION $i_1,\ldots,i_m$ into 1,…,m. FOR EXAMPLE, FOR THE 2x2 MATRIX **A** = [a_{ij}], THE DETERMINANT IS $a_{11}a_{22} – a_{12}a_{21}$.

SOME FEATURES OF DETERMINANTS ARE THE FOLLOWING, WHERE IT IS ASSUMED THAT ALL MATRICES ARE NONSINGULAR (INVERTIBLE):

1. IF **A** AND **B** ARE n x n MATRICES, THEN |**AB**| = |**A**||**B**|.
2. |**A**'| = |**A**|.
3. |$\mathbf{A}^{-1}$| = 1/|**A**|.

A MATRIX IS NONSINGULAR IF AND ONLY IF ALL OF ITS EIGENVALUES ARE NONZERO, I.E., ITS DETERMINANT IS NONZERO.

ALL OF THE EIGENVALUES OF A SYMMETRIC MATRIX ARE REAL.

THE RANK OF A MATRIX, **A**, IS THE NUMBER OF NONZERO EIGENVALUES OF THE SYMMETRIC MATRIX **AA**'.

SINCE THE EIGENVALUE / EIGENVECTOR PAIR λ, **b** SATISFY

Ab = λ**b**,

IT FOLLOWS THAT

$\mathbf{Ab} - \lambda\mathbf{Ib} = \mathbf{0}$

OR

$(\mathbf{A} - \lambda\mathbf{I})\mathbf{b} = \mathbf{0}$.

FROM THE DISCUSSION EARLIER ABOUT SOLUTIONS TO A SYSTEM OF SIMULTANEOUS EQUATIONS, THIS SYSTEM HAS A SOLUTION ONLY IF THE MATRIX $\mathbf{A} - \lambda\mathbf{I}$ IS SINGULAR.

IF $(\mathbf{A} - \lambda\mathbf{I})$ WERE NONSINGULAR, THEN WE COULDPREMULTIPLY THE PRECEDING EXPRESSION BY $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ AND OBTAIN $\mathbf{b} = \mathbf{0}$. BUT **b** IS ASSUMED TO BE NONZERO, HENCE $(\mathbf{A} - \lambda\mathbf{I})$ MUST BE SINGULAR. HENCE AN EIGENVALUE OF **A** IS A NUMBER λ SUCH THAT $|\mathbf{A} - \lambda\mathbf{I}| = 0$.

EIGENVALUES ARE ALSO CALLED CHARACTERISTIC VALUES OR CHARACTERISTIC ROOTS OR LATENT ROOTS, AND THE JUST-PRECEDING EQUATION IS CALLED THE CHARACTERISTIC EQUATION.

EIGENVECTORS SPAN THE VECTOR SPACE SPANNED BY THE ROWS OR COLUMNS OF A MATRIX. EIGENVALUES ARE MEASURES OF LENGTH ALONG EIGENVECTORS, AND THE DETERMINANT IS A MEASURE OF VOLUME. FOR EXAMPLE, CONSIDER THE CASE OF AN n x n IDENTITY MATRIX. THE EIGENVECTORS ARE UNIT VECTORS POINTING ALONG n ORTHOGONAL AXES. THE LENGTH OF EACH EIGENVECTOR IS ONE (THE CORRESPONDING EIGENVALUE), AND THE PRODUCT OF ALL OF THE LENGTHS (EIGENVALUES) IS THE n-DIMENSIONAL VOLUME OF THE n-DIMENSIONAL HYPERCUBE SPANNED BY THE n EIGENVECTORS.

THERE ARE A NUMBER OF NUMERICAL METHODS FOR EVALUATING THE DETERMINANT AND FINDING THE EIGENVALUES AND EIGENVECTORS. THESE WILL NOT BE DISCUSSED IN THIS COURSE. SOFTWARE FOR FINDING EIGENVALUES AND EIGENVECTORS IS AVAILABLE FROM MANY SOURCES, BOTH COMMERCIAL (E.G., MATLAB) AND FREE (E.G., *NUMERICAL RECIPES* BY WILLIAM H. PRESS ET AL. AND R).

THE EIGENVECTORS AND EIGENVALUES OF A MATRIX HAVE A CONCEPTUALLY SIMPLE GEOMETRIC INTERPRETATION. MULTIPLICATION OF A VECTOR BY A

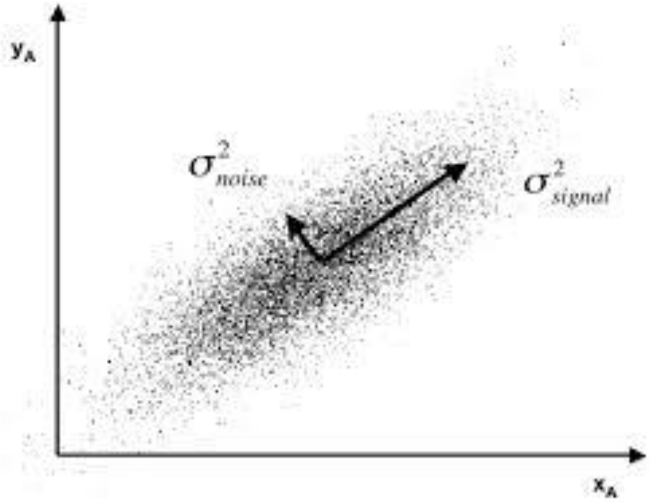
MATRIX CORRESPONDS TO MAKING A LINEAR TRANSFORMATION OF THE VECTOR. THE EIGENVECTORS ARE AN ORTHONORMAL SET OF VECTORS SPANNING THE SPACE IN WHICH THE VECTOR IS LOCATED, AND SO THE VECTOR MAY BE REPRESENTED AS A LINEAR COMBINATION OF THE EIGENVECTORS. THE EIGENVALUES ARE FACTORS BY WHICH EACH EIGENVECTOR IS MULTIPLIED, WHEN THE TRANSFORMATION IS APPLIED TO EACH EIGENVECTOR (I.E., WHEN THE EIGENVECTOR IS MULTIPLIED BY THE MATRIX). THESE FACTORS APPLY TO EACH TERM OF THE LINEAR COMBINATION OF EIGENVECTORS THAT REPRESENTS THE VECTOR.

AS A VERY SIMPLE EXAMPLE, IF THE VECTOR IS ONE OF THE EIGENVECTORS, THEN THE MATRIX MULTIPLICATION SIMPLY MULTIPLIES THE VECTOR BY THE CORRESPONDING EIGENVALUE.

PRINCIPAL COMPONENTS ANALYSIS

EIGENVECTORS AND EIGENVALUES ARE DEFINED FOR ANY SQUARE MATRIX. IF THE MATRIX IS A VARIANCE MATRIX OF A MULTIVARIATE RANDOM VARIABLE, THEN THE EIGENVECTORS AND EIGENVALUES HAVE AN INSIGHTFUL INTERPRETATION. THE EIGENVECTOR HAVING THE LARGEST EIGENVALUE IS THE LINEAR COMBINATION (OF THE RANDOM VARIABLES) HAVING MAXIMAL VARIANCE, AND THE EIGENVALUE IS THE VARIANCE. THIS EIGENVECTOR IS CALLED THE FIRST PRINCIPAL COMPONENT. THE EIGENVECTOR HAVING THE SECOND-LARGEST EIGENVALUE IS THE LINEAR COMBINATION, ORTHOGONAL TO THE PREVIOUS LINEAR COMBINATION, HAVING MAXIMAL VARIANCE, AND THE EIGENVALUE IS THE VARIANCE. THE EIGENVECTOR IS CALLED THE SECOND PRINCIPAL COMPONENT. THE REMAINING EIGENVECTORS, DEFINED IN SIMILAR FASHION AND RANKED IN ORDER OF THE MAGNITUDES OF THE EIGENVALUES, DEFINE THE REMAINING PRINCIPAL COMPONENTS.

A GRAPHICAL ILLUSTRATION OF THE PRINCIPAL COMPONENTS IS PRESENTED IN FIGURE 10.



PRINCIPAL COMPONENTS ANALYSIS (PCA) IS THE PROCESS OF IDENTIFYING ALL OF THE EIGENVECTORS AND EIGENVALUES AND THEN SELECTING A SUBSET OF THE LARGEST OF THEM THAT REPRESENTS A LARGE PORTION OF THE TOTAL VARIANCE (REPRESENTED BY THE SUM OF THE EIGENVALUES, WHICH, IN THE CASE OF A COVARIANCE MATRIX, ARE ALL POSITIVE).

PCA IS USED TO REDUCE THE DIMENSIONALITY OF A LARGE SET OF RANDOM VARIABLES. IT IS ALSO USED TO SIMPLIFY MODELS. IN FORECASTING WITH A MODEL CONTAINING EXPLANATORY VARIABLES, IT IS NECESSARY TO FORECAST THE VALUE OF EACH EXPLANATORY VARIABLE. THAT IS PROBLEMATIC FOR VARIABLES THAT ARE CORRELATED, BECAUSE THE VALUE OF EACH VARIABLE IS RELATED TO THE VALUES OF THE OTHERS. IF THE MODEL IS SPECIFIED IN TERMS OF PRINCIPAL COMPONENTS, EACH OF THE FUTURE VALUES (OF THE PRINCIPAL COMPONENTS) MAY BE SPECIFIED WITHOUT CONSIDERATION OF THE OTHERS, SINCE THEY ARE ORTHOGONAL (UNCORRELATED).

15. THE STATISTICS OF DEMOGRAPHY

THE PRECEDING SECTION DESCRIBED SOME MATHEMATICAL CONCEPTS FROM LINEAR ALGEBRA THAT ARISE IN DEMOGRAPHIC ANALYSIS, SUCH AS EIGENVECTORS AND EIGENVALUES AND PRINCIPAL COMPONENTS ANALYSIS. WE SHALL MAKE USE OF THESE RESULTS LATER.

THERE ARE A NUMBER OF STATISTICAL CONCEPTS THAT ARISE IN DEMOGRAPHIC ANALYSIS, BUT IT IS NOT SO EASY TO SUMMARIZE THEM. THE PROBLEM IS THAT DERIVATION OF THEM REQUIRES A BASIC KNOWLEDGE OF THE FIELD OF BASIC PROBABILITY AND STATISTICS. UNLIKE THE MATRIX-ALGEBRA RESULTS PRESENTED ABOVE, THESE RESULTS CANNOT BE EASILY EXPLAINED "IN A VACUUM."

IN THIS PART OF THE PRESENTATION, RESULTS WILL BE DRAWN FROM THE FOLLOWING AREAS OF PROBABILITY AND STATISTICS. WITHOUT DERIVATION:

- PROBABILITY DISTRIBUTIONS
- ESTIMATION (OF PARAMETERS OF A PROBABILITY DISTRIBUTION; THE PARAMETERS OF A STATISTICAL MODEL; OF SURVIVAL FUNCTIONS; OF THE COX PROPORTIONAL-HAZARD MODEL)
- THE GENERAL LINEAR STATISTICAL MODEL
- STOCHASTIC PROCESS (TIME SERIES) MODELS
- FORECASTING (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS, VECTOR AUTOREGRESSIVE (VAR) MODELS, STRUCTURAL MODELS, KALMAN FILTERING, BAYESIAN ESTIMATION)
- SMALL-AREA ESTIMATION

STATISTICAL COMPUTATIONS WILL BE ILLUSTRATED USING STATISTICAL SOFTWARE THAT IS AVAILABLE FREE FROM THE INTERNET, INCLUDING THE INTERNATIONAL UNION FOR THE SCIENTIFIC STUDY OF POPULATION'S *TOOLS FOR DEMOGRAPHIC ESTIMATION* (INDIRECT METHODS), UNITED NATIONS' *MORTPAK*, GARY KING'S *YOURCAST*, AND ROB HYNDMAN'S *DEMOGRAPHY IN R*, CRAN R LIBRARIES).

16. MORE ON POPULATION PROJECTION

POPULATION-TRANSITION MATRICES FOR GENERAL LIFE-CYCLE MODELS

THE REPRESENTATION OF A POPULATION PROJECTION IN MATRIX FORM, PRESENTED ABOVE, INTRODUCED THE CONCEPT OF A POPULATION-PROJECTION MATRIX. A POPULATION-PROJECTION MATRIX SPECIFIES THE PROBABILITY THAT

AN ENTITY CHANGES FROM ONE STATE TO ANOTHER, AND THE EXPECTED NUMBER OF NEW ENTITIES CREATED PER ENTITY IN EACH STATE. THE POPULATION-PROJECTION MATRIX FOR AN AGE-CLASSIFIED POPULATION IS CALLED A LESLIE MATRIX.

FOR HUMAN POPULATIONS, IT IS COMMON TO CONSIDER THE CHANGE OF AN INDIVIDUAL FROM LIVING TO DEAD AS A FUNCTION OF AGE. IN THIS CASE, THE TRANSITION PROBABILITES ARE DERIVED FROM INFORMATION IN THE LIFE TABLE. FOR MANY NONHUMAN POPULATIONS, THE AGE OF AN ENTITY IS NOT KNOWN, AND THE PROBABILITY OF TRANSITING FROM ONE STATE TO ANOTHER IS BETTER REPRESENTED IN TERMS OF STAGE OF DEVELOPMENT, RATHER THAN AGE.

THIS SECTION PRESENTS SOME ADDITIONAL INFORMATION ABOUT POPULATION-PROJECTION MATRICES, THAT APPLIES BOTH TO AGE-CLASSIFIED PROCESSES AND STAGE-CLASSIFIED PROCESSES.

POPULATION-PROJECTION MATRICES ARE GENERALIZATIONS OF STOCHASTIC TRANSITION MATRICES. A STOCHASTIC TRANSITION MATRIX SPECIFIES, FOR A PROCESS OBSERVED AT EQUALLY-SPACED TIMES, THE PROBABILITY THAT AN ENTITY IN A PARTICULAR STATE AT A PARTICULAR TIME TRANSITS TO ANOTHER STATE IN THE SUCCEEDING TIME. A POPULATION-PROJECTION MATRIX INCLUDES TRANSITION PROBABILITIES, BUT IT ALSO INCLUDES FERTILITY RATES THAT SPECIFY THE MEAN NUMBER OF NEW ENTITIES GENERATED BY AN ENTITY IN EACH STATE, PER TIME STEP.

THE POPULATION-PROJECTION MATRIX ACCOMMODATES THE FACTS THAT THE STATE OF A POPULATION IS REPRESENTED BY THE NUMBERS OF ENTITIES IN EACH STATE CATEGORY (SUCH AS AGE), AND THAT NEW ENTITIES ARE CREATED AT EACH TIME STEP.

A LESLIE MATRIX IS A POPULATION-PROJECTION MATRIX IN WHICH THE VARIOUS STATES ARE AGE CATEGORIES, AND AN INDIVIDUAL MAY TRANSITION ONLY TO THE FOLLOWING AGE CATEGORY OR TO DEATH.

IT IS USEFUL TO USE POPULATION-PROJECTION MATRICES TO DESCRIBE PROCESSES ADDITIONAL TO THE AGE-CLASSIFIED PROCESSES THAT HAVE BEEN CONSIDERED SO FAR. SPECIFICALLY, WE SHALL USE POPULATION-PROJECTION

MATRICES TO DESCRIBE PROCESSES IN WHICH THE STATE OF AN INDIVIDUAL MAY REFER EITHER TO AGE OR A MORE GENERAL STAGE OF DEVELOPMENT. FIGURES 11 AND 12 PRESENT EXAMPLES OF GRAPHICAL DESCRIPTIONS OF AGE-CLASSIFIED AND STAGE-CLASSIFIED TRANSITION PROCESSES. THESE GRAPHS ARE CALLED LIFE-CYCLE GRAPHS. THESE EXAMPLES ARE PRESENTED IN *MATRIX POPULATION MODELS*, 2nd ed., BY HAL CASWELL.

IN THE FIGURES, THE STATE OF AN ENTITY IS REPRESENTED BY A GRAPH NODE (CIRCLE), AND THE TRANSITION FROM ONE STATE TO ANOTHER STATE (OR THE CREATION OF A NEW ENTITY IN ONE STATE BY AN ENTITY IN ANOTHER STATE) IS REPRESENTED BY A DIRECTED ARC FROM THE FIRST STATE TO THE SECOND STATE. THE p_{ij} DENOTE PROBABILITIES THAT AN ENTITY TRANSITS FROM STATE i TO STATE j IN A UNIT TIME INTERVAL. THE f_{ij}, FERTILITY RATES, DENOTE THE MEAN NUMBER OF NEW ENTITIES OF TYPE j CREATED BY AN ENTITY OF TYPE i IN A UNIT TIME INTERVAL. IN THESE TWO EXAMPLES, ALL NEWLY CREATED ENTITIES ARE OF STATE 1.

FIGURE 11. EXAMPLE OF A LIFE-CYCLE GRAPH FOR AN AGE-CLASSIFIED TRANSITION PROCESS

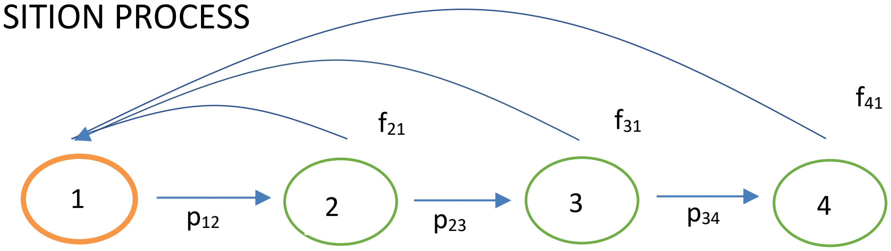
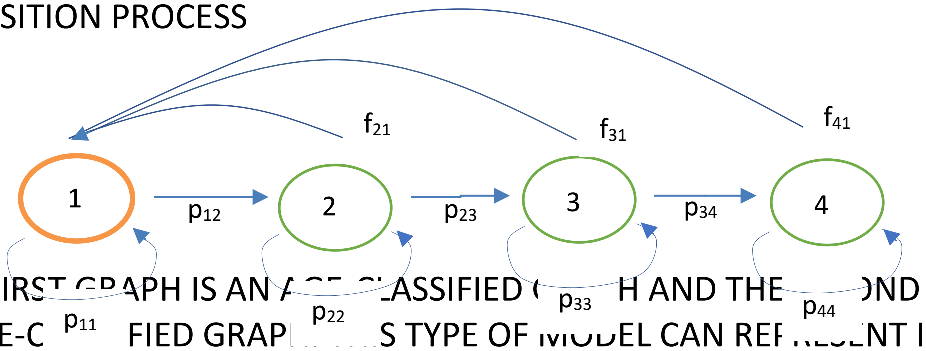


FIGURE 12. EXAMPLE OF A LIFE-CYCLE GRAPH FOR A STAGE-CLASSIFIED TRANSITION PROCESS



THE FIRST GRAPH IS AN AGE-CLASSIFIED GRAPH AND THE SECOND GRAPH IS A STAGE-CLASSIFIED GRAPH. THIS TYPE OF MODEL CAN REPRESENT INTERNAL MIGRATION, BY REPRESENTING REGIONS AS STATES.

THE POPULATION-TRANSITION MATRIX (“LESLIE MATRIX”) FOR THE AGE-CLASSIFIED LIFE-CYCLE MODEL OF FIGURE 11 IS

$$A_1 = \begin{pmatrix} 0 & f_{21} & f_{31} & f_{41} \\ p_{12} & 0 & 0 & 0 \\ 0 & p_{23} & 0 & 0 \\ 0 & 0 & p_{34} & 0 \end{pmatrix}.$$

THE POPULATION-TRANSITION MATRIX FOR THE STAGE-CLASSIFIED LIFE-CYCLE MODEL OF FIGURE 12 IS

$$A_2 = \begin{pmatrix} p_{11} & f_{21} & f_{31} & f_{41} \\ p_{12} & p_{22} & 0 & 0 \\ 0 & p_{23} & p_{33} & 0 \\ 0 & 0 & p_{34} & p_{44} \end{pmatrix}.$$

POPULATION-PROJECTION MATRICES THAT REPRESENT STAGE OR SIZE STRUCTURE OF A POPULATION LIFE-CYCLE (INSTEAD OF AGE, AS IN A LESLIE MATRIX) ARE OFTEN CALLED LEFKOVITCH MATRICES, AFTER BIOLOGIST L. P. LEFKOVITCH.

THE STATE OF A POPULATION AT A TIME t STEPS INTO THE FUTURE IS OBTAINED EXACTLY AS DESCRIBED EARLIER, BY PRE-MULTIPLYING THE POPULATION STATE VECTOR (I.E., THE VECTOR SPECIFYING THE NUMBER OF INDIVIDUALS IN EACH COMPONENT OF THE STATE VECTOR) BY THE POPULATION-PROJECTION MATRIX:

$$P_t = A^t P_0,$$

WHERE A DENOTES THE POPULATION-PROJECTION MATRIX AND P_t DENOTES THE POPULATION AT TIME t.

STABLE POPULATION

IN THE FIELD OF STOCHASTIC PROCESSES, A *STATIONARY* DISTRIBUTION IS ONE FOR WHICH THE JOINT DISTRIBUTION OF THE PROCESS FOR A SEQUENCE OF TIMES IS THE SAME IF SHIFTED BY ANY AMOUNT. THIS SAME DEFINITION APPLIES TO POPULATION PROCESSES – A POPULATION PROCESS IS STATIONARY IF THE POPULATION DISTRIBUTION DOES NOT CHANGE OVER TIME.

A POPULATION PROCESS IS SAID TO BE *STABLE* IF ITS DEMOGRAPHIC RATES (BIRTH, DEATH, IMMIGRATION) ARE UNCHANGING OVER TIME, ALTHOUGH IT MAY BE GROWING (OR SHRINKING).

ALFRED LOTKA PROVED, IN 1939, THAT IF THE FOLLOWING THREE CONDITIONS HELD, THEN A POPULATION PROCESS WOULD EVOLVE TO A STABLE POPULATION:

1.THE GROWTH RATE IN THE ANNUAL NUMBER OF BIRTHS IS CONSTANT
2. AGE-SPECIFIC DEATH RATES (I.E., THE LIFE TABLE) ARE CONSTANT
3. AGE-SPECIFIC RATES OF NET MIGRATION ARE ZERO

THE THIRD REQUIREMENT MAY BE RELAXED: A STABLE POPULATION RESULTS IF AGE-SPECIFIC MIGRATION RATES ARE CONSTANT OVER TIME.

A STABLE POPULATION IS OF INTEREST IN DEMOGRAPHY SINCE IT SHOWS THE ULTIMATE RESULT OF SPECIFIED BIRTH AND DEATH RATES, IN THE CASE OF NO MIGRATION.

THE SIGNIFICANT IMPLICATIONS OF THE PRECEDING RESULT ARE SEVERAL:

POPULATIONS WITH UNCHANGING VITAL RATES ARE STABLE
LONG-TERM POPULATION CHARACTERISTICS ARE DEFINED BY THE VITAL RATES
EVERY POPULATION’S SET OF AGE-SPECIFIC VITAL RATES IMPLIES AN UNDERLYING STABLE POPULATION THAT WILL EMERGE IF THOSE RATES REMAIN UNCHANGED

LOTKA’S RESULT IMPLIES THAT CONDITION (1) SPECIFIED ABOVE MAY BE REPLACE BY THE CONDITION:

1’. AGE-SPECIFIC FERTILITY RATES ARE CONSTANT

WE SHALL NOW DESCRIBE THE STABLE DISTRIBUTION, USING MATRIX TERMINOLOGY. THE PRESENTATION FOLLOWS THAT IN *APPLIED MATHEMATICAL DEMOGRAPHY* 3rd ed., BY KEYFITZ AND CASWELL, OR *MATRIX POPULATION MODELS*, 2nd ed., BY CASWELL.

THE MATRIX POPULATION MODEL IS

$$\boldsymbol{n}(t+1) = \boldsymbol{A}\boldsymbol{n}(t)$$

WHERE **n**(t) IS THE POPULATION VECTOR AND **A** IS A STAGE-CLASSIFIED PROJECTION MATRIX. THE MATRIX **A** IS SQUARE. LET US ASSUME THAT THE NUMBER OF ROWS (OR COLUMNS) IS s. LET US DENOTE **n**(0) AS $\mathbf{n}_0$.

WE RECALL FACTS ABOUT EIGENVALUES AND EIGENVECTORS. FOR ANY MATRIX **A**, A VECTOR **w** IS A RIGHT EIGENVECTOR OF A AND THE SCALAR λ IS THE CORRESPONDING EIGENVALUE IF

$$\boldsymbol{A}\boldsymbol{w} = \lambda\boldsymbol{w},$$

OR

$$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{w} = \boldsymbol{0},$$

WHERE **I** IS AN IDENTITY MATRIX AND **0** IS A VECTOR OF ZEROS. A NONZERO SOLUTION FOR **w** EXISTS ONLY IF THE MATRIX (**A** – λ**I**) IS SINGULAR, THAT IS

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \boldsymbol{0}.$$

ASSOCIATED WITH EACH EIGENVALUE, λ, THERE IS A LEFT EIGENVECTOR **v** SATISFYING

$$v^{*}\boldsymbol{A} = \lambda v^{*},$$

WHERE **v*** DENOTES THE COMPLEX CONJUGATE TRANSPOSE OF **v**.

FROM HERE ON, WE SHALL ASSUME THAT THE EIGENVECTORS ARE LINEARLY INDEPENDENT. THIS CONDITION WILL HOLD IF THE MATRIX A IS OF FULL RANK (I.E., NONSINGULAR), OR IF THE EIGENVALUES ARE ALL DISTINCT.

WE SHALL NOW DERIVE AN EXPRESSION FOR THE MATRIX POPULATION MODEL IN TERMS OF EIGENVALUES AND EIGENVECTORS.

SINCE THE EIGENVECTORS ARE LINEARLY INDEPENDENT, THEY FORM A BASIS FOR s-DIMENSIONAL EUCLIDEAN VECTOR SPACE, AND WE MAY WRITE THE VECTOR $\mathbf{n}_0$ AS

$$\boldsymbol{n}_0 = c_1\boldsymbol{w}_1 + c_2\boldsymbol{w}_2 + \cdots + c_s\boldsymbol{w}_s,$$

OR, IN MATRIX NOTATION,

$$\boldsymbol{n}_0 = \boldsymbol{Wc},$$

WHERE **W** DENOTES THE MATRIX WHOSE COLUMNS ARE THE VECTORS $\mathbf{w}_i$, **W** = ($\mathbf{w}_1$, $\mathbf{w}_2$,…, $\mathbf{w}_s$) AND **c**' = (c_1, c_2, …, c_s).

SINCE THE $\mathbf{w}_i$ ARE LINEARLY INDEPENDENT, THE MATRIX **W** IS NONSINGULAR AND MAY BE INVERTED. HENCE WE MAY WRITE

$$\boldsymbol{c} = \boldsymbol{W}^{-1}\boldsymbol{n}_0.$$

WE HAVE

$$\boldsymbol{n}(1) = \boldsymbol{A}\boldsymbol{n}_0 = \sum_i c_i\boldsymbol{A}\boldsymbol{w}_i = \sum_i c_i\lambda_i\boldsymbol{w}_i$$

$$\boldsymbol{n}(2) = \boldsymbol{A}\boldsymbol{n}(1) = \boldsymbol{A}\sum_i c_i\lambda_i\boldsymbol{w}_i = \sum_i c_i\lambda_i\boldsymbol{A}\boldsymbol{w}_i = \sum_i c_i\lambda_i^2\boldsymbol{w}_i,$$

AND, IN GENERAL,

$$\boldsymbol{n}(t) = \boldsymbol{A}\boldsymbol{n}(t-1) = \sum_i c_i\lambda_i^t\boldsymbol{w}_i.$$

LET US ASSUME THAT THE λ_i ARE ALL DISTINCT, AND LET US DENOTE THE ONE HAVING LARGEST ABSOLUTE VALUE AS λ_1. IN THIS CASE, AS t INCREASES, THE EXPRESSION ON THE RIGHT-HAND-SIDE OF THE PRECEDING EQUATION IS DOMINATED BY THE TERM $c_1\lambda_1^t\boldsymbol{w}_1$. HENCE, IN THE LONG TERM, THE POPULATION GROWTH RATE APPROACHES $c_1\lambda_1^t$, AND THE POPULATION STRUCTURE APPROACHES PROPORTIONALITY TO $\boldsymbol{w}_1$.

IF WE WRITE

$$c_1\lambda_1^t = c_1e^{rt}$$

THEN THE ANNUALIZED GROWTH RATE IS r = ln(λ_1). THE RATE OF GROWTH FOR THE STABLE POPULATION HAS A NUMBER OF NAMES, INCLUDING THE INTRINSIC RATE OF NATURAL INCREASE, LOTKA'S PARAMETR, LOTKA'S r, AND THE MALTHUSIAN PARAMETER.

FOR A STABLE POPULATION, THE AGE PYRAMID HAS A CONSTANT SHAPE (I.E., THE RELATIVE NUMBERS IN EACH AGE LEVEL OF THE PYRAMID REMAIN CONSTANT), BUT THE POPULATION SIZE GROWS AT RATE r.

THERE IS A SIMPLE APPROXIMATION FOR THE VALUE OF r, WHICH DOES NOT DEPEND ON PROJECTIONS. IT IS

$$r \approx \frac{\ln{(NRR)}}{\mu}$$

WHERE μ IS THE COHORT MEAN AGE OF CHILDBEARING AND NRR IS THE NATURAL REPRODUCTION RATE, DEFINED EARLIER AS

> NRR = (TOTAL NUMBER OF DAUGHTERS BORN TO COHORT MEMBERS) / (INITIAL NUMBER OF WOMEN IN THE COHORT).

THE PRECEDING DERIVATION MAY BE PRESENTED IN TERMS OF MATRIX OPERATIONS INSTEAD OF SUMS. THE POPULATION MODEL IS

$$\boldsymbol{n}(t) = \boldsymbol{A^t n_0}$$

SINCE THE MATRIX **A** HAS BEEN ASSUMED TO BE POSITIVE DEFINITE, THERE EXISTS A NONSINGULAR MATRIX **W** SUCH THAT

$$\boldsymbol{W}^{-1}\boldsymbol{AW} = \boldsymbol{\Lambda}$$

WHERE **Λ** IS A DIAGONAL MATRIX WHOSE DIAGONAL ENTRIES ARE THE EIGENVALUES λ_i.

WE HAVE

$$\boldsymbol{A} = \boldsymbol{W\Lambda W}^{-1},$$

HENCE

$$\boldsymbol{A}^2 = \boldsymbol{W\Lambda W}^{-1}\boldsymbol{W\Lambda W}^{-1} = \boldsymbol{W\Lambda}^2\boldsymbol{W}^{-1},$$

AND, IN GENERAL,

$$\boldsymbol{A}^t = \boldsymbol{W\Lambda}^t\boldsymbol{W}^{-1}.$$

THE EQUATION

$$\boldsymbol{W}^{-1}\boldsymbol{AW} = \boldsymbol{\Lambda}$$

IMPLIES

$$\boldsymbol{AW} = \boldsymbol{W\Lambda},$$

THAT IS, THE COLUMNS OF **W** ARE THE RIGHT EIGENVECTORS $\mathbf{w}_i$ OF **A**.

IT ALSO IMPLIES

$$\boldsymbol{W}^{\boldsymbol{-1}}\boldsymbol{A} = \boldsymbol{\Lambda W}^{\boldsymbol{-1}},$$

SO THAT THE ROWS OF $\mathbf{W}^{-1}$ ARE THE COMPLEX CONJUGATES OF THE LEFT EIGENVECTORS $\mathbf{v}_i$ OF **A**.

HENCE WE CAN WRITE THE POPULATION MODEL AS

$$\boldsymbol{n}(t) = \boldsymbol{A}^{\boldsymbol{t}}\boldsymbol{n}_{\boldsymbol{0}} = \boldsymbol{W}\boldsymbol{A}^{\boldsymbol{t}}\overline{\boldsymbol{V}}\boldsymbol{n}_{\boldsymbol{0}} = \sum_{\boldsymbol{i}} \lambda_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{w}_{\boldsymbol{i}}\boldsymbol{v}_{\boldsymbol{i}}^{*}\boldsymbol{n}_{\boldsymbol{0}}$$

WHERE $\boldsymbol{v}_i^*$ IS THE COMPLEX CONJUGATE TRANSPOSE OF THE LEFT EIGENVECTOR CORRESPONDING TO λ_i. THE PRODUCT $\boldsymbol{w}_i \boldsymbol{v}_i^*$ IS A MATRIX CALLED THE CONSTITUENT MATRIX OF A. THE PRODUCT $\boldsymbol{v}_i^* \boldsymbol{n}_0$ IS A SCALAR, EQUAL TO THE c_i GIVEN EARLIER.

THE STABLE POPULATION MAY BE USED AS A BASIS FOR MAKING LONG-TERM PROJECTIONS OR FORECASTS OF POPULATION SIZE AND COMPOSITION, CONDITIONAL ON THE ASSUMED VALUES OF MORTALITY, FERTILITY AND MIGRATION RATES.

17. MORE ON POPULATION-BASED FORECASTS

EARLIER, MATERIAL WAS PRESENTED ON USE OF THE BASIC SYNTHETIC-ESTIMATION METHOD FOR CONSTUCTING POPULATION-BASED FORECASTS.

THIS SECTION (OPTIONAL) PRESENTS MATERIAL ON MORE COMPLEX METHODS OF INDIRECT ESTIMATION, SUCH AS REGRESSION-TYPE ESTIMATES AND GENERAL LINEAR MODELS. THIS SECTION DRAWS ON MATERIAL PRESENTED IN THE PRESENTATION, *SMALL AREA ESTIMATION*.

18. ESTIMATION OF DEMOGRAPHIC PARAMETERS BY INDIRECT METHODS: NONPARAMETRIC METHODS

THE BASIC DEMOGRAPHIC PARAMETERS FOR WHICH ESTIMATES ARE NEEDED TO PERFORM DEMOGRAPHIC ANALYSIS INCLUDE MORTALITIES, FERTILITIES, IMMIGRATION, AND EMIGRATION. FOR BASIC APPLICATIONS, SUCH AS CONSTRUCTING A POPULATION PROJECTION FOR A COUNTRY, ESTIMATES OF THESE PARAMETERS ARE AVAILABLE FOR MANY COUNTRIES, REGIONS, AND YEARS.

IN SOME APPLICATIONS, THE REQUIRED ESTIMATES ARE NOT AVAILABLE, AND MUST BE ESTIMATED. THIS SECTION DESCRIBES SOME OF THE TOOLS THAT ARE AVAILABLE FOR THAT PURPOSE.

THIS SECTION DOES NOT DESCRIBE DIRECT METHODS OF ESTIMATING DEMOGRAPHIC PARAMETERS USING DATA AVAILABLE FROM VITAL

REGISTRATION DATA. SUCH ESTIMATES ARE TYPICALLY BASED ON DATA FOR ENTIRE POPULATIONS (FOR PARTICULAR COUNTRIES, SUBREGIONS OR YEARS), AND DO NOT REQUIRE APPLICATION OF STATISTICAL TECHNIQUES FOR ESTIMATION FROM SAMPLE DATA.

NOR DOES THIS SECTION INCLUDE DESCRIPTION OF CONSTRUCTING ESTIMATES FROM LIFE TABLES BY MEANS OF INTERPOLATION, GRADUATION OR SMOOTHING (THESE ARE MORE CUSTOMARILY ADDRESSED IN A COURSE ON ACTUARIAL METHODS THAN DEMOGRAPHY).

INDIRECT METHODS ("BRASS-TYPE ESTIMATION") FOR ESTIMATING PARTICULAR DEMOGRAPHIC PARAMETERS

EXECUTION OF THE TECHNIQUES OF DEMOGRAPHIC ANALYSIS REQUIRES HIGH-QUALITY AGE-SPECIFIC INFORMATION ABOUT MORTALITY, NATALITY, IMMIGRATION AND EMIGRATION. THIS INFORMATION IS CONSTRUCTED FROM SOURCES SUCH AS VITAL REGISTRATION SYSTEMS AND CENSUSES. SUCH DATA ARE AVAILABLE FOR ECONOMICALLY DEVELOPED COUNTRIES, BOTH AT THE COUNTRY LEVEL AND FOR INTRA-COUNTRY REGIONS. THE DATA ARE AVAILABLE FROM COUNTRY STATISTICAL AGENCIES, FROM THE UNITED NATIONS, FROM THE U.S. CENSUS BUREAU, FROM THE U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT, AND OTHER SOURCES.

WHEN ESTIMATES OF MORTALITY, NATALITY, IMMIGRATION AND EMIGRATION ARE DERIVED FROM OBSERVATIONS ON DEATHS, BIRTHS, AND MIGRATION NUMBERS, THE ESTIMATES ARE CALLED "DIRECT" ESTIMATES.

FOR SOME COUNTRIES AND REGIONS, ADEQUATE DATA ARE NOT AVAILABLE. TO PERFORM DEMOGRAPHIC ANALYSIS FOR SUCH AREAS, IT IS NECESSARY TO ESTIMATE THE REQUIRED PARAMETERS. ONE APPROACH TO GENERATION OF MORTALITY DATA, SUCH AS A LIFE TABLE, IS TO SELECT A "MODEL" LIFE TABLE THAT MATCHES THE AREA OF INTEREST ON ONE OR MORE KNOWN DEMOGRAPHIC CHARACTERISTIC. MODEL LIFE TABLES ARE AVAILABLE FROM THE U.N. AND OTHER SOURCES.

IN SOME CASES, EVEN THE BASIC DEMOGRAPHIC DATA THAT MIGHT HELP IDENTIFY AN APPROPRIATE LIFE TABLE ARE NOT AVAILABLE, AND THESE DATA

MUST BE ESTIMATED. FOR SITUATIONS IN WHICH IT IS NOT FEASIBLE TO OBTAIN THE REQUIRED INFORMATION FROM DIRECT ESTIMATES, A NUMBER OF TECHNIQUES HAVE BEEN DEVISED TO CONSTRUCT BY OTHER MEANS. THESE OTHER MEANS ARE REFERRED TO AS “INDIRECT ESTIMATES.”

MANY OF THE TECHNIQUES FOR INDIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS WERE DEVELOPED BY WILLIAM BRASS, AND INDIRECT METHODS OF ESTIMATION OF DEMOGRAPHIC PARAMETERS ARE OFTEN REFERRED TO AS “BRASS-TYPE” ESTIMATES.

THIS SECTION DEALS WITH INDIRECT METHODS FOR DEMOGRAPHIC PARAMETER ESTIMATION BASED ON REPORTS OF KIN SURVIVAL AND OF TWO CENSUSES. THE METHODS DISCUSSED HERE ARE NONPARAMETRIC ESTIMATION PROCEDURES, AS CONTRASTED TO PARAMETRIC STATISTICAL MODELS, SUCH AS REGRESSION MODELS INVOLVING ANALYTICAL REPRESENTATIONS. ESTIMATION OF DEMOGRAPHIC PARAMETERS USING PARAMETRIC STATISTICAL MODELS IS ADDRESSED IN THE NEXT SECTION OF THE PRESENTATION.

HERE FOLLOW SOME EXAMPLES OF APPLICATIONS OF INDIRECT ESTIMATION BASED ON REPORTS OF KIN SURVIVAL AND OF TWO CENSUSES. THE INFORMATION COULD BE OBTAINED FROM A CENSUS OR A SAMPLE SURVEY.

- ESTIMATION OF CHILD MORTALITY FROM INFORMATION ON CHILDREN EVER BORN AND CHILDREN SURVIVING (ESTIMATED FROM TWO SURVEY QUESTIONS TO A WOMAN: THE NUMBER OF LIVE-BORN CHILDREN THEY HAVE GIVEN BIRTH TO; AND THE NUMBER OF THOSE CHILDREN THAT HAVE SURVIVED)
- ESTIMATION OF ADULT MORTALITY USING INFORMATION ON ORPHANHOOD AND WIDOWHOOD
- ESTIMATION OF ADULT MORTALITY BY COMPARING AGE DISTRIBUTIONS AT TWO CENSUSES (ASSUMES A POPULATION THAT IS CLOSED TO MIGRATION, I.E., A POPULATION WITH ZERO NET MIGRATION IN EVERY AGE CATEGORY OF THE LIFE TABLE)
- ESTIMATION OF ADULT MORTALITY FROM INFORMATION ON THE DISTRIBUTION OF DEATHS BY AGE (ASSUMES A STABLE POPULATION, I.E., ONE THAT MAY BE GROWING, BUT FOR WHICH THE PROPORTION OF POPULATION IN EVERY AGE CATEGORY REMAINS CONSTANT)

- ESTIMATION OF FERTILITY BASED ON INFORMATION ABOUT CHILDREN EVER BORN

THE MATHEMATICAL DERIVATION OF THE FORMULAS USED TO IMPLEMENT THE INDIRECT METHODS CONSIDERED IN THIS SECTION IS COMPLICATED, AND WILL NOT BE DESCRIBED IN THIS PRESENTATION.

THE BOOK, *DEMOGRAPHY: MEASURING AND MODELING POPULATION PROCESSES* BY PRESTON, HEUVELINE AND GUILLOT PRESENTS DETAILED EXAMPLES OF SOME OF THE PRECEDING INDIRECT ESTIMATION TECHNIQUES.

A DETAILED DESCRIPTION OF A NUMBER OF INDIRECT TECHNIQUES IS PRESENTED IN THE UNITED NATIONS PUBLICATION, *MANUAL X, INDIRECT TECHNIQUES FOR DEMOGRAPHIC ESTIMATION* (DEPARTMENT OF INTERNATIONAL ECONOMIC AND SOCIAL AFFAIRS, 1983).

COMPUTER SOFTWARE FOR INDIRECT ESTIMATION

THE U.N. MAKES AVAILABLE, FOR FREE, A MICROSOFT WINDOWS-BASED PACKAGE, *MORTPAK*, WHICH IMPLEMENTS INDIRECT ESTIMATION PROCEDURES LIKE THOSE IN *MANUAL X*, AND A NUMBER OF OTHER PROCEDURES FOR DEMOGRAPHIC ANALYSIS, SUCH AS SELECTION OF LIFE TABLES AND POPULATION PROJECTION. (REF: *MORTPAK FOR WINDOWS*, VERSION 4.3 (UNITED NATIONS POPULATION DIVISION DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, POP/SW/MORTPAK/2003 15 September 2003 Update 25 April 2013). The design of the applications in *MORTPAK* as well as the program MATCH has its origins in the United States Census Bureau package, *Computer Programs for Demographic Analysis* (Arriaga, Anderson and Heligman, 1976).)

FROM THE INTRODUCTION TO THE *MORTPAK* PACKAGE DESCRIPTION:

The present volume presents a set of 20 computer programs for undertaking demographic analyses in developing countries, including empirical and model life-table construction, graduation of mortality data, mortality and fertility estimation, evaluation of census coverage and age distributions and population projections. The 20 demographic procedures included have been selected by the Population Division as useful for evaluating demographic data from censuses and surveys and preparing reliable estimates of demographic parameters. These procedures incorporate techniques for evaluation and estimation of demographic data, particularly those techniques that incorporate the United Nations model life-table system (United Nations, 1982) and generalized stable population equations (Preston and Coale, 1982).

…When selecting a new application from the menu, a window in table form presents a brief description of the procedures, categorized according to their major functions: life-table and stable population construction, model life table construction, graduation of mortality data, indirect mortality estimation, indirect fertility estimation, other estimation procedures and population projections. The package emphasizes mortality estimation, reflecting the larger number of techniques available and the further advanced mortality estimation is, compared to that of other demographic components. (Of the nine chapters in the United Nations manual on *Indirect Techniques for Demographic Estimation* (United Nations, 1983), five are dedicated solely. and two partially, to mortality analysis.)

INDIRECT METHODS FOR SELECTING ENTIRE MODEL LIFE TABLES

THE PRECEDING DISCUSSION FOCUSED ON ESTIMATION OF PARTICULAR DEMOGRAPHIC PARAMETERS. A COMMON TASK IN DEMOGRAPHIC ANALYSIS IS TO SELECT AN ENTIRE MODEL LIFE TABLE, BASED ON VALUES OF A FEW DEMOGRAPHIC PARAMETERS.

THE *MORTPAK* PACKAGE CONTAINS TWO PROGRAMS THAT SELECT MODEL LIFE TABLES BASED ON A FEW PARAMETER VALUES. THESE PROGRAMS, DESCRIBED IN THE *MORTPAK* MANUAL, ARE CALLED *BESTFT* AND *MATCH*.

HERE FOLLOWS AN EXTRACT FROM THE DESCRIPTION OF *BESTFT* PROVIDED IN THE *MORTPAK* MANUAL:

Purpose of procedure: To find the one-, two- or three-component United Nations or Coale-Demeny model life table which best fits one or more probabilities of dying (q(x,n) values) or m(x,n) given as input.

Description of technique: Using least squares criteria, the United Nations model life table of a given pattern is found which best fits one or more q(x,n) values given as input. Simply, the procedure is one of graduation with respect to a standard. When only one q(x,n) value is given, this program presents results identical to that of the procedure MATCH. The one-component model life table (i.e., those presented in United Nations, 1982, annex I) is presented, as well as the adjusted two- and three-component tables. However, at least two q(x,n) values must be given for estimation of the two-component table and at least three values for the three-component table. In place of the United Nations model, an alternative model supplied by the user can be given as input and the best fit of the empirical data to that model will be calculated (for a more detailed description of the methodology, see United Nations, 1982, chap. IV). Starting with version 4.3, new United Nations or new Coale-Demeny models can be selected. The level of these new models is determined by the closest fit to the input data before the best fit regressions are applied. The new models also permit m(x,n) to be selected as input.

The model life table is constructed from choices specified by the user. The choices are: User-defined model; UN Latin American model; UN Chilean; UN South Asian; UN Far East Asian; UN General; Coale-Demeny West; Coale-Demeny North; Coale-Demeny East; Coale-Demeny South. If “User-defined” is selected, the user is supplying the average pattern of mortality to be used as a model (see user-defined model q(x,n) values below). The United Nations principal component equations are then used to adjust this pattern to the desired mortality level.

HERE FOLLOWS AN EXTRACT FROM THE DESCRIPTION OF *MATCH* PROVIDED IN THE *MORTPAK* MANUAL:

Purpose of procedure: Calculates and prints out United Nations, Coale-Demeny or user-designated model life tables corresponding to given levels of mortality. As the user-designated model can be a mortality pattern specific to a certain population, MATCH can generate a country-specific model life table system.

Description of technique: The user must designate the model pattern (any of the five United Nations, four Coale-Demeny patterns or an external model supplied by the user) and the sex desired. The United Nations principal component equations (United Nations, 1982, p. 8) or Coale-Demeny regression equations (Coale-Demeny, 1966, p. (21)) are then used with an iterative procedure to find the model corresponding to a given level of mortality. The iterative procedure is described in United Nations, 1982 (pp. 22-23). However, because of potential extrapolation problems, model life tables are calculated only when life expectancy at birth is between 20 and 80 years. When a user-defined pattern is used, it is permitted to go up to 90 years. The mortality level is specified by the user by designating a mortality value for one of four life table functions ($_nm_x$, $_nq_x$, lx or ex) for any one of the age groups. The iterative procedure is carried out by the procedure MATCH, which calls the procedures LIFTB and, when necessary, ICM for construction of the model life table itself. The model life table is presented as computer output; the life table columns are as given in the description of the procedure LIFTB. Starting with 4.3, while this application still contains the traditional (i.e. unchanged) model life tables from previous versions, new enhanced life tables were added. The new model life tables have age patterns which were updated and have been expanded to include life expectancies at birth to age 100. When matching on q(x,n) or m(x,n), the value of "n" can now be selected as input. If “n” is zero or blank, its value will default to the interval between the current and next age group.

THIS SECTION OF THE PRESENTATION INCLUDES DEMONSTRATION OF THE *MORTPAK* COMPUTER SOFTWARE TO PERFORM INDIRECT ESTIMATION OF DEMOGRAPHIC PARAMETERS.

19. ESTIMATION OF DEMOGRAPHIC PARAMETERS BY INDIRECT METHODS: PARAMETRIC METHODS

SURVIVAL FUNCTION, FAILURE DENSITY FUNCTION, FORCE OF MORTALITY

MUCH OF THE DISCUSSION IN THE FIRST PART OF THE PRESENTATION INVOLVED USE OF PARAMETERS PRESENTED IN A LIFE TABLE. THE KEY PARAMETER OF THE TABLE IS THE SURVIVORSHIP FUNCTION, OR NUMBER OF SURVIVORS REMAINING OUT OF AN INITIAL COHORT AFTER A SPECIFIED NUMBER (x) OF YEARS.

IN A LIFE TABLE, A SIZE IS SPECIFIED FOR THE YOUNGEST COHORT. THAT NUMBER IS CALLED THE RADIX OF THE TABLE, AND IT IS USUALLY A LARGE NUMBER, SUCH AS 100,000. THE PARAMETER ℓ_x SPECIFIES THE NUMBER OF SURVIVORS AT TIME x OUT OF THE RADIX, ℓ_0. THE VALUE OF THE RADIX IS

ARBITRARY – SOME DEMOGRAPHIC PARAMETERS DEPEND ON IT, AND SOME DO NOT. FOR ANALYTICAL WORK, IT IS CONVENIENT TO ELIMINATE CONSIDERATION OF THE RADIX, AND WORK WITH THE *PROPORTION* OF SURIVORS AT TIME x RATHER THAN THE *NUMBER* OF SURVIVORS AT TIME x. THE PROPORTION OF SURVIVORS, DENOTED P_x, IS SIMPLY ℓ_x DIVIDED BY THE RADIX, THAT IS:

$$P_x = \ell_x/\ell_0.$$

THE PROPORTION OF SURVIVORS AS A FUNCTION OF TIME, x, IS CALLED THE SURVIVORSHIP FUNCTION (OR SURVIVAL FUNCTION). (WHEN DEALING WITH PROBABILITY DISTRIBUTIONS, THE SURVIVAL FUNCTION CORRESPONDING TO A RANDOM VARIABLE X WILL BE DENOTED AS $S_X(x)$. THE NOTATION P_x IS USED WHEN NO DISTRIBUTION IS SPECIFIED.)

THE LIFE TABLE IS CALLED A “NONPARAMETRIC” REPRESENTATION OF THE RELATIONSHIP OF SURVIVAL TO AGE, SINCE IT DOES NOT INVOLVE CONSIDERATION OF PARAMETERS (CONSTANTS) THAT MIGHT SPECIFY AN ANALYTICAL FORM FOR THAT RELATIONSHIP IN THE POPULATION. THIS TERMINOLOGY IS CONFUSING, SINCE THE QUANTITIES PRESENTED IN A LIFE TABLE (SUCH AS NUMBER OF SURVIVORS, OR PROBABILITY OF DEATH IN AN AGE INTERVAL) ARE COMMONLY REFERRED TO AS PARAMETERS OF THE LIFE TABLE.

THE LIFE TABLE IS VERY GENERAL. IT DESCRIBES THE RELATIONSHIP OF THE NUMBER OF SURVIVORS TO AGE, NO MATTER WHAT THE NATURE OF THAT RELATIONSHIP. SINCE THE NUMBER OF AGE CATEGORIES IN A LIFE TABLE IS ON THE ORDER OF 20, HOWEVER, IT IS A SOMEWHAT CUMBERSOME REPRESENTATION. IF THE TABLE INVOLVES POPULATION FEATURES ADDITIONAL TO AGE, SUCH AS SEX, RACE, OR REGION, THE NUMBER OF PARAMETERS DEFINING THE TABLE BECOMES QUITE LARGE. FOR SOME APPLICATIONS, SUCH AS DETERMINATION OF INSURANCE RATES OR MAKING POPULATION PROJECTIONS, THIS IS NOT A SERIOUS PROBLEM. IN OTHER APPLICATIONS, SUCH AS CLINICAL TRIALS, INTEREST FOCUSES ON ESTIMATION OF MORTALITY AS A FUNCTION OF MANY VARIABLES, AND THE LIFE-TABLE REPRESENTATION IS, IN GENERAL, NOT AN EFFICIENT OR CONVENIENT ONE.

IN THE PRECEDING SECTION, A PROCEDURE (THE KAPLAN-MEIER PRODUCT-LIMIT ESTIMATOR) WAS DESCRIBED FOR CONSTRUCTING A NONPARAMETRIC ESTIMATE

FOR THE SUVIVAL FUNCTION. THE SIGNIFICANT ADVANTAGE OF A NONPARAMETRIC ESTIMATE IS THAT IT CAN REPRESENT COMPLICATED SHAPES. A DISADVANTAGE IS THAT IT DOES NOT TAKE INTO ACCOUNT KNOWN FEATURES OF THE SURVIVAL FUNCTION, SUCH AS HIGH INFANT MORTALITY, THE "BUMP" IN MORTALITY IN THE EARLY 20'S, AND THE LOG-LINEARLY INCREASING MORTALITY RATE AFTER AGE 40. IT DOES NOT TAKE INTO ACCOUNT THAT THE MORTALITY RATES IN DIFFERENT AGE CATEGORIES ARE INTERRELATED.

FOR SOME PURPOSES, IT DOES NOT MATTER A LOT THAT THE NONPARAMETRIC ESTIMATOR DOES NOT TAKE INTO ACCOUNT INTERRELATIONSHIPS AMONG MORTALITY RATES AT DIFFERENT AGES. FOR OTHER APPLICATIONS, SUCH AS CONSTRUCTION OF CONDITIONAL FORECASTS, IT MAY MATTER VERY MUCH.

THE ESTIMATION MODEL SHOULD TAKE INTO ACCOUNT WHAT IS KNOWN ABOUT THE SURVIVAL FUNCTION, APART FROM SAMPLE DATA. THE PRIMARY MEANS BY WHICH SUCH INFORMATION IS SPECIFIED IS TO SPECIFY A STRUCTURAL MODEL, THAT IS, A MODEL WHOSE FORM AND FEATURES CAN ACCOMMODATE WHAT IS KNOWN ABOUT SURVIVAL FUNCTIONS.

OVE THE YEARS, A VARIETY OF STRUCTURAL MODELS HAVE BEEN CONSIDERED FOR REPRESENTING SURVIVAL AND MORTALITY FUNCTIONS. THIS SECTION WILL DESCRIBE SOME OF THEM. SOME OF THESE MODELS ARE PARAMETRIC MODELS, AND SOME ARE SEMI-PARAMETRIC (I.E., SOME PARAMETERS MAY BE ESTIMATED IN A "DISTRIBUTION FREE" WAY, AND OTHERS MAY RELATE TO DISTRIBUTION STRUCTURE).

IN ORDER TO DETERMINE EFFICIENT PARAMETRIC REPRESENTATIONS OF THE RELATIONSHIP OF MORTALITY TO OTHER VARIABLES, IT IS USEFUL TO UTILIZE THE METHODS OF PROBABILITY AND STATISTICS, AND THAT APPROACH WILL BE TAKEN.

DEMOGRAPHIC APPLICATIONS FOCUS ON EVENTS SUCH AS DEATH, FAILURE, RELAPSE, ACQUIRING A DISEASE, OR CHANGE IN STATUS. WHILE A PROBABILITY DISTRIBUTION MAY BE SPECIFIED TO DESCRIBE THE OCCURRENCE OF SUCH A FAILURE EVENT, DESCRIPTION OF THE PROCESS UNDER STUDY IS OFTEN EXPRESSED IN TERMS OF SURVIVAL THAN IN TERMS OF FAILURE (OR MORTALITY).

THE DISCUSSION THAT FOLLOWS IS DESCRIBED IN GREATER DETAIL IN THE BOOK *STATISTICAL METHODS FOR SURVIVAL DATA ANALYSIS*, 2nd ed., BY ELISA T. LEE (WILEY, 1992). SEE ALSO *SURVIVAL MODELS AND DATA ANALYSIS* BY REGINA C. ELANDT-JOHNSON AND NORMAN JOHNSON (WILEY, 1980); AND *MATRIX POPULATION MODELS*, 2nd ed., BY HAL CASWELL (OXFORD UNIVERSITY PRESS, 2018).

LET T DENOTE A RANDOM VARIABLE THAT IS A FAILURE TIME. LET F(T) DENOTE THE CUMULATIVE PROBABILITY DISTRIBUTION FUNCTION OF T, AND f(t) DENOTE ITS PROBABILITY DENSITY FUNCTION (WHICH IS ASSUMED TO EXIST). THE PROBABILITY DISTRIBUTION OF T MAY BE CHARACTERIZED BY ANY OF THE THREE FOLLOWING FUNCTIONS: THE SURVIVORSHIP FUNCTION, THE PROBABILITY DISTRIBUTION OF AGE AT DEATH; AND THE HAZARD FUNCTION (AGE-SPECIFIC MORTALITY RATE). THESE FUNCTIONS, AND THE RELATIONSHIPS AMONG THEM, WILL NOW BE DESCRIBED.

IN THE REMAINDER OF THIS SECTION, WE SHALL USE t (FOR "time") TO DENOTE THE AGE OF AN INDIVIDUAL, RATHER THAN x. THE REASON FOR DOING THIS IS THAT WE WILL EVENTUALLY USE THE SYMBOL x TO DENOTE COVARIATES THAT AFFECT MORTALITY.

1. THE SURVIVORSHIP FUNCTION (OR SURVIVAL FUNCTION OR CUMULATIVE SURVIVAL RATE):

S(t) = P(an individual survives longer than t) = P(T > t).

THIS IS ALSO EQUAL TO

1 – P(an individual fails before time t) = 1 – F(t).

IN TERMS OF THE LIFE-TABLE NOTATION, USING ℓ_t TO DENOTE THE NUMBER OF SURVIVORS AT AGE t, THE FUNCTION S(t) IS THE FUNCTION ℓ_t DIVIDED BY THE RADIX OF THE LIFE TABLE:

$$S(t) = \ell_t/\ell_0.$$

A GRAPH OF S(t) IS CALLED A SURVIVAL CURVE. S(t) IS A NONDECREASING FUNCTION WITH S(t) = 1 FOR t = 0 AND S(t) = 0 FOR t = 1.

2. THE PROBABILITY DENSITY FUNCTION:

$$f(t) = \lim_{\Delta \mathrm{f} \to \infty} \frac{P(an\ individual\ dying\ in\ the\ interval\ (t, t + \Delta t)}{\Delta t} = \frac{dF}{dt}.$$

3. THE HAZARD FUNCTION (OR INSTANTANEOUS FAILURE RATE, FORCE OF MORTALITY, MORTALITY RATE, AGE-SPECIFIC MORTALITY RATE, OR AGE-SPECIFIC FAILURE RATE):

$$h(t) = \lim_{\Delta \mathrm{f} \to \infty} \frac{P(an\ individual\ of\ age\ t\ fails\ in\ the\ interval\ (t, t + \Delta t)}{\Delta t}$$

$$= \frac{f(t)}{(1 - F(t))}$$

THE SYMBOL μ IS OFTEN USED TO DENOTE THE HAZARD FUNCTION, RATHER THAN h.

THE CUMULATIVE HAZARD FUNCTION IS DEFINED AS

$$H(t) = \int_0^t h(z)dz.$$

IT CAN BE SHOWN THAT

$$H(t) = -\ln\ S(t).$$

THIS IS EASY TO PROVE:

WE HAVE

$$h(t) = f(t)/S(t)$$

AND

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t).$$

SUBSTITUTING THIS LAST EXPRESSION INTO THE PRECEDING ONE YIELDS

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\ln\bigl(S(t)\bigr).$$

INTEGRATING BOTH SIDES FROM ZERO TO t AND USING S(0)=1, WE OBTAIN

$$-\int_0^t h(z)dz = \ln\bigl(S(t)\bigr)$$

OR

$$H(t) = -\ln\bigl(S(t)\bigr),$$

THE DESIRED RESULT.

WE ALSO HAVE

$$S(t) = exp[[-H(t)] = exp[[-\int_0^t h(z)dz]$$

AND

$$f(t) = h(t)\exp[-H(t)].$$

THE FUNCTION H(t) CAN TAKE ON ANY VALUE FROM ZERO TO INFINITY.

IN ENGINEERING APPLICATIONS, SUCH AS RELIABILITY ANALYSIS, THE FAILURE DISTRIBUTION IS OFTEN WELL REPRESENTED BY THE EXPONENTIAL DISTRIBUTION, WHICH HAS DISTRIBUTION FUNCTION

F(t) = 1 – exp(-ct) (t>=0),

DENSITY FUNCTION

f(t) = F'(t) = c exp(-ct) (t>=0).

AND HAZARD FUNCTION

h(t) = f(t)/(1 – F(t)) = c exp(-ct) / exp(-ct) = c.

THAT IS, FOR AN ITEM FOR WHICH THE FAILURE TIMES ARE REPRESENTED BY AN EXPONENTIAL DISTRIBUTION, THE FAILURE RATE IS A CONSTANT, AND ITS VALUE IS THE PARAMETER OF THE DISTRIBUTION.

THE FAILURE TIMES (TIMES OF DEATH) FOR HUMAN BEINGS DO NOT FOLLOW AN EXPONENTIAL DISTRIBUTION (SINCE THE HAZARD RATE INCREASES WITH AGE), BUT THE SURVIVAL FUNCTION AND HAZARD FUNCTION DO POSSESS SOME DISTINGUISHING FEATURES. A GRAPH OF THE HAZARD FUNCTION FOR HUMAN BEINGS HAS WHAT IS DESCRIBED AS A "BATHTUB" SHAPE: HIGH INITIALLY, THEN LOW FOR A LONG TIME, AND THEN BECOMING HIGH AGAIN.

IF ANY ONE OF THE THREE FUNCTIONS, SURVIVORSHIP FUNCTION, PROBABILITY DENSITY FUNCTION, OR HAZARD FUNCTION, IS KNOWN, THE OTHER TWO CAN BE DERIVED. THE FORMULAS ARE AS FOLLOWS:

IF S(t) IS KNOWN, THEN

$$f(t) = -\frac{dS(t)}{dt}$$

$$h(x) = -\frac{dln\big(S(t)\big)}{dt}.$$

IF f(x) IS KNOWN, THEN

$$S(t) = \int_t^\infty f(z)dz$$

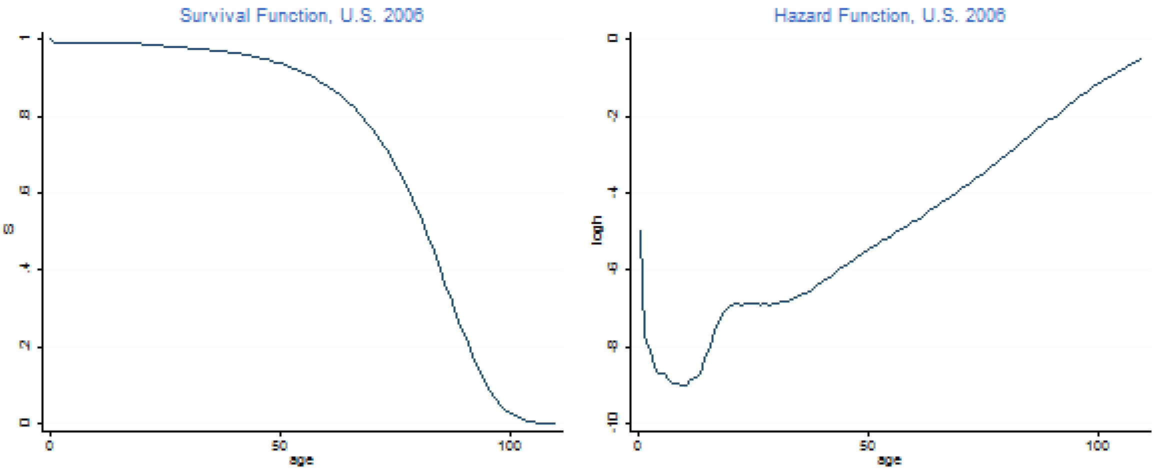
$$h(t) = \frac{f(t)}{\int_t^\infty f(z)dz}.$$

IF h(t) IS KNOWN, THEN

$$S(t) = \exp\,[\int_{t}^{\infty} h(z)dz]$$

$$f(t) = h(t)\exp\left[\int_{t}^{\infty} h(z)dz\right].$$

FIGURE 13 SHOWS A SURVIVAL FUNCTION AND A HAZARD FUNCTION FOR A TYPICAL HUMAN POPULATION.



NOTE THAT THE HAZARD FUNCTION IS PLOTTED USING A LOGARITHMIC SCALE FOR THE ORDINATE. THE INTERESTING OBSERVATION IS THAT ABOVE AGE 40, THE CURVE IS A RELATIVELY STRAIGHT LINE. THAT IS, A PLOT OF THE LOGARITHM OF THE HAZARD FUNCTION IS RELATIVE STRAIGHT, ON A LINEAR SCALE FOR THE ORDINATE. BECAUSE OF THIS FEATURE, MOST ANALYTICAL WORK IS EXPRESSED IN TERMS OF THE HAZARD FUNCTION, RATHER THAN THE SURVIVAL FUNCTION.

THE PRECEDING DISCUSSION HAS INTRODUCED BASIC THEORETICAL CONCEPTS ABOUT THE SURVIVAL FUNCTION AND THE HAZARD FUNCTION. IN THAT DISCUSSION, NO PARTICULAR FORM WAS ASSUMED FOR THESE FUNCTIONS. WE SHALL NEXT CONSIDER PARAMETRIC SPECIFICATIONS FOR THESE FUNCTIONS, BUT BEFORE DOING SO, WE SHALL DISCUSS A NONPARAMETRIC METHOD FOR ESTIMATING THE SURVIVAL FUNCTION.

NONPARAMETRIC ESTIMATION OF THE SURVIVAL FUNCTION

IN SOME CONTEXTS, SUCH AS PRODUCTION OF A LIFE TABLE FROM A COUNTRY'S VITAL REGISTRATION RECORDS, THERE IS MUCH DATA, AND ELABORATE METHODS OF SMOOTHING (OR, AS IT IS CALLED IN DEMOGRAPHY, GRADUATION) MAY BE APPLIED. IN OTHER CONTEXTS, SUCH AS CLINICAL TRIALS, THE AMOUNT OF DATA MAY BE LIMITED. IN SUCH CASES, SIMPLE NONPARAMETRIC METHODS MAY BE APPLIED TO ESTIMATE THE SURVIVAL FUNCTION. THIS SECTION SUMMARIZES THE MOST POPULAR METHOD, THE KAPLAN-MEIER PRODUCT-LIMIT METHOD FOR ESTIMATING A SURVIVAL FUNCTION. FOR DETAILS, REFER TO THE LEE BOOK.

THE KAPLAN-MEIER METHOD MAY BE USED TO ESTIMATE A SINGLE SURVIVAL FUNCTION FROM SAMPLE DATA, OR TO ESTIMATE SURVIVAL FUNCTIONS FOR SUBSAMPLES. SURVIVAL DISTRIBUTIONS FOR INDEPENDENT SAMPLES MAY BE COMPARED USING THE KOLMOGOROV-SMIRNOV TEST.

SUPPOSE THAT WE HAVE A SAMPLE OF n INDIVIDUALS, FOR WHOM WE OBSERVE SURVIVAL TIMES (FAILURE TIMES, TIME FROM INITIATION OF OBSERVATION UNTIL DEATH). LET $t_1, t_2,\ldots,t_n$ DENOTE THE SURVIVAL TIMES. ORDER THE SAMPLE IN INCREASING ORDER AND DENOTE THE SURVIVAL TIMES OF THE ORDERED SAMPLE AS $t_{(1)}, t_{(2)},\ldots,t_{(n)}$. THAT IS, $t_{(1)} \le t_{(2)} \le \ldots \le t_{(n)}$. WHEN ALL OF THE SURVIVAL TIMES ARE KNOWN, THE SURVIVORSHIP FUNCTION MAY BE ESTIMATED AS

$$\hat{S}\left(t_{(i)}\right) = \frac{n-i}{n} = 1 - \frac{1}{n}.$$

IF TWO OR MORE OBSERVATIONS ARE TIED, THE LARGEST VALUE OF (n-i)/n IS USED. SINCE ALL OBSERVATIONS ARE ALIVE AT THE BEGINNING OF THE STUDY AND NO ONE SURVIVES LONGER THAN $t_{(n)}$, WE HAVE

$$\hat{S}\left(t_{(0)}\right) = 1$$

AND

$$\hat{S}(t_n) = 0$$

HENCE WE SEE THAT IF SURVIVAL TIMES ARE OBSERVED FOR ALL INDIVIDUALS, THE ESTIMATION PROCESS IS SIMPLE.

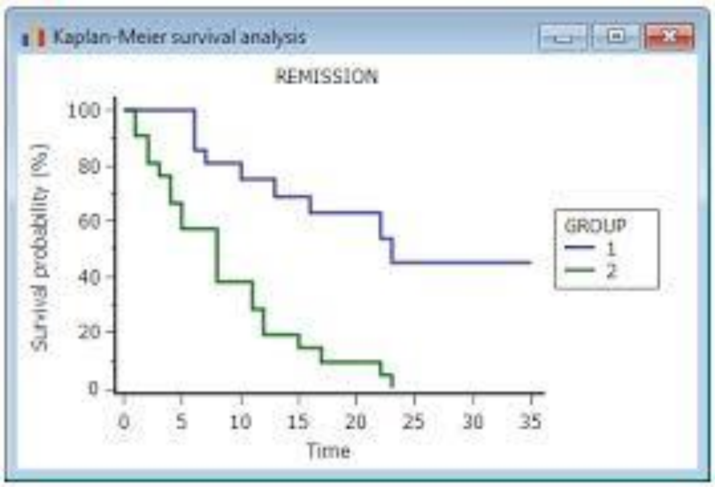
IN MANY APPLICATIONS, IT IS NOT POSSIBLE TO OBSERVE THE SURVIVAL TIMES FOR ALL INDIVIDUALS OF A SAMPLE, BECAUSE SOME LEAVE THE STUDY AND SOME MAY STILL BE ALIVE WHEN THE STUDY IS CONCLUDED. SUCH OBSERVATIONS ARE CALLED “CENSORED” OBSERVATIONS.

FOR THE CASE IN WHICH CENSORED OBSERVATIONS ARE PRESENT, THE MAXIMUM-LIKELIHOOD ESTIMATOR OF THE SURVIVAL FUNCTION IS

$$\hat{S}(t) = \prod_{t(r) \le t} \frac{n-r}{n-r+1}$$

WHERE r RUNS THROUGH THOSE POSITIVE INTEGERS FOR WHICH t(r)≤t AND t(r) IS UNCENSORED. THE PRECEDING ESTIMATOR IS CALLED THE KAPLAN-MEIER PRODUCT-LIMIT ESTIMATOR OF THE SURVIVAL FUNCTION.

FIGURE 14 SHOWS A TYPICAL ESTIMATED SURVIVAL FUNCTION. IT IS SIMPLY THE COMPLEMENT OF THE EMPIRICAL DISTRIBUTION FUNCTION.



THE REASON WHY THE KAPLAN-MEIER ESTIMATOR IS CALLED A PRODUCT-LIMIT ESTIMATOR IS THAT FOR EACH TIME INTERVAL, SAY A YEAR, THE ESTIMATOR, $\hat{S}(t)$, MAY BE REPRESENTED AS THE PRODUCT OF THE OBSERVED SURVIVAL RATE

FOR THE CURRENT YEAR, p_t, TIMES THE ESTIMATED SURVIVAL RATE FROM THE BEGINNING OF THE STUDY THROUGH THE PREVIOUS YEAR, $\hat{S}(t-1)$:

$$\hat{S}(t) = \hat{S}(t-1)pt.$$

PARAMETRIC REPRESENTATIONS OF THE HAZARD FUNCTION

BOTH THE SURVIVAL FUNCTION AND THE HAZARD FUNCTION ARE RATHER COMPLICATED FUNCTIONS OF AGE. MUCH EFFORT HAS BEEN EXPENDED ON TRYING TO FIND EFFICIENT PARAMETRIC REPRESENTATIONS. IN GENERAL, THE HAZARD FUNCTION HAS A SIMPLER FORM THAN THE SURVIVAL FUNCTION, AND PARAMETRIC MODELS ARE DEVELOPED IN TERMS OF THE HAZARD FUNCTION INSTEAD OF THE SURVIVAL FUNCTION.

THE DISCUSSION THAT FOLLOWS IS PRESENTED IN GREATER DETAIL IN *DEMOGRAPHY* BY PRESTON ET AL.

IN 1825, GOMPERTZ IDENTIFIED THAT THE FOLLOWING EXPONENTIAL FUNCTION WAS A GOOD FIT TO MORTALITY RATE FOR ADULTS OVER FORTY:

$$h(t) = \alpha e^{\beta t}$$

THIS RESULT IMPLIES THAT THE LOGARITHM OF THE DEATH RATE IS A LINEAR FUNCTION OF AGE (t):

$$\ln\big(h(t)\big) = \ln(\alpha) + \beta t.$$

IN 1860, MAKEHAM SUGGESTED ADDING A CONSTANT TO THE EXPONENTIAL FUNCTION:

$$h(x) = \gamma + \alpha e^{\beta}.$$

IT WAS OBSERVED THAT DEATH RATES TEND TO DECREASE MORE SLOWLY THAN THE EXPONENTIAL FORMS INDICATE. IN 1932, PERKS SUGGESTED THE LOGISTIC FUNCTION TO REPRESENT MORTALITY:

$$h(t) = \frac{\beta\gamma^t}{1 + \beta\gamma^t}.$$

THIS FORMULA IMPLIES

$$\frac{h(t)}{1 - h(t)} = \beta\gamma^t,$$

OR

$$\ln\left(\frac{h(t)}{1 - h(t)}\right) = \ln(\beta) + \ln(\gamma)\, t.$$

IF p DENOTES A PROBABILITY, THEN p/(1-p) IS THE ODDS. THE LOGIT FUNCTION OF p IS THE LOGARITHM OF THE ODDS, OR LOG-ODDS:

$$logit(p) = \ln\left(\frac{p}{1 - p}\right).$$

PERKS' RESULT STATES THAT THE LOGIT OF h(x) IS A LINEAR FUNCTION OF t.

IN GENERAL, BECAUSE OF THE IRREGULAR NATURE OF THE CURVE, THE FITTING OF MORTALITY FUNCTIONS REQUIRES A SUBSTANTIAL NUMBER OF PARAMETERS.

HERE FOLLOWS AN EXAMPLE OF AN EIGHT-PARAMETER CURVE TO REPRESENT THE PROBABILITY OF DYING BETWEEN AGE x AND x + 1, DEVELOPED BY HELIGMAN AND POLLARD IN 1980:

$$\frac{{}_1q_x}{{}_1p_x} = A^{x+B^C} + De^{-E(ln(x)-ln\,(F)^2} + GH^x.$$

THE THREE TERMS ADDRESS, RESPECTIVELY, INFANT MORTALITY, THE "ACCIDENT HUMP" OF YOUNG ADULTS, AND MORTALITY IN OLD AGE.

ALTHOUGH THIS FUNCTIONAL FORM MAY PROVIDE A GOOD FIT TO A WIDE VARIETY OF MORTALITY CURVES, IT IS NOT A USEFUL REPRESENTATION FOR CONDITIONAL FORECASTING, FOR TWO MAIN REASONS: (1) THE PARAMETERS

DO NOT RELATE DIRECTLY TO THE BASIC DETERMINANTS (PROPERTIES, ESSENTIAL STRUCTURAL FEATURES) OF A POPULATION PROCESS, SUCH AS GROWTH RATES, FERTILITY RATES, MORTALITY RATES, AND THE INTERRELATIONSHIPS AMONG MORTALITY RATES AT DIFFERENT AGES; AND (2) THE INTERRELATIONSHIPS AMONG THE PARAMETERS ARE UNKNOWN. IF IT IS DESIRED TO CONSTRUCT A FORECAST CONDITIONAL ON A CHANGE IN ONE PARAMETER, IT IS NOT KNOWN TO WHAT EXTENT THE OTHER PARAMETERS SHOULD BE CHANGED, TO MAINTAIN DESIRED FEATURES OF THE POPULATION PROCESS (SUCH AS LIFE EXPECTANCY), OR TO MAINTAIN CONSISTENCY AMONG THEM (SUCH AS THE SIMILARITY OF MORTALITIES IN ADJACENT AGE CATEGORIES).

SEMIPARAMETRIC MODELS OF THE HAZARD FUNCTION

WE SHALL NOW IDENTIFY SOME MODELS THAT TAKE INTO ACCOUNT KNOWN FEATURES OF MORTALITY FUNCTIONS.

IN MANY APPLICATIONS, THE HAZARD FUNCTION FOR SUBGROUPS OF A POPULATION (E.G., MALES AND FEMALES, OR SIMILAR COUNTRIES) ARE SEEN TO BE PROPORTIONAL. THIS WAS OBSERVED BY LEHMANN IN 1953. IN 1972, D. R. COX SUGGESTED A CLASS OF MULTIPICATIVE MODELS THAT HAS BECOME VERY POPULAR FOR DESCRIBING MORTALITY AS A FUNCTION OF VARIABALES THAT MAY AFFECT MORTALITY (SUCH AS SUBGROUP MEMBERSHIP). IT IS REFERRED TO AS THE COX PROPORTIONAL-HAZARDS MODEL. THIS MODEL WILL NOW BE DESCRIBED.

SUPPOSE THAT **x** DENOTES A VECTOR OF COVARIATES (OBSERVED VARIABLES THAT MAY AFFECT MORTALITY, BUT ARE NOT AFFECTED BY IT), AND DO NOT VARY WITH AGE (TIME, t). WE WRITE THE HAZARD FUNCTION, CONDITIONAL ON THESE VARIABLES, AS h(t |**x**)

WE SUPPOSE THAT THE HAZARD FUNCTION h(t |$\underline{x}$) MAY BE WRITTEN AS:

$$h(t|\boldsymbol{x}) = h_0(t)g(\boldsymbol{x}).$$

COX DISCUSSED MODELS FOR WHICH THE FUNCTION g(**x**) DEPENDS ON A NUMBER OF PARAMETERS, DENOTED BY **β**, THROUGH AN EXPONENTIAL FUNCTION:

$$g(\boldsymbol{x};\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\boldsymbol{x}),$$

SO THAT

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x})$$

OR

$$ln\ (h(t|\boldsymbol{x})) = ln\ (h_0(t)) + \exp(\boldsymbol{\beta}'\boldsymbol{x}).$$

AN ADVANTAGE OF THE PRECEDING FORM IS THAT THE QUANTITY $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ IS ALWAYS POSITIVE (SO THAT $h(t|\boldsymbol{x})$ IS ALWAYS POSITIVE, AS IT MUST BE).

IT SHOULD BE NOTED THAT THE PRECEDING MODEL IS MULTIPLICATIVE NOT JUST IN TERMS OF THE FACTOR $\exp(\boldsymbol{\beta}'\boldsymbol{x})$, BUT THAT THIS FACTOR ITSELF MAY BE REPRESENTED AS A PRODUCT:

$$\exp(\boldsymbol{\beta}'\boldsymbol{x}) = \exp(\sum_{i=1}^{k}\beta_i x_i) = \exp(\beta_1 x_1)\exp\ (\beta_2 x_2)\ldots\exp(\beta_k x_k).$$

THAT IS, THE EFFECTS OF THE COVARIATES ARE MULTIPICATIVE.

NOTE THAT A KEY FEATURE OF THE PROPORTIONAL-HAZARDS MODEL IS THAT THE MULTIPLICATIVE FACTOR DEPENDS ONLY ON THE COVARIATES, AND NOT ON AGE. THE FACTOR IS THE SAME FOR ALL AGES. IN ESSENCE, THE PROBLEM OF ESTIMATING THE BASELINE HAZARD FUNCTION (BOTH ITS MEAN LEVEL AND SHAPE) HAS BEEN SEPARATED FROM THE PROBLEM OF ESTIMATING THE EFFECTS OF COVARIATES ON IT.

IN TERMS OF THE SURVIVAL FUNCTION, THE PRECEDING MODEL IMPLIES

$$S(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\boldsymbol{x})}.$$

THE GOMPERTZ HAZARD FUNCTION MAY BE WRITTEN AS

$$h(t) = \alpha e^{\beta t}.$$

A MULTIPLICATIVE-MODEL FORM OF THE GOMPERTZ MODEL IS OBTAINED BY REPLACING α WITH α exp(**β'x**):

$$h(t) = \alpha e^{\beta t} \exp(\boldsymbol{\beta}'\boldsymbol{x}).$$

WHILE THE COX PROPORTIONAL-HAZARDS MODEL IS VERY USEFUL, IT IS LIMITED TO INTRODUCING DEVIATIONS FROM A "BASELINE" HAZARD FUNCTION, $h_0(t)$, SUCH AS IN THE PRECEDING GOMPERTZ EXAMPLE. THE PROBLEM OF FINDING A USEFUL REPRESENTATION OF THE BASELINE HAZARD FUNCTION REMAINS.

TO ADDRESS THIS ISSUE, A PARAMETRIC FORM MUST BE ASSUMED FOR THE BASELINE HAZARD FUNCTION, SUCH AS THE GOMPERTZ FUNCTION IN THE PRECEDING EXAMPLE. THE DIFFICULTY OF FINDING AN EFFICIENT PARAMETRIC REPRESENTATION FOR A HAZARD FUNCTION HAS BEEN NOTED. A GENERAL SOLUTION TO THIS PROBLEM IS TO CONSIDER A SEMI-PARAMETRIC MODEL, IN WHICH THE BASELINE HAZARD FUNCTION IS SPECIFIED BY HAZARD LEVELS FOR A NUMBER OF AGE CATEGORIES (SUCH AS EIGHT), AND THE MULTIPLICATIVE FACTOR IS REPRESENTED AS A REGRESSION-TYPE PARAMETRIC MODEL, SUCH AS $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ IN THE PRECEDING DISCUSSION.

THIS APPROACH HAS BEEN EXAMINED IN DETAIL IN THE BOOK, *DEMOGRAPHIC FORECASTING*, BY FEDERICO GIROSI AND GARY KING.

IN THE PRECEDING APPROACH, ALL OF THE PARAMETERS ARE ESTIMATED TOGETHER, AS A SINGLE OPERATION. A CONCEPTUAL SHORTCOMING OF THE PRECEDING MODEL IS THAT THE MEAN MORTALITY LEVELS FOR THE VARIOUS AGE CATEGORIES ARE ESTIMATED WITH NO CONSIDERATION OF THE FACT THAT THEY REPRESENT SUCCESSIVE SEGMENTS OF A SINGLE SURVIVAL DISTRIBUTION – THEY ARE IN ESSENCE ESTIMATED AS THE MEANS OF A NUMBER OF UNRELATED CATEGORIES. IN FACT, THEY ARE INTIMATELY INTERRELATED – THE MORTALITIES OF ADJACENT AGE CATEGORIES ARE SIMILAR; IF ONE CHANGES IN SOME WAY, OTHERS ARE LIKELY TO CHANGE AS WELL. THE UNIVARIATE MODEL CONSIDERED ABOVE DOES ALLOW FOR CORRELATIONS AMONG PARAMETERS, BUT IT DOES NOT RECOGNIZE THE FUNDAMENTAL ASPECT OF THE CATEGORIES AS BEING SEGMENTS OF A SINGLE SURVIVAL DISTRIBUTION.

A WAY OF ADDRESSING THIS SHORTCOMING IS TO VIEW THE CATEGORY MEANS AS COMPONENTS OF A MULTIVARIATE RANDOM VARIABLE, AND TO ESTIMATE THEM USING MULTIVARIATE ANALYSIS. IN THE PRECEDING, THE NUMBER OF AGE CATEGORIES HAS NOT BEEN SPECIFIED. IN PRACTICE, TO REPRESENT AN ARBITRARY MORTALITY CURVE, IT HAS BEEN SEEN THAT ABOUT EIGHT AGE CATEGORIES ARE REQUIRED. THIS IS A LARGE NUMBER OF PARAMETERS, AND IT IS USEFUL TO ADDRESS THE ISSUE OF WHETHER A SMALLER NUMBER OF PARAMETERS COULD SUFFICE.

IN FACT, A METHOD OF CONSTRUCTING A BASELINE HAZARD FUNCTION WITH A SMALLER NUMBER OF PARAMETERS IS AFFORDED BY USING THE METHOD OF PRINCIPAL COMPONENTS. WITH THAT APPROACH, THE BASELINE HAZARD FUNCTION IS ESTIMATED BY ESTIMATED A SMALL NUMBER OF PRINCIPAL COMPONENTS (SUCH AS THREE), RATHER THAN A LARGE NUMBER OF AGE CATEGORIES (SUCH AS EIGHT).

MUCH MORE IMPORTANT THAN THE NUMBER OF PARAMETERS USED TO ESTIMATE THE BASELINE HAZARDS FUNCTION, HOWEVER, IS RECOGNITION OF THE INTERRELATIONSHIPS AMONG THE MORTALITIES AT DIFFERENT AGES. THIS IS RECOGNIZED IN THE PRECEDING APPROACH BY USING THE METHOD OF PRINCIPAL COMPONENTS TO SPECIFY A MORTALITY CURVE. THE VERY SIGNIFICANT ADVANTAGE OF CHARACTERIZING A MORTALITY FUNCTION USING PRINCIPAL COMPONENTS IS THE FACT THAT THEY ARE BOTH STATISTICALLY AND GEOMETRICALLY INDEPENDENT – ONE OF THEM MAY BE CHANGED INDEPENDENTLY OF THE OTHER. THIS FEATURE IS ESSENTIAL FOR FORECASTING APPLICATIONS (OR PROJECTION OR SIMULATION EXERCISES) IN WHICH IT IS DESIRED TO CONSTRUCT CONDITIONAL FORECASTS (OR PROJECTIONS OR SIMULATIONS) CONDITIONAL ON CHANGES TO CERTAIN VARIABLES THAT AFFECT MORTALITY.

THE APPROACH OF ESTIMATING HAZARD FUNCTIONS USING THE METHOD OF PRINCIPAL COMPONENTS WAS POPULARIZED BY LEE AND CARTER, AND IT IS OFTEN REFERRED TO AS THE LEE-CARTER METHOD. THIS APPROACH IS DESCRIBED IN THE GIROSI-KING BOOK.

THE GIROSI-KING METHOD MAY BE USED WITH OR WITHOUT CONSIDERATION OF COVARIATES (ADDITIONAL TO AGE, WHICH IS ALWAYS INCLUDED).

CORRECT ESTIMATION OF GENERAL PARAMETRIC STATISTICAL MODELS REQUIRES A LEVEL OF COMPETENCE IN STATISTICAL INFERENCE THAT IS BEYOND THE BASIC KNOWLEDGE ASSUMED FOR THIS COURSE. MUCH COMPUTER SOFTWARE, BOTH COMMERCIAL AND FREE, IS AVAILABLE IN SUPPORT OF THIS FUNCTION. THIS TOPIC IS NOT ADDRESSED FURTHER IN THIS COURSE, EXCEPT FOR SOME DISCUSSION OF THE GIROSI-KING YourCast FORECASTING SOFTWARE, IN THE NEXT SECTION.

20. STOCHASTIC PROCESSES IN DEMOGRAPHY APPLICATIONS

THE STATISTICAL FIELD OF STOCHASTIC PROCESSES (TIME SERIES ANALYSIS) HAS SEEN LIMITED APPLICATION TO THE FIELD OF DEMOGRAPHY. IT HAS BEEN USED TO A LIMITED DEGREE TO MODEL DEMOGRAPHIC RATES, SUCH AS MORTALITY. FORECASTS OF POPULATION ARE MADE USING THE POPULATION-PROJECTION MODEL DESCRIBED ABOVE, SUBSTITUTING THE ESTIMATED RATES IN THE PROJECTION MATRICES. WHILE THIS PROCEDURE MAY PRODUCE REASONABLE FORECASTS, IT DOES NOT LEND ITSELF TO ASSESSMENT OF BIAS OR MEAN-SQUARED ERROR OF THE FORECASTS.

REFERENCE TEXTS ON THE APPLICATION OF STOCHASTIC PROCESS THEORY TO DEMOGRAPHY

THE FOLLOWING SUMMARIZES THE COVERAGE OF STOCHASTIC PROCESSES IN SELECTED BOOKS ON MATHEMATICAL DEMOGRAPHY:

STOCHASTIC PROCESSES IN DEMOGRAPHY APPLICATIONS BY SUDDHENDA BISWAS AND G. L. SRIWASTAV (NEW CENTRAL BOOK AGENCY, 2006). THIS BOOK PRESENTS A HIGHLY MATHEMATICAL DESCRIPTION OF THE APPLICATION OF STATISTICAL MODELING TO THE ESTIMATION OF DEMOGRAPHIC PARAMETERS. IN THE AREA OF STOCHASTIC PROCESSES, THE BOOK INCLUDES A SHORT SECTION ON THE MATHEMATICAL PROPERTIES OF A CONTINUOUS-TIME BIRTH AND DEATH PROCESS, BUT NO CONSIDERATION OF DISCRETE-TIME TIME SERIES MODELS.

APPLIED MATHEMATICAL DEMOGRAPHY, 3rd ed., BY NATHAN KEYFITZ AND HAL CASWELL (SPRINGER, 2005 (FIRST EDITION 1977). THIS BOOK DOES NOT DISCUSS REPRESENTATION OF POPULATION OR POPULATION PARAMETERS AS A STOCHASTIC PROCESS, AND DOES NOT EVEN CONTAIN THE TERM "STOCHASTIC PROCESS" IN THE INDEX.

MATRIX POPULATION MODELS 2nd ed., BY HAL CASWELL (OXFORD UNIVERSITY PRESS, 2018). THIS BOOK CONTAINS A 72-PAGE CHAPTER, "ENVIRONMENTAL STOCHASTICITY," WHICH CONSIDERS RANDOM VARIATION IN VITAL RATES; AND A 52-PAGE CHAPTER, "DEMOGRAPHIC STOCHASTICITY," WHICH CONSIDERS RANDOM VARIATION ASSOCIATED WITH SAMPLING VARIATION IN INDIVIDUALS, ASSUMING THE VITAL RATES AS FIXED. THE BOOK DOES NOT ADDRESS THE ISSUES OF ESTIMATION OF TIME-SERIES MODELS OR FORECASTING FROM SUCH MODELS. IT CONTAINS A TWO-PAGE DISCUSSION OF "SHORT-TERM STOCHASTIC FORECASTS," MENTIONING THAT ARMA MODELS HAVE BEEN USED TO MODEL SURVIVAL AND FERTILITY TRENDS IN THE U.S. POPULATION.

STATISTICAL DEMOGRAPHY AND FORECASTING, BY JUHA M. ALHO AND BRUCE D. SPENCER (SPRINGER, 2005). THIS BOOK CONTAINS A 28-PAGE CHAPTER, "APPROACHES TO FORECASTING DEMOGRAPHIC RATES," WHICH DESCRIBES THE APPLICATION OF SOME STANDARD TIME-SERIES MODELS TO FORECASTING OF DEMOGRAPHIC RATES, INCLUDING ARIMA MODELS, REGRESSION CURVE-FITTING MODELS, HARVEY-TYPE STRUCTURAL MODELS, AND ARCH / GARCH MODELS. THE TREATMENT IS VERY GENERAL. FOR EXAMPLE, IT DOES NOT ADDRESS THE ISSUE THAT MORTALITY RATES IN ADJACENT AGE CATEGORIES OF A LIFE TABLE ARE CLOSE TO EACH OTHER.

DEMOGRAPHIC FORECASTING BY FEDERICO GIROSI AND GARY KING (PRINCETON UNIVERSITY PRESS, 2008). THIS BOOK PRESENTS A METHODOLOGY FOR ESTIMATING AGE-SPECIFIC MORTALITY, WHICH ADDRESSES THE IMPORTANT ISSUES THAT MORTALITIES IN ADJACENT AGE CATEGORIES ARE SIMILAR, AND TIME-SERIES DATA ARE OFTEN LIMITED OR NON-EXISTENT. THE METHODOLOGY IS BASED ON BAYESIAN ESTIMATION AND PRINCIPAL COMPONENTS ANALYSIS. THIS PRESENTATION IS A SUBSTANTIAL IMPROVEMENT OVER PREVIOUS METHODS.

LIMITATIONS ON THE USE OF STOCHASTIC-PROCESS THEORY IN DEMOGRAPHY

NONE OF THE PRECEDING TEXTS ADDRESSES THE ISSUE OF FORECASTING POPULATION, ONLY FORECASTING POPULATION PARAMETERS. THERE ARE SEVERAL REASONS FOR THIS APPROACH. THESE INCLUDE:

1. THE CONSTRUCTION OF TIME-SERIES MODELS REQUIRES A SUBSTANTIAL AMOUNT OF TIME-SERIES DATA. SUCH DATA ARE AVAILABLE FOR MANY INDUSTRIALIZED COUNTRIES, BUT NOT FOR MANY OTHER COUNTRIES.
2. MANY DEMOGRAPHIC PARAMETERS ARE INTERRELATED, REQUIRING THE USE OF MULTIVARIATE MODELS. MULTIVARIATE TIME SERIES MODELS (SUCH AS THE VAR MODELS ROUTINELY USED IN FINANCE) TYPICALLY REQUIRE EVEN MORE DATA THAN UNIVARIATE ONES. FOR A LARGE NUMBER OF DEPENDENT VARIABLES (SUCH AS AGE-SPECIFIC MORTALITIES), THE NUMBER OF MODEL PARAMETERS BECOMES VERY LARGE. FOR MANY DEMOGRAPHIC APPLICATIONS, THERE IS INSUFFICIENT DATA TO SUPPORT THE DEVELOPMENT OF STANDARD TIME-SERIES-ANALYSIS MODELS.
3. MOST STATISTICAL TIME-SERIES MODELS ARE DESIGNED TO PRODUCE SHORT-TERM FORECASTS, NOT INTERMEDIATE OR LONG-TERM FORECASTS. FOR LONG-TERM FORECASTS, IT IS MORE USEFUL TO LOOK AT THE PROPERTIES OF STABLE POPULATIONS, OR THE ASSUMED LONG-TERM PROPERTIES OF A DEMOGRAPHIC-TRANSITION MODEL.
4.

THE EFFECT OF TAKING STOCHASTIC VARIATION INTO ACCOUNT

THE REMAINDER OF THIS SECTION WILL SUMMARIZE SOME RESULTS FROM CASWELL'S MATRIX POPULATION MODELS. THE FOLLOWING SECTION WILL DESCRIBE SOME ASPECTS OF THE GIROSI-KING APPROACH.

CASWELL PRESENTS MUCH MATERIAL ON SENSITIVITY ANALYSIS, AND NOTE THAT IF THE MAGNITUDE OF RANDOM VARIATION IS LOW, THE RESULTS FOR SENSITIVITY ANALYSIS FOR DETERMINISTIC MODELS AND STOCHASTIC MODELS ARE SIMILAR.

THE ESSENTIAL DIFFERENCE BETWEEN DETERMINISTIC ("EXPECTED-VALUE ANALYSIS") MODELS AND STOCHASTIC MODELS IS INDICATED BY JENSEN'S INEQUALITY. LET US DENOTE THE GROWTH RATE OF MEAN POPULATION SIZE BY $\ln(E(R))$, AND THE TIME-AVERAGE GROWTH RATE FOR A SINGLE REALIZATION OF THE PROCESS AS $E(\ln(R))$. THEN BY JENSEN'S INEQUALITY,

$$E(\ln(R)) \leq \ln(E(R))).$$

THAT IS, THE RESULTS FROM THE EXPECTED-VALUE ANALYSIS ARE GREATER THAN OR EQUAL TO THE MEAN OF THE STOCHASTIC MODEL.

FROM A TAYLOR-SERIES EXPANSION, IT CAN BE SHOWN THAT

$$E(\ln(R))) \approx \ln(E(R))) - V(R)/(2E(R)^2),$$

WHERE $V(R)$ DENOTES THE VARIANCE OF R. THIS EXPRESSION SHOWS THAT AS THE VARIANCE INCREASES, THE DIFFERENCE BETWEEN THE TWO RESULTS INCREASES.

(FOR A RANDOM VARIABLE HAVING A LOGNORMAL DISTRIBUTION, THE EXACT RESULT IS

$$E(X) = \exp(E(\ln(X) + V(\ln(X))/2).)$$

IN MANY APPLICATIONS, THE VARIANCE OF YEAR-TO-YEAR VALUES OF DEMOGRAPHIC PARAMETERS, SUCH AS MORTALITY, IS SUBSTANTIAL. THIS FACT HAS TWO IMPLICATIONS:

- IN SOME APPLICATIONS, THE DIFFERENCE BETWEEN RESULTS FROM DETERMINISTIC MODELS AND STOCHASTIC MODELS CAN BE SUBSTANTIAL
- THE WIDTH OF CONFIDENCE INTERVALS CAN BE VERY LARGE

FOR STABLE POPULATIONS, IT IS ASSUMED THAT THE VALUES OF DEMOGRAPHIC PARAMETERS REMAIN CONSTANT. FOR NON-STABLE POPULATIONS, THE BIRTH RATES AND DEATH RATES MUST CONVERGE QUICKLY, OR THE PROCESS GOES TO EXTINCTION OR EXPLODES. LONG-RANGE POPULATION FORECASTS DEPEND

ESSENTIALLY ON ASSUMPTIONS ABOUT STABILITY AND COMPLETION OF THE DEMOGRAPHIC TRANSITION, NOT ON WHETHER A DETERMINISTIC OR STOCHASTIC POPULATION MODEL IS USED AS A BASIS FOR THEM.

21. FORECASTING OF DEMOGRAPHIC PARAMETERS

THE TERM "DEMOGRAPHIC FORECASTING" MAY REFER TO FORECASTING OF ANY SORT OF DEMOGRAPHIC QUANTITY, SUCH AS POPULATION, OR A DEMOGRAPHIC PARAMETER SUCH AS A MORTALITY RATE.

IN THIS SECTION, THE TERM "DEMOGRAPHIC ESTIMATION" REFERS TO ESTIMATION OF THE POPULATION COMPONENTS OF BIRTH RATES, DEATH RATES AND MIGRATION RATES. MIGRATION RATES ARE SET BY GOVERNMENT POLICY, AND ARE USUALLY NOT ESTIMATED USING MATHEMATICAL MODELS (ALTHOUGH SOME RESEARCH HAS INVESTIGATED THE RELATIONSHIP OF MIGRATION RATES TO ECONOMIC VARIABLES).

THE PRESENTATION HERE FOLLOWS THE TEXT, *DEMOGRAPHIC FORECASTING*, BY GIROSI AND KING. WE SHALL REFER TO THIS TEXT AS *GK*.

THE GIROSI-KING METHOD

GK ADDRESS THE ESTIMATION OF MORTALITY RATES, BUT THE METHODOLOGY MAY BE APPLIED TO ESTIMATION OF FERTILITY RATES AS WELL.

GK POINT OUT SHORTCOMINGS OF PREVIOUS APPROACHES TO DEMOGRAPHIC ESTIMATION. THESE INCLUDE:

- FOR MANY COUNTRIES AND REGIONS, THERE IS VERY LITTLE DATA ON WHICH TO BASE SAMPLE-BASED DIRECT ESTIMATES
- SOME APPROACHES FORECAST MORTALITY FOR EACH AGE INDEPENDENTLY. MORTALITY RATES FOR NEARBY AGE CATEGORIES SHOULD BE SIMILAR, BUT THE APPROACH OF FORECASTING THE AGE CATEGORIES INDEPENDENTLY RESULTS IN FORECASTS THAT DO NOT

PRESERVE THIS FEATURE. GK SHOW THAT THE LEE-CARTER APPROACH POSSESSES THIS SHORTCOMING.

GK POINT OUT THAT THE ESTABLISHED APPROACH OF BASING ESTIMATES ON PRINCIPAL COMPONENTS WORKS WELL, AND THEY INCLUDE THIS FEATURE IN THEIR APPROACH. (THE PRINCIPAL COMPONENTS METHOD WORKS WELL FOR TWO MAIN REASONS: (1) IT RESULTS IN MODELS THAT ARE “PARSIMONIOUS” WITH RESPECT TO NUMBER OF PARAMETERS; AND (2) THE PRINCIPAL COMPONENTS ARE ORTHOGONAL, SO THAT IN SPECIFYING FUTURE VALUES OF EXPLANATORY VARIABLES, THE PRINCIPAL COMPONENTS MAY BE SPECIFIED INDEPENDENTLY.)

THE *GK* APPROACH RECOGNIZES AND ADDRESSES THE REQUIREMENT THAT MORTALITIES IN NEXT-ADJACENT AGE CATEGORIES ARE SIMILAR. ONE WAY OF ADDRESSING THIS REQUIREMENT WOULD BE TO REPRESENT THE AGE-SPECIFIC MORTALITIES AS A MULTIVARIATE RANDOM VARIABLE. THAT APPROACH TENDS TO INVOLVE A LOT OF PARAMETERS (SUCH AS CORRELATIONS AMONG THE VARIABLES), AND FOR THE LARGE NUMBER OF AGE CATEGORIES, WOULD BE CUMBERSOME.

THE *GK* APPROACH DOES NOT REPRESENT THE VECTOR OF MORTALITIES BY AGE AS A MULTIVARIATE RANDOM VARIABLE. IT REPRESENTS EACH MORTALITY BY AGE AS A UNIVARIATE RANDOM VARIABLE, BUT IT INTRODUCES A “SMOOTHNESS” CONSTRAINT THAT TENDS TO MAINTAIN THE SMOOTHNESS. THIS CONSTRAINT IS TO MINIMIZA THE SUM OF SQUARES OF THE DIFFERENCES OF MORTALITIES IN NEXT-ADJACENT AGE CATEGORIES.

THE *GK* METHODOLOGY IS BASED ON BAYESIAN ESTIMATION. THAT APPROACH IS PARTICULARY USEFUL IN THIS APPLICATION, SINCE IT ADDRESSES THE FACT THAT MANY COUNTRIES HAVE VERY LIMITED AMOUNTS OF DATA, SUCH AS A LIFE TABLE FOR JUST ONE OR TWO YEAR, OR EVEN NO YEARS. (THE “SMOOTHNESS CONSTRAINT” IS INCLUDED IN ONE OF THE PRIOR DISTRIBUTIONS SPECIFIED IN THE MODEL.)

ALTHOUGH SIMPLE IN CONCEPT, THE *GK* METHODOLOGY IS SOMEWHAT COMPLICATED TO DESCRIBE IN DETAIL, AND THE DETAILS WILL NOT BE PRESENTED IN THIS PRESENTATION.

THE BASICS OF THE *GK* METHODOLGY WILL BE PRESENTED BY MEANS OF A DEMONSTRATION OF THE YourCast SOFTWARE.

FOR ADDITIONAL MATERIAL ABOUT FORECASTING, SEE THE ARTICLE, *A SURVEY OF METHODS FOR FORECASTING, POLICY ANALYSIS AND PLANNING, WITH EXAMPLES IN R* BY JOSEPH GEORGE CALDWELL (8 JULY 2019).

FORECASTING BASED ON THE ASSUMPTION OF A DEMOGRAPHIC TRANSITION

THE FORECASTING MODELS DISCUSSED IN THE PRECEDING SECTION (ARIMA MODELS) ARE APPROPRIATE FOR SHORT-TERM FORECASTING, NOT FOR MEDIUM- OR LONG-TERM FORECASTING. THE ARIMA MODEL IS ESSENTIALLY A HIGH-PASS FILTER, THAT REMOVES INFORMATION ABOUT LEVEL AND LONG-TERM FEATURES OF THE TIME SERIES. FOR MEDIUM- AND LONG-TERM FORECASTS, IT IS NECESSARY TO UTILIZE A MODEL THAT INCLUDES INFORMATION ABOUT THE DISTRIBUTION (LIKELIHOOD FUNCTION) OF DEMOGRAPHIC PARAMETERS (MORTALITY, FERTILITY, IMMIGRATION, EMIGRATION; OR GROWTH RATES) IN THE MEDIUM- AND LONG-TERM.

THE LONG-TERM NATURE OF POPULATION GROWTH MAY BE SPECIFIED BY A MODEL SUCH AS THE “DEMOGRAPHIC TRANSITION” MODEL, WHICH SPECIFIES VALUES OF THE DEMOGRAPHIC RATES FROM THE PRESENT TIME UNTIL THE TIME IN THE FUTURE WHEN THEY BECOME EQUAL. FOR A POPULATION TO CONTINUE IN THE LONG TERM, BIRTH AND DEATH RATES MUST EVOLVE TO BEING ESSENTIALLY THE SAME, FOR AS LONG AS THEY ARE SUBSTANTIALLY DIFFERENT, POPULATION GROWTH OR DECLINE OCCURS AT AN EXPONENTIAL RATE, AND THE POPULATION QUICKLY GOES EXTINCT OR EXPLODES.

TO CONSTRUCT A LONG-TERM POPULATION PROJECTION OR FORECAST, SPECIFY THE BIRTH, DEATH, AND MIGRATION RATES CORRESPONDING TO AN ASSUMED DEMOGRAPHIC TRANSITION MODEL, AND GENERATE THE PROJECTION OR FORECAST USING THE PROJECTION METHODOLOGY ALREADY DISCUSSED.

FORECASTING ON THE BASIS OF A STABLE POPULATION

A STABLE POPULATION IS AN ESTIMATE OF THE LONG-TERM STATE OF A POPULATION IF BIRTH, DEATH AND MIGRATION RATES REMAIN CONSTANT. THE STABLE POPULATION HENCE REPRESENTS AN ESTIMATE (PROJECTION OR FORECAST) OF THE LONG-TERM POPULATION SIZE AND COMPOSITION, CONDITIONAL ON THE ASSUMED RATES.

22. COLLECTION OF DEMOGRAPHIC DATA

AS DISCUSSED EARLIER, MUCH DEMOGRAPHIC DATA IS AVAILABLE FROM OFFICIAL SOURCES, SUCH AS NATIONAL STATISTICS OFFICES AND THE UNITED NATIONS. THESE DATA ARE COMPILDED FROM VITAL REGISTRATION RECORDS, CENSUSES AND DEMOGRAPHIC SURVEYS.

IN SOME APPLICATIONS, DATA ARE COLLECTED FROM SPECIAL-PURPOSE SAMPLE SURVEYS. SIGNIFICANT RESOURCES ARE AVAILABLE TO ASSIST THE COLLECTION, STORAGE, AND PROCESSING SUCH DATA. PRINCIPAL AMONG THESE RESOURCES ARE THE CDC Epi Info AND U.S. CENSUS CSPro SYSTEMS. THESE SYSTEMS WERE DESCRIBED BRIEFLY EARLIER IN THE PRESENTATION.

THE USER GUIDE, SOFTWARE, AND OTHER SUPPORTING MATERIAL FOR Epi Info ARE AVAILABLE FROM INTERNET WEBSITE https://www.cdc.gov/epiinfo/support/downloads.html .

DOCUMENTATION AND SOFTWARE DOWNLOADS FOR THE CSPro SYSTEM ARE AVAILABLE FROM INTERNET WEBSITE https://www.census.gov/data/software/cspro.html .

THIS SECTION ADDRESSES DESCRIPTION AND USE OF ONE OF THE Epi Info OR CSPro SYSTEMS.

23. SPECIAL TOPICS

THE PRECEDING SECTIONS HAVE DISCUSSED THE MAJOR ASPECTS OF DEMOGRAPHIC ANALYSIS. BECAUSE OF THE SURVEY NATURE OF THIS PRESENTATION, THERE ARE A NUMBER OF ASPECTS THAT HAVE NOT BEEN

ADDRESSED, OR NOT COVERED IN MUCH DETAIL. THIS SECTION OF THE PRESENTATION INCLUDES DISCUSSION OF SPECIAL TOPICS OF INTEREST TO A PARTICULAR AUDIENCE, OR ADDITIONAL DETAIL, OR ADDITIONAL DEMONSTRATION OF COMPUTER SOFTWARE.

24. DISCUSSION OF REFERENCES AND OTHER RESOURCES

THIS SECTION INCLUDES DISCUSSION OF SOME OF THE REFERNCES LISTED BELOW, AND ADDITIONAL DISCUSSION OF SOURCES FOR DATA AND COMPUTER SOFTWARE.

25. COMPLETION OF COURSE EVALUATION FORM

IT IS DESIRED TO OBTAIN FEEDBACK ON OPINIONS ABOUT THE COURSE, FROM ATTENDEES. PLEASE TAKE TIME TO COMPLETE THE COURSE EVALUATION FORM, AND HAND IT IN. THANK YOU!

REFERENCES

Texts on Demographic Analysis

Primary Reference Texts for Course

Swanson, David A. and Jacob S. Siegel, *The Methods and Materials of Demography*, 2nd ed., Emerald Group Publishing, 2004. (This text is available online at https://demographybook.weebly.com/uploads/2/7/2/5/27251849/david_a._swanson_jacob_s._siegel_the_methods_and_materials_of_demography_second_edition__2004.pdf , and will be used as the main text for the first part of the course.)

Shryock, Henry S., Jacob S. Siegel, *The Methods and Materials of Demography*, condensed first edition by Edward G. Stockwell, Academic Press, 1976 (also issued by US Government Printing Office, 1971, 4th printing June 1980); comprehensive coverage of classical demography, available at low price.

Girosi, Federico and Gary King, *Demographic Forecasting*, Princeton University Press, 2008 (This text is available online at https://gking.harvard.edu/files/gking/files/prelims_1.pdf , and will be used as a primary text for the second part of the course.)

Additional Texts on Demographic Analysis

Preston, Samuel, Patrick Heuveline, Michel Guillot, *Demography: Measuring and Modeling Population Processes*, Wiley-Blackwell, 2000

Wachter, Kenneth W., *Essential Demographic Methods*, Harvard University Press, 2014

Lundquist, Jennifer Hickes, Douglas L. Anderton and David Yaukey, *Demography: The Study of Human Population*, 4th ed., Waveland Press, 2014

Weeks, John R, *Population: An Introduction to Concepts and Issues*, 12th ed., Wadsworth Publishing, 2015

Swanson, David A., *The Frontiers of Applied Demography*, Springer, 2016

Anderson, Barbara A., *World Population Dynamics: An Introduction to Demography*, Pearson, 2014

Goldstone, Jack A., Eric P. Kaufmann, Monica Duffy Toft, *Political Demography: How Population Changes Are Reshaping International Security and National Politics*, Oxford University Press, 2011

Rowland, Donald T., *Demographic Methods and Concepts*, Oxford University Press, 2003

Poston Jr, Dudley L. and Leon F. Bouvier, *Population and Society: An Introduction to Demography*, 2nd ed., Cambridge University Press 2017

Yaukey, David, *Demography, The Study of Human Population*, St. Martin's Press, 1985

Harper, Sarah, *Demography: A Very Short Introduction*, Oxford University Press, 2018

Weinstein, Jay and Vijayan K. Pillai, *Demography: The Science of Population*, 2nd ed., Rowman and Littlefield Publishers, 2015

Dorling, Danny and Stuart Gietel-Basten*, Why Demography Matters*, Polity, 2017

Hoque, M. Nazrul, Beverly Pecotte and Mary A. McGehee, *Applied Demography and Public Health in the 21st Century*, Springer, 2016

Palmore, James A. and Robert W. Gardner, *Measuring Mortality, Fertility and Natural Increase: A Self-Teaching Guide to Elementary Measures*, East-West Center, 1994

Pollard, A. H., Farhat Yusuf and G. N. Pollard, *Demographic Techniques*, Pergamon Press, 1990

Burch, Thomas K., *Model-Based Demography: Essays on Integrating Data, Technique and Theory*, Springer, 2017

National Research Council 2000. *Beyond Six Billion: Forecasting the World's Population*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9828 .

Additional Texts on Mathematical Demography

Keyfitz, Nathan and Hal Caswell, *Applied Mathematical Demography (Statistics for Biology and Health)* 3rd ed., Springer, 2005

Caswell, Hal, *Matrix Population Models: Construction, Analysis and Interpretation* 2nd ed., Oxford University Press (Sinauer Associates imprint), 2018.

Keyfitz, Nathan and John A. Beekman, *Demography Through Problems*, Springer, 1984

Alho, Juha and Bruce Spencer, *Statistical Demography and Forecasting*, Springer, 2005

Caswell, Hal, *Sensitivity Analysis: Matrix Methods in Demography and Ecology*, Springer, 2019

Biswas, Suddhendu and G. L. Sriwastav, *Stochastic Processes in Demography Applications*, New Central Book Agency (India), revision of 2006 (original edition published in 1988).

Demography-Related Books (Books on Population Issues)

Livi-Bacci, M,. *A Concise History of World Population*, 6th ed., Wiley-Blackwell, 2017

Nelson, John Carl, *Historical Atlas of the Eight Billion: World Population History 3000 BCE to 2020*, CreateSpace Independent Publishing Platform, 2014

McEvedy, Colin and Richard M. Jones, *Atlas of World Population History*, 1978

Hardin, Garrett, *Living Within Limits: Ecology, Economics and Population Taboos*, Oxford University Press, 1993

Cohen, Joel E., *How Many People Can the Earth Support?,* W. W. Norton, 1995

Pimentel, David and Marcia Pimentel, eds., *Food, Energy, and Society*, revised edition, University of Colorado Press, 1996

Diamond, Jared, *Collapse: How Societies Choose to Fail or Succeed*, Viking, 2005

Ehrlich, Paul R. and Anne H. Ehrlich, *The Population Explosion*, Touchstone, 1990

Ehrlich, Paul R., *The Population Bomb*, Sierra Club – Ballantine Books, 1968

Leakey, Richard and Roger Lewin, *The Sixth Extinction: Patterns of Life and the Future of Humankind*, Anchor Books, 1995

Brown, Lester R., *Plan B 4.0: Mobilizing to Save Civilization*, W. W. Norton, 2009

Brown, Lester R., et al., *State of the World*, W. W. Norton, 1995 (and other years)

Brown, Lester R., et al., *Vital Signs: The Environmental Trends that are Shaping Our Future*, W. W. Norton, 1998 (and other years)

Brown, Lester R. and Hal Kane, *Full House: Reassessing the Earth's Population Carrying Capacity*, W. W. Norton, 1994

Simon, Julian, *Population Matters: People, Resources, Environment and Immigration*, Transaction Publishers, 1990

Foot, David K. with Daniel Stoffman, *Boom, Bust & Echo: How to Profit from the Coming Demographic Shift*, Macfarlane, Walter and Ross, Toronto, Canada, 1996

ARIMA-Type Statistical Time Series Analysis

Box, G. E. P., and Gwilym Jenkins, *Time Series Analysis, Forecasting Control*, first edition Holden-Day, 1970, latest edition is 5th edition by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, Wiley, 2016.

Tsay, Ruey S., *Multivariate Time Series Analysis with R and Financial Applications*, Wiley, 2014.

Lütkepohl, Helmut, *New Introduction to Multiple Time Series Analysis*, Springer, 2006

Hamilton, James D., *Time Series Analysis*, Princeton University Press, 1994

Cryer, Jonathan D. and Kung-Sik Chan, *Time Series Analysis with Applications in R*, Springer, 2008 (univariate)

Structural-Type Time Series Analysis, Emphasizing Kalman Filtering

Harvey, Andrew C., *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, 1989

Durbin, J. and S. J. Koopman, *Time Series Analysis by State Space Methods*, 2nd edition, Oxford University Press, 2012

Small-Area Estimation

Rao, J. N. K., Small Area Estimation, Wiley, 2003

Computer Software References

Manual X Indirect Techniques for Demographic Estimation, United Nations Department of International Economic and Social Affairs, Population Studies No. 81, 1983

Moultrie, Tom, Rob Dorrington, Allan Hill, Kenneth Hill, Ian Timaeus and Basia Zaba, eds., *Tools for Demographic Estimation*, Paris: International Union for the Scientific Study of Population, 2013, posted at Internet website http://demographicestimation.iussp.org/ . (This collection of software is an updating of the *Manual X* content.)

Zlotnik, Hania, *Computer Programs for Demographic Estimation: A User's Guide*, Committee on Population and Demography Report No. 11, National Academy Press, 1981

Press, William H., Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., Cambridge University Press, 2007

R

Adler, Joseph, *R in a Nutshell* 2nd ed., O'Reilly, 2012

Crawley, Michael J., *The R Book*, Wiley, 2007

Sawitzki, Günther, *Computational Statistics, An Introduction to R*, CRC Press, 2009

Cornillon, Pierre-Andre, Arnaud Guyader, François Hussen, Nicolas Jégou, Julie Josse, Maela Kloareg, Eric Matzner-Løber and Laurent Rouvière, *R for Statistics*, CRC Press, 2012

Demographic Data

World Population Prospects, United Nations Department of Economic and Social Affairs, 2019

SYLLABUS

Syllabus for course, *Demographic Analysis*, by Joseph George Caldwell

Part 1. Basic Demography (“traditional,” “classical,” rates, proportions and algebra, simple matrix arithmetic (no general matrix algebra, calculus or inferential statistics)

1. Administrative items; course outline; course evaluation form; course text; course software; lecture notes; course-related resources (texts, articles, software, data)
2. Definitions and scope of demography and demographic analysis; basic (“classical,” “traditional”) demography vs. advanced (“modern”) demography; uses of demographic data and demographic analysis
3. Demographic data: definition, uses, sources (UN, US Census, USAID/DHS, World Bank).
4. The mathematics of basic demography (algebra, rates, proportions, growth; vectors and matrices, vector and matrix arithmetic).
5. Computer software for basic demography (“Office” software, geographic information system (GIS), DAPPS, Spectrum, MortPak, R)
6. Population static characteristics (size, geographic distribution, and composition (distribution by factors other than or additional to geography; population pyramid (distribution by age and sex))
7. Population-based estimates (standardized rates, ratio estimates, age-sex-specific proportions or rates, synthetic estimates, small-area estimates)
8. Population dynamics (population grouped by age, period and cohort; Lexis diagram; mortality; life tables; natality / fertility; survival rates; growth rates; migration; period-based descriptors; cohort-based descriptors)
9. Population projections and forecasts (growth models, the balancing equation, cohort-component method, transition matrices, multi-state models; multiple decrement life tables); computer software for making population projections; sources of demographic data (country vital statistics, UN, WHO, FAO, USAID DHS); model life tables
10. Population-based estimates (mathematical basis for population-based estimates; computer software for making population-based forecasts (*Spectrum)*; sources of substantive-field / application-area data (economics,

education, health, environment, urban planning, agriculture, politics, insurance)

11. Geographic information systems (GIS); GRASS and QGIS; GIS in R; sources of GIS data
12. Summary of Part 2.

Part 2. Advanced Demography ("modern" demography, "statistical" demography)

1. The mathematics of advanced demography (vector and matrix algebra (rank, determinants, vector spaces, eigenvectors and eigenvalues, principal components)); the statistics of demography (distributions, estimation (survival functions), statistical models, forecasting (autoregressive integrated moving average (ARIMA) models, VAR models, structural models, Kalman filtering, Bayesian estimation); small-area estimation
2. More on population projection (transition matrices, Leslie matrices, multistate transition matrices, Lefkovitch matrices (stage vs. age); Kalman filtering)
3. More on population-based forecasts (model-based estimates, small-area estimation)
4. Estimation of demographic parameters; direct methods (vital registration records, census); indirect nonparametric methods (Brass estimation); indirect parametric methods (failure density function, force of mortality; survival function, Kaplan-Meier product-limit estimate; Cox proportional-hazard-function model). Computer software for indirect estimation (*Tools for Demographic Estimation*), United Nations' *MortPak*, Gary King's *YourCast*, Rob Hyndman's *Demography in R,* CRAN R libraries).
5. Stochastic processes in demography applications
6. Forecasting of demographic parameters (without and with covariates, principal components, singular-value decomposition (Lee-Carter), ARIMA models, structural models, Bayesian estimation); demographic transition; stable population
7. Collection of demographic data (registration data, census, demographic surveys; data entry software (CSPro, Epi Info); *Software Tools for Demographic Estimation*)
8. Special topics (optional)
9. Discussion of references and other resources
10. Completion of course evaluation form

COURSE EVALUATION FORM

Demographic Analysis
Training Evaluation and Learning Self-Assessment

Training delivered to (Client Name)
(Date(s))

by Joseph George Caldwell, PhD (Statistics)

Dear Participant:

We appreciate your attendance and are interested in your comments in order to assess the value of the course and improve it. Please answer the following questions, adding additional comments as necessary, and return the form in the attached envelope. Thank you.

Date(s) attended________________ Location of course_________________________________

Summary Comments

Course Content and Level of Detail

1. Overall, how useful do you consider the information to your work?
   Not very useful__ Somewhat useful__ Very Useful__
2. Overall, was the material presented in sufficient detail? Yes__ No__
3. Overall, was the material presented in too much detail? Yes__ No__
4. Overall, was the material presented too advanced? Yes__ No__
5. Overall, was the material presented too elementary? Yes__ No__
6. Were there some topics on which you would have preferred more discussion?
   Yes__ No__
   If so, which ones?____
7. Was the course well-matched to your mathematical / statistical background? Yes__ No__
6. Was the course well-matched to your computer background? Yes__ No__
7. Would you prefer a course with less mathematical content? Yes__ No__
8. Would you prefer a course with more emphasis on use of computer software (examples, practice)?
   Yes__ No__
9. Do you have ready access to statistical program packages? Yes__ No__
   If so, please list: R __, Stata__ SPSS__ SAS__ Excel__ Other (specify)______________ If so, do you have access to user's manuals? Yes__ No__
   To supplementary reference texts? Yes__ No__

Course Delivery

1. Was the presenter effective (knowledgeable, clear, engaging, articulate, logical, interesting, responsive)? Yes__ No__
2. Were the presentations (explanations, examples) effective (relevant, clear, well-organized)?
   Yes__ No__

3. Were the visual aids (notes, projections, flip-charts, black/whiteboards) helpful?
 Yes__ No__
4. Were the course notes sufficiently detailed? Yes__ No__
5. Were your questions addressed satisfactorily? Yes__ No__
6. Was the course too long (too many hours or days)? Yes__ No__
7. Were the sessions too long per day? Yes__ No__
8. Were the breaks sufficient in number and length for you to continue essential job functions?
 Yes__ No__
9. Could you see the speaker and the presented material well? Yes__ No__
10. Could you hear and understand the speaker well? Yes__ No__

Facilities / Environment

1. Was the course location convenient? Yes__ No__
2. Was the environment conducive to learning (comfortable, quiet, free from distractions)?
 Yes__ No__
3. Was the seating arrangement satisfactory (comfortable chairs, nearness to presenter, table space, lighting)? Yes__ No__
4. Was the air-conditioning satisfactory? Yes__ No__
5. Was the equipment (projector, boards, microphone, etc.) satisfactory? Yes__ No__
6. Were the snacks and meals satisfactory? Yes__ No__
7. Was parking adequate? Yes__ No__

Learning Self-Assessment (Overall Rating of Training)

Please rate your knowledge and skills, before and after the training, in the areas listed in the table.

Rating Scale: 1 = Low 3 = Medium 5 = High

Before Training					**Self-Assessment of Knowledge and Skills Related to:**	**After Training**				
1	2	3	4	5	Assessment of appropriate demographic analysis technique to use	1	2	3	4	5
1	2	3	4	5	Ability to access reference materials (course text and software sources)	1	2	3	4	5
1	2	3	4	5	Construction of population projection (using *Spectrum*)	1	2	3	4	5
1	2	3	4	5	Construction of population-based forecast (using *Spectrum*)	1	2	3	4	5
1	2	3	4	5	Construction of population-based forecast (using *Spectrum DemProj* for the population forecast, but not for the population-based forecast)	1	2	3	4	4
1	2	3	4	5	Forecasting of demographic components (mortality, fertility) using *YourCast*	1	2	3	4	5
1	2	3	4	5	Accessing sources of demographic data (mortality, fertility, migration, life tables)	1	2	3	4	5
1	2	3	4	5	Accessing sources of application data (e.g., economic, educational, health, environmental, agricultural)	1	2	3	4	5
1	2	3	4	5	Ability to document analysis	1	2	3	4	5

1	2	3	4	5	Ability to describe and defend assumptions, estimates and forecasts	1	2	3	4	5
1	2	3	4	5	Ability to assess reliability of estimates (standard errors, confidence intervals)	1	2	3	4	5
1	2	3	4	5	Ability to assess validity of estimates (bias)	1	2	3	4	5
1	2	3	4	5	Ability to conduct sensitivity analysis	1	2	3	4	5
1	2	3	4	5	Ability to explain demographic analysis to others	1	2	3	4	5

Assessment of the Trainer

Please compare the trainer's *overall teaching ability* to that of others you have had before, and rate on the following scale. Factors to consider are:

1. Ability to describe concepts clearly.
2. Understanding of course material.
3. Ability to explain concepts and procedures in an interesting manner.
4. Is well organized.
5. Is responsive to questions (encourages them and responds well (quickly and clealy))
6. Encourages participation in class; encourages / elicits questions.
7. Presents good examples.
8. Uses class time efficiently.
9. States learning objectives clearly at the beginning of each session.
10. Adjusts pace of presentation to student ability to absorb material.

Rating (circle your choice)	**Description of relative ability**
1	Best of all instructors I have had in other courses
2	Top 25%
3	Middle 50%
4	Worst 25%
5	Worst of all instructors I have had in other courses

Supplementary Evaluation

Was the training relevant to your present or anticipated future job requirements? Yes__ No__

What particular skill or knowledge did you acquire as a result of this course?__________________

Please assess the extent to which the training will make a difference in your ability to do your job:

No difference__   Some difference__   Big difference__

What will you do differently in your work as a result of this training?___________________________

What were particular strengths of the training?____

What were particular weaknesses of the training?____

Was an appropriate amount of material covered during the training course? Yes__ No__

Was too much material covered, or too little? _________

Was the course length too long or too short? _________

Would you prefer training spread over more days, with shorter sessions (e.g., mornings only) Yes__ No__ If yes, please specify ____

What are the three most important things you learned in the training?_________________________

How can the training be improved?____
Are there specific changes would you recommend to improve the course (make it more relevant to your job, or make it more effective)?____

What additional training do you require to do your job better?_______________________________

What other topics or courses would you like to take?_____________________________________

What is your mathematical background?____

What is your background in statistical theory?____

What is your background in using database or statistical program package software (e.g., R, Stata, SPSS, SAS, Excel, Access, CSPro, EpiInfo)?____

Did the course meet your expectations? Yes__ No__

Would you recommend this course to others in your profession? Yes__ No__

Please rate the following statements on a scale from 1(disagree strongly) to 5 (agree strongly).

1	2	3	4	5	The level of difficulty was about right
1	2	3	4	5	My technical background was not a good match to the subject matter of the course
1	2	3	4	5	My technical background was not a good match to the technical level of the course
1	2	3	4	5	The content of the course is relevant to my job
1	2	3	4	5	I can apply the knowledge and skills that I gained in this course in my work
1	2	3	4	5	To be more useful, this course requires a background in basic statistical theory
1	2	3	4	5	To be more useful, this course requires a background in Stata
1	2	3	4	5	The trainer actively involved me in the learning process
1	2	3	4	5	As a result of the training, I feel more confident about my ability to conduct small-area estimation
1	2	3	4	5	As a result of the training, I feel more confident about my general statistical skills

Please provide an overall assessment of the course, using the following table:

Element	Poor	Fair	Good	Excellent	Comment
Quality of instruction					
Relevance of material					
Organization of course					
Participation/Discussion					
Interest of Material					
Facility Conditions					
Equipment Conditions					
Overall Evaluation					

Additional Comments:____

Name (optional)____

Organization (optional)____

FndID(208)
FndTitle(DEMOGRAPHIC ANALYSIS: LECTURE NOTES)
FndDescription(DEMOGRAPHIC ANALYSIS: LECTURE NOTES)
FndKeywords(demographic analysis; population projection; population-based forecasts; short course)