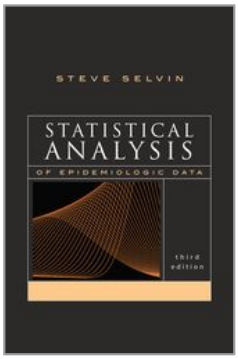

Statistical Analysis of Epidemiologic Data
Steve Selvin

Print publication date: 2004
Print ISBN-13: 9780195172805
Published to Oxford Scholarship Online: September 2009
DOI: 10.1093/acprof:oso/9780195172805.001.0001

Life Table Analysis: An Introduction

Steve Selvin

DOI:10.1093/acprof:oso/9780195172805.003.11

Abstract and Keywords

A life table is a systematic description of mortality experience. A cohort life table is constructed from the mortality of individuals followed from birth of the first to the death of the last member of a group. This chapter discusses the construction of a life table, life table survival function, life table hazard function, and competing risks.

Keywords: life tables, mortality, survival function, hazard function, competing risks

A life table is a systematic description of mortality experience. A *cohort life table* is constructed from the mortality experience of individuals followed from the birth of the first to the death of the last member of a group. Such life tables are, for example, constructed from animal and insect data. For human populations, it is obviously not practical to construct a life

table by following a cohort of individuals from birth until all have died. Instead, a life table is constructed from current mortality rates. These rates do not apply to past populations and undoubtedly will not apply to future populations. Nevertheless, mortality patterns can be clearly identified in a *current life table*, and the comparisons among current life tables calculated for different groups is a basic strategy for describing and summarizing mortality experience.

The first formal life tables were developed independently by Edmund Halley (1693) and John Graunt (1662). By the beginning of the twentieth century, life tables were routinely computed as part of an emerging awareness of the importance of mortality statistics. The first official United States life table published in 1900 showed the mean length of life for white males was 46.6 years and for white females was 48.7 years.

(p.336) COMPLETE, CURRENT LIFE TABLE: CONSTRUCTION

The word “complete” when applied to a life table means that ages are recorded in one-year intervals. The word “current” means that the life table is constructed from current mortality rates from specific population data. The actual construction of a complete, current life table is mechanical and consists of seven basic elements:

1. *Age interval* (x to $x + 1$): Each age interval consists of one year (age denoted by x) except the last age interval, which is left open ended (for example, 90^+ years).
2. *Number alive* (l_x): The symbol l_x represents the number of individuals alive at exactly age x. The number l_x is the size of the life table population-at-risk at the beginning of the interval x. The number alive at age 0 (l_0) is set at some arbitrary value, such as 100,000, and occasionally called the *radix*.
3. *Deaths* (d_x): The symbol d_x represents the number of individuals who died between the ages of x and $x + 1$.
4. *Probability of death* (q_x): The symbol q_x represents the conditional probability that an individual who is

alive at age x dies before age $x + 1$. That is, P(death before age $x + 1$ | alive at age x) = q_x and $q_x = d_x/l_x$. The probability of death within a specific age interval is related to a hazard rate. The complementary probability is $p_x = 1 - q_x$ and is the conditional probability that an individual who is alive at age x survives to age $x + 1$.

5. *Years lived* (L_x): The symbol L_x represents total cumulative time lived by the entire cohort between the ages of x to $x + 1$. Each individual alive at age x contributes to the total time lived—either one year if an individual lives the entire year or the proportion of the year lived if the individual dies in the interval. The value of L_x is the life table total per-son-years-at-risk for the interval x to $x + 1$

6. *Total time lived* (T_x): The symbol T_x represents total time lived beyond age x by all individuals alive at age x. The total time lived beyond age x is $T_x = L_x + L_{x+1} + L_{x+2} + \cdots$. The value T_x is primarily a computational step in the life table construction.

7. *Expectation of life* (e_x): The symbol e_x represents the mean number of additional years lived by those individuals alive at age x, computationally $e_x = T_x/l_x$.

The following relationships are direct consequences of these seven definitions:

a. Number dying in the interval x to $x + 1 = d_x = q_x l_x = l_x - l_{x+1}$

b. Number surviving at age $x + 1 = l_{x+1} = (1 - q_x)l_x = p_x l_x = l_x - d_x$

(p.337)

c. Probability of dying in the interval x to $x + 1 = q_x = (l_x - l_{x+1})/l_x = d_x/l_x$

d. Probability of surviving from x to $x + 1 = p_x = 1 - q_x = (l_x - d_x)/l_x$

These definitions apply to a complete life table (age intervals of one year).

The total person-years-at-risk for the interval x to $x + 1$ includes one year of survival for each person who did not die during the interval. Individuals who died contribute the proportion of the year they were alive to the total time lived. The mean time contributed by those who died in the interval x to $x + 1$ (denoted) is close to 0.5 for all ages except the first few years of life. For years 0 to 3, the values of are , , and (determined empirically [1]). The distribution of survival times of those who died in the first years of life is skewed toward the beginning of the age interval because most deaths occur among the youngest children. This property is particularly true of deaths in the age interval 0 to 1 year where the vast majority of deaths occur within the first month of life. Therefore, the mean contribution of time lived by those who died to the total years lived is particularly low for the first age interval. For ages 2 and 3 years, the mean is slightly less than 0.5 year. For all other one year age intervals, the mean value of is essentially 0.5 year. A mean of occurs when individuals die randomly throughout the one-year age interval, producing an mean contribution of 0.5 year.

The value takes on particular importance in calculating the person-years of life for a life table because , which is the life table person-years-at-risk for most age intervals x to $x + 1$ ($x >$ 3). Using L_x, the life table age-specific mortality rate becomes $r_x = d_x/L_x$ and provides a link to observed age-specific current mortality rates. The life table person-years calculation does not differ in principle from the approximate person-years calculation discussed in Chapter 1.

The starting point for construction of a life table is a set of age-specific probabilities of death (denoted q_x). These conditional probabilities are derived by equating the life table age-specific mortality rates (denoted r_x) to the current age-specific mortality rates calculated from the population of interest (denoted R_x). The current age-specific mortality rate R_x is the number of deaths from age x to age $x + 1$, divided by the person-years-at-risk for a specific population during a specific calendar year. In symbols,

$$\text{life table mortality rate} = r_x = \frac{d_x}{L_x} = R_x = \text{ observed mortality rate or}$$

$$r_x = \frac{d_x}{(l_x - d_x) + \bar{a}_x d_x} = \frac{q_x}{1 - (1 - \bar{a}_x)q_x} = R_x$$

(p.338) and solving for q_x gives

$$q_x = \frac{R_x}{1 + (1 - \bar{a}_x)R_x}.$$

A set of observed current age-specific mortality rates (R_x) produce a set of life table probabilities (q_x). The probabilities q_x generate the rest of the life table functions (l_x, d_x, L_x, T_x, and e_x) with one exception.

The person-years of life (L_x) for the last interval cannot be calculated directly because a value for is not generally available. The individuals who are alive at the start of the last interval all die ($q_{x'} = 1.0$) so that $l_{x'} = d_{x'}$, where x' denotes the final age in the life table (for example, if the last interval is 90^+, then $x' = 90$) but is certainly greater than 0.5 year. Again equating the life table mortality rate to the observed current mortality rate for this last age interval yields a value of $L_{x'}$ because

$$\text{life table mortality rate} = r_{x'} = \frac{d_{x'}}{L_{x'}} = \frac{l_{x'}}{L_{x'}} = R_{x'}.$$

Solving for $L_{x'}$ yields

$$L_{x'} = \frac{l_{x'}}{R_{x'}},$$

where, to repeat, $R_{x'}$ comes from the observed current mortality data and $l_{x'}$ from the life table. For example, if and (for male, California 1980 mortality rates; Table 11–1), then person-years of total additional life lived by those who reached age 90. Therefore, an observed set of current age-specific mortality rates is all that is needed to calculate a complete, current life table. Incidentally, the average additional years lived by those individuals (males) who reached the age of 90 is years.

Specifically, consider the age interval 65 to 66 for white males from the 1980 California data:

$$q_{65} = \frac{R_{65}}{1 + (1 - \bar{a}_x)R_{65}} = \frac{0.0284}{1 + 0.5(0.0284)} = 0.0280$$

$$\text{where } R_{65} = \frac{2{,}097}{73{,}832} = 0.0284$$

(p.339) $d_{65} = l_{65}\, q_{65} = 69{,}728(0.0280) = 1{,}953$

$L_{65} = (l_{65} - d_{65}) + 0.5d_{65} = (69{,}728 - 1{,}953) + 0.5(1.953) = 68{,}752$

$$T_{65} = L_{65} + L_{66} + \cdots + L_{90+}$$
$$= 68{,}752 + 66{,}757 + \cdots + 9{,}126 + 41{,}617 = 1{,}011{,}356$$

$$e_{65} = \frac{T_{65}}{l_{65}} = \frac{1{,}011{,}356}{69{,}728} = 14.504.$$

Thus, all individuals who reach the age of 65 can expect to live an additional 14.5 years.

These five steps applied to each age interval, starting at age 0, sequentially produce the entire current life table from current mortality rates (R_x) and an arbitrary starting value (l_0). Two complete, current life tables are given in Tables 11–1 and 11–2 for male and female residents of California based on mortality rates from the year 1980.

The expected number of years of life remaining after age x is an effective and popular summary of the entire mortality pattern described by a life table (e_x; last column in each of Tables 11–1 and 11–2). The expectation of life is no more than a special mean value and is calculated in the same way as most mean values, where

$$\text{mean years of life remaining} = e_x = \frac{\text{total years lived beyond age } x}{\text{number of individuals age } x} = \frac{T_x}{l_x}.$$

The expected years of life from birth ($e_0 = T_0/l_0$) are $e_0 = 69.61$ years for males and $e_0 = 76.93$ years for females, based on 1980 California mortality rates.
Expectations of life at birth are compared among countries and among groups within a country as a reflection of overall the mortality experience. The United States mean years of remaining life time e_0 has steadily increased over the last century, and the difference between males and females has also remarkably increased (Table 11–3). The mean years of remaining life time has a geometric interpretation related to a survival function. The expectation of life (e_0) is approximately equal to the area under the life table survival curve. In Chapter 12, this property is discussed in detail.

Another single summary value that indicates the overall mortality pattern found in a life table is the median age at death [2]. The median age at death is that age where half the life table population is alive and half have died. It is simply that age where exactly $0.5l_0$ individuals are alive (and dead). It is a straightforward calculation to find this median age in a life table. The l_x-column is (p.340)

Table 11–1. California 1980 population of white males

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
0–1	129,602	2166	1671.3	0.01647	1,647	100,000	98,518	6,960,692	69.61
1–2	117,753	123	104.5	0.00104	103	98,353	98,295	6,862,175	69.77
2–3	115,003	73	63.5	0.00063	62	98,251	98,217	6,763,880	68.84
3–4	113,314	60	53.0	0.00053	52	98,188	98,161	6,665,663	67.89
4–5	110,822	41	37.0	0.00037	36	98,136	98,118	6,567,502	66.92
5–6	110,548	55	49.8	0.00050	49	98,100	98,076	6,469,384	65.95
6–7	106,857	42	39.3	0.00039	39	98,051	98,032	6,371,308	64.98
7–8	112,184	58	51.7	0.00052	51	98,013	97,988	6,273,276	64.00
8–9	116,423	44	37.8	0.00038	37	97,962	97,944	6,175,288	63.04
9–10	132,952	52	39.1	0.00039	38	97,925	97,906	6,077,344	62.06
10–11	134,266	48	35.7	0.00036	35	97,887	97,869	5,979,438	61.09
11–12	128,938	60	46.5	0.00047	46	97,852	97,829	5,881,569	60.11
12–13	125,502	52	41.4	0.00041	41	97,806	97,786	5,783,740	59.13
13–14	128,212	82	64.0	0.00064	63	97,766	97,735	5,685,954	58.16
14–15	132,775	129	97.2	0.00097	95	97,703	97,656	5,588,219	57.20
15–16	143,600	233	162.3	0.00162	158	97,608	97,529	5,490,563	56.25

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
16–17	151,840	290	191.0	0.00191	186	97,450	97,357	5,393,034	55.34
17–18	157,365	400	254.2	0.00254	247	97,264	97,141	5,295,677	54.45
18–19	159,476	415	260.2	0.00260	252	97,017	96,891	5,198,535	53.58
19–20	171,235	416	242.9	0.00243	235	96,765	96,648	5,101,644	52.72
20–21	173,682	418	240.7	0.00240	232	96,530	96,414	5,004,996	51.85
21–22	172,656	436	252.5	0.00252	243	96,298	96,177	4,908,582	50.97
22–23	176,544	400	226.6	0.00226	217	96,056	95,947	4,812,405	50.10
23–24	175,732	410	233.3	0.00233	223	95,838	95,726	4,716,458	49.21
24–25	174,780	409	234.0	0.00234	223	95,615	95,503	4,620,731	48.33
25–26	173,214	393	226.9	0.00227	216	95,391	95,283	4,525,228	47.44
26–27	169,980	400	235.3	0.00235	224	95,175	95,063	4,429,944	46.55
27–28	168,369	366	217.4	0.00217	206	94,951	94,848	4,334,881	45.65
28–29	157,189	330	209.9	0.00210	199	94,745	94,646	4,240,033	44.75
29–30	162,394	346	213.1	0.00213	201	94,547	94,446	4,145,387	43.84
30–31	161,191	329	204.1	0.00204	192	94,345	94,249	4,050,941	42.94
31–32	154,874	355	229.2	0.00229	216	94,153	94,045	3,956,692	42.02
32–33	162,136	338	208.5	0.00208	196	93,937	93,840	3,862,647	41.12

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
33–34	163,065	305	187.0	0.00187	175	93,742	93,654	3,768,807	40.20
34–35	127,624	267	209.2	0.00209	196	93,567	93,469	3,675,153	39.28
35–36	128,890	296	229.7	0.00229	214	93,371	93,264	3,581,684	38.36
36–37	127,933	302	236.1	0.00236	220	93,157	93,047	3,488,420	37.45
37–38	127,923	334	261.1	0.00261	242	92,937	92,816	3,395,373	36.53
38–39	109,718	281	256.1	0.00256	237	92,695	92,576	3,302,557	35.63
39–40	108,168	325	300.5	0.00300	277	92,458	92,319	3,209,981	34.72
40–41	104,314	338	324.0	0.00324	298	92,180	92,031	3,117,662	33.82
41–42	100,059	342	341.8	0.00341	314	91,882	91,725	3,025,630	32.93
42–43	97,330	344	353.4	0.00353	323	91,569	91,407	2,933,905	32.04
43–44	92,394	356	385.3	0.00385	351	91,246	91,070	2,842,497	31.15
44–45	91,741	431	469.8	0.00469	426	90,895	90,682	2,751,427	30.27
45–46	92,331	438	474.4	0.00473	428	90,469	90,255	2,660,745	29.41
46–47	88,150	522	592.2	0.00590	532	90,041	89,775	2,570,491	28.55
47–48	90,475	559	617.9	0.00616	551	89,509	89,233	2,480,716	27.71
48–49	90,095	650	721.5	0.00719	639	88,958	88,638	2,391,483	26.88
49–50	97,275	696	715.5	0.00713	630	88,318	88,003	2,302,845	26.07

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
50–51	98,008	734	748.9	0.00746	654	87,688	87,361	2,214,841	25.26
51–52	93,134	825	885.8	0.00882	768	87,034	86,650	2,127,480	24.44
52–53	94,496	875	926.0	0.00922	795	86,267	85,869	2,040,830	23.66
53–54	93,239	1,010	1,083.2	0.01077	921	85,472	85,011	1,954,960	22.87
54–55	96,443	1,126	1,167.5	0.01161	981	84,551	84,060	1,869,949	22.12
55–56	97,763	1,197	1,224.4	0.01217	1,017	83,569	83,061	1,785,889	21.37
56–57	96,823	1,272	1,313.7	0.01305	1,077	82,552	82,014	1,702,829	20.63
57–58	96,189	1,334	1,386.9	0.01377	1,122	81,475	80,914	1,620,815	19.89
58–59	98,518	1,553	1,576.4	0.01564	1,257	80,353	79,724	1,539,901	19.16
59–60	96,154	1,564	1,626.6	0.01613	1,276	79,096	78,458	1,460,177	18.46
60–61	88,552	1,472	1,662.3	0.01649	1,283	77,820	77,179	1,381,719	17.76
61–62	83,814	1,684	2009.2	0.01989	1,522	76,537	75,776	1,304,541	17.04
62–63	81,464	1,763	2,164.1	0.02141	1,606	75,014	74,211	1,228,766	16.38
63–64	76,317	1,871	2,451.6	0.02422	1,778	73,408	72,519	1,154,554	15.73
64–65	75,505	2,032	2,691.2	0.02656	1,902	71,630	70,679	1,082,035	15.11
65–66	73,832	2,097	2,840.2	0.02801	1,953	69,728	68,752	1,011,356	14.50
66–67	69,480	2,121	3,052.7	0.03007	2,038	67,776	66,757	942,604	13.91

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
67–68	65,690	2,130	3,242.5	0.03191	2,098	65,738	64,689	875,847	13.32
68–69	62,557	2,256	3,606.3	0.03542	2,254	63,640	62,513	811,159	12.75
69–70	57,412	2,327	4,053.2	0.03973	2,439	61,386	60,166	748,646	12.20
70–71	53,926	2,205	4,088.9	0.04007	2,362	58,947	57,766	688,479	11.68
71–72	50,402	2,376	4,714.1	0.04606	2,606	56,585	55,282	630,713	11.15
72–73	47,213	2,342	4,960.5	0.04840	2,613	53,979	52,673	575,431	10.66
73–74	42,931	2,233	5,201.4	0.05070	2,604	51,366	50,064	522,759	10.18
74–75	39,611	2,300	5,806.5	0.05643	2,751	48,762	47,386	472,694	9.69
75–76	36,306	2,408	6,632.5	0.06420	2,954	46,011	44,534	425,308	9.24
76–77	33,386	2,251	6,742.3	0.06523	2,808	43,057	41,653	380,774	8.84
77–78	30,141	2,102	6,973.9	0.06739	2,712	40,249	38,892	339,121	8.43
78–79	26,432	2,272	8,595.6	0.08241	3,094	37,536	35,990	300,229	8.00
79–80	26,264	2,093	7,969.1	0.07664	2,640	34,443	33,123	264,239	7.67
80–81	21,846	1,958	8,962.7	0.08578	2,728	31,803	30,439	231,117	7.27
81–82	18,868	1,947	10,319.1	0.09813	2,853	29,075	27,648	200,677	6.90
82–83	16,653	1,802	10,820.9	0.10265	2,692	26,222	24,876	173,029	6.60
83–84	14,825	1,751	11,811.1	0.11153	2,624	23,530	22,218	148,153	6.30

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
84–85	13,137	1,689	12,856.8	0.12080	2,525	20,906	19,643	125,935	6.02
85–86	11,350	1,622	14,290.7	0.13338	2,452	18,380	17,155	106,292	5.78
86–87	9,442	1,426	15,102.7	0.14042	2,237	15,929	14,811	89,137	5.60
87–88	8,047	1,198	14,887.5	0.13856	1,897	13,692	12,744	74,327	5.43
88–89	6,091	1,072	17,599.7	0.16176	1,908	11,795	10,841	61,583	5.22
89–90	5,382	897	16,666.7	0.15385	1,521	9,887	9,126	50,742	5.13
90^+	17,346	3,487	20,102.6	1.00000	8,366	8,366	41,617	41,617	4.97

(*) = rate per 100,000 person-years-at-risk.

(p.341) (p.342) (p.343)

Table 11-2. California 1980 population of white females

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
0–1	123,342	1,635	1,325.6	0.01310	1,310	100,000	98,821	7,693,461	76.93
1–2	111,520	64	57.4	0.00057	57	98,690	98,658	7,594,641	76.95
2–3	109,200	41	37.5	0.00038	37	98,633	98,613	7,495,983	76.00
3–4	108,749	22	20.2	0.00020	20	98,596	98,586	7,397,370	75.03
4–5	105,698	41	38.8	0.00039	38	98,576	98,557	7,298,784	74.04
5–6	105,801	37	35.0	0.00035	34	98,538	98,521	7,200,227	73.07
6–7	101,630	37	36.4	0.00036	36	98,504	98,486	7,101,706	72.10
7–8	106,850	32	29.9	0.00030	29	98,468	98,453	7,003,220	71.12
8–9	110,410	32	29.0	0.00029	29	98,438	98,424	6,904,767	70.14
9–10	127,237	33	25.9	0.00026	26	98,410	98,397	6,806,342	69.16
10–11	128,916	33	25.6	0.00026	25	98,384	98,372	6,707,945	68.18
11–12	124,123	32	25.8	0.00026	25	98,359	98,347	6,609,573	67.20
12–13	119,672	28	23.4	0.00023	23	98,334	98,322	6,511,227	66.22
13–14	123,652	48	38.8	0.00039	38	98,311	98,292	6,412,905	65.23
14–15	127,869	68	53.2	0.00053	52	98,273	98,247	6,314,613	64.26
15–16	139,122	98	70.4	0.00070	69	98,220	98,186	6,216,366	63.29

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
16–17	146,318	93	63.6	0.00064	62	98,151	98,120	6,118,180	62.33
17–18	150,163	132	87.9	0.00088	86	98,089	98,046	6,020,059	61.37
18–19	152,382	121	79.4	0.00079	78	98,003	97,964	5,922,014	60.43
19–20	162,203	138	85.1	0.00085	83	97,925	97,883	5,824,050	59.47
20–21	162,313	118	72.7	0.00073	71	97,842	97,806	5,726,167	58.52
21–22	162,709	104	63.9	0.00064	62	97,771	97,739	5,628,360	57.57
22–23	167,087	96	57.5	0.00057	56	97,708	97,680	5,530,621	56.60
23–24	168,874	121	71.7	0.00072	70	97,652	97,617	5,432,940	55.64
24–25	168,959	119	70.4	0.00070	69	97,582	97,548	5,335,324	54.68
25–26	168,414	110	65.3	0.00065	64	97,513	97,481	5,237,776	53.71
26–27	165,167	141	854	0.00085	83	97,450	97,408	5,140,295	52.75
27–28	164,403	123	74.8	0.00075	73	97,366	97,330	5,042,887	51.79
28–29	154,062	137	88.9	0.00089	86	97,294	97,250	4,945,557	50.83
29–30	158,102	135	85.4	0.00085	83	97,207	97,166	4,848,307	49.88
30–31	157,975	134	84.8	0.00085	82	97,124	97,083	4,751,141	48.92
31–32	153,534	134	87.3	0.00087	85	97,042	97,000	4,654,058	47.96
32–33	160,016	157	98.1	0.00098	95	96,957	96,910	4,557,058	47.00

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
33–34	160,299	127	79.2	0.00079	77	96,862	96,824	4,460,149	46.05
34–35	125,826	144	114.4	0.00114	111	96,785	96,730	4,363,324	45.08
35–36	126,747	158	124.7	0.00125	120	96,675	96,614	4,266,594	44.13
36–37	125,960	155	123.1	0.00123	119	96,554	96,495	4,169,980	43.19
37–38	127,942	161	125.8	0.00126	121	96,436	96,375	4,073,485	42.24
38–39	109,358	169	154.5	0.00154	149	96,314	96,240	3,977,110	41.29
39–40	106,481	196	184.1	0.00184	177	96,166	96,077	3,880,870	40.36
40–41	103,828	171	164.7	0.00165	158	95,989	95,910	3,784,793	39.43
41–42	99,325	205	206.4	0.00206	198	95,831	95,732	3,688,883	38.49
42–43	96,380	228	236.6	0.00236	226	95,633	95,520	3,593,151	37.57
43–44	93,276	256	274.5	0.00274	261	95,407	95,276	3,497,631	36.66
44–45	92,873	258	277.8	0.00277	264	95,146	95,014	3,402,355	35.76
45–46	92,183	246	266.9	0.00267	253	94,882	94,755	3,307,341	34.86
46–47	88,595	274	309.3	0.00309	292	94,629	94,483	3,212,586	33.95
47–48	91,046	323	354.8	0.00354	334	94,337	94,170	3,118,103	33.05
48–49	89,588	384	428.6	0.00428	402	94,003	93,802	3,023,934	32.17
49–50	97,274	398	409.2	0.00408	382	93,601	93,409	2,930,132	31.30

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
50–51	98,371	449	456.4	0.00455	425	93,218	93,006	2,836,722	30.43
51–52	95,717	474	495.2	0.00494	458	92,794	92,565	2,743,716	29.57
52–53	99,570	557	559.4	0.00558	515	92,335	92,078	2,651,152	28.71
53–54	101,653	687	675.8	0.00674	618	91,820	91,511	2,559,074	27.87
54–55	105,815	675	637.9	0.00636	580	91,202	90,912	2,467,563	27.06
55–56	108,657	737	678.3	0.00676	613	90,622	90,316	2,376,651	26.23
56–57	106,689	784	734.8	0.00732	659	90,009	89,680	2,286,336	25.40
57–58	106,142	842	793.3	0.00790	706	89,350	88,997	2,196,656	24.58
58–59	107,384	929	865.1	0.00861	764	88,644	88,263	2,107,659	23.78
59–60	103,981	1007	968.4	0.00964	847	87,881	87,457	2,019,396	22.98
60–61	97,063	964	993.2	0.00988	860	87,034	86,604	1,931,939	22.20
61–62	93,115	1,033	1,109.4	0.01103	951	86,174	85,698	1,845,335	21.41
62–63	90,046	1,070	1,188.3	0.01181	1,007	85,223	84,720	1,759,637	20.65
63–64	86,916	1,141	1,312.8	0.01304	1,098	84,216	83,667	1,674,917	19.89
64–65	85,726	1,282	1,495.5	0.01484	1,234	83,118	82,501	1,591,250	19.14
65–66	86,996	1,387	1,594.3	0.01582	1,295	81,884	81,237	1,508,749	18.43
66–67	83,258	1,400	1,681.5	0.01668	1,344	80,589	79,917	1,427,513	17.71

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
67–68	79,961	1,428	1,785.9	0.01770	1,403	79,245	78,544	1,347,595	17.01
68–69	78,039	1,485	1,902.9	0.01885	1,467	77,842	77,109	1,269,052	16.30
69–70	74,389	1,617	2,173.7	0.02150	1,642	76,375	75,554	1,191,943	15.61
70–71	70,163	1,614	2,300.4	0.02274	1,700	74,733	73,883	1,116,389	14.94
71–72	67,599	1,816	2,686.4	0.02651	1,936	73,033	72,065	1,042,506	14.27
72–73	65,045	1,813	2,787.3	0.02749	1,954	71,097	70.120	970,441	13.65
73–74	60,676	1,905	3,139.6	0.03091	2,137	69,143	68,074	900,320	13.02
74–75	57,975	1,889	3,258.3	0.03206	2,148	67,006	65,931	832,246	12.42
75–76	54,912	1,995	3,633.1	0.03568	2,314	64,857	63,700	766,315	11.82
76–77	51,217	2,089	4,078.7	0.03997	2,500	62,543	61,293	702,615	11.23
77–78	48,251	1,993	4,130.5	0.04047	2,430	60,043	58,828	641,322	10.68
78–79	43,234	2,344	5,421.7	0.05279	3,041	57,613	56,093	582,494	10.11
79–80	47,158	2,399	5,087.2	0.04961	2,707	54,572	53,218	526,401	9.65
80–81	39,462	2,318	5,874.0	0.05706	2,960	51,865	50,385	473,183	9.12
81–82	36,295	2,416	6,656.6	0.06442	3,151	48,905	47,330	422,798	8.65
82–83	31,875	2,360	7,403.9	0.07140	3,267	45,755	44,121	375,468	8.21
83–84	30,470	2,535	8,319.7	0.07987	3,394	42,488	40,791	331,347	7.80

x to (x + 1)	Population	Deaths	R_x *	q_x	d_x	l_x	l_x	T_x	e_x
84–85	27,904	2,540	9,102.6	0.08706	3,404	39,094	37,392	290,556	7.43
85–86	24,712	2,458	9,946.6	0.09475	3,382	35,690	34,000	253,163	7.09
86–87	21,302	2,383	11,186.7	0.10594	3,423	32,309	30,597	219,164	6.78
87–88	19,402	2,120	10,926.7	0.10361	2,993	28,886	27,389	188,567	6.53
88–89	14,905	1,993	13,371.4	0.12533	3,245	25,893	24,270	161,177	6.22
89–90	13,873	1,900	13,695.7	0.12818	2,903	22,648	21,196	136,907	6.05
90^+	47,650	8,131	17,064.0	1.00000	19,745	19,745	115,710	115,710	5.86

(*) = rate per 100,000 person-years-at-risk.

(p.344) (p.345) (p.346) (p.347) (p.348)

Table 11–3. Expectation of life for white males and females in the United States (1900–99)

Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	1999
Male	46.6	48.6	54.5	59.7	62.1	66.5	67.4	68.0	70.7	71.8	73.9
Female	48.7	52.0	55.6	63.5	66.6	72.2	74.1	75.6	78.1	78.8	79.4
Difference	2.1	3.4	1.1	3.8	4.5	5.7	6.7	7.6	7.4	7.0	5.5

Source: National Center for Health Statistics, U.S. Department of Health and Human Services

searched to find the age interval that contains $0.5l_0$ individuals (denoted to). Then, using linear interpolation (other more elegant interpolation schemes are certainly possible) gives an approximate median value of

$$\text{median} = \hat{x}_{0.5} = \tilde{x} + \frac{l_{\tilde{x}} - 0.5l_0}{l_{\tilde{x}} - l_{\tilde{x}+1}}.$$

From the California life table for male residents in 1980 (Table 11–1), the age interval containing $0.5l_0 = 50{,}000$ individuals is age 73 to 74 ($l_{73} = 51{,}366$ and $l_{74} = 48{,}762$), and

$$\hat{x}'_{0.5} = 73 + \frac{51{,}366 - 50{,}000}{51{,}366 - 48{,}762} = 73.525$$

is the median age at death [2]. For the females, the age interval containing the median is 80 to 81 (Table 11–3), giving

$$\hat{x}_{0.5} = 80 + \frac{51{,}865 - 50{,}000}{51{,}865 - 48{,}905} = 80.630.$$

That is, half the women experiencing the 1980 pattern of mortality risk will live beyond the age of 80.6 years. Another simple and effective estimate of the median age at death is the lower end point of the interval containing the value $0.5l_0$ individuals. From the California data, the values 73 years for males and 80 years for females also estimate the median age at death.

The crude mortality rate calculated from a life table is the total number of persons who died divided by the total number of person-years lived by the entire life table population, or

$$\text{crude mortality rate} = \frac{\text{total deaths}}{\text{total person-years}} = \frac{\Sigma d_x}{T_0} = \frac{l_0}{T_0}.$$

(p.349) The life table crude mortality rate is the reciprocal of the expected years of life at birth

$$\text{crude mortality rate} = \frac{l_0}{T_0} = \frac{1}{e_0}$$

or

$$e_0 = \frac{T_0}{l_0} = \frac{1}{l_0/T_0} = \frac{1}{\text{crude mortality rate}}.$$

Referring to the life table for males (Table 11–1), the crude mortality rate is 100,000/6,960,692 = 0.01437 or 1,437 deaths per 100,000 person-years and 1/0.01437 = 69.607 years of life are expected to be lived by a newborn male infant who experiences exactly the 1980 age-specific mortality rates. A life table formally illustrates the expected relationship that survival time (mean remaining lifetime) is inversely related to risk (rate of death). Three assumptions are implicit in constructing and interpreting a life table. The life table calculation assumes that the same number of births occur each year (l_0 constant). The deaths are also assumed to be randomly distributed within each interval for ages greater than three (thus, $a_i = 0.5$), and the population size does not change (the number of births equals the number of deaths each year and no immigration or emigration occurs). When a population conforms to this last property, it is called a *stationary population*. For example, a perfectly stationary population requires the risk of death for a 60-year-old in the year 2000 to be the same as the risk of death experienced by an individual born in 2000 when that person is 60 years old (year 2060). Although stationary human populations do not exist, in many cases population changes are sufficiently small so that postulating that the life table mortality pattern derived from an approximately stationary population is not misleading, making a life table a useful tool to describe and compare human mortality experience.

Aside: Consider three consecutive time intervals where a death can occur in any one these intervals with

probabilities q_1, q_2 and q_3. Therefore (Table 11–4), where the probability $p_i = P$(surviving t o interval $i + 1$ | survived up to the interval i) The

Table 11–4. Computation of the survival probability for three consecutive time intervals

Interval	P(death in interval i)	P(survive interval i)	P(survive beyond interval i)
1	q_1	$p_1 = 1 - q_1$	p_1
2	q_2	$p_2 = 1 - q_2$	$p_1 \times p2$
3	q_3	$p_3 = 1 - q_3$	$p_1 \times p_2 \times p3$

(p.350) probability of surviving beyond a given interval is the product of these conditional probabilities. For example, the probability of surviving beyond interval 3 is $p_1 \times p_2 \times p_3$. This product-rule applies to any number of intervals and is a fundamental part of the description of survival in a life table.

LIFE TABLE SURVIVAL FUNCTION

A fundamental summary statistic derived from a life table is a survival function (introduced in Chapter 1). As before, the symbol $S(x)$ represents the probability of surviving beyond age x. Two identical ways of computing a survival function $S(x)$ from a life table are

$$S(x) = \frac{l_x}{l_0}$$

$$S(x) = \prod_{i=0}^{x-1} (1 - q_i) = \prod_{i=0}^{x-1} p_i.$$

These two calculations are identical because

$$S(x) = \prod_{i=0}^{x-1} p_i = \prod_{i=0}^{x-1} \frac{l_{i+1}}{l_i} = \frac{l_1}{l_0}\frac{l_2}{l_1}\frac{l_3}{l_2}\frac{l_4}{l_3} \cdots \frac{l_{x-1}}{l_{x-2}}\frac{l_x}{l_{x-1}} = \frac{l_x}{l_0}$$

where $p_i = l_{i+1}/l_i$, is the conditional probability of surviving from age i to age $i + 1$ given that the individual is alive at the beginning of the age interval. Also, $S(0) = l_0/l_0 = 1$, which is a property of a survival function in general.

The life table survival functions for the male (solid line) and female (dotted line) for the 1980 California populations are displayed in Figure 11–1 (top). A small but sharp decrease in $S(x)$ caused by high rates of infant mortality in the first year of life is followed by a slight and gradual decrease in the probability of survival until about ages 60 or 70. After about age 60, the $S(x)$ curve begins to fall rapidly. The probability of living beyond 90 years of age, for example, is given by the values $S(90) = 0.084$ for males and $S(90) = 0.197$ for females (females are 2.4 times more likely than males to live beyond the age of 90). This pattern of survival is typically observed in modern human populations.

LIFE TABLE HAZARD FUNCTION

The slope of the survival function or the derivative of $S(x)$ at the point $x(dS(x)/dx)$ measures the impact of mortality on a population at exactly age x (Chapter 1). The slope indicates the

rate of change described by the curve representing the (p.351)

probability of surviving beyond a particular point (instantaneous intensity of mortality). To reflect risk, this slope is divided by the probability of surviving beyond age *x*. Analogous to the definition of a rate (Chapter 1), when the instantaneous slope of the survival function at age *x* is measured relative to the probability of surviving beyond age *x*, the previous definition of a rate emerges, given as

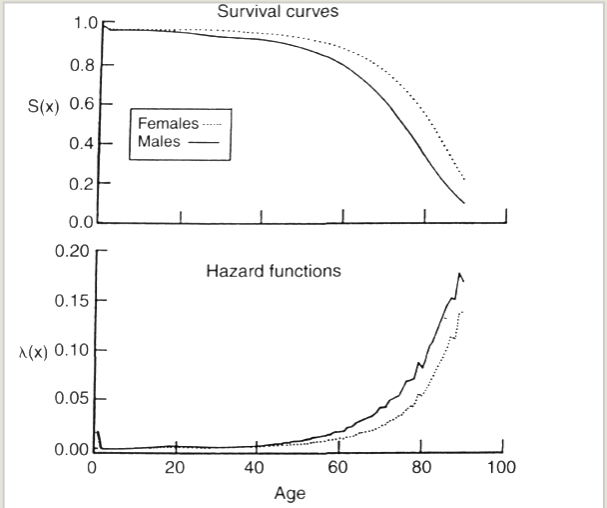

Figure 11–1. Survival curves and hazard functions from the life table for white males and females, California, 1980.

$$\lambda(x) = -\frac{dS(x)/dx}{S(x)},$$

where $\lambda(x)$ represents the hazard rate at age *x* and the negative sign makes it a positive quantity. Because a hazard rate is an instantaneous quantity, it must be approximated when the survival function $S(x)$ is estimated from a life table.

To calculate the hazard rate from a complete life table, it is necessary to make two approximations to estimate this theoretical quantity. The slope of the survival curve at the

midpoint of the interval x to $x + 1$ is approximately $S(x + 1) - S(x)$, and the value of the survival curve at the midpoint $x + 1/2$ is approximately $[S(x + 1) + S(x)]/2$. These two approximations are exact if the survival (p.352) curve is a straight line and a good approximation over a one-year interval. Combining these two quantities gives an approximate expression for the hazard rate at age $x + 1/2$ of

$$\lambda(x + 1/2) = -\frac{dS(x + 1/2)/dx}{S(x + 1/2)} \approx -\frac{S(x + 1) - S(x)}{[S(x + 1) + S(x)]/2}.$$

This expression in terms of the number of persons alive at age $x(l_x)$ is

$$\lambda(x + 1/2) \approx -\frac{l_{x+1} - l_x}{(l_{x+1} + l_x)/2} = -2\,\frac{p_x - 1}{p_x + 1} = \frac{2q_x}{p_x + 1}$$

because $S(x) = l_x/l_0$ and $p_x = l_x{+}1/l_x$.
Since $\log(p) \approx 2(p - 1)/(p + 1)$ for $p > 0.7$, then $\lambda(x + 1/2) \approx -\log(p_x)$ provides a useful approximation of the hazard rate from the life table–generated values of p_x. An approximate expression for the hazard rate at age x is the mean of the hazard rates at age $x - 1/2$ and $x + 1/2$, or

$$\lambda(x) \approx \frac{-[\log(p_{x-1/2}) + \log(p_{x+1/2})]}{2}.$$

A further simplification is achieved by using yet another approximation, because $\log(p) \approx p - 1$ for $p > 0.9$, then $\lambda(x + 1/2) \approx -\log(p_x) \approx q_x$, and, as before, an approximation for the hazard rate at age x is

$$\lambda(x) \approx \frac{q_{x-1/2} + q_{x+1/2}}{2}$$

for age intervals with low probabilities of death ($q_x < 0.1$). Thus, a hazard rate is not very different from the life table conditional probability of death when q_x is small (the usual case).

Another view of the hazard rate $\lambda(x + 1/2)$ comes from the fact that a hazard rate is an instantaneous age-specific rate. An average age-specific rate from a life table is

$$\text{life table mortality rate} = \frac{d_x}{l_x - 0.5d_x}.$$

For a small interval (say, one year), the age-specific life table mortality rate is approximately equal to the hazard rate at the middle of an age interval, or (p.353)

$$\lambda(x + 1/2) \approx \text{life table mortality rate} = \frac{d_x}{l_x - 0.5d_x}.$$

Two other versions of this expression are

$$\lambda(x + 1/2) \approx \frac{q_x}{1 - 0.5q_x} = \frac{2q_x}{p_x + 1}.$$

The last expression is the same as the previous expression for the hazard rate derived from different considerations. Again, if d_x is small relative to l_x ($p_x \approx 1$), then $\lambda(x + 1/2) \approx q_x$. In general, an approximate life table hazard rate is

$$\lambda(x + \delta_x) \approx \frac{d_x}{\delta_x(l_x - 0.5d_x)}$$

where δ_x represents the age interval width. The accuracy of this approximate hazard rate decreases as the age interval width δ_x increases.

The hazard functions (a continuous series of hazard rates) calculated from the California 1980 life tables for males and

females are plotted in Figure 11–1 (bottom). Details of the mortality pattern are clearly seen from these hazard functions. For example, an inconsistency in the rise of the hazard function for the older age groups is obvious and undoubtedly due to the lack of reliability in reporting of age among older individuals (about 80 years or so).

The shape of the hazard curve observed for the 1980 California life table populations is typical of most modern human populations over the entire age span. After the first year of life, the next 60 years are characterized by a slightly increasing hazard function followed by a sharp increase. However, hazard functions in other contexts take on other shapes. A population subject to only accidental (random) deaths unrelated to age, for example, could have a mortality pattern with a constant hazard function (a horizontal line). A hazard function and a survival function are inversely related because high rates of mortality imply low probabilities of survival. The exact mathematical relationship is described in Chapter 12, and complete discussions are found in more technical texts on survival analysis (e.g., [2]).

Life tables can be constructed from small sets of data where the probability of death q_x is estimated directly from the observed data. The principles are the same as those described, but the issue of sampling variation cannot be ignored. The life table functions (q_x, l_x, and the rest) are estimated quantities subject to sampling variation. Huge numbers of individuals make up the California data used to create the male and female life tables (Tables 11–1 and 11–2), so the precision of the estimates (p.354) is not much of an issue. For a life table based on a small number of individuals, however, the sampling variability of the estimated quantities should be taken into account. Expressions for the variances of life table estimates are usually based on assuming that the probabilities of death can be accurately described by binomial distributions (these expressions are discussed in detail elsewhere [1]).

Eleven individuals received a special treatment. The times to death (in months) are 1, 3, 8, 8, 18, 23, 29, 35, 38, 41, 48. A life table, based on 10-month intervals, summarizes these 11 observations (Table 11–5). The size of the sample used to

construct this life table is small (n = 11), making the variability of the estimates an issue (described in a following section) and, once again, categorizing a continuous variable (in this case, survival time) is not an ideal way to proceed. Small sets of survival data are better analyzed by other approaches (presented in Chapters 12 and 13).

An instructive application of a life table involves a calculation showing the consequences of lower hazard rates on a specific population. Suppose a hazard rate is reduced uniformly by a proportion c. In symbols, $\lambda(x) = c\lambda_0(x)$, where $\lambda_0(x)$ is either a known or an estimated hazard function. Construction of a life table based on such a reduced hazard function provides a rich description of the resulting mortality experience. Figure 11–2 (top) shows three hypothetical hazard functions based on the 1980 California, white male mortality rates ($\lambda_0(x)$, top curve) where c is set at 0.75, 0.50, and 0.25 (the next three curves). The logarithms of the hazard rates more clearly show the detail of these curves (Figure 11–2, bottom). The logarithms of a set of proportional hazard rates produce parallel lines. The associated life table constructed from the $c\lambda(x)$-values describes the impact of the lower hazard rates. That is, the rates $c\lambda(x)$ produce the probabilities q_x, which, in turn, produce the rest of the life table functions and allow a complete description of the three hypothetical populations in terms of survival probabilities and expected years of remaining life time.

To summarize the life tables constructed from the three reduced hazard functions, the proportion of individuals alive at ages 65, 75, and 85 years ($S(x) = l_x/l_0$ = survival probabilities) and the expected length of life at birth (e_0-values)

Table 11–5. Life table for small set of survival data

Interval	Midpoint	d_x	l_x	q_x	p_x	l_x	S(x)	SE*	λ(x + 5)
0–10	5	4	11	0.364	0.636	1000	S(10) = 0.636	0.145	0.044
10–20	15	1	7	0.143	0.857	636	S(20) = 0.545	0.150	0.015
20–30	25	2	6	0.333	0.667	545	S(30) = 0.364	0.145	0.040
30–40	35	2	4	0.500	0.500	364	S(40) = 0.182	0.116	0.067
40–50	45	2	2	1.000	0.000	182	S(50) = 0.182	0.116	0.200

(*) = standard error of $S(x)$ from Greenwood's expression (next section).

(p.355) for the "proportional populations" are estimated (Table 11–6). It is unlikely that a decrease in mortality would be exactly proportional throughout the entire life span (exactly proportional hazards rates), nevertheless, a simple description of the impact of decreasing mortality risk is created by using life table summary values. The percentage of older individuals increases strikingly as age-specific mortality decreases. For example, about 46% of the 1980 California males are

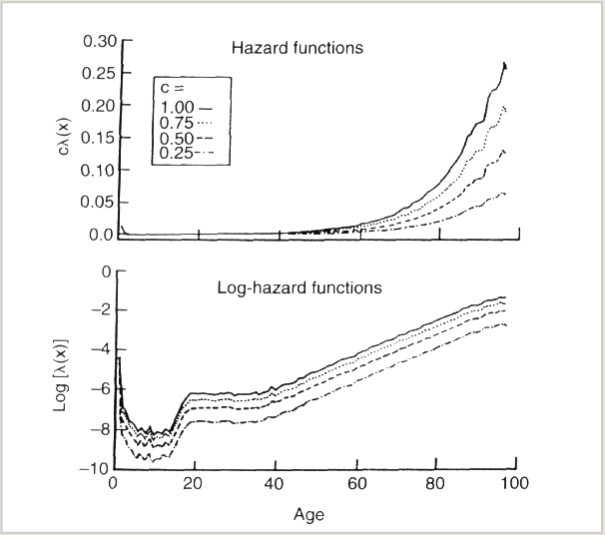

Figure 11–2. Hazard functions and the logarithm of the hazard functions for three hypothetical patterns of mortality based on the white male mortality rates, California, 1980.

Table 11–6. Impact on the 1980 California male population from three hypothetically reduced hazard rates

Hazard	% ≥ 65 years*	% ≥ 75 years*	% ≥ 85 years*	Expectation (e_0)
1.00*λ	69.7	46.0	18.3	69.6
0.75*λ	78.7	58.7	30.0	74.7
0.50*λ	85.3	70.2	45.1	79.8
0.25*λ	92.3	83.8	67.4	87.4

(*) % ≥ x years represents 100 × times $S(x)$.

(p.356) older than 75 years, but, when the hazard rate is reduced by a factor of 4 (c = 0.25), this value increases to an estimated

84%. The expected length of life at birth correspondingly increases about 18 years from 69.6 to 87.4 years.
The effects on a population from changes in a hazard rate are not always clear. As the illustration shows, the impact of a hazard rate is easily interpreted in terms of life table summary statistics. A decrease in a hazard rate becomes a less abstract expression of risk when translated, for example, into an increase in the number of individuals exceeding a specific age or into an increase in the expected years of remaining life.

LIFE TABLES: THREE APPLICATIONS OF LIFE TABLE TECHNIQUES

Calculating a Survival Probability from Follow-up Data

The evaluation of a treatment for a chronic disease usually involves the assessment of survival times or, perhaps, remission times. The probability of surviving five years after receiving a particular treatment is a frequently used measure of the efficacy of a drug or surgical procedure. The severity of a cancer is sometimes expressed in terms of a five-year survival probability. The probability of surviving a specific period of time is clearly a critical element in describing a wide range of kinds of risk.

Survival data are typically collected over a period of time and recorded from a series of cohorts (one cohort for each year of follow-up, for example). This follow-up pattern of data collection allows an efficient estimate of the five-year survival probability or, in general, an estimate of the survival function associated with the sampled population. Follow-up data [3] concerning the survival after diagnosis of kidney cancer patients illustrate the estimation of a survival probability (Table 11–7).

The complete display of the sampled data shows the cohorts formed each year of the study as new patients are diagnosed. The symbol x denotes the years survived after the kidney cancer is diagnosed. The column labeled l_x contains the count of the newly diagnosed patients alive at the beginning of the time interval x to $x + 1$. The number of deaths in each interval is represented by d_x. The possibility exists that patients are “lost to follow-up” during the time period covered by the

study. The count of patients lost during a specific time interval x is denoted u_x. The last column in the table contains the counts of the patients *censored or withdrawn* from study (denoted w_x). Individuals are said to be withdrawn when they are no longer relevant to further calculations. For example, consider the 1950 cohort of 19 patients. Five patients died the first year, one the second year, two were lost (one each year). The remaining 11 patients contribute (p.357) information about the second year of survival but no information about the third year or beyond because they were only observed for a maximum of two years (1950 and 1951). The 11 ($w_2 = 11$) remaining members from the 1950 cohort alive at the end of the second year are said to be withdrawn from study. They either survived or died after 1951, but this information is not part of the collected data. All that is known about the 11 patients is that they were alive at the closing date of the study.

The relevant counts for the four possible outcomes (l_x, d_x, u_x, and w_x) are displayed by single year from diagnosis of kidney cancer (Table 11–7) for each year after diagnosis for each cohort. The survival experiences of theses 126 kidney cancer patients from the six cohorts are efficiently tabulated into a single summary table (Table 11–8).

The number of individuals who are at risk at the beginning of the interval l_x+1 result from what occurred in the previous interval (d_x, u_x, and w_x), or $l_x+1 = l_x - d_x - u_x - w_x$. For example, after the second year of follow-up, 38 individuals are available to estimate the survival during the third year. These individuals remain from the 60 who started the second year because 5 died, 6 were lost, and 11 were withdrawn, giving $60 - 5 - 6 - 11 = 38$ study participants.

Table 11–7. Calculation of a survival probability: data

Year	x to x + 1	l_x	d_x	u_x	w_x
1946	0–1	9	4	1	—
	1–2	4	0	0	—
	2–3	4	0	0	—

Year	x to x + 1	l_x	d_x	u_x	w_x
	3–4	4	0	0	—
	4–5	4	0	0	—
	5–6	4	0	0	4
1947	0–1	18	7	0	—
	1–2	11	0	0	—
	2–3	11	1	0	—
	3–4	10	2	2	—
	4–5	6	0	0	6
1948	0–1	21	11	0	—
	1–2	10	1	2	—
	2–3	7	0	0	—
	3–4	7	0	0	7
1949	0–1	34	12	0	—
	1–2	22	3	3	—
	2–3	16	1	0	15
1950	0–1	19	5	1	—
	1–2	13	1	1	11
1951	0–1	25	8	2	15

(p.358)

Table 11-8. Calculation of a survival probability from cohort data: summary data

x to (x + 1)	l_x	d_x	u_x	w_x
0–1	126	47	4	15
1–2	60	5	6	11
2–3	38	2	0	15
3–4	21	2	2	7
4–5	10	0	0	6

x to (x + 1)	l_x	d_x	u_x	w_x
5–6	4	0	0	4

If all study participants entered the study on the first day, were followed for at least five years and no one was lost, then the estimated five-year survival probability would be the number who lived more than five years, divided by the number who started the study. For most survival data, however, individuals typically enter the study at different times during the study period. The fact that the data are collected sequentially over time makes it necessary to sequentially piece together the follow-up information for a five-year period. It is also likely that during the course of collecting follow-up data, some individuals die from causes other than the one being investigated. Somewhat pragmatically, these individuals are usually classified as lost (u_x is increased), which does not bias subsequent calculations as long as these deaths are completely unrelated to the disease under study.

Notice that 15 kidney cancer patients in the 1951 cohort were withdrawn after one year. Clearly, each of these 15 patients were observed (survived) only a part of the year. If the exact survival times where known, then the total person-years-at-risk would be the sum of these individual times. When this information is not available, estimates of survival time must be adjusted to compensate for the incomplete nature of these observations. One approach is to assume that each person withdrawn during an interval, on the average, contributes one-half the length of the time interval $a_i = 0.5$ to the total survival time. That is, it is assumed that individuals come into the study randomly throughout the follow-up period, implying they are censored randomly from observation. If this were the case, then attributing an amount of time equal to one-half the interval length for each person withdrawn is “on the average” correct. A similar assumption is usually made about individuals lost from follow-up. Thus, individuals lost from follow-up are also assumed to contribute half the interval of survival time during the period they were lost. An estimate of the probability of death (q_x) that accounts for the two sources of incomplete information is made by reducing the number of persons beginning the interval (l_x) to compensate for the

unobserved time from individuals lost (u_x) and censored (w_x) during the interval. (p.359) Specifically, $l'_x = l_x - 0.5u_x - 0.5w_x$, and l'_x is called the *effective number of persons at risk* in the interval and the probability of death within the interval is then estimated by

$$q_x = \frac{d_x}{l'_x}.$$

The adjusted persons at risk (l'_x) better reflects the situation underlying the collection of follow-up data.

An alternate view of this adjustment comes from noting that the observed number of deaths each year since diagnosis is understated because lost and withdrawn individuals are not followed for, on the average, half an interval. Deaths occurring among these individuals that did not occur during the follow-up period are not recorded. An estimate of the additional number of "missing deaths" is $0.5(u_x + w_x)q_x$. Adding these "missing deaths" to the number of observed deaths gives an estimate of the probability of death as

$$q_x = \frac{d_x + 0.5(u_x + w_x)q_x}{l_x},$$

and solving this expression for q_x produces the same probability of death as before ($q_x = d_x/l'_x$).

Employing the value q_x to estimate the probability of death among those who were lost and withdrawn implies that these individuals do not differ in their mortality experience from those remaining in the study. This assumption may not be tenable in some situations. For example, it might be that lost individuals are more likely to have survived or, perhaps, more likely to have died. A suitable q_x should be used under these conditions. A more subtle implication of employing the estimate l'_x is the implicit assumption that mortality experience is unrelated to the reasons that an individual is withdrawn from follow-up.

Analogous to the life table calculation, the survival probabilities are

$$\hat{P}_k = \prod_{x=0}^{k-1} p_x,$$

where, as before, $p_x = 1 - q_x$ is the conditional probability of surviving the interval x. The estimate is the probability of surviving beyond the kth time interval. Estimating these probabilities (p_x) from the kidney cancer data produces a survival probability (P_k) for

each year of follow-up (Table 11–9). The estimated five-year survival probability is with standard error = 0.060.
(p.360)

Table 11–9. Calculation of a five-year survival probability from tabled data: calculations

Interval	l_x	d_x	l'_x	q_x	p_x		Πp_x	SE
0–1	126	47	116.5	0.403	0.597	$\hat{P}_1$	0.597	0.045
1–2	60	5	51.5	0.097	0.903	$\hat{P}_2$	0.539	0.048
2–3	38	2	30.5	0.066	0.934	$\hat{P}_3$	0.503	0.051
3–4	21	2	16.5	0.121	0.879	$\hat{P}_4$	0.442	0.060
4–5	10	0	7.0	0.000	1.000	$\hat{P}_5$	0.442	0.060
5–6	4	0	2.0	0.000	1.000	$\hat{P}_6$	0.442	0.060

The estimated variance of the distribution of the estimate of P_k (last column of Table 11–9) comes from the expression

$$\text{variance}(\hat{P}_k) = \hat{P}_k^2 \sum_{x=0}^{k-1} \frac{q_x}{l'_x p_x}.$$

This variance estimate is sometimes referred to as “Greenwood’s formula” after Major M. Greenwood, an early contributor to biostatistics [4]. As usual, an estimate of the variance is necessary to test hypotheses or construct confidence intervals to assess the impact of sampling variation on an estimated survival probability. Another estimate of the five-year survival probability is the number of individuals who survived more than five years, divided by the number of individuals who began the study at least five years previously, as mentioned. Only the 1946 cohort can be used to estimate the five-year survival probability in this manner (n = 9) because the other cohorts contain only individuals with less than five years of follow-up time. This five-year survival probability is 4/9 = 0.444 with a standard error of 0.166 (assuming the lost individual survived). Using all available data (n = 126) rather than a single cohort (n = 9) produces a more precise estimate of the five-year survival probability (ratio of standard errors = 0.166/0.060 = 2.7 in the kidney cancer example). The cost of this increased precision, however, is a possible bias caused by assuming that the mortality experience over the study period is similar enough among all cohorts that combining data for all years accurately reflects the overall mortality experience of the sampled population.

Another useful summary of survival data is an estimate of the mean survival time. For the kidney cancer data, the calculation of a mean value is complicated by the fact that the time of death is not known for all participating individuals. For the data recorded on the 126 kidney cancer patients (56 died and 64 censored), the mean survival time is 3.523 years. Mean survival time estimates are discussed in Chapter 12.

(p.361)

Table 11–10. WCGS data: a body-mass greater than the 75th-percentile

x to (x + 1)	l_x	d_x	w_x	q_x		SE
0–1	871	6	0	0.0069	0.993	0.0028
1–2	865	8	21	0.0094	0.984	0.0043
2–3	836	16	19	0.0194	0.965	0.0063
3–4	801	9	23	0.0114	0.954	0.0072
4–5	769	11	14	0.0144	0.940	0.0082
5–6	744	12	19	0.0163	0.925	0.0092
6–7	713	18	46	0.0261	0.901	0.0106
7–8	649	9	195	0.0163	0.886	0.0115
8–9	445	5	431	0.0218	0.867	0.0141
9–10	9	0	9	0.0000	0.867	0.0141

Survival patterns experienced by different groups are frequently summarized and compared using a specific survival probability. Two such groups from the WCGS data (Appendix A) are those with high values of the body mass index (greater than the 75th percentile) and those with smaller body mass values (less than the 75th percentile). The data and the calculated “survival” probabilities (in this case, “survival” means time free from a coronary event) are given in Tables 11–10 and 11–11.

The comparison of these survival probabilities (versus , Table 11–10 versus Table 11–11) shows a lower probability (higher risk) of “survival” for those individuals with a high body mass index. For example, the estimated five-year “survival” probability is for individuals with high values of body mass index compared to observed for individuals with “normal” values. The standard errors (Greenwood’s formula) associated with the distributions of these estimates (0.0082 and 0.0042, respectively) indicate that this difference is not likely to have occurred by chance. From a more formal prospective, the

Table 11–11. WCGS data: a body mass less than the 75th percentile

x to (x + 1)	l_x	d_x	w_x	q_x	$\hat{p}_i$	SE
0–1	2283	9	4	0.0039	0.996	0.0013
1–2	2270	20	24	0.0089	0.987	0.0024
2–3	2226	23	50	0.0104	0.977	0.0032
3–4	2153	18	41	0.0084	0.969	0.0037
4–5	2094	18	37	0.0087	0.960	0.0042
5–6	2039	27	61	0.0134	0.947	0.0048
6–7	1951	14	99	0.0074	0.940	0.0051
7–8	1838	22	502	0.0139	0.927	0.0057
8–9	1314	12	1271	0.0177	0.911	0.0073
9–10	31	0	31	0.0000	0.911	0.0073

(p.362) confidence intervals only slightly overlap. For the greater than 75th percentile group, the approximate 95% confidence interval based on is (0.924, 0.956), and for the less than 75th percentile group the confidence interval based on is (0.953, 0.969). Figure 11–3 is a plot of the “survival” probabilities for the two sources of data.

Three assumptions about the structure of the sampled population are typically made to calculate a survival probability using life table techniques. First, all lost and censored subjects are assumed to contribute, on the average, half the survival information of an individual followed for a complete year (or complete time interval). Second, it is assumed that the data collected for a number of calendar time periods can be combined to maximize the number of observations available to calculate the probability of death. Unbiased estimates of survival probabilities occur only when these cohorts experience the same pattern of mortality during the follow-up period (again, the absence of interaction permits the data to be summarized). In terms of the kidney cancer data, the individuals who entered the study in 1947, for example, are assumed to have the same pattern of mortality as the patients who entered in 1951, thereby allowing the data from both cohorts to be combined to calculate more precisely the probability of surviving the first year after diagnosis. The third assumption is that individuals lost and withdrawn from observation have the same probability of death as the individuals remaining in the follow-up study. This conjecture is probably the most

(p.363)

tenuous when applied to individuals lost from observation. Situations certainly arise where lost individuals experience a different survival pattern and the estimated probability q_x is affected. For example, if all individuals classified as lost actually survived, then

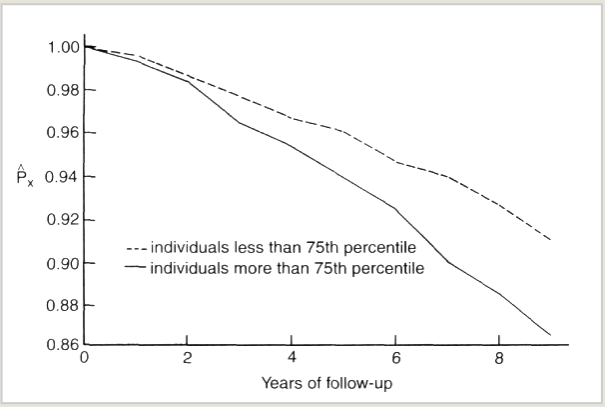

Figure 11–3. Survival probabilities for individuals with a body mass index less than and greater than the 75th percentile (WCGS data).

$$q'_x = \frac{d_x + 0.5w_x q'_x}{l_x} \quad \text{or} \quad q'_x = \frac{d_x}{l_x - 0.5w_x}$$

or, if all individuals lost in fact died, then

$$q''_x = \frac{d_x + u_x + 0.5w_x q''_x}{l_x} \quad \text{or} \quad q''_x = \frac{d_x + u_x}{l_x - 0.5w_x}.$$

The probabilities q'_x and q''_x represent the extremes in terms of the impact of the lost individuals on the calculation of the probability of death. These probabilities applied to the kidney cancer data yield five-year survival probabilities of if all lost patients survived and if all lost patients died compared to when all lost patients are assumed to have the same probability of death as those not lost to follow-up (lost at random). The range 0.328 to 0.454 provides limits to the potential bias caused by differences in the survival between individuals lost and individuals not lost to follow-up.

Life Table Measures of Specific Causes of Death

Hundreds of causes of death act within human populations. Two approaches based on life table methods provide an opportunity to isolate specific issues influencing the pattern of human mortality. These methods help resolve two questions:

1. What is the age structure throughout the life span associated with specific causes of death?
2. How does the probability of death from a specific cause change when other causes are “eliminated” from the population?

The first question is addressed by applying a *multiple-cause life table* (also called a *multiple-decrement life table*). The second question is addressed by a *competing risk analysis*.

A multiple-cause life table is similar to the single cause life table but describes the mortality patterns of a number of diseases simultaneously. The mechanics of constructing a multiple cause life table are defined and illustrated with data consisting of California resident males who died during 1980. The (p.364) causes of death data come from death certificates classified according to the ninth revision of the International Classification of Diseases (ICD9) [5]. These deaths are classified into four categories: deaths from lung cancer (ICD9, code 162), deaths from ischemic heart disease (ICD9, codes 410 to 414), deaths from motor vehicle accidents (ICD9, codes E810 to E819), and deaths from all other causes. Also necessary are age-specific population counts. The 1980 U.S. Census counts of California male residents are used to calculate rates. The following life table is abridged, which means that the lengths of the age intervals are not consistently one year. Most age intervals are five-year lengths (represented as δ_x; for example, $\delta_{60} = 5$ years). Although the example uses five-year age intervals, survival data can be organized into intervals of any lengths, even vary lengths. Regardless of the choice of the interval length, the principles of constructing a life table remain unchanged. Clearly, some technical details differ.

The basic components required to construct a multiple cause life table are the person-years-at-risk (typically approximated

by age-specific midyear populations) and the age-, cause-specific numbers of deaths. The notation is

D_x = total number of deaths in the age interval x to $x + \delta_x$,

$D_x^{(i)}$ = number of deaths from the ith cause in the age interval x to $x + \delta_x$, and

P_x = total person-years-at-risk in the age interval x to $x + \delta_x$.

These three quantities for male residents of California (1980) are recorded as part of the vital records system and U.S. Census counts (Table 11–12).

Average age-specific mortality rates calculated from Table 11–11 are $R_x = D_x/P_x$ for the age interval x to $x + \delta_x$ and, similar to the single-cause, complete life table,

$$q_x = \frac{\delta_x R_x}{1 + (1 - \bar{a}_x)\delta_x R_x}$$

is the estimated conditional probability of death within the interval x to $x + \delta_x$, where δ_x is the length of the interval starting at age x. That is, $q_x = P(\text{death between } x \text{ and } x + \delta_x \mid \text{alive at age } x)$. These probabilities are analogous to those calculated in the single-cause life table but applied to age intervals with length of δ_x years. For example, the estimated probability of death for individuals age 60 before age 65 is

$$q_{60} = \frac{5(0.0199)}{1 + 0.5(5)0.0199} = 0.0949 \qquad \text{where} \qquad R_{60} = \frac{9{,}319}{467{,}607} = 0.0199.$$

(p.365)

Table 11–12. Deaths from four causes: California, male residents, 1980

Age	P_x Population	D_i^1 Lung cancer	D_i^2 IHD*	D_i^3 Motor*	D_i^4 All other	D_x Total
0–1	193,310	1	2	3	2,507	2,513
1–4	515,150	1	3	58	375	437
5–9	843,750	0	2	90	195	287
10–14	915,240	0	1	80	248	329
15–19	1,091,684	3	1	523	1,162	1,689
20–24	1,213,068	4	6	965	1,507	2,482
25–29	1,132,811	3	13	627	1,665	2,308
30–34	1,008,606	12	63	437	1,547	2,059
35–39	776,545	36	136	277	1,371	1,820
40–44	629,452	85	306	201	1,510	2,102

Age	P_x Population	$D_x^{(1)}$ Lung cancer	$D_x^{(2)}$ IHD*	$D_x^{(3)}$ Motor*	$D_x^{(4)}$ All other	D_x Total
45–49	578,420	225	567	197	2,115	3,104
50–54	578,795	445	1050	150	3,163	4,808
55–59	573,119	786	1807	147	4,663	7,403
60–64	467,607	1059	2528	129	5,603	9,319
65–69	378,259	1297	3328	97	7,014	11,736
70–74	269,849	1266	3815	89	7,423	12,593
75–79	175,580	941	3793	99	7,508	12,341
80–84	95,767	557	3452	44	6,202	10,255
85+	78,832	430	5249	61	8,222	13,962
Total	11,515,844	7,151	26,122	4,274	64,000	101,547

(*) IHD = ischemic heart disease; motor = motor vehicle accidents.

To “fine tune” these calculations, values other than $a_x = 0.5$ can be used, but they have little influence on the final calculations for data covering the entire life span.

To estimate the cause-specific conditional probabilities of death (denoted $q_x^{(i)}$), the q_x values are distributed proportionally (prorated) by the observed numbers of death. Because

$$q_x^{(i)} = \frac{\delta_x D_x^{(i)}}{P_x + 0.5\delta_x D_x} \quad \text{and} \quad q_x = \frac{\delta_x D_x}{P_x + 0.5\delta_x D_x},$$

then

$$q_x^{(i)} = \frac{D_x^{(i)}}{D_x} q_x.$$

The value $q_x^{(i)}$ is the age- and cause-specific conditional probability of death from cause i before age $x + \delta_x$ for individuals alive at age x. Continuing the illustration (p.366) for the age interval 60 to 65, the probability of dying from lung cancer before age 65 for individuals alive at age 60 is

$$q_{60}^{(\text{lung})} = \frac{1059}{9319} 0.0949 = 0.0108.$$

These conditional probabilities for the 1980 California data are given in Table 11–13.

Because all causes of death are included, the probability . The -probabilities calculated from the California mortality data indicate that the cause-specific conditional probabilities for lung cancer increase rapidly after age 40 until about age 70

and then increase less rapidly for the older ages. The probabilities for ischemic heart disease also increase sharply at about age 70 but are generally associated with older individuals (shifted to the right). The conditional probabilities describing deaths from motor vehicle accidents , however, increase until ages 20 to 25, decrease and remain fairly constant until age 70 where they too sharply increase. The cause-specific probability curves describing these three causes of death are displayed in Figure 11–4 (smoothed; see Chapter 1). As before, these curves roughly approximate the hazard function associated with each specific cause of death.

Table 11–13. Conditional probabilities: California, male residents, 1980

Age	q_x Total	$q^{(1)}$ Lung cancer	$q^{(2)}$ IHD	$q^{(3)}$ Motor	$q^{(4)}$ All others
0–1	0.01292	0.00001	0.00001	0.00002	0.01289
1–4	0.00339	0.00001	0.00002	0.00045	0.00291
5–9	0.00170	0.00000	0.00001	0.00053	0.00115
10–14	0.00180	0.00000	0.00001	0.00044	0.00135
15–19	0.00771	0.00001	0.00000	0.00239	0.00530
20–24	0.01018	0.00002	0.00002	0.00396	0.00618
25–29	0.01014	0.00001	0.00006	0.00275	0.00731
30–34	0.01016	0.00006	0.00031	0.00216	0.00763
35–39	0.01165	0.00023	0.00087	0.00177	0.00878
40–44	0.01656	0.00067	0.00241	0.00158	0.01190
45–49	0.02648	0.00192	0.00484	0.00168	0.01804

Age	q_x Total	$q^{(1)}$ Lung cancer	$q^{(2)}$ IHD	$q^{(3)}$ Motor	$q^{(4)}$ All others
50–54	0.04069	0.00377	0.00889	0.00127	0.02677
55–59	0.06256	0.00664	0.01527	0.00124	0.03941
60–64	0.09492	0.01079	0.02575	0.00131	0.05707
65–69	0.14397	0.01591	0.04082	0.00119	0.08604
70–74	0.20896	0.02101	0.06330	0.00148	0.12317
75–79	0.29891	0.02279	0.09187	0.00240	0.18185
80–84	0.42235	0.02294	0.14217	0.00181	0.25543
85+	1.00000	0.03080	0.37595	0.00437	0.58888

(p.367)

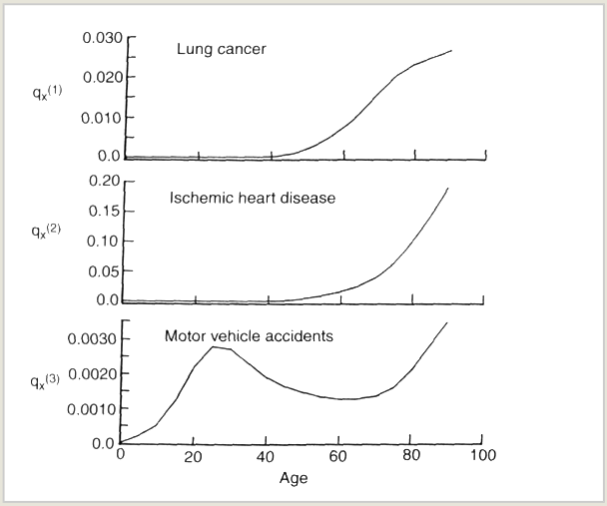

Figure 11–4. Cause-specific probabilities of death for three specific diseases (lung cancer, ischemic heart disease, and motor vehicle accidents) for males, California, 1980.

Again parallel to the single cause life table, an arbitrary number of individuals (l_0) can be distributed according to the conditional probabilities of death to produce a distribution of the number of life table deaths with a pattern of age-specific mortality described by the $q_x^{(i)}$-values (Table 11–13).

Starting with an arbitrary number of individuals (for example, $l_0 = 1{,}000{,}000$), the life table deaths (Table 11–14) result from applying the relationship

$$d_x^{(i)} = l_x q_x^{(i)},$$

where, as before, l_x represents the number of persons alive at the beginning of age interval x. For example, the life table number of persons age 60 dying from lung cancer between age 60 to 65 is $d_{60}^{(\text{lung})} = 802{,}800(0.0108) = 8{,}659$.

An additional table calculated by accumulating the deaths in each cause-specific category is also part of a description of the life table population. These (p.368)

Table 11–14. Life table “deaths” from four causes: California, male residents, 1980

Age	l_x Total	$d'_{x,1}$ Lung cancer	$d'_{x,2}$ IHD*	$d'_{x,3}$ Motor	$d'_{x,4}$ All other
0–1	1,000,000	5	10	15	12,885
1–4	987,084	8	23	444	2,869
5–9	983,740	0	12	524	1,136
10–14	982,069	0	5	429	1,329
15–19	980,305	13	4	2,339	5,197
20–24	972,751	16	24	3,849	6,012
25–29	962,850	13	55	2,651	7,040
30–34	953,091	56	296	2,054	7,272
35–39	943,412	217	821	1,673	8,280
40–44	932,421	624	2,248	1,476	11,091
45–49	916,982	1,760	4,435	1,541	16,543
50–54	892,703	3,362	7,933	1,133	23,896
55–59	856,379	5,689	13,078	1,064	33,748
60–64	802,800	8,659	20,671	1,055	45,814

Age	I_x Total	$d_x^{(1)}$ Lung cancer	$d_x^{(2)}$ IHD*	$d_x^{(3)}$ Motor	$d_x^{(4)}$ All other
65–60	726,601	11,560	29,663	865	62,517
70–74	621,996	13,066	39,374	919	76,611
75–79	492,026	11,214	45,203	1,180	89,476
80–84	344,954	7,913	49,042	625	88,111
85+	199,263	6,137	74,913	871	117,343

sums represent the number of individuals who reach age x and ultimately die from a specific cause. In symbols,

$$D_x^{(i)} = d_x^{(i)} + d_{x+\delta_x}^{(i)} + \cdots + d_{x'}^{(i)}.$$

To illustrate, $D_{60}^{(\text{lung})} = 8{,}659 + 11{,}560 + \cdots + 7{,}913 + 6{,}137 = 58{,}550$ is the number of individuals who reach age 60 who eventually die from lung cancer. Again from the California data, these accumulated life table deaths are given in Table 11–15.

The cumulative numbers of deaths provide the values necessary to calculate the probability of death before age x for each cause. That is, for the ith cause

$$F_x^{(i)} = 1 - \frac{D_x^{(i)}}{D_0^{(i)}}$$

is the probability of dying before age x. Among individuals dying from lung cancer, the probability of dying before age 60 is, for example,

$$F_{60}^{(\text{lung})} = 1 - \frac{58{,}550}{70{,}313} = 0.1673,$$

(p.369)

Table 11–15. Number of life table deaths after age *x* based on rates from California, male residents, 1980

Age	$D_i^{(1)}$ Lung cancer	$D_i^{(2)}$ IHD	$D_i^{(3)}$ Motor	$D_i^{(4)}$ All other
0–1	70,313	287,809	24,707	617,171
1–4	70,308	287,799	24,691	604,285
5–9	70,301	287,776	24,248	601,416
10–14	70,301	287,765	23,723	600,280
15–19	70,301	287,759	23,295	598,951
20–24	70,287	287,755	20,955	593,754
25–29	70,271	287,731	17,106	587,742
30–34	70,259	287,676	14,455	580,702
35–39	70,202	287,380	12,401	573,430
40–44	69,985	286,558	10,728	565,151
45–49	69,360	284,311	9,251	554,059
50–54	67,601	279,876	7,711	537,516

Age	$D_x^{(1)}$ Lung cancer	$D_x^{(2)}$ IHD	$D_x^{(3)}$ Motor	D_x^{*} All other
55–59	64,239	271,943	6,577	513,620
60–64	58,550	258,865	5,513	479,872
65–69	49,891	238,194	4,459	434,058
70–74	38,330	208,531	3,594	371,540
75–79	25,264	169,157	2,676	294,929
80–84	14,050	123,955	1,496	205,454
85+	6,137	74,913	871	117,343

or about 17% of the lung cancer deaths occur before age 60. Table 11–16 contains the cumulative probabilities of death (the $F_x^{(i)}$-probabilities) for the California 1980 data.

The age structure for each cause of death throughout the life span is apparent from the $F^{(i)}$-probabilities, and the patterns for each cause of death is evident. For example, 78% of all motor vehicle accident deaths occur by age 60, while 17% of lung cancer deaths occur by age 60. These cumulative probability distributions are displayed in Figure 11–5, and a few representative summary values are given in Table 11–17.

The cumulative distributions reveal distinct patterns of mortality associated with three specific causes. Motor vehicle accidents, as expected, have the greatest impact at the younger ages (median age among automobile deaths is 36.4 years) while, perhaps less expected, the ischemic heart disease is associated with the older ages, producing a median age at death of 78.8 years.

Lifetime Probability of Death

A multiple-cause life table allows a direct calculation of the lifetime probability of death from a specific cause, which is an occasionally used description of risk. **(p.370)**

Table 11–16. Cumulative distributions for four causes of death based on rates from California, male residents, 1980

Age	$F_1^{(1)}$ Lung cancer	$F_2^{(2)}$ IHD	$F_1^{(3)}$ Motor	$F_1^{(4)}$ All other
0–1	0.00000	0.00000	0.00000	0.00000
1–4	0.00007	0.00004	0.00062	0.02088
5–9	0.00018	0.00012	0.01859	0.02553
10–14	0.00018	0.00016	0.03980	0.02737
15–19	0.00018	0.00017	0.05716	0.02952
20–24	0.00037	0.00019	0.15184	0.03794
25–29	0.00060	0.00027	0.30764	0.04768
30–34	0.00078	0.00046	0.41494	0.05909
35–39	0.00158	0.00149	0.49809	0.07087
40–44	0.00467	0.00435	0.56580	0.08429
45–49	0.01355	0.01216	0.62555	0.10226
50–54	0.03858	0.02757	0.68792	0.12906
55–59	0.08640	0.05513	0.73379	0.16778
60–64	0.16730	0.10057	0.77685	0.22246
65–69	0.29045	0.17239	0.81954	0.29670

Age	$F_1^{(1)}$ Lung cancer	$F_2^{(2)}$ IHD	$F_1^{(3)}$ Motor	$F_1^{(4)}$ All other
70–74	0.45486	0.27545	0.85453	0.39799
75–79	0.64069	0.41226	0.89171	0.52213
80–84	0.80018	0.56932	0.93946	0.66710
85+	0.91272	0.73971	0.96476	0.80987

The probability of dying from a specific cause is estimated by the number of people who died of that cause divided by the number of persons who could have died (those at risk). The expected numbers of deaths after a specific age produce this probability (Table 11–15). The first row in the table contains the total number of individuals ultimately dying from each cause over the entire life span. Since 1,000,000 males make up the 1980 California life table "population-at-risk" (sum of the first row of Table 11–15), then P(dying from lung cancer) = 70,313/1,000,000 = 0.070

P(dying from ischemic heart disease) = 287,809/1,000,000 = 0.288

P(dying from a motor vehicle accident) = 24,707/1,000,000 = 0.025

are the lifetime probabilities of dying from any one of the three specific causes and

P(dying from other causes) = 617,171/1,000,000 = 0.617.

Each row in the table allows the calculation of the lifetime probability associated with individuals of a specific age. For example, for males age 60, the (p.371)

lifetime probability of dying from lung cancer is

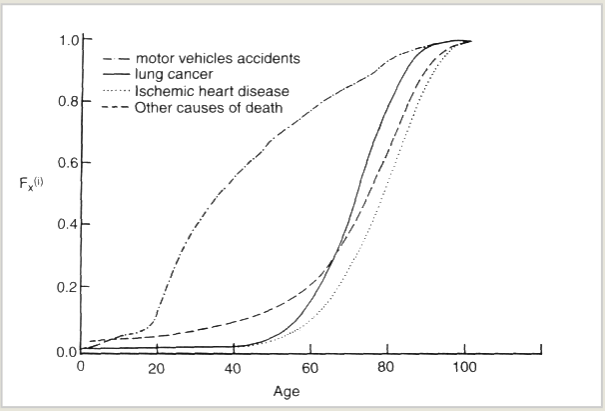

Figure 11–5. Cumulative probabilities of age at death for three specific diseases (lung cancer, ischemic heart disease, and motor vehicle accidents) for males, California, 1980.

58,550/802,800 = 0.073 where 802,800 individuals are alive at the beginning of the age interval 60–65 (the sum of the row age 60–65 of Table 11–15), and 58,550 is the number who died from lung cancer after age 60. Three cause-specific conditional probabilities for the 1980 California data are
P(dying from lung cancer after age 60) = 58,550/802,800 = 0.073

P(dying from ischemic heart disease after age 60) = 258,865/802,800 = 0.322

P(dying from a motor vehicle accident after age 60) = 5,513/802,800 = 0.007

P(dying from other causes after age 60) = 479,872/802,800 = 0.598.

Table 11–17. Median age (as well as 25th and 75th percentiles) at death

	25th percentile	Median	75th percentile
Lung cancer	64.5	72.0	78.8

	25th percentile	Median	75th percentile
Ischemic heart disease	69.3	78.8	86.4
Motor vehicle accidents	24.2	36.4	58.1
Other causes	63.2	75.0	83.8
All causes	63.4	74.6	83.9

(p.372) The disease-specific cumulative probability of death is related to the lifetime probability of death from a specific cause. The probability $1 - F_i^{(x)}$ is the conditional probability of death after age x among those who ultimately die from cause i. The lifetime probability of death from a specific cause i is the conditional probability of death from cause i for all individuals who reach age x. That is, the first probability is P(death after age x | death from cause i) and the second is P(death from cause i | death after age x) Specifically, for lung cancer deaths and age 60, P(death after 60 | death from lung cancer) = 58,550/70,313 = 0.833 and P(death from lung cancer | death after age 60) = 58,550/802,800 = 0.073. The relationship between these two probabilities is (Bayes' theorem)

$$P(\text{death cause } i \mid \text{death after age } x) = \frac{P(\text{death after age } x \mid \text{death cause } i)P(\text{death cause } i)}{P(\text{death after age } x)}.$$

COMPETING RISKS

British statistician William Fair (1875) was among the first to discuss the impact on the risk of a specific disease while other risks operated in a population. This problem was also explored by the early French mathematicians Bernoulli and D'Alembert and later by a British actuary Makeham. The issues are neatly summarized by a simple example given by Berkson and Elveback [6]:

Two marksmen shoot at a range of targets under conditions in which, if a target is struck, it instantly drops from view so that it cannot be struck again. Represent the striking rate of marksman 1, that is the crude probability of a hit when he is firing alone, as Q_1 and similarly the rate of marksman 2 when he is firing alone as Q_2. The probability when one risk operates alone is called the net risk or rate and is represented by upper case Q; when it operates together with another risk it is called the crude risk or rate and is represented by lower case q.

Suppose N targets are exposed and marksman 1 shoots first, followed by marksman 2: Rate for 1 is $q_1 = Q_1$ Rate for 2 is $q_2 = (1 - Q_1)Q_2$ Total rate is $q = q_1 + q_2 = Q_1 + Q_2 - Q_1 Q_2$ Suppose marksman 2 shoots first, followed by marksman 1, then: Rate for 2 is $q_2 = Q_2$ Rate for 1 is $q_1 = (1 - Q_2)Q_1$ Total rate is $q = q_1 + q_2 = Q_1 + Q_1 - Q_1 Q_2$.

(p.373)

It is seen that the total crude rate with both marksmen shooting is the same, whichever marksmen shoots first and assuming independence of the net probabilities Q_1 and Q_2, this will be true in general. Regardless of the ordering of the shooting or whether the two marksman shoot together, the total crude rate is given by the “total rate,” which, of course, can be derived as the complement of the product of the probabilities, $P_1 = 1 - Q_1$ and $P_2 = 1 - Q_2$, of not being struck (survival rate). [That is, $1 - (1 - Q_1)(1 - Q_2) = q$.]

If, from independent trials, we know Q_1, the net rate of marksman 1, and have a record of q, the crude rate when both shot together, we can derive the net rate Q_2 from “total rate”:

$$Q_2 = \frac{q - Q_1}{1 - Q_1}.$$

Rarely are the net probabilities Q_1 or Q_2 known, but the crude probabilities q_1, q_2 and q can be estimated from data. Manipulation of these estimated probabilities, under specific conditions, allows estimation of the net probabilities.
The following description of the mechanics of competing risks is focused on only two causes of death, and only a single time interval is considered. These two restrictions do not affect the principles underlying the competing risk argument (mathematicians say, “there is no loss of generality”) and simplify notation and explanations.

The formal definitions of the two central probabilities are as follows:

> *Crude probability:* q_i = the probability that an individual who is alive at the start of the interval dies from cause i when cause j is present, sometimes called the *mixed probability of death*.
>
> *Net probability:* Q_i = the probability that an individual who is alive at the start of the interval dies from cause i when cause j is not present, sometimes called the *pure probability of death*.

The marksman example shows a relationship between the net and crude probabilities, but it is not much use unless one of the net probabilities is known. To estimate the net probabilities, further statistical structure is required. First, it

is assumed that the net probabilities are described by two exponential functions, where λ_1 and λ_2 are hazard rates associated with causes 1 and 2, respectively, and are constant within the time period under consideration. The exponential model applied to survival analysis is explored in more detail in the next chapter. Under these conditions, the net probabilities of death are and . Cause 2 can be thought of as a specific cause of death and cause 1 as all other causes combined. For example, cause 2 could be death from coronary heart disease, and cause 1 could be all other causes. The net probability Q_1 (p.374) describes the probability of death uninfluenced by death from cause 2. Specifically, Q_1 would be the probability of death as if the influence from coronary heart disease was "eliminated."

Second, it is assumed that the probability of surviving the time interval is $P(\text{surviving}) = P_1P_2 = (1 - Q_1)(1 - Q_2) = (e^{-\lambda_1})(e^{-\lambda_2}) = e^{-\lambda_1+-\lambda_2} = e^{-\lambda}$, where $\lambda = \lambda_1 + \lambda_2$. That is, cause 1 and cause 2 are statistically independent. Although death from cause 1 is mutually exclusive of death from cause 2, it is still required that the mechanisms underlying these two events act independently. In terms of the marksman example, independence means that the hits and misses of one marksman do not influence the hits and misses of the other marksman, and vice versa. Equivalently, cause of death 1 is assumed to be unrelated in any way to cause of death 2. Independence of causes of death is definitely not a realistic assumption for some diseases, particularly certain chronic diseases. For example, smoking-related diseases such as lung cancer and emphysema are not independent. The influence of the dependency among causes of death on the estimate of the net probabilities has not been extensively studied.

These two assumptions (exponential net probabilities and independence) make it possible to estimate the probability of death from one cause when the other cause is "eliminated" from consideration (net probability). To estimate the net probability of death, a bit of algebra relates the crude and net probabilities. The crude probability of death (denoted q), death from either cause 1 or 2, is $P(\text{death}) = q = 1 - P_1 P_2 = 1 - e$

$^{-\lambda}$. The crude probability has the same form as both net probabilities (exponential). Furthermore,

$$(1-q)^{\frac{\lambda_i}{\lambda}} = e^{-\lambda_i} = P_i \qquad \text{giving} \qquad Q_i = 1 - P_i = 1 - (1-q)^{\frac{\lambda_i}{\lambda}}.$$

This relationship allows the estimation of the net probabilities from the crude probabilities because the ratio of the two constant hazard rates λ_i/λ is estimated by d_i/d, where d_i represents the number of deaths from cause i and $d = d_1 + d_2$ represents the total number of deaths from both causes in the time interval being considered. The estimated net probability of death from cause i with cause j "eliminated" is then estimated by

$$\hat{Q}_i = 1 - \left(1 - \frac{d}{l}\right)^{\frac{d_i}{d}},$$

where l individuals are at risk from both causes of death.
An alternative estimate of the net probability of death can be derived from intuitive considerations that do not directly involve an exponential risk model. Individuals can be classified into three categories: (*1*) died of cause 1, (*2*) died of cause 2, or (*3*) lived through the interval. A death from cause 2 can be considered (p.375) as a person "lost to follow-up" with respect to calculations for cause 1. When the possibility of death from cause 2 is "eliminated," deaths from cause 1 are under counted because the individuals previously "lost to follow-up" are now at risk. That is, the directly estimated probability is too small because a proportion of the individuals who would have died from cause 2 and become "lost" can now die from cause 1. Those who would have died from cause 2 are now exposed to risk, on the average, for half the interval so that $0.5d_2$ represents an additional "effective" number of individuals at risk when cause 2 is "eliminated." The value

$0.5d_2 Q_1$ estimates the number of additional deaths from cause 1 among the individuals who would have died from cause 2, if it were present. Therefore, "correcting" the number of deaths d_1 gives

$$\hat{Q}'_1 = \frac{d_1 + 0.5d_2\hat{Q}'_1}{l}$$

and solving for the net probability Q' yields the estimate

$$\hat{Q}'_1 = \frac{d_1}{l - 0.5d_2}.$$

The probability is an alternative estimate of the net probability of death from cause 1 based on l individuals at risk, where, again, d_1 represents the observed number of deaths from cause 1 and d_2 represents the observed number of deaths from cause 2. The estimated net probability is greater than crude probability q_1 because additional individuals are at risk and die from cause 1 when cause 2 is "eliminated." In general,

$$\text{net probability} = \hat{Q}'_1 = \frac{d_i}{l - 0.5d_j} \geq \frac{d_i}{l} = \hat{q}_i = \text{crude probability}.$$

In fact, the assumptions underlying both estimates of the net probabilities are essentially the same. Furthermore, for many applications, the crude probability and the net probability hardly differ. The expression for indicates why. For and to differ substantially from q_i, the number of deaths from the competing cause must be a large proportion of the individuals at risk (d_j has to be substantial relative to l), which is not

usually the case for human mortality. In fact, it is usually quite the opposite. The number of deaths d_j is almost always much less than .

The estimation of the net probabilities (exponential and intuitive) are illustrated by a small subset of data from a large study of the effects of smoking on the risk of coronary heart disease (CHD) mortality (Hammond and Horn [7] and (p.376)

Table 11–18. Competing risks: deaths after 44 months of follow-up for ages 60–65

	Nonsmokers	Smokers
CHD = d_1	552	921
Other = d_2	714	1095
Population	20278	21594
Crude	0.0272	0.0427
Exponential	0.0277	0.0438
Intuitive	0.0277	0.0438

reported in [6]; Table 11–18). As expected, the net probabilities of death from CHD for smokers and nonsmokers increase, but slightly, when competing causes of death are “eliminated.” Occasionally the argument is put forth that increases in cancer incidence in the last half of the twentieth century, at least in part, are due to the decrease in mortality from infectious diseases. This conjecture is based on the idea that deaths from infectious diseases operate early in life, thereby eliminating a proportion of individuals who would die from cancer later in life. In short, it was thought that deaths from infectious diseases were a substantial competing risk. National mortality data collected for the years 1900 to 1950 reflect on this question (Table 11–19).

Using competing risk estimates, the net probabilities (infectious disease eliminated) show no reason to believe that mortality from infectious disease plays an appreciable role in the observed rate of cancer mortality. Estimation of the crude and net probabilities are essentially identical for all six decades. That is, under the conditions for a competing risk

calculation, “eliminating” infectious disease as a cause of death does not appreciably influence the national mortality pattern over the years 1900 to 1950.

Table 11–19. Competing risks: infectious disease deaths by year for the United States, 1900–50

Year	1900	1910	1920	1930	1940	1950
Infection	240,077	225,565	191,958	137,971	90,239	60,370
Total deaths	1,308,056	1,356,535	1,382,887	1,394,611	1,422,161	1,472,842
Population**	76,094	92,407	106,466	123,188	132,122	151,683
Crude*	315.50	244.10	180.30	112.00	68.30	39.80
Net*	318.25	245.91	181.48	112.64	68.67	39.99
Net*	318.24	245.90	181.48	112.64	68.67	39.99

(*) Crude and net probabilities multiplied by 100,000.

(**) U.S. Census counts divided by 100,000.

(p.377)

Table 11–20. Expected years of remaining life time with specific competing causes of death "eliminated," California, male residents, 1980

	All causes	CVD*	IHD*	Lung cancer*	Motor*
e_0	70.92	80.63	73.79	71.80	71.81

(*) = the cause of death eliminated (cause *j*).

Net probabilities can be calculated from specific causes of death and summarized with life table functions. The exponential-based expression for a net probability of death from cause *i* at age *x* using life table deaths is

$$Q_{x,i} = 1 - (1 - q_x)^{\frac{d_x^{(i)}}{d_x}},$$

where represents life table deaths from the cause *i* in the age interval *x* to *x* + 1 and represents the total life table deaths in that interval. The net life table probabilities $Q_{x,i}$ reflect the effect of mortality at age *x* from cause *i* with cause *j* "eliminated"; these probabilities can be used to calculate other life table functions, particularly the expectation of life (denoted $e_{x,i}$). For example, if all deaths from cardiovascular disease (caused *j*) are "eliminated" and a life table based on all other deaths (cause *i*) is constructed from the net probabilities, then an estimate of the total years of life lost attributable to cardiovascular disease is found by comparing the increase in "net" expected years of life to the same value calculated when all causes of death are operating (80.6 years versus 70.9 years; Table 11–20).

Table 11–20 gives the expected years of life at birth for California males in 1980. Estimates based on a life table constructed from the crude probabilities of death q_x (no causes of death "eliminated") produce the usual mean years of remaining life time (e_0 = 70.9 years). Also, included in Table 11–20 are the expected years of life when four causes of death

—cardiovascular disease (CVD), ischemic heart disease (IHD), lung cancer, and motor vehicle accidents (motor)—are each "eliminated." Four life tables were constructed (not shown) from four sets of age-specific Q_x,i-probabilities; they produce the mean years of life remaining at age $x(e_x,i)$.

The effect of cardiovascular disease on the total mortality picture is clear. The life table competing risk calculation indicates that the expected years of life would be increased about 10 years if cardiovascular disease was "eliminated" as a risk of death (from $e_0 = 70.92$ to $e_{0.\mathrm{CVD}} = 80.63$ years). A three-year increase would result if deaths from ischemic heart disease were "eliminated." Less of an impact on the expected years of life is observed (about a one-year increase) when lung cancer or motor vehicle accidents are "eliminated" as causes of death. The word eliminated is in quotes as a reminder that the calculations are based on two assumptions—exponential survival probabilities and independence of causes of death.