
Contents

I	Severity, Frequency, and Aggregate Loss	1
1	Basic Probability	3
1.1	Functions and moments	3
1.2	Percentiles	7
1.3	Conditional probability and expectation	8
1.4	Moment and probability generating functions	10
1.5	The empirical distribution	11
	Exercises	12
	Solutions	19
2	Parametric Distributions	27
2.1	Scaling	27
2.2	Transformations	28
2.3	Common parametric distributions	30
2.3.1	Uniform	30
2.3.2	Beta	30
2.3.3	Exponential	31
2.3.4	Weibull	32
2.3.5	Gamma	32
2.3.6	Pareto	33
2.3.7	Single-parameter Pareto	34
2.3.8	Lognormal	35
2.4	The linear exponential family	35
2.5	Limiting distributions	38
	Exercises	40
	Solutions	42
3	Variance	45
3.1	Additivity	45
3.2	Normal approximation	46
3.3	Bernoulli shortcut	47
	Exercises	48
	Solutions	49
4	Mixtures and Splices	51
4.1	Mixtures	51
4.1.1	Discrete mixtures	51
4.1.2	Continuous mixtures	53
4.1.3	Frailty models	54
4.2	Conditional Variance	55
4.3	Splices	57
	Exercises	61
	Solutions	66

5 Policy Limits	75
Exercises	78
Solutions	80
6 Deductibles	85
6.1 Ordinary and franchise deductibles	85
6.2 Payment per loss with deductible	85
6.3 Payment per payment with deductible	87
Exercises	90
Solutions	103
7 Loss Elimination Ratio	111
Exercises	112
Solutions	118
8 Risk Measures and Tail Weight	127
8.1 Coherent risk measures	127
8.2 Value-at-Risk (VaR)	129
8.3 Tail-Value-at-Risk (TVaR)	131
8.4 Tail Weight	136
Exercises	139
Solutions	141
9 Other Topics in Severity Coverage Modifications	149
Exercises	152
Solutions	158
10 Bonuses	169
Exercises	170
Solutions	172
11 Discrete Distributions	177
11.1 The $(a, b, 0)$ class	177
11.2 The $(a, b, 1)$ class	180
Exercises	182
Solutions	186
12 Poisson/Gamma	195
Exercises	196
Solutions	200
13 Frequency— Exposure & Coverage Modifications	205
13.1 Exposure modifications	205
13.2 Coverage modifications	205
Exercises	207
Solutions	210
14 Aggregate Loss Models: Compound Variance	215
14.1 Introduction	215
14.2 Compound variance	216
Exercises	219
Solutions	226

15 Aggregate Loss Models: Approximating Distribution	237
Exercises	240
Solutions	248
16 Aggregate Loss Models: The Recursive Formula	257
Exercises	260
Solutions	263
17 Aggregate Losses—Aggregate Deductible	267
Exercises	272
Solutions	278
18 Aggregate Losses: Miscellaneous Topics	285
18.1 Coverage modifications	285
18.2 Analytic results	286
18.2.1 The Erlang distribution function	286
18.2.2 Negative binomial/exponential compound models	288
18.2.3 Compound Poisson models	289
18.3 Discretizing	289
18.3.1 Method of rounding	290
18.3.2 Method of local moment matching	290
Exercises	292
Solutions	297
II Empirical Models	307
19 Review of Mathematical Statistics	309
19.1 Estimator quality	309
19.1.1 Bias	310
19.1.2 Consistency	311
19.1.3 Variance and mean square error	312
19.2 Hypothesis testing	313
19.3 Confidence intervals	315
Exercises	317
Solutions	322
20 The Empirical Distribution for Complete Data	329
20.1 Individual data	329
20.2 Grouped data	330
Exercises	331
Solutions	333
21 Variance of Empirical Estimators with Complete Data	335
21.1 Variance	335
21.1.1 Individual data	335
21.1.2 Grouped data	336
Exercises	339
Solutions	342
22 Kaplan-Meier and Nelson-Åalen Estimators	347
22.1 Kaplan-Meier Product Limit Estimator	348

22.2 Nelson-Åalen Estimator	352
Exercises	353
Solutions	362
23 Estimation of Related Quantities	371
23.1 Moments and related quantities	371
23.1.1 Complete individual data	371
23.1.2 Grouped data	371
23.1.3 Incomplete data	373
23.2 Range probabilities	375
23.3 Deductibles and limits	376
23.4 Inflation	377
Exercises	378
Solutions	382
24 Variance of Kaplan-Meier and Nelson-Åalen Estimators	387
Exercises	390
Solutions	396
25 Kernel Smoothing	405
25.1 Density and distribution	405
25.1.1 Uniform kernel	406
25.1.2 Triangular kernel	411
25.1.3 Other symmetric kernels	417
25.1.4 Kernels using two-parameter distributions	418
25.2 Moments of kernel-smoothed distributions	419
Exercises	422
Solutions	426
26 Approximations for Large Data Sets	433
Exercises	439
Solutions	442
III Parametric Models	447
27 Method of Moments	449
27.1 Introductory remarks	449
27.2 The method of moments for various distributions	450
27.2.1 Exponential	450
27.2.2 Gamma	450
27.2.3 Pareto	451
27.2.4 Lognormal	452
27.2.5 Uniform	452
27.2.6 Other distributions	453
27.3 Fitting other moments, and incomplete data	453
Exercises	456
Solutions	463
28 Percentile Matching	475
28.1 Smoothed empirical percentile	475

28.2	Percentile matching for various distributions	476
28.2.1	Exponential	476
28.2.2	Weibull	477
28.2.3	Lognormal	477
28.2.4	Other distributions	478
28.3	Percentile matching with incomplete data	479
28.4	Matching a percentile and a moment	480
	Exercises	481
	Solutions	487
29	Maximum Likelihood Estimators	495
29.1	Defining the likelihood	496
29.1.1	Individual data	497
29.1.2	Grouped data	498
29.1.3	Censoring	499
29.1.4	Truncation	499
29.1.5	Combination of censoring and truncation	500
	Exercises	501
	Solutions	510
30	Maximum Likelihood Estimators—Special Techniques	517
30.1	Cases where the Maximum Likelihood Estimator equals the Method of Moments Estimator	517
30.1.1	Exponential distribution	517
30.2	Parametrization and Shifting	518
30.2.1	Parametrization	518
30.2.2	Shifting	519
30.3	Transformations	519
30.3.1	Lognormal distribution	519
30.3.2	Inverse exponential distribution	520
30.3.3	Weibull distribution	521
30.4	Special distributions	521
30.4.1	Uniform distribution	521
30.4.2	Pareto distribution	522
30.4.3	Beta distribution	523
30.5	Bernoulli technique	524
30.6	Estimating q_x	527
	Exercises	528
	Solutions	542
31	Variance Of Maximum Likelihood Estimators	557
31.1	Information matrix	557
31.1.1	Calculating variance using the information matrix	557
31.1.2	Asymptotic variance of MLE for common distributions	561
31.1.3	True information and observed information	564
31.2	The delta method	566
31.3	Confidence Intervals	568
31.3.1	Normal Confidence Intervals	568
31.3.2	Non-Normal Confidence Intervals	569
	Exercises	571
	Solutions	580

32 Fitting Discrete Distributions	591
32.1 Poisson distribution	591
32.2 Negative binomial	592
32.3 Binomial	592
32.4 Fitting $(a, b, 1)$ class distributions	593
32.5 Adjusting for exposure	596
32.6 Choosing between distributions in the $(a, b, 0)$ class	596
Exercises	599
Solutions	606
33 Hypothesis Tests: Graphic Comparison	613
33.1 $D(x)$ plots	613
33.2 p - p plots	614
Exercises	616
Solutions	619
34 Hypothesis Tests: Kolmogorov-Smirnov	623
34.1 Individual data	623
34.2 Grouped data	626
Exercises	628
Solutions	635
35 Hypothesis Tests: Anderson-Darling	641
Exercises	642
Solutions	643
36 Hypothesis Tests: Chi-square	647
36.1 Definition of chi-square statistic	647
36.2 Degrees of freedom	649
36.3 Other requirements for the chi-square test	651
36.4 Data from several periods; handling the models from Section 32.5	653
Exercises	655
Solutions	671
37 Likelihood Ratio Algorithm, Schwarz Bayesian Criterion	679
37.1 Likelihood Ratio algorithm	679
37.2 Schwarz Bayesian Criterion	682
Exercises	683
Solutions	686
IV Credibility	691
38 Limited Fluctuation Credibility: Poisson Frequency	695
Exercises	701
Solutions	708
39 Limited Fluctuation Credibility: Non-Poisson Frequency	715
Exercises	718
Solutions	721

40 Limited Fluctuation Credibility: Partial Credibility	727
Exercises	728
Solutions	734
41 Bayesian Methods—Discrete Prior	737
Exercises	741
Solutions	753
42 Bayesian Methods—Continuous Prior	771
42.1 Calculating posterior and predictive distributions	771
42.2 Recognizing the posterior distribution	775
42.3 Loss functions	776
42.4 Interval estimation	777
42.5 The linear exponential family and conjugate priors	778
Exercises	779
Solutions	786
43 Bayesian Credibility: Poisson/Gamma	799
Exercises	800
Solutions	809
44 Bayesian Credibility: Normal/Normal	813
Exercises	816
Solutions	818
45 Bayesian Credibility: Bernoulli/Beta	821
45.1 Bernoulli/beta	821
45.2 Negative binomial/beta	824
Exercises	825
Solutions	828
46 Bayesian Credibility: Exponential/Inverse Gamma	831
Exercises	835
Solutions	838
47 Bühlmann Credibility: Basics	841
Exercises	845
Solutions	850
48 Bühlmann Credibility: Discrete Prior	857
Exercises	861
Solutions	878
49 Bühlmann Credibility: Continuous Prior	895
Exercises	896
Solutions	908
50 Bühlmann-Straub Credibility	921
50.1 Bühlmann-Straub model: Varying exposure	921
50.2 Hewitt model: Generalized variance of observations	922
Exercises	925
Solutions	930

51 Exact Credibility	937
Exercises	937
Solutions	941
52 Bühlmann As Least Squares Estimate of Bayes	945
52.1 Regression	945
52.2 Graphic questions	946
52.3 $\text{Cov}(X_i, X_j)$	948
Exercises	949
Solutions	953
53 Empirical Bayes Non-Parametric Methods	955
53.1 Uniform exposures	956
53.2 Non-uniform exposures	958
53.2.1 No manual premium	958
53.2.2 Manual premium	964
Exercises	965
Solutions	972
54 Empirical Bayes Semi-Parametric Methods	983
54.1 Poisson model	983
54.2 Non-Poisson models	987
54.3 Which Bühlmann method should be used?	989
Exercises	991
Solutions	997
V Simulation	1003
55 Simulation—Inversion Method	1005
Exercises	1009
Solutions	1018
56 Number of Data Values to Generate	1027
Exercises	1031
Solutions	1033
57 Simulation—Applications	1037
57.1 Actuarial applications	1037
57.2 Statistical analysis	1039
57.3 Risk measures	1039
Exercises	1041
Solutions	1049
58 Bootstrap Approximation	1057
Exercises	1060
Solutions	1063
VI Practice Exams	1067
1 Practice Exam 1	1069

2 Practice Exam 2	1079
3 Practice Exam 3	1089
4 Practice Exam 4	1097
5 Practice Exam 5	1105
6 Practice Exam 6	1115
7 Practice Exam 7	1125
8 Practice Exam 8	1135
9 Practice Exam 9	1145
10 Practice Exam 10	1155
Appendices	1165
A Solutions to the Practice Exams	1167
Solutions for Practice Exam 1	1167
Solutions for Practice Exam 2	1179
Solutions for Practice Exam 3	1190
Solutions for Practice Exam 4	1203
Solutions for Practice Exam 5	1215
Solutions for Practice Exam 6	1226
Solutions for Practice Exam 7	1238
Solutions for Practice Exam 8	1249
Solutions for Practice Exam 9	1260
Solutions for Practice Exam 10	1273
B Solutions to Old Exams	1291
B.1 Solutions to CAS Exam 3, Spring 2005	1291
B.2 Solutions to SOA Exam M, Spring 2005	1295
B.3 Solutions to CAS Exam 3, Fall 2005	1298
B.4 Solutions to SOA Exam M, Fall 2005	1301
B.5 Solutions to Exam C/4, Fall 2005	1305
B.6 Solutions to CAS Exam 3, Spring 2006	1315
B.7 Solutions to CAS Exam 3, Fall 2006	1318
B.8 Solutions to SOA Exam M, Fall 2006	1321
B.9 Solutions to Exam C/4, Fall 2006	1324
B.10 Solutions to Exam C/4, Spring 2007	1333
C Exam Question Index	1343

Lesson 22

Kaplan-Meier and Nelson-Åalen Estimators

Reading: *Loss Models* Third Edition 14.1

Exams routinely feature questions based on the material in this lesson.

When conducting a study, we often do not have complete data, and therefore cannot use raw empirical estimators. Data may be incomplete in two ways:

1. No information at all is provided for certain ranges of data. Examples would be:
 - An insurance policy has a deductible d . If a loss is for an amount d or less, it is not submitted. Any data you have regarding losses is conditional on the loss being greater than d .
 - You are measuring amount of time from disablement to recovery, but the disability policy has a six-month elimination period. Your data only includes cases for which disability payments were made. If time from disablement to recovery is less than six months, there is no record in your data.

When data are not provided for a range, the data is said to be **truncated**. In the two examples just given, the data are *left truncated*, or *truncated from below*. It is also possible for data to be truncated from above, or right truncated. An example would be a study on time from disablement to recovery conducted on June 30, 2009 that considers only disabled people who recovered by June 30, 2009. For a group of people disabled on June 30, 2006, this study would truncate the data at time 3, since people who did not recover within 3 years would be excluded from the study.

2. The exact data point is not provided; instead, a range is provided. Examples would be:
 - An insurance policy has a policy limit u . If a loss is for an amount greater than u , the only information you have is that the loss is greater than u , but you are not given the exact amount of the loss.
 - In a mortality study on life insurance policyholders, some policyholders surrender their policy. For these policyholders, you know that they died (or will die) some time after they surrender their policy, but don't know the exact amount of death.

When a range of values rather than an exact value is provided, the data is said to be **censored**. In the two examples just given, the data are *right censored*, or *censored from above*. It is also possible for data to be censored from below, or left censored. An example would be a study of smokers to determine the age at which they started smoking in which for smokers who started below age 18 the exact age is not provided.

We will discuss techniques for constructing data-dependent estimators in the presence of left truncation and right censoring. Data-dependent estimators in the presence of right truncation or left censoring are beyond the scope of the syllabus.¹

¹However, parametric estimators in the presence of right truncation or left censoring are not excluded from the syllabus. We will study parametric estimators in Lessons 27–30.

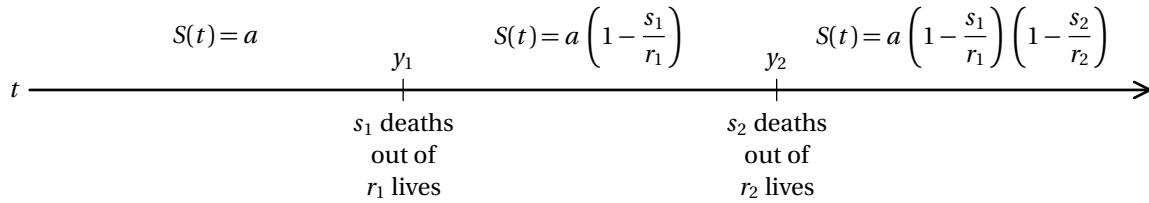


Figure 22.1: Illustration of the Kaplan-Meier product limit estimator. The survival function is initially a . After each event time, it is reduced in the same proportion as the proportion of deaths in the group.

22.1 Kaplan-Meier Product Limit Estimator

The first technique we will study is the *Kaplan-Meier product limit estimator*. We shall discuss its use for estimating survival distributions for mortality studies, but it may be used just as easily to estimate $S(x)$, and therefore $F(x)$, for loss data. To motivate it, consider a mortality study starting with n lives. Suppose that right before time y_1 , we have somehow determined that the survival function $S(y_1^-)$ is equal to a . Now suppose that there are r_1 lives in the study at time y_1 . Note that r_1 may differ from n , since lives may have entered or left the study between inception and time y_1 . Now suppose that at time y_1 , s_1 lives died. See Figure 22.1 for a schematic. The proportion of deaths at time y_1 is s_1/r_1 . Therefore, it is reasonable to conclude that the conditional survival rate past time y_1 , given survival to time y_1 , is $1 - s_1/r_1$. Then the survival function at time y_1 should be multiplied by this proportion, making it $a(1 - s_1/r_1)$. The same logic is repeated at the second event time y_2 in Figure 22.1, so that the survival function at time y_2 is $a(1 - s_1/r_1)(1 - s_2/r_2)$.

Suppose we have a study where the event of interest, say death, occurs at times $y_j, j \geq 1$. At each time y_j , there are r_j individuals in the study, out of which s_j die. Then the Kaplan-Meier estimator of $S(t)$ sets $S_n(t) = 1$ for $t < y_1$. Then recursively, at the j th event time y_j , $S_n(y_j)$ is set equal to $S_n(y_{j-1})(1 - s_j/r_j)$, with $y_0 = 0$. For t in between event times, $S_n(t) = S_n(y_j)$, where y_j is the latest event time no later than t . The Kaplan Meier product limit formula is

Kaplan-Meier Product Limit Estimator

$$S_n(t) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right), \quad y_{j-1} \leq t < y_j \tag{22.1}$$

r_i is called the *risk set* at time y_i . It is the set of all individuals subject to the risk being studied at the event time. If entries or withdrawals occur at the same time as a death—for example, if 2 lives enter at time 5, 3 lives leave, and 1 life dies—the lives that leave *are* in the risk set, while the lives that enter *are not*.

EXAMPLE 22A In a mortality study, 10 lives are under observation. One death apiece occurs at times 3, 4, and 7, and two deaths occur at time 11. One withdrawal apiece occurs at times 5 and 10. The study concludes at time 12.

Calculate the product limit estimate of the survival function.

ANSWER: We calculate the survival function $S_{10}(t)$ for $0 \leq t \leq 12$ recursively in the following table.

j	Time y_j	Risk Set r_j	Deaths s_j	Survival Function $S_{10}(t)$ for $y_j \leq t < y_{j+1}$
1	3	10	1	$(10 - 1)/10 = 0.9000$
2	4	9	1	$S_{10}(4^-) \times (9 - 1)/9 = 0.8000$
3	7	7	1	$S_{10}(7^-) \times (7 - 1)/7 = 0.6857$
4	11	5	2	$S_{10}(11^-) \times (5 - 2)/5 = 0.4114$

$S_{10}(t) = 1$ for $t < 3$. In the above table, y_5 should be construed to equal 12. □

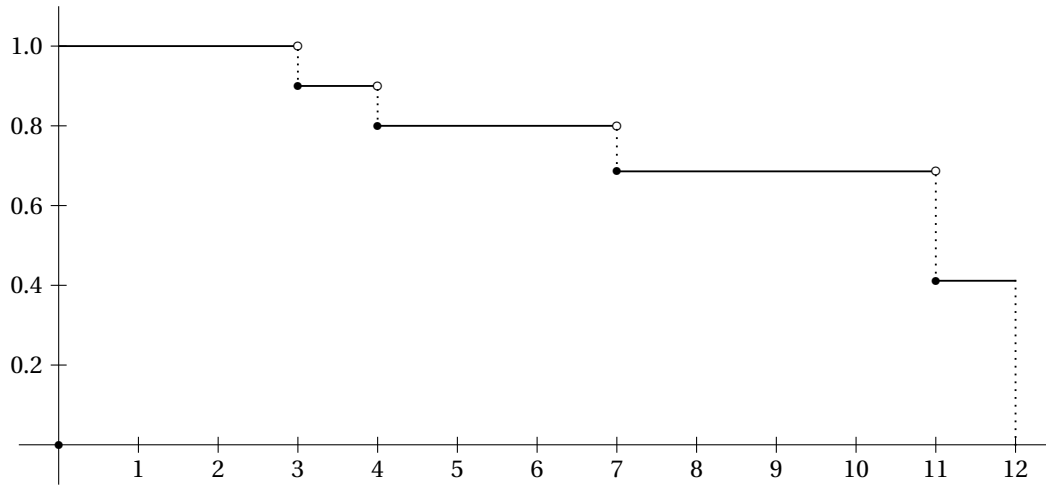


Figure 22.2: Graph of $y = S_{10}(x)$ computed in Example 22A

We plot the survival function of Example 22A in Figure 22.2. Note that the estimated survival function is *constant* between event times, and for this purpose, only the event we are interested in—death—counts, not withdrawals. This means, for example, that whereas $S_{10}(7) = 0.6857$, $S_{10}(6.999) = 0.8000$, the same as $S_{10}(4)$. The function is discontinuous. By definition, if X is the survival time random variable, $S(x) = \Pr(X > x)$. This means that if you want to calculate $\Pr(X \geq x)$, this is $S(x^-)$, which may not be the same as $S(x)$.

EXAMPLE 22B Assume that you are given the same data as in Example 22A. Using the product limit estimator, estimate:

1. the probability of a death occurring at any time greater than 3 and less than 7.
2. the probability of a death occurring at any time greater than or equal to 3 and less than or equal to 7.

ANSWER: 1. This is $\Pr(3 < X < 7) = \Pr(X > 3) - \Pr(X \geq 7) = S(3) - S(7^-) = 0.9 - 0.8 = \boxed{0.1}$.

2. This is $\Pr(3 \leq X \leq 7) = \Pr(X \geq 3) - \Pr(X > 7) = S(3^-) - S(7) = 1 - 0.6857 = \boxed{0.3143}$. □

Example 22A had withdrawals but did not have new entries. New entries are treated as part of the risk set after they enter. The next example illustrates this, and also illustrates another notation system used in the textbook. In this notation system, each individual is listed separately. d_i indicates the entry time, u_i indicates the withdrawal time, and x_i indicates the death time. Only one of u_i and x_i is listed.

EXAMPLE 22C You are given the following data from a mortality study:

i	d_i	x_i	u_i
1	0	—	7
2	0	5	—
3	2	—	8
4	5	7	—

Estimate the survival function using the product-limit estimator.

ANSWER: There are two event times, 5 and 7. At time 5, the risk set includes individuals 1, 2, and 3, but not individual 4. New entries tied with the event time do not count. So $S_4(5) = 2/3$. At time 7, the risk set includes

individuals 1, 3, and 4, since withdrawals tied with the event time do count. So $S_4(7) = (2/3)(2/3) = 4/9$. The following table summarizes the results:

j	y_j	r_j	s_j	$S_4(t)$ for $y_j \leq t \leq y_{j+1}$
1	5	3	1	2/3
2	7	3	1	4/9

In any time interval with no withdrawals or new entries, if you are not interested in the survival function within the interval, you may merge all event times into one event time. The risk set for this event time is the number of individuals at the start of the interval, and the number of deaths is the total number of deaths in the interval. For example, in Example 22A, to calculate $S_{10}(4)$, rather than multiplying two factors for times 3 and 4, you could group the deaths at 3 and 4 together, treat the risk set at time 4 as 10 and the number of deaths as 2, and calculate $S_{10}(4) = 8/10$.

These principles apply equally well to estimating severity with incomplete data.

EXAMPLE 22D An insurance company sells two types of auto comprehensive coverage. Coverage A has no deductible and a maximum covered loss of 1000. Coverage B has a deductible of 500 and a maximum covered loss of 10,000. The company experiences the following loss sizes:

Coverage A: 300, 500, 700, and three claims above 1000

Coverage B: 700, 900, 1200, 1300, 1400

Let X be the loss size.

Calculate the Kaplan-Meier estimate of the probability that a loss will be greater than 1200 but less than 1400, $\Pr(1200 < X < 1400)$.

ANSWER: We treat the loss sizes as if they're times! And the "members" of Coverage B enter at "time" 500. The inability to observe a loss below 500 for Coverage B is analogous to a mortality study in which members enter the study at time 500. The loss sizes above 1000 for Coverage A are treated as withdrawals; they are censored observations, since we know those losses are greater than 1000 but don't know exactly what they are.

The Kaplan-Meier table is shown in Table 22.1. We will explain below how we filled it in.

Table 22.1: Survival function calculation for Example 22D

j	Loss Size y_j	Risk Set r_j	Losses s_j	Survival Function $S_{11}(t)$ for $y_j \leq t < y_{j+1}$
1	300	6	1	5/6
2	500	5	1	2/3
3	700	9	2	14/27
4	900	7	1	4/9
5	1200	3	1	8/27
6	1300	2	1	4/27
7	1400	1	1	0

At 300, only coverage A claims are in the risk set; coverage B claims are truncated from below. Thus, the risk set at 300 is 6. Similarly, the risk set at 500 is 5; remember, new entrants are not counted at the time they enter, only after the time, so even though the deductible is 500, coverage B losses do not count at 500. So we have that $S_{11}(500) = \left(\frac{5}{6}\right)\left(\frac{4}{5}\right) = \frac{2}{3}$.

At 700, 4 claims from coverage A (the one for 700 and the 3 censored ones) and all 5 claims from coverage B are in the risk set, making the risk set 9. Similarly, at 900, the risk set is 7. So $S_{11}(900) = \left(\frac{2}{3}\right)\left(\frac{7}{9}\right)\left(\frac{6}{7}\right) = \frac{4}{9}$.

At 1200, only the 3 claims 1200 and above on coverage B are in the risk set. So $S_{11}(1200) = \left(\frac{4}{9}\right)\left(\frac{2}{3}\right) = \frac{8}{27}$. Similarly, $S_{11}(1300) = \left(\frac{8}{27}\right)\left(\frac{1}{2}\right) = \frac{4}{27}$.

The answer to the question is $\Pr_{11}(X > 1200) - \Pr_{11}(X \geq 1400) = S_{11}(1200) - S_{11}(1400^-)$. $S_{11}(1200) = \frac{8}{27}$. But $S_{11}(1400^-)$ is not the same as $S_n(1400)$. In fact, $S_{11}(1400^-) = S_{11}(1300) = \frac{4}{27}$. The final answer is then $\Pr_{11}(1200 < X < 1400) = \frac{8}{27} - \frac{4}{27} = \boxed{\frac{4}{27}}$. \square

If all lives remaining in the study die at the last event time of the study, then S can be estimated as 0 past this time. It is less clear what to do if the last observation is censored. The two extreme possibilities are

1. to treat it as if it were a death, so that $S(t) = 0$ for $t \geq y_k$, where y_k is the last observation time of the study.
2. to treat it as if it lives forever, so that $S(t) = S(y_k)$ for $t \geq y_k$.

A third option is to use an exponential whose value is equal to $S(y_k)$ at time y_k .

EXAMPLE 22E In example 22A, you are to use the Kaplan-Meier estimator, with an exponential to extrapolate past the end of the study.

Determine $S_{10}(15)$.

ANSWER: $S_{10}(12) = S_{10}(11) = 0.4114$, as determined above. We extend exponentially from the end of the study at time 12. In other words, we want $e^{-12/\theta} = 0.4114$, or $\theta = -\frac{12}{\ln 0.4114}$. Then $S_{10}(15) = \exp\left(\frac{15 \ln 0.4114}{12}\right) = 0.4114^{15/12} =$

$\boxed{0.3295}$. \square

Notice in the above example that using an exponential to go from year 12 to year 15 is equivalent to raising the year 12 value to the 15/12 power. In general, if u is the ending time of the study, then exponential extrapolation sets $S_n(t) = S_n(u)^{t/u}$ for $t > u$.

If a study has no members before a certain time—in other words, the study starts out with 0 individuals and the first new entries are at time y_0 —then the estimated survival function is conditional on the estimated variable being greater than y_0 . There is simply no estimate for values less than y_0 . For example, if Example 22D is changed so that Coverage A has a deductible of 250, then the estimates are for $S_{11}(x | X > 250)$, and $\Pr_{11}(1200 < X < 1400 | X > 250) = 4/27$. It is not possible to estimate the unconditional survival function in this case.

Note that the letter k is used to indicate the number of unique event times. There is a released exam question in which they expected you to know that that is the meaning of k .



Quiz 22-1 You are given the following information regarding six individuals in a study:

d_j	u_j	x_j
0	5	—
0	4	—
0	—	3
1	3	—
2	—	4
3	5	—

Calculate the Kaplan-Meier product-limit estimate of $S(4.5)$.

Now we will discuss another estimator for survival time.

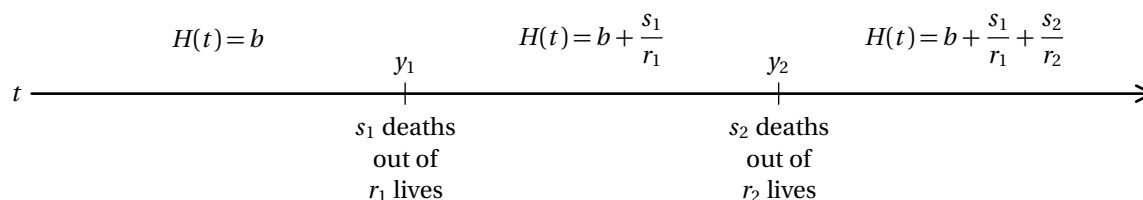


Figure 22.3: Illustration of the Nelson-Åalen estimator of cumulative hazard function. The cumulative hazard function is initially b . After each event time, it is incremented by the proportion of deaths in the group.

22.2 Nelson-Åalen Estimator

The Nelson-Åalen estimator estimates the cumulative hazard function. The idea is simple. Suppose the cumulative hazard rate before time y_1 is known to be b . If at that time s_1 lives out of a risk set of r_1 die, that means that the hazard at that time y_1 is s_1/r_1 . Therefore the *cumulative* hazard function is increased by that amount, s_j/r_j , and becomes $b + s_1/r_1$. See Figure 22.3. The Nelson-Åalen estimator sets $\hat{H}(0) = 0$ and then at each time y_j at which an event occurs, $\hat{H}(y_j) = \hat{H}(y_{j-1}) + s_j/r_j$. The formula is:

Nelson-Åalen Estimator

$$\hat{H}(t) = \sum_{i=1}^{j-1} \frac{s_i}{r_i}, \quad y_{j-1} \leq t < y_j \quad (22.2)$$

EXAMPLE 22F In a mortality study on 98 lives, you are given that

- (i) 1 death occurs at time 5
- (ii) 2 lives withdraw at time 5
- (iii) 3 lives enter the study at time 5
- (iv) 1 death occurs at time 8

Calculate the Nelson-Åalen estimate of $H(8)$.

ANSWER: The table of risk sets and deaths is

j	Time y_j	Risk Set r_j	Deaths s_j	NA estimate $\hat{H}(y_j)$
1	5	98	1	$\frac{1}{98}$
2	8	98	1	$\frac{1}{98} + \frac{1}{98}$

At time 5, the original 98 lives count, but we don't remove the 2 withdrawals or count the 3 new entrants. At time 8, we have the original 98 lives minus 2 withdrawals minus 1 death at time 5 plus 3 new entrants, or $98 - 2 - 1 + 3 = 98$ in the risk set.

$$\hat{H}(8) = \frac{1}{98} + \frac{1}{98} = \boxed{\frac{1}{49}}$$

□

To estimate the survival function using Nelson-Åalen, exponentiate the Nelson-Åalen estimate; $\hat{S}(x) = e^{-\hat{H}(x)}$. In the above example, the estimate would be $\hat{S}(8) = e^{-1/49} = 0.9798$. This will always be higher than the Kaplan-Meier estimate, except when $\hat{H}(x) = 0$ (and then both estimates of S will be 1). In the above example, the Kaplan-Meier estimate would be $(\frac{97}{98})^2 = 0.9797$.

Table 22.2: Summary of Formulas in this Lesson

Kaplan-Meier Product Limit Estimator	$\hat{S}(t) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right), \quad y_{j-1} \leq t < y_j$	(22.1)
Nelson-Åalen Estimator	$\hat{H}(t) = \sum_{i=1}^{j-1} \frac{s_i}{r_i}, \quad y_{j-1} \leq t < y_j$	(22.2)
Exponential extrapolation	$\hat{S}(t) = \hat{S}(t_0)^{t/t_0} \quad t \geq t_0$	

Everything we said about extrapolating past the last time, or conditioning when there are no observations before a certain time, applies equally well to $\hat{S}(t)$ estimated using Nelson-Åalen.



Quiz 22-2 In a mortality study on 10 lives, 2 individuals die at time 4 and 1 individual at time 6. The others survive to time 10.

Using the Nelson-Åalen estimator, estimate the probability of survival to time 10.

Exercises

22.1. [160-F86:2] The results of using the product-limit (Kaplan-Meier) estimator of $S(x)$ for a certain data set are:

$$\hat{S}(x) = \begin{cases} 1.0, & 0 \leq x < a \\ \frac{49}{50}, & a \leq x < b \\ \frac{1,911}{2,000}, & b \leq x < c \\ \frac{36,309}{40,000}, & c \leq x < d \end{cases}$$

Determine the Nelson-Åalen estimate of $S(c)$.

- (A) $e^{-23/250}$ (B) $e^{-93/1000}$ (C) $e^{-19/200}$ (D) $e^{-97/1000}$ (E) $e^{-1/10}$

22.2. [160-S88:15] You are given the following for a complete data study:

- No simultaneous deaths occur.
- One third of the original entrants are surviving after k deaths at time y_k .
- The Nelson-Åalen estimate of $H(y_k) = 0.95$.

Determine k .

- (A) 2 (B) 4 (C) 6 (D) 8 (E) 10

22.3. [160-S90:14] You are given the following regarding a 2 year mortality study:

- (i) Ten lives enter the study at the beginning.
- (ii) One additional life enters at each of the following times: 0.8, 1.0.
- (iii) One life terminates at time 1.5.
- (iv) One death occurs at each of the following times: 0.2, 0.5, 1.3, 1.7

Calculate the product limit estimate of $S(2)$.

22.4. [160-F90:16] You are given the following regarding a 1 year mortality study:

- (i) 25 lives entered the study at the beginning.
- (ii) n lives entered at time 0.4.
- (iii) There were no withdrawals.

(iv)	Age	Number
	At Death	Of Deaths
	0.25	4
	0.50	2
	0.75	3
	1.00	4

- (v) The product limit estimate of $S(1)$ was 0.604.

Determine n .

- (A) 8 (B) 11 (C) 15 (D) 19 (E) 25

22.5. [160-83-94:11] For a complete data study, you are given:

- (i) There is only one death at each death point.
- (ii) $H(x)$ is estimated by the Nelson-Åalen method.
- (iii) $\hat{H}(y_7) = 0.3726$, where y_7 denotes the time at which the seventh death occurs.

Calculate the product limit estimate of $S(y_7)$.

- (A) 0.66 (B) 0.67 (C) 0.68 (D) 0.69 (E) 0.70

22.6. [160-83-97:9] You are given that:

- (i) 100 people enter a mortality study at time 0.
- (ii) At time 6, 15 people leave.
- (iii) 10 deaths occur before time 6.
- (iv) 3 deaths occur between time 6 and time 10.

Calculate the product limit estimate of $S(10)$.

22.7. [160-F87:14] You are given the following data from a clinical study:

Time	Event
0.0	20 new entrants
1.1	1 death
1.5	9 terminations
2.3	1 death
3.0	1 new entrant
3.2	1 death
4.7	1 termination
6.0	2 deaths

Calculate the absolute difference between the product limit estimate of $S(6)$ and the Nelson-Åalen estimate of $S(6)$.

- (A) 0.01 (B) 0.03 (C) 0.05 (D) 0.08 (E) 0.11

22.8. [160-F87:18] In a mortality study with no censored or truncated data, the Nelson-Åalen estimator of the cumulative hazard function is calculated. There are no ties for death times. You obtain:

$$\hat{H}(y_{10}) = 0.669 \text{ and}$$

$$\hat{H}(y_{11}) = 0.769.$$

Calculate $\hat{H}(y_2)$.

- (A) 0.103 (B) 0.108 (C) 0.113 (D) 0.118 (E) 0.123

22.9. [160-S87:14] In a mortality study, the following observations are made:

- (i) x persons die, 1 withdraws and 1 enters at time $t = 1$.
- (ii) y persons die and 1 enters at $t = 2$.
- (iii) 1 person dies at $t = 3$.

Based on these observations, three values of $\hat{H}(t)$, the Nelson-Åalen estimate of the cumulative hazard function at time t are:

$$\hat{H}(1.5) = 0.20$$

$$\hat{H}(2.5) = 0.45$$

$$\hat{H}(3.5) = 0.55$$

Determine $x + y$.

- (A) 3 (B) 4 (C) 5 (D) 6 (E) 7

22.10. [160-S87:15] You are given the following data from a mortality study:

Individual	Time At Entry	Time At Termination
1	0	—
2	0	2 (censored)
3	0	3 (death)
4	2	5 (death)
5	4	—

Calculate the Nelson-Åalen estimate of the cumulative hazard function, $\hat{H}(5)$.

22.11. [160-F89:13] In a mortality study on n individuals, you are given:

- (i) The first 2 deaths occur at times y_1 and y_2 .
- (ii) The product limit estimate of $S(y_2)$ is not zero.
- (iii) The sum of the product limit estimate of $S(y_2)$ and the Nelson-Åalen estimate of $H(y_2) = 17/16$.
- (iv) All withdrawals occur within (y_1, y_2) .

Determine the number of withdrawals.

- (A) 2 (B) 3 (C) 4 (D) 5 (E) 6

22.12. [160-S90:12] A mortality study involves a group of n individuals. One individual apiece dies at times y_1 and y_2 . No withdrawals occur before time y_2 .

You calculate the Nelson-Åalen estimator of the cumulative hazard function at time y_2 , $\hat{H}(y_2) = 0.1144$.

Determine the product limit estimate of $S(y_2)$.

- (A) 0.86 (B) 0.87 (C) 0.88 (D) 0.89 (E) 0.90

22.13. [160-S91:17] 16 individuals are observed in a mortality study. No withdrawals occur before time 12. The product limit estimator of $S(12)$ is 0.9375.

Calculate the Nelson-Åalen estimate of $S(12)$.

- (A) 0.9337 (B) 0.9356 (C) 0.9375 (D) 0.9394 (E) 0.9413

22.14. [160-81-96:11] In a mortality study, n individuals are observed. No withdrawals occur. 2 deaths occur at time y_1 and 1 death occurs at time y_2 . The Nelson-Åalen estimate of $H(y_2)$ is 1.0.

Calculate the product limit estimate of $S(y_2)$.

- (A) 0.25 (B) 0.33 (C) 0.37 (D) 0.40 (E) 0.50

22.15. [160-82-96:10] You are given the following product limit estimates from a mortality study:

Time (y_t)	10	12	15
No. of deaths	1	2	1
$S_n(y_t)$	0.72	0.60	0.50

There were no other deaths, and no new entrants, at any time between 10 and 15.

Calculate the number of withdrawals occurring in the time interval $[12, 15)$.

- (A) 0 (B) 1 (C) 2 (D) 3 (E) 4

22.16. In a mortality study:

- (i) At time 130, there are two deaths.
- (ii) The product limit estimate of $S(130)$ is 0.8247.
- (iii) After time 128 but before time 130, 5 lives leave and no lives die.
- (iv) At time 128, there are 247 lives, of which one died.

Determine the product limit estimate of $S(128)$.

22.17. Auto liability insurance is offered with three policy limits: 25,000, 50,000, and 100,000. You experience the following payments on these coverages:

25,000 limit:	5,000, 10,000, 20,000, 25,000, 25,000
50,000 limit:	10,000, 30,000, 50,000, 50,000, 50,000
100,000 limit:	5,000, 10,000, 25,000, 50,000, 100,000

You are given:

- (i) When the policy limit was paid, the actual loss was greater than the policy limit.
- (ii) The underlying loss distribution does not vary with policy limit.

Using the Product Limit Estimator, determine the probability that a loss will be greater than 60,000 before any policy limit is applied.

22.18. For 10 policies, the length of time from receipt of policy application to policy issue is as follows:

15 15 17 20 21 25 25 27 31 35

For 5 additional policies, the applications were withdrawn on days 12, 16, 18, 20, and 20 without the policy being issued.

Let X be the length of time from application to policy issue.

Using the product limit estimator, estimate $\Pr(17 \leq X \leq 24)$.

22.19. On an automobile liability coverage, you experience the following losses:

50,000 15,000 20,000 80,000 30,000 12,000

Let X be claim size.

Estimate $\Pr(30,000 \leq X \leq 50,000)$ using the Nelson-Åalen estimator.

22.20. You are combining data from two auto collision coverages. Coverage A has a deductible of 500 and a limit of 10,000, and coverage B has a deductible of 1000 and a limit of 100,000. Reported claim sizes (including the deductible but capped by the limit) are as follows:

Coverage A: 600, 2000, 5000, 10,000, 10,000, 10,000

Coverage B: 2000, 6000, 20,000, 100,000, 100,000

Estimate the probability of a loss greater than 30,000 given that the loss is greater than 500, using the Nelson-Åalen estimator.

22.21. You are studying the length of time from hiring an agent to regular termination. Regular termination means termination for causes other than death or disability. For a group of 100 agents, you have the following data:

Year	Regular Termination	Termination due to Death or Disability
1	38	1
2	16	2
3	10	2
4	8	3

The study ended at the end of the fourth year.

All terminations in the above study occurred at the end of each year.

Use the Kaplan-Meier estimator, extending it past the study's end with an exponential curve.

Estimate the probability of not terminating with a regular termination before the end of the sixth year.

22.22. [4-S00:4] For a mortality study with right-censored data, you are given:

Time y_i	Number of Deaths s_i	Number at Risk r_i
5	2	15
7	1	12
10	1	10
12	2	6

Calculate $\hat{S}(12)$ based on the Nelson-Åalen estimate for $\hat{H}(12)$.

- (A) 0.48 (B) 0.52 (C) 0.60 (D) 0.65 (E) 0.67

22.23. [C4 Sample:2] The number of employees leaving a company for all reasons is tallied by the number of months since hire. The following data was collected for a group of 50 employees hired one year ago:

Number of Months Since Hire	Number Leaving the Company
1	1
2	1
3	2
5	2
7	1
10	1
12	1

Determine the Nelson-Åalen estimate of the cumulative hazard at the sixth month since hire.

Note: Assume that employees always leave the company after a whole number of months.

22.24. [C4 Sample:22] An insurance company wishes to estimate its four-year agent retention rate using data on all agents hired during the last six years. You are given:

- Using the Product-Limit estimator, the company estimates the proportion of agents remaining after 3.75 years of service as $\hat{S}(3.75) = 0.25$.
- One agent resigned between 3.75 and 4 years of service.
- Eleven agents have been employed longer than the agent who resigned between 3.75 and 4 years of service.
- Two agents have been employed for six years.

Determine the Product-Limit estimate of $S(4)$.

22.25. [4-F00:4] You are studying the length of time attorneys are involved in settling bodily injury lawsuits. T represents the number of months from the time an attorney is assigned such a case to the time the case is settled.

Nine cases were observed during the study period, two of which were not settled at the conclusion of the study. For those two cases, the time spent up to the conclusion of the study, 4 months and 6 months, was recorded instead. The observed values of T for the other seven cases are as follows:

1 3 3 5 8 8 9

Estimate $\Pr(3 \leq T \leq 5)$ using the Product-Limit estimator.

- (A) 0.13 (B) 0.22 (C) 0.36 (D) 0.40 (E) 0.44

22.26. [4-S01:4] You are given the following times of first claim for five randomly selected auto insurance policies observed from time $t = 0$:

1 2 3 4 5

You are later told that one of the five states given is actually the time of policy lapse, but you are not told which one.

The smallest Product-Limit estimate of $S(4)$, the probability that the first claim occurs after time 4, would result if which of the given times arose from the lapsed policy?

- (A) 1 (B) 2 (C) 3 (D) 4 (E) 5

22.27. [4-F01:19] For a mortality study of insurance applicants in two countries, you are given:

(i)

	Country A		Country B	
y_i	s_i	r_i	s_i	r_i
1	20	200	15	100
2	54	180	20	85
3	14	126	20	65
4	22	112	10	45

- (ii) r_i is the number at risk over the period (y_{i-1}, y_i) . Deaths during the period (y_{i-1}, y_i) are assumed to occur at y_i .
- (iii) $S^T(t)$ is the Product-Limit estimate of $S(t)$ based on the data for all study participants.
- (iv) $S^B(t)$ is the Product-Limit estimate of $S(t)$ based on the data for study participants in Country B.

Determine $|S^T(4) - S^B(4)|$.

- (A) 0.06 (B) 0.07 (C) 0.08 (D) 0.09 (E) 0.10

22.28. [4-F02:4] In a study of claim payment times, you are given:

- (i) The data were not truncated or censored.
- (ii) At most one claim was paid at any one time.
- (iii) The Nelson-Åalen estimate of the cumulative hazard function, $H(t)$, immediately following the second paid claim, was $23/132$.

Determine the Nelson-Åalen estimate of the cumulative hazard function, $H(t)$, immediately following the fourth paid claim.

- (A) 0.35 (B) 0.37 (C) 0.39 (D) 0.41 (E) 0.43

22.29. [4-F02:25] The claim payments on a sample of ten policies are:

2 3 3 5 5⁺ 6 7 7⁺ 9 10⁺
 + indicates that the loss exceeded the policy limit

Using the Product-Limit estimator, calculate the probability that the loss on a policy exceeds 8.

- (A) 0.20 (B) 0.25 (C) 0.30 (D) 0.36 (E) 0.40

22.30. [4-F03:40] You are given the following about 100 insurance policies in a study of time to policy surrender:

- (i) The study was designed in such a way that for every policy that was surrendered, a new policy was added, meaning that the risk set, r_j , is always equal to 100.
- (ii) Policies are surrendered only at the end of a policy year.
- (iii) The number of policies surrendered at the end of each policy year was observed to be:
 - 1 at the end of the 1st policy year
 - 2 at the end of the 2nd policy year
 - 3 at the end of the 3rd policy year
 - ⋮
 - n at the end of the n^{th} policy year
- (iv) The Nelson-Åalen empirical estimate of the cumulative distribution function at time n , $\hat{F}(n)$, is 0.542.

What is the value of n ?

- (A) 8 (B) 9 (C) 10 (D) 11 (E) 12

22.31. [4-F04:4] For observation i of a survival study:

- d_i is the left truncation point
- x_i is the observed value if not right censored
- u_i is the observed value if right censored

You are given:

Observation (i)	d_i	x_i	u_i
1	0	0.9	—
2	0	—	1.2
3	0	1.5	—
4	0	—	1.5
5	0	—	1.6
6	0	1.7	—
7	0	—	1.7
8	1.3	2.1	—
9	1.5	2.1	—
10	1.6	—	2.3

Determine the Kaplan-Meier Product-Limit estimate, $S_{10}(1.6)$.

- (A) Less than 0.55
- (B) At least 0.55, but less than 0.60
- (C) At least 0.60, but less than 0.65
- (D) At least 0.65, but less than 0.70
- (E) At least 0.70

22.32. [C-S05:3] You are given:

- (i) A mortality study covers n lives.
- (ii) None were censored and no two deaths occurred at the same time.
- (iii) $t_k =$ time of the k^{th} death.
- (iv) A Nelson-Åalen estimate of the cumulative hazard rate function is $\hat{H}(t_2) = \frac{39}{380}$.

Determine the Kaplan-Meier product-limit estimate of the survival function at time t_3 .

- (A) Less than 0.56
- (B) At least 0.56, but less than 0.58
- (C) At least 0.58, but less than 0.60
- (D) At least 0.60, but less than 0.62
- (E) At least 0.62

Additional released exam questions: C-F06:14,20,31, C-S07:38

Solutions

22.1. We can back out $1 - \frac{s_j}{r_j}$ at each point, since $S(y_j) = S(y_{j-1}) \left(1 - \frac{s_j}{r_j}\right)$. Numbering the three times corresponding to a , b , and c as 1, 2, and 3 respectively, we have:

$$\begin{aligned} \frac{49}{50} &= 1 - \frac{s_1}{r_1} \Rightarrow \frac{s_1}{r_1} = \frac{1}{50} \\ \frac{\frac{1911}{2000} - \frac{49}{50}}{\frac{49}{50}} &= \frac{39}{40} = 1 - \frac{s_2}{r_2} \Rightarrow \frac{s_2}{r_2} = \frac{1}{40} \\ \frac{\frac{36,309}{40,000} - \frac{1911}{2000}}{\frac{1911}{2000}} &= \frac{19}{20} = 1 - \frac{s_3}{r_3} \Rightarrow \frac{s_3}{r_3} = \frac{1}{20} \end{aligned}$$

By equation (22.2),

$$\begin{aligned} \hat{H}(c) &= \frac{1}{50} + \frac{1}{40} + \frac{1}{20} = \frac{20 + 25 + 50}{1000} = \frac{95}{1000} = \frac{19}{200} \\ \hat{S}(c) &= e^{-19/200} \quad (\text{C}) \end{aligned}$$

22.2. The only way I can see to do this is trial and error. Trying $n = 3$ and $k = 2$ deaths, we get $\hat{H}(y_2) = \frac{1}{3} + \frac{1}{2} \neq 0.95$. For $n = 6$ and $k = 4$ deaths, we get $\hat{H}(y_4) = \frac{1}{6} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} = 0.95$. So the answer is **4**, (B).

22.3. Setting up the usual table:

y_j	r_j	s_j	$S_{10}(y_j)$
0.2	10	1	9/10
0.5	9	1	8/10
1.3	10	1	72/100
1.7	8	1	63/100

So the answer is $63/100 = \mathbf{0.63}$.

22.4. We can use the shortcut of grouping all deaths together for times above 0.4, since there were no entries or withdrawals afterwards. The first risk set is 25; the risk set after time 0.4 is $25 - 4 + n = 21 + n$. So:

$$\begin{aligned}\frac{21}{25} \frac{12+n}{21+n} &= 0.604 \\ 1 - \frac{9}{21+n} &= 0.604 \left(\frac{25}{21} \right) = 0.7190 \\ \frac{9}{21+n} &= 0.2810 \\ n &= \frac{9}{0.2810} - 21 = \boxed{11} \quad (\text{B})\end{aligned}$$

22.5. We are given that $\sum_{j=0}^6 \frac{1}{n-j} = 0.3726$. To help determine n , we estimate that the middle term of the sum is approximately equal to the average; in other words $\frac{1}{n-3} \approx \frac{0.3726}{7}$ or $n \approx 22$. In fact, plugging 22 in for n in the sum works. So $n = 22$ and $\hat{S}(t_7) = \frac{15}{22}$ (the product limit estimate is the empirical estimate since it is a complete data study) = $\boxed{0.68}$. (C)

22.6. The risk set for the first 10 deaths is 100. The risk set for the second 3 deaths is $100 - 15 - 10 = 75$. So $S_{100}(10) = \binom{90}{100} \binom{72}{75} = \boxed{0.864}$.

22.7. The r_j 's and s_j 's are:

y_j	1.1	2.3	3.2	6.0
r_j	20	10	10	8
s_i	1	1	1	2

$$S_n(6) = \left(\frac{19}{20} \right) \left(\frac{9}{10} \right) \left(\frac{9}{10} \right) \left(\frac{3}{4} \right) = 0.577125$$

$$\hat{H}(6) = \frac{1}{20} + \frac{1}{10} + \frac{1}{10} + \frac{1}{4} = 0.5$$

$$\hat{S}(6) = e^{-0.5} = 0.606531$$

$$0.606531 - 0.577125 = \boxed{0.03} \quad (\text{B})$$

22.8. Since there are no ties for death times and no censored or truncated data, the Nelson-Åalen estimator reduces to

$$\hat{H}(y_t) = \sum_{i=1}^t \frac{1}{n-t+1}$$

where n is the original study population and is equal to r_1 . This means that

$$\hat{H}(y_t) - \hat{H}(y_{t-1}) = \frac{1}{n-t+1}.$$

We use this to back out r_{11} , which is $n - 10$, and then to calculate $r_1 = n$ and $r_2 = n - 1$.

$$\begin{aligned}0.1 &= \hat{H}(y_{11}) - \hat{H}(y_{10}) = \frac{1}{r_{11}} \\ r_{11} &= 10 \\ r_2 &= 10 + (11 - 2) = 19 \quad \text{and } r_1 = 20 \\ \hat{H}(y_2) &= \frac{1}{20} + \frac{1}{19} = \boxed{0.103} \quad (\text{A})\end{aligned}$$

22.9. Since $\hat{H}(1.5) = \hat{H}(1) = \frac{s_1}{r_1}$, we have

$$0.20 = \frac{s_1}{r_1} = \frac{x}{r_1} \quad (*)$$

and since $\hat{H}(2.5) = \hat{H}(2) = \hat{H}(1) + \frac{s_2}{r_2}$, we have

$$0.45 - 0.20 = 0.25 = \frac{s_2}{r_2} = \frac{y}{r_2}, \text{ and } r_2 = r_1 - x \quad (**)$$

and since $\hat{H}(3.5) = \hat{H}(3) = \hat{H}(2) + \frac{s_3}{r_3}$, we have

$$0.55 - 0.45 = 0.10 = \frac{1}{r_2 - y + 1} \quad (***)$$

Using equation (***),

$$\begin{aligned} r_2 - y + 1 &= 10 \\ r_2 &= 9 + y \end{aligned}$$

and plugging into equation (**),

$$\begin{aligned} \frac{y}{9 + y} &= 0.25 \\ 2.25 + 0.25y &= y \\ y &= 3, r_2 = 12 \\ r_1 &= 12 + x \end{aligned}$$

Now, using equation (*),

$$\begin{aligned} \frac{x}{12 + x} &= 0.20 \\ x &= 2.4 + 0.2x \\ x &= 3 \\ x + y &= \boxed{6} \quad (\mathbf{D}) \end{aligned}$$

22.10. There are two event times, 3 and 5.

At time 3, the risk set consists of individuals 1, 3, and 4. 2 left earlier, and 5 has not entered yet.

At time 5, the risk set consists of individuals 1, 4, and 5. 2 left earlier, and 3 died earlier.

Accordingly, we have the following table.

y_j	r_j	s_j
3	3	1
5	3	1

Using the Nelson-Åalen formula:

$$\hat{H}(5) = \frac{1}{3} + \frac{1}{3} = \boxed{\frac{2}{3}}.$$

22.11. Let r_j be the risk set at time y_j .

$$\begin{aligned} \left(\frac{r_1-1}{r_1}\right)\left(\frac{r_2-1}{r_2}\right) + \left(\frac{1}{r_1} + \frac{1}{r_2}\right) &= \frac{17}{16} \\ \frac{1}{r_1 r_2}(r_1 r_2 - r_2 - r_1 + 1 + r_1 + r_2) &= \frac{17}{16} \\ \frac{r_1 r_2 + 1}{r_1 r_2} &= \frac{17}{16} \\ r_1 r_2 &= 16 \end{aligned}$$

$r_2 < r_1$ and $r_2 \neq 1$ (since $S_n(y_2) \neq 0$), so the only possible factorization of 16 into $r_1 r_2$ is $r_1 = 8$, $r_2 = 2$. There were $8 - 2 - 1 = \boxed{5}$ withdrawals. **(D)**

22.12. We must calculate $n \cdot \frac{1}{n} + \frac{1}{n-1} = 0.1144$. $\frac{1}{n}$ is about 0.0572 (half of 0.1144), so n is about 18. Experimenting, $\frac{1}{18} + \frac{1}{17} = 0.1144$, so $n = 18$. $S_n(y_2) = \frac{17}{18} \frac{16}{17} = \boxed{0.89}$. **(D)**

22.13. Since no withdrawals occur, the deaths can be grouped. If s is the number of deaths before 12, $0.9375 = \hat{S}(12) = \frac{16-s}{16}$, so $s = 1$. Switching to Nelson-Åalen, $\hat{H}(12) = \frac{1}{16}$; exponentiating to get the estimate of the survival function, $\hat{S}(12) = e^{-1/16} = \boxed{0.9394}$. **(D)**

22.14. We set up the Nelson-Åalen formula:

$$\frac{2}{n} + \frac{1}{n-2} = 1$$

In reality, since n has to be an integer, it is probably fastest to use trial and error; n must be at least 3 (otherwise $n - 2 \leq 0$), and by trying the values 3 and 4 you quickly see that $n = 4$. If trial and error doesn't appeal to you, you can solve the quadratic:

$$\begin{aligned} 2n - 4 + n &= n(n - 2) \\ n^2 - 5n + 4 &= 0 \\ n &= 4 \end{aligned}$$

The risk sets are then 4 (for the first 2 deaths) and 2 (for the final death). Then:

$$S_n(y_2) = \binom{2}{4} \binom{1}{2} = \frac{1}{4} = \boxed{0.25} \quad \text{(A)}$$

22.15. Since $S_n(y_t) = S_n(y_{t-1})(r_t - s_t)/r_t$, we have

$$\frac{S_n(y_t)}{S_n(y_{t-1})} = \frac{r_t - s_t}{r_t}.$$

We use this equation at times 12 and 15:

$$\begin{aligned} \frac{0.60}{0.72} &= \frac{r_{12} - 2}{r_{12}}, \quad \text{so } r_{12} = 12 \\ \frac{0.50}{0.60} &= \frac{r_{15} - 1}{r_{15}}, \quad \text{so } r_{15} = 6 \end{aligned}$$

There were $12 - 6 - 2 = \boxed{4}$ withdrawals. **(E)**

22.16. There were $247 - 1 - 5 = 241$ lives at time 130, so

$$\begin{aligned}\hat{S}(130) &= \hat{S}(128) \left(\frac{239}{241} \right) \\ 0.8247 &= \hat{S}(128) \left(\frac{239}{241} \right) \\ \hat{S}(128) &= \frac{241(0.8247)}{239} = \boxed{0.8316}\end{aligned}$$

22.17. The claims above the policy limits are censored observations. Policies with a 25,000 limit are not in the risk set for claims higher than 25,000. We have

y_j	r_j	s_j
5,000	15	2
10,000	13	3
20,000	10	1
25,000	9	1
30,000	6	1
50,000	5	1

$$\begin{aligned}S_n(25,000) &= \frac{8}{15} \\ S_n(30,000) &= \frac{8}{15} \left(\frac{5}{6} \right) = \frac{4}{9} \\ S_n(60,000) &= \hat{S}_n(50,000) = \frac{4}{9} \left(\frac{4}{5} \right) = \boxed{\frac{16}{45}}\end{aligned}$$

22.18.

$$\begin{aligned}\hat{S}(15) &= \frac{12}{14} = \frac{6}{7} = 0.8571 \\ \hat{S}(17) &= \left(\frac{6}{7} \right) \left(\frac{10}{11} \right) = \frac{60}{77} \\ \hat{S}(20) &= \left(\frac{60}{77} \right) \left(\frac{8}{9} \right) = 0.6926 \\ \hat{S}(21) &= 0.6926 \left(\frac{5}{6} \right) = 0.5772 \\ \Pr(17 \leq X \leq 24) &= 0.8571 - 0.5772 = \boxed{0.2799}\end{aligned}$$

22.19. We need $\hat{S}(30,000^-) - \hat{S}(50,000) = \hat{S}(20,000) - \hat{S}(50,000)$. There are 3 claims less than or equal to 20,000.

$$\begin{aligned}\hat{H}(20,000) &= \frac{1}{6} + \frac{1}{5} + \frac{1}{4} = 0.61667 & \hat{S}(20,000) &= e^{-0.61667} = 0.53974 \\ \hat{H}(50,000) &= \frac{1}{6} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2} = 1.45 & \hat{S}(50,000) &= e^{-1.45} = 0.2346 \\ 0.53974 - 0.23457 &= \boxed{0.3052}\end{aligned}$$

22.20. The conditioning on the loss being greater than 500 is automatic, since we have no data for losses below 500.

y_j	r_j	s_j	$\hat{H}(y_j)$
600	6	1	0.1667
2000	10	2	0.3667
5000	8	1	0.4917
6000	7	1	0.6345
20000	3	1	0.9679

$$\hat{S}(30,000) = e^{-0.9679} = \boxed{0.3799}.$$

22.21.

$$S_{100}(4) = \left(\frac{62}{100}\right) \left(\frac{45}{61}\right) \left(\frac{33}{43}\right) \left(\frac{23}{31}\right) = 0.2604.$$

Using an exponential to go from the fourth to the sixth year is equivalent to raising the fourth year value to the 6/4 power. So $S_{100}(6) = 0.2604^{6/4} = \boxed{0.1329}$.

22.22. The Nelson-Åalen estimate of $\hat{H}(12)$ is

$$\hat{H}(12) = \frac{2}{15} + \frac{1}{12} + \frac{1}{10} + \frac{2}{6} = 0.65$$

Then $\hat{S}(12) = e^{-0.65} = \boxed{0.5220}$. (B)

22.23. Since there is no censoring, we have

y_i	r_i	s_i	$\hat{H}(y_i)$
1	50	1	$1/50 = 0.02$
2	49	1	$0.02 + 1/49 = 0.04041$
3	48	2	$0.04041 + 2/48 = 0.08207$
5	46	2	$0.08207 + 2/46 = \boxed{0.12555}$

22.24. To go from time 3.75 to time 4, since only one agent resigned in between, we multiply $\hat{S}(3.75)$ by $\frac{r_i - s_i}{r_i}$, where $s_i = 1$ for the one agent who resigned and r_i is the risk set at the time that agent resigned. Since 11 agents were employed longer, the risk set is $r_i = 11 + 1 = 12$ (counting the agent who resigned and the 11 who were employed longer). If we let y_i be the time of resignation, since nothing happens between y_i and 4,

$$\hat{S}(4) = \hat{S}(y_i) = 0.25 \left(\frac{11}{12}\right) = \boxed{0.2292}.$$

The fact 2 agents were employed for 6 years is extraneous.

22.25. The product-limit estimator up to time 5, taking the 2 censored observations at 4 and 6 into account, is:

y_i	r_i	s_i	$\hat{S}(y_i)$
1	9	1	8/9
3	8	2	6/9
5	5	1	$(6/9)(4/5) = 24/45$

$$\widehat{\Pr}(3 \leq T \leq 5) = \hat{S}(3^-) - \hat{S}(5) = \frac{8}{9} - \frac{24}{45} = \frac{16}{45} = \boxed{0.3556} \quad (\text{C})$$

22.26. You can calculate all five possibilities, but let's reason it out. If the lapse occurred at time 5, 4 claims occurred; otherwise, only 3 claims occurred, so one would expect the answer to be **5**, (E).

22.27. Since there is no censoring (in every case, $r_{i+1} = r_i - s_i$), the products telescope, and the product-limit estimator becomes the empirical estimator.

$$S^T(4) = \frac{(112 - 22) + (45 - 10)}{200 + 100} = \frac{125}{300} = 0.417$$

$$S^B(4) = \frac{45 - 10}{100} = 0.35$$

$$S^T(4) - S^B(4) = \mathbf{0.067} \quad (\text{B})$$

22.28. We have $\frac{1}{n} + \frac{1}{n-1} = \frac{23}{132}$ which is a quadratic, but since n must be an integer, it is easier to approximate the equation as

$$\frac{2}{n - 1/2} \approx \frac{23}{132}$$

$$n - \frac{1}{2} \approx \frac{264}{23} = 11.48$$

so $n = 12$. Then $\frac{23}{132} + \frac{1}{10} + \frac{1}{9} = \mathbf{0.3854}$. (C)

22.29. Through time 5 there is no censoring, so $\hat{S}(5) = \frac{6}{10}$ (6 survivors out of 10 original lives). Then $\hat{S}(7) = \left(\frac{6}{10}\right)\left(\frac{3}{5}\right)$ (three survivors from 5 lives past 5), so $\hat{S}(7) = 0.36$. There are no further claims between 7 and 8, so the answer is **0.36**. (D)

22.30.

$$\hat{H}(n) = -\ln(1 - \hat{F}(n)) = 0.78$$

$$\sum_{i=1}^n \frac{i}{100} = 0.78$$

$$\frac{n(n+1)}{2} = 78$$

This quadratic can be solved directly, or by trial and error; approximate the equation with $\frac{(n+0.5)^2}{2} = 78$ making $n + 0.5$ around 12.5, and we verify that **12** works. (E)

22.31. The x_i 's are the events. d_i 's are entry times into the study, and u_i 's are withdrawal, or censoring, times. Every member of the study is counted in the risk set for times in the interval $(d_i, u_i]$.

Before time 1.6, there are 2 event times, 0.9 and 1.5. (The other x_i 's are 1.7 and 2.1, which are past 1.6.)

At time 0.9, the risk set consists of all entrants before 0.9, namely $i = 1$ through 7, or 7 entries. There are no withdrawals or deaths before 0.9, so the risk set is 7.

At time 1.5, the risk set consists of all entrants before 1.5, or $i = 1$ through 8, minus deaths or withdrawals before time 1.5: the death at 0.9 and the withdrawal at 1.2, leaving 6 in the risk set. Note that entrants at time 1.5 are not counted in the risk set and withdrawals at time 1.5 are counted.

The standard table with y_j 's, r_j 's, and s_j 's looks like this:

y_j	r_j	s_j	$\hat{S}(y_j)$
0.9	7	1	6/7
1.5	6	1	5/7

The Kaplan-Meier estimate is then $\left(\frac{6}{7}\right)\left(\frac{5}{6}\right) = \frac{5}{7} = \mathbf{0.7143}$, or (E).

22.32. We must calculate n . Either you observe that the denominator 380 has divisors 19 and 20, or you estimate

$$\frac{2}{n - 0.5} \approx \frac{39}{380}$$

and you conclude that $n = 20$, which you verify by calculating $\frac{1}{20} + \frac{1}{19} = \frac{39}{380}$. The Kaplan-Meier estimate is the empirical complete data estimate since no one is censored; after 9 deaths, the survival function is $(20 - 9)/20 =$

0.55. (A)

Quiz Solutions

22-1. The risk set is 5 at time 3, since the entry at 3 doesn't count. The risk set is 4 at time 4, after removing the third and fourth individuals, who left at time 3. The estimate of $S(4.5)$ is $(4/5)(3/4) =$ **0.6**.

22-2. The risk sets are 10 at time 4 and 8 at time 6. Therefore

$$\hat{H}(10) = \frac{2}{10} + \frac{1}{8} = 0.325$$

$$\hat{S}(10) = e^{-0.325} =$$
 0.7225

