a/s/m
SOA Exam C
Study Manual

With ***Study****Plus*+

Study*Plus*+ gives you digital access* to:

- Flashcards & Formula Sheet
- Actuarial Exam & Career Strategy Guides
- Technical Skill eLearning Tools
- Samples of Supplemental Textbooks
- And more!

**See inside for keycode access and login instructions*

18th Edition, Third Printing

Abraham Weishaus, Ph.D., F.S.A., CFA, M.A.A.A.

NO RETURN IF OPENED

a/s/m
Actuarial Study Materials
Learning Made Easier

TO OUR READERS:

Please check A.S.M.'s web site at **www.studymanuals.com** for errata and updates. If you have any comments or reports of errata, please e-mail us at **mail@studymanuals.com**.

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 . . . 11
1.5 The empirical distribution . . . 13
Exercises . . . 14
Solutions . . . 21

2 Parametric Distributions **31**
2.1 Scaling . . . 31
2.2 Transformations . . . 33
2.3 Common parametric distributions . . . 35
2.3.1 Uniform . . . 36
2.3.2 Beta . . . 37
2.3.3 Exponential . . . 37
2.3.4 Weibull . . . 38
2.3.5 Gamma . . . 39
2.3.6 Pareto . . . 40
2.3.7 Single-parameter Pareto . . . 41
2.3.8 Lognormal . . . 42
2.4 The linear exponential family . . . 42
2.5 Limiting distributions . . . 45
Exercises . . . 47
Solutions . . . 50

3 Variance **53**
3.1 Additivity . . . 53
3.2 Normal approximation . . . 54
3.3 Bernoulli shortcut . . . 56
Exercises . . . 57
Solutions . . . 58

4 Mixtures and Splices **61**
4.1 Mixtures . . . 61
4.1.1 Discrete mixtures . . . 61
4.1.2 Continuous mixtures . . . 63
4.1.3 Frailty models . . . 64
4.2 Conditional Variance . . . 65
4.3 Splices . . . 67
Exercises . . . 71
Solutions . . . 78

5 Policy Limits **89**
Exercises . . . 91
Solutions . . . 95

6 Deductibles **99**
6.1 Ordinary and franchise deductibles . . . 99
6.2 Payment per loss with deductible . . . 99
6.3 Payment per payment with deductible . . . 101
Exercises . . . 105
Solutions . . . 116

7 Loss Elimination Ratio **125**
Exercises . . . 126
Solutions . . . 132

8 Risk Measures and Tail Weight **143**
8.1 Coherent risk measures . . . 143
8.2 Value-at-Risk (VaR) . . . 145
8.3 Tail-Value-at-Risk (TVaR) . . . 148
8.4 Tail Weight . . . 153
8.5 Extreme value distributions . . . 155
Exercises . . . 156
Solutions . . . 160

9 Other Topics in Severity Coverage Modifications **167**
Exercises . . . 171
Solutions . . . 177

10 Bonuses **189**
Exercises . . . 190
Solutions . . . 192

11 Discrete Distributions **197**
11.1 The $(a, b, 0)$ class . . . 197
11.2 The $(a, b, 1)$ class . . . 201
Exercises . . . 205
Solutions . . . 210

12 Poisson/Gamma **221**
Exercises . . . 222
Solutions . . . 227

13 Frequency— Exposure & Coverage Modifications **231**
13.1 Exposure modifications . . . 231
13.2 Coverage modifications . . . 231
Exercises . . . 233
Solutions . . . 239

14 Aggregate Loss Models: Compound Variance **245**
14.1 Introduction . . . 245
14.2 Compound variance . . . 246

Exercises . . . 249
Solutions . . . 259

15 Aggregate Loss Models: Approximating Distribution 271
Exercises . . . 274
Solutions . . . 283

16 Aggregate Losses: Severity Modifications 295
Exercises . . . 296
Solutions . . . 304

17 Aggregate Loss Models: The Recursive Formula 313
Exercises . . . 317
Solutions . . . 321

18 Aggregate Losses—Aggregate Deductible 327
Exercises . . . 332
Solutions . . . 339

19 Aggregate Losses: Miscellaneous Topics 347
19.1 Exact Calculation of Aggregate Loss Distribution . . . 347
19.1.1 Normal distribution . . . 347
19.1.2 Exponential and gamma distributions . . . 348
19.1.3 Compound Poisson models . . . 351
19.2 Discretizing . . . 351
19.2.1 Method of rounding . . . 352
19.2.2 Method of local moment matching . . . 352
Exercises . . . 354
Solutions . . . 356

20 Supplementary Questions: Severity, Frequency, and Aggregate Loss 361
Solutions . . . 365

II Empirical Models 373

21 Review of Mathematical Statistics 375
21.1 Estimator quality . . . 375
21.1.1 Bias . . . 376
21.1.2 Consistency . . . 378
21.1.3 Variance and mean square error . . . 378
21.2 Hypothesis testing . . . 379
21.3 Confidence intervals . . . 381
Exercises . . . 385
Solutions . . . 391

22 The Empirical Distribution for Complete Data 399
22.1 Individual data . . . 399
22.2 Grouped data . . . 400
Exercises . . . 401
Solutions . . . 404

23 Variance of Empirical Estimators with Complete Data **407**
23.1 Individual data . . . 407
23.2 Grouped data . . . 408
Exercises . . . 411
Solutions . . . 414

24 Kaplan-Meier and Nelson-Åalen Estimators **419**
24.1 Kaplan-Meier Product Limit Estimator . . . 420
24.2 Nelson-Åalen Estimator . . . 424
Exercises . . . 427
Solutions . . . 438

25 Estimation of Related Quantities **447**
25.1 Moments . . . 447
25.1.1 Complete individual data . . . 447
25.1.2 Grouped data . . . 447
25.1.3 Incomplete data . . . 450
25.2 Range probabilities . . . 453
25.3 Deductibles and limits . . . 453
25.4 Inflation . . . 454
Exercises . . . 455
Solutions . . . 460

26 Variance of Kaplan-Meier and Nelson-Åalen Estimators **467**
Exercises . . . 470
Solutions . . . 478

27 Kernel Smoothing **489**
27.1 Density and distribution . . . 489
27.1.1 Uniform kernel . . . 490
27.1.2 Triangular kernel . . . 495
27.1.3 Other symmetric kernels . . . 502
27.1.4 Kernels using two-parameter distributions . . . 503
27.2 Moments of kernel-smoothed distributions . . . 505
Exercises . . . 507
Solutions . . . 514

28 Mortality Table Construction **525**
28.1 Individual data based methods . . . 525
28.1.1 Variance of estimators . . . 529
28.2 Interval-based methods . . . 530
Exercises . . . 535
Solutions . . . 544

29 Supplementary Questions: Empirical Models **551**
Solutions . . . 554

III Parametric Models **559**

30 Method of Moments **561**

30.1 Introductory remarks . . . 561
30.2 The method of moments for various distributions . . . 562
30.2.1 Exponential . . . 562
30.2.2 Gamma . . . 562
30.2.3 Pareto . . . 563
30.2.4 Lognormal . . . 564
30.2.5 Uniform . . . 565
30.2.6 Other distributions . . . 565
30.3 Fitting other moments, and incomplete data . . . 566
Exercises . . . 569
Solutions . . . 578

31 Percentile Matching **591**
31.1 Smoothed empirical percentile . . . 591
31.2 Percentile matching for various distributions . . . 592
31.2.1 Exponential . . . 592
31.2.2 Weibull . . . 593
31.2.3 Lognormal . . . 594
31.2.4 Other distributions . . . 594
31.3 Percentile matching with incomplete data . . . 595
31.4 Matching a percentile and a moment . . . 597
Exercises . . . 597
Solutions . . . 606

32 Maximum Likelihood Estimators **617**
32.1 Defining the likelihood . . . 619
32.1.1 Individual data . . . 619
32.1.2 Grouped data . . . 620
32.1.3 Censoring . . . 621
32.1.4 Truncation . . . 622
32.1.5 Combination of censoring and truncation . . . 623
Exercises . . . 624
Solutions . . . 634

33 Maximum Likelihood Estimators—Special Techniques **645**
33.1 Cases for which the Maximum Likelihood Estimator equals the Method of Moments Estimator . . . 645
33.1.1 Exponential distribution . . . 645
33.2 Parametrization and Shifting . . . 646
33.2.1 Parametrization . . . 646
33.2.2 Shifting . . . 647
33.3 Transformations . . . 647
33.3.1 Lognormal distribution . . . 648
33.3.2 Inverse exponential distribution . . . 648
33.3.3 Weibull distribution . . . 649
33.4 Special distributions . . . 650
33.4.1 Uniform distribution . . . 650
33.4.2 Pareto distribution . . . 651
33.4.3 Beta distribution . . . 652
33.5 Bernoulli technique . . . 653

33.6 Estimating q_x . . . 656
Exercises . . . 658
Solutions . . . 675

34 Variance Of Maximum Likelihood Estimators 693
34.1 Information matrix . . . 693
34.1.1 Calculating variance using the information matrix . . . 693
34.1.2 Asymptotic variance of MLE for common distributions . . . 697
34.1.3 True information and observed information . . . 702
34.2 The delta method . . . 704
34.3 Confidence Intervals . . . 707
34.3.1 Normal Confidence Intervals . . . 707
34.3.2 Non-Normal Confidence Intervals . . . 708
34.4 Variance of Exact Exposure Estimate of $\hat{q}_j$ . . . 710
Exercises . . . 711
Solutions . . . 722

35 Fitting Discrete Distributions 735
35.1 Poisson distribution . . . 735
35.2 Negative binomial . . . 736
35.3 Binomial . . . 736
35.4 Fitting $(a, b, 1)$ class distributions . . . 738
35.5 Adjusting for exposure . . . 740
35.6 Choosing between distributions in the $(a, b, 0)$ class . . . 741
Exercises . . . 744
Solutions . . . 753

36 Hypothesis Tests: Graphic Comparison 763
36.1 $D(x)$ plots . . . 763
36.2 p–p plots . . . 764
Exercises . . . 766
Solutions . . . 770

37 Hypothesis Tests: Kolmogorov-Smirnov 775
37.1 Individual data . . . 775
37.2 Grouped data . . . 780
Exercises . . . 782
Solutions . . . 788

38 Hypothesis Tests: Chi-square 795
38.1 Introduction . . . 795
38.2 Definition of chi-square statistic . . . 798
38.3 Degrees of freedom . . . 801
38.4 Other requirements for the chi-square test . . . 803
38.5 Data from several periods . . . 805
Exercises . . . 807
Solutions . . . 824

39 Likelihood Ratio Test and Algorithm, Penalized Loglikelihood Tests 835
39.1 Likelihood Ratio Test and Algorithm . . . 835
39.2 Schwarz Bayesian Criterion and Akaike Information Criterion . . . 840

Exercises . . . 841
Solutions . . . 847

40 Supplementary Questions: Parametric Models 855
Solutions . . . 860

IV Credibility 867

41 Classical Credibility: Poisson Frequency 869
Exercises . . . 874
Solutions . . . 884

42 Classical Credibility: Non-Poisson Frequency 891
Exercises . . . 894
Solutions . . . 897

43 Classical Credibility: Partial Credibility 903
Exercises . . . 904
Solutions . . . 910

44 Bayesian Methods—Discrete Prior 915
Exercises . . . 919
Solutions . . . 934

45 Bayesian Methods—Continuous Prior 953
45.1 Calculating posterior and predictive distributions . . . 953
45.2 Recognizing the posterior distribution . . . 958
45.3 Loss functions . . . 959
45.4 Interval estimation . . . 960
45.5 The linear exponential family and conjugate priors . . . 961
Exercises . . . 961
Solutions . . . 969

46 Bayesian Credibility: Poisson/Gamma 985
Exercises . . . 986
Solutions . . . 995

47 Bayesian Credibility: Normal/Normal 999
Exercises . . . 1003
Solutions . . . 1004

48 Bayesian Credibility: Bernoulli/Beta 1009
48.1 Bernoulli/beta . . . 1009
48.2 Negative binomial/beta . . . 1012
Exercises . . . 1013
Solutions . . . 1017

49 Bayesian Credibility: Exponential/Inverse Gamma 1021
Exercises . . . 1025
Solutions . . . 1028

50 Bühlmann Credibility: Basics **1031**
Exercises . . . 1036
Solutions . . . 1043

51 Bühlmann Credibility: Discrete Prior **1051**
Exercises . . . 1056
Solutions . . . 1076

52 Bühlmann Credibility: Continuous Prior **1097**
Exercises . . . 1101
Solutions . . . 1113

53 Bühlmann-Straub Credibility **1127**
Exercises . . . 1129
Solutions . . . 1134

54 Exact Credibility **1141**
Exercises . . . 1143
Solutions . . . 1148

55 Bühlmann As Least Squares Estimate of Bayes **1153**
55.1 Regression . . . 1153
55.2 Graphic questions . . . 1155
55.3 Cov(X_i, X_j) . . . 1157
Exercises . . . 1158
Solutions . . . 1165

56 Empirical Bayes Non-Parametric Methods **1169**
56.1 Uniform exposures . . . 1170
56.2 Non-uniform exposures . . . 1172
Exercises . . . 1179
Solutions . . . 1186

57 Empirical Bayes Semi-Parametric Methods **1199**
57.1 Poisson model . . . 1199
57.2 Non-Poisson models . . . 1203
57.3 Which Bühlmann method should be used? . . . 1204
Exercises . . . 1206
Solutions . . . 1215

58 Supplementary Questions: Credibility **1223**
Solutions . . . 1228

V Simulation 1235

59 Simulation—Inversion Method **1237**
Exercises . . . 1242
Solutions . . . 1251

60 Simulation—Special Techniques **1261**

60.1 Mixtures . . . 1261
60.2 Multiple decrements . . . 1262
60.3 Simulating $(a, b, 0)$ distributions . . . 1265
60.4 Normal random variables: the polar method . . . 1267
Exercises . . . 1270
Solutions . . . 1276

61 Number of Data Values to Generate 1283
Exercises . . . 1288
Solutions . . . 1291

62 Simulation—Applications 1297
62.1 Actuarial applications . . . 1297
62.2 Statistical analysis . . . 1299
62.3 Risk measures . . . 1299
Exercises . . . 1301
Solutions . . . 1314

63 Bootstrap Approximation 1325
Exercises . . . 1330
Solutions . . . 1334

64 Supplementary Questions: Simulation 1339
Solutions . . . 1342

VI Practice Exams 1347

1 Practice Exam 1 1349

2 Practice Exam 2 1361

3 Practice Exam 3 1371

4 Practice Exam 4 1381

5 Practice Exam 5 1391

6 Practice Exam 6 1401

7 Practice Exam 7 1413

8 Practice Exam 8 1423

9 Practice Exam 9 1433

10 Practice Exam 10 1445

11 Practice Exam 11 1455

12 Practice Exam 12 1467

13 Practice Exam 13 1479

Appendices 1489

A Solutions to the Practice Exams **1491**
Solutions for Practice Exam 1 . . . 1491
Solutions for Practice Exam 2 . . . 1503
Solutions for Practice Exam 3 . . . 1516
Solutions for Practice Exam 4 . . . 1529
Solutions for Practice Exam 5 . . . 1542
Solutions for Practice Exam 6 . . . 1553
Solutions for Practice Exam 7 . . . 1565
Solutions for Practice Exam 8 . . . 1578
Solutions for Practice Exam 9 . . . 1591
Solutions for Practice Exam 10 . . . 1604
Solutions for Practice Exam 11 . . . 1616
Solutions for Practice Exam 12 . . . 1629
Solutions for Practice Exam 13 . . . 1643

B Solutions to Old Exams **1661**
B.1 Solutions to CAS Exam 3, Spring 2005 . . . 1661
B.2 Solutions to CAS Exam 3, Fall 2005 . . . 1665
B.3 Solutions to CAS Exam 3, Spring 2006 . . . 1669
B.4 Solutions to CAS Exam 3, Fall 2006 . . . 1673

C Cross Reference from *Loss Models* **1677**

D Exam Question Index **1679**

Lesson 24

Kaplan-Meier and Nelson-Åalen Estimators

Reading: *Loss Models* Fourth Edition 12.1

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 time 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]

[1]However, parametric estimators in the presence of right truncation or left censoring are not excluded from the syllabus. We will study parametric estimators in Lessons 30–33.

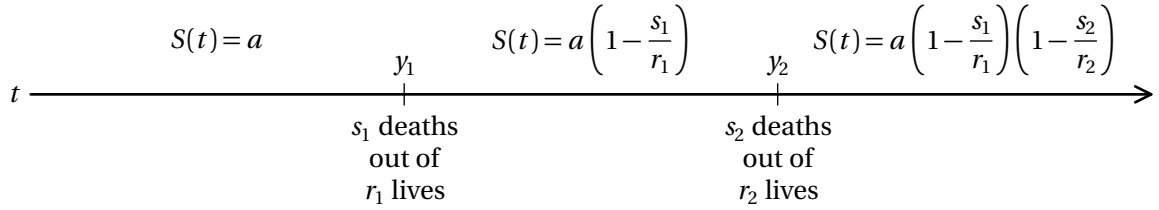

Figure 24.1: Illustration of the Kaplan-Meier product limit estimator. The survival function is initially a. After each event time, it is reduced in the same proportion as the proportion of deaths in the group.

24.1 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 24.1 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 24.1, 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

$$S_n(t) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right), \qquad y_{j-1} \leq t < y_j \tag{24.1}$$

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 24A 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: In this example, the event of interest is death. The event times are the times of death: 3, 4, 7, and 11. We label these events y_i. The number of deaths at the four event times are 1, 1, 1, and 2 respectively. We label these numbers s_i. That leaves us with calculating the risk set at each event time.

At time 3, there are 10 lives under observation. Therefore, the first risk set, the risk set for time 3, is $r_1 = 10$.

At time 4, there are 9 lives under observation. The life that died at time 3 doesn't count. Therefore, $r_2 = 9$.

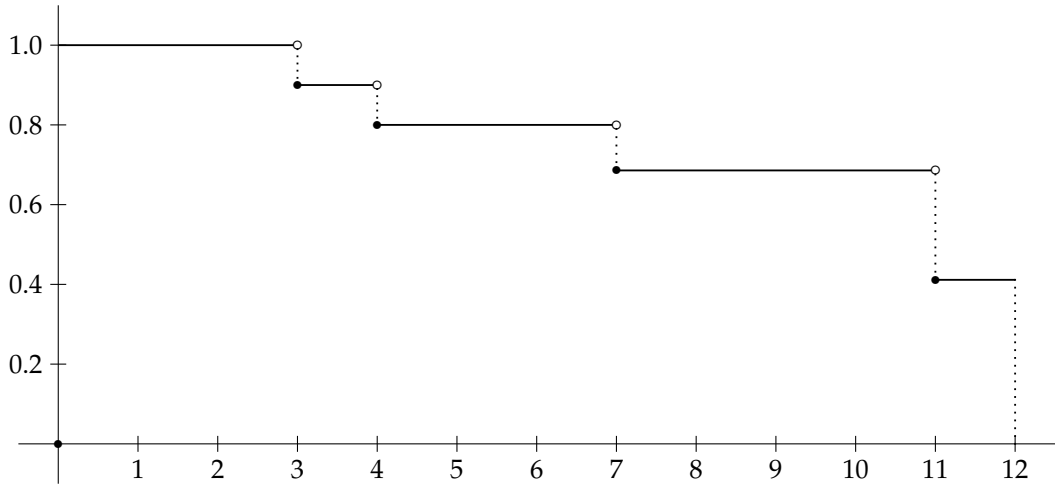

Figure 24.2: Graph of $y = S_{10}(x)$ computed in Example 24A

At time 7, there are 7 lives under observation. The lives that died at times 3 and 4, and the life that withdrew at time 5, don't count. Therefore, $r_3 = 7$.

At time 11, the lives that died at times 3, 4, and 7 aren't in the risk set. Nor are the lives that withdrew at times 5 and 10. That leave 5 lives in the risk set. $r_4 = 5$.

We now calculate the survival function $S_{10}(t)$ for $0 \le t \le 12$ recursively in the following table, using formula (24.1).

j	Time y_j	Risk Set r_j	Deaths s_j	Survival Function $S_{10}(t)$ for $y_j \le t < y_{j+1}$
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. □

We plot the survival function of Example 24A in Figure 24.2. 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 $\Pr(X \ge x)$, this is $S(x^-)$, which may not be the same as $S(x)$.

Example 24B Assume that you are given the same data as in Example 24A. 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 $\Pr(3 < X < 7) = \Pr(X > 3) - \Pr(X \ge 7) = S(3) - S(7^-) = 0.9 - 0.8 = \boxed{\mathbf{0.1}}$.

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

Example 24A 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 24C 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 $y_j \le t < y_{j+1}$
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 24A, 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 24D 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 24.1. We will explain below how we filled it in.

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

Table 24.1: Survival function calculation for Example 24D

j	Loss Size y_j	Risk Set r_j	Losses s_j	Survival Function $S_{11}(t)$ for $y_j \le t < y_{j+1}$
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

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 $\Pr_{11}(X > 1200) - \Pr_{11}(X \ge 1400) = S_{11}(1200) - S_{11}(1400^-)$. $S_{11}(1200) = \frac{8}{27}$. But $S_{11}(1400^-)$ *is not the same as* $S_{11}(1400)$. In fact, $S_{11}(1400^-) = S_{11}(1300) = \frac{4}{27}$, while $S_{11}(1400) = 0$. The final answer is then $\Pr_{11}(1200 < X < 1400) = \frac{8}{27} - \frac{4}{27} = \boxed{\frac{4}{27}}$. □

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 $t \ge y_k$, 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 $t \ge y_k$.

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

Example 24E In example 24A, 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 $S_{10}(15) = \exp\left(\frac{15 \ln 0.4114}{12}\right) = 0.4114^{15/12} = \boxed{\mathbf{0.3295}}$. □

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 24D is changed so that Coverage A has a deductible of 250, then the estimates are for $S_{11}(x \mid X > 250)$, and $\Pr_{11}(1200 < X < 1400 \mid X > 250) = 4/27$. It is not possible to estimate the unconditional survival function in this case.

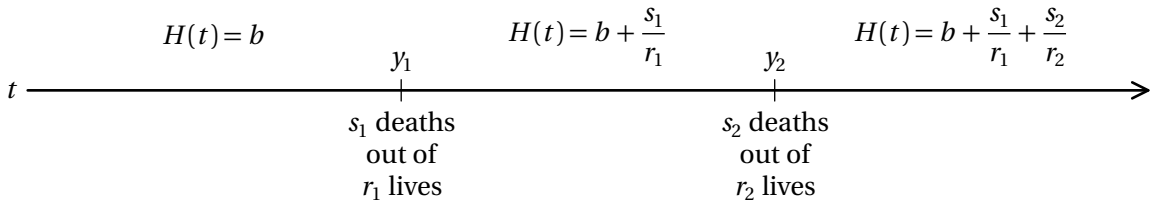

Figure 24.3: Illustration of the Nelson-Åalen estimator of cumulative hazard function. The cumulative hazard function is initially b. After each event time, it is incremented by the proportion of deaths in the group.

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 24-1 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.

24.2 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 24.3. 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

$$\hat{H}(t) = \sum_{i=1}^{j-1} \frac{s_i}{r_i}, \qquad y_{j-1} \le t < y_j \tag{24.2}$$

Example 24F 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 $H(8)$.

Table 24.2: Summary of Formulas in this Lesson

Kaplan-Meier Product Limit Estimator

$$\hat{S}(t) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right), \qquad y_{j-1} \le t < y_j \tag{24.1}$$

Nelson-Åalen Estimator

$$\hat{H}(t) = \sum_{i=1}^{j-1} \frac{s_i}{r_i}, \qquad y_{j-1} \le t < y_j \tag{24.2}$$

Exponential extrapolation

$$\hat{S}(t) = \hat{S}(t_0)^{t/t_0} \qquad t \ge t_0$$

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.

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

□

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 $\left(\frac{97}{98}\right)^2 = 0.9797$.

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 24-2 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.

Calculator Tip

Usually it is easy enough to calculate the Kaplan-Meier product limit estimator by directly multiplying $1 - s_j/r_j$. If you need to calculate several functions of s_j and r_j at once, such as both the Kaplan-Meier and the Nelson-Åalen estimator, it may be faster to enter s_j/r_j into a column of the TI-30XS/B Multiview's data table, and the function ln(1 − L1). The Kaplan-Meier estimator is a product, whereas the statistics registers only include sums, so it is necessary to log each factor, and then exponentiate the sum in the statistics register. Also, the sum is always of the entire column, so you must not have extraneous rows. If you need to calculate the estimator at two times, enter the rows needed for the earlier time, calculate the estimate, then add the additional rows for the second time.

Example 24G Seven times of death were observed:

5 6 6 8 10 12 15

In addition, there was one censored observation apiece at times 6, 7, and 11.

Calculate the absolute difference between the product-limit and Nelson-Åalen estimates of $S(10)$.

Answer: Only times up to 10 are relevant; the rest should be omitted. The r_i's and s_i's are

y_i	r_i	s_i
5	10	1
6	9	2
8	5	1
10	4	1

Here is the sequence of steps on the calculator:

Clear table	data data 4	L1 L2 L3 L1[1]=
Enter s_i/r_i in column 1	1 ÷ 10 ▼ 2 ÷ 9 ▼ 1 ÷ 5 ▼ 1 ÷ 4 enter	L1 L2 L3 0.2222 0.2 0.25 L1[5]=
Enter formula for Kaplan-Meier in column 2	▶ data ▶ 1 ln 1 - data 1) enter	L1 L2 L3 0.1 −0.105 0.2222 −0.251 0.2 −0.223 0.25 −0.288 L2[1]= −0.10536051..
Calculate statistics registers Clear display	2nd [stat]2 (Select L1 as first variable and L2 as second) ▼ ▼ enter clear clear	2-Var:L1,L2 1:n=4 2:x̄=0.1930555556 3↓Sx=0.065322089

Calculator Tip

Extract sum x (statistic 8) and sum y (statistic 10) from table	2nd [e^x] (−) 2nd [stat]38 ▶ = 2nd [e^x] 2nd [stat]3 (Press ▼ 9 times to get to A)	2-Var:L1,L2 8↑ ∑x=0.77222222 9:∑ x^2=0.1618827 A↓ ∑y=⁻0.86750056
Calculate difference of estimates	enter	$e^{-\sum x} - e^{\sum y}$ 0.041985293

The answer is **0.041985293**. Notice that the negative of the Nelson-Åalen estimator was exponentiated, but no negative sign is used for the sum of the logs of the factors of the product-limit estimator. □

Exercises

Kaplan-Meier

24.1. [**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 $S(2)$.

24.2. [**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 At Death	Number 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

Exercises continue on the next page . . .

24.3. [**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 $S(10)$.

24.4. You are given the following data from a mortality study on 10 lives:

d_i	x_i	u_i
0	—	8
0	—	12
5	11	—
8	—	23
11	21	—
12	21	—
17	—	23
18	—	28
21	33	—
21	—	24

Calculate the estimated discrete failure rate function at 21 using the Kaplan-Meier estimator.

24.5. In a mortality study starting with 50 lives:

(i) There are 2 new entrants at time 5 and 4 new entrants at time 10.
(ii) There are 3 withdrawals at time 5 and 1 withdrawal apiece at times 7, 9, 10, and 12.
(iii) One death apiece occurs at time 3, 5, 7, and 11.

Calculate the product-limit estimate of $H(11)$.

24.6. [**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

24.7. 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 $S(128)$.

Exercises continue on the next page . . .

24.8. 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 $\Pr(17 \le X \le 24)$.

24.9. [**1999 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 $S(4)$.

24.10. [**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 $\Pr(3 \le T \le 5)$ using the Product-Limit estimator.

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

24.11. [**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 times 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

Exercises continue on the next page . . .

24.12. [**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

24.13. [**4-F02:25**] The claim payments on a sample of ten policies are:

$$2 \quad 3 \quad 3 \quad 5 \quad 5^+ \quad 6 \quad 7 \quad 7^+ \quad 9 \quad 10^+$$

+ 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

Exercises continue on the next page . . .

24.14. **[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

24.15. In a mortality study on 10 lives, two lives die at times 6 and 9. One life leaves the study at time 7 and another life leaves the study at time 10. The remaining six lives remain in the study until time 12, at which time the study ends.

Estimate the probability of survival to time 20 using the Kaplan-Meier product limit estimator with an exponential tail correction.

24.16. 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 that a regular termination does not occur within the first six years.

Exercises continue on the next page . . .

24.17. [**C-S07:38**] You are given:

(i) All members of a mortality study are observed from birth. Some leave the study by means other than death.
(ii) $s_3 = 1$, $s_4 = 3$
(iii) The following Kaplan-Meier product-limit estimates were obtained:
$S_n(y_3) = 0.65$, $S_n(y_4) = 0.50$, $S_n(y_5) = 0.25$.
(iv) Between times y_4 and y_5, six observations were censored.
(v) Assume no observations were censored at the times of deaths.

Determine s_5.

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

Nelson-Åalen

24.18. [**160-F86:2**] The results of using the product-limit (Kaplan-Meier) estimator of $S(x)$ for a certain data set are:

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

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) $e^{-1/10}$

24.19. [**160-S88:15**] You are given the following for a complete data study:

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

Determine k.

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

24.20. [**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

Exercises continue on the next page . . .

24.21. [**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

24.22. [**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:

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

24.23. [**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

Exercises continue on the next page . . .

24.24. 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, $\hat{H}(5)$.

24.25. [**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

24.26. [**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

24.27. [**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

24.28. [**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

Use the following information for questions 24.29 and 24.30:

A bowling player has achieved the following scores on the last 10 games he played:

106 170 132 89 122 74 138 95 102 150

He is currently playing an eleventh game. You find it necessary to leave the game early. When you leave the game, he has scored 100 so far. You do not know how many frames are left for the game.

24.29. Using the Nelson-Åalen estimator, estimate the probability that his score for this game will be greater than 125.

Exercises continue on the next page . . .

24.29–30. *(Repeated for convenience) Use the following information for questions 24.29 and 24.30:*

A bowling player has achieved the following scores on the last 10 games he played:

106 170 132 89 122 74 138 95 102 150

He is currently playing an eleventh game. You find it necessary to leave the game early. When you leave the game, he has scored 100 so far. You do not know how many frames are left for the game.

24.30. Using the Nelson-Åalen estimator, estimate the probability that his score for a future game will be greater than 125.

24.31. **[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

24.32. **[1999 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.

24.33. **[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

Exercises continue on the next page . . .

24.34. [**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
$\vdots$
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

24.35. [**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_9.

(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

24.36. [**C-F06:14**, **C Sample Question #258**] For the data set

200 300 100 400 X

you are given:

(i) $k = 4$
(ii) $s_2 = 1$
(iii) $r_4 = 1$
(iv) The Nelson-Åalen Estimate $\hat{H}(410) > 2.15$

Determine X.

(A) 100 (B) 200 (C) 300 (D) 400 (E) 500

Exercises continue on the next page . . .

24.37. **[C-F06:20, C Sample Question #264]** You are given:

(i) The following data set:

2500 2500 2500 3617 3662 4517 5000 5000 6010 6932 7500 7500

(ii) $\hat{H}_1(7000)$ is the Nelson-Åalen estimate of the cumulative hazard rate function calculated under the assumption that all of the observations in (i) are uncensored.

(iii) $\hat{H}_2(7000)$ is the Nelson-Åalen estimate of the cumulative hazard rate function calculated under the assumption that all occurrences of the values 2500, 5000 and 7500 in (i) reflect right-censored observations and that the remaining observed values are uncensored.

Calculate $|\hat{H}_1(7000) - \hat{H}_2(7000)|$.

(A) Less than 0.1
(B) At least 0.1, but less than 0.3
(C) At least 0.3, but less than 0.5
(D) At least 0.5, but less than 0.7
(E) At least 0.7

24.38. **[C-F06:31, C Sample Question #274]** For a mortality study with right censored data, you are given the following:

Time	Number of Deaths	Number at Risk
3	1	50
5	3	49
6	5	k
10	7	21

You are also told that the Nelson-Åalen estimate of the survival function at time 10 is 0.575.

Determine k.

(A) 28 (B) 31 (C) 36 (D) 44 (E) 46

24.39. In a mortality study starting with 50 lives:

(i) There is 1 death apiece at times 5, 12, 17
(ii) There is 1 new entrant apiece at times 7, 12
(iii) There is 1 withdrawal apiece at times 13, 17
(iv) The study ends at time 20

Survival rates are estimated using the Nelson-Åalen estimator.

Estimate the probability of death before time 25 using exponential extrapolation.

Solutions

24.1. 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 $63/100 = \boxed{\mathbf{0.63}}$.

24.2. 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:

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

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

24.4. The discrete failure rate function is $1 - S(y_j)/S(y_{j-1})$, and with Kaplan-Meier $S(y_j)/S(y_{j-1}) = 1 - s_j/r_j$, so we just have to calculate the number of events and risk set at time 21. There are 2 events at time 21. The risk set is all entrants before time 21, or 8, minus 1 death at time 11 and 2 censored observations at times 8 and 12, or $8 - 1 - 2 = 5$. So $h(21) = 2/5 = \boxed{\mathbf{0.4}}$.

24.5. The risk set at time 3 is 50.

At time 5, the withdrawals count but not the new entry. The risk set is affected by one death at time 3, so it is 49.

At time 7, we consider the 2 new entrants at time 5 and the 3 withdrawals at time 5, so the risk set is $49 - 1 - 3 + 2 = 47$.

At time 11, we consider 3 more withdrawals (times 7, 9, 10) and 4 new entrants, so the risk set is $47 - 1 - 3 + 4 = 47$.

The table of risk sets is then:

y_j	r_j	s_j	$\hat{S}(y_j)$
3	50	1	0.98
5	49	1	0.96
7	47	1	0.939574
11	47	1	0.919584

Then $\hat{H}(11) = -\ln 0.919584 = \boxed{\mathbf{0.0838}}$.

24.6. 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:

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

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

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

$$\hat{S}(130) = \hat{S}(128)\left(\frac{239}{241}\right)$$
$$0.8247 = \hat{S}(128)\left(\frac{239}{241}\right)$$
$$\hat{S}(128) = \frac{241(0.8247)}{239} = \boxed{\mathbf{0.8316}}$$

24.8.

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

24.9. 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,

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

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

24.10. 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

$$\widehat{\Pr}(3 \le T \le 5) = \hat{S}(3^-) - \hat{S}(5) = \frac{8}{9} - \frac{24}{45} = \frac{16}{45} = \boxed{\mathbf{0.3556}} \qquad \textbf{(C)}$$

24.11. 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 $\boxed{\mathbf{5}}$, **(E)**.

24.12. 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.

$$S^T(4) = \frac{(112-22)+(45-10)}{200+100} = \frac{125}{300} = 0.417$$
$$S^B(4) = \frac{45-10}{100} = 0.35$$
$$S^T(4) - S^B(4) = \boxed{\mathbf{0.067}} \qquad \textbf{(B)}$$

24.13. 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 $\boxed{\mathbf{0.36}}$. **(D)**

24.14. 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 $\left(\frac{6}{7}\right)\left(\frac{5}{6}\right) = \boxed{\mathbf{\frac{5}{7} = 0.7143}}$, or **(E)**.

24.15. The risk set at time 6 is all 10 lives; the risk set at time 9 is 8 lives, since the lives that died at time 6 or left at time 7 aren't included.

y_i	s_i	r_i
6	1	10
9	1	8

The estimate of survival to the end of the study is

$$\hat{S}(12) = \left(\frac{9}{10}\right)\left(\frac{7}{8}\right) = 0.7875$$

Extrapolating to time 20,

$$\hat{S}(20) = \hat{S}(12)^{20/12} = 0.7875^{20/12} = \boxed{\mathbf{0.67156}}$$

24.16.

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

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 $S_{100}(6) = 0.2604^{6/4} = \boxed{\mathbf{0.1329}}$.

24.17. s_3 is extraneous. From $S_n(y_3)$ and $S_n(y_4)$, we have

$$\begin{aligned} 0.50 &= 0.65\frac{r_4 - s_4}{r_4} \\ \frac{10}{13} &= \frac{r_4 - 3}{r_4} \\ r_4 &= 13 \end{aligned}$$

Then $r_5 = r_4 - s_4 - 6 = 13 - 3 - 6 = 4$. From $S_n(y_5)$, we have

$$\begin{aligned} 0.25 = 0.50\frac{r_5 - s_5}{r_5} &= \frac{4 - s_5}{4} \\ 1 - \frac{s_5}{4} &= 0.5 \\ s_5 &= \boxed{\mathbf{2}} \qquad \textbf{(B)} \end{aligned}$$

24.18. 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:

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

By equation (24.2),

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

24.19. 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 $\boxed{\mathbf{4}}$, **(B)**.

24.20. 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) $= \boxed{\mathbf{0.68}}$. **(C)**

24.21. 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

$$S_n(6) = \left(\frac{19}{20}\right)\left(\frac{9}{10}\right)\left(\frac{9}{10}\right)\left(\frac{3}{4}\right) = 0.577125$$
$$\hat{H}(6) = \frac{1}{20} + \frac{1}{10} + \frac{1}{10} + \frac{1}{4} = 0.5$$
$$\hat{S}(6) = e^{-0.5} = 0.606531$$
$$0.606531 - 0.577125 = \boxed{\textbf{0.03}} \qquad \textbf{(B)}$$

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

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

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 $r_2 = n - 1$.

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

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

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

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

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

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

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

Using equation (***),

$$r_2 - y + 1 = 10$$

$$r_2 = 9 + y$$

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 (*),

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

24.24. 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:

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

24.25. 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 $8 - 2 - 1 = \boxed{\mathbf{5}}$ withdrawals. **(D)**

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

24.27. 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, $\hat{S}(12) = e^{-1/16} = \boxed{\mathbf{0.9394}}$ **(D)**

24.28. 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 $n-2 \le 0$), 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:

$$S_n(y_2) = \left(\frac{2}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{4} = \boxed{\mathbf{0.25}} \qquad \textbf{(A)}$$

24.29. The sorted data is: 74, 89, 95, 102, 106, 122, 132, 138, 150, 170. We want $\Pr(X > 125 \mid X > 100) = S(125)/S(100)$. Since Nelson-Åalen is a cumulative sum with $\hat{H}(125) = \hat{H}(100) + \sum_{100 < y_i \le 125} s_i/r_i$, we only need to sum up s_i/r_i between 100 and 125. The risk set at 102 is 7; at 106 it's 6; and at 122 it's 5. So

$$\hat{H}(125) - \hat{H}(100) = \frac{1}{7} + \frac{1}{6} + \frac{1}{5} = 0.509524$$

and $\widehat{\Pr}(X > 125 \mid X > 100) = e^{-0.509524} = \boxed{\mathbf{0.6008}}$.

24.30. We now have 10 observations plus the censored observation of 100, so we calculate the cumulative hazard rate using risk sets of 11 at 74, 10 at 89, and 9 at 95. The risk sets at 102, 106, and 122 are the same as in the previous exercise, so we'll add the sum computed there, 0.509524, to the sum of the quotients from the lowest three observations.

$$\hat{H}(125) = \frac{1}{11} + \frac{1}{10} + \frac{1}{9} + 0.509524 = 0.811544$$

and $\widehat{\Pr}(X > 125) = e^{-0.811544} = \boxed{\mathbf{0.4442}}$.

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

24.32. 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	$0.08207 + 2/46 =$ **0.12555**

24.33. 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 $\frac{23}{132} + \frac{1}{10} + \frac{1}{9} = \boxed{\mathbf{0.3854}}$. **(C)**

24.34.

$$\hat{H}(n) = -\ln\left(1 - \hat{F}(n)\right) = 0.78$$
$$\sum_{i=1}^{n} \frac{i}{100} = 0.78$$
$$\frac{n(n+1)}{2} = 78$$

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 $\boxed{\mathbf{12}}$ works. **(E)**

24.35. 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 $(20 - 9)/20 = \boxed{\mathbf{0.55}}$. **(A)**

24.36. k is the number of distinct observation points, so X must be one of the other 4 values, eliminating (E).

You want to make $\hat{H}(410)$ as high as possible by (iv), so you want to make the risk set as small as possible. Thus 100 offers the best opportunity, and works. **(A)** They had to state (ii) or else 200 would also work. I'm not sure why (iii) is needed.

24.37. The difference between the two estimates is that the first one will have, in the sum, terms for 2500 and 5000, while the second one will not. Those terms are $\frac{3}{12}$ (at 2500, risk set is 12 and 3 events) and $\frac{2}{6}$ (at 5000, risk set is 6 and 2 events). The sum is $\frac{1}{4} + \frac{1}{3} = \boxed{\mathbf{\frac{7}{12} = 0.58333}}$. **(D)**

24.38. The Nelson-Åalen estimate of $H(10)$ is $-\ln 0.575 = 0.5534$. Then

$$\frac{1}{50} + \frac{3}{49} + \frac{7}{21} = 0.4146$$
$$0.4146 + \frac{5}{k} = 0.5534$$
$$\frac{5}{k} = 0.5534 - 0.4146 = 0.1388$$
$$k = \frac{5}{0.1388} = \boxed{\mathbf{36}} \qquad \textbf{(C)}$$

24.39. The risk sets may be calculated recursively. At time 5, $r_1 = 50$. At time 12, $r_2 = 50 - 1 + 1 = 50$, where the new entrant at time 12 is tied with death at that time and therefore doesn't count. From time 12 to time 17, we subtract the death at time 12, add a new entrant at time 12, and subtract one withdrawal at time 13, leading to $r_3 = 50 - 1 + 1 - 1 = 49$, where the withdrawal at time 17 is tied with death at that time and therefore doesn't count.

$$\hat{H}(20) = \frac{1}{50} + \frac{1}{50} + \frac{1}{49} = 0.060408$$
$$\hat{S}(20) = e^{-\hat{H}(20)} = e^{-0.060408} = 0.941380$$

Extrapolating to time 25,

$$1 - \hat{S}(25) = 1 - 0.941380^{25/20} = 1 - 0.927270 = \boxed{\mathbf{0.072730}}$$

Note that exponentially extrapolating the survival function is equivalent to linearly extrapolating the cumulative hazard function, or increasing it prorata. In this case, $\hat{H}(25) = (25/20)\hat{H}(20)$.

Quiz Solutions

24-1. 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 $(4/5)(3/4) = \boxed{\mathbf{0.6}}$.

24-2. The risk sets are 10 at time 4 and 8 at time 6. Therefore

$$\hat{H}(10) = \frac{2}{10} + \frac{1}{8} = 0.325$$
$$\hat{S}(10) = e^{-0.325} = \boxed{\mathbf{0.7225}}$$

Practice Exam 2

1. Losses for an insurance coverage have the following cumulative distribution function:

$F(0) = 0$
$F(1{,}000) = 0.2$
$F(5{,}000) = 0.4$
$F(10{,}000) = 0.9$
$F(100{,}000) = 1$

with linear interpolation between these values.

Calculate the hazard rate at 9,000, $h(9{,}000)$.

(A) 0.0001 (B) 0.0004 (C) 0.0005 (D) 0.0007 (E) 0.0010

2. You are given the following data on loss sizes:

Loss Amount	Number of Losses
0– 1000	5
1000– 5000	4
5000–10000	3

An ogive is used as a model for loss sizes.

Determine the fitted median.

(A) 2000 (B) 2200 (C) 2500 (D) 3000 (E) 3083

3. In a mortality study on 5 individuals, death times were originally thought to be 1, 2, 3, 4, 5. It then turned out that one of these five observations was a censored observation rather than an actual death.

Let t_i be the time of the censored observation.

Determine the value of t_i for which variance of the Nelson-Åalen estimator of $H(4)$ is minimized.

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

4. For an insurance coverage, the number of claims per year follows a Poisson distribution. Claim size follows a Pareto distribution with $\alpha = 3$. Claim counts and claim sizes are independent.

The methods of classical credibility are used to determine premiums. The standard for full credibility is that actual aggregate claims be within 5% of expected aggregate claims 95% of the time. Based on this standard, 10,000 exposure units are needed for full credibility, where an exposure unit is a year of experience for a single insured.

Determine the expected number of claims per year.

(A) Less than 0.45
(B) At least 0.45, but less than 0.50
(C) At least 0.50, but less than 0.55
(D) At least 0.55, but less than 0.60
(E) At least 0.60

5. Which of the following statements is true?

(A) If data grouped into 7 groups are fitted to an inverse Pareto, the chi-square test of goodness of fit will have 5 degrees of freedom.
(B) The Kolmogorov-Smirnov statistic may be used to test the fit of a discrete distribution.
(C) The critical values of the Kolmogorov-Smirnov statistic do not require adjustment for estimated parameters.
(D) The critical values of the Kolmogorov-Smirnov statistic do not vary with sample size.
(E) The critical values of the chi-square statistic do not vary with sample size.

6. The amount of travel time to work for an employee is denoted by T. Given μ, $T - \mu - 0.5$ follows a beta distribution with $\theta = 1$ and $a = b = 2$. The parameter μ varies by employee and is uniformly distributed on $[15, 17]$.

For a randomly selected employee, the employee's travel time to work on one day is 16.

Calculate the Bühlmann credibility prediction of travel time to work for this employee.

(A) 16.1 (B) 16.3 (C) 16.5 (D) 16.7 (E) 16.9

7. For two classes of insureds of equal size, A and B, claim counts and claim sizes have the following distribution:

Claim Count	Probability A	Probability B
0	0.4	0.3
1	0.3	0.3
2	0.2	0.2
3	0.1	0.2

Claim Size	Probability A	Probability B
100	0.5	0.6
200	0.5	0.4

Claim counts and claim sizes are independent.

For a randomly selected insured, aggregate losses are 200.

Calculate the variance of predictive aggregate losses for the next period for this insured.

(A) 25,589 (B) 25,899 (C) 25,918 (D) 26,155 (E) 26,174

8. A class takes an exam. Half the students are good and half the students are bad. For good students, grades are distributed according to the probability density function

$$f(x) = \frac{4}{100}\left(\frac{x}{100}\right)^3 \qquad 0 \le x \le 100$$

For bad students, grades are distributed according to the probability density function

$$f(x) = \left(\frac{2}{100}\right)\left(\frac{x}{100}\right) \qquad 0 \le x \le 100$$

The passing grade is 65.

Determine the average grade on this exam for a passing student.

(A) 84.8 (B) 84.9 (C) 85.0 (D) 85.1 (E) 85.2

Exam questions continue on the next page . . .

9. In a mortality study on five lives, death times were 6, 7, 9, 15, and 30. Using the empirical distribution, $S(10)$ is estimated as 0.4.

To approximate the mean square error of the estimate, the bootstrap method is used. Five bootstrap samples are:

1. 6, 7, 7, 9, 30
2. 9, 6, 30, 7, 9
3. 30, 6, 6, 15, 7
4. 6, 15, 7, 9, 9
5. 30, 9, 6, 15, 15

Calculate the bootstrap approximation of the mean square error of the estimate.

(A) 0.032 (B) 0.034 (C) 0.036 (D) 0.038 (E) 0.040

10. An insurance coverage covers two types of insureds, A and B. There are an equal number of insureds in each class. Claim sizes in each class follow a Pareto distribution. Claim counts and claim sizes for insureds in each class have the following distributions:

Claim counts

	A	B
0	0.9	0.8
1	0.1	0.2

Size of claims (Pareto parameters)

	A	B
α	3	3
θ	50	60

Within each class, claim size and claim counts are independent.

Calculate the Bühlmann credibility to assign to 2 years of data.

(A) 0.01 (B) 0.02 (C) 0.03 (D) 0.04 (E) 0.05

11. An auto collision coverage is sold with deductibles of 500 and 1000. You have the following information for total loss size (including the deductible) on 86 claims:

Deductible 1000

Loss size	Number of losses
1000–2000	20
Over 2000	10

Deductible 500

Loss size	Number of losses
500–1000	32
Over 1000	24

Ground up underlying losses for both deductibles are assumed to follow an exponential distribution with the same parameter. You estimate the parameter using maximum likelihood.

For policies with an ordinary deductible of 500, determine the fitted average total loss size (including the deductible) for losses on which non-zero claim payments are made.

(A) 671 (B) 707 (C) 935 (D) 1171 (E) 1207

Exam questions continue on the next page . . .

12. You are given the following data from a 2-year mortality study.

Year j	Entries n_j	Withdrawals w_j	Deaths d_j
1	1000	100	33
2	500	100	c

Withdrawals and new entries occur uniformly over each year..

The actuarial estimate of q_1, the conditional probability of death in the second year given survival in the first year, is 0.03.

Determine c.

(A) 26 (B) 32 (C) 35 (D) 38 (E) 41

13. The number of claims per year on an insurance coverage has a binomial distribution with parameters $m = 2$ and Q. Q varies by insured and is distributed according to the following density function:

$$f(q) = 42q(1-q)^5 \qquad 0 \le q \le 1$$

An insured submits 1 claim in 4 years.

Calculate the posterior probability that for this insured, Q is less than 0.25.

(A) 0.52 (B) 0.65 (C) 0.70 (D) 0.76 (E) 0.78

14. You simulate a random variable with probability density function

$$f(x) = \begin{cases} -2x & -1 \le x \le 0 \\ 0 & \text{otherwise} \end{cases}$$

using the inversion method.

You use the following random numbers from the uniform distribution on $[0, 1]$:

0.2 0.4 0.3 0.7

Calculate the mean of the simulated observations.

(A) -0.7634 (B) -0.6160 (C) -0.2000 (D) 0.6160 (E) 0.7634

15. You are given a sample of 5 claims:

$$2, 3, 4, x_1, x_2$$

with $x_2 > x_1$. This sample is fitted to a Pareto distribution using the method of moments. The resulting parameter estimates are $\hat{\alpha} = 47.71$, $\hat{\theta} = 373.71$.

Determine x_1.

(A) 6.0 (B) 6.6 (C) 7.0 (D) 7.6 (E) 8.0

Exam questions continue on the next page . . .

16. The number of claims per year on a policy follows a Poisson distribution with parameter Λ. Λ has a uniform distribution on $(0, 2)$.

An insured submits 5 claims in one year.

Calculate the Bühlmann credibility estimate of the number of claims for the following year.

(A) 1.6 (B) 1.8 (C) 2.0 (D) 2.5 (E) 3.5

17. You are given:

(i) Annual claim counts follow a Poisson distribution with mean λ.
(ii) λ varies by insured. The distribution over all insureds is normal with mean 0.6 and variance 0.04.
(iii) An insured is selected at random and claim counts over 3 years are simulated for this insured by first simulating λ and then simulating each year's claim counts.
(iv) All simulations are done using the inversion method.

Use the following random numbers from the uniform distribution on $[0, 1)$ in order to perform the simulations:

0.28 0.82 0.13 0.94

Determine the total number of simulated claims over three years.

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

18. A study is performed on the amount of time on unemployment. The records of 10 individuals are examined. 7 of the individuals are not on unemployment at the time of the study. The following is the number of weeks they were on unemployment:

5, 8, 10, 11, 17, 20, 26

Three individuals are still on unemployment at the time of the study. They have been unemployed for the following number of weeks:

5, 20, 26

Let T be the amount of time on unemployment.

Using the Kaplan-Meier estimator with exponential extrapolation past the last study time, estimate $\Pr(20 \le T \le 30)$.

(A) 0.17 (B) 0.21 (C) 0.28 (D) 0.32 (E) 0.43

19. Annual claim counts per risk are binomial with parameters $m = 2$ and Q. Q varies by risk uniformly on $(0.25, 0.75)$.

For a risk selected at random, determine the probability of no claims.

(A) 0.14 (B) 0.25 (C) 0.26 (D) 0.27 (E) 0.28

Exam questions continue on the next page . . .

20. The distribution of auto insurance policyholders by number of claims submitted in the last year is as follows:

Number of claims	Number of insureds
0	70
1	22
2	6
3	2
Total	100

The number of claims for each insured is assumed to follow a Poisson distribution.

Use semi-parametric empirical Bayes estimation methods, with unbiased estimators for the variance of the hypothetical mean and the expected value of the process variance, to calculate the expected number of claims in the next year for a policyholder with 2 claims in the last year.

(A) Less than 0.52
(B) At least 0.52, but less than 0.57
(C) At least 0.57, but less than 0.62
(D) At least 0.62, but less than 0.67
(E) At least 0.67

21. X is a random variable. Simulation is used to estimate $F_X(500)$. Fifty pseudorandom values are generated. Of these fifty values, twenty values are less than or equal to 500.

Estimate the number of pseudorandom values that need to be generated in order to have 95% confidence that the estimate of $F_X(500)$ is within 5% of the true value.

(A) Less than 1600
(B) At least 1600, but less than 1800
(C) At least 1800, but less than 2000
(D) At least 2000, but less than 2200
(E) At least 2200

22. For an insurance, the number of claims per year for each risk has a Poisson distribution with mean Λ. Λ varies by risk according to a gamma distribution with mean 0.5 and variance 1. Claim sizes follow a Weibull distribution with $\theta = 5$, $\tau = \frac{1}{2}$. Claim sizes are independent of each other and of claim counts.

Determine the variance of aggregate claims.

(A) 256.25 (B) 312.50 (C) 350.00 (D) 400.00 (E) 450.00

23. You are given the following claims data from an insurance coverage with policy limit 10,000:

1000, 2000, 2000, 2000, 4000, 5000, 5000

There are 3 claims for amounts over 10,000 which are censored at 10,000.

You fit this experience to an exponential distribution with parameter $\theta = 6{,}000$.

Calculate the Kolmogorov-Smirnov statistic for this fit.

(A) 0.11 (B) 0.13 (C) 0.15 (D) 0.18 (E) 0.19

Exam questions continue on the next page . . .

24. For an insurance coverage, you observe the following claim sizes:

400, 1100, 1100, 3000, 8000

You fit the loss distribution to a lognormal with $\mu = 7$ using maximum likelihood.

Determine the mean of the fitted distribution.

(A) Less than 2000
(B) At least 2000, but less than 2500
(C) At least 2500, but less than 3000
(D) At least 3000, but less than 3500
(E) At least 3500

25. In a mortality study performed on 5 lives, ages at death were

70, 72, 74, 75, 75

Estimate $S(75)$ using kernel smoothing with a uniform kernel with bandwidth 4.

(A) 0.2 (B) 0.3 (C) 0.4 (D) 0.5 (E) 0.6

26. For an insurance coverage, the number of claims per year follows a Poisson distribution with mean θ. The size of each claim follows an exponential distribution with mean 1000θ. Claim count and size are independent given θ.

You are examining one year of experience for four randomly selected policyholders, whose claims are as follows:

Policyholder #1	2000, 4000, 4000, 7000
Policyholder #2	4000
Policyholder #3	2000, 3000
Policyholder #4	1000, 4000, 5000

You use maximum likelihood to estimate θ.

Determine the variance of aggregate losses based on the fitted distribution.

(A) Less than 48,000,000
(B) At least 48,000,000, but less than 50,000,000
(C) At least 50,000,000, but less than 52,000,000
(D) At least 52,000,000, but less than 54,000,000
(E) At least 54,000,000

27. You are given:

(i) The annual number of claims for each risk follows a Poisson distribution with parameter Λ.
(ii) Λ varies by insured according to a gamma distribution with $\alpha = 3$ and $\theta = 0.1$.
(iii) Claim sizes follow a Pareto distribution with $\alpha = 3$ and $\theta = 20{,}000$.
(iv) Claim sizes are independent of claim counts.
(v) Your department handles only claims with sizes below 10,000.

Determine the variance of the annual number of claims handled per risk in your department.

(A) 0.159 (B) 0.176 (C) 0.226 (D) 0.232 (E) 0.330

Exam questions continue on the next page . . .

28. Past data on aggregate losses for two group policyholders is given in the following table.

Group		Year 1	Year 2
A	Total losses	1000	1200
	Number of members	40	50
B	Total losses	500	600
	Number of members	20	40

Calculate the credibility factor used for Group A's experience using non-parametric empirical Bayes estimation methods.

(A) Less than 0.40
(B) At least 0.40, but less than 0.45
(C) At least 0.45, but less than 0.50
(D) At least 0.50, but less than 0.55
(E) At least 0.55

29. For an insurance coverage, claim size follows a Pareto distribution with parameters $\alpha = 4$ and θ. θ varies by insured and follows a normal distribution with $\mu = 3$ and $\sigma = 1$.

Determine the Bühlmann credibility to be assigned to a single claim.

(A) 0.05 (B) 0.07 (C) 0.10 (D) 0.14 (E) 0.18

30. You are given the following information regarding loss sizes:

d	Mean excess loss $e(d)$	$F(d)$
0	3000	0.0
500	2800	0.1
10,000	2600	0.8

Determine the average payment per loss for a policy with an ordinary deductible of 500 and a maximum covered loss of 10,000.

(A) Less than 1600
(B) At least 1600, but less than 1800
(C) At least 1800, but less than 2000
(D) At least 2000, but less than 2200
(E) At least 2200

Exam questions continue on the next page . . .

31. A claims adjustment facility adjusts all claims for amounts less than or equal to 10,000. Claims for amounts greater than 10,000 are handled elsewhere.

In 2002, the claims handled by this facility fell into the following ranges:

Size of Claim	Number of Claims
Less than 1000	100
1000– 5000	75
5000–10000	25

The claims are fitted to a parametric distribution using maximum likelihood.

Which of the following is the correct form for the likelihood function of this experience?

(A) $\big(F(1000)\big)^{100}\big(F(5000)-F(1000)\big)^{75}\big(F(10{,}000)-F(5000)\big)^{25}$

(B) $\dfrac{\big(F(1000)\big)^{100}\big(F(5000)-F(1000)\big)^{75}\big(F(10{,}000)-F(5000)\big)^{25}}{\big(F(10{,}000)\big)^{200}}$

(C) $\dfrac{\big(F(1000)\big)^{100}\big(F(5000)-F(1000)\big)^{75}\big(F(10{,}000)-F(5000)\big)^{25}}{\big(1-F(10{,}000)\big)^{200}}$

(D) $\big(F(1000)\big)^{100}\big(F(5000)-F(1000)\big)^{75}\big(F(10{,}000)-F(5000)\big)^{25}\big(F(10{,}000)\big)^{200}$

(E) $\big(F(1000)\big)^{100}\big(F(5000)-F(1000)\big)^{75}\big(F(10{,}000)-F(5000)\big)^{25}\big(1-F(10{,}000)\big)^{200}$

32. For a sample from an exponential distribution, which of the following statements is false?

(A) If the sample has size 2, the sample median is an unbiased estimator of the population median.

(B) If the sample has size 2, the sample median is an unbiased estimator of the population mean.

(C) If the sample has size 3, the sample mean is an unbiased estimator of the population mean.

(D) If the sample has size 3, 1.2 times the sample median is an unbiased estimator of the population mean.

(E) The sample mean is a consistent estimator of the population mean.

33. The median of a sample is 5. The sample is fitted to a mixture of two exponential distributions with means 3 and $x > 3$, using percentile matching to determine the weights to assign to each exponential.

Which of the following is the range of values for x for which percentile matching works?

(A) $3 < x < 4.3281$.

(B) $3 < x < 7.2135$.

(C) $4.3281 < x < 7.2135$.

(D) $x > 4.3281$.

(E) $x > 7.2135$.

Exam questions continue on the next page . . .

34. A random variable X has the probability density function

$$f(x) = \frac{32e^{-4/x}}{x^4}$$

$\bar{X}$ is the sample mean of 100 observations of X.

Using the normal approximation, estimate $\Pr(\bar{X} < 2.5)$.

(A) 0.6915 (B) 0.8413 (C) 0.9332 (D) 0.9772 (E) 0.9938

35. Losses follow a lognormal distribution with $\mu = 3$, $\sigma = 0.5$.

Calculate the Value at Risk measure at security level $p = 95\%$.

(A) Less than 40
(B) At least 40, but less than 45
(C) At least 45, but less than 50
(D) At least 50, but less than 55
(E) At least 55

Solutions to the above questions begin on page 1503.

Answer Key for Practice Exam 2

1	C	11	E	21	E	31	B
2	A	12	B	22	D	32	A
3	D	13	D	23	D	33	E
4	E	14	A	24	A	34	E
5	E	15	C	25	B	35	C
6	A	16	C	26	E		
7	D	17	C	27	C		
8	D	18	D	28	E		
9	A	19	D	29	A		
10	A	20	E	30	D		

Practice Exam 2

1. **[Lesson 1]** $F(9{,}000) = 0.4 + \left(\frac{9{,}000-5{,}000}{10{,}000-5{,}000}\right)(0.9 - 0.4) = 0.8$. Then $S(9{,}000) = 0.2$. The hazard rate is $\frac{f(x)}{S(x)}$. The density function is the derivative of F, which at 9,000 is the slope of the line from 5,000 to 10,000, which is $\frac{0.9-0.4}{10{,}000-5{,}000} = 0.0001$. The answer is

$$h(9{,}000) = \frac{f(9{,}000)}{S(9{,}000)} = \frac{0.0001}{0.2} = \boxed{\textbf{0.0005}} \qquad \textbf{(C)}$$

2. **[Lesson 22]** The distribution function at 1000, $F(1000)$, is $\frac{5}{12}$, and $F(5000) = \frac{9}{12}$. By definition, the median is the point m such that $F(m) = 0.5$. The ogive linearly interpolates between 1000 and 5000. Thus we solve the equation

$$\frac{m - 1000}{5000 - 1000} = \frac{0.5 - \frac{5}{12}}{\frac{9}{12} - \frac{5}{12}}$$

$$m - 1000 = \frac{1}{4}(4000) = 1000$$

$$m = \boxed{\textbf{2000}} \qquad \textbf{(A)}$$

3. **[Lesson 26]** The variance is $\sum \frac{1}{r_i^2}$, where r_i is the i^{th} risk set. Since there are originally 5 individuals, and one individual drops out from the risk set at each event time, the first four risk sets, if there were no censored observation, would be {5, 4, 3, 2}. If none of these are censored, the variance is $1/5^2 + 1/4^2 + 1/3^2 + 1/2^2$. If one of these is censored, the variance is the sum of three terms instead of four, making it lower. Of the four reciprocals of {5, 4, 3, 2}, 1/2 is the largest. Removing 1/2 reduces the variance the most. The risk set of size 2 corresponds to the fourth event, time 4. By making $\boxed{\textbf{4}}$ the censored observation, the r_i's are {5, 4, 3}, minimizing $\sum 1/r_i^2$. **(D)**

4. **[Lesson 41]** The credibility formula in terms of expected number of claims requires $1 + \text{CV}_s^2$, or the second moment divided by the first moment squared of the severity distribution. For a Pareto with

$\alpha = 3$, this is

$$\frac{\frac{2\theta^2}{(2)(1)}}{\left(\frac{\theta}{2}\right)^2} = 4$$

The expected number of claims needed for full credibility is then $\left(\frac{1.96}{0.05}\right)^2(4) = 6146.56$, so we have

$$6146.56 = 10{,}000\lambda$$
$$\lambda = \boxed{\mathbf{0.614656}} \qquad \textbf{(E)}$$

5. [**Lessons 37, 38**] (A) is false because the number of degrees of freedom is $n - 1$ minus the number of parameters estimated. Here $n = 7$ and the inverse Pareto has 2 parameters, so there are 4 degrees of freedom.

(B) is false, as indicated on page 780.

(C) is false, as indicated on page 779, where it says that the indicated critical values only work when the distribution is completely specified, not when parameters have been estimated.

(D) is false; in fact, the critical values get divided by $\sqrt{n}$.

(E) is true.

6. [**Lesson 52**] For the beta, the mean is 0.5, so the hypothetical and overall means are calculated as

$$\mathbf{E}[T - \mu - 0.5 \mid \mu] = 0.5$$
$$\mathbf{E}[T - \mu \mid \mu] = 0.5 + 0.5 = 1$$
$$\mathbf{E}[T \mid \mu] = \mathbf{E}[T - \mu \mid \mu] + \mu = \mu + 1 \quad \text{(This is the hypothetical mean.)}$$
$$\mathbf{E}[T] = \mathbf{E}\big[\mathbf{E}[T \mid \mu]\big] = 1 + \mathbf{E}[\mu] = 17 \quad \text{(This is the overall mean.)}$$

The variance of the hypothetical means is the variance of μ. Since μ is uniformly distributed on $[15, 17]$, its variance is $a = \frac{(17-15)^2}{12} = \frac{1}{3}$.

The process variance is $\text{Var}(T \mid \mu)$. The conditional variable $T \mid \mu$ is a shifted beta, and variance is unaffected by shifting. The variance of an unshifted beta is $\frac{ab}{(a+b)^2(a+b+1)} = \frac{4}{4^2(5)} = \frac{1}{20} = 0.05$. So the process variance is the constant 0.05, and the expected value of the process variance is also $v = 0.05$.

The Bühlmann credibility is $Z = \frac{1/3}{1/3+0.05} = \frac{20}{23}$. The Bühlmann prediction of travel time is $\left(\frac{20}{23}\right)(16) + \left(\frac{3}{23}\right)(17) = \boxed{\mathbf{16\frac{3}{23}}}$. **(A)**

7. [**Lesson 44**] The probability of 200 is $0.3(0.5)+0.2(0.5^2) = 0.2$ from A and $0.3(0.4)+0.2(0.6^2) = 0.192$ from B. The posterior probability of A is $\frac{0.2}{0.392}$ and the posterior probability of B is $\frac{0.192}{0.392}$.

The mean aggregate loss in class A is

$$\mathbf{E}[S] = \mathbf{E}[N]\,\mathbf{E}[X] = [0.3 + 0.2(2) + 0.1(3)](150) = 150$$

and the variance is

$$\text{Var}(S) = \mathbf{E}[N]\,\text{Var}(X) + \text{Var}(N)\,\mathbf{E}[X]^2$$
$$\text{Var}(N) = \left(0.3(1^2) + 0.2(2^2) + 0.1(3^2)\right) - 1^2 = 1$$
$$\text{Var}(X) = (200 - 100)^2(0.5)(0.5) = 2500$$
$$\text{Var}(S) = (1)(2500) + (1)(150^2) = 25{,}000$$

The mean and variance in class B are:

$$\begin{aligned}
\mathbf{E}[S] &= \mathbf{E}[N]\,\mathbf{E}[X] = [0.3 + 0.2(2) + 0.2(3)](140) = (1.3)(140) = 182 \\
\text{Var}(N) &= \left(0.3(1^2) + 0.2(2^2) + 0.2(3^2)\right) - 1.3^2 = 1.21 \\
\text{Var}(X) &= (200 - 100)^2(0.6)(0.4) = 2400 \\
\text{Var}(S) &= (1.3)(2400) + (1.21)(140^2) = 26{,}836
\end{aligned}$$

Let I be the indicator variable for the class (A or B). The predictive variance of aggregate losses is

$$\begin{aligned}
\text{Var}(S) &= \mathbf{E}\big[\text{Var}(S \mid I)\big] + \text{Var}\big(\mathbf{E}[S \mid I]\big) \\
&= \frac{0.2}{0.392}(25{,}000) + \frac{0.192}{0.392}(26{,}836) + (182 - 150)^2\left(\frac{0.2}{0.392}\right)\left(\frac{0.192}{0.392}\right) \\
&= 12{,}755.10 + 13{,}144.16 + 255.89 = \boxed{\mathbf{26{,}155}} \qquad \textbf{(D)}
\end{aligned}$$

8. **[Lesson 44]** We need to calculate $\mathbf{E}[X \mid X > 65]$, where X is the grade. By definition

$$\mathbf{E}[X \mid X > 65] = \frac{\int_{65}^{100} x f(x)\mathrm{d}x}{1 - F(65)}$$

$f(x)$ is an equally weighted mixture of the good and bad students, and therefore is

$$f(x) = \frac{1}{2}\left(\frac{4}{100}\left(\frac{x}{100}\right)^3 + \frac{2}{100}\left(\frac{x}{100}\right)\right)$$

First we calculate the denominator of $\mathbf{E}[X \mid X > 65]$.

$$\begin{aligned}
F(x) &= \int_0^x f(u)\mathrm{d}u = 0.5\left(\left(\frac{x}{100}\right)^4 + \left(\frac{x}{100}\right)^2\right) \\
F(65) &= 0.5(0.65^4 + 0.65^2) = 0.300503 \\
1 - F(65) &= 1 - 0.300503 = 0.699497
\end{aligned}$$

Then we calculate the numerator.

$$\begin{aligned}
\int_{65}^{100} x f(x)\mathrm{d}x &= \int_{65}^{100} 0.5\left(\frac{4x^4}{100^4} + \frac{2x^2}{100^2}\right)\mathrm{d}x \\
&= 0.5\left(\frac{4x^5}{5(100^4)} + \frac{2x^3}{3(100^2)}\right)_{65}^{100} \\
&= 0.5\left(\frac{400}{5} - \frac{4(65^5)}{5(100^4)} + \frac{200}{3} - \frac{2(65^3)}{3(100^2)}\right) \\
&= 0.5(80 - 9.2823 + 66.6667 - 18.3083) = 59.5380 \\
\mathbf{E}[X \mid X > 65] &= \frac{59.5380}{0.699497} = \boxed{\mathbf{85.1155}} \qquad \textbf{(D)}
\end{aligned}$$

An alternative way to solve this problem is to use the tabular form for Bayesian credibility that we studied in Lesson 44. The table would look like this:

	Good Students	Bad Students	
Prior probabilities	0.50	0.50	
Likelihood of experience	0.821494	0.5775	
Joint probabilities	0.410747	0.28875	0.699497
Posterior probabilities (p_i)	0.587203	0.412797	
65+$e(65)$	86.084252	83.737374	
$[65 + e(65)] \times p_i$	50.548957	34.566511	**85.1155**

The second line, the likelihood, is derived as follows:

$$F(65) = \int_0^{65} f(x)\,\mathrm{d}x = \begin{cases} (65/100)^4 & \text{for good students} \\ (65/100)^2 & \text{for bad students} \end{cases}$$

so the likelihood of more than 65 is $1 - 0.65^4 = 0.821494$ for good students and $1 - 0.65^2 = 0.5775$ for bad students.

The fifth line, the average grade of those with grades over 65, is derived as follows for good students:

$$\begin{aligned} 65 + e(65) &= 65 + \frac{\int_{65}^{100} S(x \mid \text{Good})\,\mathrm{d}x}{S(65 \mid \text{Good})} \\ &= 65 + \frac{\int_{65}^{100} \left(1 - \left(\frac{x}{100}\right)^4\right)\mathrm{d}x}{1 - 0.65^4} \\ &= 65 + \frac{35 - \left(\frac{x^5}{5(100^4)}\right)\Big|_{65}^{100}}{1 - 0.65^4} \\ &= 65 + \frac{35 - \frac{100^5 - 65^5}{5(100^4)}}{1 - 0.65^4} = 86.084252 \end{aligned}$$

and for bad students, replace the 4's with 2's and the 5's with 3's to obtain $65 + \frac{35 - \frac{100^3 - 65^3}{2(100^2)}}{1 - 0.65^2} = 83.737374$.

9. **[Lesson 63]** The estimates of $S(10)$ from these five samples are the proportion of numbers above 10, which are 0.2, 0.2, 0.4, 0.2, 0.6 respectively, so the bootstrap approximation is

$$\frac{(0.2 - 0.4)^2 + (0.2 - 0.4)^2 + (0.4 - 0.4)^2 + (0.2 - 0.4)^2 + (0.6 - 0.4)^2}{5} = \boxed{\mathbf{0.032}} \qquad \textbf{(A)}$$

10. **[Lesson 51]** Let μ_A be the hypothetical mean for A, v_A the process variance for A, and use the same notation with subscripts B for B. For process variance, we will use the compound variance formula to compute the process variances. In the case of Class A (with N and X being frequency and severity respectively):

$$\mathbf{E}[N] = 0.1 \qquad \mathrm{Var}(N) = 0.1(0.9) = 0.09$$

$$\mathbf{E}[X] = \frac{\theta}{\alpha - 1} = \frac{50}{2} = 25 \qquad \mathbf{E}[X^2] = \frac{2\theta^2}{(\alpha - 1)(\alpha - 2)} = \frac{2(50^2)}{(2)(1)} = 2500$$

so $\mathrm{Var}(X) = 2500 - 625$. A similar calculation is done for Class B. So using the compound variance formula for v_i, we have

$$\mu_A = 0.1(25) = 2.5 \qquad v_A = 0.1(2500 - 625) + 0.09(25^2) = 243.75$$

$$\mu_B = 0.2(30) = 6 \qquad v_B = 0.2(3600 - 900) + 0.16(30^2) = 684$$
$$\text{VHM} = (6 - 2.5)^2\left(\frac{1}{4}\right) = 3.0625 \qquad \text{EPV} = \frac{1}{2}(243.75 + 684) = 463.875$$

For 2 years of experience, the credibility is

$$Z = \frac{N\,\text{VHM}}{N\,\text{VHM} + \text{EPV}} = \frac{(2)(3.0625)}{(2)(3.0625) + 463.875} = \boxed{\mathbf{0.013032}} \qquad \textbf{(A)}$$

11. **[Lesson 32]** The likelihood function (either using the fact that the exponential is memoryless, or else writing it all out and canceling out the denominators) is

$$L(\theta) = \left(1 - e^{-(1000/\theta)}\right)^{20}\left(e^{-(1000/\theta)}\right)^{10}\left(1 - e^{-(500/\theta)}\right)^{32}\left(e^{-(500/\theta)}\right)^{24}$$

Logging and differentiating:

$$l(\theta) = 20\ln\left(1 - e^{-1000/\theta}\right) + 32\ln\left(1 - e^{-500/\theta}\right) - \frac{10{,}000 + 12{,}000}{\theta}$$
$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = \frac{-20{,}000e^{-1000/\theta}}{\theta^2(1 - e^{-1000/\theta})} + \frac{-16{,}000e^{-500/\theta}}{\theta^2(1 - e^{-500/\theta})} + \frac{22{,}000}{\theta^2} = 0$$

Multiply through by $\frac{\theta^2}{1000}$, and set $x = e^{-500/\theta}$. We obtain

$$22 - \frac{20x^2}{1 - x^2} - \frac{16x}{1 - x} = 0$$
$$\frac{22(1 - x^2) - 20x^2 - 16x(1 + x)}{1 - x^2} = 0$$
$$22 - 22x^2 - 20x^2 - 16x - 16x^2 = 0$$
$$58x^2 + 16x - 22 = 0$$
$$x = \frac{-16 + \sqrt{16^2 + 4(22)(58)}}{116} = 0.493207$$
$$\theta = \frac{-500}{\ln x} = 707.3875$$

An easier way to do this would be to make the substitution $x = e^{-500/\theta}$ right in $L(\theta)$, and to immediately recognize that $1 - x^2 = (1 - x)(1 + x)$. This would avoid the confusing differentiation step:

$$l(\theta) = 20\ln(1 - x^2) + 44\ln x + 32\ln(1 - x)$$
$$= 20\ln(1 - x) + 20\ln(1 + x) + 44\ln x + 32\ln(1 - x)$$
$$= 52\ln(1 - x) + 20\ln(1 + x) + 44\ln x$$
$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = -\frac{52}{1 - x} + \frac{20}{1 + x} + \frac{44}{x} = 0$$
$$-52(x)(1 + x) + 20(x)(1 - x) + 44(1 - x^2) = 0$$
$$-52x^2 - 52x - 20x^2 + 20x + 44 - 44x^2 = 0$$
$$116x^2 + 32x - 44 = 0$$

which is double the quadratic above, and leads to the same solution for θ.

Using the fact that the exponential distribution is memoryless, the average total loss size for a 500 deductible is $500 + 707.3875 = \boxed{\mathbf{1207.3875}}$. **(E)**

12. [**Lesson 28**] The conditional probability of death in the second year, q_1, is estimated by $\frac{d_2}{e_2}$, number of deaths over the exposure in the second year. Year 2 starts with $1000 - 100 - 33 = 867$ lives, and since withdrawals and new entries occur uniformly, we add half the new entries and subtract half the withdrawals to arrive at $e_2 = 867 + 0.5(500 - 100) = 1067$. Then

$$0.03 = \hat{q}_1 = \frac{c}{1067}$$
$$c = \boxed{\mathbf{32}} \qquad \textbf{(B)}$$

13. [**Lesson 48**] The prior distribution is a beta distribution with $a = 2$, $b = 6$. (In general, in a beta distribution, a is 1 more than the exponent of q and b is 1 more than the exponent of $1 - q$.) The number of claims is binomial, which means that 2 claims are possible each year. Of the 8 possible claims in 4 years, you received 1 and didn't receive 7. Thus $a' = 2 + 1 = 3$ and $b' = 6 + 7 = 13$ are the parameters of the posterior beta. The density function for the posterior beta is

$$\begin{aligned}\pi(q|\mathbf{x}) &= \frac{\Gamma(3+13)}{\Gamma(3)\Gamma(13)} q^2(1-q)^{12} \\ &= 1365q^2(1-q)^{12}\end{aligned}$$

since $\frac{15!}{2!12!} = \frac{(15)(14)(13)}{2} = 1365$. We must integrate this function from 0 to 0.25 to obtain $\Pr(Q < 0.25)$. It is easier to integrate if we change the variable, by setting $q' = 1 - q$. Then we have

$$\begin{aligned}\Pr(Q < 0.25) &= 1365 \int_{0.75}^{1} (1-q')^2 q'^{12} \mathrm{d}q' \\ &= 1365 \int_{0.75}^{1} (q'^{12} - 2q'^{13} + q'^{14}) \mathrm{d}q' \\ &= \left. \frac{q'^{13}}{13} - \frac{2q'^{14}}{14} + \frac{q'^{15}}{15} \right|_{0.75}^{1} \\ &= 1365(0.0007326 - 0.0001730) = \boxed{\mathbf{0.7639}} \qquad \textbf{(D)}\end{aligned}$$

14. [**Lesson 59**] The distribution function is

$$F(x) = \int_{-1}^{x} -2u \, \mathrm{d}u = \left. -u^2 \right|_{-1}^{x} = 1 - x^2$$

Inverting,

$$\begin{aligned} u &= 1 - x^2 \\ 1 - u &= x^2 \\ x &= -\sqrt{1-u} \end{aligned}$$

It is necessary to use the negative square root, since the simulated observation must be between -1 and 0. So

$$\begin{aligned} x_1 &= -\sqrt{1-0.2} = -0.8944 \\ x_2 &= -\sqrt{1-0.4} = -0.7746 \\ x_3 &= -\sqrt{1-0.3} = -0.8367 \\ x_4 &= -\sqrt{1-0.7} = -0.5477 \end{aligned}$$

$$\frac{-0.8944 - 0.7746 - 0.8367 - 0.5477}{4} = \boxed{\mathbf{-0.7634}} \qquad \textbf{(A)}$$

15. [**Lesson 30**] We write the moment equations for the first and second moments:

$$\begin{aligned}
\frac{2+3+4+x_1+x_2}{5} &= \frac{\theta}{\alpha-1} \\
9 + x_1 + x_2 &= 5\left(\frac{373.71}{47.71-1}\right) = 40 \\
\frac{2^2+3^2+4^2+x_1^2+x_2^2}{5} &= \frac{2\theta^2}{(\alpha-1)(\alpha-2)} \\
29 + x_1^2 + x_2^2 &= 5\left(\frac{2(373.71^2)}{(46.71)(45.71)}\right) = 654
\end{aligned}$$

We use the first equation to solve for x_2 in terms of x_1, and plug that into the second equation and solve.

$$\begin{aligned}
x_2 &= 31 - x_1 \\
29 + x_1^2 + (31 - x_1)^2 &= 654 \\
29 + 2x_1^2 - 62x_1 + 961 &= 654 \\
2x_1^2 - 62x_1 + 336 &= 0 \\
x_1^2 - 31x_1 + 168 &= 0 \\
x_1 &= \frac{31 \pm \sqrt{31^2 - 4(168)}}{2} \\
&= \frac{31 \pm 17}{2} = 7 \text{ or } 24
\end{aligned}$$

Since x_2 is higher than x_1, $x_1 = \boxed{7}$. **(C)**

16. [**Lesson 51**] The hypothetical mean is Λ. The expected hypothetical mean $\mu = \mathbf{E}[\Lambda] = 1$ (the mean of the uniform distribution). The process variance is Λ. The expected process variance EPV, or the expected value of Λ, is 1. The variance of the hypothetical mean is $\text{Var}(\Lambda)$. For a uniform distribution on $(0, \theta)$, the variance is $\frac{\theta^2}{12}$, so the variance is VHM $= \frac{1}{3}$. The Bühlmann K is therefore $\frac{1}{1/3} = 3$. $Z = \frac{1}{1+3} = 0.25$. The credibility premium is $\frac{1}{4}(5) + \frac{3}{4}(1) = \boxed{2}$. **(C)**

17. [**Section 60.1**] By SOA rounding rules, since $\Phi(0.58) = 0.7190$ and $\Phi(0.59) = 0.7224$, 0.58 is the closest inverse to 0.72 and $\Phi^{-1}(0.28) = -0.58$. Then $\lambda = 0.6 - 0.58\sqrt{0.04} = 0.484$. For the Poisson distribution,

$$\begin{aligned}
p_0 &= e^{-0.484} = 0.6163 \\
p_1 &= 0.484e^{-0.484} = 0.2983 & F(1) &= 0.6163 + 0.2983 = 0.9146 \\
p_2 &= \frac{0.484^2}{2}e^{-0.484} = 0.0722 & F(2) &= 0.9146 + 0.0722 = 0.9868
\end{aligned}$$

Thus $0.82 \to 1$, $0.13 \to 0$, $0.94 \to 2$, and the sum of claim counts is $1 + 0 + 2 = \boxed{3}$. **(C)**

18. **[Lesson 24 and section 25.2]** First we calculate $\hat{S}(26)$.

y_i	r_i	s_i	$S_{10}(y_i)$
5	10	1	0.9
8	8	1	0.7875
10	7	1	0.675
11	6	1	0.5625
17	5	1	0.45
20	4	1	0.3375
26	2	1	0.16875

So $\hat{S}(26) = 0.16875$. To extrapolate, we exponentiate $\hat{S}(26)$ to the $\frac{30}{26}$ power, as discussed in example 24E on page 423:

$$\hat{S}(30) = 0.16875^{30/26} = 0.12834.$$

$\Pr(20 \le T \le 30) = S(20^-) - S(30)$, since the lower endpoint is included. But $S(20^-) = S(17) = 0.45$. So the answer is $0.45 - 0.12834 = \boxed{\mathbf{0.3217}}$. **(D)**

19. **[Section 4.1 and Lesson 11]** The density of the uniform distribution is the reciprocal of the range $(1/2)$, or 2. We integrate p_0 for the binomial, or $(1-q)^2$, over the uniform distribution.

$$\begin{aligned}\Pr(N=0) &= \int_{0.25}^{0.75} (2)(1-q)^2 \mathrm{d}q \\ &= -2\left(\frac{(1-q)^3}{3}\right)\Bigg|_{0.25}^{0.75} \\ &= \left(\frac{2}{3}\right)(0.75^3 - 0.25^3) = \boxed{\mathbf{0.270833}} \qquad \textbf{(D)}\end{aligned}$$

20. **[Lesson 57]** We estimate μ, v, and a:

$$\hat{\mu} = \widehat{\text{EPV}} = \bar{x} = \frac{22(1) + 6(2) + 2(3)}{100} = \frac{40}{100} = 0.4$$

We will calculate s^2, the unbiased sample variance, by calculating the second moment, subtracting the square of the sample mean (which gets us the empirical variance) and then multiplying by $\frac{n}{n-1}$ to turn it into the sample variance.

$$\begin{aligned}\widehat{\text{VHM}} + \widehat{\text{EPV}} = s^2 &= \frac{100}{99}\left(\frac{22(1^2) + 6(2^2) + 2(3^2)}{100} - \bar{x}^2\right) \\ &= \frac{100}{99}(0.64 - 0.4^2) = 0.4848 \\ \widehat{\text{VHM}} &= 0.4848 - 0.4 = 0.0848 \\ Z = \frac{\text{VHM}}{\text{VHM} + \text{EPV}} &= \frac{0.0848}{0.4848} \\ P_C = 0.4 + \frac{0.0848}{0.4848}(1.6) &= \boxed{\mathbf{0.68}} \qquad \textbf{(E)}\end{aligned}$$

21. **[Lesson 61]** $F_X(500)$ is approximately 0.4 based on the fifty runs. The standard deviation of $\hat{F}_X(500)$ is therefore $\sqrt{(0.4)(0.6)/n}$. We want the half-width of the confidence interval equal to 5% of 0.4, or

$$\begin{aligned} 1.96\sqrt{(0.4)(0.6)/n} &= 0.05(0.4) \\ \frac{0.9602}{0.02} &= \sqrt{n} \\ n &= 2304.96 \end{aligned}$$

2305 runs are needed. **(E)**

22. **[Lessons 12 and 14]** Let N be claim counts, X claim size, S aggregate claims. N is a gamma mixture of a Poisson, or a negative binomial. The gamma has parameters α and θ (not the same θ as the Weibull) such that

$$\begin{aligned} \alpha\theta &= 0.5 \\ \alpha\theta^2 &= 1 \end{aligned}$$

implying $\beta = \theta = 2$, $r = \alpha = 0.25$, so

$$\begin{aligned} \mathbf{E}[N] &= r\beta = 0.5 \\ \text{Var}(N) &= r\beta(1+\beta) = 0.25(2)(3) = 1.5 \end{aligned}$$

The Weibull has mean[1]

$$\mathbf{E}[X] = \theta\Gamma(1+2) = (5)(2!) = (5)(2) = 10$$

and second moment

$$\mathbf{E}\left[X^2\right] = \theta^2\Gamma(1+2^2) = (5^2)(4!) = (25)(24) = 600$$

and therefore $\text{Var}(X) = 600 - 10^2 = 500$. By the compound variance formula

$$\text{Var}(S) = (0.5)(500) + (1.5)(10^2) = \boxed{\mathbf{400}} \qquad \textbf{(D)}$$

23. **[Lesson 37]** We set up a table for the empirical and fitted functions. Note that we do not know the empirical function at 10,000 or higher due to the policy limit.

x	$F_n(x^-)$	$F_n(x)$	$F(x)$	Largest difference
1000	0	0.1	$1 - e^{-1/6} = 0.1535$	0.1535
2000	0.1	0.4	$1 - e^{-1/3} = 0.2835$	0.1835
4000	0.4	0.5	$1 - e^{-2/3} = 0.4866$	0.0866
5000	0.5	0.7	$1 - e^{-5/6} = 0.5654$	0.1346
10,000	0.7		$1 - e^{-5/3} = 0.8111$	0.1111

Inspection indicates that the maximum difference occurs at 2000 and is **0.1835**. **(D)**

24. **[Lesson 33]** See the discussion of transformations and the lognormal Example 33C on page 648, and the paragraph before the example. *The shortcut for lognormals must be adapted for this question since μ is given; it is incorrect to calculate the empirical mean and variance as if μ were unknown.* We will log each of the

[1] In general, $\Gamma(n) = (n-1)!$ for n an integer

claim sizes and fit them to a normal distribution. You may happen to know that for a normal distribution, the MLE's of μ and σ are independent, so given μ, the MLE for σ will be the same as if μ were not given. Moreover, the MLE for σ for a normal distribution is the sample variance divided by n (rather than by $n-1$). If you know these two facts, you can calculate the MLE for σ on the spot. If not, it is not too hard to derive. The likelihood function (omitting the constant $\sqrt{2\pi}$) is (where we let $x_i =$ the *log* of the claim size)

$$L(\sigma) = \frac{1}{\sigma^5}\prod_{i=1}^{5} e^{-\frac{(x_i-7)^2}{2\sigma^2}}$$

$$l(\sigma) = -5\ln\sigma - \sum_{i=1}^{5}\frac{(x_i-7)^2}{2\sigma^2}$$

$$\frac{\mathrm{d}l}{\mathrm{d}\sigma} = -\frac{5}{\sigma} + \frac{\sum_{i=1}^{5}(x_i-7)^2}{\sigma^3} = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{5}(x_i-7)^2}{5}$$

We calculate

$$\sigma^2 = \frac{(\ln 400-7)^2 + 2(\ln 1100-7)^2 + (\ln 3000-7)^2 + (\ln 8000-7)^2}{5} = 1.1958$$

The mean of the lognormal is

$$\exp\left(\mu + \sigma^2/2\right) = \exp(7 + 1.1958/2) = \boxed{\textbf{1994}} \qquad \textbf{(A)}$$

25. [**Lesson 27**] The kernel survival function for a uniform kernel is a straight line from 1 to 0, starting at the observation point minus the bandwidth and ending at the observation point plus the bandwidth. From the perspective of 74, we reverse orientation; the kernel survival for 74 *increases* as the observation increases. Therefore, the kernels are 0 at 70, $\frac{1}{8}$ at 72 (which is $\frac{1}{8}$ of the way from 71 to 79), $\frac{3}{8}$ at 74 (which is $\frac{3}{8}$ of the way from 71 to 79), and $\frac{1}{2}$ at 75 (which is $\frac{1}{2}$ of the way from 71 to 79). Each point has a weight of $\frac{1}{n} = \frac{1}{5}$. We therefore have:

$$\frac{1}{5}\left(\frac{1}{8} + \frac{3}{8} + (2)\left(\frac{1}{2}\right)\right) = \boxed{\textbf{0.30}} \qquad \textbf{(B)}$$

26. [**Lesson 32**] The likelihood function is the product of

$$e^{-\theta}\frac{\theta^{n_i}}{n_i!}$$

for the number of claims n_i for the 4 individuals, times the product of

$$\frac{1}{1000\theta}e^{-x_i/1000\theta}$$

for each of the 10 claim sizes x_i. These get multiplied together to form the likelihood function. We have

$$\sum n_i = 4 + 1 + 2 + 3 = 10$$

and

$$\sum \frac{x_i}{1000\theta} = \frac{2000+4000+4000+7000+4000+2000+3000+1000+4000+5000}{1000\theta} = \frac{36}{\theta}$$

If we ignore the constants, the likelihood function is:

$$\begin{aligned} L(\theta) &= e^{-4\theta}\theta^{10}\frac{1}{\theta^{10}}e^{-36/\theta} \\ l(\theta) &= -4\theta - \frac{36}{\theta} \\ \frac{\mathrm{d}l}{\mathrm{d}\theta} &= -4 + \frac{36}{\theta^2} = 0 \\ \theta &= 3 \end{aligned}$$

To complete the answer to the question, use equation (14.2), or better, since number of claims is Poisson, equation (14.4). Let S be aggregate losses. Using either formula, we obtain $\mathrm{Var}(S) = 3\big(2(3000^2)\big) =$ **54,000,000** for the fitted distribution. **(E)**

27. **[Lessons 12 and 13]** The mixed number of claims for all risks is negative binomial with $r = 3$, $\beta = 0.1$. However, this must be adjusted for severity modification; only $F(10{,}000) = 1 - \left(\frac{20{,}000}{30{,}000}\right)^3 = \frac{19}{27}$ of claims are handled by your department, where F is the distribution function of a Pareto. The modification is to set $\beta = \frac{19}{27}(0.1)$. The variance is then $r\beta(1+\beta) = 3\left(\frac{1.9}{27}\right)\left(1 + \left(\frac{1.9}{27}\right)\right) =$ **0.2260**. **(C)**

28. **[Subsection 56.2]** We apply formulas (56.5) and (56.6).

$$\begin{aligned} \bar{x}_1 &= \frac{1000+1200}{40+50} = 24\tfrac{4}{9} \\ \bar{x}_2 &= \frac{500+600}{20+40} = 18\tfrac{1}{3} \\ \bar{x} &= \frac{1000+1200+500+600}{40+50+20+40} = \frac{3300}{150} = 22 \\ \hat{v} &= \frac{40\left(\frac{1000}{40} - 24\frac{4}{9}\right)^2 + 50\left(\frac{1200}{50} - 24\frac{4}{9}\right)^2 + 20\left(\frac{500}{20} - 18\frac{1}{3}\right)^2 + 40\left(\frac{600}{40} - 18\frac{1}{3}\right)^2}{2} \\ &= 677\tfrac{7}{9} \\ \hat{a} &= \frac{1}{150 - \frac{1}{150}(90^2+60^2)}\left(90(24\tfrac{4}{9}-22)^2 + 60(18\tfrac{1}{3}-22)^2 - (677\tfrac{7}{9})(1)\right) \\ &= \frac{666\frac{2}{3}}{72} = 9.2593 \\ \hat{k} &= \frac{\hat{v}}{\hat{a}} = \frac{677\frac{7}{9}}{9.2593} = 73.2 \\ \hat{Z}_1 &= \frac{90}{90+\hat{k}} = \frac{90}{90+73.2} = \boxed{\textbf{0.5515}} \qquad \textbf{(E)} \end{aligned}$$

29. **[Lesson 52]** For aggregate losses, the mean given θ is $\frac{\theta}{3}$ and the variance given θ is $\frac{2\theta^2}{6} - \left(\frac{\theta}{3}\right)^2 = \frac{2\theta^2}{9}$. Then

$$\mathrm{EPV} = \frac{2}{9}\,\mathbf{E}\left[\theta^2\right] = \frac{2}{9}(\mu^2+\sigma^2) = \frac{2}{9}(3^2+1^2) = \frac{20}{9}$$

$$\text{VHM} = \frac{1}{9}\,\text{Var}(\theta) = \frac{1}{9}\sigma^2 = \frac{1}{9}$$
$$Z = \frac{\text{VHM}}{\text{VHM} + \text{EPV}} = \boxed{\frac{1}{21}} \qquad \textbf{(A)}$$

30. [**Lesson 9**] Let X be loss size. Since $F(0) = 0$, $\mathbf{E}[X] = e(0) = 3000$. Then

$$\begin{aligned}
\mathbf{E}[X] &= \mathbf{E}[X \wedge d] + e(d)\big(1 - F(d)\big) \\
3000 &= \mathbf{E}[X \wedge 500] + e(500)\big(1 - F(500)\big) \\
&= \mathbf{E}[X \wedge 500] + (2800)(0.9) \\
\mathbf{E}[X \wedge 500] &= 480 \\
3000 &= \mathbf{E}[X \wedge 10{,}000] + e(10{,}000)\big(1 - F(10{,}000)\big) \\
&= \mathbf{E}[X \wedge 10{,}000] + (2600)(0.2) \\
\mathbf{E}[X \wedge 10{,}000] &= 2480 \\
\mathbf{E}[X \wedge 10{,}000] - \mathbf{E}[X \wedge 500] &= 2480 - 480 = \boxed{\textbf{2000}} \qquad \textbf{(D)}
\end{aligned}$$

An alternative which is slightly shorter is

$$\begin{aligned}
\mathbf{E}[(X - 500)_+] &= e(500)\big(1 - F(500)\big) = 2800(0.9) = 2520 \\
\mathbf{E}[(X - 10{,}000)_+] &= 2600(0.2) = 520 \\
\mathbf{E}[(X - 500)_+] - \mathbf{E}[(X - 10{,}000)_+] &= 2520 - 520 = \boxed{\textbf{2000}}
\end{aligned}$$

31. [**Lesson 32**] The claims are truncated, not censored, at 10,000. The probability of seeing any claim is $F(10{,}000)$. Any likelihood developed before considering this condition must be divided by this condition.

The likelihood of each of the 100 claims less than 1000, if not for the condition, is $F(1000)$. The conditional likelihood, conditional on the claim being below 10,000, is $\frac{F(1000)}{F(10{,}000)}$.

The likelihood of each of the 75 claims between 1000 and 5000, if not for the condition, is $F(5000) - F(1000)$. The conditional likelihood is $\frac{F(5000)-F(1000)}{F(10{,}000)}$.

The likelihood of each of the 25 claims between 5000 and 10,000, if not for the condition, is $F(10{,}000) - F(5000)$. The conditional likelihood is $\frac{F(10{,}000)-F(5000)}{F(10{,}000)}$.

Multiplying all these 200 likelihoods together we get answer **(B)**.

32. [**Lesson 21**] The sample mean is an unbiased estimator of the population mean, and if the population variance is finite (as it is if it has an exponential distribution), the sample mean is a consistent estimator of the population mean. (C) and (E) are therefore true. For a sample of size 2, the sample median is the sample mean, so (B) is true. (D) is proved in *Loss Models*. That leaves (A). **(A)** is false, because (B) is true and the median of an exponential is not the mean. In fact, it is the mean times $\ln 2$. So the sample median, which is an unbiased estimator of the mean, and therefore has an expected value of θ, does not have expected value $\theta \ln 2$, the value of the median.

33. [**Lesson 31**] For a mixture F is the weighted average of the F's of the individual distribution. The median of the mixture F is then the number m such that

$$wF_1(x) + (1 - w)F_2(x) = 0.5$$

where w is the weight. Here, it is more convenient to use survival functions. (The median is the number m such that $S(m) = 0.5$) $m = 5$. We have:

$$\begin{aligned} we^{-5/3} + (1-w)e^{-5/x} &= 0.5 \\ w(e^{-5/3} - e^{-5/x}) &= 0.5 - e^{-5/x} \\ w &= \frac{0.5 - e^{-5/x}}{e^{-5/3} - e^{-5/x}} \end{aligned}$$

In order for this procedure to work, w must be between 0 and 1. Note that since $x > 3$, $-\frac{5}{3} < -\frac{5}{x}$, so the denominator is negative. For $w > 0$, we need

$$\begin{aligned} 0.5 - e^{-5/x} &< 0 \\ e^{-5/x} &> 0.5 \\ -5/x &> \ln 0.5 \\ 5/x &< \ln 2 \\ x &> \frac{5}{\ln 2} = 7.2135 \end{aligned}$$

For $w < 1$, we need

$$\begin{aligned} e^{-5/x} - 0.5 &< e^{-5/x} - e^{-5/3} \\ 0.5 &> e^{-5/3} = 0.1889 \end{aligned}$$

and this is always true. So percentile matching works when $\boxed{x > 7.2135}$. **(E)**

34. [**Section 3.2**] We recognize X as inverse gamma with $\alpha = 3$, $\theta = 4$. Then $\mathbf{E}[X] = \frac{4}{3-1} = 2$ and $\mathbf{E}\left[X^2\right] = \frac{4^2}{(2)(1)} = 8$, so $\text{Var}(X) = 8 - 2^2 = 4$, and $\bar{X}$ has mean 2 and variance 0.04. The normal approximation gives

$$\Pr(\bar{X} < 2.5) = \Phi\left(\frac{2.5-2}{\sqrt{0.04}}\right) = \Phi(2.5) = \boxed{\mathbf{0.9938}} \qquad \textbf{(E)}$$

35. [**Lesson 8**] The 95th percentile of a normal distribution with parameters $\mu = 3$, $\sigma = 0.5$ is $3 + 1.645(0.5) = 3.8225$. Exponentiating, the 95th percentile of a lognormal distribution is $e^{3.8225} = \boxed{\mathbf{45.718}}$. **(C)**

9 78- 1- 63588- 206- 3

ASM Study Manual
for SOA Exam C