

AD-A161 322

PIECEWISE GEOMETRIC ESTIMATION OF A SURVIVAL FUNCTION
(U) FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS
G M MINNACK ET AL APR 85 FSU-STATISTICS-M700

1/1

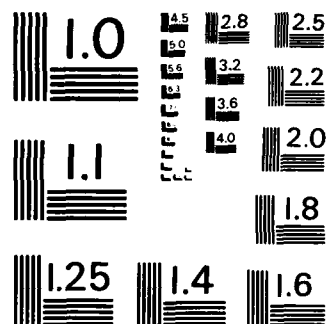
UNCLASSIFIED

AFOSR-TR-85-0953 F49620-02-K-0007

F/B 12/1

NL

										END			
										FILED			
										DTIC			



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR

2-3

(2)

Piecewise Geometric Estimation of a Survival Function

by

Gillian M. Mimmack and Frank Proschan

FSU Statistics Report M700
AFOSR Technical Report No. 176

April, 1985

The Florida State University
Department of Statistics
Tallahassee, Florida 32306

AD-A161 322

DTIC
ELECTE
NOV 21 1985

Research supported by the Air Force Office of Scientific Research, AFSC, USAR,
under Grant ~~XXXXXXXXXX~~.

F49620-82-K-0007 ✓

Key Words: Piecewise Geometric, piecewise Exponential, failure rate estimator,
life distribution estimator, incomplete data.

AMS Subject Classification Numbers: 62N05, 62G05.

DTIC FILE COPY

85 11 15 041

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER AFOSR-TR- AFOSR-176	2. NAME ACCESSION NO. AD-A161312	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and subtitle) Piecewise Geometric Estimation of a Survival Function		5. TYPE OF REPORT & PERIOD COVERED <i>technical</i>
6. AUTHOR(s) Gillian Mimmack and Frank Proschan		7. PERFORMING ORG. REPORT NUMBER
8. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Tallahassee, FL 32306 Florida State University		9. CONTRACT OR GRANT NUMBER(s) F47620- AFOSR 82-K-0007
10. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Bolling Air Force Base, D.C. 20332		11. PROGRAM ELEMENT, PROJECT, TASK AREA, & WORK UNIT NUMBERS QWCRF 2304 A5
12. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		13. REPORT DATE April, 1985
14. DISTRIBUTION STATEMENT (of this report) Unlimited		15. NUMBER OF PAGES 42
16. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from report)		17. SECURITY CLASS. (of this report)
18. SUPPLEMENTARY NOTES		19a. DECLASSIFICATION/DOWNGRADING SCHEDULE

Approved for public release;
distribution unlimited.

18. SUPPLEMENTARY NOTES

19. KEY WORDS

Piecewise Geometric, piecewise Exponential, failure rate estimator, life distribution estimator, incomplete data.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We describe a procedure that uses incomplete data to estimate failure rate and survival functions. Although the procedure is designed for discrete distributions, it applies in the continuous case also.

Our description is expository and therefore contains no proofs: they are provided by Mimmack (1985).

The procedure is based on the assumption of a piecewise constant failure rate. The resultant survival function estimator is a piecewise geometric function, denoted the Piecewise Geometric Estimator (PEGE). The PEGE is the discrete version of the piecewise exponential estimators proposed independently by Kitchin, Langberg and Proschan (1983) and Whittemore and Keller (1983), and it is a generalization of an estimator of Umholtz (1984) who considers complete data taken from an exponential distribution.

20. ABSTRACT (Continued)

The PEGE is consistent and asymptotically normal under conditions more general than those of the model of random censorship. Although the PEGE and the widely used Kaplan-Meier estimator (KME) are asymptotically equivalent and generally interlace, the PEGE is expected to perform better than the KME in terms of small sample properties.

The PEGE is attractive to users because it is computationally simple and realistic in that it decreases at every possible failure time: it therefore not only has the appearance of a survival function, but also provides a realistic estimate of the failure rate function. The KME, in contrast, is a step function.

Kaplan-Meier estimator (KME) is a step function.

Section 1: Introduction and Summary

The problem of estimating survival probabilities from incomplete data is well known in the fields of reliability, medicine, biometry and actuarial science. The general situation is described as follows. The variable of interest is the lifespan of some unit: the investigator wishes to estimate the probability of survival beyond any given time. To this end, n identical units are placed "on test". Each item is either observed until failure, resulting in an uncensored observation, or is removed from the test before failure, resulting in a censored observation. Thus the data available consist of a number of lifelengths and a number of truncated lifelengths: the statistical problem is to estimate the probability distribution of the lifelengths.

The various statistical approaches to the problem can generally be classified according to the restrictiveness of the model assumed and the type of information utilized. At one extreme are purely parametric procedures, which involve assuming that the underlying life distribution belongs to a specific parametric family. These procedures utilize interval information. The Bayesian estimator described by Susarla and Van Ryzin (1976) makes allowance for both parametric and nonparametric models: the type of information utilized depends on the assumptions about the prior distribution. As our approach to the problem is neither parametric nor Bayesian, we do not consider these procedures further but concentrate on nonparametric procedures.

Nonparametric procedures range in sophistication from the well-known actuarial estimator, which is a step function constructed from ordinal information alone, to the piecewise polynomial estimators of Whittemore and Keller (1983) that utilize interval information. The most widely used nonparametric estimators are those of Kaplan and Meier (1958) and Nelson (1969). These

Approved For	
NIS	☒
DIS	☐
SECRET	☐
Justification	
By	
Date	
Classification Codes	
1. ☐ 1. ☐ 2. ☐ 3. ☐ 4. ☐ 5. ☐ 6. ☐ 7. ☐ 8. ☐ 9. ☐ 10. ☐ 11. ☐ 12. ☐ 13. ☐ 14. ☐ 15. ☐ 16. ☐ 17. ☐ 18. ☐ 19. ☐ 20. ☐ 21. ☐ 22. ☐ 23. ☐ 24. ☐ 25. ☐ 26. ☐ 27. ☐ 28. ☐ 29. ☐ 30. ☐ 31. ☐ 32. ☐ 33. ☐ 34. ☐ 35. ☐ 36. ☐ 37. ☐ 38. ☐ 39. ☐ 40. ☐ 41. ☐ 42. ☐ 43. ☐ 44. ☐ 45. ☐ 46. ☐ 47. ☐ 48. ☐ 49. ☐ 50. ☐ 51. ☐ 52. ☐ 53. ☐ 54. ☐ 55. ☐ 56. ☐ 57. ☐ 58. ☐ 59. ☐ 60. ☐ 61. ☐ 62. ☐ 63. ☐ 64. ☐ 65. ☐ 66. ☐ 67. ☐ 68. ☐ 69. ☐ 70. ☐ 71. ☐ 72. ☐ 73. ☐ 74. ☐ 75. ☐ 76. ☐ 77. ☐ 78. ☐ 79. ☐ 80. ☐ 81. ☐ 82. ☐ 83. ☐ 84. ☐ 85. ☐ 86. ☐ 87. ☐ 88. ☐ 89. ☐ 90. ☐ 91. ☐ 92. ☐ 93. ☐ 94. ☐ 95. ☐ 96. ☐ 97. ☐ 98. ☐ 99. ☐ 100. ☐ 101. ☐ 102. ☐ 103. ☐ 104. ☐ 105. ☐ 106. ☐ 107. ☐ 108. ☐ 109. ☐ 110. ☐ 111. ☐ 112. ☐ 113. ☐ 114. ☐ 115. ☐ 116. ☐ 117. ☐ 118. ☐ 119. ☐ 120. ☐ 121. ☐ 122. ☐ 123. ☐ 124. ☐ 125. ☐ 126. ☐ 127. ☐ 128. ☐ 129. ☐ 130. ☐ 131. ☐ 132. ☐ 133. ☐ 134. ☐ 135. ☐ 136. ☐ 137. ☐ 138. ☐ 139. ☐ 140. ☐ 141. ☐ 142. ☐ 143. ☐ 144. ☐ 145. ☐ 146. ☐ 147. ☐ 148. ☐ 149. ☐ 150. ☐ 151. ☐ 152. ☐ 153. ☐ 154. ☐ 155. ☐ 156. ☐ 157. ☐ 158. ☐ 159. ☐ 160. ☐ 161. ☐ 162. ☐ 163. ☐ 164. ☐ 165. ☐ 166. ☐ 167. ☐ 168. ☐ 169. ☐ 170. ☐ 171. ☐ 172. ☐ 173. ☐ 174. ☐ 175. ☐ 176. ☐ 177. ☐ 178. ☐ 179. ☐ 180. ☐ 181. ☐ 182. ☐ 183. ☐ 184. ☐ 185. ☐ 186. ☐ 187. ☐ 188. ☐ 189. ☐ 190. ☐ 191. ☐ 192. ☐ 193. ☐ 194. ☐ 195. ☐ 196. ☐ 197. ☐ 198. ☐ 199. ☐ 200. ☐ 201. ☐ 202. ☐ 203. ☐ 204. ☐ 205. ☐ 206. ☐ 207. ☐ 208. ☐ 209. ☐ 210. ☐ 211. ☐ 212. ☐ 213. ☐ 214. ☐ 215. ☐ 216. ☐ 217. ☐ 218. ☐ 219. ☐ 220. ☐ 221. ☐ 222. ☐ 223. ☐ 224. ☐ 225. ☐ 226. ☐ 227. ☐ 228. ☐ 229. ☐ 230. ☐ 231. ☐ 232. ☐ 233. ☐ 234. ☐ 235. ☐ 236. ☐ 237. ☐ 238. ☐ 239. ☐ 240. ☐ 241. ☐ 242. ☐ 243. ☐ 244. ☐ 245. ☐ 246. ☐ 247. ☐ 248. ☐ 249. ☐ 250. ☐ 251. ☐ 252. ☐ 253. ☐ 254. ☐ 255. ☐ 256. ☐ 257. ☐ 258. ☐ 259. ☐ 260. ☐ 261. ☐ 262. ☐ 263. ☐ 264. ☐ 265. ☐ 266. ☐ 267. ☐ 268. ☐ 269. ☐ 270. ☐ 271. ☐ 272. ☐ 273. ☐ 274. ☐ 275. ☐ 276. ☐ 277. ☐ 278. ☐ 279. ☐ 280. ☐ 281. ☐ 282. ☐ 283. ☐ 284. ☐ 285. ☐ 286. ☐ 287. ☐ 288. ☐ 289. ☐ 290. ☐ 291. ☐ 292. ☐ 293. ☐ 294. ☐ 295. ☐ 296. ☐ 297. ☐ 298. ☐ 299. ☐ 300. ☐ 301. ☐ 302. ☐ 303. ☐ 304. ☐ 305. ☐ 306. ☐ 307. ☐ 308. ☐ 309. ☐ 310. ☐ 311. ☐ 312. ☐ 313. ☐ 314. ☐ 315. ☐ 316. ☐ 317. ☐ 318. ☐ 319. ☐ 320. ☐ 321. ☐ 322. ☐ 323. ☐ 324. ☐ 325. ☐ 326. ☐ 327. ☐ 328. ☐ 329. ☐ 330. ☐ 331. ☐ 332. ☐ 333. ☐ 334. ☐ 335. ☐ 336. ☐ 337. ☐ 338. ☐ 339. ☐ 340. ☐ 341. ☐ 342. ☐ 343. ☐ 344. ☐ 345. ☐ 346. ☐ 347. ☐ 348. ☐ 349. ☐ 350. ☐ 351. ☐ 352. ☐ 353. ☐ 354. ☐ 355. ☐ 356. ☐ 357. ☐ 358. ☐ 359. ☐ 360. ☐ 361. ☐ 362. ☐ 363. ☐ 364. ☐ 365. <	

3

Bv _____
Dated _____
_____ Series

A-1

A-1

The PEGE is attractive to users because it is computationally simple and realistic in that it decreases at every possible failure time: it therefore not only has the appearance of a survival function, but also provides a realistic estimate of the failure rate function. The KME, in contrast, is a step function.

estimators are also step functions constructed from ordinal information. Their properties are described by Efron (1967), Breslow and Crowley (1974), Petersen (1977), Aalen (1976, 1978), Kitchin, Langberg and Proschan (1983), Nelson (1972), Fleming and Harrington (1979), and Chen, Hollander and Langberg (1982).

One of the by-products of the estimation process is an estimate of the failure rate function: here, another issue is raised. It is evident that survival function estimators that are step functions do not provide useful failure rate function estimators: Miller (1981) mentions smoothing the Kaplan-Meier estimator for this reason and summarizes the development of other survival function estimators that may be obtained by considering a special case of the regression model of Cox (1972). These estimators generally correspond to failure rate function estimators that are step functions and utilize at most part (but not all) of the interval information contained in the data. Whittemore and Keller (1983) give several more refined failure rate function estimators that are step functions and utilize full interval information. They also describe even more complex estimators that utilize full interval information: however, these are not computationally convenient compared with their simpler estimators. It seems, from their work, that a successful rival of the Kaplan-Meier estimator should be only marginally more complex than it (so as to be computationally convenient and yet yield a useful failure rate function estimator) and also should utilize more than ordinal information.

In Section 2, we propose an estimator that not only provides a reasonable failure rate function estimator but also utilizes interval information. Moreover, it is computationally simple. Our estimator is a discrete counterpart of two versions of a continuous estimator proposed independently by Kitchin, Langberg and Proschan (1983) and Whittemore and Keller (1983). The motivation for the construction of our estimator is the same as that of the former authors, and our

model is the discrete version of theirs: in contrast, the latter authors assume the more restrictive model of random censorship and obtain their estimator by the method of maximum likelihood. This provides an alternative method of deriving our estimator.

The remaining sections are concerned with properties of our estimator. As this presentation is expository, proofs are omitted: Mimmack (1985) provides proofs.

In Section 3, we explore the asymptotic properties of our estimator under increasingly restrictive models. Our estimator is strongly consistent and asymptotically normal under conditions more general than those typically assumed.

Section 4 deals with the relationships among our estimator, the Kaplan-Meier estimator, and the above-mentioned estimator of Kitchin et al. and Whittemore and Keller. The section ends with an example using real data.

In Section 5, we continue the comparison of the new estimator and the Kaplan-Meier estimator: since the properties of the new estimator are expected to resemble those of its continuous counterparts, we discuss the implications of simulation studies designed to investigate the small sample behaviour of these estimators. We also present the results of a Monte Carlo pilot study designed to investigate the small sample properties of our estimator.

Section 2: Preliminaries.

In this section we formulate the problem in statistical terms and define our estimator.

Let X denote the lifelength of a randomly chosen unit, where X has distribution function G . Suppose that n identical items are placed on test. The resultant sample consists of the pairs $(Z_1, \delta_1), \dots, (Z_n, \delta_n)$, where Z_i represents

the time for which unit i is observed and δ_i indicates whether unit i fails while under observation or is removed from the test before failure. Symbolically, for $i = 1, \dots, n$, we have:

X_i = lifelength of unit i , where X_i has distribution G ,

Y_i = time to censorship of unit i ,

$Z_i = \min(X_i, Y_i)$,

$\delta_i = I(X_i \leq Y_i)$.

$(X_1, Y_1), \dots, (X_n, Y_n)$ are assumed to be independent random pairs. Elements of a pair X_i and Y_i , where $i = 1, \dots, n$, are not assumed to be independent.

We assume that the lifelength and censoring random variables are discrete. Let $X = \{x_1, x_2, \dots\}$ denote the set of possible values of X and $Y = \{y_1, y_2, \dots\}$ denote the union of the sets of possible values of $Y_1, Y_2, \dots$, where $Y \subseteq X$. The survival probabilities of interest are denoted $P(X > x_k)$, $k = 1, 2, \dots$, where $P(X > x_k) = \bar{G}(x_k) = 1 - G(x_k)$, $k = 1, 2, \dots$.

It is evident that this formulation differs from that of the model of random censorship which is generally assumed in the literature, and in particular, by Whittemore and Keller (1983). These authors assume that the lifelength and censoring random variables are continuous, that the corresponding pairs X_i and Y_i , where $i = 1, 2, \dots$, are independent, and that the censoring random variables are identically distributed. Although Kaplan and Meier (1958) assume only independence between corresponding lifelength and censoring random variables, Breslow and Crowley (1974), Petersen (1977), Aalen (1975, 1978), and others -- all of whom describe the properties of the Kaplan-Meier estimator -- assume also that the censoring random variables are identically distributed. Our formulation is the discrete counterpart of that of Kitchin, Langberg and Proschan (1983): likewise, our estimator is the discrete counterpart of theirs.

Before describing our estimator, we give the notation required.

Let n_1 be the random number of distinct uncensored observations in the sample and let $t_1 < t_2 < \dots < t_{n_1}$ denote these distinct observed failure times, with $t_0 \equiv 0$. Let n_2 be the random number of distinct censored observations in the sample and let $s_1 < s_2 < \dots < s_{n_2}$ denote these times, with $s_0 \equiv 0$.

Let D_i be the number of failures observed at time t_i :

$$D_i = \sum_{j=1}^n I(Z_j = t_i, \delta_j = 1) \quad \text{for } i = 1, \dots, n_1.$$

Let C_i be the number of censored observations equal to s_i :

$$C_i = \sum_{j=1}^n I(Z_j = s_i, \delta_j = 0) \quad \text{for } i = 1, \dots, n_2.$$

Let $\bar{F}_n(t) = 1 - F_n(t)$ denote the proportion of observations that exceed t :

$$\bar{F}_n(t) = \frac{1}{n} \sum_{j=1}^n I(Z_j > t) \quad \text{for } t \in [0, \infty).$$

Let $F_n^1(t)$ denote the proportion of failures observed at or before t :

$$F_n^1(t) = \frac{1}{n} \sum_{j=1}^n I(Z_j \leq t, \delta_j = 1) \quad \text{for } t \in [0, \infty).$$

Let T_i be a measure of the total time on test in the interval

$(t_{i-1}, t_i]$:

$$T_i = \#\{m: t_{i-1} < x_m \leq t_i\} (n\bar{F}_n(t_i) + D_i) \\ + \sum_{k: t_{i-1} < s_k \leq t_i} \#\{m: t_{i-1} < x_m \leq s_k\} C_k \quad \text{for } i = 1, \dots, n_1,$$

where $\#A$ denotes the cardinality of the set A .

(If failure and censoring random variables are lattice random variables, then T_i is the total time on test in $(t_{i-1}, t_i]$. In general, however, T_i increases by one unit whenever an item on test survives an interval of the form $(x_{j-1}, x_j]$, where $t_{i-1} < x_{j-1} < x_j \leq t_i$, irrespective of the distance between x_{j-1} and x_j .)

We now construct our estimator. Expressing the survival function $\bar{G}$ in terms of the failure rates $P(X = x_k | X \geq x_k)$, $k = 1, 2, \dots$, we have:

$$(2.1) \quad P(X > x_k) = \prod_{j=1}^k [1 - P(X = x_j | X \geq x_j)] \quad \text{for } k = 1, 2, \dots$$

It is evident from (2.1) that we may estimate our survival function at x_k from estimates of the failure rates at $x_1, x_2, \dots, x_k$. In the experimental situation, failures are not observed at all the times $x_1, x_2, \dots$ so specific information about the failure rates at many of the possible failure times is not available. Having observed failures at $t_1, t_2, \dots, t_{n_1}$, we find it simple to estimate the failure rates $P(X = t_i | X \geq t_i)$, $i = 1, \dots, n_1$. However, the question of how to estimate the failure rates at the intervening possible failure times requires special consideration.

One approach -- that of Kaplan and Meier (1958), Nelson (1969) and others -- is to estimate the failure rates at these intervening times as zero since no failures are observed then. However, not observing failures at some possible failure times may be a result of being in an experimental situation rather than evidence of very small failure rates at these times, so we discard this approach and consider nonzero estimates.

It is reasonable to assume that the underlying process possesses an element of continuity in that adjacent failure rates do not differ radically

from one another. Thus we consider using the estimate of the failure rate at t_i to estimate the failure rate at each of the possible failure times between t_{i-1} and t_i , where $i = 1, \dots, n_1$. We are therefore assuming that our approximating distribution has a constant failure rate between the times at which failures are observed -- that is:

$$(2.2) \quad \hat{P}(X = x_k | X \geq x_k) = \hat{q}_i \quad \text{for } t_{i-1} < x_k \leq t_i, i = 1, \dots, n_1,$$

where

$$q_i = P(X = t_i | X \geq t_i) \quad \text{for } i = 1, \dots, n_1.$$

Substituting (2.2) into (2.1), we obtain:

$$(2.3) \quad \hat{P}(X > x_k) = (1 - \hat{q}_i)^{\#\{m: t_{i-1} < x_m \leq x_k\}} \prod_{j=1}^{i-1} (1 - \hat{q}_j)^{\#\{m: t_{j-1} < x_m \leq t_j\}}$$

$$\text{for } t_{i-1} < x_k \leq t_i, i = 1, \dots, n_1.$$

We note that the property of having constant failure rate on X characterizes a family of geometric distributions defined on X . In particular, the failure rates $q_1, \dots, q_{n_1}$ identify n_1 geometric distributions $G_1, \dots, G_{n_1}$ defined on X . The survival functions, $\bar{G}_1, \dots, \bar{G}_{n_1}$, have the geometric form:

$$(2.4) \quad \bar{G}_i(x_k) = (1 - q_i)^k \quad \text{for } k = 1, 2, \dots \text{ and } i = 1, \dots, n_1.$$

Inspection of (2.3) and (2.4) reveals that our estimating function is constructed from the geometric survival functions $\bar{G}_1, \dots, \bar{G}_{n_1}$, where $\bar{G}_i$ is used in the interval $(t_{i-1}, t_i]$, $i = 1, \dots, n_1$. Consequently, the estimator (2.3) is called the Piecewise Geometric Estimator (PEGE).

It remains to define estimators of the failure rates $q_1, \dots, q_{n_1}$. This was originally done by separately obtaining the maximum likelihood estimators of the parameters of n_1 truncated geometric distributions: the procedure is outlined at a later stage because it utilizes the geometric structure of (2.3) and therefore provides further motivation for the name "PEGE". A more straightforward but less appealing approach is to obtain the maximum likelihood estimates of $q_1, \dots, q_{n_1}$ directly: denoting by L the likelihood of the sample, we have:

$$L = \left\{ \prod_{i=1}^{n_1} [P(X=t_i)]^{D_i} \right\} \left\{ \prod_{j=1}^{n_2} [P(X>s_j)]^{C_j} \right\}.$$

Substituting (2.3) into this expression and differentiating yields the unique maximum likelihood estimates

$$\hat{q}_i = D_i/T_i, \quad i = 1, \dots, n_1.$$

Substituting $\hat{q}_1, \dots, \hat{q}_{n_1}$ into (2.3), we finally obtain our estimator, formally defined as follows.

Definition 2.1. The Piecewise Geometric Estimator (PEGE) of the survival function of the lifelength random variable X is defined as follows:

$$\hat{P}(X > x_k) = \begin{cases} 1 & \text{for } x_k \leq 0 \text{ or } n_1 = 0. \\ (1 - D_i/T_i)^{\#\{m: t_{i-1} < x_m \leq x_k\}} \prod_{j=1}^{i-1} (1 - D_j/T_j)^{\#\{m: t_{j-1} < x_m \leq t_j\}} & \text{for } t_{i-1} < x_k \leq t_i, i = 1, \dots, n_1, n_1 > 0. \\ \prod_{j=1}^{n_1} (1 - D_j/T_j)^{\#\{m: t_{j-1} < x_m \leq t_j\}} & \text{for } x_k > t_{n_1}, n_1 > 0. \end{cases}$$

The alternative derivation of the PEGE emphasizes its geometric structure. it turns out that $\hat{q}_1, \dots, \hat{q}_{n_1}$ defined above are maximum likelihood estimators of the parameters of the truncated geometric distributions $G_1^*, \dots, G_{n_1}^*$ defined below.

For $i = 1, \dots, n_1$ we formulate the following definitions:

Let $N_i = \#\{m: t_{i-1} < x_m \leq t_i\}$ be the number of possible times of failure in the interval $(t_{i-1}, t_i]$ and let X_i^* be the number of possible times of failure that a unit of age t_{i-1} survives -- that is,

$X_i^* =$ number of trials to failure of a unit of age t_{i-1} ,

where the possible values of X_i^* are assumed to be $1, 2, \dots, N_i, N_i^+$. The distribution G_i^* of X_i^* is then given by:

$$\bar{G}_i^*(k) = (1-q_i)^k \quad \text{for } k = 1, 2, \dots, N_i,$$

$$\bar{G}_i^*(N_i^+) = 0.$$

The information available for estimating q_i consists of $n\bar{F}_n(t_{i-1})$ observations on X_i^* : of these, D_i are equal to N_i , $n\bar{F}_n(t_i)$ are equal to N_i^+ , and for all s_j in the interval $(t_{i-1}, t_i]$, C_j are equal to the number $\#\{m: t_{i-1} < x_m \leq s_j\}$. The resultant maximum likelihood estimator of q_i is precisely $\hat{q}_i$ defined above.

It is evident that the estimators $\hat{q}_1, \dots, \hat{q}_{n_1}$ have the form of the usual maximum likelihood estimator of a geometric parameter -- that is,

$$\text{Estimated failure rate} = \frac{\text{number of failures observed}}{\text{total time on test}}.$$

Moreover, we note that this is the form of the failure rate estimators in the intervals $(t_0, t_1], \dots, (t_{n_1}, \infty)$ defined for the Piecewise Exponential Estimator (PEXE) of Kitchen, Langberg and Proschan (1983). In terms of our notation (modified for continuity), the PEXE is defined as follows:

$$(2.5) \quad P^*(X > t) = \begin{cases} 1 & \text{for } t < 0 \text{ or } n_1 = 0, \\ \exp[-(t-t_{i-1})\hat{\lambda}_i] \prod_{j=1}^{i-1} \exp[-(t_j-t_{j-1})\hat{\lambda}_j] & \text{for } t_{i-1} < t \leq t_i, i = 1, \dots, n_1, n_1 > 0, \\ \prod_{j=1}^{n_1} \exp[-(t_j-t_{j-1})\hat{\lambda}_j] & \text{for } t > t_{n_1}, n_1 > 0, \end{cases}$$

where

$$\begin{aligned} \hat{\lambda}_i &= 1/\gamma_i & \text{for } i = 1, \dots, n_1, \\ \gamma_i &= \int_{t_{i-1}}^{t_i} n\bar{F}_n(u) du & \text{for } i = 1, \dots, n_1. \end{aligned}$$

For $i = 1, \dots, n_1$, λ_i is the failure rate in the interval $(t_{i-1}, t_i]$ and γ_i is the total time on test in this interval.

The PEXE is a piecewise exponential function because its construction is based on the assumption of constant failure rate between observed failures: just as a constant discrete failure rate characterizes a geometric distribution so a constant continuous failure rate characterizes an exponential distribution. Thus the PEGE is the discrete counterpart of the PEXE.

Returning to our introductory discussion about the desirable features of survival function estimators, we now compare the PEGE with other estimators in terms of these and other features.

First, the PEGE is intuitively pleasing because it reflects the continuity inherent in any life process. The Kaplan-Meier and other estimators that are step functions do not have this property.

Second, we note that the PEGE utilizes interval information from both censored and uncensored observations. It is therefore more sophisticated than the Kaplan-Meier and Nelson estimators. Moreover, none of the estimators of Whittemore and Keller utilizes more information than does the PEGE.

Third, the PEGE provides a simple, useful estimator of the failure rate function. While this estimator is naïve compared with the nonlinear estimators of Whittemore and Keller, the PEGE has the advantage of being simple enough to calculate by hand -- moreover it requires only marginally more computational effort than does the Kaplan-Meier estimator.

Regarding the applicability of the PEGE, we note that use of the PEGE is not restricted to discrete distributions because it can be easily modified by linear interpolation or by being defined as continuous wherever necessary. This is theoretically justified by the fact that the integer part of an exponential random variable has a geometric distribution: by defining the PEGE to be continuous, we are merely defining a variant of the PEXE. The properties of this estimator follow immediately from those of the PEXE.

Finally, apart from being intuitively pleasing, the form of the PEGE allows reasonable estimates of both the survival function and its percentiles. The Kaplan-Meier estimator is known to overestimate because of its step function form. We show in a later section that the PEGE tends to be less than the Kaplan-Meier estimator, and therefore the PEGE may be more accurate than the Kaplan-Meier estimator. Whittemore and Keller give some favourable indications in this respect. They define three survival function estimators that have constant failure rate between observed failure times. One of these is the PEXE, modified for ties in the data: the form of the failure rate estimator is the same as the form of the PEGE failure rate estimator -- specifically,

for $i = 1, \dots, n_1$:

$$(2.6) \text{ Estimated failure rate in } (t_{i-1}, t_i] = \frac{\text{number of failures observed at } t_i}{\text{total time on test during } (t_{i-1}, t_i]} .$$

The second of these estimators is defined instead on intervals of the form

$[t_{i-1}, t_i)$: for $i = 1, \dots, n_1$, the failure rate estimator has the form:

$$(2.7) \text{ Estimated failure rate in } [t_{i-1}, t_i) = \frac{\text{number of failures observed at } t_{i-1}}{\text{total time on test during } [t_{i-1}, t_i)} .$$

The third of these estimators is obtained from the average of the two failure rate estimators described by (2.6) and (2.7).

In a simulation study to investigate the small sample properties of these three estimators, Whittemore and Keller find that the first estimator tends to underestimate the survival function while the second tends to overestimate the survival function. From these results, we expect the PEGE to underestimate the survival function and its percentiles. Whittemore and Keller do not record further results for the first two estimators: however, they do indicate that, in terms of bias at extreme percentiles, variance and mean square error, the third estimator tends to be better than the Kaplan-Meier estimator.

The implications for the discrete version of the third estimator are that, in terms of bias, variance and mean square error, it will compare favourably with the Kaplan-Meier estimator. An unanswered question is whether the performance of this estimator is so superior to the performance of the PEGE as to warrant the additional computational effort required for the former.

Section 3: Asymptotic Properties of the PEGE.

This section treats the asymptotic properties of the PEGE and of the corresponding failure rate function estimator. The properties of primary

interest are those of consistency and asymptotic normality: secondary issues are asymptotic bias and asymptotic correlation.

Initially considering a very general model, we obtain the limiting function of the PEGE and show that the sequences $\{\hat{P}_n(X > x_k)\}_{k=1}^{\infty}$ and $\{\hat{P}_n(X = x_k | X \geq x_k)\}_{k=1}^{\infty}$ converge in distribution to Gaussian sequences. We then explore the effects of making various assumptions about the lifelength and censoring random variables. Under the most general model, the PEGE is not consistent and the failure rate estimators are not asymptotically uncorrelated: a sufficient condition for consistency is independence between corresponding lifelength and censoring random variables, and a sufficient condition for asymptotically independent failure rate estimators is that the censoring random variables be identically distributed. However, it is not necessary to impose both of these conditions in order to ensure both consistency and asymptotic independence of the failure rate estimators: relaxing the condition of independent lifelength and censoring random variables, we give conditions under which both desirable properties are obtained.

Before investigating the asymptotic properties of the PEGE, we describe the theoretical framework of the problem, give some notation, and present a preliminary result that facilitates the exploration of the asymptotic properties of the PEGE.

The probability space $(\Omega, \mathcal{A}, P)$ on which all of the lifelength and censoring random variables are defined is envisaged as the infinite product probability space that may be constructed in the usual way from the sequence of probability spaces corresponding to the sequence of independent random pairs (X_1, Y_1) , (X_2, Y_2) , Thus Ω consists of all possible sequences of pairs of outcomes corresponding to pairs of realizations in $X \times Y$: the first member of each pair

corresponds to failure at a particular time and the second member of each pair corresponds to censorship at a particular time -- that is, for each ω in Ω , $k = 1, 2, \dots$ and $j = 1, 2, \dots$:

$(X_i, Y_i)(\omega) = (X_i(\omega), Y_i(\omega)) = (x_k, y_j)$ if the i^{th} element of the infinite sequence ω is the pair of outcomes corresponding to failure at x_k and censorship at y_j .

The argument ω is omitted wherever possible.

Two conditions are imposed on the random pairs $(X_1, Y_1), (X_2, Y_2), \dots$:

(A1) There is a distribution function F such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P(Z_i \leq x_k) = F(x_k) \quad \text{for } k = 1, 2, \dots$$

(A2) There is a subdistribution function F^1 such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P(Z_i \leq x_k, \delta_i = 1) = F^1(x_k) \quad \text{for } k = 1, 2, \dots$$

It is evident that a sufficient condition for (A1) and (A2) is that the censoring random variables be identically distributed.

Definitions of symbols used in this section are given below. Assumptions (A1) and (A2) ensure the existence of the limits defined.

Let $P_{ki} = P(Z_i = x_k, \delta_i = 1)$ for $k = 1, 2, \dots$ and $i = 1, \dots, n$,

$R_{ki} = P(Z_i = x_k, \delta_i = 0)$ for $k = 1, 2, \dots$ and $i = 1, \dots, n$,

$P_k = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_{ki} = F^1(x_k) - F^1(x_{k-1})$ for $k = 1, 2, \dots$,

$R_k = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n R_{ki}$ for $k = 1, 2, \dots$.

The proposition below is fundamental: it asserts that, with probability one, as the sample size increases to infinity, at least one failure is observed at every possible value of the lifelength random variable. First, we need a definition.

Definition 3.1. Let $\Omega^* \subset \Omega$ be the set of infinite sequences which contain, for each possible failure time, at least one element corresponding to the outcome of observing failure at that time -- that is,

$$\Omega^* = \{\omega: (\forall k)(\exists n) X_n(\omega) = x_k, Y_n(\omega) \geq x_k\}.$$

Proposition 3.2. $P(\Omega^*) = 1$.

The proposition is proven by showing that the set of infinite sequences that do not contain at least one element corresponding to the outcome of observing failure at each possible failure time x_k has probability zero -- that is,

$$P\left(\lim_{n \rightarrow \infty} \bigcap_{i=1}^n \{X_i = x_k, Y_i \geq x_k\}^c\right) = 0 \text{ for } k = 1, 2, \dots$$

As the pairs $(X_1, Y_1), (X_2, Y_2), \dots$ are independent, this is equivalent to proving the following equality:

$$(3.1) \quad \lim_{n \rightarrow \infty} \prod_{i=1}^n (1 - P(X_i = x_k, Y_i \geq x_k)) = 0 \quad \text{for } k = 1, 2, \dots$$

Since $\prod_{i=1}^{\infty} (1 - p_i) = 0$ if and only if $\sum_{i=1}^{\infty} p_i = \infty$, where $\{p_i\}_{i=1}^{\infty}$ is any sequence of probabilities, and since (A2) implies that

$$\sum_{i=1}^{\infty} P(X_i = x_k, Y_i \geq x_k) = \infty \text{ for } k = 1, 2, \dots, \text{ we have (3.1).}$$

The importance of the preceding proposition lies in the simplifications it allows. It turns out that, on Ω^* and for n large enough, the PEGE may be expressed in simple terms of functions that have well-known convergence properties. Since $P(\Omega^*) = 1$, we need consider the asymptotic properties of the PEGE on Ω^* alone: these properties are easily obtained from those of the well-known functions.

In order to express the PEGE in this convenient way, we view the estimation procedure in an asymptotic context.

Suppose ω is chosen arbitrarily from Ω^* . Then, for each k , there is an N (depending on k and ω) such that $X_i(\omega) = x_j$ and $Y_i(\omega) \geq x_j$ for $j = 1, \dots, k$ and some $i \leq N$. Consequently, for $n \geq N$, the smallest k distinct observed failure times $t_1, \dots, t_k$ are merely $x_1, \dots, x_k$, and, since the set of possible censoring times is contained in X , the smallest k distinct observed times are also $x_1, \dots, x_k$. The first k intervals between observed failure times are simply $(0, x_1]$, $(x_1, x_2]$, $\dots$, $(x_{k-1}, x_k]$, and the function $T_{i,n}$ defined on the i^{th} interval is given by the number of units on test just before the end of the i^{th} interval -- that is,

$$(3.2) \quad T_{i,n} = n\bar{F}_n(x_i^-) = n\bar{F}_n(x_{i-1}) \quad \text{for } i = 1, \dots, k \text{ and } n \geq N.$$

Likewise, we express the function $D_{i,n}$ defined on the i^{th} interval in terms of the empirical subdistribution function F_n^1 as follows.

$$(3.3) \quad D_{i,n} = n[F_n^1(x_i) - F_n^1(x_{i-1})] \quad \text{for } i = 1, \dots, k \text{ and } n \geq N.$$

As the PEGE is a function of $D_{i,n}$ and $T_{i,n}$, it can be expressed in terms of the empirical functions F_n and F_n^1 . Specifically, on Ω^* , for any choice of k , there is an N such that

$$\hat{P}_n(X > x_k) = \prod_{i=1}^k \left(1 - \frac{F_n^1(x_i) - F_n^1(x_{i-1})}{\bar{F}_n(x_{i-1})} \right) \text{ for } n \geq N.$$

Consequently, taking the limit of each side and using Proposition 3.2, we have:

$$P \left[\lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \lim_{n \rightarrow \infty} \prod_{i=1}^k \left(1 - \frac{F_n^1(x_i) - F_n^1(x_{i-1})}{\bar{F}_n(x_{i-1})} \right) \text{ for } k = 1, 2, \dots \right] = 1.$$

In exploring the asymptotic behaviour of the PEGE, therefore, we consider the behaviour of the limiting sequence of the sequence

$$\left\{ \prod_{i=1}^k \left(1 - \frac{F_n^1(x_i) - F_n^1(x_{i-1})}{\bar{F}_n(x_{i-1})} \right) \right\}_{k=1}^{\infty}.$$

The proofs of the results that follow are omitted in the interest of brevity. The most general model we consider is that in which only conditions (A1) and (A2) are imposed. The following theorem identifies the limits of the sequences $\{\hat{P}_n(X=x_k | X \geq x_k)\}_{n=1}^{\infty}$ and $\{\hat{P}_n(X > x_k)\}_{n=1}^{\infty}$ for $k = 1, 2, \dots$ and establishes that the sequences $\{\hat{P}_n(X=x_k | X \geq x_k)\}_{k=1}^{\infty}$ and $\{\hat{P}_n(X > x_k)\}_{k=1}^{\infty}$ converge to Gaussian sequences.

Theorem 3.3.

(i) With probability 1,

$$\lim_{n \rightarrow \infty} \hat{P}_n(X=x_k | X \geq x_k) = \frac{F^1(x_k) - F^1(x_{k-1})}{\bar{F}(x_{k-1})} \text{ for } k = 1, 2, \dots.$$

(ii) With probability 1,

$$\lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \prod_{i=1}^k \left(1 - \frac{F^1(x_i) - F^1(x_{i-1})}{\bar{F}(x_{i-1})} \right) \text{ for } k = 1, 2, \dots.$$

(iii) Let $k_1, \dots, k_M$ be M arbitrarily chosen integers such that $k_1 < k_2 < \dots < k_M$. Then

$$(\hat{P}_n(X=x_{k_1} | X \geq x_{k_1}), \dots, \hat{P}_n(X=x_{k_M} | X \geq x_{k_M})) \text{ is } AN(\underline{\mu}^*, \frac{1}{n}\Sigma^*),$$

where

$$\underline{\mu}^* = (P_{k_1}/\bar{F}(x_{k_1-1}), \dots, P_{k_M}/\bar{F}(x_{k_M-1})).$$

$$\Sigma^* = \{\sigma_{qr}^*\}_{q=1, \dots, M, r=1, \dots, M}$$

$$\sigma_{qr}^* = P_{k_q} P_{k_r} \sum_{i=1}^{q-1} \sum_{j=1}^{r-1} (\sigma_{k_i, k_j} + \sigma_{k_M+k_i, k_j} + \sigma_{k_i, k_M+k_j} + \sigma_{k_M+k_i, k_M+k_j}) / (\bar{F}(x_{k_q-1}) \bar{F}(x_{k_r-1}))^2$$

$$+ P_{k_r} \sum_{i=1}^{r-1} (\sigma_{k_M+k_q, k_i} + \sigma_{k_M+k_q, k_M+k_i}) / ((\bar{F}(x_{k_r-1}))^2 \bar{F}(x_{k_q-1}))$$

$$+ P_{k_q} \sum_{i=1}^{q-1} (\sigma_{k_i, k_r} + \sigma_{k_M+k_i, k_r}) / (\bar{F}(x_{k_r-1}) (\bar{F}(x_{k_q-1}))^2)$$

$$+ \sigma_{k_M+k_q, k_r} / (\bar{F}(x_{k_q-1}) \bar{F}(x_{k_r-1})) \text{ for } q \leq r.$$

$$\sigma_{qr} = \begin{cases} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_{k_q, i} (1 - P_{k_q, i}) & \text{for } q = r, q = 1, \dots, M. \\ -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_{k_q, i} P_{k_r, i} & \text{for } q < r, q = 1, \dots, M, r = 1, \dots, M. \\ -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_{k_{q-M}, i} R_{k_r, i} & \text{for } q = M+1, \dots, 2M, r = 1, \dots, M. \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n R_{k_{q-M}, i} (1 - R_{k_{q-M}, i}) & \text{for } q = r, q = M+1, \dots, 2M. \\ -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n R_{k_{q-M}, i} R_{k_{r-M}, i} & \text{for } q < r, q = M+1, \dots, 2M, r = M+1, \dots, 2M. \end{cases}$$

(iv) Let $k_1, \dots, k_M$ be M arbitrarily chosen integers such that $k_1 < k_2 < \dots < k_M$. Then

$$(\hat{P}_n(X > x_{k_1}), \dots, \hat{P}_n(X > x_{k_M})) \text{ is AN}(\mu^{**}, \frac{1}{n}\Sigma^{**}),$$

where

$$\mu^{**} = \left(\prod_{i=1}^{k_1} (1 - P_i / \bar{F}(x_{i-1})) , \dots , \prod_{i=1}^{k_M} (1 - P_i / \bar{F}(x_{i-1})) \right).$$

$$\Sigma^{**} = \{\sigma_{qr}^{**}\}_{q=1, \dots, M, r=1, \dots, M}$$

$$\sigma_{qr}^{**} = \prod_{i=1}^{k_q} (1 - P_i / \bar{F}(x_{i-1})) \prod_{j=1}^{k_r} (1 - P_j / \bar{F}(x_{j-1})) \sum_{\ell=1}^{k_q} \sum_{m=1}^{k_r} \sigma_{\ell m}^* / [(1 - P_\ell / \bar{F}(x_{\ell-1})) (1 - P_m / \bar{F}(x_{m-1}))]$$

for $q \leq r$.

It is evident from the theorem above that the PEGE is a strongly consistent estimator of the underlying survival function if and only if

$$(3.4) \quad \frac{F^1(x_k) - F^1(x_{k-1})}{\bar{F}(x_{k-1})} = \frac{P(X=x_k)}{P(X \geq x_k)} \quad \text{for } k = 1, 2, \dots$$

The theorems below give conditions under which this equality holds. As for correlation, it is evident from the structure of the PEGE that any two elements of the sequence $\{\hat{P}_n(X > x_k)\}_{k=1}^{\infty}$ are correlated. Consequently the matrix Σ^{**} cannot be reduced to a diagonal matrix under even the most stringent conditions. However it turns out that, under certain conditions, the asymptotic correlation between pairs of the sequence $\{\hat{P}_n(X=x_k | X \geq x_k)\}_{k=1}^{\infty}$ is zero -- that is, Σ^* is a diagonal matrix.

The following theorem shows that independence between lifelength and censoring random variables results in strongly consistent (and therefore

asymptotically unbiased) estimators. However any pair in the sequence $\{\hat{P}_n(X=x_k | X \geq x_k)\}_{k=1}^{\infty}$ is asymptotically correlated in this case. Since the matrices Σ^* and Σ^{**} have the same form as in the theorem above, they are not explicitly defined below.

Theorem 3.4.

Suppose

(i) the random variables X_i and Y_i are independent for $i = 1, 2, \dots$

and

(ii) there is a distribution function H such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P(Y_i \leq x_k) = H(x_k) \quad \text{for } k = 1, 2, \dots$$

Then

$$(iii) \quad F^1(x_k) = \sum_{i=1}^k P(X=x_i) \bar{H}(x_{i-1}) \text{ and } \bar{F}(x_k) = P(X > x_k) \bar{H}(x_k) \quad \text{for } k = 1, 2, \dots$$

(iv) With probability 1,

$$\lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \bar{G}(x_k) \quad \text{for } k = 1, 2, \dots$$

$$(v) \quad (\hat{P}_n(X=x_{k_1} | X \geq x_{k_1}), \dots, \hat{P}_n(X=x_{k_M} | X \geq x_{k_M})) \text{ is } AN(\underline{\mu}^*, \frac{1}{n} \Sigma^*),$$

where $k_1 < k_2 < \dots < k_M$ are arbitrarily chosen integers and

$$\underline{\mu}^* = (P(X=x_{k_1} | X \geq x_{k_1}), \dots, P(X=x_{k_M} | X \geq x_{k_M})).$$

$$(vi) \quad (\hat{P}_n(X > x_{k_1}), \dots, \hat{P}_n(X > x_{k_M})) \text{ is } AN(\underline{\mu}^{**}, \frac{1}{n} \Sigma^{**}),$$

where $k_1 < k_2 < \dots < k_M$ are arbitrarily chosen integers and

$$\underline{\mu}^{**} = (P(X > x_{k_1}), \dots, P(X > x_{k_M})).$$

A sufficient condition for (A1), (A2) and assumption (ii) of the preceding theorem is that the censoring random variables be identically distributed. In this case the failure rate estimators are asymptotically independent and the matrix Σ^{**} is somewhat simplified. The conditions of the following corollary define the model of random censorship widely assumed in the literature.

Corollary 3.5.

Suppose

- (i) the random variables X_i and Y_i are independent for $i = 1, 2, \dots$
and
(ii) the random variables $Y_1, Y_2, \dots$ are identically distributed.

Then

- (iii) with probability 1, $\lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \bar{G}(x_k)$ for $k = 1, 2, \dots$,
(iv) $(\hat{P}_n(X = x_{k_1} | X \geq x_{k_1}), \dots, \hat{P}_n(X = x_{k_M} | X \geq x_{k_M}))$ is $AN(\underline{\mu}^*, \frac{1}{n} \Sigma^*)$,

where

$$\underline{\mu}^* = (P(X = x_{k_1} | X \geq x_{k_1}), \dots, P(X = x_{k_M} | X \geq x_{k_M})),$$

$$\Sigma^* = \{\sigma_{qr}^*\}_{q=1, \dots, M}^{r=1, \dots, M}$$

$$\sigma_{qr}^* = \begin{cases} P(X = x_{k_q} | X \geq x_{k_q}) P(X > x_{k_q} | X \geq x_{k_q}) / \bar{F}(x_{k_q-1}) & \text{for } q = r. \\ 0 & \text{for } q \neq r. \end{cases}$$

- (v) $(\hat{P}_n(X > x_{k_1}), \dots, \hat{P}_n(X > x_{k_M}))$ is $AN(\underline{\mu}^{**}, \frac{1}{n} \Sigma^{**})$,

where

$$\underline{\mu}^{**} = (P(X > x_{k_1}), \dots, P(X > x_{k_M})),$$

$$\Sigma^{**} = \{\sigma_{qr}^{**}\}_{\substack{q=1, \dots, M \\ r=1, \dots, M}}$$

$$\sigma_{qr}^{**} = P(X > x_{k_q}) P(X > x_{k_r}) \sum_{i=1}^k P(X = x_i | X \geq x_i) / [\bar{F}(x_{i-1}) P(X > x_i | X \geq x_i)] \text{ for } q \leq r.$$

Having dealt with the most restrictive case in which the lifelength and censoring random variables are assumed to be independent, we now consider relaxing this condition. It turns out that independence between corresponding lifelength and censoring random variables is not necessary for asymptotic independence between pairs of the sequence of failure rate estimators: if the censoring random variables are assumed to be identically distributed but not necessarily independent of the corresponding lifelength random variables, then the failure rate estimators are asymptotically independent. However both the survival function and failure rate estimators are asymptotically biased. The following corollary expresses these facts formally.

Corollary 3.6.

Suppose

- (i) the random variables $Y_1, Y_2, \dots$ are identically distributed.

Then

- (ii) $P_k = P(Z = x_k, \delta = 1)$ and $\bar{F}(x_k) = P(Z > x_k)$ for $k = 1, 2, \dots$,
 (iii) $(\hat{P}_n(X = x_{k_1} | X \geq x_{k_1}), \dots, \hat{P}_n(X = x_{k_M} | X \geq x_{k_M}))$ is $AN(\underline{\mu}^*, \frac{1}{n} \Sigma^*)$,

where

$$\underline{\mu}^* = (P_{k_1} / \bar{F}(x_{k_1-1}), \dots, P_{k_M} / \bar{F}(x_{k_M-1})),$$

$$\Sigma^* = \{\sigma_{ij}^*\}_{i=1, \dots, M}^{j=1, \dots, M}$$

$$\sigma_{ij}^* = \begin{cases} P_{k_i} (1 - P_{k_i} / \bar{F}(x_{k_i-1})) / (\bar{F}(x_{k_i-1}))^2 & \text{for } i = j. \\ 0 & \text{for } i \neq j. \end{cases}$$

$$(iv) \quad (\hat{P}_n(X > x_{k_1}), \dots, \hat{P}_n(X > x_{k_M})) \text{ is } AN(\mu^{**}, \frac{1}{n} \Sigma^{**}),$$

where

$$\mu^{**} = \left(\prod_{i=1}^{k_1} (1 - P_i / \bar{F}(x_{i-1})) , \dots , \prod_{i=1}^{k_M} (1 - P_i / \bar{F}(x_{i-1})) \right),$$

$$\Sigma^{**} = \{\sigma_{j\ell}^{**}\}_{j=1, \dots, M}^{\ell=1, \dots, M}$$

$$\sigma_{j\ell}^{**} = \prod_{i=1}^{k_j} (1 - P_i / \bar{F}(x_{i-1})) \prod_{m=1}^{k_\ell} (1 - P_m / \bar{F}(x_{m-1})) \sum_{r=1}^{k_j} P_r / [(\bar{F}(x_{r-1}))^2 (1 - P_r / \bar{F}(x_{r-1}))]$$

for $j \leq \ell$.

The corollaries above give sufficient (rather than necessary) conditions for the two desirable properties of (i) consistency and (ii) asymptotic independence between pairs of the sequence of failure rate estimators $\{\hat{P}_n(X=x_k | X \geq x_k)\}_{k=1}^\infty$. The final corollaries show that both of the conditions of Corollary 3.5 are not necessary for these two desirable properties: the conditions specified in these corollaries are not so stringent as to require that corresponding censoring and lifelength random variables be independent (as in Corollary 3.5), but rather that they be related in a certain way.

Corollary 3.7.

If the random variables $Y_1, Y_2, \dots$ are identically distributed, then with probability 1,

$$\lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \bar{G}(x_k) \quad \text{for } k = 1, 2, \dots$$

if and only if

$$P(Y_i \geq x_k | X = x_k) = P(Y_i \geq x_k | X \geq x_k) \quad \text{for } k = 1, 2, \dots \text{ and } i = 1, 2, \dots$$

Corollary 3.8.

Suppose

- (i) the random variables $Y_1, Y_2, \dots$ are identically distributed
and
(ii) $P(Y_i \geq x_k | X = x_k) = P(Y_i \geq x_k | X \geq x_k)$ for $k = 1, 2, \dots$ and $i = 1, 2, \dots$

Then

$$(iii) \quad (\hat{P}_n(X = x_{k_1} | X \geq x_{k_1}), \dots, \hat{P}_n(X = x_{k_M} | X \geq x_{k_M})) \text{ is } AN(\underline{\mu}^*, \frac{1}{n} \Sigma^*),$$

where

$$\underline{\mu}^* = (P(X = x_{k_1} | X \geq x_{k_1}), \dots, P(X = x_{k_M} | X \geq x_{k_M})),$$

$$\Sigma^* = \{\sigma_{ij}^*\}_{i=1, \dots, M}^{j=1, \dots, M}$$

$$\sigma_{ij}^* = \begin{cases} P(X = x_{k_i} | X \geq x_{k_i}) P(X > x_{k_i} | X \geq x_{k_i}) / \bar{F}(x_{k_i-1}) & \text{for } i = j. \\ 0 & \text{for } i \neq j. \end{cases}$$

$$(iv) \quad (\hat{P}_n(X > x_{k_1}), \dots, \hat{P}_n(X > x_{k_M})) \text{ is } AN(\underline{\mu}^{**}, \frac{1}{n} \Sigma^{**}),$$

where

$$\underline{\mu}^{**} = (P(X > x_{k_1}), \dots, P(X > x_{k_M})),$$

$$\Sigma^{**} = \{\sigma_{j\ell}^{**}\}_{j=1, \dots, M}^{\ell=1, \dots, M}$$

$$\sigma_{j\ell}^{**} = P(X > x_{k_j}) P(X > x_{k_\ell}) \sum_{i=1}^{k_j} P(X = x_i | X \geq x_i) / [\bar{F}(x_{i-1}) P(X > x_i | X \geq x_i)] \text{ for } j \leq \ell.$$

The last two corollaries are of special interest because they deal with consistency and asymptotic independence in the case of dependent lifelength and censoring random variables -- a situation that is not generally considered despite its obvious practical significance. Desu and Narula (1977), Langberg, Proschan and Quinzi (1980) and Kitchin, Langberg and Proschan (1983) consider the continuous version of the model specified in the last two corollaries.

The condition specifying the relationship between lifelength and censoring random variables is in fact a mild one: re-expressing it, we have the following condition.

$$\frac{P(X=x_k | X \geq x_k, Y_i \geq x_k)}{P(X \geq x_k | X \geq x_k, Y_i \geq x_k)} = \frac{P(X=x_k)}{P(X \geq x_k)} \quad \text{for } k = 1, 2, \dots \quad \text{and } i = 1, 2, \dots$$

This condition specifies that the failure rate among those under observation at any particular age is the same as the failure rate of the whole population of that age. It is evident both intuitively and mathematically that this is a fundamental assumption inherent in the process of estimating a life distribution from incomplete data: if this assumption could not be made, the data available would be deemed inadequate for estimating the life distribution. Formally, it is the fact that the condition is both necessary and sufficient for consistency that indicates that it is minimal for the estimation process. It is clear, therefore, that the last two corollaries play an important role in estimation in the context of a practical model more general than the statistically convenient, but unnecessarily restrictive, model of random censorship.

Section 4: The PEGE Compared with Rivals.

In Section 1 we motivate the construction of the PEGE by describing some desirable properties of nonparametric survival function estimators and then mentioning that the commonly used estimator of Kaplan and Meier (1958) does

not fare well in terms of these properties. We now compare the PEGE with the Kaplan-Meier estimator.

We begin with the most obvious desirable features of survival function estimators and then consider statistical and mathematical properties. In comparing the two estimators, we find that the issue of continuity arises and that the PEXE enters the comparison. The section ends with an example using real data. The subsequent section continues the comparison: we discuss the results of simulation studies.

The Kaplan-Meier estimator (KME) of the survival function of the life-length random variable X is defined as follows:

$$\tilde{P}(X > t) = \begin{cases} 1 & \text{for } n_1 = 0 \text{ or } t < t_1, n_1 \geq 1. \\ \prod_{j=1}^{i-1} (1 - D_j / n\bar{F}_n(t_j^-)) & \text{for } t_{i-1} \leq t < t_i, i = 2, \dots, n_1, n_1 \geq 2. \\ \prod_{j=1}^{n_1} (1 - D_j / n\bar{F}_n(t_j^-)) & \text{for } t \geq t_{n_1}, n_1 \geq 1. \end{cases}$$

To the prospective user of a survival function estimator, two fundamental questions are, firstly, does the estimating function have the appearance of a survival function, and secondly, is it easy to compute?

Considering the second question first, we observe that calculating the PEGE involves only marginally more effort than calculating the KME. Therefore, both estimators are accessible to users equipped with only hand calculators.

The first question is a deeper one. If the sample is small or if there are many ties among the uncensored observations in a large sample, the KME has only a few steps and consequently appears unrealistic. The PEGE, in contrast, reflects the continuity inherent in any life process by decreasing

at every possible failure time, not only at the observed failure times. As the number of distinct uncensored observations increases, both the PEGE and the KME become smoother: the many steps of the KME do allow it the appearance of a survival function, except possibly at the right extreme -- there is no way of extrapolating very far beyond the range of observation if the KME is used. (There are several ways of extrapolating from the PEGE.) At face value, therefore, the PEGE is at least as attractive as the KME.

A related consideration is whether the estimator provides a realistic estimate of the failure rate function. The KME, being a step function, does not. The seriousness of this omission becomes more apparent when the KME failure rate function is examined from a user's point of view: if an item of age t has a (perhaps large) chance of failing at its age, then claiming that a slightly older (or slightly younger) item cannot fail at its age seems unreasonable, particularly when it becomes evident that the claim is made on the grounds that none of the items on test happened to fail just after (or just before) time t . Intuitively -- or from a frequentist's point of view -- the very fact that one of the items on test failed at time t makes it less likely that another item in the sample will fail soon after t because the observed failure times should be scattered along the appropriate range according to the distribution function. Clearly, then, the gaps between observed failure times are a result of the fact that the sample is finite and are not indicative of zero (or very small) failure rates.

The PEGE, on the other hand, is constructed so that a failure at time t , say, affects the failure rate in the gap before t . Thus the PEGE compensates for the lack of observations at the possible (but unobserved) failure times. The resultant failure rate function, being a step function, is still naïve, but

it does at least take into account the continuity of life processes and it does provide reasonable estimates of the failure rates at all possible failure times.

A more aesthetic -- but none the less important -- issue is that of information loss. Here the PEGE is again at an advantage. Although interval information about the uncensored observations is used in spacing out the successive values of the KME, the failure rate estimators utilize only ordinal information. Moreover, the only information utilized from the censored observations is their positioning relative to the uncensored observations. Thus the information lost by the KME is of both the ordinal and interval types. In contrast, the PEGE failure rate estimators use interval information from all the observations: in particular, the positions of censored observations are taken into account precisely. In terms of information usage, then, the PEGE is far more desirable than the KME.

An apparently attractive feature of the KME is that its values are invariant under monotone transformation of the scale of measurement. The PEGE is not invariant under even linear transformation. However, in the light of the discussion about information loss, it is evident that the KME's invariance, and the PEGE's lack thereof, are results of their levels of sophistication rather than properties that can be used for comparison.

Having noted that the step function form of the KME is not pleasing, we now point out that it is also responsible for a statistical defect, namely, that the KME tends to overestimate the underlying survival function and its percentiles. The fact that the KME consistently overestimates suggests that its form is inappropriate. Some indications about the bias of the PEGE are given by considering the relationship between the PEGE and the KME.

Under certain conditions (for example, if there are no ties among the uncensored observations), the PEGE and the KME interlace: within each failure interval, the PEGE crosses the KME once from above. This is not true in

general, however. It turns out that the KME may have large steps in the presence of ties. In the case of the PEGE, however, the effect of the ties is damped and the PEGE decreases slowly relative to the KME. In general, therefore, it is possible to relate the PEGE and the KME only in a one-sided fashion: specifically, the PEGE at any observed failure time is larger than the KME at that time. Examples have been constructed to show that, in general, the PEGE cannot be bounded from above by the KME. The following theorem relates $\hat{P}$ (the PEGE) and $\tilde{P}$ (the KME).

Theorem 4.1.

- (i) $\hat{P}(X > t_i) \geq \tilde{P}(X > t_i)$ for $i = 1, \dots, n_1$.
- (ii) If $n_n^-(t_{j-1}) / (n_n^-(t_{j-1}) + W_{j-1}) \leq D_j / D_{j-1}$ for $j = 2, \dots, i$,

where W_j denotes the number of censored observations at t_j for $j = 1, \dots, n_1$, then $\hat{P}(X > t_i) \leq \tilde{P}(X > t_i)$ for $i = 1, \dots, n_1$.

It is evident that the condition in (ii) is met if there are no ties among the uncensored observations: this is likely if the sample is small. From the relationships in the theorem, we infer that the bias of the PEGE is likely to be of the same order of magnitude as that of the KME. Further indications about bias are given later.

Having considered some of the practical and physical features of the PEGE and the KME, we turn briefly to asymptotic properties -- briefly because the PEGE and the KME are asymptotically equivalent -- that is,

$$P[(\forall k) \lim_{n \rightarrow \infty} \hat{P}_n(X > x_k) = \lim_{n \rightarrow \infty} \tilde{P}_n(X > x_k)] = 1.$$

The practical implication of this is that there is little reason for strong preference of either the PEGE or the KME if the sample is very large.

51

We now compare the models assumed in using the KME and the PEGE. In the many studies of the KME, the most general model includes the assumption of independence between corresponding life and censoring random variables. Our most general model does not include this assumption. However this difference is not important because the assumption of independence is used only to facilitate the derivation of certain asymptotic properties of the KME: in fact, the definition of the KME does not depend on this assumption, and the KME and the PEGE are asymptotically equivalent under the conditions of the most general model of the PEGE. Therefore this assumption is not necessary for using the KME.

The other difference between the models assumed is that the PEGE is designed specifically for discrete life and censoring distributions while the Kaplan-Meier model makes no stipulations about the supports of these distributions. However, distinguishing between continuous and discrete random variables in this context is merely a statistical convention -- in fact, time to occurrence of some event is always measured along a continuous scale, and the set of observable values is always countable because it is defined by the precision of measurement. Since the process of estimating a life distribution requires measurements, it always entails the assumption of a discrete distribution: whether the support of the estimator is continuous or discrete depends on the way the user perceives the scale of measurement. In practice, therefore, there are no differences between the models underlying the PEGE and the KME: the PEGE is appropriate whenever the KME is, and vice versa.

Having pointed out that the PEGE may be used for estimating continuous survival functions, and having introduced the PEXE as the continuous counterpart of the PEGE, we compare the two. First we note that the PEXE is the

continuous version of the PEGE because the construction of each is based on the assumption of constant failure rate between distinct observed failure times. The forms of the estimators differ because of the difference in the ways of expressing discrete and continuous survival functions in terms of failure rates. The PEGE and the PEXE are equally widely applicable since a minor modification of the PEXE can be made to allow for ties. (This estimator is defined in Whittemore and Keller (1983).)

The relationship between the PEGE and the modified PEXE, and their positioning relative to the KME, is summarized by the following theorem and the succeeding relationship.

Theorem 4.2.

Let $P^{**}(X>t)$ denote the modified PEXE of the survival probability $P(X>t)$ for $t > 0$.

- (i) $\hat{P}(X>t) < P^{**}(X>t)$ for $t > 0$.
- (ii) If $n\bar{F}_n(t_{j-1})/(n\bar{F}_n(t_{j-1}) + W_{j-1}) \leq D_j/D_{j-1}$ for $j = 2, \dots, i$, where W_j denotes the number of censored observations at t_j for $j = 1, \dots, n_1$, then $P^{**}(X>t_i) \leq \tilde{P}(X>t_{i-1})$ for $i = 1, \dots, n_1$.

From Theorems 4.1(i) and 4.2(i), we have:

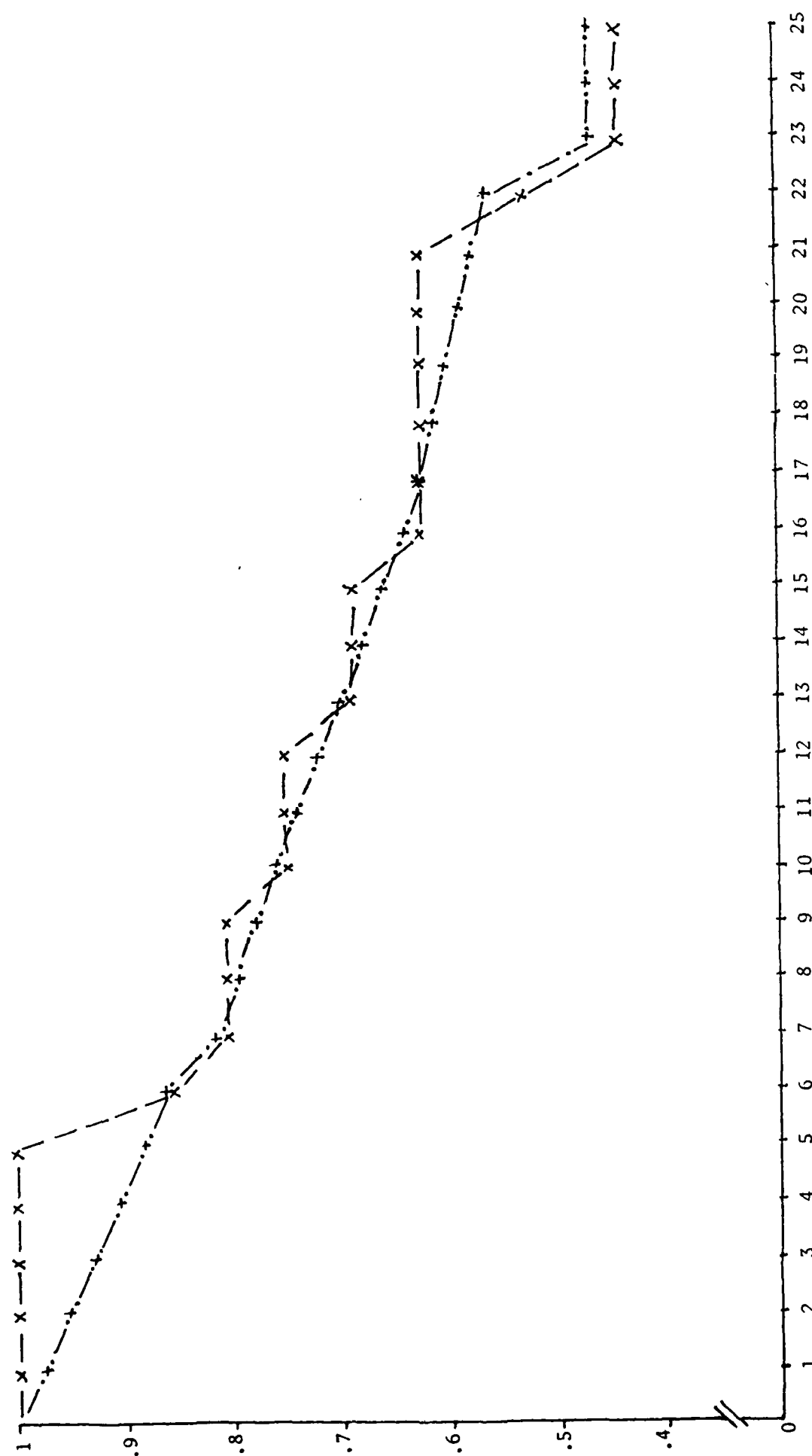
$$\tilde{P}(X>t_i) \leq \hat{P}(X>t_i) < P^{**}(X>t_i) \quad \text{for } i = 1, \dots, n_1.$$

Consequently, if the condition in (ii) above is met (as it is when there are no ties among the uncensored observations), both the PEGE and the PEXE interlace with the KME: in each interval of the form $(t_{i-1}, t_i]$, the PEGE and the PEXE cross the KME once from above. Practical experience suggests that the condition in (ii) above is not a stringent one: even though this condition is violated in many of the data sets considered to date, the PEGE and the

PEXE still interlace with the KME in the manner described. Another indication from practical experience is that the difference between the PEXE and the PEGE is negligible, even in small samples.

Finally, we present an example using the data of Freireich et al. (1963). The data are the remission times of 21 leukemia patients who have received 6 MP (a mercaptopurine used in the treatment of leukemia). The ordered remission times in weeks are: 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+. The PEGE and the KME are presented in Figure 4.1. (Since the PEGE and the PEXE differ by at most .09, only the PEGE appears.) The graphs illustrate the smoothness of the PEGE in contrast with the jagged outline of the KME. The KME and the PEGE interlace even though the condition in Theorems 4.1(ii) and 4.2(ii) is violated. Since the PEGE is only slightly above the KME at the observed failure times and the PEGE crosses the KME early in each failure interval, the KME is considerably larger than the PEGE by the end of each interval. This behaviour is typical. We infer that the PEGE certainly does not overestimate: it may even tend to underestimate.

We conclude that the PEGE (and the modified PEXE) have significant advantages over the KME, particularly in the cases of large samples containing many ties and small samples. It is only in the case of a large sample spread over a large range that the slight increase in computational effort required for the PEGE might merit using the KME because the PEGE and the KME are likely to be very similar.



X—X KME
+ - + PEGE

TIME TO REMISSION (WEEKS)

FIGURE 4.1. THE KME AND THE PEGE FOR LEUKEMIA DATA OF FREIREICH ET AL (1963).

Section 5: Small Sample Properties of the PEGE.

In this section we give some indications of the small sample properties of the PEGE by considering three simulation studies. In the first study, Kitchin (1980) compares the small sample properties of the PEXE with those of the KME. In the second study, Whittemore and Keller (1983) consider the small sample behaviour of a number of estimators: we extract the results for the KME and a particular version of the PEXE. In the third study, we make a preliminary comparison of the KME and the PEGE. We expect the behaviour of the piecewise exponential estimators to resemble that of the PEGE because piecewise exponential estimators are continuous versions of the PEGE and, moreover, piecewise exponential estimators and the PEGE are similar when the underlying life distribution is continuous.

The piecewise exponential estimator considered by Whittemore and Keller is denoted $\hat{F}_{Q_4}$. It is constructed by averaging the PEXE failure rate function estimator with a variant of the PEXE failure rate function estimator -- that is, $\hat{F}_{Q_4}$ is the same as the PEXE except that the PEXE failure rate estimators $\lambda_1^-, \dots, \lambda_{n_1}^-$ are replaced by the failure rate estimators $\lambda_1^*, \dots, \lambda_{n_1}^*$ defined as follows:

$$\lambda_i^* = \frac{1}{2}(\lambda_i^- + \lambda_{i-1}^+) \quad \text{for } i = 1, \dots, n_1,$$

where

$$\lambda_i^- = D_i / \text{Total time on test in } (t_{i-1}, t_i] \quad \text{for } i = 1, \dots, n_1,$$

$$\lambda_i^+ = D_i / \text{Total time on test in } [t_i, t_{i+1}) \quad \text{for } i = 0, \dots, n_1-1,$$

$$\lambda_{n_1}^+ = \begin{cases} D_{n_1} / \text{Total time on test in } [t_{n_1}, \infty) & \text{if } \max_{1 \leq i \leq n} Z_i > t_{n_1} \\ 0 & \text{otherwise.} \end{cases}$$

Although Whittemore and Keller include in their study the two estimators $\hat{F}_{Q_1}$ and $\hat{F}_{Q_2}$ constructed from $\lambda_1^-, \dots, \lambda_{n_1}^-$ and $\lambda_1^+, \dots, \lambda_{n_1}^+$ respectively, they present the results for the hybrid estimator $\hat{F}_{Q_4}$ alone because they find that $\hat{F}_{Q_1}$ tends to be negatively biased and $\hat{F}_{Q_2}$ tends to be positively biased.

The same model is assumed in all three studies. The model is that of random censorship: corresponding life and censoring random variables are independent and the censoring random variables are identically distributed. Whittemore and Keller generate 200 samples in each of the $6 \times 3 \times 4 = 72$ situations that result from considering six life distributions (representing failure rate functions that are constant, linearly increasing, exponentially increasing, decreasing, U-shaped, and discontinuous), three levels of censoring ($P(Y < X) \approx 0, .55, .76$), and four sample sizes ($n=10, 25, 50, 100$). Kitchin obtains 1000 samples in each of a variety of situations: he considers four life distributions (Exponential, Weibull with parameter 2, Weibull with parameter $\frac{1}{2}$, and Uniform), three levels of censoring ($P(Y < X) = 0, .5, .67$), and four sample sizes ($n=10, 20, 50$). Kitchin's study is broader than that of Whittemore and Keller in that Kitchin considers Exponential, Weibull and Uniform censoring distributions while Whittemore and Keller consider only Exponential censoring distributions. Kitchin apparently produces the greater variety of sampling conditions because his results vary slightly according to the model, while Whittemore and Keller find so much similarity in the results from the various distributions that they record only the results from the Weibull distribution.

The conclusions we draw from the two studies are similar. Regarding mean squared error (MSE), both Kitchin and Whittemore and Keller find that, in general:

(i) The MSE of the exponential estimator is smaller than that of the KME.

(ii) As the level of censoring increases, the increase in the MSE is smaller for the exponential estimator than for the KME.

Kitchin reports that (i) and (ii) are not always true of the PEXE and the KME: the exceptional cases occur in the tails of the distributions.

The conclusions about bias are not so straightforward. Whittemore and Keller find that the PEXE tends to be negatively biased while Kitchin reports that the bias of the PEXE is a monotone increasing function of time: examining his figures, we find that the bias tends to be near zero at some point between the 40th and 60th percentiles except when the life and censoring distributions are Uniform. (In this case, the bias is positive only after the 90th percentile.) We conclude that Whittemore and Keller merely avoid detailed discussion of bias. Regarding the hybrid estimator, we find in the figures recorded some suggestions of the tendencies observed in the PEXE -- specifically, monotone increasing bias and a tendency for underestimation when the sample size is small and censoring is heavy. Whether this behaviour is typical of the PEXE also remains to be seen.

In considering the magnitude of the bias of the estimators, we find the following.

(i) Both Kitchin and Whittemore and Keller report that the bias of the KME is negligible except in the right tail of the distribution and in the case of a very small sample ($n=10$) and heavy censoring.

(ii) The PEXE is considerably more biased than the KME.

(iii) The bias of $\hat{F}_{Q_4}$ is negligible except in the case of a very small sample and heavy censoring.

(iv) The bias of each estimator increases as the censoring becomes heavier and it decreases as the sample size increases.

In view of these two studies, we conclude, firstly, that the PEGE is likely to compare favourably with the KME in terms of MSE, and secondly, that the PEGE is likely to be considerably more biased than the KME. We expect that the discrete counterpart of $\hat{F}_{Q_4}$ performs well in terms of both MSE and bias. Since the bias of this estimator is likely to be small, adjustment for its presumed tendency to increase monotonically is deemed an unnecessary complication.

In the pilot study we generate three collections of data, each consisting of 100 samples of size 10, from independent Geometric life and censoring distributions. In each case the life distribution has parameter $p = \exp(-.1)$. The censoring distributions are chosen so as to produce three levels of censoring: setting $p = \exp(-\lambda)$, where $\lambda = .00001, .1, .3$, yields the censoring probabilities $P(Y < X) = 0, .475, .711$ respectively.

The conventions followed for extrapolation in the range beyond the largest observed failure time are as follows:

$$\tilde{P}(X > k) = \begin{cases} \tilde{P}(X > t_{n_1}) & \text{for } t_{n_1} \leq k < s_{n_2} \\ 0 & \text{for } k \geq s_{n_2} \geq t_{n_1} \end{cases}$$

$$\hat{P}(X > k) = \hat{P}(X > t_{n_1}) (1 - \hat{q}_{n_1})^{k - t_{n_1}} \quad \text{for } k \geq t_{n_1}.$$

This definition of the KME rests on the assumption that the largest observation is uncensored, while the definition of the PEGE results from assuming that the failure rate after the largest observed failure time is the same as the failure rate in the interval $(t_{n_1-1}, t_{n_1}]$.

Our conventions for extrapolation differ from those of Kitchen and of Whittemore and Keller. Consequently our results involving right-hand tail probabilities differ from theirs: a preliminary indication is that our extrapolation procedures result in estimators that are more realistic than theirs.

Although the size of the study precludes reaching more than tentative conclusions, we observe several tendencies.

Tables 1(a), 2(a) and 3(a) contain the estimated bias and mean squared error (MSE) for the KME and the PEGE of $P(X > k)$ for $k = \xi_p$, where ξ_p is the p^{th} percentile of the underlying life distribution and $p = 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 99$. From these tables we make the following observations.

(i) The MSE of the PEGE is generally smaller than that of the KME. The exceptions occur in the right-hand tail of the distribution under conditions of moderate and heavy censoring.

(ii) The MSE of each estimator increases as censoring increases.

(iii) The disparity in the MSE of the two estimators becomes more marked as the censoring increases -- that is, the MSE of the PEGE increases by relatively little as the censoring increases, except in the right-hand tail.

(iv) The difference in the MSE of the two estimators is smallest near the median of the distribution.

(v) Both the KME and the PEGE generally exhibit negative bias: the magnitude of the bias of each estimator is greatest around the median of the distribution.

(vi) The magnitude of the bias of the KME is consistently smaller than that of the PEGE only when there is no censoring. Under conditions of moderate and heavy censoring, the KME is less biased than the PEGE only at percentiles to the left of the median: to the right of the median, the PEGE is considerably less biased than the KME.

(vii) As censoring increases, the magnitude of the bias of the KME increases faster than does that of the PEGE.

Tables 1(b), 2(b) and 3(b) contain the estimated bias and MSE for the Kaplan-Meier (KM) and piecewise geometric (PG) estimators of the percentiles ξ_p , $p = 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 99$. From these tables we make the following observations.

(i) With a few exceptions, the PG percentile estimator is less biased than the KM percentile estimator.

(ii) Both estimators tend to be negatively biased.

(iii) At each level of censoring, the bias of the PG percentile estimator is negligible for percentiles smaller than the 70th, and it is acceptably small for larger percentiles, except perhaps the 99th percentile. In contrast, the KM percentile estimators are almost unbiased only for percentiles smaller than the 60th: to the right of the 60th percentile the bias tends to be very much larger than that of the PG estimators. This tendency is particularly noticeable in the case of heavy censoring.

(iv) The MSE of the PG percentile estimator is smaller than that of the KM percentile estimator only in certain ranges, viz: $p \leq 70$ for heavy censoring, $p \leq 40$ for moderate censoring, and $5 \leq p \leq 95$ for no censoring. Since the PG percentile estimator is almost unbiased outside these ranges, the large MSE must be the result of having large variance.

On the basis of the observations involving the survival function estimators, we conclude that the small sample behaviour of the PEGE resembles that of the PEXE: specifically, when there is little or no censoring, the PEGE compares favourably with the KME in terms of MSE but not in terms of bias. We expect that this is true irrespective of the level of censoring when the sample size is larger. It remains to be seen whether inversion of this general behaviour is typical when the sample size is very small and censoring is heavy.

It is evident that increased censoring affects the bias and the MSE of PEGE less than it affects the bias and the MSE of the KME.

Our conclusions about the percentile estimators are even more tentative because of the lack of results involving the behaviour of percentile estimators. The fact that the PG percentile estimator is almost unbiased even in the presence of heavy censoring, and even as far to the right as the 95th percentile, is of considerable interest because the KM extrapolation procedures are clearly inadequate for estimating extreme right percentiles.

Regarding the MSE, we note that, under conditions of moderate or heavy censoring, any estimator of the larger percentiles is expected to vary considerably because there are likely to be very few observations in this range. The ad hoc extrapolation procedure for the KM is expected to cause the estimators of the extreme right percentiles to exhibit large negative bias and little variation. In view of these considerations and the accuracy of the PG percentile estimators, we conclude that the fact that the MSE of the PG percentile estimator of the larger percentiles is greater than that of the KM percentile estimator is not evidence of a breakdown in the reliability and efficiency of the PG percentile estimator.

The general indications of our pilot study are that the PEGE and the discrete version of $\hat{F}_{Q_4}$ are attractive alternatives to the KME. In view of the resemblance between the properties of the PEGE and those of the PEXE, the results for $\hat{F}_{Q_4}$ portend well for the new discrete estimator: we expect it to be almost unbiased and to be not only more efficient than the KME but also more stable under increased censoring. Moreover, we expect the corresponding percentile estimator to have these desirable properties also because it is likely to behave at least as well as the PG percentile estimator.

The properties involving relative efficiency are of considerable importance because relative efficiency is a measure of the relative quantities of information utilized by the estimators being compared. This interpretation of relative efficiency, and the fact that heavy censoring is often encountered in engineering problems, makes $\hat{F}_{Q_4}$ and its discrete counterpart even more attractive.

Table 1. Results of Pilot Study Using 100 Samples of Size 10, Geometric ($p = \exp(-.1)$) Life Distribution, Geometric ($p = \exp(-.00001)$) Censoring Distribution and $P(Y < X) \approx 0$.

(a) Survival Function Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	-.0184	-.0018	.0078	.0101
5	-.0184	-.0018	.0078	.0101
10	-.0137	.0123	.0118	.0145
20	-.0172	.0092	.0161	.0182
30	-.0253	-.0053	.0194	.0225
40	-.0293	-.0118	.0255	.0279
50	-.0351	-.0196	.0271	.0278
60	-.0347	-.0159	.0223	.0257
70	-.0318	-.0185	.0176	.0212
80	-.0283	-.0187	.0108	.0133
90	-.0199	-.0167	.0047	.0060
95	-.0096	-.0028	.0028	.0049
99	.0029	-.0011	.0006	.0009

(b) Percentile Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	.00	.63	.00	1.63
5	.21	.63	.35	1.63
10	-.21	-.37	1.69	1.37
20	-.08	-.32	3.00	2.38
30	.28	-.10	4.48	3.88
40	-.16	-.79	6.20	5.71
50	.53	-.08	9.57	9.72
60	-.62	-1.31	13.70	14.37
70	-1.35	-2.28	20.43	22.82
80	-1.87	-3.34	35.23	35.96
90	-2.29	-4.87	82.53	95.37
95	-2.20	-1.53	130.22	140.17
99	-5.01	-18.53	577.47	481.19

Table 2. Results of Pilot Study Using 100 Samples of Size 10, Geometric ($p = \exp(-.1)$) Life Distribution, Geometric ($p = \exp(-.1)$) Censoring Distribution and $P(Y < X) = .475$.

(a) Survival Function Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	-.0223	-.0018	.0077	.0101
5	-.0223	-.0018	.0077	.0101
10	-.0207	.0106	.0124	.0157
20	-.0215	.0094	.0170	.0208
30	-.0282	-.0042	.0244	.0300
40	-.0432	-.0037	.0407	.0502
50	-.0509	-.0230	.0475	.0601
60	-.0564	-.0442	.0430	.0634
70	-.0553	-.0800	.0333	.0603
80	-.0368	-.0707	.0229	.0413
90	-.0060	-.0590	.0124	.0151
95	.0082	-.0401	.0082	.0049
99	.0149	-.0091	.0033	.0001

(b) Percentile Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	.00	.80	.00	3.36
5	.19	.80	.33	3.36
10	-.34	-.20	1.66	2.76
20	-.09	.08	3.69	5.36
30	.38	.80	7.40	9.84
40	.10	.64	12.62	17.24
50	.77	1.43	20.97	25.21
60	-.20	.62	34.24	37.26
70	-.67	-1.44	64.85	36.28
80	-.88	-2.73	128.02	52.21
90	-1.23	-8.92	302.31	121.66
95	-.60	-14.92	561.06	264.70
99	-2.30	-31.92	1497.30	1060.98

Table 3. Results of Pilot Study Using 100 Samples of Size 10, Geometric ($p = \exp(-.1)$) Life Distribution, Geometric ($p = \exp(-.3)$) Censoring Distribution and $P(Y < X) \approx .711$.

(a) Survival Function Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	-.0230	-.0018	.0077	.0101
5	-.0230	-.0018	.0077	.0101
10	-.0370	.0033	.0171	.0185
20	-.0582	-.0273	.0301	.0508
30	-.0714	-.0479	.0437	.0704
40	-.1150	-.1011	.0705	.1257
50	-.1232	-.1443	.0709	.1382
60	-.1006	-.2421	.0594	.1273
70	-.0702	-.2286	.0456	.0711
80	-.0347	-.1775	.0321	.0341
90	.0032	-.0907	.0187	.0082
95	.0173	-.0498	.0125	.0025
99	.0206	-.0091	.0043	.0001

(b) Percentile Estimators.

Percentile	Estimated Bias		Estimated MSE	
	PEGE	KME	PEGE	KME
1	.10	.87	.52	3.27
5	.24	.87	.68	3.27
10	-.41	-.13	1.37	2.53
20	-.08	.52	3.22	7.86
30	.29	.76	7.19	8.82
40	-.20	-.10	15.16	9.56
50	.48	.16	28.06	10.86
60	-.47	-2.38	50.99	16.66
70	-.78	-4.91	90.72	36.07
80	-1.11	-8.54	167.67	84.44
90	-1.68	-15.53	357.58	252.63
95	-1.25	-21.53	619.71	474.99
99	-3.34	-38.53	1508.06	1496.01

REFERENCES

- Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. Scandinavian Journal of Statistics 3, 15-27.
- Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. Annals of Statistics 6, 534-545.
- Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimators under random censorship. Annals of Statistics 2, 437-453.
- Chen, Y.Y., Hollander, M. and Langberg, N. (1982). Small-sample results for the Kaplan-Meier estimator. Journal of the American Statistical Association 77, 141-144.
- Cox, D.R. (1972). Regression models and life tables. Journal of the Royal Statistical Society, Series B 34, 187-202.
- Desu, M.M. and Narula, S.C. (1977). Reliability estimation under competing causes of failure. The Theory and Applications of Reliability I, ed. by I. Shimi and C.P. Tsokos. Academic Press, New York.
- Efron, B. (1967). The two sample problem with censored data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV. University of California Press, Berkeley, California, 831-853.
- Fleming, T.R. and Harrington, D.P. (1979). Nonparametric estimation of the survival distribution in censored data. Technical Report No. 8, Section of Medical Research Statistics, Mayo Clinic, Rochester, Minnesota.
- Freireich, E.J. et al. (1963). The effect of 6-Mercaptopurine on the duration of steroid-induced remission in acute leukemia. Blood 21, 699-716.
- Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.
- Kitchin, J. (1980). A new method for estimating life distributions from incomplete data. Unpublished doctoral dissertation, Florida State University.
- Kitchin, J., Langberg, N. and Proschan, F. (1983). A new method for estimating life distributions from incomplete data. Statistics and Decisions 1, 241-255.

- Teagberg, N., Proschan, F. and Quinzi, A.J. (1980). Estimating dependent life lengths, with applications to the theory of competing risks. Annals of Statistics 8, No. 6.
- Miller, R.G. (1981). Survival analysis. Wiley, New York.
- Mimmack, G.M. (1985). Piecewise geometric estimation of a survival function. Unpublished doctoral dissertation, Florida State University.
- Nelson, W. (1969). Hazard plotting for incomplete failure data. Journal of Quality Technology 1, 27-52.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. Technometrics 14, 945-966.
- Peterson, A.V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical sub-survival functions. Journal of the American Statistical Association 72, 854-858.
- Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. Journal of the American Statistical Association 71, 897-902.
- Umholtz, R.L. (1984). Estimation of the Exponential parameter for discrete data. Report, Aberdeen Proving Ground.
- Whittemore, A.S. and Keller, J.B. (1983). Survival estimation with censored data. Stanford University Technical Report No. 69.

END

FILMED

1-86

DTIC