Estimating and Simulating Nonhomogeneous Poisson Processes

Larry Leemis
Department of Mathematics
The College of William & Mary
Williamsburg, VA 23187–8795 USA
757–221–2034
E-mail: leemis@math.wm.edu
May 23, 2003

Outline

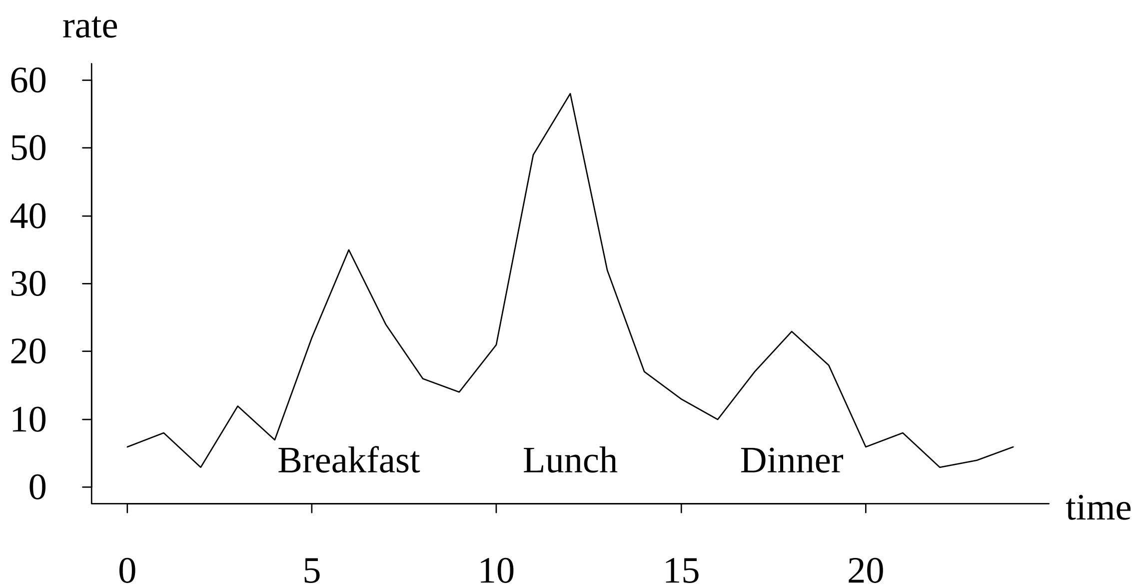
1. Motivation
2. Probabilistic properties
3. Estimating $\Lambda(t)$ from k realizations on $(0, S]$
4. Estimating $\Lambda(t)$ from overlapping realizations
5. Software
6. Conclusions

Note: Portions of this work are with Brad Arkin (RST Corporation), Andy Glen (United States Military Academy), John Drew (William & Mary), and Diane Evans (Rose–Hulman). Part of this work was supported by an NSF grant supporting the UMSA (Undergraduate Modeling and Simulation Analysis) REU (Research Experience for Undergraduates) at The College of William & Mary.

1. Motivation

Although easy to estimate and simulate, HPPs and renewal processes do not allow for varying rates. The use of an NHPP is often more appropriate for modeling.

Example: Customer arrivals to a fast food restaurant



Other examples:

Cyclone arrival times in the Arctic Sea (Lee, Wilson, and Crawford, 1991)

Database transaction times (Lewis and Shedler, 1976)

Calls for on-line analysis of electrocardiograms at a hospital in Houston (Kao and Chang, 1988)

Respiratory cancer deaths near a steel complex in Scotland (Lawson, 1988)

Repairable systems: blood analyzers, fan motors, power supplies, turbines (Nelson, 1988)

2. Probabilistic properties

Notation

t	time
$N(t)$	number of events by time t
$\lambda(t)$	instantaneous arrival rate at time t (intensity function)
$\Lambda(t) = \int_0^t \lambda(\tau) d\tau$	cumulative intensity function

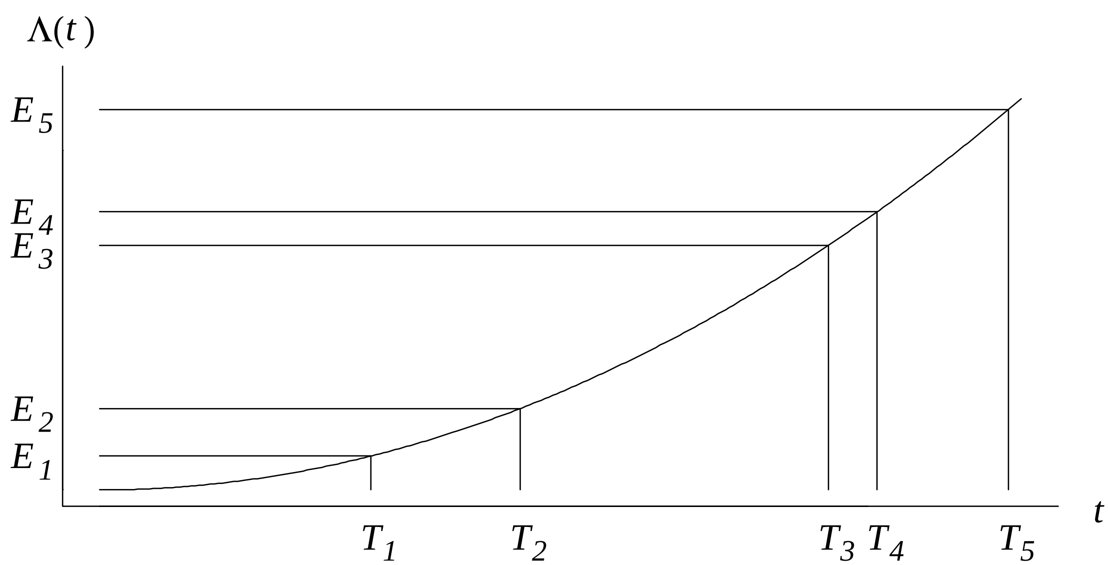
Property 1

$$\Pr\left(N(b) - N(a) = n\right) = \frac{\left[\int_a^b \lambda(\tau) d\tau\right]^n e^{-\int_a^b \lambda(\tau) d\tau}}{n!} \qquad n = 0, 1, \ldots$$

Property 2

$$E[N(t)] = \Lambda(t)$$

Property 3 (Çinlar, 1975) If $E_1, E_2, \ldots$ are event times in a *unit* HPP then $\Lambda^{-1}(E_1), \Lambda^{-1}(E_2), \ldots$ are event times in an NHPP with cumulative intensity function $\Lambda(t)$.



3. Estimating $\Lambda(t)$ from k realizations on $(0, S]$

Data

t	time
$(0, S]$	time interval where observations are collected
k	number of realizations collected
$n_1, n_2, \ldots, n_k$	number of observations per realization
$t_{(1)}, t_{(2)}, \ldots, t_{(n)}$	superposition of observations
$n = \sum_{i=1}^{k} n_i$	total number of observations collected

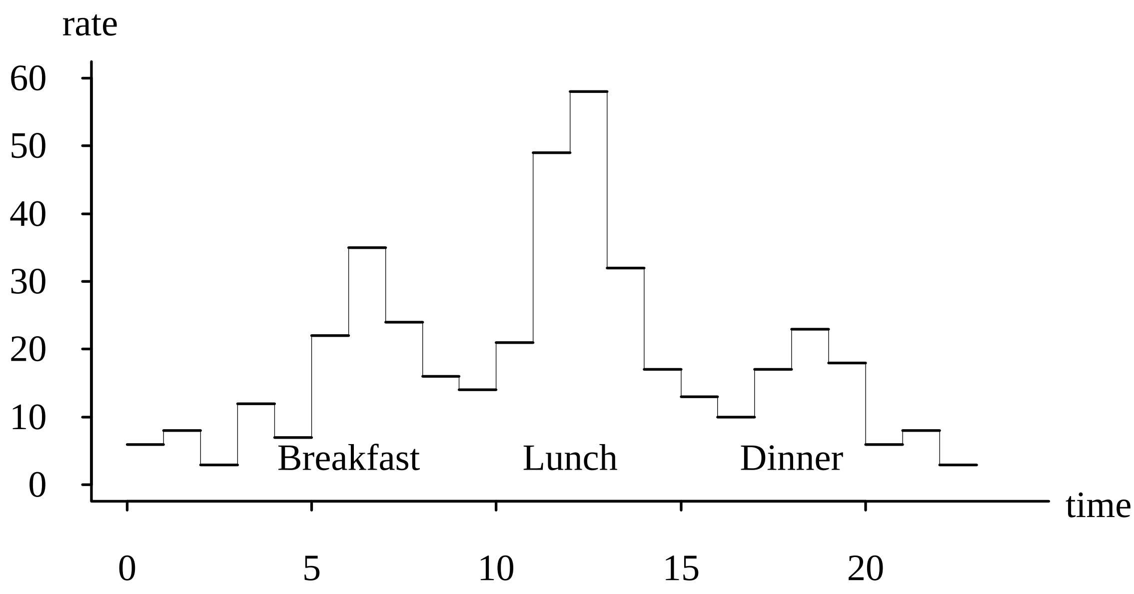
Intuitive solution (Law and Kelton, 2000): partition time axis and let $\lambda(t)$ be piecewise constant.

Problems

(a) Determining cell width

- Small cell width — sampling variability
- Large cell width — miss trend

(b) Subjective due to arbitrary parameters

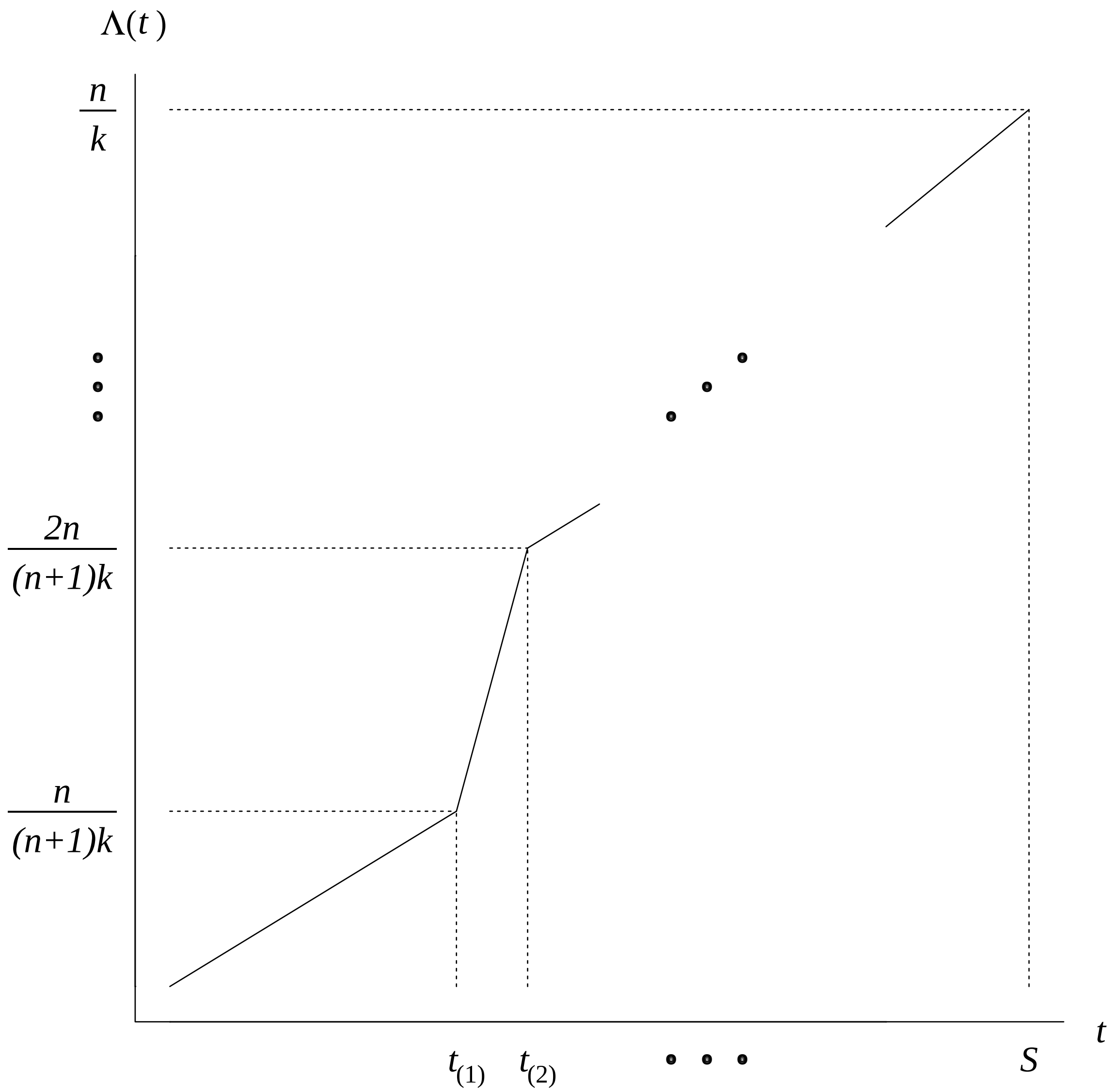


Piecewise-linear nonparametric cumulative intensity function estimator

$$\hat{\Lambda}(t) = \frac{in}{(n+1)k} + \left[\frac{n\left(t - t_{(i)}\right)}{(n+1)\,k\left(t_{(i+1)} - t_{(i)}\right)}\right] \qquad t_{(i)} < t \leq t_{(i+1)}$$

for $i = 0, 1, 2, \ldots, n$.

Rationale: $\hat{\Lambda}(S) = n/k$



Properties

- Handles ties as expected
- Consistency
$$\lim_{k\to\infty} \hat{\Lambda}(t) = \Lambda(t)$$
- Confidence interval (asymptotically exact) for $\Lambda(t)$
$$\hat{\Lambda}(t) \pm z_{\alpha/2}\sqrt{\frac{\hat{\Lambda}(t)}{k}}$$
- Variate generation straightforward

  Input:
  Number of observed arrivals n
  Number of active realizations k
  Superpositioned observations $t_{(1)}, t_{(2)}, \ldots, t_{(n+1)}$
  Output:
  Event times $T_1, T_2, \ldots, T_{i-1}$ on $(0, S]$

$i \leftarrow 1$ [initialize variate counter]
generate $U_i \sim U(0,1)$ [generate initial random number]
$E_i \leftarrow -\log_e(1 - U_i)$ [generate initial exponential variate]
while $E_i < n/k$ **do**
  begin
    $m \leftarrow \left\lfloor \frac{(n+1)kE_i}{n} \right\rfloor$ [set $m \ni \hat{\Lambda}(t_{(m)}) < E_i \le \hat{\Lambda}(t_{(m+1)})$]
    $T_i \leftarrow t_{(m)} + [t_{(m+1)} - t_{(m)}]\left(\frac{(n+1)kE_i}{n} - m\right)$
    [generate event time]
    $i \leftarrow i + 1$ [increment variate counter]
    generate $U_i \sim U(0,1)$ [generate next random number]
    $E_i \leftarrow E_{i-1} - \log_e(1 - U_i)$ [generate next HPP event time]
  end

4. Estimating $\Lambda(t)$ from overlapping realizations

Data

t	time
$(0, S]$	time interval where observations are collected
r	# time intervals where the # realizations is fixed
$(s_j, s_{j+1}]$	interval $j+1$, $j = 0, 1, \ldots, r-1$
k_{j+1}	# realizations observed on $(s_j, s_{j+1}]$, $j = 0, 1, \ldots, r-1$
n_{j+1}	number of observations on $(s_j, s_{j+1}]$
$t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$	superposition of observations, s_0, s_1, ..., s_r

Note: $s_0 = 0$, $s_r = S$

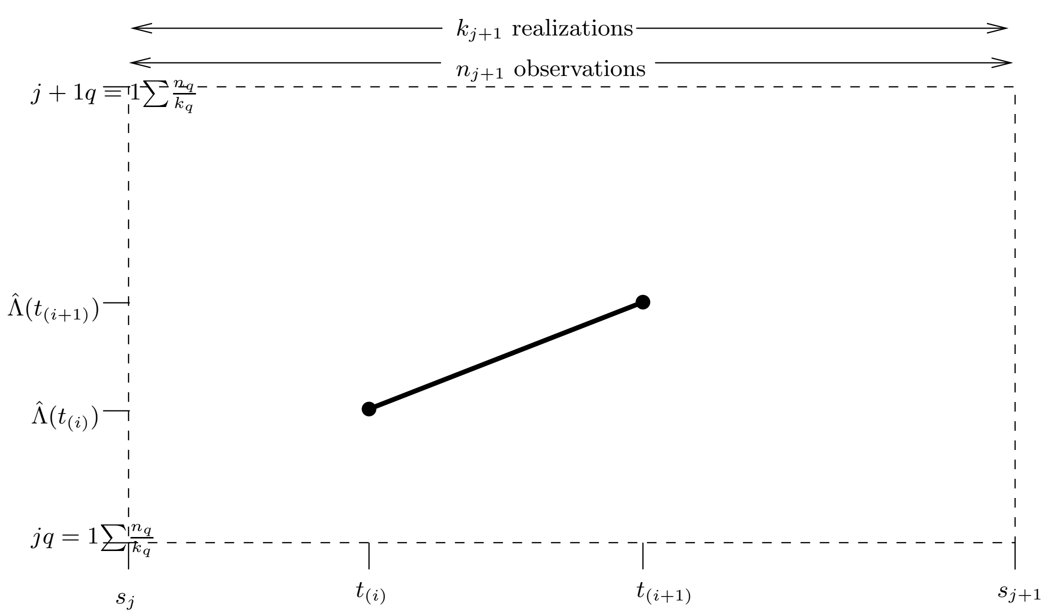
$$\hat{\Lambda}(t) = \sum_{q=1}^{j} \frac{n_q}{k_q} + \frac{\left(i - \sum_{q=1}^{j}(n_q+1)\right) n_{j+1}}{(n_{j+1}+1)\, k_{j+1}} \;+\; \left[\frac{n_{j+1}\left(t - t_{(i)}\right)}{(n_{j+1}+1)\, k_{j+1}\left(t_{(i+1)} - t_{(i)}\right)}\right],$$

$$t_{(i)} < t \le t_{(i+1)}; \qquad i = 0, 1, 2, \ldots, n+r-1,$$

$$s_j < t \le s_{j+1}; \qquad j = 0, 1, \ldots, r-1,$$

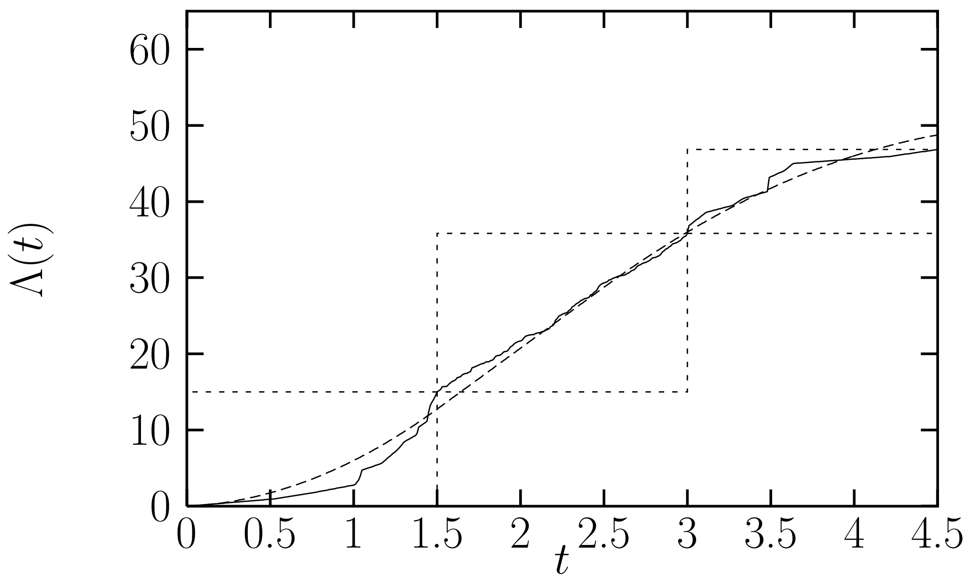
Rationale: $\hat{\Lambda}(s_{j+1}) = \sum_{q=1}^{j+1} n_q / k_q$

Single segment of $\hat{\Lambda}(t)$ in the $(j+1)$st region:



Example: Lunchwagon arrivals (Klein and Roberts, 1984)

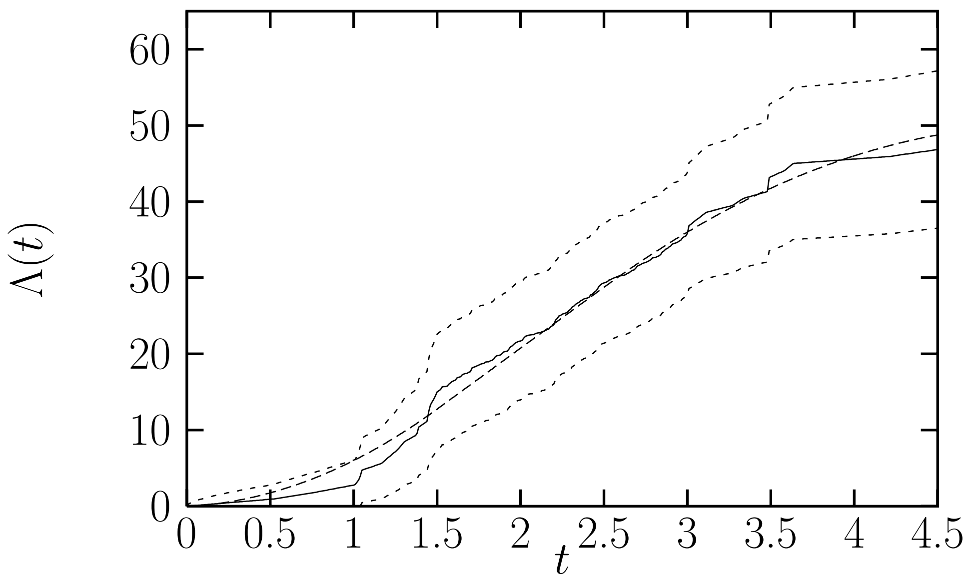
Depiction of three regions for lunchwagon arrivals from 10:00 AM to 2:30 PM for $k_1 = 1$, $k_2 = 12$, and $k_3 = 1$; $s_1 = 1.5$, $s_2 = 3$, and $s_3 = 4.5$.



An asymptotically exact $100(1-\alpha)\%$ confidence interval for $\Lambda(t)$

$$\left|\Lambda(t) - \hat{\Lambda}(t)\right| < z_{\alpha/2}\sqrt{\frac{\hat{\Lambda}(t) - \hat{\Lambda}(s_j)}{k_{j+1}} + \sum_{q=1}^{j} \frac{\hat{\Lambda}(s_q) - \hat{\Lambda}(s_{q-1})}{k_q}}$$

Parent cumulative intensity function, nonparametric estimator, and 95% confidence bands for lunchwagon arrivals from 10:00 AM to 2:30 PM for $k_1 = 1$, $k_2 = 12$, and $k_3 = 1$; $s_1 = 1.5$, $s_2 = 3$, and $s_3 = 4.5$.



Variate Generation

Input:

Number of partitions r

Number of active realizations $k_1, k_2, \ldots, k_r$

Number of observed arrivals per partition $n_1, n_2, \ldots, n_r$

Superpositioned values $t_{(0)}, t_{(1)}, \ldots, t_{(n+r)}$

Output:

Event times $T_1, T_2, \ldots, T_{i-1}$ on $(0, S]$

$i \leftarrow 1$ [initialize variate counter]

$j \leftarrow 0$ [initialize region counter]

MAX $\leftarrow \sum_{q=1}^{r} n_q/k_q$ [set MAX to $\hat{\Lambda}(S)$]

generate $U_i \sim U(0,1)$ [generate initial random number]

$E_i \leftarrow -\log_e(1-U_i)$ [generate initial exponential variate]

while $E_i <$ MAX **do**

  begin

    while $E_i > \sum_{q=1}^{j+1} n_q/k_q$ **do** [update j if necessary]

      begin

        $j \leftarrow j+1$ [increment region counter]

      end

    $m \leftarrow \left\lfloor \frac{(n_{j+1}+1)k_{j+1}\left(E_i - \sum_{q=1}^{j} n_q/k_q\right)}{n_{j+1}} \right\rfloor + \sum_{q=1}^{j}(n_q+1)$

[set $m \ni \hat{\Lambda}(t_{(m)}) < E_i \leq \hat{\Lambda}(t_{(m+1)})$]

    $T_i \leftarrow t_{(m)} + [t_{(m+1)} - t_{(m)}]\left(\frac{(n_{j+1}+1)k_{j+1}\left(E_i - \sum_{q=1}^{j} n_q/k_q\right)}{n_{j+1}} - \left(m - \sum_{q=1}^{j}(n_q+1)\right)\right)$

[generate event time]

    $i \leftarrow i+1$ [increment variate counter]

    generate $U_i \sim U(0,1)$ [generate next random number]

    $E_i \leftarrow E_{i-1} - \log_e(1-U_i)$ [generate next HPP event time]

  end

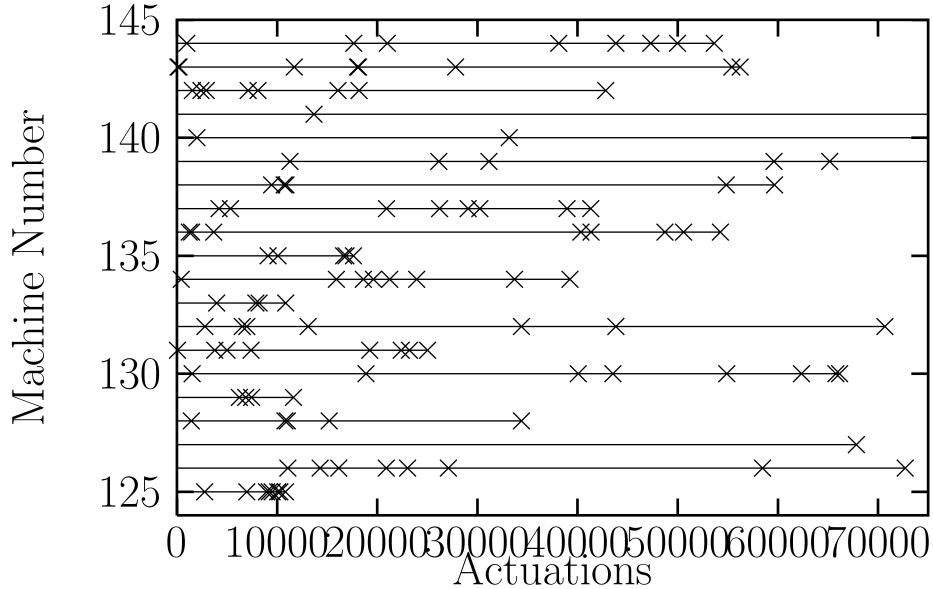
Example 1: Monte Carlo evaluation of the confidence interval for $\Lambda(t)$

Coverages in the lunchwagon example (nominal coverage 0.95; 100,000 replications; $k_1 = 1, k_2 = 12, k_3 = 1$; $s_0 = 0, s_1 = 1.5, s_2 = 3, s_3 = 4.5$).

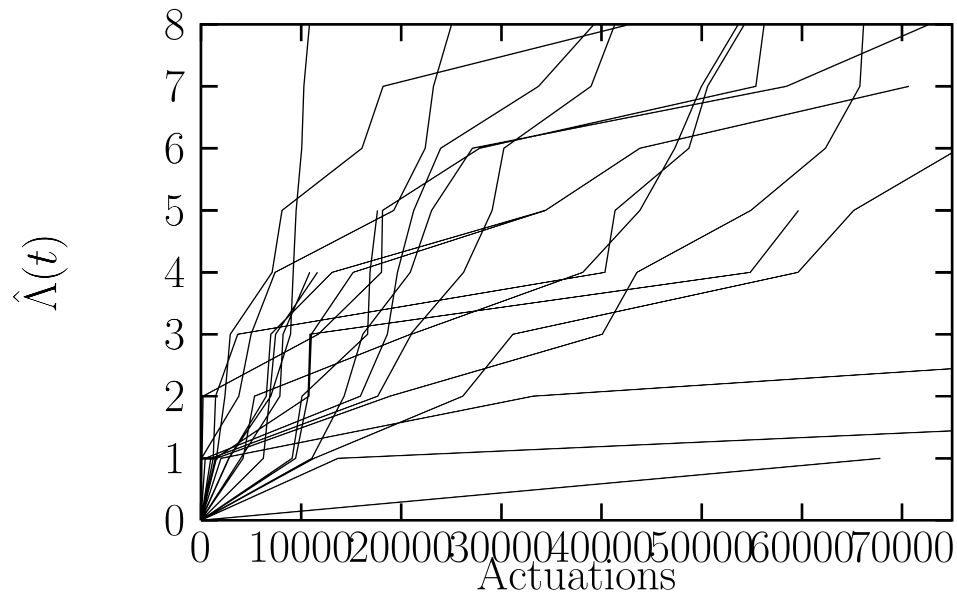
Time	Actual Coverage	Misses High	Misses Low
0.90	0.9501	0.0013	0.0487
1.35	0.9386	0.0048	0.0566
1.80	0.9505	0.0200	0.0296
2.25	0.9466	0.0196	0.0339
2.70	0.9498	0.0174	0.0329
3.15	0.9509	0.0295	0.0196
3.60	0.9498	0.0251	0.0251
4.05	0.9517	0.0167	0.0316

Example 2: Failure times for 20 copy machines (Zaino and Berke 1992)

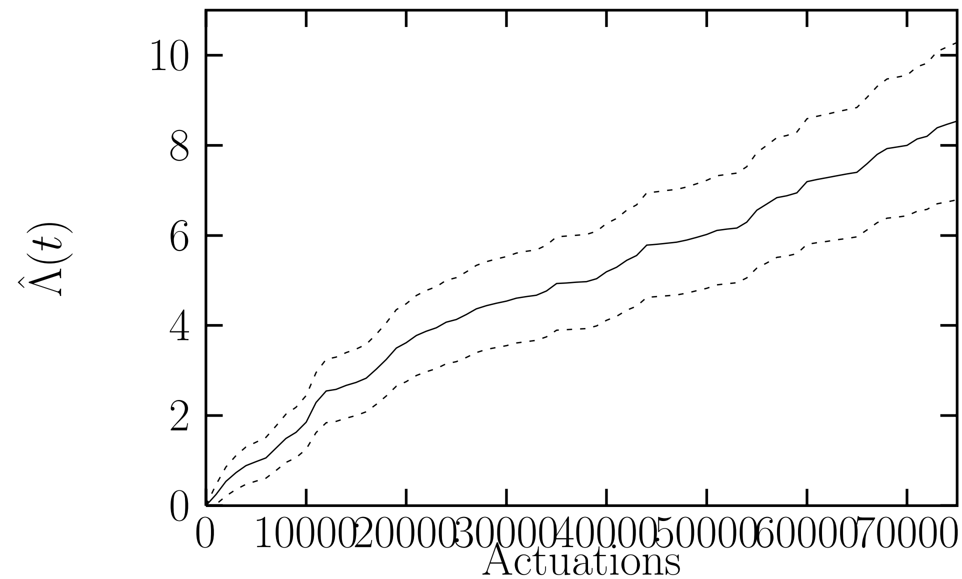
Failure times



$\hat{\Lambda}(t)$ for each machine



$\hat{\Lambda}(t)$ for the copy machine failure times

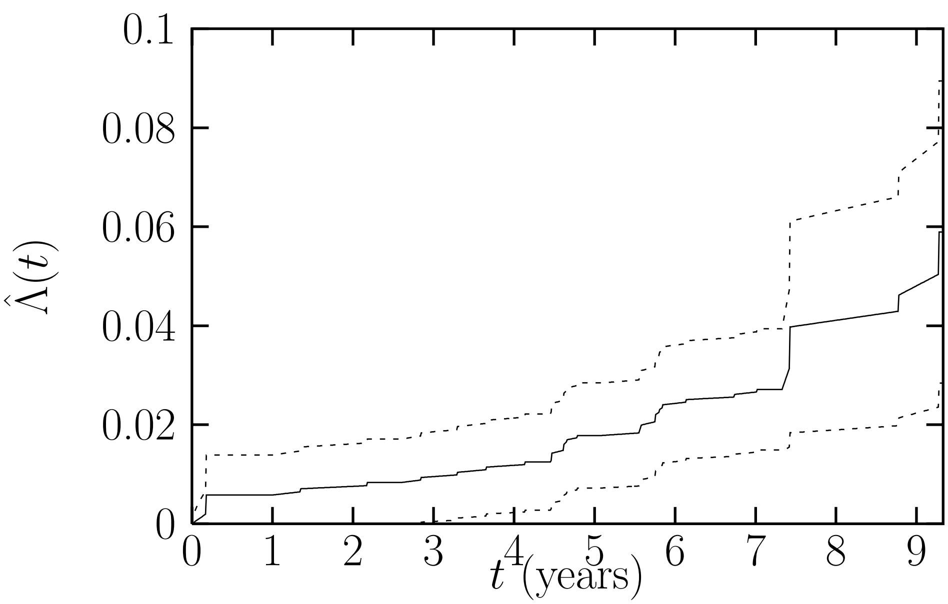


Example 3: Failure times of heat pump compressors (Nelson 1990)

The compressors are located in five separate buildings, each under repair contract for a time span $(a, b]$, indicated by the boldface values. The data set consists of $n = 28$ failure times, and yields $r = 29$ regions.

Bldg	Num of Comp	**Entry time**, Failure Times, **Exit time**
B	164	**2.59**, 3.30, 4.62, 4.62, 5.75, 5.75, 7.42, 7.42, 8.77, 9.27,
D	356	**4.45**, 4.47, 4.47, 5.56, 5.57, 5.80, 6.13, 7.02, **7.0**
E	458	**1.00**, 2,85, 4.65, 4.79, 5.85, 6.73, **7.33**
H	149	**0.00**, 0.17, 0.17, 1.34, **5.09**
K	195	**0.00**, 2.17, 3.65, 4.14, **4.14**

$\hat{\Lambda}(t)$ for the heat pump compressor failure times



5. Software

Civilization advances by extending the number of important operations which we can perform without thinking about them.
—Alfred North Whitehead (1861–1947)

APPL (A Probability Programming Language) is a Maple-based language with data structures for discrete and continuous random variables and algorithms for their manipulation.

Example 1: Let $X_1, X_2, \ldots, X_{10}$ be independent and identically distributed U(0,1) random variables. Find

$$\Pr\left(4 < \sum_{i=1}^{10} X_i < 6\right)$$

Typical approaches

- Central limit theorem
- Simulation

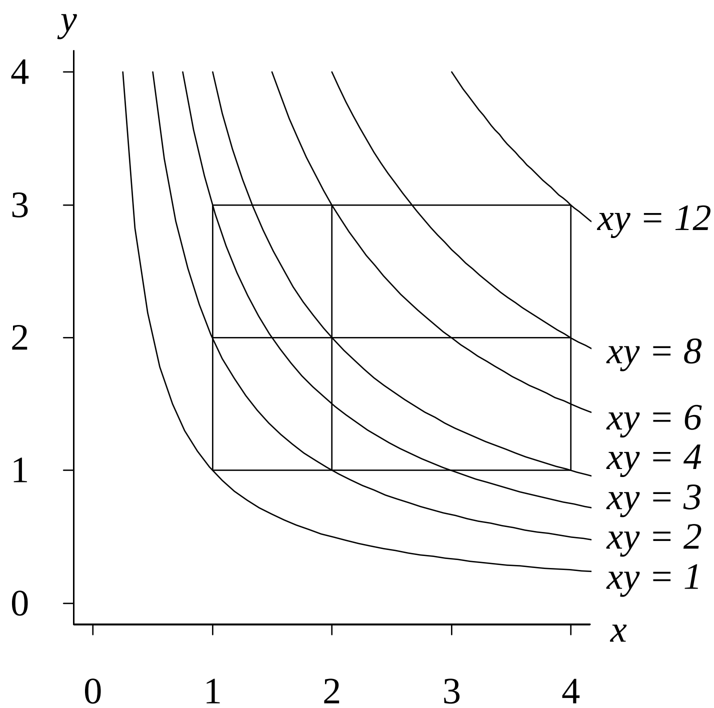
```
n := 10;
X := UniformRV(0, 1);
Y := Convolution(X, n);
CDF(Y, 6) - CDF(Y, 4);
```

$$\frac{655177}{907200}$$

Example 2:

$$X \sim Triangular(1, 2, 4)$$
$$Y \sim Triangular(1, 2, 3)$$

Find the distribution of $V = XY$.




```
X := TriangularRV(1, 2, 4);
Y := TriangularRV(1, 2, 3);
V := Product(X, Y);
```

which returns the probability density function of `V` as

$$f_V(v) = \begin{cases} -\frac{4}{3}v + \frac{2}{3}\ln v + \frac{2v}{3}\ln v + \frac{4}{3} & 1 < v \le 2 \\ -8 + \frac{14}{3}\ln 2 + \frac{7v}{3}\ln 2 + \frac{10}{3}v - 4\ln v - \frac{5v}{3}\ln v & 2 < v \le 3 \\ -4 + \frac{14}{3}\ln 2 + \frac{7v}{3}\ln 2 + 2v - 2\ln v - v\ln v - 2\ln 3 - \frac{2v}{3}\ln 3 & 3 < v \le 4 \\ \frac{44}{3} - 14\ln 2 - \frac{7v}{3}\ln 2 - \frac{8}{3}v - 2\ln 3 + \frac{22}{3}\ln v - \frac{2v}{3}\ln 3 + \frac{4v}{3}\ln v & 4 < v \le 6 \\ \frac{8}{3} - 8\ln 2 - \frac{4v}{3}\ln 2 - \frac{2}{3}v + \frac{4}{3}\ln v + \frac{v}{3}\ln v + 4\ln 3 + \frac{v}{3}\ln 3 & 6 < v \le 8 \\ -8 + 8\ln 2 + \frac{2v}{3}\ln 2 + \frac{2}{3}v + 4\ln 3 - 4\ln v + \frac{v}{3}\ln 3 - \frac{v}{3}\ln v & 8 < v < 12 \end{cases}$$

Example 3: Kolmogorov–Smirnov test statistic (all parameters known)
Defining formula:

$$D_n = \sup_x |F(x) - F_n(x)|,$$

Computational formula:

$$D_n = \max_{i=1,2,\ldots,n} \left\{ \left| \frac{i-1}{n} - x_{(i)} \right|, \left| \frac{i}{n} - x_{(i)} \right| \right\}$$

The cdf for the test statistic is (Birnbaum, 1952)

$$P\left(D_n < \frac{1}{2n} + v\right) = n! \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \cdots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v}$$

$$g(u_1, u_2, \ldots, u_n)\, du_n \ldots du_2\, du_1$$

for $0 \le v \le \frac{2n-1}{2n}$, where

$$g(u_1, u_2, \ldots, u_n) = 1$$

for $0 \le u_1 \le u_2 \le \cdots \le u_n$.

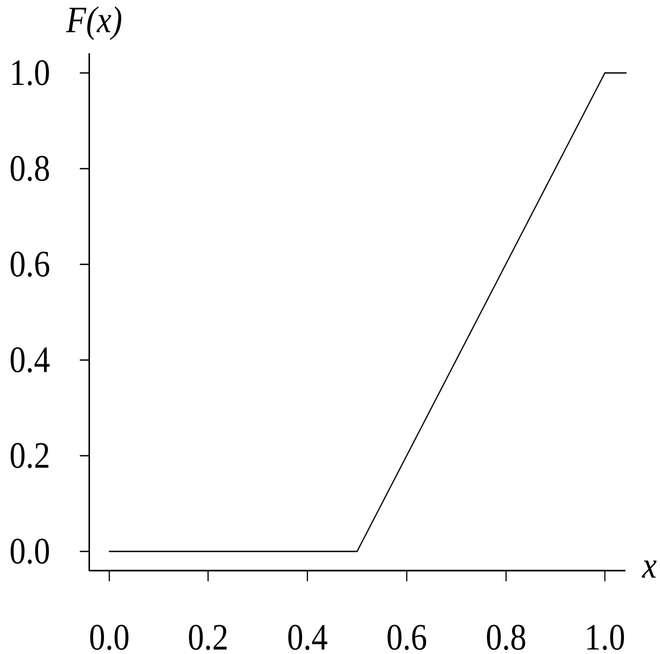
CASE I: $n = 1$

$$F_{D_1}(t) = \Pr(D_1 \le t) = \begin{cases} 0 & t \le \frac{1}{2} \\ 2t - 1 & \frac{1}{2} < t < 1 \\ 1 & t \ge 1 \end{cases}$$

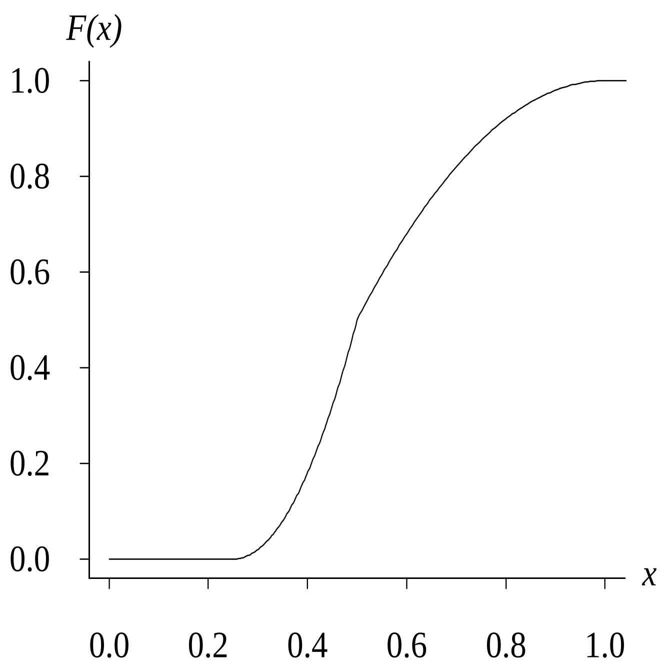
CASE II: $n = 2$

$$F_{D_2}(t) = \Pr(D_2 \le t) = \begin{cases} 0 & t \le \frac{1}{4} \\ 8\left(t - \frac{1}{4}\right)^2 & \frac{1}{4} < t < \frac{1}{2} \\ 1 - 2(1-t)^2 & \frac{1}{2} < t < 1 \\ 1 & t \ge 1 \end{cases}$$

CASE I: $n = 1$

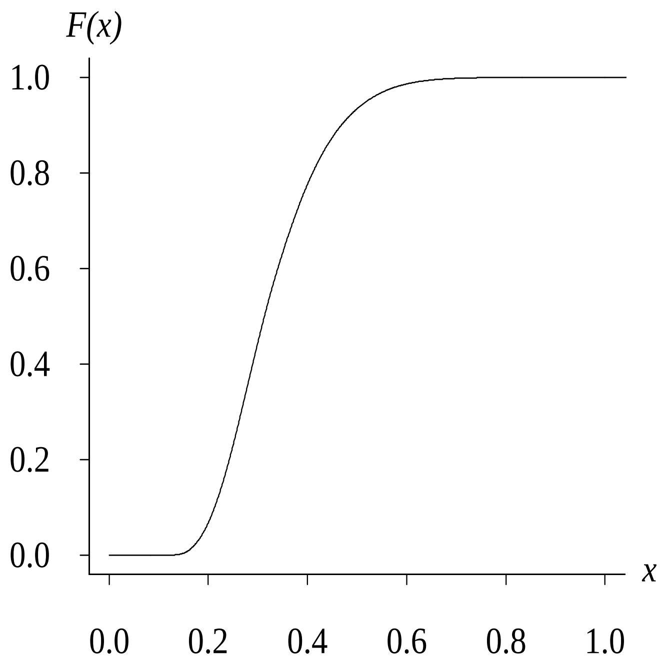


CASE II: $n = 2$



Goal: `X := KSRV(n);`

CASE III: $n = 6$



$$F_{D_6}(y) = \begin{cases} 0 & y < \frac{1}{12} \\ 46080\,y^6 - 23040\,y^5 + 4800\,y^4 - \frac{1600}{3}\,y^3 + \frac{100}{3}\,y^2 - \frac{10}{9}\,y + \frac{5}{324} & \frac{1}{12} \le y < \frac{1}{6} \\ 2880\,y^6 - 4800\,y^5 + 2360\,y^4 - \frac{1280}{3}\,y^3 + \frac{235}{9}\,y^2 + \frac{10}{27}\,y - \frac{5}{81} & \frac{1}{6} \le y < \frac{1}{4} \\ 320\,y^6 + 320\,y^5 - \frac{2600}{3}\,y^4 + \frac{4240}{9}\,y^3 - \frac{785}{9}\,y^2 + \frac{145}{27}\,y - \frac{35}{1296} & \frac{1}{4} \le y < \frac{1}{3} \\ -280\,y^6 + 560\,y^5 - \frac{1115}{3}\,y^4 + \frac{515}{9}\,y^3 + \frac{1525}{54}\,y^2 - \frac{565}{81}\,y + \frac{5}{16} & \frac{1}{3} \le y < \frac{5}{12} \\ 104\,y^6 - 240\,y^5 + 295\,y^4 - \frac{1985}{9}\,y^3 + \frac{775}{9}\,y^2 - \frac{7645}{648}\,y + \frac{5}{16} & \frac{5}{12} \le y < \frac{1}{2} \\ -20\,y^6 + 32\,y^5 - \frac{185}{9}\,y^3 + \frac{175}{36}\,y^2 + \frac{3371}{648}\,y - 1 & \frac{1}{2} \le y < \frac{2}{3} \\ 10\,y^6 - 38\,y^5 + \frac{160}{3}\,y^4 - \frac{265}{9}\,y^3 - \frac{115}{108}\,y^2 + \frac{4651}{648}\,y - 1 & \frac{2}{3} \le y < \frac{5}{6} \\ -2\,y^6 + 12\,y^5 - 30\,y^4 + 40\,y^3 - 30\,y^2 + 12\,y - 1 & \frac{5}{6} \le y < 1 \\ 1 & y \ge 1. \end{cases}$$

Example 4:. Let $X_1, X_2, \ldots, X_{10}$ be iid geometric(1 / 4) random variables (parameterized from 1). Find the mean and variance of $X_{(2)}$.

```
Y := OrderStat(GeometricRV(1 / 4), 10, 2);
Mean(Y);
Variance(Y);
```

yielding

$$\mu = \frac{305836589056}{239921705947}$$

and

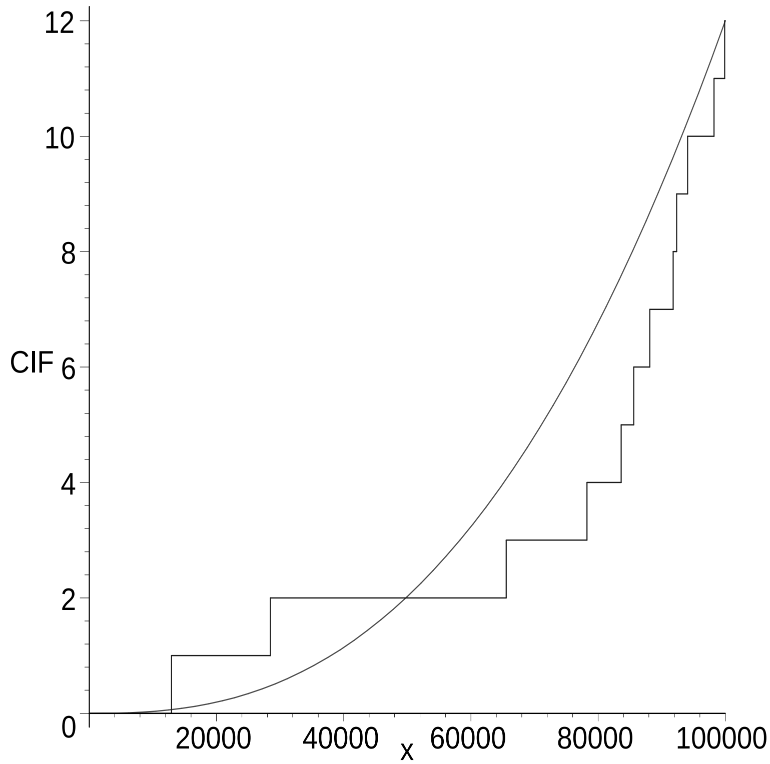
$$\sigma^2 = \frac{13969984570474695 22944}{52329477258653395 60619}$$

Example 5:. Fit a power law process to the odometer failure data over $(0, 100\,000]$:

12,942	28,489	65,561	78,254	83,639	85,603
88,143	91,809	92,360	94,078	98,231	99,900.

```
CarFailures := [12942, 28489, 65561, 78254,
  83639, 85603, 88143, 91809, 92360, 94078,
  98231, 99900];
X := WeibullRV(lambda, kappa);
hat := MLENHPP(X, CarFailures,
               [lambda, kappa], 100000);
PlotEmpVsFittedCIF(X, Sample, [lambda = hat[1],
  kappa = hat[2]], 0, 100000);
```

$$\hat{\lambda} \cong 0.000026317 \qquad \hat{\kappa} \cong 2.56800$$



6. Conclusions

(a) Nonparametric estimation and simulation for NHPPs is straightforward.

(b) Collecting data across overlapping intervals does not pose any significant problems.

(c) Once coded, this approach requires less effort than a parametric renewal process in order to simulate the observations.

(d) There may be potential for a "probability package" analogous to "statistical packages" such as SAS, SPSS, or S-Plus.

(e) I am searching for situations where an "exact" probability calculation is needed (typically not the case in economics, OR, engineering, classical statistics; possibly the case in biology, chemistry, physics).

Bibliography

Arkin, B., and Leemis, L., "Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process from Overlapping Intervals", *Management Science*, 46, 7, 2000, 989–998.

Drew, J.H., Glen, A.G., Leemis, L.M., "Computing the Cumulative Distribution Function of the Kolmogorov-Smirnov Statistic", *Computational Statistics and Data Analysis*, 34, 1, 2000, 1–15.

Evans, D.L. and Leemis, L.M., "Input Modeling Using a Computer Algebra System", in *Proceedings of the 2000 Winter Simulation Conference*, J.A. Joines, R.R. Barton, K. Kang, P.A. Fishwick, eds., Institute of Electrical and Electronics Engineers, Orlando, Florida, 2000, 577–586.

Glen, A.G., Leemis, L.M., and Drew, J.H., "A Generalized Univariate Change-of-Variable Transformation Technique", *INFORMS Journal on Computing*, 9, 3, 1997, 288–295.

Glen, A.G., Evans, D.L., and Leemis, L.M., "APPL: A Probability Programming Language", *The American Statistician*, 55, 2, 2001, 156–166.

Glen, A.G., Leemis, L.M., and Drew, J.H., "Computing the Distribution of the Product of Two Continuous Random Variables", forthcoming, *Computational Statistics and Data Analysis*.

Leemis, L.M., "Nonparametric Estimation of the Intensity Function for a Nonhomogeneous Poisson Process", *Management Science*, 37, 7, 1991, 886–900.