STAT331
Accuracy of Asymptotic and Simulated Approximations

We've discussed 2 ways of approximating processes of interest: the asymptotic process and the simulated version based on a multiplier CLT. Let's examine how well these perform for approximating the quantiles of the sup of a process. For simplicity, suppose we want to get an approximate 95% confidence band for the cumulative hazard function $\Lambda(t)$, for $0 \leq t \leq \tau$. We saw in Unit 15 that

$$\sqrt{n}[\hat{\Lambda}(\cdot) - \Lambda(\cdot)]/\sigma(\tau) \xrightarrow{\mathcal{L}} W\left[\frac{\sigma^2(\cdot)}{\sigma^2(\tau)}\right]$$

over $[0, \tau]$, where $W(\cdot)$ is a Wiener process, $\hat{\Lambda}(\cdot)$ is the Nelson-Aalen estimator, and

$$\sigma^2(t) \;=\; \int_0^t \frac{\lambda(u)}{[1-F(u)][1-G(u)]}du \;.$$

To get an $(1-\alpha)$ approximate confidence band for $\Lambda(\cdot)$, we used the fact that

$$\sup_{0\leq t\leq\tau} |\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]/\sigma(\tau)| \xrightarrow{\mathcal{L}} \sup_{0\leq u\leq 1} |W(u)| \;. \tag{18.1}$$

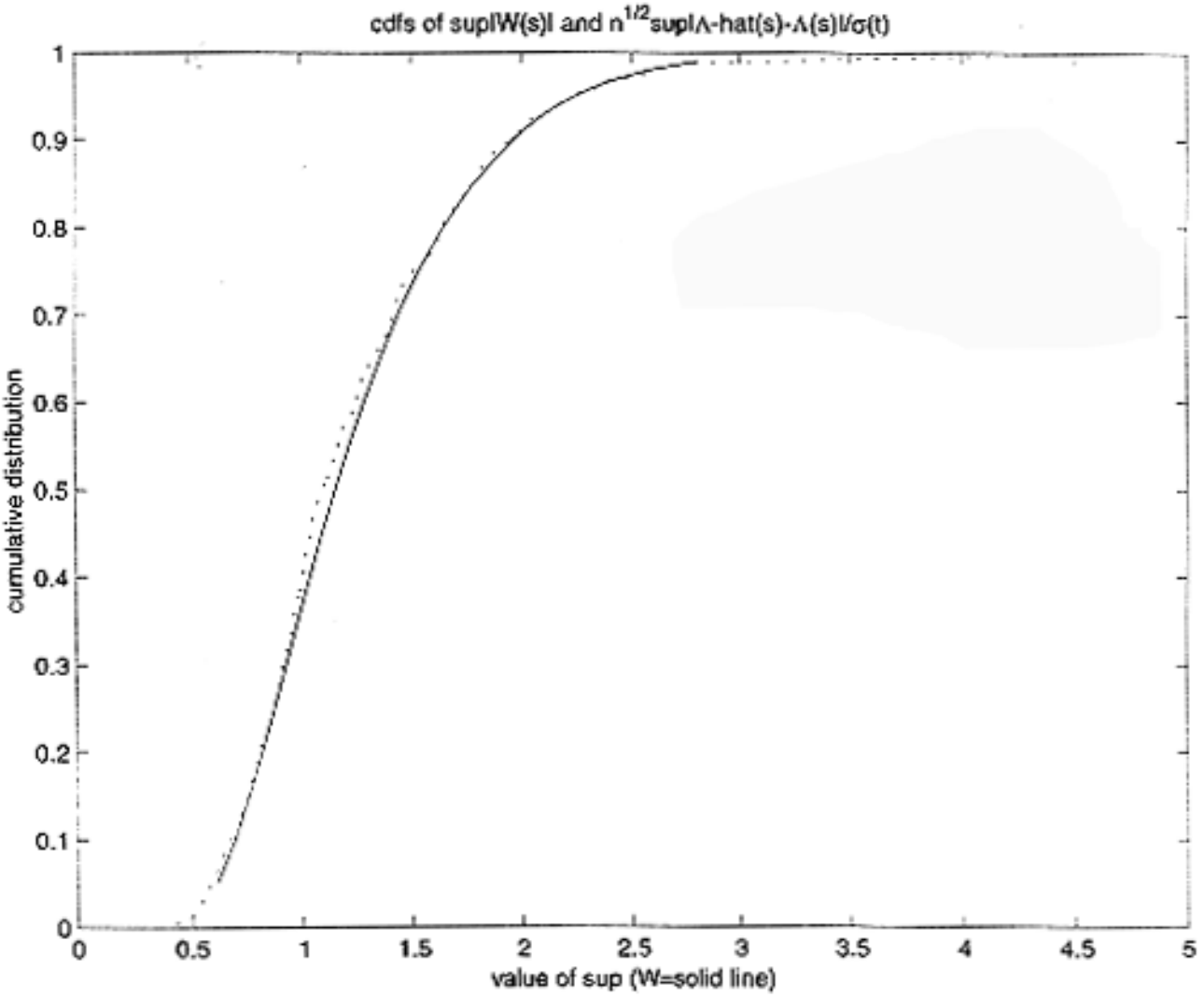
The right-hand side was used to find the constant that exceeded $(1-\alpha)\%$ of the distribution of this random variable and the left-hand side was inverted to find the band. Let's first examine how well the left-hand side of (18.1) approximates it's limiting Wiener distribution. To do this, suppose $n = 100, \tau = 1, T$ is NE(1), and C is NE(λ_c), where λ_c is selected as $1/3$ to give a 25% chance of censoring.

We generated M=500 sets of data of this type and for each one computed

$$\sup_{0\le t\le\tau} |\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]/\sigma(\tau)| \ .$$

We then plotted the c.d.f. of these 500 sups and compare this to the c.d.f. of $\sup_{0\le u\le 1} |W(u)|$. The results are shown in Figure 18-1. Note that the 2 c.d.f.s are quite similar. This suggests that using a quantile, say c_α, from the sup of a Wiener process to form a confidence band will yield a good approximation to the corresponding quantile of the exact distribution of $\sup_{0\le t\le\tau} |\sqrt{n}[\hat{\Lambda}(t)-\Lambda(t)]/\sigma(\tau)|$. MATLAB code for the program generating Figure 18-1 is given in program nelson1.m on the course webpage.

Figure 18-1


cdfs of sup|W(s)| and n^1/2 sup|Λ-hat(s)-Λ(s)|/σ(t)
cumulative distribution
value of sup (W=solid line)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Accuracy of Simulated Process: Now let's consider how well the simulation approach introduced in this section approximates the exact distribution of

$$\sup_{0\le t\le\tau} |\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]/\sigma(\tau)| \ .$$

Note that

$$\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]/\sigma(\tau) = \frac{\sqrt{n}}{\sigma(\tau)} \sum_{i=1}^{n} \int_0^t \frac{1[\overline{Y}(u) > 0]}{\overline{Y}(u)} dM_i(u)$$

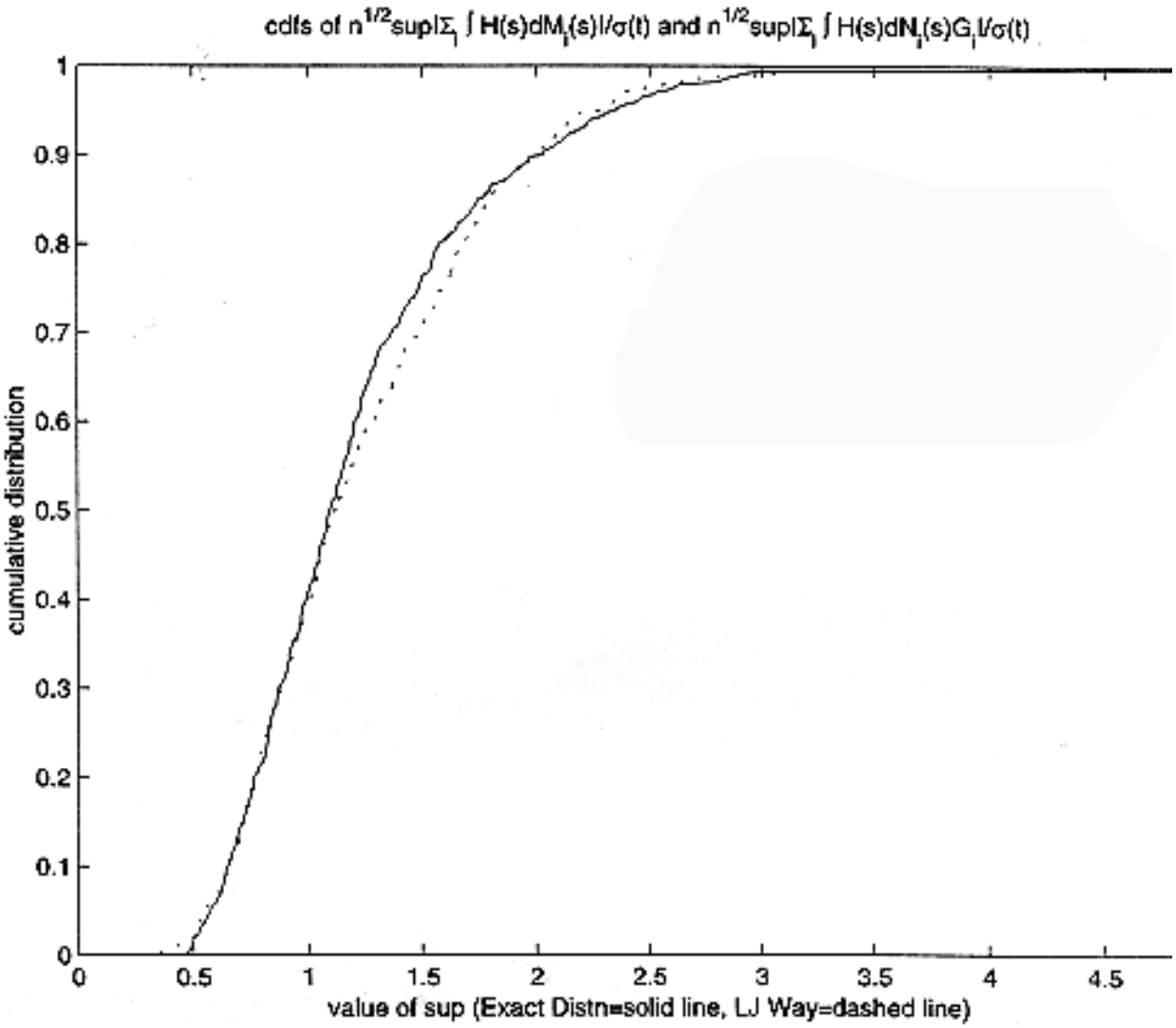
$$= \frac{\sqrt{n}}{\sigma(\tau)} \sum_{i=1}^{n} \int_0^t H(u) dM_i(u) \ ,$$

where $H(u) = \frac{1[\overline{Y}(u)>0]}{\overline{Y}(u)}$. To illustrate the performance, we generated a single data set of size n=100, where T is NE(1) and C is NE(λ_c), where, as before, τ=1 and λ_c is selected to be 1/3 to give a 25% chance of censoring. Then, for this single data set, we generated M=1000 sets of n=100 N(0,1) random variables $G_1, \cdots, G_{100}$ and for each set computed

$$\sup_{0\le t\le\tau} |\sqrt{n} \sum_{i=1}^{n} \int_0^t H(u) dN_i(u) G_i|/\sigma(\tau) \ . \tag{18.2}$$

The graph on the following page gives the c.d.f. of the 1000 sups generated according to (18.2), along with the c.d.f. of $\sup_{0\le t\le\tau} |\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]|/\sigma(\tau) = \sqrt{n} \sup_{0\le t\le\tau} |\sum_i \int H(t) dM_i(t)|/\sigma(\tau)$. Note the close agreement of these two curves. This suggests that using the simulation approach for approximating a quantile of $\sup_{0\le t\le\tau} |\sqrt{n}[\hat{\Lambda}(t) - \Lambda(t)]/\sigma(\tau)|$ will yield a reasonable approximation. The MATLAB program nelson2, located on the course webpage, generates a dataset and then uses the simulation approach to assess the accuracy of the latter.

Figure 18-2


cdfs of $n^{1/2}\sup|\Sigma_i \int H(s)dM_i(s)|/\sigma(t)$ and $n^{1/2}\sup|\Sigma_i \int H(s)dN_i(s)G_i|/\sigma(t)$
cumulative distribution
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
value of sup (Exact Distn=solid line, LJ Way=dashed line)

Program nelson1.m

Note: Before using the program nelson1.m.txt, you must remove the '.txt'.

This program compares the finite-sample distribution of

$$W_n \stackrel{def}{=} \sup_{0 \le t \le \tau} \mid \sqrt{n} \frac{[\hat{\Lambda}(t) - \Lambda(t)]}{\sigma(\tau)} \mid$$

to that of

$$W \stackrel{def}{=} \sup_{0 \le u \le 1} \mid W(u) \mid \ .$$

Recall that the sequence of random variables $\{W_n\}$ converges in distribution to W as $n \to \infty$. Thus, the comparison assesses the adequacy of the large-sample approximation to the distribution of W_n.

The program assumes that the underlying survival times $T_1, T_2, \cdots, T_n$ are i.i.d. exponentials with parameter $\lambda = 1$, and that the underlying censoring variables $C_1, C_2, \cdots, C_n$ are i.i.d. exponential with parameter lamc. Thus, $U_i = min\{T_i, C_i\}$ is exponential with parameter 1+lamc and $\delta_i = 1[T_i \le C_i]$ is Bernoulli with $P[\delta_i = 1]$ =lamc/(1+lamc).

Rather than specifying lamc, the user specifies τ and the probability that an observation is censored [i.e., 1/(1+lamc)]. In addition, the user specifies the sample size n and the number of simulations, nsims. The output consists of the input variables plus:

probtau=the probability of an uncensored observation between 0 and τ

prob10 and prob05= the proportions of W_n that exceed the 90^{th} and 95^{th} percentiles of the distribution of W

c10 and c05= the 90^{th} and 95^{th} percentiles of W_n, and

a plot of the c.d.f. of W versus the empirical c.d.f. of the nsims values of W_n.

Program nelson2.m

Note: Before using the program nelson2.m.txt, you must remove the '.txt'.

This program compares the finite-sample distribution of

$$W_n \stackrel{def}{=} \sup_{0 \le t \le \tau} \mid \sqrt{n} \frac{[\hat{\Lambda}(t) - \Lambda(t)]}{\sigma(\tau)} \mid$$

to that of

$$W_n^* = \sup_{0 \le t \le \tau} \mid \frac{\sqrt{n}}{\sigma(\tau)} \sum_{i=1}^{n} \int_0^t H(u) dN_i(u) G_i \mid,$$

where $G_1, \cdots, G_n$ are i.i.d. N(0,1) random variables.

The finite-sample distribution of W_n is obtained by ordinary simulation.

As in nelson1.m, it is assumed that the T_i are i.i.d. NE(1), and the C_i are i.i.d. NE(lamc), where lamc is chosen to give the desired amount of censoring. The user specifies n, τ, the probability of censoring, and the number of simulations used with W_n^*. The user also specifies the number of simulations used to find the distribution of W_n.

The output is a plot of the c.d.f. of W_n versus the empirical c.d.f. of W_n^*.

Program nelson3.m

Note: Before using the program nelson3.m.txt, you must remove the '.txt'.

This program plots simulated sample paths from the process

$$\sqrt{n}[\hat{\Lambda}(\cdot) - \Lambda(\cdot)] \ .$$

The program assumes that the underlying survival times $T_1, T_2, \cdots, T_n$ are i.i.d. exponentials with parameter $\lambda = 1$, and that the underlying censoring variables $C_1, C_2, \cdots, C_n$ are i.i.d. exponential with parameter lamc. Thus, $U_i = min\{T_i, C_i\}$ is exponential with parameter 1+lamc and $\delta_i = 1[T_i \leq C_i]$ is bernoulli with $P[\delta_i = 1]$ =lamc/(1+lamc).

Rather than specifying lamc, the user specifies τ and the probability that an observation is censored [i.e., 1/(1+lamc)]. In addition, the user specifies the sample size n and the number of sample paths.