Exam 1 is Wed. Feb. 15. **You are allowed 6 sheets of notes and a calculator.** The exam covers HW1-3 and Q1-3. Numbers refer to types of problems on exam.

In this class $\log(t) = \ln(t) = \log_e(t)$ while $\exp(t) = e^t$.

Let $T \geq 0$ be a nonnegative random variable.

Then the **cumulative distribution function** (cdf) $F(t) = P(T \leq t)$. Since $T \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

The probability density function (**pdf**) $f(t) = F'(t)$.

The **survival function** $S(t) = P(T > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

The **hazard function** $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

The **cumulative hazard function** $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

1) Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, be able to find the other 4 quantities for $t > 0$. See HW1: 1,3. **Know that each quantity is nonnegative.**

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1-F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tip: if $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

Tip: For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$is easier than integrating $h(t)$.

Know that if $T \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(T) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process, but constant hazard means that a used product is as good as a new product.

Know that if $T \sim$ Weibull(λ, γ) where $\lambda > 0$ and $\gamma > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is the most important distribution in reliability analysis.

2) Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of T: $P(T \leq t(p)) = F(t(p)) = p$ so $1-p = S(t(p)) = P(T > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used. See HW1, 4,5.

The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

If none of the survival times are censored, then the **empirical survival function** = (number of individual with survival times $> t$)/(number of individuals) $= a/n =$
$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^n I(T_i > t) = \hat{p}_t =$ sample proportion of lifetimes $> t$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let d_i = number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

$\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^m d_i = n$.

3) Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let $a_0 = n$ and $a_i = \sum_{i=1}^n I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{i=1}^n I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$. See HW2, 1.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

4) See HW2, 1. Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1 - \hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)].$$

5) See HW2, 1. Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n+4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1-\tilde{p}_{t_c}]}{n+4}} = \tilde{p}_{t_c} \pm 1.96 SE[\tilde{p}_{t_c}].$$

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ are the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_o = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$). Suppose that the following quantities are known: d_j = # deaths in I_j,
c_j = # of censored survival times in I_j,
n_j = # at risk in I_j = # who were alive and not yet censored at the start of I_j (at time t_{j-1}).
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

6) The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. If $t_m \neq \infty$, take $\hat{S}_L(t) = 0$ for $t \geq t_m$. **To graph** $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k}$$

for $k = 1, ..., m-1$.

7) Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output. See HW2 2b).

```
interval survival survival SE  or interval  survival survival SE
0     50   1.00         0             0     50   0.7594    0.0524
50   100   0.7594    0.0524          50    100   0.5889    0.0608
100  200   0.5889    0.0608         100    200   0.5253    0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the left output, and for the right endpoint for the right output. For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

8) See HW2 2d). A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96 \; SE[\hat{S}_L(t_i)].$$

9) Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 columns need to be given but the last 3 columns may need to be filled in. You may be given a table with all but a few entries filled. See HW3, 1.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j-d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0=0,t_1)$	d_1	c_1	n_1	$n_1-\frac{c_1}{2}$	$\frac{n'_1-d_1}{n'_1}$	$\hat{S}_L(t_o)=\hat{S}_L(0)=1$
$[t_1,t_2)$	d_2	c_2	n_2	$n_2-\frac{c_2}{2}$	$\frac{n'_2-d_2}{n'_2}$	$\hat{S}_L(t_1)=\hat{S}_L(t_0)\frac{n'_1-d_1}{n'_1}$
$[t_2,t_3)$	d_3	c_3	n_3	$n_3-\frac{c_3}{2}$	$\frac{n'_3-d_3}{n'_3}$	$\hat{S}_L(t_2)=\hat{S}_L(t_1)\frac{n'_2-d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1},t_k)$	d_k	c_k	n_k	$n_k-\frac{c_k}{2}$	$\frac{n'_k-d_k}{n'_k}$	$\hat{S}_L(t_{k-1})=\hat{S}_L(t_{k-2})\frac{n'_{k-1}-d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2},t_{m-1})$	d_{m-1}	c_{m-1}	n_{m-1}	$n_{m-1}-\frac{c_{m-1}}{2}$	$\frac{n'_{m-1}-d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2})=\hat{S}_L(t_{m-3})\frac{n'_{m-2}-d_{m-2}}{n'_{m-2}}$
$[t_{m-1},t_m=\infty)$	d_m	c_m	n_m	$n_m-\frac{c_m}{2}$	$\frac{n'_m-d_m}{n'_m}$	$\hat{S}_L(t_{m-1})=\hat{S}_L(t_{m-2})\frac{n'_{m-1}-d_{m-1}}{n'_{m-1}}$

10) Also get a 95% CI from output like that below. See HW2 2c).

```
time survival SDF_LCL SDF_UCL
0    1.0      1.0     1.0
50   0.7594   0.65666 0.86213 so the 95% CI for S(50) is (0.65666,0.86213)
```

Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \le Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let d_i = number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \ge 2$, then there are **ties**.

11) Let $n_i = \sum_{j=1}^n I(t_{(j)} \ge t_i)$ = # at risk at t_i = # alive and not yet censored just before t_i. Let d_i = # of events (deaths) at t_i. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and $\hat{S}_K(t_i) = \prod_{k=1}^i (1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i})$. $\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \le t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \le t \le t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

12) Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below. See HW3, 3a).

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

13) Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i), SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $(0.7782, 1)$. See HW3.3c).

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

14) In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$. See HW3.2de).

15) Let $P(Y \leq t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. See HW3.2b) and c).

```
Quartile estimates
Percent point estimate lower upper
75         .           220.0  .      CI not given
50       210.00        63.00 1296.00 t(.5) approx 210 and 95%CI is (63,1296)
25        63.00        18.00 195.00  t(.25) approx 63 and 95% CI is (18,195)
```

16) R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

17) If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t) dt = \int_0^\infty S_Y(t) dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.