Math 403 Exam 3 is Wed. Nov. 29. **You are allowed 15 sheets of notes and a calculator.** The exam covers HW7–10, and Q7–10. Numbers refer to types of problems on exam. Bring Exam1 1 review pages 1–2 to all exams.

79) For the **individual risk model**, the aggregate loss = total loss = $S = \sum_{i=1}^n X_i$ Assume the X_i are iid unless told otherwise: then $E(S) = nE(X)$ and $V(S) = nV(X)$. Sometimes $S = \sum_{i=1}^n X_i$ has a nice distribution. See 21).

80) For the **collective risk model**, $S = \sum_{i=1}^N X_i$. The distribution of S is called a compound distribution with N the primary distribution and X the secondary distribution. Assume the X_i are iid and $X_i \perp\!\!\!\perp N$ unless told otherwise: then $E(S) = E(N)E(X)$ and $V(S) = E(N)V(X) + [E(X)]^2V(N)$. Note that $S = 0$ if $N = 0$.

81) For both 79) and 80), often $S \sim AN(\mu = E(S), \sigma^2 = V(S))$. Then use the normal approximation to find i) $P(a < S < b) \approx P\left(\frac{a-\mu}{\sigma} < Z < \frac{a-\mu}{\sigma}\right)$ where $<$ can be replaced by $\leq$ unless S is discrete and a continuity correction is desired. ii) $\pi_p(S) = VaR_P(S) \approx \mu + \sigma z_p$. Often $f_S(x)$ will be used for a pdf when S is continuous and for a pmf $f_S(x) = P(S = x)$ when S is discrete. Let $S_S(x)$ and $F_S(x)$ be the survival function and cdf of S.

82) **Reinsurance** is insurance for aggregate losses that occur for an insurance company and guards against a bad year. (Reinsurance or) insurance on aggregate losses, subject to an aggregate deductible d, is called **stop–loss insurance**. The expected cost of this insurance is the **net stop-loss premium** = $E[(S-d)_+] = E(S) - E[S \wedge d]$. Get $E(S)$ from 79) or 80).

83) If S is continuous, then $E[(S-d)_+] = \int_d^\infty S_S(x)dx = \int_d^\infty (x-d)f_S(x)dx$, and $E[(S \wedge d)_+] = \int_0^d S_S(x)dx = \int_0^d xf_S(x)dx = dS_S(d)$.

84) If S is discrete, then $E[(S-d)_+] = \sum_{x>d}(x-d)f_S(x)$.

85) **Know** Suppose $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp ... \perp\!\!\!\perp X_n$ where $n = 3$ is common and the pmf $f_i(x) = P(X_i = x)$ is given for a few values of x. Find the pmf of $S = \sum_{i=1}^n X_i$ by using a tree diagram. The numbers on the branches of the tree add to s_i. Multiply the probabilities corresponding to the numbers on the branches to get $P(X_1 = x_1, ..., X_n = x_n)$ for that branch. Then accumulate probabilities from all branches that have $S = s_i = k$ to get $P(S = k)$. Alternatively, list the n values from left to right, change the rightmost values quickest. See the example below where X_i takes on the values 0 and 1, and $n = 3$.

X_1	X_2	X_3	$S = s_i$	$P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$
0	0	0	0	$f_1(0)f_2(0)f_3(0) = a_1$
0	0	1	1	$f_1(0)f_2(0)f_3(1) = a_2$
0	1	0	1	$f_1(0)f_2(1)f_3(0) = a_3$
0	1	1	2	$f_1(0)f_2(1)f_3(1) = a_4$
1	0	0	1	$f_1(1)f_2(0)f_3(0) = a_5$
1	0	1	2	$f_1(1)f_2(0)f_3(1) = a_6$
1	1	0	2	$f_1(1)f_2(1)f_3(0) = a_7$
1	1	1	3	$f_1(1)f_2(1)f_3(1) = a_8$

k	0	1	2	3
$P(S = k)$	a_1	$a_2 + a_3 + a_5$	$a_4 + a_6 + a_7$	a_8

86) Sometimes S is discrete and want $E[(S-d)_+] = E(S) - E[S \wedge d]$. Suppose S

takes on values $s_0, s_1, s_2, ...,$ where often $s_i = hi$ for some positive integer h and $d = s_i$ where i is small, often 1 or 2. Then $E(S \wedge s_i) = \sum_{k=0}^{\infty} \min(s_k, s_i) P(S = s_k) =$
$s_0 P(S = s_0) + s_1 P(S = s_1) + \cdots + s_{i-1} P(S = s_{i-1}) + \sum_{k=i}^{\infty} s_i P(S = s_k) =$
$s_0 P(S = s_0) + s_1 P(S = s_1) + \cdots + s_{i-1} P(S = s_{i-1}) + s_i P(S \geq s_i) =$
$s_0 P(S = s_0) + s_1 P(S = s_1) + \cdots + s_{i-1} P(S = s_{i-1}) +$
$s_i[1 - P(S = s_0) - P(S = s_1) - \cdots - P(S = s_{i-1})]$. In particular, if $d = s_1$ then $E(S \wedge s_1) = s_0 P(S = s_0) + s_1[1 - P(S = s_0)]$, and if $d = s_2$, then
$E(S \wedge s_2) = s_0 P(S = s_0) + s_1 P(S = s_1) + s_2[1 - P(S = s_0) - P(S = s_1)]$.

This technique is called a **convolution method.** Could start the numbering at $s_1, s_2, ...$ (at s_1 instead of s_0). Assume the X_i are iid and take on values $x_0, x_1, ..., x_m$ where m is small. Then find $E(X)$. Many variants are possible, and sometimes several combinations of $N, X_1, ..., X_N$ will result in $S = s_k$. See HW7 2, 3.

I) Suppose $S = \sum_{i=1}^{N} X_i$. Then $E(S) = E(N)E(X)$. **Usually N will be Poisson, binomial, negative binomial or geometric.** Then $s_0 = 0$.

Assume $x_0 > 0$. Then $P(S = 0) = P(N = 0)$, and $P(S = s_1) = P(N = 1, X = s_1) = P(N = 1)P(X = s_1)$ where $s_1 = x_0$. Note that $s_2 = \min(2x_0, x_1)$.

Use these facts to find $E[(S \wedge s_2)]$. Often x_i will be a multiple of $s_1 = x_0$ if $x_0 > 0$: S takes on values $s_i = (i)s_1$ for $i = 1, 2, ...$, but you only need to find s_0, s_1, and s_2. Often $s_i = i$ for $i = 0, 1, 2, ...$

II) Suppose $S = \sum_{i=1}^{n} X_i$ where n is small, often 3. Then $E(S) = nE(X)$. Then the smallest value of S is $s_0 = nx_0$. Then $P(S = s_0) = [P(X = x_0)]^n$.

Often $x_i = i$ and $s_i = i$ for $i = 0, 1, ...$

87)

j	x	$F_S(x)$	$E[(S - jh)_+]$	
0	0	$F_S(0) = 0$	$E(S)$	
1	h	$F_S(h)$	$E(S - h)_+ =$	$E(S) - h(1 - F_S(0))$
2	2h	$F_S(2h)$	$E(S - 2h)_+ =$	$E(S - h)_+ - h(1 - F_S(h))$
3	3h	$F_S(3h)$	$E(S - 3h)_+ =$	$E(S - 2h)_+ - h(1 - F_S(2h))$
4	4h	$F_S(4h)$	$E(S - 4h)_+ =$	$E(S - 3h)_+ - h(1 - F_S(3h))$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Suppose $E(S)$ is given where S is discrete and $P(S = kh) > 0$ for some integer $h > 0$ and $k = 0, 1, 2, ...$ Assume $P(S = x) = 0$ for all other values of x. If $d = jh$ where j is a nonnegative integer, then $E[(S - d)_+] = h \sum_{m=1}^{\infty} (1 - F_S[(m + j)h])$, and there is a recursion $E([S - (j + 1)h]_+] = E[(S - jh)_+] - h[1 - F_S(jh)]$. The above table replaces j by $j - 1$. Given a partially filled table similar to the one above, you should be able to find the missing value or next value of $E[(S - jh)_+]$. Also, if $P(S = a) > 0, P(S = b) > 0$ and $P(a < S < b) = 0$, for $a < d < b$ use linear interpolation to find $E[(S - d)_+] = \dfrac{b - d}{b - a} E[(S - a)_+] + \dfrac{d - a}{b - a} E[(S - b)_+]$.

88) Suppose $S_j = \sum_{i=1}^{N_j} X_{ij}$ has a compound Poisson distribution with $N_j \sim$ Poisson(λ_j) and X_{ij} has cdf $F_j(x)$ for $j = 1, ..., n$ where $S_1 \perp\!\!\!\perp S_2 \perp\!\!\!\perp ... \perp\!\!\!\perp S_n$. Then $S = \sum_{i=1}^{n} S_i = \sum_{k=1}^{N} W_k$ has a compound Poisson distribution where $N \sim$ Poisson($\lambda = \sum_{i=1}^{n} \lambda_i$) and W_k has cdf $F_W(x) = \displaystyle\sum_{j=1}^{n} \frac{\lambda_j}{\lambda} F_j(x)$, an n–point mixture of the X_{ij} distributions, $j = 1, ..., n$.

89) Suppose X has a mixture distribution or mixed distribution where parameter Λ is a RV. Hence $X|\Lambda = \lambda$ has a conditional pdf or pmf $f_{X|\Lambda}(x|\lambda)$ where Λ has marginal or unconditional pdf or pmf $f_\Lambda(\lambda)$. Then the marginal or unconditional pdf or pmf of X is $f_X(x) = \int_{-\infty}^{\infty} f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda$ if Λ is continuous, and $f_X(x) = \sum_\lambda f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)$ if Λ is discrete. Note the value x is fixed.

90) Suppose $f_X(x)$ is defined on 0, 1, 2, ..., m where $m = \infty$ is possible. Then $S = \sum_{i=1}^N X_i$ is discrete. Let $f_0 = P(X = 0)$. Then $P(S = 0)$ is tabled below.

distribution of N	$f_S(0) = P(S = 0)$
Pois(λ)	$\exp[\lambda(f_0 - 1)]$
bin(q, m)	$[1 + q(f_0 - 1)]^m$
NB(β, r)	$[1 + \beta(1 - f_0)]^{-r}$
Geom(β)	$[1 + \beta(1 - f_0)]^{-1}$

91) There is a recursion. Under the conditions of 90), $f_S(x) = P(S = x) = \dfrac{\sum_{y=1}^{x \wedge m} \left(a + \frac{by}{x}\right) f_X(y) f_S(x - y)}{1 - a f_X(0)}$ for $x = 1, 2, ...$, where N is from an $(a, b, 0)$ distribution. Typically x is small, so $x \wedge m = x$.

92) Suppose $P(X > 0) = 1$ and $v = P(X > d) = S_X(d)$. Let N = number of claims when there is no deductible and let N_{new} be the number of claims when there is a deductible d. Often $N = N^L$ and $N_{new} = N^P$.

distribution of N	of N_{new}
Pois(λ)	Pois($v\lambda$)
bin(q, m)	bin(vq, m)
NB(β, r)	NB($v\beta, r$)
Geom(β)	Geom $v\beta$)

93) Under the conditions of 92) suppose insurance with deductible d_1 is changed to insurance with deductible d_2. Let γ be the parameter that is revised (λ, q or β). Then $\gamma_{new} = \dfrac{S_X(d_2)}{S_X(d_1)} \gamma$. Note that γ_{new} decreases if $d_2 > d_1$ and increases if $d_2 < d_1$. When γ_{new} decreases, there are more payments of 0 and fewer positive payments.

94) Under the conditions of 92), let $N = N^L$ and $N_{new} = N^P$. Then $S = \sum_{i=1}^{N^L} Y_i^L = \sum_{i=1}^{N^P} Y_i^P$. Then using the per loss basis, $E(S) = E(N^L)E(Y^L)$ and $V(S) = E(N^L)V(Y^L) + [E(Y^L)]^2 V(N^L)$. It is assumed that N^L does not change under a coverage modification (usually a change in deductible), but N^P does. Using the per payment basis, $E(S) = E(N^P)E(Y^P)$, can be useful if $E(Y^P) = e_X(d)$ has a useful formula. See 60).

STATISTICS 95) Suppose that a RV W has a parametric distribution that has a vector of parameters $\boldsymbol{\theta}$ that can take on values in the *parameter space* Θ. Often $\Theta = \{\boldsymbol{\theta} | f(w|\boldsymbol{\theta})$ is a pdf or pmf$\}$.

96) Let $E(\hat{\theta}) = E(\hat{\theta}|\theta) = E_\theta(\hat{\theta})$ be the expected value of the estimator $\hat{\theta}$ when the true parameter is θ.

97) The estimator $\hat{\theta}$ is an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for all θ (often for all $\theta \in \Theta$).

98) The **bias** of an estimator $\hat{\theta}$ of θ is $\text{bias}_{\hat{\theta}}(\theta) = \text{E}(\hat{\theta}) - \theta = \text{E}(\hat{\theta} - \theta)$. Note that an unbiased estimator has $\text{bias}_{\hat{\theta}}(\theta) \equiv 0$ (the bias is 0 for all θ).

99) Let $\hat{\theta}_n$ be an estimator of θ based on a sample of size n (often n is suppressed).

Then $\hat{\theta}_n$ is an **asymptotically unbiased** estimator of θ if $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$ for all θ. Note that unbiased estimators are asymptotically unbiased.

100) An estimator $\hat{\theta}$ is a **consistent estimator** of θ if for all $\delta > 0$ and for any $\theta \in \Theta$, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \delta) = 0$. Equivalently, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \delta) = 1$.

101) The **mean square error** of an estimator $\hat{\theta}_n$ of θ is $MSE_{\hat{\theta}_n}(\theta) = E[(\hat{\theta}_n - \theta)^2] = V(\hat{\theta}_n) + [\text{bias}_{\hat{\theta}_n}(\theta)]^2$.

102) The estimator $\hat{\theta}_n$ is a consistent estimator of θ if i) $MSE_{\hat{\theta}_n}(\theta) \to 0$ as $n \to \infty$, or if ii) $E(\hat{\theta}_n) \to \theta$ (so the bias $\to 0$) and $V(\hat{\theta}_n) \to 0$ as $n \to \infty$.

103) Let $\hat{\theta}_1 = \hat{\theta}_{1,n}$ and $\hat{\theta}_2 = \hat{\theta}_{2,n}$ be two estimators of θ. If $MSE_{\hat{\theta}_1}(\theta) \le MSE_{\hat{\theta}_2}(\theta)$ for all $\theta \in \Theta$, then $\hat{\theta}_1$ is a "better" estimator than $\hat{\theta}_2$, according to the MSE criterion.

104) The unbiased sample variance $S_U^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. If $X_1, ..., X_n$ are iid with $V(X_i) = \sigma^2$, then $E(S_U^2) = \sigma^2$.

105) A biased estimator of $V(X_i) = \sigma^2$ is $S_E^2 = \dfrac{n-1}{n} S_U^2 = \dfrac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. S_E^2 is the variance of the empirical distribution, and $E(S_E^2) = \dfrac{n-1}{n}\,\sigma^2$.

106) A *point estimator* $\hat{\theta}_n$ gives a single value (point) as an estimate. An *interval estimator* gives an interval of reasonable values.

107) A $100(1-\alpha)\%$ confidence interval (CI) (L, U) for θ satisfies $P(L < \theta < U) \ge 1-\alpha$ for all θ.

A large sample $100(1-\alpha)\%$ CI (L_n, U_n) for θ satisfies $P(L_n < \theta < U_n) \to 1-\delta \ge 1-\alpha$ for all θ.

108) Often for CIs, $t_{\alpha/2,n-1}$ of $z_{\alpha/2}$ is an upper cutoff or upper percentile: $P(T > t_{\alpha/2,n-1}) = \alpha/2$ if $T \sim t_{n-1}$ and $P(Z > z_{\alpha/2}) = \alpha/2$ if $Z \sim N(0,1)$.

The same notation was used for a percentile: $P(T \le t_{\alpha/2,n-1}) = \alpha/2$ and $P(Z \le z_{\alpha/2}) = \alpha/2$. Hence context must be used to determine whether $t_{\alpha/2,n-1}$ and $z_{\alpha/2}$ are upper cutoffs or percentiles.

109) If RV X comes from a parametric distribution with parameter $\boldsymbol{\theta}$, then say $X \sim PD(\boldsymbol{\theta})$. If $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, use $X \approx PD(\hat{\boldsymbol{\theta}})$ to estimate quantities in point 0) of exam 1 review such as $F(x), E(X), V(X), S(x), e_X(d) = E(Y^P), \pi_p(X) = VaR_P(X), E(X \wedge d)$, and $TVaR_P(X)$.

110) In a test of hypotheses, $H_0 : \theta \in \Theta_0$ is the *null hypothesis* and $H_1 : \theta \in \Theta_1$ is the *alternative hypothesis*. Reject H_0 if the test statistic is in a critical region (often $(-\infty, a]$ or $[a, \infty)$). Finite boundaries of a critical region are called *critical values* (eg a).

111) The *p–value* is the probability that a test statistic takes on a value that is less in agreement (more extreme) with the null hypothesis than the observed value of the test statistic. For an α level test, reject H_0 if p–value $< \alpha$. Fail to reject H_0 if p–value $> \alpha$.

112) A *type I error* occurs if the test rejects H_0 when H_0 is true. The significance **level** of the test is $\alpha = \max_{\theta \in \Theta_0} P(\text{reject H}_0|\text{H}_0 \text{ is true})$. Typically

$\alpha = \max_{\theta \text{ a critical point}} P(\text{reject H}_0|\text{H}_0 \text{ is true})$.

113) Let $X_1, ..., X_n$ be iid random variables from a distribution with cdf F, mean μ and variance σ^2. Let $x_1, ..., x_n$ be the observed values of the X_i. The distribution of the

RV D is the *empirical distribution* if D is a discrete RV with the following pmf.

x	x_1	x_2	$\cdots$	x_n
$P(D = x)$	$1/n$	$1/n$	$\cdots$	$1/n$

Then $E(D) = \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $V(D) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$. Note that $E(D) = \overline{x}$ is the observed sample mean, and $V(D)$ is the observed empirical sample variance. **Often "observed"** is omitted. If the x_i are not distinct, then let k_j = number of $x_i = x_j$, then $P(D = x_j) = k/n$, but this just combines columns in the above table that have $x_i = x_j$.

114) Let the unbiased sample variance $S_U^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Let the empirical sample variance $S_E^2 = \frac{n-1}{n} S_U^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. Then under the conditions of (113), $E(S_U^2) = \sigma^2$ and $E(S_E^2) = \frac{n-1}{n}\sigma^2$.

115) The empirical estimators of quantities like $F(x), S(x), H(x)$, and $f(x)$ will be denoted by $F_n(x), S_n(x), H_n(x)$, and $f_n(x)$. Other estimators will be denoted as $\hat{F}(x), \hat{S}(x)$, $\hat{H}(x)$, and $\hat{f}(x)$. When the RVs X_i are used, the estimators are statistics (RVs). The observed values use the x_i. Hence as a statistic (random variable), the empirical cdf $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x) = \frac{\text{number of } \mathrm{X_i} \leq \mathrm{x}}{n}$. The observed value of the statistic (empirical cdf) is $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(x_i \leq x) = \frac{\text{number of } \mathrm{x_i} \leq \mathrm{x}}{n}$, a nondecreasing step function that can be plotted. Here the indicator random variable $W_i = I(X_i \leq x) = 1$ if $X_i \leq x$ and $W_i = I(X_i \leq x) = 0$ if $X_i > x$. Hence under the conditions of 113), the W_i are iid Bernoulli($q = F(x)$) RVs. Fix x. By the CLT, $\sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, F(x)(1 - F(x)) = N(0, S(x)(1 - S(x))$. So for fixed x, $F_n(x) \sim AN(F(x), S(x)(1 - S(x))$.

116) The empirical survival function $S_n(x) = 1 - F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i > x) = \frac{\text{number of } \mathrm{X_i} > \mathrm{x}}{n}$. The empirical cumulative hazard function $H_n(x) = -\ln(S_n(x))$.

Get the observed values by replacing X_i by x_i. Hence the observed value of $S_n(x)$ is $S_n(x) = \frac{1}{n}\sum_{i=1}^n I(x_i > x) = \frac{\text{number of } \mathrm{x_i} > \mathrm{x}}{n}$. The (observed) empirical pdf or pmf is the pmf $f_n(x) = \frac{\text{number of } \mathrm{x_i} = \mathrm{x}}{n}$, and is best when the underlying distribution of the X_i is discrete.

117) Let $y_1 < y_2 < \cdots < y_k$ be the k distinct values of $x_1, ..., x_n$ that appear in a sample of size $n \geq k$. Let s_j = number of times y_j appears in the sample with $\sum_{j=1}^k s_j = n$. Let $r_j = \sum_{i=j}^k s_j$ = number of observations $\geq y_j$. So $r_1 = n$ and $r_k = s_k$.

118) The *order statistics* are $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The observed order statistics are $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Given a small data set, order the data from smallest to largest and make the following table. It is often useful to get the column of y_j first. For complete data, $r_j = r_{j-1} - s_{j-1} = \sum_{i=j}^k s_j$ with $r_1 = n$.

j	y_j	s_j	r_j
1	y_1	s_1	$r_1 = n = \sum_{i=1}^k s_j$
2	y_2	s_2	$r_2 = r_1 - s_1 = \sum_{i=2}^k s_j$
3	y_3	s_3	$r_3 = r_2 - s_2 = \sum_{i=3}^k s_j$
4	y_4	s_4	$r_4 = r_3 - s_3 = \sum_{i=4}^k s_j$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k-1$	y_{k-1}	s_{k-1}	$r_{k-1} = r_{k-2} - s_{k-2} = \sum_{i=k-1}^k s_j$
k	y_k	s_k	$r_k = r_{k-1} - s_{k-1} = s_k$

119) Given a table as in 118), be able to find $F_n(y_j) = \dfrac{\sum_{i=1}^j s_i}{n} = 1 - \dfrac{r_{j+1}}{n}$ where $r_{k+1} = 0$.

$$F_n(x) = \begin{cases} 0, & x < y_1 \\ 1 - \frac{r_j}{n}, & y_{j-1} \le x < y_j, \ j = 2, ..., k \\ 1, & y_k \le x \end{cases}$$

$$F_n(x) = \begin{cases} 0 = 1 - \frac{n}{n}, & x < y_1 \\ 1 - \frac{r_2}{n}, & y_1 \le x < y_2 \\ 1 - \frac{r_3}{n}, & y_2 \le x < y_3 \\ \vdots & \vdots \\ 1 - \frac{r_{k-1}}{n}, & y_{k-2} \le x < y_{k-1} \\ 1 - \frac{r_k}{n}, & y_{k-1} \le x < y_k \\ 1 = 1 - \frac{0}{n}, & y_k \le x \end{cases}$$

120) Given a table as in 118), be able to find the **Nelson Aalen** estimator $\hat{H}(x)$ of the cumulative hazard rate function $H(x)$. This estimate is a step function with $\hat{H}(y_j) = \displaystyle\sum_{i=1}^j \frac{s_i}{r_i} = \sum_{i=1}^{j-1} \frac{s_i}{r_i} + \frac{s_j}{r_j} = \hat{H}(y_{j-1}) + \frac{s_j}{r_j}$ with $\hat{H}(y_1) = \dfrac{s_1}{r_1}$.

$$\hat{H}(x) = \begin{cases} 0, & x < y_1 \\ \hat{H}(y_1) = 0 + \frac{s_1}{r_1}, & y_1 \le x < y_2 \\ \hat{H}(y_2) = \hat{H}(y_1) + \frac{s_2}{r_2}, & y_2 \le x < y_3 \\ \vdots & \vdots \\ \hat{H}(y_{k-2}) = \hat{H}(y_{k-3}) + \frac{s_{k-2}}{r_{k-2}}, & y_{k-2} \le x < y_{k-1} \\ \hat{H}(y_{k-1}) = \hat{H}(y_{k-2}) + \frac{s_{k-1}}{r_{k-1}}, & y_{k-1} \le x < y_k \\ \hat{H}(y_k) = \hat{H}(y_{k-1}) + \frac{s_k}{r_k}, & y_k \le x \end{cases}$$

121) An alternative to $F_n(x)$ is $\hat{F}(x) = 1 - \exp(-\hat{H}(x))$.

122) For **grouped data**, the complete data $x_1, ..., x_n$ are not known but is known how many observations x_i fall in groups $(c_0, c_1], (c_1, c_2], (c_2, c_3], ..., (c_{k-2}, c_{k-1}], (c_{k-1}, c_k]$ where use $[c_0$ if $x = c_0$ is possible and $\infty)$ if $c_k = \infty$. Let n_j = number of observations falling in $(c_{j-1}, c_j]$ where $\sum_{i=1}^k n_i = n$. Then $F_n(c_j) = \dfrac{1}{n} \displaystyle\sum_{i=1}^j n_i = \dfrac{\text{number of } x_i \le c_j}{n}$ for $j = 1, ..., k$. Often only the middle two columns of the table below are given.

j	interval	n_j	$F_n(c_j) = \frac{1}{n}\sum_{i=1}^j n_i$
1	$[0 = c_0, c_1]$	n_1	$F_n(c_1) = n_1/n$
2	$(c_1, c_2]$	n_2	$F_n(c_2) = (n_1 + n_2)/n$
3	$(c_2, c_3]$	n_3	$F_n(c_3) = (n_1 + n_2 + n_3)/n$
4	$(c_3, c_4]$	n_4	$F_n(c_4) = (n_1 + \cdots + n_4)/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k-1$	$(c_{k-2}, c_{k-1}]$	n_{k-1}	$F_n(c_{k-1}) = \sum_{i=1}^{k-1} n_i/n$
k	$(c_{k-1}, c_k]$	n_k	$F_n(c_k) = 1 = \sum_{i=1}^{k} n_i/n$

123) For grouped data, an **ogive** $F_n(x)$ is obtained by connecting the values of $F_n(c_j)$ in 122) with straight lines where $F_n(0) = 0$ for a nonnegative RV X. Since linear interpolation is used, the ogive is continuous with $F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j)$ for $c_{j-1} \le x \le c_j$.

124) For the grouped data table of 122), the empirical density function (pdf for continuous data, histogram of pmf for discrete data) is the right continuous **histogram** or relative frequency histogram $f_n(x) = \frac{n_j}{n(c_j - c_{j-1})}$ for $c_{j-1} \le x < c_j$.

125) For the ogive and histogram using grouped data of 122), a different **uniform distribution** is assumed for each interval, where the height of the uniform distribution is given by $f_n(x)$ of 124). Take $f_n(x)$ to be the pdf for each interval if told to assume a uniform distribution on each interval.

126) If given intervals $c_0 - -c_1, c_1 - -c_2, ..., c_{k-1} - -c_k$ (without the open or closed parentheses or brackets,) assume $F_n(c_j) = \frac{1}{n}\sum_{i=1}^j n_i$ as in 122).

Estimation for Modified Data

127) An observation x_i is **truncated** at d, or left truncated at d, or truncated below at d, if x_i is not recorded if $x_i \le d$, but is recorded if $x_i > d$. For example, if there is a deductible d, and the loss $x_i \le d$, the insured policy holder will not report the loss since no benefit will be paid. Assume the loss x_i is reported if $x_i > d$ to get the benefit.

128) An observation x_i is **censored** at u, or censored above at u, or right censored at u, if x_i is recorded as u for $x_i \ge u$ and recorded as x_i for $x_i \le u$. For example, if there is a maximum payment u, then $w_i = x_i \vee u = \max(x_i, u)$ is the censored value of x_i.

129) Suppose $X \ge 0$, $x =$ time, and $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_k = [x_0, x_1) \cup [x_1, x_2) \cup \cdots \cup [x_{k-1}, x_k)$ where $I_i = [x_{i-1}, x_i)$, $x_0 = 0$ and $x_k = \infty$. Let $p_i = P($surviving through $I_i|$ survived at the start of $I_i) = P(X > x_i | X > x_{i-1}) = \frac{S_X(x_i)}{S_X(x_{i-1})}$. Then $S_X(x_j) = \prod_{i=1}^j p_i$ where $S_X(0) = S_X(x_0) = 1$. Often use $\hat{p}_i = 1 - \frac{\text{number with an event in I}_\text{i}}{\text{number with potential for an event in I}_\text{i}}$. Often the event is "dying" or "failing". If x is not time, the event could be an x_j (such as a loss) in the interval $I_i = [x_{i-1}, x_i)$.

130) The jth observation needs a truncation point d_j. Use $d_j = 0$ if the observation is not truncated. If the observation is truncated, then $d_j > 0$. The jth observation

$$= \begin{cases} x_j, & x_j \text{ not censored} \\ u_j, & x_j \text{ censored at u}_\text{j}. \end{cases}$$

Hence the jth observation could be truncated ($d_j > 0$) or not truncated ($d_j = 0$). Suppose that there are m observations x_i or u_i where L is the number of x_i. Let $y_1 < y_2 < \cdots < y_k$

be the k unique values of the x_i where $k \leq L$. Let s_j be the number of times the uncensored observation y_j appears in the sample = number of $x_i = y_j$. The number of d_i's is equal to m.

131) Let r_j be the size of the risk set (number at risk or under observation) at value y_j. Often the value is time or age. The risk set includes observations with truncation values $d_i < y_j$ **and** either $x_i \geq y_j$ or censored at values $u_i \geq y_j$. Hence
r_j = number of x_i's $\geq y_j$ + number of u_i's $\geq y_j$ − number of d_i's $\geq y_j$
= number with x_i or u_i values $\geq y_j$ ignoring truncation − number not in risk set because truncation value $d_i \geq y_j$, and
r_j = number of d_i's $< y_j$ − number of x_i's $< y_j$ − number of u_i's $< y_j$
= number who entered study before (value) y_j − number who have left study (eg due to death or censoring) by (value) y_j.
Also $r_j = r_{j-1}$ + number of $d_i \in [y_{j-1}, y_j)$ $-s_{j-1}$ − number of $u_i \in [y_{j-1}, y_j)$ with $r_0 = 0$, $s_0 = 0$ and $y_0 = 0$.

132) **Know**:

j	y_j	s_j	r_j
1	y_1	s_1	r_1
2	y_2	s_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k-1$	y_{k-1}	s_{k-1}	r_{k-1}
k	y_k	s_k	r_k

Given the above table, as in 120),

a) the **Nelson Aalen** estimator $\hat{H}(x) = \hat{H}(y_{j-1}) = \sum_{i=1}^{j-1} \frac{s_i}{r_i}$ for $y_{j-1} \leq x < y_j$ with $\hat{H}(y_1) = \frac{s_1}{r_1}$. This estimator is a nondecreasing step function. Note that $\hat{H}(y_j) = \sum_{i=1}^{j} \frac{s_i}{r_i} = \hat{H}(y_{j-1}) + \frac{s_j}{r_j}$.

b) The **Kaplan Meier product limit estimator**, or Kaplan Meier (KM) estimator, or product limit (PL) estimator $S_n(x) = S_n(y_{j-1}) = \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right)$ for $y_{j-1} \leq x < y_j$ where $S_n(0) = 1$ and $S_n(x) = 1$ for $0 \leq x < y_1$. This estimator is a nonincreasing step function. Note that $S_n(y_j) = \prod_{i=1}^{j} \left(1 - \frac{s_i}{r_i}\right) = S_n(y_{j-1}) \left(1 - \frac{s_j}{r_j}\right)$. $\quad H_n(x) = -\ln(S_n(x))$.

133) Given a table of i, d_i, x_i, and u_i, be able to make a table of j, y_j, s_j, and r_j as in 132). Often $d_1 \leq d_2 \leq \cdots \leq d_m$, and for the L $d_i = 0$,
x_1 or $u_1 \leq x_2$ or $u_2 \leq \cdots \leq x_L$ or u_L.

134) Let $S^* = S_n(y_k) = \prod_{i=1}^{k} \left(1 - \frac{s_i}{r_i}\right)$. Can define $S_n(x) = S^*$ for $x > y_k$, or $S_n(x) = 0$ for $x > y_k$ (especially if $s_k = r_k$ so $S^* = 0$). Alternatively, the text uses $S_n(x) = S^*$ for $y_k \leq x < w$ and $S_n(x) = 0$ or $S_n(x) = S^*$ or $S_n(x) = (S^*)^{x/w}$ for $x \geq w$ where w is the largest of the x_i and u_i (the largest observed censored or uncensored survival value (time) from the data).

135) An alternative to $S_n(x)$ is $\hat{S}(x) = e^{-\hat{H}(x)} = \exp(-\hat{H}(x))$. Let $S^* = \hat{S}(y_k)$. Can define $\hat{S}(x) = S^*$ for $x > y_k$, or $\hat{S}(x) = 0$ for $x > y_k$. Alternatively, the text uses $\hat{S}(x) = S^*$ for $y_k \le x < w$ and $\hat{S}(x) = 0$ or $\hat{S}(x) = S^*$ or $\hat{S}(x) = (S^*)^{x/w}$ for $x \ge w$ where w is the largest of the x_i and u_i (the largest observed censored or uncensored survival value (time) from the data).

136) Suppose $d_{(1)} = \min(d_1, ..., d_m) > 0$. Then $\hat{S}(0) = S_n(0) = 1$, but there is not enough information to define $S_n(x)$ or $\hat{S}(x)$ for $x \in (0, d_{(1)})$. So $S_n(x)$ and $\hat{S}(x)$ are defined for $x > d_{(1)}$.

137) For the empirical estimators $F_n(x)$ and $S_n(x)$ with complete data, $\hat{V}(F_n(x)) = \hat{V}(S_n(x)) = \dfrac{S_n(x)F_n(x)}{n} = \dfrac{S_n(x)(1 - S_n(x))}{n}$, and $\widehat{Cov}(F_n(x), F_n(y)) = \dfrac{F_n(x)(F_n(y) - F_n(x))}{n}$ where $x < y$. Since x and y are fixed, it might be useful to use t or z as the dummy variable, eg $F_n(z)$.

138) The empirical distributions are discrete.

a) $P(a < X \le b) = F(b) - F(a) = S(a) - S(b) = P(X \le b) - P(X \le a)$.

b) $P(a \le X \le b) = F(b) - F(a-) = S(a-) - S(b) = P(X \le b) - P(X < a)$.

c) $P(a \le X < b) = F(b-) - F(a-) = S(a-) - S(b-) = P(X < b) - P(X < a)$.

d) $P(a < X < b) = F(b-) - F(a) = S(a) - S(b-) = P(X < b) - P(X \le a)$.

So, for example, $P(a < X \le b) \approx F_n(b) - F_n(a) = S_n(a) - S_n(b)$.

139) Recall that $E(X) = \int_0^\infty S(x)dx$, and $E(X \wedge d) = \int_0^d S(x)dx$. Hence $E(X) \approx$ area under the step function $S_n(x)$ or $\hat{S}(x)$, while $E(X \wedge d) \approx$ area under the step function $S_n(x)$ or $\hat{S}(x)$ on the interval $[0, d]$.

140) **Know**: **Greenwood's approx.** for $V(S_n(x))$ where $S_n(x)$ is the KMPL estimator is $\hat{V}(S_n(y_j)) = \hat{V}(S_n(x)) = [S_n(y_j)]^2 \sum_{i=1}^{j} \dfrac{s_i}{r_i(r_i - s_i)}$ where $y_j \le x < y_{j+1}$.

(Also $\hat{V}(S_n(x)) = [S_n(x)]^2 \sum_{i: y_i \le x} \dfrac{s_i}{r_i(r_i - s_i)}$.) (Using 137) for **complete data** is easier.)

The KMPL estimator is **unbiased**: $E(S_n(x)) = S(x)$.

141) Let ${}_tp_x = P(X > x+t | X > x)$, and let ${}_tq_x = 1 - {}_tp_x = P(x < X \le x+t | X > x)$. Let $p_x = {}_1p_x$, and $q_x = {}_1q_x$. For a mortality study, ${}_tp_x =$ P(someone age x survives at least another t years), while ${}_tq_x =$ P(someone age x survives dies in the next t years).

142) Let $y > x$. Then ${}_{y-x}q_x = P(x < X \le y | X > x)$ and ${}_{y-x}p_x = P(X > y | X > x)$.

143) Let $y > x$. For complete data ${}_{y-x}\hat{q}_x = \dfrac{S_n(x) - S_n(y)}{S_n(x)}$, and ${}_{y-x}\hat{p}_x = \dfrac{S_n(y)}{S_n(x)}$. Let n be the number in the initial sample, let n_x be the number alive (with values $>$) x, and let n_y be the number alive at age y. Then $\hat{V}({}_{y-x}\hat{q}_x | n_x) = \hat{V}({}_{y-x}\hat{p}_x | n_x) = \dfrac{(n_x - n_y)n_y}{n_x^3}$. Note that n is the number initially at risk (at age 0) and the subscript in $S_n(x)$. Similarly, n_x is the number at risk at age x and n_y is the number at risk at age y.

144) **Know**: The **approx.** for $V(\hat{H}(x))$ where $\hat{H}(x)$ is the Nelson Aalen estimator is $\hat{V}(\hat{H}(y_j)) = \hat{V}(\hat{H}(x)) = \sum_{i=1}^{j} \dfrac{s_i}{r_i^2} = \hat{V}(\hat{H}(y_{j-1})) + \dfrac{s_j}{r_j^2}$ where $y_j \le x < y_{j+1}$.

(Also $\hat{V}(\hat{H}(x)) = \sum_{i:y_i \le x} \frac{s_i}{r_i^2}$.)

145) Let $y > x$. For modified data, it is still true that $_{y-x}\hat{q}_x = \frac{S_n(x) - S_n(y)}{S_n(x)}$, and $_{y-x}\hat{p}_x = \frac{S_n(y)}{S_n(x)}$. But if $y_{a-1} \le x < y_a$ and $y_{j-1} \le y < y_j$, then $_{y-x}\hat{p}_x = \frac{S_n(y)}{S_n(x)} = \prod_{i=a}^{j-1} \left(1 - \frac{s_i}{r_i}\right)$. Then $\hat{V}(\,_{y-x}\hat{q}_x) = \hat{V}(\,_{y-x}\hat{p}_x) = [\,_{y-x}\hat{p}_x]^2 \sum_{i=a}^{j-1} \frac{s_i}{r_i(r_i - s_i)}$.

From a table like 132), computations are like KMPL 132b) and 140), but start at y_a instead of y_1.

146) Let z_p be the $1-\alpha/2$ percentile $z_{1-\alpha/2}$ = the upper $\alpha/2$ percentile $z_{\alpha/2}$, using bad notation. So $P(Z \le z_p) = 1 - \alpha/2$ and $P(Z > z_p) = \alpha/2$.

CI	90%	95%	99%
z_p	1.645	1.96	2.576

147) Using Greenwood's approx. 140), a linear $100(1-\alpha)$% CI for $S(x)$ is $S_n(x) \pm z_p\sqrt{\hat{V}(S_n(x))}$.

148) **Know:** The **log transformed** $100(1-\alpha)$% **CI** for $S(x)$ is $([S_n(x)]^{1/U}, [S_n(x)]^U)$ where $U = \exp\left(\frac{z_p\sqrt{\hat{V}(S_n(x))}}{S_n(x)\ln(S_n(x))}\right)$.

149) Using the Nelson Aalen estimator and 144), a linear $100(1-\alpha)$% CI for $H(x)$ is $\hat{H}(x) \pm z_p\sqrt{\hat{V}(\hat{H}(x))}$.

150) **Know:** The **log transformed** $100(1-\alpha)$% **CI** for $H(x)$ is $\left(\frac{\hat{H}(x)}{U}, [\hat{H}(x)]U\right)$ where $U = \exp\left(\frac{z_p\sqrt{\hat{V}(\hat{H}(x))}}{\hat{H}(x)}\right)$.

151) Let $p(y_j) = s_j/n$ be the probability assigned to y_j by the empirical distribution where $s_j = n_j$ = (number of $x_i = y_j$) for $j = 1, ..., k$. A **kernel density estimator** of the pdf (kernel smoothing) is $\hat{f}(x) = \sum_{j=1}^k p(y_j)k_{y_j}(x) = \sum_{i=1}^n \frac{1}{n}k_{x_i}(x)$. The area under the pdf $k_{y_j}(x)$ is 1.

152) Let b be the bandwidth of the kernel. a) The **uniform kernel** $k_y(x) = \frac{1}{2b}$, $\; y - b \le x \le y + b$, and $k_y(x) = 0$, otherwise.

b) For the **triangular kernel** the height of the triangle is $1/b$, and the base goes from $y-b$ to $y+b$: $\quad k_y(x) = \frac{x - y + b}{b^2}$, $\; y - b \le x \le y$,
$k_y(x) = \frac{y + b - x}{b^2}$, $\; y \le x \le y + b$,
and $k_y(x) = 0$, otherwise.

153) For the uniform kernel,
$\hat{f}(x) = \frac{1}{2nb}\sum_{i=1}^n I(|x_i - x| \le b) = \frac{1}{2nb}\sum_{i=1}^n I(x_i \in [x-b, x+b]) =$
$\frac{1}{2nb}$ (# $x_i \in [x-b, x+b]$).