ESTIMATION FOR MODIFIED DATA

- **Definition 14.1** – An observation is **truncated from below** (also called left truncated) at d if when it is below d it is not recorded, but when it is above d it is recorded at its observed value.

  An observation is **truncated from above** (also called right truncated) at u if when it is above u it is not recorded, but when it is below u it is recorded at its observed value.

  An observation is **censored from below** (also called left censored) at d if when it is below d it is recorded as being equal to d, but when it is above d it is recorded at its observed value.

  An observation is **censored from above** (also called right censored) at u if when it is above u it is recorded as being equal to u, but when it is below u it is recorded at its observed value.

- **Comments**:

  - **Truncation** - In insurance, truncation from below can happen when there is a deductible: A policyholder will not report a claim whose value is below the deductible. However the knowledge of "small" claims (number and amounts) can be important for a correct evaluation of the policy risk.

- **Censoring** – Let y be the "correct" value, c the censoring point and x the available data.

  - Censoring from below
  $$x = \begin{cases} c & y \le c \\ y & y > c \end{cases}$$

  - Censoring from above
  $$x = \begin{cases} y & y < c \\ c & y \ge c \end{cases}$$

  - In insurance **censoring from above is quite usual**. If a policy pays no more than 10000 euro for a claim and if the insurance company only records the payments made, any time a loss is above 10000 euro the amount of the claim will be unknown but we will know that a payment of 10000 euro has happened.
  - The censoring points could be known or random. Random censoring occurs for instance when a policyholder decides to surrender his policy (data set D1). In any case we will know the censoring points that can differ from observation to observation.

  o From a statistical point of view, **truncation is a more severe limitation than censoring**.
  o When nothing else is said, truncation will mean left truncation and censoring right censoring.

- **The individual data**

  - For each observation 3 facts are needed:

    - Truncation point (if any)

    - The value of the observation

    - A flag to indicate if the observation was or was not censored

  - Notation (*Loss Models*)

    - d_j – truncation point. If there was no truncation $d_j = 0$ (assuming that we are dealing

      with a positive valued variable);

    - The observation

      - x_j if not censored

      - u_j if censored

    - This notation is not usual. However, we will follow *Loss Models* notation.

Main objectives of the chapter

- To estimate the survival function using complete data or using censored and truncated data

- To estimate the cumulative hazard function using censored and truncated data and to use the estimated cumulative hazard function to estimate the survival function

- To obtain confidence intervals (for a given x) for both the survival and the cumulative hazard functions (different methods)

- To generalize this approach to cover two situations

  o Kernel estimation – Method to estimate the **density** (distribution) function of a **continuous** random variable.

  o Approximation techniques – how to deal with aggregated data and large data sets, namely when we are only interested by the survival function values at some point (usually the end of each year or quarter)

The Kaplan-Meier estimator

- How to estimate the survival function using censored and truncated data?

- The first step is to summarize the information in a useful manner:

  - Let $y_1 < y_2 < \cdots < y_k$ be the k unique values that appear in the sample of **uncensored values**; Obviously $k \leq n_1$ where $n_1 \leq n$ is the number of uncensored values.

  - Let s_j be the number of times the uncensored observation y_j appears in the sample, $j = 1,2,\cdots,k$. Obviously $s_j = \#\{x_i = y_j\}$ and $\sum_{j=1}^{k} s_j = n_1$.

  - Let r_j be the **risk set** at y_j. At y_j we have under observation the "individuals" whose observation or censoring point are greater than or equal to y_j and whose truncation point is less than y_j (if the truncation point is greater than or equal to y_j, the "individual" is not yet under observation). Formally,

  $$r_j = \#\{d_i < y_j\} - \#\{x_i < y_j\} - \#\{u_i < y_j\} \quad \text{or} \quad r_j = \#\{x_i \geq y_j\} + \#\{u_i \geq y_j\} - \#\{d_i \geq y_j\}. \quad \square$$

- **Example 14.1** – Using Data Set D2, calculate the r_j values.

```
> #read data - usual notation
> d=c(rep(0,30),0.3,0.7,1.0,1.8,2.1,2.9,2.9,3.2,3.4,3.9)
> # w corresponds to the column "last observed". Merges x and u
> w=c(0.1,0.5,0.8,0.8,1.8,1.8,2.1,2.5,2.8,2.9,2.9,3.9,4.0,4.0,
+   4.1,4.8,4.8,4.8,rep(5.0,14),4.1,3.1,3.9,5.0,4.8,4.0,5.0,5.0)
>cs=c(rep(0,3),1,rep(0,5),rep(1,2),0,1,0,0,1,rep(0,16),1,1,
+   rep(0,3),1,0,0) #value 1 = "died"
> # Loss Models notation
> x=w[cs==1]
> u=w[cs==0]
> y=sort(unique(x))
> # or y=as.numeric(names(table(x))); s=as.numeric(table(x))
> r=rep(0,length(y)); s=r;
> for(i in 1:length(y)) {
+   s[i]=sum(x==y[i]); r[i]=sum(d<y[i])-sum(x<y[i])-sum(u<y[i]);
+   # r[i]=sum(d<y[i])-sum(w<y[i]); another option u is useless
+   }
> y; s; r
[1] 0.8 2.9 3.1 4.0 4.1 4.8
[1] 1 2 1 2 1 1
[1] 30 26 26 26 23 21
```

- Basic idea of Kaplan-Meier estimator: Start with $S(0) = 1$ (usual assumption about the r.v.) and, at each value y_j, estimate the conditional probability of survival (not having experienced the event), $\pi_j = \Pr(X > y_j \mid X \geq y_j)$. Once conditional probabilities have been estimated we use

$$S(y_j) = \Pr(X > y_j) = \frac{\Pr(X > y_j)}{\Pr(X \geq y_j)} \times \Pr(X \geq y_j) = \pi_j \times \Pr(X \geq y_j)$$

  and we will assume that the survival function is constant between y_{j-1} and y_j, $j = 1, 2, \cdots, k$.

- The estimates are

  o $\hat{\pi}_j = \dfrac{r_j - s_j}{r_j}$. $\qquad\qquad\qquad$ $(r_j - s_j)$ – number of ind. *surviving* at y_j and r_j – risk set

  o $\hat{\Pr}(X \geq y_j) = \hat{\Pr}(X > y_{j-1}) = \hat{S}(y_{j-1})$. $\qquad$ No events in the sample between y_{j-1} and y_j

  o $\hat{S}(y_j) = \hat{\pi}_j \times \hat{S}(y_{j-1}) = \hat{\pi}_j \times \hat{\pi}_{j-1} \times \hat{S}(y_{j-2}) = \hat{\pi}_j \times \hat{\pi}_{j-1} \times \cdots \times \hat{\pi}_1 \times \hat{S}(0) = \prod_{i=1}^{J} \hat{\pi}_i$

- Then

$$
S_n(t) = \begin{cases} 1 & t < y_1 \\ \displaystyle\prod_{i=1}^{j-1}\left(\frac{r_i - s_i}{r_i}\right) & y_{j-1} \le t < y_j \quad j = 2,3,\cdots,k \\ \displaystyle\prod_{i=1}^{k}\left(\frac{r_i - s_i}{r_i}\right) \ \text{or}\ 0\ \text{or} \ .... & y_k \le t \end{cases} = \begin{cases} 1 & t < y_1 \\ \displaystyle\prod_{i:y_i \le t}\left(\frac{r_i - s_i}{r_i}\right) & y_1 \le t \le y_k \\ \displaystyle\prod_{i=1}^{k}\left(\frac{r_i - s_i}{r_i}\right) \ \text{or}\ 0\ \text{or} \ .... & t > y_k \end{cases}
$$

- Let us discuss the estimation of the survival function when $t \ge y_k$.

  - $S_n(y_k) = \displaystyle\prod_{i=1}^{k}\left(\frac{r_i - s_i}{r_i}\right)$

  - If $s_k = r_k$, then $S(t) = 0$ for $t \ge y_k$ can make sense (after y_k the risk set is empty. However some previously censored observation can still be *alive* or *survivors* can exist among the population after y_k).

- o If $s_k < r_k$ we know that at least $(r_k - s_k)$ individuals survive at time $t = y_k$ but there is no empirical data to complete the survival function. 3 options are available:

  - Keep the survival function at its last value, $S_n(t) = \prod_{i=1}^{k}\left(\dfrac{r_i - s_i}{r_i}\right)$, $t \geq y_k$.

  - Keep the survival function at its last value until the last censoring time is reached and then declare the function to be 0,

  $$S_n(t) = \begin{cases} \prod_{i=1}^{k}\left(\dfrac{r_i - s_i}{r_i}\right) & y_k \leq t < w \\ 0 & w \leq t \end{cases} \qquad \text{where } w = \text{last censoring time}$$

  We are imposing that no one survives after time w and a step transition to 0.

  - Use an exponential curve to reduce the value of the survival function from its current value to zero. For instance, let $s* = \prod_{i=1}^{k}\left(\dfrac{r_i - s_i}{r_i}\right)$ and w = max (last censoring time, y_k) and use

  $S_n(t) = \exp\left(\dfrac{t}{w}\ln s*\right) = (s*)^{t/w}$, $t \geq w$. We can combine this approach with the previous one i.e. we choose an upper value and declare the function to be 0 when t is greater than this value.

- **Example 14.2** – Determine the Kaplan-Meier estimate for Data Set D2.

  Following example 14.1

  ```
  > pihat=(r-s)/r
  > Sn=cumprod(pihat)
  > Sn
  [1] 0.9666667 0.8923077 0.8579882 0.7919891 0.7575548 0.7214807
  ```

  Alternatively
  # read data - Data Set D2 - usual notation – see example 14.1 – d   w   cs
  ```
  > library(survival)
  Loading required package: splines
  > fit <- survfit(Surv(d,w,cs)~1) # kaplan-Meier and Greenwood by
                                    default
  > summary(fit)
  ```

```
Call: survfit(formula = Surv(d, w, 1 - cs) ~ 1)
```

time	n.risk	n.event	entered	censored	survival	std.err	lower 95% CI
0.8	30	1	0	1	0.967	0.0328	0.905
2.9	26	2	2	0	0.892	0.0589	0.784
3.1	26	1	0	0	0.858	0.0659	0.738
4.0	26	2	0	1	0.792	0.0755	0.657
4.1	23	1	0	1	0.758	0.0797	0.616
4.8	21	1	0	3	0.721	0.0837	0.575

upper 95% CI
1.000
1.000
0.997
0.955
0.931
0.906

```
> plot(fit)
```

Note: As we will see, the confidence intervals (CI) are calculated using a different method.
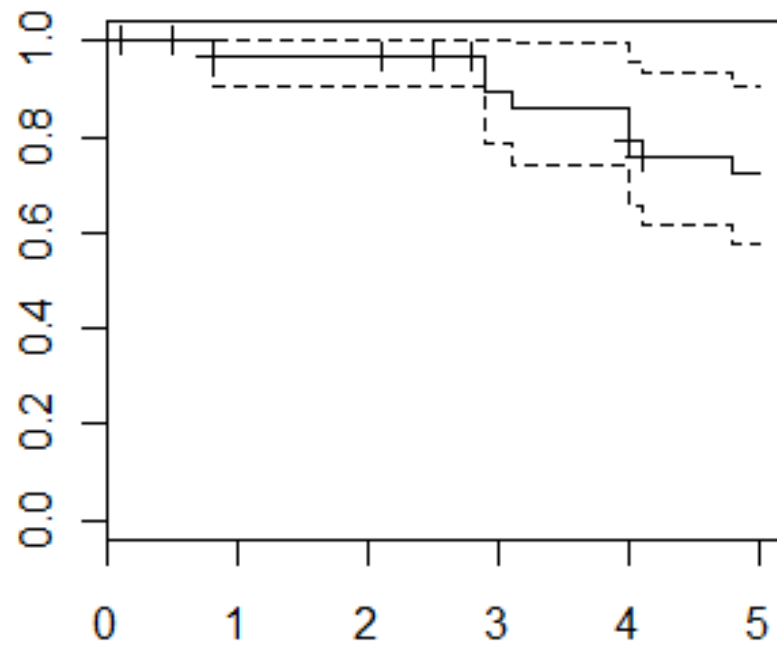
The Nelson-Aalen estimator

- Aim: To estimate the cumulative hazard rate. Remember from previous chapter that, if $H(t)$ is differentiable, $H(t) = \int_{-\infty}^{t} h(u)\,du$ where $h(u) = f(u)/S(u)$

- Then our estimate (estimator) is $\hat{H}(t) = \sum_{i:y_i \leq t} \dfrac{s_i}{r_i}$

- The **Nelson-Aalen estimator** is

$$\hat{H}(t) = \begin{cases} 0 & 0 \leq t < y_1 \\ \sum_{i=1}^{j-1} \dfrac{s_i}{r_i} & y_{j-1} \leq t < y_j \quad j = 2,3,\cdots,k \\ \sum_{i=1}^{k} \dfrac{s_i}{r_i} & y_k \leq t \end{cases} = \begin{cases} 0 & t < y_1 \\ \sum_{i:y_i \leq t} \dfrac{s_i}{r_i} & y_1 \leq t \leq y_k \\ \sum_{i=1}^{k} \dfrac{s_i}{r_i} & t > y_k \end{cases}$$

- As we have already seen, we can also use the Nelson-Aalen estimator to get another estimator of the survival function $\hat{S}(t) = e^{-\hat{H}(t)}$ and for $t > y_k$ we can return to the previous discussion.

- **Example 14.3** – Determine the Nelson-Aalen estimate of the survival function for data set D2.
  See examples 14.1 and 14.2

```
> Hn=cumsum(s/r)
> Hn
[1] 0.03333333 0.11025641 0.14871795 0.22564103 0.26911929
0.31673833
> Sn_H=exp(-Hn)
> Sn_H
[1] 0.9672161 0.8956045 0.8618122 0.7980045 0.7640521 0.7285214
```

Means, variances and interval estimation

- The first part of this section refers to complete data and has already been presented. Now we will consider censored and/or truncated data.

- Our main concern is to approximate the variance of the Kaplan-Meier estimator (or of the Nelson-Aalen estimator) of $S(t)$ to calculate confidence bands – a confidence interval for each value of *t*.

- Approximate confidence interval for $S(t)$: $S_n(t) \pm z_{\alpha/2} \times \sqrt{\hat{\text{var}}\, S_n(t)}$ keeping in mind that the lower limit must be greater than or equal to 0 and the upper limit lesser than or equal to 1. Question: How to estimate $\text{var}\, S_n(t)$?

- Kaplan-Meier estimator of $S(t)$: $S_n(t) = \prod_{i=1}^{j-1}\left(\dfrac{r_i - s_i}{r_i}\right)$, $y_{j-1} \leq t < y_j$, $j = 2,3,\cdots,k$.

$$S_n(y_j) = \prod_{i=1}^{j}\left(\frac{r_i - s_i}{r_i}\right)$$

- The usual estimate for the variance is given by **Greenwood's** formula: $\hat{\text{var}}\left(S_n(y_j)\right) \approx S_n(y_j)^2 \times \sum_{i=1}^{j}\dfrac{s_i}{r_i\left(r_i - s_i\right)}$

  See *Loss Models* for a deduction of Greenwood's approximation or use the delta method. □

- Greenwood's formula can be written as $\hat{\text{var}}\left(S_n(t)\right) \approx S_n(t)^2 \times \sum_{i:y_i \leq t}\dfrac{s_i}{r_i\left(r_i - s_i\right)}$

- **Example 14.12** – Using Data Set D1, estimate the variance of $S_{30}(3)$ both directly and using Greenwood's formula. Do the same for $_2\hat{q}_3$.

Solution:

$$S_n(3) = \frac{27}{30} = 0.9; \qquad\qquad _2\hat{q}_3 = \frac{S_n(3) - S_n(5)}{S_n(3)} = \frac{(27/30) - (23/30)}{(27/30)} = \frac{4}{27} \approx 0.1481$$

Directly – There is no censoring or truncation

$$\hat{var}(S_n(3)) = \frac{S_n(3) \times (1 - S_n(3))}{n} = \frac{(27/30) \times (3/30)}{30} = \frac{81}{30^3} = 0.003$$

$$\hat{var}(_2\hat{q}_3 \mid S(3) = S_n(3)) = \frac{1}{n_3^3} n_5 (n_3 - n_5) = \frac{1}{27^3} \times 23 \times (27 - 23) = \frac{92}{27^3} \approx 0.0047$$

Greenwood's formula

$r_1 = 30;\ s_1 = 1;\ r_2 = 29;\ s_2 = 2$

$$\hat{var}(S_n(3)) \approx \left(\frac{27}{30}\right)^2 \times \left(\frac{1}{30 \times 29} + \frac{2}{29 \times 27}\right) = \left(\frac{27}{30}\right)^2 \times \frac{27 + 2 \times 30}{30 \times 29 \times 27} = \frac{81}{30^3} = 0.003$$

To apply Greenwood's formula to estimate $\hat{var}(_2\hat{q}_x \mid S(3) = S_n(3))$ we must consider a sub-sample of size n_3 (only the 27 observations greater than 3 are relevant) and estimate $1 - S(5)$ using this

subsample. With this sub-sample (index 1 refers to the value 3.1, index 2 to the value 4.0 and index 3 to the value 4.8) $r_1 = 27$; $s_1 = 1$; $r_2 = 26$; $s_2 = 1$; $r_3 = 25$; $s_3 = 2$;

$$\text{vâr}\left(_2\hat{q}_3 \mid S(3) = S_n(3)\right) \approx \left(\frac{23}{27}\right)^2 \times \left(\frac{1}{27 \times 26} + \frac{1}{26 \times 25} + \frac{2}{25 \times 23}\right) = \frac{92}{27^3}$$

- **Example 14.13** – Repeat example 14.12, this time using all 40 observations in Data Set D2 and the incomplete information due to censoring and truncation.

  From examples 14.1 and 14.2 we have

j	1	2	3	4	5	6
y_j	0.8	2.9	3.1	4.0	4.1	4.8
r_j	30	26	26	26	23	21
s_j	1	2	1	2	1	1
S_n	0.9667	0.8923	0.8580	0.7920	0.7576	0.7215

$$S_n(3) = \prod_{j=1}^{2}\left(\frac{r_j - s_j}{r_j}\right) = \prod_{y_j \leq 3}\left(\frac{r_j - s_j}{r_j}\right) = \frac{29}{30} \times \frac{24}{26} = 0.8923$$

$$_2\hat{q}_3 = 1 - \frac{S_n(5)}{S_n(3)} = 1 - \frac{0.7215}{0.8923} \approx 0.1914$$

Directly – not possible due to censoring and truncation

Greenwood's formula

$r_1 = 30$; $s_1 = 1$; $r_2 = 26$; $s_2 = 2$

$$\text{vâr}\left(S_n(3)\right) \approx 0.8923077^2 \times \left(\frac{1}{30 \times 29} + \frac{2}{26 \times 24}\right) = 0.003467152$$

To apply the Greenwood's formula to estimate $\text{var}\left(_2\hat{q}_x \mid S(3) = S_n(3)\right)$ we will keep the subscripts

$r_3 = 26$; $s_3 = 1$; $r_4 = 26$; $s_4 = 2$; $r_5 = 23$; $s_5 = 1$; $r_6 = 21$; $s_6 = 1$;

$$\text{vâr}\left(_2\hat{q}_3 \mid S(3) = S_n(3)\right) = \text{vâr}\left(1 - \frac{S_n(5)}{S_n(3)} \mid S(3) = S_n(3)\right) = \text{vâr}\left(\frac{S_n(5)}{S_n(3)} \mid S(3) = S_n(3)\right)$$

$$\approx \left(\frac{S_n(5)}{S_n(3)}\right)^2 \times \sum_{i:3 < y_i \le 5} \frac{s_i}{r_i(r_i - s_i)}$$

$$= \left(\frac{0.7215}{0.8923}\right)^2 \times \left(\frac{1}{26 \times 25} + \frac{2}{26 \times 24} + \frac{1}{23 \times 22} + \frac{1}{21 \times 20}\right)$$

$$= 0.005950$$

A new approach to obtain confidence intervals for $S(t)$

- To guarantee that the limits of the confidence interval for $S_n(t)$ are bounded by 0 and 1, we can use an alternative methodology.

- The idea is to construct confidence intervals for a function $g(S(t))$ (monotonic and unbounded) and then to use the inverse of this function to get a confidence interval for $S(t)$.

- To be more explicit let us consider $g(S(t)) = \ln(-\ln S(t))$ and, for each given value of t, we will use $g(S_n(t)) = \ln(-\ln S_n(t))$ to estimate $g(S(t))$. Note that $-\infty < g(S(t)) < +\infty$.

  □ Question: Why is $g(S(t)) = \ln(-\ln S(t))$ a suitable function?

- To obtain a confidence interval for $g(S(t))$, we use the delta method, i.e.

$$E\big(g(S_n(t))\big) \approx g(S(t)) + E\big(S_n(t) - S(t)\big)\, g'(S(t)) = g(S(t))$$

For each t, $S_n(t)$ is an unbiased estimator of $S(t)$

$$\text{var}(g(S_n(t))) \approx \big(g'(S(t))\big)^2 \text{var}\big(S_n(t) - S(t)\big) = \big(g'(S(t))\big)^2 \text{var}\big(S_n(t)\big)$$

$$se(g(S_n(t))) = \sqrt{\hat{\text{var}}(g(S_n(t)))} \approx \big|g'(S_n(t))\big| \sqrt{\hat{\text{var}}\big(S_n(t)\big)}$$

19

Note: $g(x) = \ln(-\ln x)$ then $g'(x) = \dfrac{1}{x \ln x}$ and, for $0 < S_n(t) < 1$, $g'(S_n(t)) = \dfrac{1}{S_n(t) \ln S_n(t)}$, i.e.

$$\left| g'(S_n(t)) \right| = -\dfrac{1}{S_n(t) \ln S_n(t)}.$$

- The bounds of the approximate confidence interval for $g(S(t))$ (level $1 - \alpha$) are given by

$$g(S_n(t)) \pm z_{\alpha/2}\, se\big(g(S_n(t))\big) \quad \text{i.e.} \quad \ln(-\ln S_n(t)) \pm z_{\alpha/2}\, \dfrac{\sqrt{v\hat{a}r(S_n(t))}}{S_n(t) \times \ln S_n(t)}$$

- Now, using this result, we will construct a confidence interval for $S(t)$.

    - $g(S(t)) = \ln(-\ln S(t)) \Leftrightarrow S(t) = \exp(-\exp(g(S(t))))$

    - Using this result we get $UB = \big(S_n(t)\big)^{U}$ and $LB = \big(S_n(t)\big)^{1/U}$ with $U = \exp\left(z_{\alpha/2}\, \dfrac{\sqrt{v\hat{a}r(S_n(t))}}{S_n(t) \times \ln S_n(t)} \right)$

    - Then the confidence interval is $\left(\big(S_n(t)\big)^{1/U} ; \big(S_n(t)\big)^{U} \right)$. This interval will always be inside the

        range 0 to 1 and is referred to as the **log-transformed interval**.

- **Example 14.14** – Obtain the log-transformed confidence interval for $S(3)$ as in example 14.13.

Direct Method

$\left(0.8923077 - 1.96 \times \sqrt{0.003467152}\,;\, 0.8923077 + 1.96 \times \sqrt{0.003467152}\right)$, i.e. (0.7769; 1.0077) and then (0.7769; 1.00)

Log-transformed interval

$$U = \exp\left(1.96 \times \frac{\sqrt{0.003467152}}{0.8923077 \times \ln 0.8923077}\right) = 0.321389$$

$\left(0.8923077^{1/0.321389}\,;\, 0.8923077^{0.321389}\right)$, i.e. (0.7015; 0.9640)

Confidence intervals for the cumulative hazard function

- Similar results are available using the Nelson-Aalen estimator. The procedure and the set of hypotheses are similar with the ones used in the previous situation, namely that the risk sets are known (not random) and that there is conditional independence among the S_i.

- We will assume that the number of "deaths" at y_i, S_i, follows approximately a Poisson distribution with parameter $r_i h(y_i)$. Similar (but not equal) results can be obtained using a binomial distribution.

- Assuming a Poisson distribution, the variance of S_i, given r_i, is given by $r_i h(y_i)$ and, since $h(y_i)$ is estimated using s_i / r_i, we could use the approximation $\hat{var}(S_i) = r_i \times s_i / r_i = s_i$.

- Then, assuming independence among the S_i / r_i, we get

$$\hat{var}\left(\hat{H}(y_j)\right) = \hat{var}\left(\sum_{i=1}^{j} \frac{s_i}{r_i}\right) = \sum_{i=1}^{j} \frac{\hat{var}(s_i)}{r_i^2} = \sum_{i=1}^{j} \frac{s_i}{r_i^2} \qquad \text{or} \qquad \hat{var}\left(\hat{H}(t)\right) = \sum_{i:y_i \le t} \frac{s_i}{r_i^2}$$

- A direct (linear) confidence interval is then given by $\hat{H}(t) \pm z_{\alpha/2} \sqrt{\hat{var}\left(\hat{H}(t)\right)}$

- A log-transformed confidence interval is given by $\left(\hat{H}(t)U_L;\hat{H}(t)U_U\right)$ where

$$U_L = \exp\left(-\frac{z_{\alpha/2}\sqrt{\mathrm{v\hat{a}r}\left(\hat{H}(t)\right)}}{\hat{H}(t)}\right) \text{ and } U_U = \exp\left(+\frac{z_{\alpha/2}\sqrt{\mathrm{v\hat{a}r}\left(\hat{H}(t)\right)}}{\hat{H}(t)}\right)$$

- **Example 14.15** – Construct an approximate 95% confidence interval for $H(3)$ by each formula using all 40 observations in Data set D2.

  Let us take advantage of Example 14.13.

  $$\hat{H}(3) = \sum_{i:y_i\leq 3}\frac{s_i}{r_i} = \sum_{i=1}^{2}\frac{s_i}{r_i} = \frac{1}{30} + \frac{2}{26} = 0.11026 \text{ and } \mathrm{v\hat{a}r}\left(\hat{H}(3)\right) = \sum_{i=1}^{2}\frac{s_i}{r_i^2} = 0.0040697$$

  Direct confidence interval: (-0.01478; 0.2352393) → (0; 0.2352393)

  $$U_L = \exp\left(-1.96\frac{\sqrt{0.0040697}}{0.11026}\right) = 0.321727 \qquad\qquad U_U = \exp\left(1.96\frac{\sqrt{0.0040697}}{0.11026}\right) = 3.108225$$

  Log-transformed confidence interval: (0.035472; 0.342702)
  ☐Question: How to use these results in order to get CI for $S(3)$?

KERNEL DENSITY MODELS

- Although the empirical distribution converges to the distribution of the random variable, as $n \to \infty$, a main point remains: for finite samples the empirical distribution is always discrete, even if the underlying variable is **continuous**. This problem is more annoying when the sample size is moderate.
- Our aim is to smooth, using non parametric methods (i.e. ignoring the functional form of the density), the empirical distribution to obtain an estimate of the continuous density (or distribution) function.

- **Definition 14.2** – A kernel density estimator of a distribution function is

$$\hat{F}(x) = \sum_{j=1}^{k} p(y_j) K_{y_j}(x)$$

And the estimator of the density function is

$$\hat{f}(x) = \sum_{j=1}^{k} p(y_j) k_{y_j}(x).$$

The function $k_y(x)$ is called the **kernel**.

- **Comments**

  o The kernel is a non-negative real-valued integrable function satisfying $\int_{-\infty}^{+\infty} k_y(x)\,dx = 1$ to

  guarantee that the kernel method originates a density function. We will also have,

  $$K_y(x) = \int_{-\infty}^{x} k_y(u)\,du \,.$$

  □Question: How can we guarantee that $\hat{f}(x)$ is a density function?

  o In much cases we impose that $\int_{-\infty}^{+\infty} x\,k_y(x)\,dx = y$, that is the expected value is unchanged by the

  kernel.

  o $p(y_j)$ is the probability assigned to the value y_j, $j = 1,2,\cdots,k$, by the empirical distribution. : If

  all the sample values are unique we get $p(y_j) = 1/n$ and then $\hat{F}(x) = \sum_{i=1}^{n}(1/n)\,K_{x_i}(x)$ and

  $\hat{f}(x) = \sum_{i=1}^{n}(1/n)\,k_{x_i}(x)$ respectively.

- **Definition 14.3** (using a different notation)
  - Uniform kernel:

$$k_y(x) = (2b)^{-1} I(|x-y| \le b) = (2b)^{-1} I(y-b \le x \le y+b) = \begin{cases} 0 & x < y-b \\ 1/(2b) & y-b \le x \le y+b \\ 0 & x > y+b \end{cases}$$

  - Triangular kernel: $k_y(x) = \dfrac{b-|y-x|}{b^2} I(|y-x|/b \le 1) = \begin{cases} 0 & x < y-b \\ (x-y+b)/b^2 & y-b \le x \le y \\ (y+b-x)/b^2 & y \le x \le y+b \\ 0 & x > y+b \end{cases}$

  - Gamma kernel: $k_y(x) = \dfrac{x^{\alpha-1} e^{-x\alpha/y}}{(y/\alpha)^\alpha \, \Gamma(\alpha)} I_{(0;+\infty)}(x)$

    Gamma density with mean y and variance y^2/α. The lesser α the smoother the kernel.

    How to choose α? One can use $\alpha = \sqrt{n} \sqrt{(\hat{\mu}'_4 / \hat{\mu}'^2_2) - 1}$ (Typo in the book)

    Remember that $\hat{\mu}'_k = \sum y_j^k \, p(y_j)$

- Comments:
  - b is called the bandwidth . The higher is b the smoother will be the kernel density.
  - The first and second kernels are symmetric around y. In symmetric kernels the bandwidth is usually much more important than the choice of a particular kernel.
  - The third kernel is asymmetric and α plays a role similar to the bandwidth. Note that the gamma kernel can be used only with positive valued random variables.

- How to get $K_y(x)$?

  - $$K_y(x) = \int_{-\infty}^{x} k_y(u)\,du$$

  - For example in the uniform case,

    $$K_y(x) = \begin{cases} 0 & x < y-b \\ \int_{y-b}^{x} \dfrac{1}{2b}\,du & y-b \le x \le y+b \\ 1 & x > y+b \end{cases} = \begin{cases} 0 & x < y-b \\ \dfrac{x-y+b}{2b} & y-b \le x \le y+b \\ 1 & x > y+b \end{cases}$$

- In the remaining of the course we will follow Definition 14.2. However this is not the standard definition of a kernel density estimator. For a standard presentation, see Wasserman (2004).

  A kernel is any smooth function K such that $K(x) \geq 0$, $\int_{-\infty}^{+\infty} K(x)\,dx = 1$, $\int_{-\infty}^{+\infty} x\,K(x)\,dx = 0$ and

  $$\sigma_K^2 = \int_{-\infty}^{+\infty} x^2\ K(x)\,dx < \infty.$$

  Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is

  defined to be $\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$.

  Examples of kernels are:

    - The Gaussian kernel: $K(u) = (2\pi)^{-1/2}\, e^{-u^2/2}$

    - The Epanechnikov kernel: $K(u) = \frac{3}{4 \times \sqrt{5}}\left(1 - \frac{u^2}{5}\right) I\left(|u| < \sqrt{5}\right)$

    - The uniform kernel: $K(u) = \frac{1}{2} I\left(|u| \leq 1\right)$

    - The triangular kernel: $K(u) = \left(1 - |u|\right) I\left(|u| \leq 1\right)$

All these kernels act symmetrically around each sample point. In this setup the choice of a particular kernel is generally much less important than the choice of the bandwidth. They are methods to approximate the "best" choice of the bandwidth (see Wasserman (2004)).

- Let us return to definition 14.12 that will always be used.

- **Example 14.16** – Determine the kernel density estimate for Example 13.2 using each of the three kernels.

  We will use only the uniform kernel with $b=0.1$ and $b=1.0$ (try $b=0.5$ and get the results for the other situations)

  Sample $(1.0; 1.3; 1.5; 1.5; 2.1; 2.1; 2.1; 2.8)$

y_j	1.0	1.3	1.5	2.1	2.8
$p(y_j)$	1/8	1/8	2/8	3/8	1/8

Bandwith b=0.1 then $1/(2b)=5$

1.0		0.9	1.1
1.3		1.2	1.4
1.5	$\rightarrow$	1.4	1.6
2.1		2.0	2.2
2.8		2.7	2.9

$$\hat{f}(x) = \begin{cases} 5/8 & 0.9 < x < 1.1 \\ 5/8 & 1.2 < x < 1.4 \\ 10/8 & 1.4 < x < 1.6 \\ 15/8 & 2.0 < x < 2.2 \\ 5/8 & 2.7 < x < 2.9 \\ 0 & \text{otherwise} \end{cases}$$

Discuss the problem related to the intervals limit

Bandwith b=1.0 then $1/(2b) = 0.5$

1.0		0.0	2.0
1.3		0.3	2.3
1.5	$\rightarrow$	0.5	2.5
2.1		1.1	3.1
2.8		1.8	3.8

$$\hat{f}(x) = \begin{cases} 1/16 & 0 < x < 0.3 \\ 2/16 & 0.3 < x < 0.5 \\ 4/16 & 0.5 < x < 1.1 \\ 7/16 & 1.1 < x < 1.8 \\ 8/16 & 1.8 < x < 2.0 \\ 7/16 & 2.0 < x < 2.3 \\ 6/16 & 2.3 < x < 2.5 \\ 4/16 & 2.5 < x < 3.1 \\ 1/16 & 3.1 < x < 3.8 \\ 0 & \text{otherwise} \end{cases}$$

APPROXIMATIONS FOR LARGE DATA SETS

- When there are large amounts of data and we only are interested by some values of the survival function, constructing the Kaplan-Meier estimate may require more sorting and counting than can be justified by the results (in computer's age this sentence is questionable). However approximation techniques can be useful to deal with aggregated data.

- Let us assume that there is a set of **given** intervals with $c_0 < c_1 < \cdots < c_k$ and let:
  - d_j number of observations that are left truncated somewhere in the interval $[c_j ; c_{j+1})$. Note that d_0 includes the number of non-truncated observations;
  - u_j number of right censored observations somewhere in $(c_j ; c_{j+1}]$;
  - x_j number of uncensored observations that experiment the event ("died") in $(c_j ; c_{j+1}]$.

- Comments:
  - The intervals endpoints differ because truncation is possible at the left end of the first interval but not at the right end of the last interval (the reverse is true for censoring).
  - Each observation originates one d_j and one u_j or (exclusive) x_j.

- Now some assumptions must be made about the location of the values in each interval; among many possibilities we will present 2 methods.

- **First method** (the most used)

  - Hypothesis:

    - All truncations arrive at the beginning of the intervals

    - All censoring at the end of the intervals

    - It does not matter when the uncensored observations fall. Let us just assume that none of them are at the endpoints of the intervals.

  - The intuition

    - $\hat{S}(c_0) = 1;\ r_0 = d_0$

    - $\hat{S}(c_1) = 1 - x_0 / r_0;\ r_1 = d_0 - u_0 - x_0 + d_1 = (d_0 + d_1) - (u_0 + x_0)$

    - $\hat{S}(c_2) = \hat{S}(c_1) \times (1 - x_1 / r_1);\ r_2 = \cdots$

o The method

- $r_0 = d_0$; $\hat{S}(c_0) = 1$

- For $j = 1, 2, \cdots, k-1$

$$r_j = \sum_{i=0}^{j} d_i - \sum_{i=0}^{j-1}(u_i + x_i) = r_{j-1} + d_j - (x_{j-1} + u_{j-1});$$

$$\hat{S}(c_j) = \prod_{i=0}^{j-1}(1 - x_i / r_i) = \hat{S}(c_{j-1}) \times (1 - x_{j-1} / r_{j-1});$$

o Conditional probabilities

$$q_{c_j} = \Pr(X \le c_{j+1} \mid X > c_j) = \frac{S(c_j) - S(c_{j+1})}{S(c_j)} \qquad \text{is estimated by}$$

$$\hat{q}_j = 1 - \frac{\hat{S}(c_{j+1})}{\hat{S}(c_j)} = 1 - \frac{\prod_{i=0}^{j}(1 - x_i / r_i)}{\prod_{i=0}^{j-1}(1 - x_i / r_i)} = 1 - \left(1 - \frac{x_j}{r_j}\right) = \frac{x_j}{r_j}$$

- **Second method**

  o Hypothesis:

    ▪ The truncation and the censoring points occur uniformly through each interval. At the midpoint of the interval we will assume that half of the truncations and half of the censoring have occurred;

    ▪ Uncensored observations occur at the midpoints of the intervals.

  o The method

    ▪ $r_0 = (d_0 - u_0)/2;\ \hat{S}(c_0) = 1$

    ▪ For $j = 1, 2, \cdots, k-1$

$$r_j = \frac{(d_j - u_j)}{2} + \sum_{i=0}^{j-1}(d_i - u_i - x_i) = r_{j-1} + \frac{(d_j - u_j)}{2}d_j + \frac{d_{j-1}}{2} - \frac{u_{j-1}}{2} - x_{j-1}$$

$$\hat{S}(c_j) = \prod_{i=0}^{j-1}(1 - x_i/r_i) = \hat{S}(c_{j-1}) \times (1 - x_{j-1}/r_{j-1});$$

○ Conditional probabilities

The same formula applies.

$$\hat{q}_j = 1 - \frac{\hat{S}(c_{j+1})}{\hat{S}(c_j)} = 1 - \frac{\prod_{i=0}^{j}(1 - x_i/r_i)}{\prod_{i=0}^{j-1}(1 - x_i/r_i)} = 1 - \left(1 - \frac{x_j}{r_j}\right) = \frac{x_j}{r_j}$$

- Note that we can combine hypothesis from both methods when appropriate.

Final comments

- Note that the goal of all the estimation procedures is to deduce the probability distribution for the variable of interest in the absence of censoring and truncation.

- In the language of *Actuarial Mathematics* (Bowers et al.) these are called single decrement probabilities and are denoted q'_j

- In life insurance setup the censoring rates are sometimes as important as the mortality rates. A superscript identifies the decrement of interest: q'^d_j for death and q'^w_j for withdrawal.

- **Example 14.17** – Estimate single decrement probabilities using data set D2 and the methods of this section. Make reasonable assumptions.

Death

$j = 0,1,2,3,4$ and $c_j = j$. The columns d_j, u_j and x_j are obvious. The point is to determine the risk sets.

The truncation points follow two patterns: 30 values are known to be exactly 0 (no truncation) and the other 10 fall "randomly" in each interval. We will then consider the non-truncated observations and use the second method for the truncated ones.

A similar situation occurs with the censoring points: we know that 17 observations are truncated at $t = 5$ and the other censored observations fall "randomly".

$$r_0 = 30 + (d_0 - 30)/2 - u_0/2 = 29.5$$

$$r_j = \sum_{i=0}^{j-1}(d_i - u_i - x_i) + d_j/2 - u_j/2 \qquad j = 1,2,3$$

$$r_4 = \sum_{i=0}^{3}(d_i - u_i - x_i) + d_4/2 - (21-17)/2$$

$$q'_j = x_j / r_j$$

j	c_j	c_{j+1}	d_j	u_j	x_j	r_j	q'^d_j
0	0	1	32	3	1	29.5	0.0339
1	1	2	2	2	0	28.0	0.0000
2	2	3	3	3	2	28.0	0.0714
3	3	4	3	3	3	26.0	0.1154
4	4	5	0	21	2	21.0	0.0952

Withdrawal

The intervals are the same as in the previous case. The same happens to column d_j. Now column u_j corresponds to the previous column x_j (and x_j to former column u_j) except for the last interval.

In fact we must split the previously censored observations in 2 cases: when censoring is due to the policy withdrawal which is now the event of interest or when censoring is due to the ending of the

observation period (the observation remains censored). The observations for which we observe a death are now censored observation (the policyholder dies before withdrawal). The problem only arrives for $j = 4$. As we can see, the 19 censored observations correspond to 2 death (assumed falling "randomly" in the interval) and 17 observation whose time of observation expire at $t = 5$ (belonging to the risk set). Then $r_4 = \sum_{i=0}^{3}(d_i - u_i - x_i) + d_4/2 - 2/2$

j	c_j	c_{j+1}	d_j	u_j	x_j	r_j	$q_j'^w$
0	0	1	32	1	3	29.5	0.0339
1	1	2	2	0	2	28.0	0.0000
2	2	3	3	2	3	28.0	0.0714
3	3	4	3	3	3	26.0	0.1154
4	4	5	0	19	4	21.0	0.0952

Appendix 0 – Answer to challenging questions

☐ Why $r_j = \#\{d_i < y_j\} - \#\{x_i < y_j\} - \#\{u_i < y_j\}$ is equivalent to $r_j = \#\{x_i \geq y_j\} + \#\{u_i \geq y_j\} - \#\{d_i \geq y_j\}$?

Let us split our sample in 2 parts: n_x observations of non-truncated values and n_u observations of truncated values. As it is obvious:

$$n = n_x + n_u; \quad \#\{x_i \geq y_j\} = n_x - \#\{x_i < y_j\}; \quad \#\{u_i \geq y_j\} = n_u - \#\{u_i < y_j\}; \quad \#\{d_i \geq y_j\} = n - \#\{d_i < y_j\}$$

Then

$$r_j = \#\{d_i < y_j\} - \#\{x_i < y_j\} - \#\{u_i < y_j\} = n - \#\{d_i \geq y_j\} - n_x + \#\{x_i \geq y_j\} - n_u + \#\{u_i \geq y_j\}$$

$$= -\#\{d_i \geq y_j\} + \#\{x_i \geq y_j\} + \#\{u_i \geq y_j\}$$

☐ Challenging question: Why is $g(S(t)) = \ln(-\ln S(t))$ a suitable function?

The idea is to get a function whose values are not bounded. As it is straightforward to verify $g(S(t)) = \ln(-\ln S(t))$ fulfills this requirement.

$$0 < S_n(t) < 1 \Leftrightarrow -\infty < \ln S_n(t) < 0 \Leftrightarrow 0 < -\ln S_n(t) < \infty \Leftrightarrow -\infty < \ln(-\ln S_n(t)) < \infty$$

☐Question: How to use these results in order to get a CI for $S_n(3)$?

Remember that $S(3)$ can be estimated by $\hat{S}(3) = \exp(-\hat{H}(3))$ and apply this transformation to the bounds of the confidence intervals.

Direct method: (0.7904; 1)

Log-transformed: (0.7099; 0.9652)

☐Question: How can we guarantee that $\hat{f}(x)$ is a density function?

Remember that $\hat{f}(x) = \sum_{j=1}^{k} p(y_j) k_{y_j}(x)$

- $\hat{f}(x)$ has to be non-negative which is obvious since $p(y_j) \geq 0$ and $k_{y_j}(x) \geq 0$

- We must have $\int_{-\infty}^{+\infty} \hat{f}(x)\,dx = 1$

41

$$\int_{-\infty}^{+\infty} \hat{f}(x)\,dx = \int_{-\infty}^{+\infty} \left(\sum_{j=1}^{k} p(y_j) k_{y_j}(x) \right) dx \qquad \text{definition}$$

$$= \sum_{j=1}^{k} \int_{-\infty}^{+\infty} p(y_j) k_{y_j}(x)\,dx \qquad \text{swap integration with derivation}$$

$$= \sum_{j=1}^{k} p(y_j) \int_{-\infty}^{+\infty} k_{y_j}(x)\,dx$$

$$= \sum_{j=1}^{k} p(y_j) \qquad\qquad\qquad k_{y_j}(x) \text{ is a density}$$

$$= 1$$

.

The function $k_y(x)$ is called the **kernel**.

- $p(y_j)$ is the probability assigned to the value y_j, $j = 1, 2, \cdots, k$, by the empirical distribution

- The kernel is a non-negative real-valued integrable function satisfying $\int_{-\infty}^{+\infty} k_y(x)\,dx = 1$ to guarantee that

    the kernel method originates a density function. We will also have, $K_y(x) = \int_{-\infty}^{x} k_y(u)\,du$.

Appendix 1 – Deduction 1 of Greenwood's approximation (*Loss Models*)

- Kaplan-Meier estimator of $S(t)$: $\hat{S}(t) = \prod_{i=1}^{j-1} \left(\frac{r_i - s_i}{r_i} \right)$, $y_{j-1} \le t < y_j$, $j = 2,3,\cdots,k$

  where r_i is the (cardinal) of the risk set at age y_i, i.e. the number of individuals available to experience the event and s_i the number who actually did so.

- Let us assume that y_i and r_i are fixed (and known), that is s_i is the only random quantity in expression $\left(\frac{r_i - s_i}{r_i} \right)$. Discuss these hypotheses

- Let s_i be the estimate and S_i the corresponding estimator. It is straightforward to see that

  $S_i \sim b(r_i ; p_i)$ where $p_i = \dfrac{S(y_{i-1}) - S(y_i)}{S(y_{i-1})}$, i.e., p_i is the conditional probability of dying at age y_i, given that the individual is alive at age y_{i-1}. Obviously, $E(S_i) = r_i \, p_i$ and $\text{var}(S_i) = r_i \, p_i \, (1 - p_i)$.

- Let us proof that $\left(\frac{r_i - S_i}{r_i} \right)$ is an unbiased estimator for $\Pr\left(X > y_j \mid X > y_{j-1} \right)$.

$$E\left(\frac{r_i - S_i}{r_i}\right) = \frac{r_i - r_i\, p_i}{r_i} = 1 - p_i = 1 - \frac{S(y_{i-1}) - S(y_i)}{S(y_{i-1})} = \frac{S(y_i)}{S(y_{i-1})} = \Pr\left(X > y_j \mid X > y_{j-1}\right)$$

- The variance of the estimator is

$$\mathrm{var}\left(\frac{r_i - S_i}{r_i}\right) = \frac{r_i\, p_i\,(1 - p_i)}{r_i^2} = \frac{\dfrac{S(y_{i-1}) - S(y_i)}{S(y_{i-1})} \times \dfrac{S(y_i)}{S(y_{i-1})}}{r_i} = \frac{(S(y_{i-1}) - S(y_i))S(y_i)}{r_i\, S(y_{i-1})^2}$$

- Now let us consider the Kaplan-Meier estimator of $S(y_j)$.

  o Assuming that that S_i, S_k are independent, $i \neq k$, (discuss considering the sample size) we get

$$E\left(\hat{S}(y_j)\right) = E\left(\prod_{i=1}^{j}\left(\frac{r_i - S_i}{r_i}\right)\right) = \prod_{i=1}^{j} E\left(\frac{r_i - S_i}{r_i}\right) = \prod_{i=1}^{j} \frac{S(y_i)}{S(y_{i-1})} = \frac{S(y_j)}{S(y_0)}.$$

  o Let us now establish a general result for the variance of a product of independent variables:

$$\mathrm{var}(X_1\, X_2 \cdots X_n) = E(X_1^2\, X_2^2 \cdots X_n^2) - \left(E(X_1\, X_2 \cdots X_n)\right)^2$$
$$= E(X_1^2) \times E(X_2^2) \times \cdots \times E(X_n^2) - \mu_1^2 \times \mu_2^2 \times \cdots \times \mu_n^2$$
$$= \left(\mu_1^2 + \sigma_1^2\right) \times \cdots \times \left(\mu_n^2 + \sigma_n^2\right) - \mu_1^2 \times \cdots \times \mu_n^2$$

  o Then

$$\mathrm{var}\big(\hat{S}(y_j)\big) = \mathrm{var}\left(\prod_{i=1}^{j}\left(\frac{r_i - S_i}{r_i}\right)\right)$$

$$= \prod_{i=1}^{j}\left(\frac{S(y_i)^2}{S(y_{i-1})^2} + \frac{(S(y_{i-1})-S(y_i))S(y_i)}{r_i\,S(y_{i-1})^2}\right) - \prod_{i=1}^{j}\left(\frac{S(y_i)^2}{S(y_{i-1})^2}\right)$$

$$= \prod_{i=1}^{j}\left(\frac{r_i\,S(y_i)^2 + (S(y_{i-1})-S(y_i))S(y_i)}{r_i\,S(y_{i-1})^2}\right) - \frac{S(y_j)^2}{S(y_0)^2}$$

$$= \prod_{i=1}^{j}\left(\frac{S(y_i)\big(r_i\,S(y_i) + (S(y_{i-1})-S(y_i))\big)}{r_i\,S(y_{i-1})^2}\right) - \frac{S(y_j)^2}{S(y_0)^2}$$

$$= \prod_{i=1}^{j}\left(\frac{S(y_i)^2}{S(y_{i-1})^2} \times \frac{\big(r_i\,S(y_i) + (S(y_{i-1})-S(y_i))\big)}{r_i\,S(y_i)}\right) - \frac{S(y_j)^2}{S(y_0)^2}$$

$$= \prod_{i=1}^{j}\left(\frac{S(y_i)^2}{S(y_{i-1})^2}\right) \times \prod_{i=1}^{j}\left(\frac{r_i\,S(y_i) + (S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}\right) - \frac{S(y_j)^2}{S(y_0)^2}$$

$$= \frac{S(y_j)^2}{S(y_0)^2} \times \prod_{i=1}^{j}\left(1 + \frac{(S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}\right) - \frac{S(y_j)^2}{S(y_0)^2}$$

$$= \frac{S(y_j)^2}{S(y_0)^2} \times \left(\prod_{i=1}^{j}\left(1 + \frac{(S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}\right) - 1\right)$$

o Noting that $(1+a_1)\times(1+a_2)\times\cdots\times(1+a_n)\approx 1+a_1+a_2+\cdots+a_n$ when the a_i, $i=1,2,\cdots,n$, are small (the cross product will be even smaller) , we can approximate this unpleasant formula by

$$\text{var}\left(\hat{S}(y_j)\right)=\frac{S(y_j)^2}{S(y_0)^2}\times\left(\prod_{i=1}^{j}\left(1+\frac{(S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}\right)-1\right)$$

$$\approx\frac{S(y_j)^2}{S(y_0)^2}\times\left(1+\sum_{i=1}^{j}\frac{(S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}-1\right)$$

$$=\frac{S(y_j)^2}{S(y_0)^2}\times\left(\sum_{i=1}^{j}\frac{(S(y_{i-1})-S(y_i))}{r_i\,S(y_i)}\right)$$

o Now, we can obtain an estimate for the variance. To do so, remember that $E\left(\frac{r_i-S_i}{r_i}\right)=\frac{S(y_i)}{S(y_{i-1})}$

and consequently, $\left(\frac{r_i-s_i}{r_i}\right)$ is a good estimate for $\frac{S(y_i)}{S(y_{i-1})}$ or that we can use $r_i/(r_i-s_i)$ as an estimate for $S(y_{i-1})/S(y_i)$.

o **The Greenwood's approximation** for the variance of the Kaplan-Meier estimator is

$$\hat{\text{var}}\left(\hat{S}(y_j)\right) \approx S_n(y_j)^2 \times \sum_{i=1}^{j} \frac{\left(\dfrac{r_i}{r_i - s_i} - 1\right)}{r_i} = S_n(y_j)^2 \times \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)}$$

Appendix 2 – Deduction 2 of Greenwood's approximation (using delta method)

- **Delta method** – If the random variation of Y around μ is small (for example if μ is the mean of Y and $\operatorname{var} Y$ has order $1/n$ we can use

$$g(Y) \approx g(\mu) + (Y - \mu)\, g'(\mu) + (1/2)(Y - \mu)^2\, g''(\mu) + \cdots$$

Considering that $\mu = E(Y)$, using the first order expansion and taking expectations

$$E\big(g(Y)\big) \approx g(\mu)$$
$$\operatorname{var}\big(g(Y)\big) \approx \big(g'(\mu)\big)^2 \operatorname{var}(Y)$$

- Kaplan-Meier estimate of $S(t)$: $\hat{S}(t) = \prod_{i=1}^{j-1}\left(\dfrac{r_i - s_i}{r_i}\right)$, $y_{j-1} \le t < y_j$, $j = 2, 3, \cdots, k$

  where r_i is the (cardinal) of the risk set at age y_i, i.e. the number of individuals available to experience the event and s_i the number who actually did so.

- Let us assume that y_i and r_i are fixed (and known), that is S_i is the only random quantity in expression $\left(\dfrac{r_i - S_i}{r_i}\right)$. It is straightforward to see that $S_i \sim b(r_i\,; p_i)$ where p_i is the conditional probability of dying at age y_i, given that the individual is alive at age y_{i-1}.

  Obviously, $E(S_i) = r_i\, p_i$ and $\operatorname{var}(S_i) = r_i\, p_i\, (1 - p_i)$

- Using the delta method ($g(.) = \ln(.)$) we get $\mathrm{var}(\ln \hat{S}(t)) \approx \dfrac{1}{(S(t))^2} \, \mathrm{var}\!\left(\hat{S}(t)\right)$, i.e.

$$\mathrm{var}\!\left(\hat{S}(t)\right) \approx (S(t))^2 \, \mathrm{var}(\ln \hat{S}(t))$$

- $\mathrm{var}(\ln \hat{S}(t)) = \mathrm{var}\!\left(\ln \prod_{i=1}^{j-1}\!\left(\dfrac{r_i - S_i}{r_i} \right) \right) = \mathrm{var}\!\left(\sum_{i=1}^{j-1}\ln\!\left(\dfrac{r_i - S_i}{r_i} \right) \right) = \sum_{i=1}^{j-1}\mathrm{var}\!\left(\ln\!\left(\dfrac{r_i - S_i}{r_i} \right) \right)$ assuming that S_i, S_k

  are independent, $i \neq k$, (discuss considering the sample size)

- Let us use the delta method to approximate $\mathrm{var}\!\left(\ln\!\left(\dfrac{r_i - S_i}{r_i} \right) \right)$.

$$g(S_i) = \ln\!\left(\frac{r_i - S_i}{r_i} \right), \; E(S_i) = \mu_i = r_i\,p_i \text{ and then } g'(\mu_i) = \frac{(-1/r_i)}{(r_i - \mu_i)/r_i} = \frac{-1}{(r_i - \mu_i)}$$

$$\mathrm{var}\!\left(\ln\!\left(\frac{r_i - S_i}{r_i} \right) \right) \approx \left(\frac{-1}{(r_i - \mu_i)} \right)^2 \mathrm{var}(S_i) = \frac{r_i\,p_i\,(1 - p_i)}{(r_i - r_i\,p_i)^2} = \frac{p_i}{r_i\,(1 - p_i)}$$

- Then

49

$$\text{var}\left(\hat{S}(t)\right) \approx (S(t))^2 \, \text{var}(\ln \hat{S}(t)) = (S(t))^2 \sum_{i=1}^{j-1} \text{var}\left(\ln\left(\frac{r_i - S_i}{r_i}\right)\right) \approx (S(t))^2 \sum_{i=1}^{j-1} \frac{p_i}{r_i(1-p_i)}$$

- The estimate of the variance is then

$$\hat{\text{var}}\left(\hat{S}(t)\right) \approx (S(t))^2 \sum_{i=1}^{j-1} \frac{s_i}{r_i(r_i - s_i)}$$

Which is Greenwood's approximation

Note that $\hat{p}_i = s_i / r_i$

Appendix 3 – Intuition for the Nelson-Aalen estimator (*Loss Models*)

- Aim: To estimate the cumulative hazard rate

- Let

    o $r(t)$ be the risk set at time *t*

    o $h(t)$ be the hazard rate function

    o $s(t)$ be the expected total number of observed "death" prior to time *t,* i.e. $s(t) = \int_0^t r(u)\,h(u)\,du$

- As $\dfrac{d}{dt} s(t) = r(t)\,h(t)$, then $\dfrac{d\,s(t)}{r(t)} = h(t)\,dt$ and integrating both sides yields $\int_o^t \dfrac{d\,s(u)}{r(u)} = \int_0^t h(u)\,du = H(t)$,

    i.e. $H(t) = \int_o^t \dfrac{d\,s(u)}{r(u)}$.

- To estimate $H(t)$ we must replace the theoretical continuous variables $s(u)$ and $r(u)$ by the discrete

    observed values. Consequently $\hat{H}(t) = \sum_{t_i \le t} \dfrac{s_i}{r_i}$ $\qquad$ $t_i \le t$ means $i : y_i \le t$

Appendix 4 – Kaplan-Meier estimation and Greenwood's approximation

- The formulae for truncated and censored data, when applied to complete individual data, originate the same estimates as in chapter 13.

- Remember that, with complete data: $\quad r_j = r_{j-1} - s_{j-1}; \quad r_1 = n; \quad r_j = n - \sum_{i=1}^{j-1} s_i$ and

$$S_n(x) = \begin{cases} 1 & x < y_1 \\ \dfrac{r_j}{n} & y_{j-1} \leq x < y_j \quad j = 2,3,\cdots,k \\ 0 & y_k \leq x \end{cases}$$

- Kaplan-Meier estimation $S_n(t) = \begin{cases} 1 & t < y_1 \\ \displaystyle\prod_{i=1}^{j-1}\left(\dfrac{r_i - s_i}{r_i}\right) & y_{j-1} \leq t < y_j \quad j = 2,3,\cdots,k \\ \displaystyle\prod_{i=1}^{k}\left(\dfrac{r_i - s_i}{r_i}\right) \text{ or } 0 \text{ or } .... & y_k \leq t \end{cases}$

52

- Let us show that the Kaplan-Meier procedure $S_n(y_j) = \prod_{i=1}^{j}\left(\dfrac{r_i - s_i}{r_i}\right)$ originates the "classical" result

$$S_n(y_j) = \frac{r_{j+1}}{n} \text{ for } j = 1, 2, \cdots, k-1$$

$$S_n(y_j) = \prod_{i=1}^{j}\left(\frac{r_i - s_i}{r_i}\right) = \frac{r_1 - s_1}{r_1} \times \frac{r_2 - s_2}{r_2} \times \cdots \times \frac{r_j - s_j}{r_j} = \frac{r_2}{r_1} \times \frac{r_3}{r_2} \times \cdots \times \frac{r_{j+1}}{r_j} = \frac{r_{j+1}}{r_1} = \frac{r_{j+1}}{n} \text{ since } r_j - s_j = r_{j+1}$$

- Variance estimate (chapter 13): $\widehat{\mathrm{var}}(S_n^*(y_j)) = \dfrac{S_n(y_j)\left(1 - S_n(y_j)\right)}{n}$

$$\frac{S_n(y_j)\left(1 - S_n(y_j)\right)}{n} = \frac{1}{n} \times \frac{r_{j+1}}{n} \times \left(1 - \frac{r_{j+1}}{n}\right) = \frac{1}{n^3} \times r_{j+1} \times (n - r_{j+1})$$

- Greenwood's approximation: $\widehat{\mathrm{var}}\left(S_n^*(y_j)\right) \approx \left(S_n(y_j)\right)^2 \sum_{i=1}^{j} \dfrac{s_i}{r_i(r_i - s_i)}$

$$\sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)} = \sum_{i=1}^{j}\left(\frac{1}{r_i - s_i} - \frac{1}{r_i}\right) = \sum_{i=1}^{j}\left(\frac{1}{r_i - s_i}\right) - \sum_{i=1}^{j}\left(\frac{1}{r_i}\right) = \sum_{i=1}^{j}\left(\frac{1}{r_{i+1}}\right) - \sum_{i=1}^{j}\left(\frac{1}{r_i}\right) = \frac{1}{r_{j+1}} - \frac{1}{r_1}$$

$$\left(S_n(y_j)\right)^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)} = \left(\frac{r_{j+1}}{n}\right)^2 \times \left(\frac{1}{r_{j+1}} - \frac{1}{n}\right) = \left(\frac{r_{j+1}}{n}\right)^2 \times \frac{(n - r_{j+1})}{n \times r_{j+1}} = \frac{1}{n^3} \times r_{j+1} \times (n - r_{j+1})$$

Appendix 5 – Log transformed confidence interval - deduction

- The idea is to construct confidence intervals for a function $g\left(S(t)\right)$ and then to use the inverse of this function to get a confidence interval for $S(t)$. We will use the function $\theta = g\left(S(t)\right) = \ln\left(-\ln S(t)\right)$ and, for each given value of t, we will use $T = g\left(S_n(t)\right) = \ln\left(-\ln S_n(t)\right)$ to estimate θ.

- To obtain a confidence interval for θ, we use the delta method
   - The Delta Method uses Taylor series to approximate the mean and variance of the estimator.
   - Let us develop $T = g\left(S_n(t)\right)$ around $S(t)$ and take the **first order term**
   $$g\left(S_n(t)\right) = g\left(S(t)\right) + \left(S_n(t) - S(t)\right) g'\left(S(t)\right) + \cdots, \text{ that is}$$
   $$T \approx g\left(S(t)\right) + \left(S_n(t) - S(t)\right) g'\left(S(t)\right)$$
   - Now we get
   $$E(T) \approx g\left(S(t)\right) + E\left(S_n(t) - S(t)\right) g'\left(S(t)\right) = g\left(S(t)\right), \text{ as } S_n(t) \text{ (the random variable) is an}$$
   $$\text{unbiased estimator of } S(t), \text{ for each } t.$$
   $$\mathrm{var}(T) \approx \left(g'\left(S(t)\right)\right)^2 \mathrm{var}\left(S_n(t) - S(t)\right) = \left(g'\left(S(t)\right)\right)^2 \mathrm{var}\left(S_n(t)\right)$$
   - The standard error of T is $se(T) = \sqrt{\hat{\mathrm{var}}(T)} \approx \left|g'\left(S_n(t)\right)\right| \sqrt{\hat{\mathrm{var}}\left(S_n(t)\right)}$

- Note that the derivative of the function $g(x) = \ln(-\ln x)$ is $g'(x) = \dfrac{1}{x \ln x}$ and, for

$$0 < S_n(t) < 1, \; g'(S_n(t)) = \frac{1}{S_n(t) \ln S_n(t)}, \text{ i.e. } \left| g'(S_n(t)) \right| = -\frac{1}{S_n(t) \ln S_n(t)}$$

- The limits of the approximate confidence (level $1 - \alpha$) interval for $\theta = g(S(t))$ will be given by

$$\ln(-\ln S_n(t)) \pm z_{\alpha/2} \frac{\sqrt{\text{vâr}(S_n(t))}}{S_n(t) \times \ln S_n(t)}$$

o Now, using this result, we will construct a confidence interval for $S(t)$.

- First, note that $\theta = \ln(-\ln S(t)) \Leftrightarrow S(t) = \exp(-e^{\theta})$

- Recall that $\exp(a + b) = e^a \times e^b$ and $\exp(a \times b) = \left(e^a\right)^b$

- For the upper limit we get

$$UL = \exp\left(-\exp\left(\ln(-\ln S_n(t)) + z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right)$$

$$= \exp\left(-\exp(\ln(-\ln S_n(t)))\times\exp\left(z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right)$$

$$= \exp\left(\ln S_n(t)\times\exp\left(z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right) = (S_n(t))^U$$

with $U = \exp\left(z_{\alpha/2}\dfrac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)$

- For the lower limit we get

$$LL = \exp\left(-\exp\left(\ln(-\ln S_n(t)) - z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right)$$

$$= \exp\left(-\exp(\ln(-\ln S_n(t)))\times\exp\left(-z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right)$$

$$= \exp\left(\ln S_n(t)\times\exp\left(-z_{\alpha/2}\frac{\sqrt{\hat{var}(S_n(t))}}{S_n(t)\times\ln S_n(t)}\right)\right) = (S_n(t))^{1/U}$$

- Then the confidence interval is $\left((S_n(t))^{1/U} \; ; (S_n(t))^U \right)$ with $U = \exp\left(z_{\alpha/2} \dfrac{\sqrt{\hat{\mathrm{var}}(S_n(t))}}{S_n(t) \times \ln S_n(t)} \right)$.

- This interval will always be inside the range 0 to 1 and is referred to as the **log-transformed interval**.

Appendix 6 – Confidence intervals using the cumulative hazard function - deduction

- Similar results are available for the Nelson-Aalen estimator. Let us first derive a variance estimate for this estimator. As in the previous case all results are obtained assuming that the risk sets are known (not random). We will assume that the number of "deaths" at time t_i, s_i, follows approximately a Poisson distribution with parameter $r_i\, h(t_i)$. We could use a binomial distribution but a Poisson distribution is most commonly used,
  The variance of s_i, given r_i, is given by $r_i\, h(t_i)$ and, since $h(t_i)$ is estimated using s_i / r_i, we could use the approximation $\mathrm{v\hat{a}r}(s_i) = r_i \times s_i / r_i = s_i$.

- Then, assuming independence among the s_i / r_i, we get

$$\mathrm{v\hat{a}r}\big(\hat{H}(y_j)\big) = \mathrm{v\hat{a}r}\left(\sum_{i=1}^{j} \frac{s_i}{r_i}\right) = \sum_{i=1}^{j} \frac{\mathrm{v\hat{a}r}(s_i)}{r_i^2} = \sum_{i=1}^{j} \frac{s_i}{r_i^2}$$

- A direct (linear) confidence interval is then given by $\hat{H}(t) \pm z_{\alpha/2}\sqrt{\mathrm{v\hat{a}r}\big(\hat{H}(y_j)\big)}$ where $j = \max\{i : y_i \le t\}$

- A log-transformed interval is given by $\left(\hat{H}(t)U_L ; \hat{H}(t)U_U\right)$ with $U_L = \exp\left(-\dfrac{z_{\alpha/2}\sqrt{\text{vâr}\left(\hat{H}(y_j)\right)}}{H(t)}\right)$ and

$$U_U = \exp\left(+\dfrac{z_{\alpha/2}\sqrt{\text{vâr}\left(\hat{H}(y_j)\right)}}{H(t)}\right)$$

- The derivation is similar to the one developed for the survival function but now we use $\theta = g(H(t)) = \ln(H(t))$. Now $T = g\left(\hat{H}(t)\right) = \ln\left(\hat{H}(t)\right)$.
  - ○ Confidence interval for θ using the delta method
    - ▪ $T = g\left(\hat{H}(t)\right) \approx g(H(t)) + \left(\hat{H}(t) - H(t)\right)g'(H(t))$
      $\text{var}(T) \approx \left(g'(H(t))\right)^2 \text{var}\left(\hat{H}(t)\right)$
    - ▪ Note that the derivative of the function $g(x) = \ln x$ is $g'(x) = 1/x$ and, for $0 < \hat{H}(t)$, $g'\left(\hat{H}(t)\right) = 1/\hat{H}(t)$
    - ▪ The limits of the approximate confidence (level $1-\alpha$) interval for $\theta = g(H(t))$ will be given by
      $$\ln\left(\hat{H}(t)\right) \pm z_{\alpha/2}\dfrac{\sqrt{\text{vâr}\left(\hat{H}(t)\right)}}{\hat{H}(t)}$$
  - ○ Now, using this result, we will construct a confidence interval for $S(t)$.
    - ▪ $\theta = \ln(H(t)) \Leftrightarrow H(t) = \exp(\theta)$

- For the upper limit we get

$$UL = \exp\left(\ln(\hat{H}(t)) + z_{\alpha/2} \frac{\sqrt{\text{vâr}(\hat{H}(t))}}{\hat{H}(t)} \right)$$

$$= -\exp(\ln(\hat{H}(t))) \times \exp\left(z_{\alpha/2} \frac{\sqrt{\text{vâr}(\hat{H}(t))}}{\hat{H}(t)} \right)$$

$$= \hat{H}(t) \times U_U$$

With $U_U = \exp\left(z_{\alpha/2} \dfrac{\sqrt{\text{vâr}(\hat{H}(t))}}{\hat{H}(t)} \right)$

- For the lower limit we get

$$LL = \exp\left(\ln(\hat{H}(t)) - z_{\alpha/2} \frac{\sqrt{\text{vâr}(\hat{H}(t))}}{\hat{H}(t)} \right)$$

$$= -\exp(\ln(\hat{H}(t))) \times \exp\left(-z_{\alpha/2} \frac{\sqrt{\text{vâr}(\hat{H}(t))}}{\hat{H}(t)} \right)$$

$$= \hat{H}(t) \times U_L$$

- With $U_L = \exp\left(-z_{\alpha/2} \dfrac{\sqrt{\mathrm{v\hat{a}r}\left(\hat{H}(t)\right)}}{\hat{H}(t)} \right)$

Appendix 7 – Example 14.16

Example 14.16 – Determine the kernel density estimate for Example 13.2 using each of the three kernels.

- Results using the uniform kernel with b=0.1, b=1.0 and b=0.5
- R program for other kernels (including Pareto kernel)

Uniform kernel

Sample $(1.0; 1.3; 1.5; 1.5; 2.1; 2.1; 2.1; 2.8)$

y_j	1.0	1.3	1.5	2.1	2.8
$p(y_j)$	1/8	1/8	2/8	3/8	1/8

Bandwith b=0.1 then $1/(2b) = 5$

1.0		0.9	1.1
1.3		1.2	1.4
1.5	→	1.4	1.6
2.1		2.0	2.2
2.8		2.7	2.9

$$\hat{f}(x) = \begin{cases} 5/8 & 0.9 < x < 1.1 \\ 5/8 & 1.2 < x < 1.4 \\ 10/8 & 1.4 < x < 1.6 \\ 15/8 & 2.0 < x < 2.2 \\ 5/8 & 2.7 < x < 2.9 \\ 0 & \text{otherwise} \end{cases}$$

Discuss the problem related to the intervals limit

Bandwith b=0.5 then $1/(2b) = 1$

1.0		0.5	1.5
1.3		0.8	1.8
1.5	→	1.0	2.0
2.1		1.6	2.6
2.8		2.3	3.3

$$\hat{f}(x) = \begin{cases} 3/80 & 0.5 < x < 0.8 \\ 4/8 & 0.8 < x < 1.0 \\ 20/80 & 1.0 < x < 1.5 \\ 3/80 & 1.5 < x < 1.6 \\ 12/80 & 1.6 < x < 1.8 \\ 10/80 & 1.8 < x < 2.0 \\ 9/80 & 2.0 < x < 2.3 \\ 12/80 & 2.3 < x < 2.6 \\ 7/80 & 2.6 < x < 3.3 \\ 0 & \text{otherwise} \end{cases}$$

Bandwith b=1.0 then $1/(2b) = 0.5$

1.0		0.0	2.0
1.3		0.3	2.3
1.5	→	0.5	2.5
2.1		1.1	3.1
2.8		1.8	3.8

$$\hat{f}(x) = \begin{cases} 1/16 & 0 < x < 0.3 \\ 2/16 & 0.3 < x < 0.5 \\ 4/16 & 0.5 < x < 1.1 \\ 7/16 & 1.1 < x < 1.8 \\ 8/16 & 1.8 < x < 2.0 \\ 7/16 & 2.0 < x < 2.3 \\ 6/16 & 2.3 < x < 2.5 \\ 4/16 & 2.5 < x < 3.1 \\ 1/16 & 3.1 < x < 3.8 \\ 0 & \text{otherwise} \end{cases}$$

R Program

```
# "handmade"
fhat_uniform=function(x,y,s,b){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=dunif(x,y[i]-b,y[i]+b)
  return(sum((s/n)*a))
  }
```

```
Fhat_uniform=function(x,y,s,b){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=punif(x,y[i]-b,y[i]+b)
  return(sum((s/n)*a))
    }

fhat_triangular=function(x,y,s,b){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=(x-y[i]+b)*(x<y[i])*(x>=(y[i]-b))+
                    (-x+y[i]+b)*(x>=y[i])*(x<=(y[i]+b))
  return(sum((s/n)*a))
  }

Fhat_triangular=function(x,y,s,b){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=((x-y[i]+b)^2/(2*b^2))*(x<y[i])*(x>=(y[i]-b))+
                    (1-(-x+y[i]+b)^2/(2*b^2))*(x>=y[i])*(x<=(y[i]+b))+
                    1*(x>(y[i]+b))
  return(sum((s/n)*a))
    }
```

```
fhat_gamma=function(x,y,s,alpha){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=dgamma(x,shape=alpha,rate=alpha/y[i])
  return(sum((s/n)*a))
  }

Fhat_gamma=function(x,y,s,alpha){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=pgamma(x,shape=alpha,rate=alpha/y[i])
  return(sum((s/n)*a))
   }

# Pareto1 uses logs to avoid overflow
# Using Pareto Kernel we always got too much smooth
fhat_Pareto1=function(x,y,s,alpha){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=exp(log(alpha)+alpha*log(y[i]*(alpha-1))-(alpha+1)
*log(x+y[i]*(alpha-1)))
  return(sum((s/n)*a))
  }
```

```
fhat_Pareto=function(x,y,s,alpha){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=alpha*((y[i]*(alpha-1))^alpha)/(x+y[i]*(alpha-
1))^(alpha+1)
  return(sum((s/n)*a))
  }

Fhat_Pareto=function(x,y,s,alpha){
  n=sum(s); k=length(y); a=rep(0,k);
  for(i in 1:k)a[i]=1-((y[i]*(alpha-1))/(x+y[i]*(alpha-1)))^alpha
  return(sum((s/n)*a))
    }



y=c(1.0,1.3,1.5,2.1,2.8)
s=c(1,1,2,3,1)

# Select the number of points to be used, NP, and the bounds LB and UB
NP=400; LB=0; UB=4.0;

x=(1:NP)*((UB-LB)/NP)+LB
z1=rep(NA,NP); z2=z1; z3=z1; z4=z1; z5=z1; z6=z1; z7=z1; z8=z1;
```

```
for(i in 1:400){
  z1[i]=fhat_gamma(x[i],y,s,500)
  z2[i]=Fhat_gamma(x[i],y,s,500)
  z3[i]=fhat_Pareto1(x[i],y,s,500)
  z4[i]=Fhat_Pareto(x[i],y,s,500)
  z5[i]=fhat_uniform(x[i],y,s,0.1)
  z6[i]=Fhat_uniform(x[i],y,s,0.1)
  z7[i]=fhat_triangular(x[i],y,s,0.1)
  z8[i]=Fhat_triangular(x[i],y,s,0.1)
  }
plot(x,z1,type="l",main="Gamma with alpha=500")
plot(x,z2,type="l",main="Gamma with alpha=500")
plot(x,z3,type="l",main="Pareto with alpha=500")
plot(x,z4,type="l",main="Pareto with alpha=500")
plot(x,z5,type="l",main="Uniform with bw=0.1")
plot(x,z6,type="l",main="Uniform with bw=0.1")
plot(x,z7,type="l",main="Triangular with bw=0.1")
plot(x,z8,type="l",main="Triangular with bw=0.1")
```

```r
for(i in 1:400){
  z1[i]=fhat_gamma(x[i],y,s,50)
  z2[i]=Fhat_gamma(x[i],y,s,50)
  z3[i]=fhat_Pareto(x[i],y,s,5)
  z4[i]=Fhat_Pareto(x[i],y,s,5)
  z5[i]=fhat_uniform(x[i],y,s,1.0)
  z6[i]=Fhat_uniform(x[i],y,s,1.0)
  z7[i]=fhat_triangular(x[i],y,s,1.0)
  z8[i]=Fhat_triangular(x[i],y,s,1.0)
  }
plot(x,z1,type="l",main="Gamma with alpha=50")
plot(x,z2,type="l",main="Gamma with alpha=50")
plot(x,z3,type="l",main="Pareto with alpha=5")
plot(x,z4,type="l",main="Pareto with alpha=5")
plot(x,z5,type="l",main="Uniform with bw=1.0")
plot(x,z6,type="l",main="Uniform with bw=1.0")
plot(x,z7,type="l",main="Triangular with bw=1.0")
plot(x,z8,type="l",main="Triangular with bw=1.0")

# using R function "density"
p=s/sum(s)
res=density(y,0.1/sqrt(3),kernel="r",weights=p)
```

```
res$x
res$y
plot(res)
res=density(y,1.0/sqrt(3),kernel="r",weights=p)
plot(res)
res=density(y,0.1/sqrt(6),kernel="t",weights=p)
plot(res)
res=density(y,1.0/sqrt(6),kernel="t",weights=p)
plot(res)


density(x, ...)
## Default S3 method:
density(x, bw = "nrd0", adjust = 1,
        kernel = c("gaussian", "epanechnikov", "rectangular",
                   "triangular", "biweight",
                   "cosine", "optcosine"),
        weights = NULL, window = kernel, width,
        give.Rkern = FALSE,
        n = 512, from, to, cut = 3, na.rm = FALSE, ...)
```