ESTIMATION FOR MODIFIED DATA

- **Definition 12.1 (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 € for a claim and if the insurance company only records the payments made, any time a loss is above 10000 € the amount of the claim will be unknown but we will know that a payment of 10000 € has happened.
  - The censoring points could be known (defined by the insurance policy) 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.
- From a statistical point of view, **truncation is a more severe limitation than censoring**.
- 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
    - Kernel estimation – Method to estimate the **density** (distribution) function of a **continuous** random variable.
    - 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 $k \le n_1$ where $n_1 \le n$ 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,

    $r_j = \#\{d_i < y_j\} - \#\{x_i < y_j\} - \#\{u_i < y_j\}$ or $r_j = \#\{x_i \ge y_j\} + \#\{u_i \ge y_j\} - \#\{d_i \ge y_j\}$. □

- **Example 12.1 (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
  - $\hat{\pi}_j = \frac{r_j - s_j}{r_j}$. ($r_j - s_j$) – number of ind. *surviving* at y_j and r_j – risk set
  - $\hat{\Pr}(X \geq y_j) = \hat{\Pr}(X > y_{j-1}) = \hat{S}(y_{j-1})$. No events in the sample between y_{j-1} and y_j
  - $\hat{S}(y_j) = \hat{\pi}_j \times \hat{S}(y_{j-1}) = \hat{\pi}_j \times \hat{\pi}_{j-1} \times \hat{S}(y_{j-2}) = \hat{\pi}_j \times \hat{\pi}_{j-1} \times \cdots \times \hat{\pi}_1 \times \hat{S}(0) = \prod_{i=1}^{j} \hat{\pi}_i$

- Then

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

- Let us discuss the estimation of the survival function when $t \ge y_k$.
  - $S_n(y_k) = \prod_{i=1}^{k}\left(\frac{r_i - s_i}{r_i}\right)$
  - 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).

- 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, $S_n(t) = \prod_{i=1}^{k} \left(\frac{r_i - s_i}{r_i} \right)$, $t \geq y_k$.
  - Keep the survival function at its last value until the last censoring time is reached and then declare the function to be 0,

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

    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 $s^* = \prod_{i=1}^{k} \left(\frac{r_i - s_i}{r_i} \right)$ and w = max (last censoring time, y_k) and use $S_n(t) = \exp\left(\frac{t}{w} \ln s^* \right) = (s^*)^{t/w}$, $t \geq w$. 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 12.2** (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
```

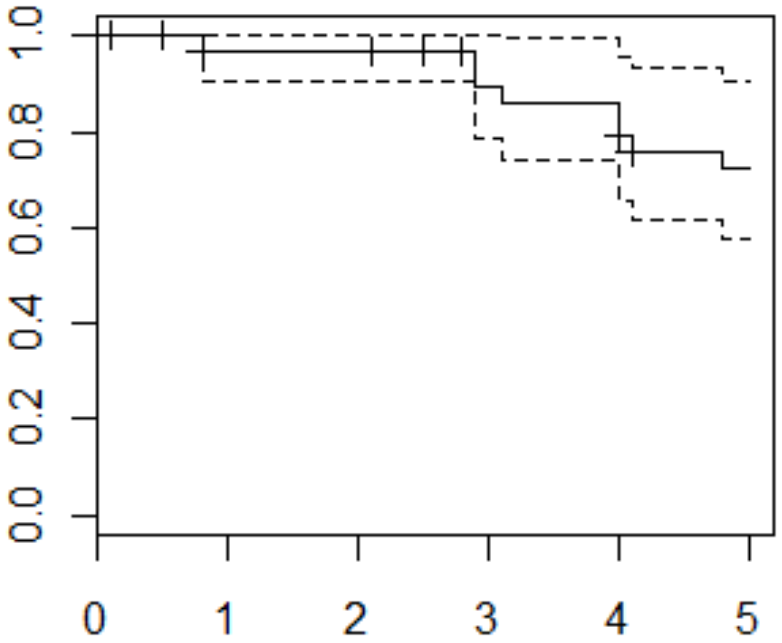
  Alternatively

  # 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.


0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5

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 $\hat{H}(t) = \sum_{i:y_i \le t} \frac{s_i}{r_i}$

- The **Nelson-Aalen estimator** is

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

- 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 12.3 (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 $S(t)$: $S_n(t) = \prod_{i=1}^{j-1} \left(\frac{r_i - s_i}{r_i} \right)$, $y_{j-1} \le t < y_j$, $j = 2,3,\cdots,k$.

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

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

  See *Loss Models* for a deduction of Greenwood's approximation or use the delta method. □
- Greenwood's formula can be written as $\hat{\text{var}}\left(S_n(t)\right) \approx S_n(t)^2 \times \sum_{i:y_i \le t} \frac{s_i}{r_i(r_i - s_i)}$

- **Example 12.9 (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:

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

  Directly – There is no censoring or truncation

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

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

  Greenwood's formula

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

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

  To apply Greenwood's formula to estimate $\text{var}({}_2\hat{q}_x \mid S(3)=S_n(3))$ 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) $r_1 = 27;\ s_1 = 1;\ r_2 = 26;\ s_2 = 1;\ r_3 = 25;\ s_3 = 2;$

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

- **Example 12.10** (14.13) – Repeat example 12.9 (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

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

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

Directly – not possible due to censoring and truncation

Greenwood's formula

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

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

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

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

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

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

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

$$= 0.005950$$

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 $g\left(S(t)\right)$ (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 $g\left(S(t)\right)=\ln\left(-\ln S(t)\right)$ and, for each given value of *t*, we will use $g\left(S_n(t)\right)=\ln\left(-\ln S_n(t)\right)$ to estimate $g\left(S(t)\right)$. Note that $-\infty < g\left(S(t)\right) < +\infty$.

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

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

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

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

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

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

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

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

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

$$g(S_n(t)) \pm z_{\alpha/2}\, se\left(g(S_n(t))\right) \quad \text{i.e.} \quad \ln(-\ln S_n(t)) \pm z_{\alpha/2} \frac{\sqrt{\hat{\text{var}}(S_n(t))}}{S_n(t) \times \ln S_n(t)}$$

- 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 $UB = (S_n(t))^U$ and $LB = (S_n(t))^{1/U}$ with $U = \exp\left(z_{\alpha/2} \frac{\sqrt{\hat{\text{var}}(S_n(t))}}{S_n(t) \times \ln S_n(t)} \right)$
  - Then the confidence interval is $\left((S_n(t))^{1/U} ; (S_n(t))^U \right)$. This interval will always be inside the range 0 to 1 and is referred to as the **log-transformed interval**.

- **Example 12.11 (14.14)** – Obtain the log-transformed confidence interval for $S(3)$ as in example 12.10 (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.
- We will use the same set of hypothesis: the y_i and the risk sets (r_i) are known (not random) and 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 $\hat{\text{var}}(S_i) = r_i \times s_i / r_i = s_i$.

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

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

The "Formulae and Tables for Examination" book uses the binomial approach and get

$$\hat{\text{var}}\left(\hat{H}(t)\right) = \sum_{i:y_i \le t} \frac{\hat{\text{var}}(S_i)}{r_i^2} = \sum_{i:y_i \le t} \frac{s_i (r_i - s_i)}{r_i^3}$$

In the remaining **we will use Loss Models formula**.

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

- A log-transformed confidence interval is given by $\left(\hat{H}(t) \times (1/U); \hat{H}(t) \times U\right)$ where

$$U = \exp\left(\frac{z_{\alpha/2} \sqrt{\hat{\text{var}}\left(\hat{H}(t)\right)}}{\hat{H}(t)}\right)$$

The formula is deduced using the Delta method with $g\left(\hat{H}(t)\right) = \ln\left(\hat{H}(t)\right)$ since $\ln\left(\hat{H}(t)\right)$ is unbounded.

- **Example 12.12 (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.

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

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

  Log-transformed confidence interval: (0.035472; 0.342702)

$$U=\exp\left(1.96\frac{\sqrt{0.0040697}}{0.11026}\right)=3.108225$$

  □Question: How to use these results in order to get CI for $S(3)$?

  Direct → (0.79038; 1)

  Log-transformed → (0.70985; 0.96515)

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 12.2** (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**
  - 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?
  - 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.
  - $p(y_j)$ 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 12.3** (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,

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

- In the remaining of the course we will follow Definition 12.2 (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)).

- **Example 12.13 (14.16)** – Determine the kernel density estimate for Example 11.2 (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	→	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 $1/(2b) = 0.5$

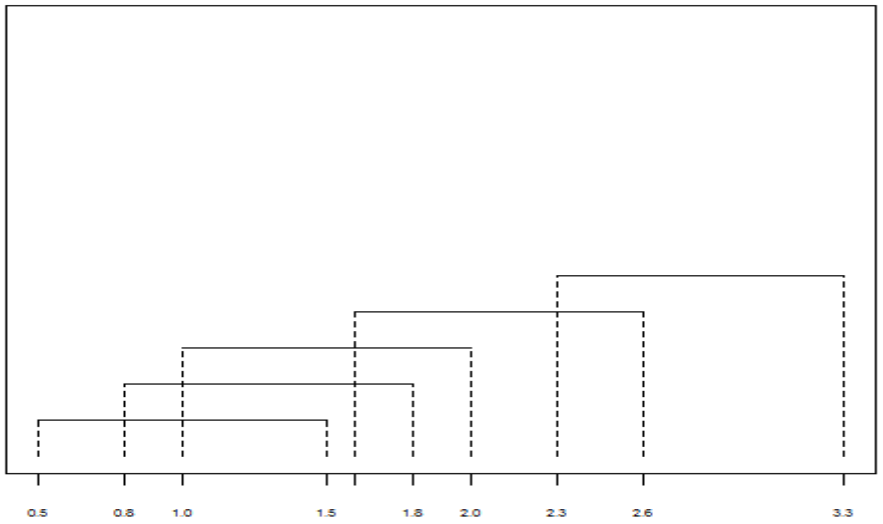
1.0		0.0	2.0
1.3		0.3	2.3
1.5	→	0.5	2.5
2.1		1.1	3.1
2.8		1.8	3.8

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

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

$$\hat{f}(x)=\begin{cases} 1/8 & 0.5<x<0.8 \\ 2/8 & 0.8<x<1.0 \\ 4/8 & 1.0<x<1.5 \\ 3/8 & 1.5<x<1.6 \\ 6/8 & 1.6<x<1.8 \\ 5/8 & 1.8<x<2.0 \\ 3/8 & 2.0<x<2.3 \\ 4/8 & 2.3<x<2.6 \\ 1/8 & 2.6<x<3.3 \\ 0 & \text{otherwise} \end{cases}$$



Using R

```
y=c(1.0,1.3,1.5,2.1,2.8); s=c(1,1,2,3,1); n=sum(s)
p_y=s/n
x=seq(0,4,by=0.025); fx=rep(NA,length(x))

# Uniform kernel
b=0.5; LU=y-b; UU=y+b
for(i in 1:length(x)) fx[i]=sum(p_y*dunif(x[i],LU,UU))
plot(x,fx,type="l",main="example 14.16 - Uniform kernel with b=0.5")

# Gamma kernel
alpha=50
for(i in 1:length(x)) fx[i]=sum(p_y*dgamma(x[i],shape=alpha,scale=y/alpha))
plot(x,fx,type="l",main="example 14.16 - gamma kernel with alpha=50")
```

- **Example (New)** – Using the data of the previous example estimate $F(2)$ using a uniform kernel with b=0.5.

  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

$$\hat{F}(2) = \frac{1}{8}\times 1 + \frac{1}{8}\times 1 + \frac{2}{8}\times 1 + \frac{3}{8}\times \frac{(2-2.1+0.5)}{1} + \frac{1}{8}\times 0$$

$$= \frac{5.2}{8}$$

APPROXIMATIONS FOR LARGE DATA SETS – We will skip this section