Expected and Relative Survival

Vincenzo Coviello
Department of Prevention ASL BA/1
Minervino Murge (Ba)
Email: coviello@mythnet.it
Outline of talk

• Estimating Expected Survival

• stexpect

• Example 1: clinical survival study

• Example 2: Population-based survival study
ESTIMATING EXPECTED SURVIVAL (1)
Definition

Expected survival is the survival in a reference population which is similar to the study cohort of patients at the start of follow-up, where the matching factors are usually age, calendar time, sex and optionally other variables (race, census).

The estimate is achieved through population mortality rate tables.
Using population mortality rates: stexpect

1. Estimates individual expected survival, the building block of the overall curve.

2. Combines these individual estimates to give the expected survival of the cohort according to three methods:
   - Ederer or “exact”
   - Hakulinen
   - Conditional or Ederer II
### Individual Expected Survival

- A 36 years old man born on 23\(^{th}\) April 1964
- Followed-up from 15\(^{th}\) June 2000 to 25\(^{th}\) October 2001

<table>
<thead>
<tr>
<th>Follow-up</th>
<th>Hazard per day</th>
<th>Cumulative hazard (( \hat{\lambda} ))</th>
<th>Survival probability exp(-( \hat{\lambda} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-Jun-2000</td>
<td>23-Apr-2001</td>
<td>0.000000155</td>
<td>0.0004836</td>
</tr>
<tr>
<td>23-Apr-2001</td>
<td>25-Oct-2001</td>
<td>0.000000161</td>
<td>0.00029785</td>
</tr>
</tbody>
</table>

Cumulative hazard from 15-Jun-2000 to 25-Oct-2001 = 0.00078145

Survival probability \( \exp(-0.00078145) \) = 0.999218855
Formulas

• Ederer and Hakulinen method:

\[ S_e(t + s) = S_e(t) \frac{\sum S_i(t + s)}{\sum S_i} \]

• Conditional or Ederer 2 method:

\[ S_e(t + s) = S_e(t) \exp \left( \frac{\sum h_i(t, s)Y_i(t)}{\sum Y_i(t)} \right) \]

where \( Y_i \) is 1 if the subject is at risk at time \( t \) and 0 otherwise.
Problems in large data sets

• To compute the above equations the time axis must be partitioned at every observed failure and censored time.

• In large data sets this episode splitting may require huge amounts of memory.
Approximation

• The range of follow-up times is partitioned in $n$ evenly spaced points. In such fixed width intervals each subject will contribute to the expected survival with a weight equals to the proportion of time for which he is observed.

Ederer - Hakulinen approximate formula:

$$S_e(t + s) = S_e(t) \frac{\sum S_i w_i (t + s)}{\sum S_i w_i}$$

where $$w_i = \frac{(t_i - t)}{s}$$
stexpect
stexpect ..., ratevar(varname) output(filename [,replace])

[ method(#) ]

They are not options

- **ratevar(varname)**: variable containing the general population mortality rates

- **output(filename [,replace])**: file where the estimates will be stored

**method(#)**: methods to be used

1 = Ederer I

2 = Ederer II or Conditional

3 = Hakulinen (default)
stexpect ..., ...

by(varlist) : up to 5 variables specifying separate groups over which the expected survival is to be calculated.

at(numlist) : analysis times at which the expected survival is to be computed.

npoints(#) : number of equally spaced points in the range of follow-up times used for the approximate estimate.
Before using stexpect one needs to

1. `stset` data using the `id()` option.

2. split follow-up time by age and calendar period.

3. merge the cohort data set with the file of reference population rates.
Example 1

Clinical Survival Study
MGUS Study

• 241 cases of Monoclonal Gammopathy of Undetermined Significance.

• **time** is in days since identification to death or occurrence of lymphoproliferative disease or to the end of the study.

• **status** is a failure/censor indicator.

Contains data from C:\Convegni2004\mgusconvegno.dta

obs: 241
vars: 12

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>int</td>
<td>%9.0g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>byte</td>
<td>%9.0g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>float</td>
<td>%9.0g</td>
<td></td>
<td>Time since Diagnosis</td>
</tr>
<tr>
<td>status</td>
<td>byte</td>
<td>%17.0g</td>
<td>status</td>
<td></td>
</tr>
</tbody>
</table>

...omitted...
Preparing the dataset

1 – stset data
   . stset time, f(status) id(id) scale(365.25)

2 – split the follow-up time by age and calendar period
   . stsplit fu, at(0(1)25)
   . gen age = agedia+fu
   . gen year = yeardiagnosis + fu

3 – merge the cohort data with a file (usrate) of reference rates
   . sort year age sex
   . merge year age sex using usrate, keep(rate) uniquus nokeep
stexpect, ratevar(rate) out(cond_example,replace) method(2)

- **rate** is the variable containing reference population rates
- **method(2)** specifies that the conditional estimate is to be computed
- **cond_example** is the output file structured as follows:

```
use cond_example,clear
list in 1/3, noobs
```

<table>
<thead>
<tr>
<th>t_exp</th>
<th>atrisk</th>
<th>Survexp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>241</td>
</tr>
<tr>
<td>.00027405</td>
<td>241</td>
<td>.99998966</td>
</tr>
<tr>
<td>.08487337</td>
<td>239</td>
<td>.9968664</td>
</tr>
</tbody>
</table>
```
Survexp saves the estimate of the expected survival. The user can define a different name for this variable:

\[ \text{stexpect [ newvarname ]}, \ldots \]

t_exp stores the times at which the function is estimated. If \( \text{at(numlist)} \) is omitted, they correspond to each survival time.

atrisk contains the number of subjects at risk at the time t_exp.
Check the validity of the results

The table below lists the results at the last five follow-up times achieved by stexpect and by the R macro survexp.

```
. list t_exp Survexp R_est in -5/1,noobs

+-------------------------------+-----------------+-------------+
<table>
<thead>
<tr>
<th>t_exp     Survexp   R_est</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.277892   .22859971   .2286</td>
</tr>
<tr>
<td>27.359343      .20821   .2082</td>
</tr>
<tr>
<td>27.712526   .20168448   .2017</td>
</tr>
<tr>
<td>28.361396   .18769732   .1877</td>
</tr>
<tr>
<td>34.105407   .07531006   .0753</td>
</tr>
</tbody>
</table>
+-------------------------------+
```
at(numlist) and by(varlist)

To illustrate these options new conditional estimates are saved in the file cond_byex:

```
stexpect, ratevar(rate) out(cond_byex, replace) ///
method(2) at(0(1)25) by(sex)
```

The file cond_byex will record the expected survival
• at the times $t_{exp} = 0, 1, 2, \ldots, 24, 25$
• for each value of byvar sex.
Output file with **by(varlist)** and **at(numlist)**

```stata
. use cond_byex,clear

. list if t_exp>20,noobs
```

<table>
<thead>
<tr>
<th>sex</th>
<th>t_exp</th>
<th>atrisk</th>
<th>Survexp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>19</td>
<td>0.24535683</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>11</td>
<td>0.22539159</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>8</td>
<td>0.2075762</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>6</td>
<td>0.18930929</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>4</td>
<td>0.17506198</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>21</td>
<td>0.45990346</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>12</td>
<td>0.434795</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>7</td>
<td>0.40512875</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>5</td>
<td>0.38333169</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>4</td>
<td>0.36152862</td>
</tr>
</tbody>
</table>
Other methods

• To estimate the expected survival according to Ederer or Hakulinen method, the follow-up time of the subjects must be set differently.

• So the expected survival of the three methods cannot be estimated sequentially, because each of them needs a different `timevar` in the `stset` statement.
• To estimate the expected survival, subjects in data set are to be considered as elements within the reference population. Fixing the follow-up of these elements at the observed times in the study cohort, as in Conditional method, is meaningless.

• Follow-up time in Ederer and Hakulinen methods actually matches the expected survival definition “The survival in a reference population which is similar to the study cohort of patients at the start of follow-up”.
Follow-up Time

• **Ederer Method**
  
The follow-up time is the same for all of the subjects and corresponds to the largest time at which an expected survival estimate is required.

• **Hakulinen Method**
  
The follow-up time is the actual censoring time for those subjects who are censored and the “maximum potential follow-up” for those who have died.
  
Find the rationale in references (3) and (4).
Ederer Method

Expected Survival until 25 years from diagnosis

1 – stset

```stset suvderer, f(status) id(id) scale(365.25)`

2 – merge with the file of reference rates

```stsplot fu,at(0(1)35)
gen age = aged+fu
gen year=yeard+fu
sort year age sex
merge year age sex using c:\data\usrate, nokeep ///
keep(rate)```
Ederer Method with stexpect

```
stexpect, ratevar(rate) out(ederer_ex, replace) ///
method(1) at(0(1)25) by(sex)
```

- method(1) tells stexpect to use the Ederer-Hakulinen formula.

- at(numlist) is not an option in this method since no failure occurs during the follow-up.
Results with Ederer Method

\[ . \text{use ederer} \_\text{ex,} \text{clear} \]
\[ . \text{list if t} \_\text{exp}<5, \text{noobs} \]

\begin{verbatim}
+----------------------------------+
<table>
<thead>
<tr>
<th>sex   t_exp   atrisk     Survexp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1       0      140           1</td>
</tr>
<tr>
<td>1       1      140   .95254107</td>
</tr>
<tr>
<td>1       2      140   .90595187</td>
</tr>
<tr>
<td>1       3      140   .86049999</td>
</tr>
<tr>
<td>1       4      140   .81635571</td>
</tr>
<tr>
<td>2       0      101           1</td>
</tr>
<tr>
<td>2       1      101   .97917553</td>
</tr>
<tr>
<td>2       2      101   .95746886</td>
</tr>
<tr>
<td>2       3      101   .93495095</td>
</tr>
<tr>
<td>2       4      101   .91170243</td>
</tr>
<tr>
<td>----------------------------------</td>
</tr>
</tbody>
</table>
\end{verbatim}

Note that the number at risk does not change.
Hakulinen’s Method

The “maximum potential follow-up time” for failed subjects is settled as the difference between the most optimistic last contact date and the enrollment date.

The MGUS study ends at August 1, 1990. So, the survival time according to Hakulinen’s method is set as:

```stata
    gen survhakulinen = cond(status, mdy(8,1,1990)-datediag, time)
    stset survhakulinen, f(status) id(id) scale(365.25)
```

Merge instructions are omitted.
Hakulinen’s Method with stexpect

\texttt{stexpect, ratevar(rate) out(hakulinen\_ex,replace) ///}
\texttt{at(0(1)25) by(sex)}

method(3) is omitted because it is the default.

Since the follow-up time has been modified, the number of subjects at risk is not the same as in the study cohort.
In the next slide a graph comparing the three estimates is shown. Here are the lines to achieve it:

```stata
use hakulinen_ex, clear
rename Survexp Hakulinen
merge sex t_exp using ederer_ex, keep(Survexp)
rename Survexp Ederer
drop _m
sort sex t_exp
merge sex t_exp using cond_byexample, keep(Survexp)
rename Survexp Conditional
twoway line Hakulinen Conditional Ederer t_exp, ///
    legend(label(1 "Hakulinen") label(2 "Conditional") ///
    label(3 "Ederer") row(1)) xla(0(5)25) ///
    by(sex, legend(pos(12))) clc(black red green)
```
The three methods often yield similar results. The Conditional estimate is of a small amount lower than the Ederer and Hakulinen estimates, which overlap completely.
Comparison of Observed vs. Expected Survival

An useful and suggestive endpoint of this analysis is to compare Kaplan-Meier survival estimates with the expected survival at a given time from various occurrence.

Graphs by sex
Example 2

Population-based Survival Study
Relative Survival

• Relative survival is the preferred measure for survival analysis based on population cancer registry data mainly because it does not depend on the information on cause of death.

• It is computed as the ratio between observed and expected survival.

• Relative survival can be estimated using **sts gen** to produce an estimate of the observed survival and **stexpect** for the expected survival.
Melanoma Data of the Finnish Cancer Registry

2145 patients with localised skin melanoma in Finland during 1975-1984.

. use melanoma,clear
(Skin melanoma, all stages, Finland 1975-94, follow-up to 1995)

. keep if year8594==0 & stage==1

Contains data from melanoma.dta
obs: 2,145
vars: 14

-------------------------------------------------------------------
storage  display      value
variable name  type  format      label  variable label
-------------------------------------------------------------------
id         int    %9.0g
sex        byte   %9.0g      sex     Sex
surv_mm    float  %9.0g       Survival time in
            completed months
status     byte   %17.0g     status  Vital status at last
date of contact
...omissis
Hakulinen’s Method for Relative Survival

- \textit{surv\_mm} is the timevar in months from diagnosis,
- \textit{status} is coded 1 or 2 if death occurs and 0 otherwise.

The analysis cutoff is set at December 31, 1995

As shown before, the survival time must be adapted to get the Hakulinen expected survival:

\begin{verbatim}
.gen surv_hak=cond(status==1|status==2, ///
(1995-yydx)*12+(12-mmdx), surv_mm)
.stset surv_hak,f(status==1 2) id(id) scale(12)
\end{verbatim}
• Data are expanded by age and calendar period:

```
stsplit fu, at(0(1)20)
replace age = age + fu
gen int year = yydx + fu
```

• File popmort with reference rates is merged with patients data:

```
sort year sex age
merge year sex age using popmort, keep(rate) nokeep
```
Estimates with and without the `np(#)` option

• In this small data set the expected survival can be estimated both using `np(#)` option

```
.stexpect, ratevar(rate) at(0(1)20) ///
   out(apprmelanhak,replace) np(100)
```

• and without using it

```
.stexpect, ratevar(rate) at(0(1)20) out(melanhak,replace)
```

These estimates are compared with the results produced by SURV3, a DOS program designed for the survival analysis based on cancer registry data.
<table>
<thead>
<tr>
<th>Time</th>
<th>np(100)</th>
<th>SURV 3</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97904</td>
<td>0.97904</td>
<td>0.97904</td>
</tr>
<tr>
<td>2</td>
<td>0.95808</td>
<td>0.95808</td>
<td>0.95808</td>
</tr>
<tr>
<td>3</td>
<td>0.93703</td>
<td>0.93704</td>
<td>0.93703</td>
</tr>
<tr>
<td>4</td>
<td>0.91594</td>
<td>0.91595</td>
<td>0.91594</td>
</tr>
<tr>
<td>5</td>
<td>0.89482</td>
<td>0.89483</td>
<td>0.89482</td>
</tr>
<tr>
<td>6</td>
<td>0.87361</td>
<td>0.87362</td>
<td>0.87361</td>
</tr>
<tr>
<td>7</td>
<td>0.85238</td>
<td>0.85238</td>
<td>0.85238</td>
</tr>
<tr>
<td>8</td>
<td>0.83113</td>
<td>0.83113</td>
<td>0.83113</td>
</tr>
<tr>
<td>9</td>
<td>0.80992</td>
<td>0.80992</td>
<td>0.80992</td>
</tr>
<tr>
<td>10</td>
<td>0.78880</td>
<td>0.78880</td>
<td>0.78880</td>
</tr>
<tr>
<td>11</td>
<td>0.76774</td>
<td>0.76773</td>
<td>0.76774</td>
</tr>
<tr>
<td>12</td>
<td>0.74673</td>
<td>0.74656</td>
<td>0.74665</td>
</tr>
<tr>
<td>13</td>
<td>0.72553</td>
<td>0.72514</td>
<td>0.72536</td>
</tr>
<tr>
<td>14</td>
<td>0.70464</td>
<td>0.70399</td>
<td>0.70434</td>
</tr>
<tr>
<td>15</td>
<td>0.68410</td>
<td>0.68322</td>
<td>0.68366</td>
</tr>
<tr>
<td>16</td>
<td>0.66359</td>
<td>0.66259</td>
<td>0.66301</td>
</tr>
<tr>
<td>17</td>
<td>0.64362</td>
<td>0.64241</td>
<td>0.64290</td>
</tr>
<tr>
<td>18</td>
<td>0.62396</td>
<td>0.62246</td>
<td>0.62302</td>
</tr>
<tr>
<td>19</td>
<td>0.60424</td>
<td>0.60232</td>
<td>0.60308</td>
</tr>
<tr>
<td>20</td>
<td>0.58382</td>
<td>0.58147</td>
<td>0.58226</td>
</tr>
</tbody>
</table>
The stexpect and SURV3 estimates differ from 12 years since diagnosis on, but always in a very small amount.

Compared with “exact” results the np(#) approximation will be always biased upward. However in this example at the end of follow-up the bias is less than 0.002.
Observed Survival

To compute a ratio between observed and expected survival, the observed survival must be estimated at the same follow-up times specified when stexpect has been used:

```stata
use melanoma, clear
keep if year8594==0 & stage==1
stset surv_mm,f(status==1 2) id(id) scale(12)
stsplit fu, at(0(1)20)
sts gen Osservata = s Hilim = ub(s) Lowlim = lb(s)
```

Confidence intervals for log(-logS(t)) can be used to estimate confidence intervals for the Relative Survival.
Merging Estimates

Only one observation at the end of each follow-up time is kept:

```
bysort _t : keep if _t==fu+1 & _n==1
```

After renaming `_t`, the file with observed estimates can be merged with the file with expected survival at the corresponding times:

```
keep _t Osservata Hilim Lowlim
rename _t t_exp
sort t_exp
merge t_exp using apprmelanhak
rename t_exp time
```
Relative Survival

gen double Relsurv = Osservata / Survexp

Confidence Intervals

replace Hilim = Hilim / Survexp

replace Lowlim = Lowlim / Survexp
The results are tabulated sideways and graphed in the next slide.

<table>
<thead>
<tr>
<th>time</th>
<th>Osservata</th>
<th>Survexp</th>
<th>Relsurv</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.97062058</td>
<td>0.97904406</td>
<td>0.9913962</td>
</tr>
<tr>
<td>2</td>
<td>0.90107875</td>
<td>0.95807621</td>
<td>0.9405084</td>
</tr>
<tr>
<td>3</td>
<td>0.82824544</td>
<td>0.9370337</td>
<td>0.8839014</td>
</tr>
<tr>
<td>4</td>
<td>0.78529246</td>
<td>0.91594289</td>
<td>0.8573596</td>
</tr>
<tr>
<td>5</td>
<td>0.74700829</td>
<td>0.89482277</td>
<td>0.8348114</td>
</tr>
<tr>
<td>6</td>
<td>0.71057591</td>
<td>0.8736134</td>
<td>0.8133757</td>
</tr>
<tr>
<td>7</td>
<td>0.67927507</td>
<td>0.85237852</td>
<td>0.7969171</td>
</tr>
<tr>
<td>8</td>
<td>0.65496515</td>
<td>0.83112553</td>
<td>0.788046</td>
</tr>
<tr>
<td>9</td>
<td>0.62972024</td>
<td>0.80992114</td>
<td>0.7775081</td>
</tr>
<tr>
<td>10</td>
<td>0.6091457</td>
<td>0.78880281</td>
<td>0.7722408</td>
</tr>
<tr>
<td>11</td>
<td>0.58434944</td>
<td>0.76773565</td>
<td>0.7611337</td>
</tr>
<tr>
<td>12</td>
<td>0.56019137</td>
<td>0.74672549</td>
<td>0.7501972</td>
</tr>
<tr>
<td>13</td>
<td>0.54832636</td>
<td>0.7255261</td>
<td>0.7557638</td>
</tr>
<tr>
<td>14</td>
<td>0.53309461</td>
<td>0.70464163</td>
<td>0.7565472</td>
</tr>
<tr>
<td>15</td>
<td>0.51557421</td>
<td>0.6840979</td>
<td>0.7536556</td>
</tr>
<tr>
<td>16</td>
<td>0.49996838</td>
<td>0.66358734</td>
<td>0.7534326</td>
</tr>
<tr>
<td>17</td>
<td>0.48209584</td>
<td>0.64362049</td>
<td>0.7490374</td>
</tr>
<tr>
<td>18</td>
<td>0.47039178</td>
<td>0.62396114</td>
<td>0.75388</td>
</tr>
<tr>
<td>19</td>
<td>0.45598877</td>
<td>0.60424111</td>
<td>0.754647</td>
</tr>
<tr>
<td>20</td>
<td>0.42710142</td>
<td>0.58381723</td>
<td>0.731567</td>
</tr>
</tbody>
</table>
twoway (lowess Relsurv time, clw(medthick) clc(black)) ///
(lowess Hilim time, clc(red) clw(medthick) clp(dash)) ///
(lowess Lowlim time, clw(medthick) clc(red) clp(dash)), ///
xla(0(2)20) yla(0(.2)1) legend(off) tlt("Melanoma Localised") ///
yti("Relative Survival") xti("Years since Diagnosis")
Period Analysis

- Period analysis is a relatively new method proposed by Brenner et al. to derive more up-to-date long-term relative survival estimates better describing the improvements in life expectancy of cancer patients.

- To obtain period survival estimates left truncated observations have to be allowed, i.e. subjects are allowed to enter the observation time after the diagnosis.
Is stexpect compatible with late entry?

• By the `enter` option in the `stset` command it is possible to deal with left truncation (late entry) in survival data.

• Internal codes of stexpect recognize the occurrence of late entry in the data and adapt its computations to this situation.

• Period estimates of relative survival can be achieved as illustrated previously.
strs

- strs is a new Stata command written by Paul Dickman and available at:
  
  http://www.pauldickman.com/rsmodel/stataColon/

- This command estimates expected and relative survival according to the Conditional Method. The applied formula is somewhat different, assuming that data are grouped in time intervals.

- In my checks stexpect,method(2) and strs estimates are very similar.
Conclusions

• stexpect is a new “st” command. It takes advantage of all of the checks and flexibilities stset allows. Its use strictly depends on a timevar suitably generated by the user.

• It does not directly estimate relative survival, but few simple instructions are required to compute it.

• Estimates are consistent (at least until now) with the output of other programs. Only the spreading of stexpect may reveal its limits and contribute to its improvement.
References


