Exercise 5.2 Reweighting nested case-control data

In these exercises, you will calculate the Kaplan-Meier weights by hand for a simple example of a nested case-control sample. You will also run software commands to compute these weights for different kinds of data that may be available in a research study. The first command "**cohort_km_weights**" is used when individual data is available for the outcome and the sampling strata variables for all cohort members from which the nested case-control sample has been taken (or is to be taken): this would be the situation where a researcher selects a nested case-control sample from an electronic cohort that they have available. The second command "**sampled_km_weights**" is used when the researcher has a nested case-control sample but the population counts within the sampling strata are only available from an external source, for example from national statistics offices.

1.
The line-plot on the next page represents the data in "**tiny_cohort.dta**" arranged by follow-up time, with a potential 1:1 NCC sample. The "*" indicates an event and "o" indicates an individual chosen as control (denoted by the indicator variable "chosen" in the data set. *For each event time* (i.e. the grey rows 3,5,6, 9,10,12), complete the following columns for the table on the right of the plot:

p=probability the potential control is sampled at this event time
q= probability the potential control is not sampled at this event time

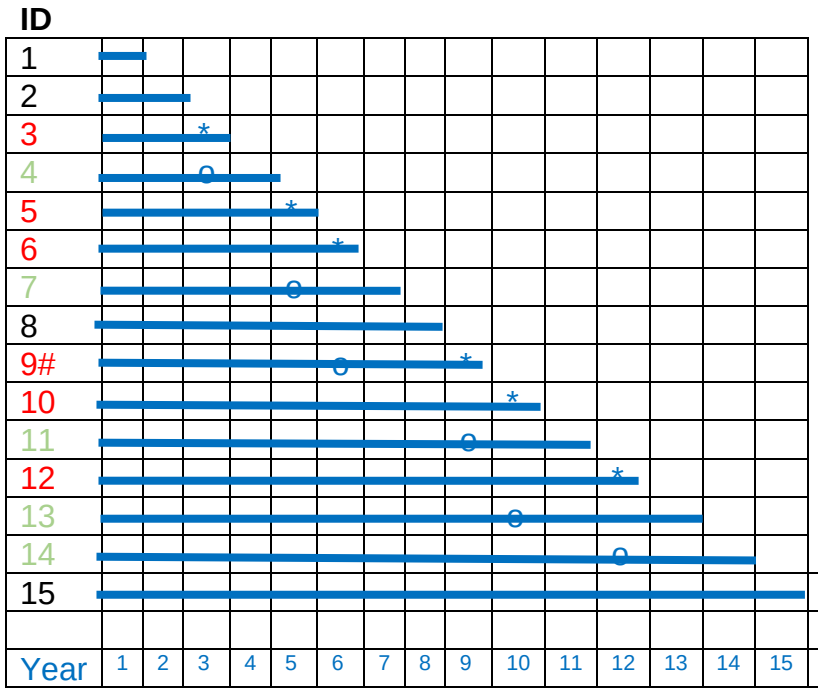
Using these values, complete the following columns for each potential control (i.e the remaining rows of the table):

Q= probability the potential control is not sampled for any event
P*= probability the individual is sampled for the study
Wt=weight

(i) Compare your weights in (i) to the weights obtained from the appropriate software command (*see the hints at the end of this docum*ent).

[Optional for later]
(ii) Repeat steps (i) –(ii) for "**tinycohort2.dta**" which is similar to "tiny_cohort" but has two shared event times (see subsequent page). Note that there are now only 4 event times to be considered, on rows 3, 5/6, 9/10, 12.

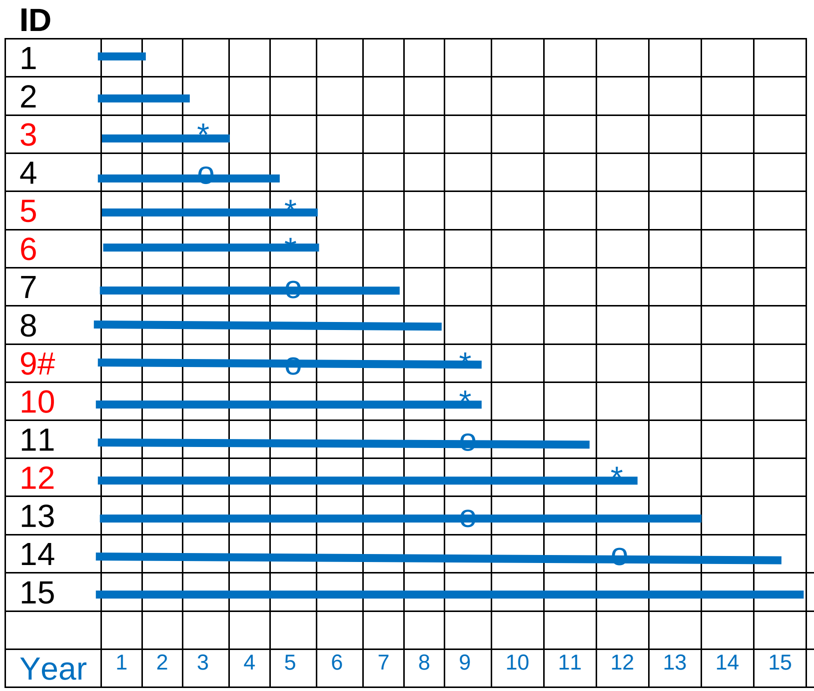
Tiny cohort



p	q	Q	P*	Wt

Note: The "" indicates an event and "o" indicates an individual chosen as control (denoted by the indicator variable "chosen" in the data set*

<u>Tiny cohort 2</u>



p	q	Q	P*	Wt

Note: The "" indicates an event and "o" indicates an individual chosen as control (denoted by the indicator variable "chosen" in the data set*

2.

(i) Use the **tiny_cohort** data set from Question1 to create a file with the risk sets and number of failures using an appropriate survival analysis command and save this as a temporary data set.
(ii) Open "**tiny_cohort**" again and select just the "chosen" NCC individuals as your case-control study data.
(iii) Use the appropriate software command and the external risk set information you saved in (i) to obtain the weights for the individuals in this NCC sample.
(iv) Compare the results from (iii) to what you obtained from the alternative weighting using the skeleton of the cohort in Question 1.

[Optional for later]
(v) Repeat the steps for the data with tied event times ("tiny_cohort2"). *NOTE: the riskset information must only have one observation for each time point.*

3.
The data set **skeleton_cohort.dta** contains "skeleton" information for the cohort of 50,000 individuals that you analysed in Exercise 3.2: just the ID, the time (t, in years) from enrolment to coronary heart disease (CHD), and an event indicator 1=CHD, 0=censored. The data set "**1-2_NCC_sample.dta"** is a 1:2 nested case-control sample drawn from the cohort with all variables recorded for the selected individuals:

id (unique identifier)
age (Age, in years)
gender (1=Male, 0=Female)
Treat (antihypertensive treatment status 1=Yes, 0=No)
Smoke (smoking status 1=Yes, 0=No)
Chol (cholesterol in mg/dL)
HDL (high-density lipoprotein in mg/dL)
SBP (systolic blood pressure in mmHg)

(i) Generate the age category variable **agecat** using categories <=49, 50-59, 60-69, 70-79, 80+ and verify that the OR from conditional logistic regression of the nested case-control data provides similar estimates to the HRs below from the full cohort which were obtained in Exercise 3.2

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
centered C~l	1.006221	.0007014	8.90	0.000	1.004847	1.007596
centered_HDL	.9756035	.001985	-12.14	0.000	.9717207	.9795018
centered_SBP	1.013027	.0010992	11.93	0.000	1.010875	1.015183
Gender	1.667794	.0836694	10.20	0.000	1.51161	1.840116
Treat	1.301126	.0680688	5.03	0.000	1.174326	1.441618
Smoke	1.729781	.0893758	10.61	0.000	1.563185	1.914131
agecat						
50-59	1.253414	.2442627	1.16	0.246	.8554883	1.836432
60-69	2.129328	.4035375	3.99	0.000	1.468683	3.087146
70-79	3.395816	.6522993	6.36	0.000	2.330439	4.94824
80-97	4.509782	.9920104	6.85	0.000	2.930339	6.940538

(ii) Use the nested case-control data to create a data set of *unique* individuals, by breaking the matching and keeping just one record per person: the case record for any case who was previously a control, and just one record for any individual sampled more than once as a control (ensure that the data contains the entry time, and the event or censoring time, of all individuals)

(iii) From the skeleton of the cohort, find the weights for all individuals and merge these to the data from (ii) using the unique ID (Note: the weight should be 1 for cases as all cases were selected).

(iv) Sketch the weighted Kaplan-Meier curve and compare to the Kaplan-Meier curve from the skeleton cohort.

(v) Conduct a weighted Cox regression of the nested case-control data and verify that the estimates are close to the estimates from the full cohort shown in (i) above.

Hints in Stata

The **stset** and **sts generate** commands can be used to get the risk set sizes
and the **cohort_km_weights** and **sampled_km_weigh**ts commands to assign weights

Hints in R
KM weights are calculated by the command **compute_km_weights()** in the
package https://github.com/nyilin/SamplingDesignTools, where the command is illustrated for both settings (using skeleton data from the whole cohort or using external counts)