Who will survive?
Dealing with censored data

Mathematics Alympiad Competition
Preliminary Round
28 november 2003

How to work on this assignment?

The Mathematics Alympiad preliminary assignment 2003/2004 consists of 2 introductory problems and a final task.

General hints

- First read through the whole assignment, in order to know what work needs to be done.
- Do not spend too long on the introductory problems. Assign tasks as possible.
- Before starting the work on the final problem as a team, share and discuss the results of the work on the introductory problems.
- For the final task it is important that you clarify and justify the advices you give.
- Make sure the result is a real advice, including every aspect that is asked for. Make sure that all ideas and details from the introductory problems are incorporated in the advice of the finale assignment, one way or another.

What you need to hand in!

A complete advice – product of the final assignment – legible and understandable for the 'client', independent from the introductory problems. The solutions to the introductory problems can be added as appendices.

Use black ink or print out your answers: they need to be photocopied!

Assessment

In assessing the product, the following aspects will be taken into account:

- the clarity of the described advice, the reasoning and the detailed example
- Completeness (is everything included?);
- the use of mathematics;
- the quality of the line of reasoning and the foundations of the choices that have been made (whether solutions have a sense of reality can also be an aspect);
- the profundity of the work
- the legibility, the structure, the lay out, the form (is it really an advice?) etc

Enjoy and good luck

Who will survive?

Arno Smit is a medical specialist who is specialised in working with heroin addicts. He is studying the use of a chemical substitute for heroin. He wants to examine the effect of this substitute on the life expectancy of the drug addicts. During a period of 10 years, he is monitoring a group of heavily addicted junkies who are using the chemical substitute.The group he is monitoring does not consist of a constant number of people during those ten years: some of the addicts pass away, some of them just don't show up anymore, others move, or they drop out for other reasons.

Fig. 1: Bus supplying free methadon to heroin addicts

In medical science, the effectivity of a treatment is indicated by the *survival probability* after a certain period of time.
In the table below, data from dr. Smit's study are presented.

Year since start of study	Number of living addicts participating in the study (at the start of the year)	Number of deceased addicts participating in the study (during the year)	Number of addicts that dropped out of the study (during the year)
[0,1> = the first year	138	9	14
[1,2>	115	11	10
[2,3>	94	15	5
[3,4>	74	10	7
[4,5>	57	9	9
[5,6>	39	7	11
[6,7>	21	3	5
[7,8>	13	1	7
[8,9>	5	2	1
[9,10>	2	1	1

Because of the premature dropout of addicts from the study, the data are not complete. In statistics we call this incomplete set of data "*censored data*". The group of people we lose track of, we call dropouts.
Notwithstanding the fact that he has to deal with censored data, doctor Smit still wants to be able to produce reliable predictions on the percentage of addicts that will survive five years.

Problem 1

Many specialists use the following method:

> *"Disregard the number of people that drop out of the study (for another reason than decease), up to and including the year you are examining. Now, calculate the percentage that survived this period of time. In this case, the survival possibility after 5 years is 42%."*

There are many more ways to calculate for the survival probability. There is also a method that states:

> *"You have to disregard <u>all</u> dropouts from the data. Using the remaining data, you determine the percentage that died within 5 years. The complementary probability is the survival probability after 5 years and this will be 21% in this case."*

Come up with at least 2 other methods that could be used. Calculate for the corresponding survival probabilities and include your solutions. Also indicate which method seems most reliable to you, and why.

Problem 2

To get more grip on the reliability of the methods from Problem 1, we will look at a set of uncensored data, meaning to say: there are no dropouts.

Year since start of study	Number of living patients participating in the study (at the start of the year)	Number of deceased patients participating in the study (during the year)
[0,1>	120	27
[1,2>	93	18
[2,3>	75	21
[3,4>	54	9
[4,5>	45	4
[5,6>	41	5
[6,7>	36	3
[7,8>	33	3
[8,9>	30	2
[9,10>	28	2

a) Calculate for the survival probability after one year of study, after two years, etc. Select an appropriate way to represent your results.

For uncensored data, we know the exact survival probabilities. For censored data, we will *never* know exactly. But: we can estimate. The reliability of the estimation depends on how you handle the number of dropouts.
For the uncensored dataset given, we will investigate the effects on the survival probabilities for censored data, by treating 'dropouts' in different ways.

b) Manipulate the uncensored dataset above to become a censored dataset. Do this in at least 3 different ways. Every alternative censored dataset should start with 120 patients at the start of the study. Devise your own numbers of dropout patients. Explain how and why you choose to do it the way you do.

c) Apply the methods from Problem 1 to each of the three censored datasets, to come up with survival probabilities after 5 years. Compare the results with the results from Problem 2a. Come up with explanations for the differences found.

d) By now, are you able to be more specific on the reliability of the methods in Problem 1? Explain why, or why not.

Final assignment

Doctor Jozien Oosterhof is a gynaecologist in a university hospital. She is studying the fertility of women who have taken the contraceptive pill for a long period of time. She wants to be able to produce reliable predictions on the probability to get pregnant, dependent on the length of time from stopping taking the pill.
She will also encounter patients who will drop out from the study, for whatever reason. She realizes this will influence the results, but she doesn't know exactly how to deal with censored data.

Fig. 2: the pill

Use what you have learned from Problems 1 and 2 to write an elaborate **advice** to dr. Oosterhof. Make sure that all ideas and details from Problems 1 and 2 are somehow incorporated in this final assignment.
Incorporate also at least one detailed example of a dataset of patients Jozien Oosterhof could possibly encounter during the study.
In the advice, consider different ways she could come up with the probability of getting pregnant after a certain period of time. Indicate clearly the advantages and the limitations of the methods used.

See to it that this advice can be read and understood independent from the assignment!