Exercise 2: A statistical process control chart

At the end of this exercise you should be able to:

a. Aggregate data into the format required for a binomial outcome

b. Create a statistical process control chart for a proportion

EpiData Analysis offers a variety of statistical process control (SPC) graphs. In this exercise we will deal with the determination of a proportion that changes over a period of time, and to what extent this variation deviates significantly from the expectation.

Let's assume that you have 1,200 observations during one year in a laboratory. Among these, 10 per cent (120) have positive result. We could determine the standard deviation and a confidence interval around the positive result or go one step further and take the average we expect for one month (12 of 100) with some measurement of uncertainty around this monthly estimate and then chart the actually observed monthly proportion. This would be a correct procedure if the expected monthly denominator is exactly one twelfth of the annual observation. However, this is rarely the case if ever and more likely is the scenario that the denominator varies in each month.

An SPC graph takes these fluctuations in the denominator into account and calculates the uncertainty as a function of the denominator in the element that is of interest (in this example the month). While there are some discussions on what is best to use, it has become customary to use 3 standard deviations as the upper and lower, so-called control limits. In the chart below, the proportion of scanty positive sputum smear microscopy results among all positive results is shown for Mongolia from the large laboratory register study over a one-year period:

Statistical Process Chart for Scanty Positive Results
Among all Positive Results, Mongolia, 2003
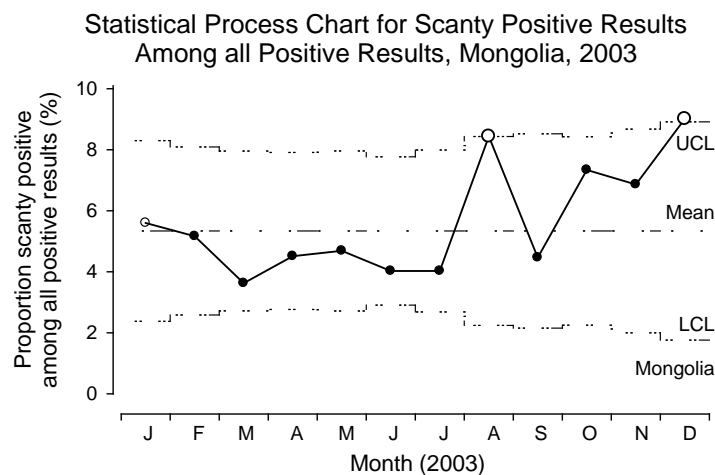
Because the denominator differs in every month, the upper and lower control limits also differ from month to month.

We will be looking at examinations (not at examinees) and determine five different proportions (see below).

The dataset for the exercise

The dataset to be used in this exercise is the cleaned dataset of the four-country laboratory study which was provided in the solution to Part C, Exercise 1, dataset `C_EX01.REC` which is included as a "supplementary required file" with the current exercise with the name `MMUZ.REC`. `MMUZ.REC` is slightly different from `C_EX01.REC` in that the coding for the registration date has been corrected in a few records with an apparent error and the field `REGYEAR` has been removed.

To simplify the task (see specifics at the end), you will limit the analysis to the laboratories in Uganda and to tuberculosis suspects presenting for a diagnostic examination.

The time periods

The Uganda dataset contains information on three years and the unit of measurement of time will be the month. Because each month will thus appear three times (in each year), a new variable must be created that gives a sequential number for each month over the three-year period.

The outcome

The outcome is the monthly proportion of some type of positive results among either all smears or among positive smears.

There are five different proportions we might be interested in: 1) positive smears of any grade among all smears, 2) scanty positive smears among all smears, 3) low-positive smears (defined as either scanty positive or 1+ positive) among all smears, 4) scanty positive smears among all positive smears, and 5) low-positive smears among all positive smears.

Aggregating the data into the correct format

What is thus needed are counts of examinations with the characteristic (a scanty, a low-positive, or any positive result) among all examinations or among all positive examinations. To this end, we aggregate the data as detailed in Part B, Exercise 3.

Making a statistical process control (SPC) chart for proportions over time

The above aggregate file is all we need to get an SPC chart, the general command for which is:

```
pchart numerator denominator [time unit]
```

Proposed procedure in writing the program

It is proposed to split up the program into five distinct sequential procedures:

1) Make the basic dataset

2) Count all smears, all positive, all low-positive smears, and all scanty positive smears

3) Aggregate the months and sum up the smears in question

4) Make the SPC charts

1) Make the basic dataset

In the basic dataset we need to create the years and the months as separate variables and make the appropriate selections:

Month and year of recording: the registers were collected reporting laboratory results during one year up to three years between January 1999 and December 2003. When making a cross-tabulation of country versus registration year (create a field for registration year from REGDATE), we see that:

Year of registration	Country				
	Moldova	Mongolia	Uganda	Zimbabwe	Total
1999	0	0	17300	0	17300
2000	0	0	18662	0	18662
2001	0	0	18088	1213	19301
2002	0	149	0	29307	29456
2003	17725	22406	0	3958	44089
Total	17725	22555	54050	34478	128808

It is thus best to sequentially number all the 60 months from beginning to the end, even if at the end we will require only the three years covered by Uganda.

Select for diagnostic examinations, country, and range of months.

Save the dataset with a new name.

2) Count all smears, all positive, all low-positive smears, and all scanty positive smears

Create four new variables that count for each record respectively all smears, all positive smears, all low-positive smears and all scanty positive smears

3) Aggregate the months and sum up the smears in question and save them to four different files

The element to be aggregated is the month and for each month one must have the relevant smears (all, all positive, all scanty positive, all low-positive).

4) Make the SPC charts

The appropriate chart type for this binomial outcome is a PChart which has the format:

```
pchart numerator denominator time
```

Task

- *Produce five PCharts to display the proportion of 1) positive smears among all smears, 2) low-positive smears among all smears, 3) scanty positive smears among all smears, 4) low-positive smears among all positive smears, and 5) scanty positive smears among all positive smears.*