

Exercise 2: A statistical process control chart

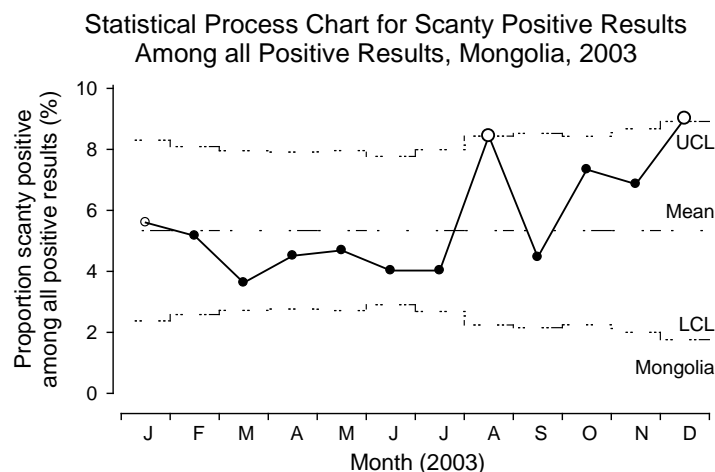
At the end of this exercise you should be able to:

- Aggregate data into the format required for a binomial outcome
- 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:



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. Make sure that you have both the REC and the CHK file in the folder in which you carry out the analysis.

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

You will be determining five different proportions.. A “low-positive” result is defined here as a result that is either scanty positive or 1+ positive. The outcome is the monthly proportion of some kind of positive results among either all smears or among positive smears.

There are thus five different proportions we might be interested in: 1) positive smears of any grade among all smears, 2) scanty positive smear among all smears, 3) low-positive smears among all smears, 4) scanty positive smears among all positive smears, and 5) low-positive smears among all positive smears.

Aggregating the data in the correct format

What is thus needed are counts of examinations with the characteristic (a low-positive result) among all examinations (any examination with at least 1 AFB). You know from tabulations, that tables aggregate individual observations. It is possible to write the same aggregate information that is shown in a table into an EpiData REC file. As an example, you have seven participants in a course from two countries, two of whom are female, and five are male, the only information collected being on COUNTRY and SEX (dataset d_ex02_example1.rec). Browsing the data file you get:

	sex	country
1	Male	Uganda
2	Male	Uganda
3	Male	Uganda
4	Female	Tanzania
5	Male	Uganda
6	Female	Uganda
7	Male	Tanzania

If you make a table:

tables sex country

you get:

Participant's sex			
Participant's country	Female	Male	Total
Tanzania	1	1	2
Uganda	1	4	5
Total	2	5	7

If you replace the command TABLES with AGGREGATE:

```
aggregate sex country
```

you get:

sex	country	N
Female	Tanzania	1
Female	Uganda	1
Male	Tanzania	1
Male	Uganda	4

You can save this in an EpiData REC file with the following options:

```
aggregate sex country /save="temp01.rec" /replace /close  
close  
read "temp01.rec"
```

and then BROWSE to see:

	sex	country	N
1	Female	Tanzania	1
2	Female	Uganda	1
3	Male	Tanzania	1
4	Male	Uganda	4

You now have a REC file with data aggregated for SEX and COUNTRY.

Let's assume now that we have one more field, BANANA, which says how many bananas each individual has in the lunch bag (dataset d_ex02_example2.rec) and browse this dataset, we have:

	sex	country	banana
1	Male	Uganda	2
2	Male	Uganda	3
3	Male	Uganda	1
4	Female	Tanzania	4
5	Male	Uganda	2
6	Female	Uganda	6
7	Male	Tanzania	3

If we ask to aggregate by SEX and COUNTRY, and making a sum of the bananas held by each aggregated group, we would write:

```
aggregate sex country /sum=banana /save="temp01.rec" /replace /close  
close
```

```
read "temp01.rec"
```

and see when browsing:

	sex	country	N	Nbanana	SUMbanana
1	Female	Tanzania	1	1	4
2	Female	Uganda	1	1	6
3	Male	Tanzania	1	1	3
4	Male	Uganda	4	4	8

We noted by browsing the individual dataset that there were 4 Ugandan males with 2, 3, 1, and 2 bananas respectively. With the `/SUM=banana` command option, we sum the bananas for these four Ugandan males and get therefore the 8 bananas in an EpiData Analysis created field `SUMbanana`, which is the total, while `Nbanana` is the count of people with the characteristic (e.g., 4 male Ugandans).

If we modify our questionnaire and ask each of 40 individuals on any of 5 days whether they have a banana with them or not (a binomial outcome which we may code as 0 for not, and 1 for having a banana) and have thus only two variables (in addition to the individual identifier), and where the day sequence of entering the data does not matter, then we would see in browsing the data (only the first 20 individuals, dataset `d_ex02_example3.rec`):

Showing value labels

	dd	ban
1	1	Has no banana
2	1	Has a banana
3	1	Has no banana
4	2	Has a banana
5	3	Has a banana
6	3	Has a banana
7	4	Has a banana
8	4	Has a banana
9	4	Has a banana
10	4	Has no banana
11	5	Has a banana
12	5	Has no banana
13	2	Has a banana
14	3	Has no banana
15	4	Has a banana
16	5	Has a banana
17	1	Has a banana
18	3	Has no banana
19	1	Has a banana
20	3	Has a banana

Showing values

	dd	ban
1	1	0
2	1	1
3	1	0
4	2	1
5	3	1
6	3	1
7	4	1
8	4	1
9	4	1
10	4	0
11	5	1
12	5	0
13	2	1
14	3	0
15	4	1
16	5	1
17	1	1
18	3	0
19	1	1
20	3	1

where `dd` is the field name for the day and `ban` the field name for possession of a banana.

A `tables` command:

```
tables ban dd
```

gives us the number of individuals who have and who do not have a banana on any of these five days:

Having a banana			
Day of observation	Has no banana	Has a banana	Total
1	2	6	8
2	2	3	5
3	6	5	11
4	3	6	9
5	4	3	7
Total	17	23	40

Aggregating the data by day and asking a SUM for the number of “yes” responses to having a banana:

```
aggregate dd /sum=ban /close
```

gives:

dd	N	Nban	SUMban
1	8	8	6
2	5	5	3
3	11	11	5
4	9	9	6
5	7	7	3

Here, `dd` is the day, `N` and `Nban` are identical and are the denominator (the number of individuals asked on that day) and `SUMban` is the count of those who have a banana on that day.

The proportions with a banana on a given day is thus $SUMban/Nban$ (or $SUMban/N$).

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

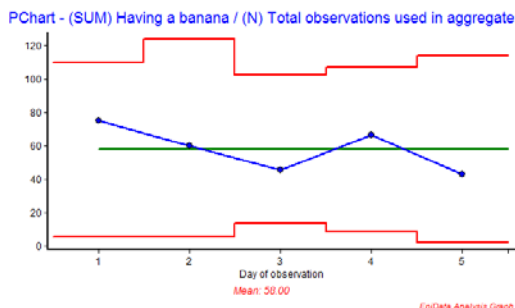
The above aggregate file is all we need to get an SPC chart, the general command for which is:

```
pchart count total [time]
```

or specifically here:

```
aggregate dd /sum=ban /close
pchart sumban N dd
```

and the graphic output is:



The mean proportion here (the green line in the middle) says that 58% of all 40 individuals had a banana, but the upper and lower control limits (the two red lines) differ for each day because the standard deviation varies with the denominator on each day.

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				Total
	Moldova	Mongolia	Uganda	Zimbabwe	
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). The aggregated data are saved into four different files

4) Merge the four files

The four aggregated files are merged (with the month used as the unique identifier).

5) Make the SPC charts

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

pchart numerator denominator time

Up to four have to be produced to get the four proportions of interest on the time axis.

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.***