

## Exercise 4: Aggregating data and saving the summary data in a file

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

- a. Aggregate data using two different approaches, and save summary data from the other approach
- b. Do some manipulations and calculations in a spreadsheet

If aggregate data are collected, there is no way to get back to the individual records. Conversely, it is easy to aggregate data from individual records. This was done in various ways in other exercises (making age groups from age, eg). One can easily find a bit more complex task.

For this exercise you must download the supplementary file `B_EX04_WORKLOAD.REC`, which is an extract from a laboratory dataset (personal communication and data courtesy Biggie Mabaera, Zimbabwe, January 24, 2005).

### First approach

You will have to create a new variable that counts the number of smears done on each examinee. We could then try to make a table of the registration date and the number of smears for each examinee, stratified by laboratory. This would give a very large table (many days in one year). If you try this, you will get an error message that there are too many categories. You will thus have to work with tables where possible and frequencies where needed, making the appropriate selection for the laboratory. This approach does not require anything new, but a somewhat innovative approach to what was learned earlier. The final manipulations are done in a spreadsheet.

### Second approach

You can use the `AGGREGATE` command of EpiData Analysis. This creates the underlying basis for a table which can be written to a new data file.

To explain what `AGGREGATE` does, we include the dataset `banana.rec` which contains just 7 observations with three variables (person's sex, person's country, and the number of bananas each person has). `AGGREGATE` has many things in common with the `FREQ` and `TABLES` commands, but it extends on these in important ways. Like `FREQ` and `TABLES`, `AGGREGATE` does calculations vertically in a dataset. Using the `banana.rec`:

```
freq sex
```

gives:

```
Female    2
Male      5
Total     7
```

and similarly:

```
aggregate sex
```

gives:

```
Female  2
Male    5
```

Note that the result is the same, but the Total is not given with AGGREGATE. If we have here the similarity of AGGREGATE with FREQ, we can look at its similarity with TABLES:

```
TABLES sex country
```

we get:

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

while with:

```
AGGREGATE sex country
```

we get:

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

The similarity is not particularly surprising because all what analysis is about (using FREQ and TABLES) is to actually aggregate observations in some meaningful way.

Where the power of AGGREGATE gets in and leaves FREQ and TABLES behind is when we want to make calculations on the variables of interest. For instance, we could ask “How many bananas do all the women and how many do all the men have together?” To calculate this we use options. For this example, we would write:

```
aggregate sex /sum=banana
```

and get:

```
sex      N      Nbanana  SUMbanana
Female  2      2          10
Male    5      5          11
```

This informs that the 2 women had 10 bananas and the 5 men had 11. Because sex is very unbalanced in the data set, we might wish to know that mean number of bananas that were possessed by each sex and while at it also get the confidence interval for the mean:

```
aggregate sex /mci=banana
```

and get:

```
sex      N      Nbanana  MEAbanana  MEAbananaLOCI  MEAbananaHICI
Female  2      2          5.00        -7.71           17.71
Male    5      5          2.20         1.16            3.24
```

We could aggregate as above by sex and country and sum up the bananas for each of the four strata:

```
aggregate sex country /sum=banana
```

and get:

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

The option `/close` closes the original `*.REC` file and opens the file with the aggregated data. To write the aggregated file into a new file you add the option:

```
aggregate var1 var2 /close /save="newfile1"
```

For our task of investigating the workload in the various laboratories, we probably want to sort on `VAR1` and `VAR2` and write the resulting output into another file that will then be the working file:

```
aggregate var1 var2 /sum=var3 /save="newfile1"
sort var1 var2
savedata "newfile2.rec" /replace
```

Finally, we might want the `NEWFILE2.REC` as a text file. This might be accomplished by going through the export function in EpiData Entry or if the aggregated file is not too large through browsing and copying the content to the clipboard which can be imported or pasted into a spreadsheet for final manipulations and calculations.

### **Tasks:**

- o The `B_EX04_WORKLOAD.REC` has been edited to contain only three laboratories (out of the original 30) and only the year 2002. Nonsensical results (e.g., first examination not recorded, followed by a valid result) have been excluded. Create a program `B_EX04.PGM` to provide you with the necessary information to determine the number of smears performed on average on each day on which at least one examinee was examined for each of the three laboratories.*
- o Use your spreadsheet program to do the necessary final calculations and save it as `B_EX04.XLS`. Summarize the result in a simple table.*