Part C. Operations Research

Part C: Operations research

Introductory note

In this Part C three operationally relevant research questions will be answered:

- Does the dataset tell us something about how diligent the work was performed in the tuberculosis microscopy laboratory?

- Is the third serial smear examination associated with an excessive amount of work for little gain?

- Is it necessary to confirm a positive smear result?

These and related questions were asked by graduates from The Union's operations research courses in fulfillment of the field component of the course. The data were collected in Moldova (Dr Dumitru Laticevschi, fifth course, Paris, 2003), Mongolia (Dr Nymadawaa Naranabat, seventh course, Paris, 2004), Uganda (Dr Achilles Katamba, fifth course, Paris, 2003), and Zimbabwe (Dr Biggie Mabaera, seventh course, Paris, 2004). Six publications have resulted from this study:

Mabaera B, Naranbat N, Dhliwayo P, Rieder H L. Efficiency of serial smear examinations in excluding sputum smear-positive tuberculosis. Int J Tuberc Lung Dis 2006;10:1030-5.

Katamba A, Laticevschi D, Rieder H L. Efficiency of a third serial sputum smear examination in the diagnosis of tuberculosis in Moldova and Uganda. Int J Tuberc Lung Dis 2007;11:659-64.

Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Sputum smear-positive tuberculosis: empiric evidence challenges the need for confirmatory smears. Int J Tuberc Lung Dis 2007;11:959-64.

Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Making pragmatic sense of data in the tuberculosis laboratory register. Int J Tuberc Lung Dis 2008;12:294-300.

Mabaera B, Naranbat N, Katamba A, Laticevschi D, Lauritsen J M, Rieder H L. Seasonal variation among tuberculosis suspects in four countries. International Health 2009;1:53-60.

Rieder H L, Lauritsen J M, Naranbat N, Katamba A, Laticevschi D, Mabaera B. Quantitative differences in sputum smear microscopy results for acid-fast bacilli by age and sex in four countries. Int J Tuberc Lung Dis 2009;13:1393-8.

With permission of the investigators, the datasets have been made publicly accessible for use in this course exactly as they have been collected.

Exercise 1: Creating a working dataset

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

- a. Combine different datasets into one combined dataset
- b. Recode 'text variables' to 'numeric variables'
- c. Remove 'undesirable' records from a dataset
- d. Correct obvious gross errors from the datasets
- e. Create a 'cleaned' final working dataset from available datasets

Moldova and Uganda worked together using the same data entry forms. You obtained `MOL_25.ZIP` and `UGA_30.ZIP`. These two files contain respectively the data files obtained from the 25 laboratories in Moldova and the data files obtained from the 30 laboratories in Uganda. In addition, each of the zip files contains the base pair of `QES` and `CHK` files (which are identical for both countries, except for the field name for the laboratory).

Mongolia and Zimbabwe worked together using the same data entry forms. You obtained `MON_31.ZIP` and `ZIM_23.ZIP`. These two files contain respectively the data files obtained from the 31 laboratories in Mongolia and the data files obtained from the 23 laboratories in Zimbabwe. In addition, each of the zip files contains the base pair of `QES` and `CHK` files (which are identical for both countries).

The two pairs of countries collected exactly the same information from the laboratory register, but their data collection forms (the `QES` files, and thus `REC` files) and `CHK` files had small differences. You can find these by inspecting the respective files. However, as you come in here as an outsider, we summarize these in the following table, and also give you the field names that the final data set combining all files should have.

Field label	Field name Moldova / Uganda	Field name Mongolia / Zimbabwe	Final Field name
Study country	--	--	country
Laboratory code	labcode / labno	laboratory	laboratory
Laboratory serial number	serno	serno	--
Registration date	labdate	regdate	regdate
Year of registration	--	--	regyear
Created unique identifier	unique	id	--
Sex of examinee	sex	sex	sex
Age (in years) of examinee	age	age	age
Reason for examination	reason	reason	reason
Result of first examination	res1	res1	result1
Result of second examination	res2	res2	result2
Result of third examination	res3	res3	result3

Omissions and commissions

In contrast to what you learned in Part A, the data entry form used only field names but had no field labels.

In both studies `SEX` and `REASON` were coded as text fields rather than numerically (but label blocks were used). The fields `RES1`, `RES2`, and `RES3` also differed slightly: a value of 4.0 did not exist in Moldova / Uganda, but denoted "Positive, not quantified" in Mongolia / Zimbabwe, while "Positive, not quantified" was coded as 8.0 in the latter but did not exist in the former. "Scanty, not quantified" was coded as 5.0 in Mongolia / Zimbabwe, but was forgotten as a possible value in Moldova / Uganda.

You could obtain the information from the `CHK` files, but the summary of the coding for the fields of relevance with the differences is as follows:

Field name	Field value Moldova / Uganda	Field value Mongolia / Zimbabwe	Value label
Sex	F	F	Female
	M	M	Male
	9	9	Unknown sex
Reason	D	D	Diagnosis
	F	F	Follow-up, month not stated
	9	9	Reason not stated
	--	1	Follow-up at 1 month
	--	2	Follow-up at 2 months
	--	3	Follow-up at 3 months
	--	4	Follow-up at 4 months
	--	5	Follow-up at 5 months
	--	6	Follow-up at 6 months
	--	7	Follow-up at 7 months
	--	8	Follow-up at 8 months or later
res1 (also res2, res3)	0.0	0.0	Negative
	0.1	0.1	Scanty, 1 AFB / 100 fields
	0.2	0.2	Scanty, 2 AFB / 100 fields
	0.3	0.3	Scanty, 3 AFB / 100 fields
	0.4	0.4	Scanty, 4 AFB / 100 fields
	0.5	0.5	Scanty, 5 AFB / 100 fields
	0.6	0.6	Scanty, 6 AFB / 100 fields
	0.7	0.7	Scanty, 7 AFB / 100 fields
	0.8	0.8	Scanty, 8 AFB / 100 fields
	0.9	0.9	Scanty, 9 AFB / 100 fields
	1.0	1.0	1+ positive
	2.0	2.0	2+ positive
	3.0	3.0	3+ positive
	--	4.0	Positive, not quantified
	--	5.0	Scanty, not quantified
	8.0	--	Positive, not quantified
	9.0	9.0	No result recorded

Tasks:

o *Create a combined dataset C_EX01_COMBINE.REC from all 107 files with a program C_EX01_COMBINE.PGM.*

Notes to the first task:

From the dataset from Moldova, drop the data for the laboratory "BND" (containing data from only 1 week) and remove one empty record.

From the dataset from Mongolia, remove the empty records

In Zimbabwe, one record has no laboratory value, but it has an ID (this is most likely attributable to some manipulation with the mouse after ID creation). You can retain this record by giving the laboratory the correct code that we know from the ID.

If you have removed all empty records (plus the one laboratory from Moldova) and you make a frequency of COUNTRY you should get:

country	N	%
MOL	17865	13.7
MON	22588	17.3
UGA	55114	42.3
ZIM	34744	26.7
Total	130311	100.0

o *Create a "cleaned" final working dataset C_EX01.REC with a program C_EX01_2_RESTRUCTURE.PGM which excludes non-sensically coded result sequences, and with all fields codes numerically (including COUNTRY and LABORATORY).*

Notes to the second task:

For the numeric coding of the COUNTRY follow the alphabet: 1 for Moldova, 2 for Mongolia, …, 4 for Zimbabwe.

For the numeric coding of the laboratories, make a frequency for each country, and then code numerically following the country notation:

Moldova laboratories:

```
if laboratory="ANR" then lab0=101
if laboratory="BLM" then lab0=102
if laboratory="BRL" then lab0=103
if laboratory="BSR" then lab0=104
if laboratory="CCE" then lab0=105
…etc
```

Mongolia laboratories:

```
if laboratory="AR_B" then lab0=201
if laboratory="BG_B" then lab0=202
if laboratory="BN_B" then lab0=203
…etc
```

Uganda laboratories:

```
if trim(laboratory)="1"  then lab0=301
if trim(laboratory)="2"  then lab0=302
if trim(laboratory)="3"  then lab0=303
…etc
```

Zimbabwe laboratories:

```
if laboratory="BY_A" then lab0=401
if laboratory="MC_A" then lab0=402
if laboratory="MC_B" then lab0=403
if laboratory="MC_C" then lab0=404
…etc
```

We also propose to correct some obvious gross errors (which are obvious from the sequence in recording what they should have been) in the registration date. In order to get a common ground, we point these out here and provide the program file commands for these (note that we made a date variable just for this manipulation here):

```
define regyear0 ####
regyear0=year(regdate)
define regyear ####
regyear=regyear0
* correct errors in year of recording
if regyear0=1990 and laboratory=301 then regyear=1999
if regyear0=1990 and laboratory=306 then regyear=1999
if regyear0=1990 and laboratory=319 then regyear=1999
if regyear0=1990 and laboratory=320 then regyear=2000
if regyear0=1990 and laboratory=410 then regyear=2002

if regyear0=2000 and laboratory=408 then regyear=2002
if regyear0=2000 and laboratory=416 then regyear=2002
if regyear0=2000 and laboratory=419 then regyear=2002

if regyear0=2004 and laboratory=211 then regyear=2003
if regyear0=2004 and laboratory=223 then regyear=2003
if regyear0=2004 and laboratory=401 then regyear=2002
if regyear0=2004 and laboratory=408 then regyear=2002
if regyear0=2004 and laboratory=412 then regyear=2003
if regyear0=2004 and laboratory=413 then regyear=2002

if regyear0=2005 and laboratory=207 then regyear=2003
if regyear0=2033 and laboratory=207 then regyear=2003
```

If you have cleaned the dataset and you make a table of COUNTRY by REGYEAR you should get:

Study country					
Year of registration	Moldova	Mongolia	Uganda	Zimbabwe	Total
1999	0	0	17308	0	17308
2000	0	0	18655	0	18655
2001	0	0	18087	1213	19300
2002	0	149	0	29307	29456
2003	17725	22406	0	3958	44089
Total	17725	22555	54050	34478	128808

Note the following on the CHK and QES files:

If you start with a REC file that is accompanied by its CHK file and then create new variables with Field values and Value labels using the LABELVALUE, EpiData Analysis takes the original CHK file and appends it with the new Field values and their Value labels when you create a new REC file. You can also define a Field label (command LABEL newvar "X").

Solution to Exercise 1: Creating a working dataset

Key Learning Points
   a. You should clean the final dataset so as to remove 'undesirable records' and correct obvious gross errors. Records removed from the dataset should be documented as well as the reason.
   b. The 'cleaned' working dataset will then be used for data analysis.

Task:

Create a combined dataset C_EX01_COMBINE.REC from all 107 files with a program C_EX01_1_COMBINE.PGM.

Solution

This is the dataset by country and year that should result from your program:

Year of registration	Country				
	Moldova	Mongolia	Uganda	Zimbabwe	Total
1999	0	0	17308	0	17308
2000	0	0	18655	0	18655
2001	0	0	18087	1213	19300
2002	0	149	0	29307	29456
2003	17725	22406	0	3958	44089
Total	17725	22555	54050	34478	128808

A possible solution is the following `C_EX01_1_COMBINE.PGM`:

```
* Part C, Exercise 1, first program

* Produce combined dataset for
* Moldova, Mongolia, Uganda, Zimbabwe
* and remove empty records

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:     Achilles Katamba, OR Paris 2003
* Zimbabwe:     Biggie Mabaera, OR Paris 2004

* Written by:    Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 03 Jun 2013

cls
close
logclose


*********************************************
* Combine original final Moldova datasets
* Create mol_1.rec

cls
logclose
```

```
close

read   "mol_01.rec"
append /file="mol_02.rec"
append /file="mol_03.rec"
append /file="mol_04.rec"
append /file="mol_05.rec"
append /file="mol_06.rec"
append /file="mol_07.rec"
append /file="mol_08.rec"
append /file="mol_09.rec"
append /file="mol_10.rec"
append /file="mol_11.rec"
append /file="mol_12.rec"
append /file="mol_13.rec"
append /file="mol_14.rec"
append /file="mol_15.rec"
append /file="mol_16.rec"
append /file="mol_17.rec"
append /file="mol_18.rec"
append /file="mol_19.rec"
append /file="mol_20.rec"
append /file="mol_21.rec"
append /file="mol_22.rec"
append /file="mol_23.rec"
append /file="mol_24.rec"
append /file="mol_25.rec"
savedata "mol_0.rec" /replace

cls
logclose
close
read   "mol_0.rec"
define country #
country=1
label country "Study country"
* Exlclude laboratory BND with 13 records
* collected during 1 week only
select labcode<>"BND"
* remove 1 empty record
select serno<>.
var drop unique serno
savedata "mol_1.rec" /replace

close
*******************************************
* Combine original final Mongolia datasets
* Create mon_1.rec

cls
logclose
close

read   "mon_01.rec"
append /file="mon_02.rec"
append /file="mon_03.rec"
append /file="mon_04.rec"
append /file="mon_05.rec"
append /file="mon_06.rec"
append /file="mon_07.rec"
append /file="mon_08.rec"
* Note: 1 record in MON_09.REC had a corrupted
* date which prevented appending.  This record
* was manually changed in EpiData from "203" to "2003"
append /file="mon_09.rec"
append /file="mon_10.rec"
append /file="mon_11.rec"
append /file="mon_12.rec"
```

```
append /file="mon_13.rec"
append /file="mon_14.rec"
append /file="mon_15.rec"
append /file="mon_16.rec"
append /file="mon_17.rec"
append /file="mon_18.rec"
append /file="mon_19.rec"
append /file="mon_20.rec"
append /file="mon_21.rec"
append /file="mon_22.rec"
append /file="mon_23.rec"
append /file="mon_24.rec"
append /file="mon_25.rec"
append /file="mon_26.rec"
append /file="mon_27.rec"
append /file="mon_28.rec"
append /file="mon_29.rec"
append /file="mon_30.rec"
append /file="mon_31.rec"
savedata "mon_0.rec" /replace
close

read "mon_0.rec"
define country #
country=2
label country "Study country"
* The following removes 10 empty records
select serno<>.
savedata "mon_1.rec" /replace

close
*********************************************
* Combine original final Uganda datasets
* Create uga_1.rec

cls
close
logclose

read    "uga_01.rec"
append /file="uga_02.rec"
append /file="uga_03.rec"
append /file="uga_04.rec"
append /file="uga_05.rec"
append /file="uga_06.rec"
append /file="uga_07.rec"
append /file="uga_08.rec"
append /file="uga_09.rec"
append /file="uga_10.rec"
append /file="uga_11.rec"
append /file="uga_12.rec"
append /file="uga_13.rec"
append /file="uga_14.rec"
append /file="uga_15.rec"
append /file="uga_16.rec"
append /file="uga_17.rec"
append /file="uga_18.rec"
append /file="uga_19.rec"
append /file="uga_20.rec"
append /file="uga_21.rec"
append /file="uga_22.rec"
append /file="uga_23.rec"
append /file="uga_24.rec"
append /file="uga_25.rec"
append /file="uga_26.rec"
append /file="uga_27.rec"
append /file="uga_28.rec"
append /file="uga_29.rec"
```

```
append /file="uga_30.rec"
savedata "uga_0.rec" /replace

cls
logclose
close
read  "uga_0.rec"
define country #
let country=3
label country "Study country"
define labcode _____
let labcode=labno
var drop labno serno
savedata "uga_1.rec" /replace

close
*********************************************
* Combine original final Zimbabwe datasets
* Create zim_1.rec

cls
logclose
close

read  "zim_01.rec"
append /file="zim_02.rec"
append /file="zim_03.rec"
append /file="zim_04.rec"
append /file="zim_05.rec"
append /file="zim_06.rec"
append /file="zim_07.rec"
append /file="zim_08.rec"
append /file="zim_09.rec"
append /file="zim_10.rec"
append /file="zim_11.rec"
append /file="zim_12.rec"
append /file="zim_13.rec"
append /file="zim_14.rec"
append /file="zim_15.rec"
append /file="zim_16.rec"
append /file="zim_17.rec"
append /file="zim_18.rec"
append /file="zim_19.rec"
append /file="zim_20.rec"
append /file="zim_21.rec"
append /file="zim_22.rec"
append /file="zim_23.rec"
savedata "zim_0.rec" /replace
close

read "zim_0.rec"
define country #
country=4
label country "Study country"
* Note: if you freq on laboratory then
* you have a lab without a code.  When you sort
* on laboratory, then you see it on the top with
* 4 dots.  Curiously, an ID was created nevertheless
* it is laboratory "MW_L"
* Thus, from the following recoding, we get
* an appropriate laboratory and can retain the record
if ID="MW_L-2002-554" then laboratory="MW_L"
* Laboratory coded as "G867" is actually "ML_L"
* Thus, from the following recoding, we get
* an appropriate laboratory and can retain the record
if laboratory="G867" then laboratory="ML_L"
savedata "zim_1.rec" /replace
```

```
close
***********************************************
* Combine 4 country sets

cls
close
logclose

cls
read "mon_1.rec"
drop serno id result pattern
savedata "montemp.rec" /replace
close

read "mol_1.rec"
define laboratory ____
laboratory=labcode
define regdate <dd/mm/yyyy>
regdate=labdate
drop labcode labdate
savedata "moltemp.rec" /replace
close

read "uga_1.rec"
define laboratory ____
laboratory=labcode
define regdate <dd/mm/yyyy>
regdate=labdate
drop labcode labdate
savedata "ugatemp.rec" /replace
close

cls
read "zim_1.rec"
drop serno id result pattern
savedata "zimtemp.rec" /replace
close

read "moltemp.rec"
append /file="montemp.rec"
append /file="ugatemp.rec"
append /file="zimtemp.rec"
labelvalue country /1="Moldova"
labelvalue country /2="Mongolia"
labelvalue country /3="Uganda"
labelvalue country /4="Zimbabwe"
savedata "c_ex01_combine.rec" /replace
close

read "c_ex01_combine.rec"
freq country

*************************
* Clean up and erase temporary session files

set echo=off
close
define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  cls
  type "Be patient, you will be alerted to completion" /h2
  erase "mon_0.chk  "
  erase "mon_0.rec  "
  erase "mon_1.chk  "
  erase "mon_1.rec  "
  erase "mol_0.chk  "
  erase "mol_0.rec  "
```

```
    erase "mol_1.chk  "
    erase "mol_1.rec  "
    erase "uga_0.chk  "
    erase "uga_0.rec  "
    erase "uga_1.chk  "
    erase "uga_1.rec  "
    erase "zim_0.chk  "
    erase "zim_0.rec  "
    erase "zim_1.chk  "
    erase "zim_1.rec  "
    erase "moltemp.chk"
    erase "moltemp.rec"
    erase "montemp.chk"
    erase "montemp.rec"
    erase "ugatemp.chk"
    erase "ugatemp.rec"
    erase "zimtemp.chk"
    erase "zimtemp.rec"
    cls
    type "All temporary files erased" /h2
  else
    type "All temporary files retained" /h2
endif
set echo=on
```

Task:

o *Create a "cleaned" final working dataset C_EX01.REC with a program C_EX01_2_RESTRUCTURE.PGM which excludes non-sensically coded result sequences, and with all fields codes numerically (including COUNTRY and LABORATORY).*

Solution

A possible solution is the following C_EX01_2_RESTRUCTURE.PGM:

```
* Part C, Exercise 1, second program

* Produce cleaned dataset for
* Moldova, Mongolia, Uganda, Zimbabwe
* Removing results with nonsensical sequence

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia: Nymadawa Naranbat, OR Paris 2004
* Uganda:   Achilles Katamba, OR Paris 2003
* Zimbabwe:  Biggie Mabaera, OR Paris 2004

* Written by:   Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 03 Jun 2013

cls
close
logclose

read "c_ex01_combine.rec"

                define res1b _
          if res1=0 then res1b="N"
if res1>0 and res1<9 then res1b="P"
          if res1=9 then res1b="9"
```

```
                      define res2b _
          if res2=0 then res2b="N"
if res2>0 and res2<9 then res2b="P"
          if res2=9 then res2b="9"

                      define res3b _
          if res3=0 then res3b="N"
if res3>0 and res3<9 then res3b="P"
          if res3=9 then res3b="9"


define sequence _____
label sequence "Sequence of serial results"
let sequence=res1b+"-"+res2b+"-"+res3b

* The following removes records with an impossible
* sequence of results
cls
select sequence<>"9-9-9"
select sequence<>"9-9-N"
select sequence<>"9-9-P"
select sequence<>"9-N-9"
select sequence<>"9-N-N"
select sequence<>"9-P-P"
select sequence<>"N-9-N"
select sequence<>"N-9-P"
select sequence<>"P-9-P"
select sequence<>"9-P-9"
select sequence<>"9-P-N"
select sequence<>"P-9-N"
select sequence<>"9-N-P"

cls
define result1 #.#
label result1 "Result of 1st examination"
let result1=res1
if res1=8.0 then result1=4.0

define result2 #.#
label result2 "Result of 2nd examination"
let result2=res2
if res2=8.0 then result2=4.0

define result3 #.#
label result3 "Result of 3rd examination"
let result3=res3
if res3=8.0 then result3=4.0

cls
define reason0 ##
if reason="D" then reason0=00
if reason="F" then reason0=10
if reason="9" then reason0=99
if reason="1" then reason0=01
if reason="2" then reason0=02
if reason="3" then reason0=03
if reason="4" then reason0=04
if reason="5" then reason0=05
if reason="6" then reason0=06
if reason="7" then reason0=07
if reason="8" then reason0=08

cls
define sex0 #
if sex="F" then sex0=1
if sex="M" then sex0=2
if sex="9" then sex0=9
```

```
cls
define lab0 ###
* Moldova laboratories
if laboratory="ANR" then lab0=101
if laboratory="BLM" then lab0=102
if laboratory="BRL" then lab0=103
if laboratory="BSR" then lab0=104
if laboratory="CCE" then lab0=105
if laboratory="CDR" then lab0=106
if laboratory="CHR" then lab0=107
if laboratory="CLR" then lab0=108
if laboratory="CMN" then lab0=109
if laboratory="CMR" then lab0=110
if laboratory="CNR" then lab0=111
if laboratory="CRR" then lab0=112
if laboratory="CTR" then lab0=113
if laboratory="DNR" then lab0=114
if laboratory="EDR" then lab0=115
if laboratory="FLR" then lab0=116
if laboratory="FRR" then lab0=117
if laboratory="HNR" then lab0=118
if laboratory="LVR" then lab0=119
if laboratory="PRB" then lab0=120
if laboratory="RZR" then lab0=121
if laboratory="SRR" then lab0=122
if laboratory="STR" then lab0=123
if laboratory="VLR" then lab0=124


cls
* Mongolia laboratories
if laboratory="AR_B" then lab0=201
if laboratory="BG_B" then lab0=202
if laboratory="BN_B" then lab0=203
if laboratory="BU_B" then lab0=204
if laboratory="BZ_B" then lab0=205
if laboratory="CH_B" then lab0=206
if laboratory="DA_B" then lab0=207
if laboratory="DD_B" then lab0=208
if laboratory="DG_B" then lab0=209
if laboratory="DU_B" then lab0=210
if laboratory="GA_B" then lab0=211
if laboratory="GS_B" then lab0=212
if laboratory="KE_B" then lab0=213
if laboratory="KH_B" then lab0=214
if laboratory="KO_B" then lab0=215
if laboratory="KU_B" then lab0=216
if laboratory="NA_B" then lab0=217
if laboratory="OR_B" then lab0=218
if laboratory="PR_B" then lab0=219
if laboratory="RE_B" then lab0=220
if laboratory="SB_B" then lab0=221
if laboratory="SK_B" then lab0=222
if laboratory="SU_B" then lab0=223
if laboratory="TU_B" then lab0=224
if laboratory="UM_B" then lab0=225
if laboratory="US_B" then lab0=226
if laboratory="UV_B" then lab0=227
if laboratory="ZA_B" then lab0=228
if laboratory="SE_B" then lab0=229
if laboratory="BK_B" then lab0=230
if laboratory="B-UB" then lab0=231


cls
* Uganda laboratories
if trim(laboratory)="1"  then lab0=301
if trim(laboratory)="2"  then lab0=302
if trim(laboratory)="3"  then lab0=303
if trim(laboratory)="4"  then lab0=304
```

```
if trim(laboratory)="5"  then lab0=305
if trim(laboratory)="6"  then lab0=306
if trim(laboratory)="7"  then lab0=307
if trim(laboratory)="8"  then lab0=308
if trim(laboratory)="9"  then lab0=309
if trim(laboratory)="10" then lab0=310
if trim(laboratory)="11" then lab0=311
if trim(laboratory)="12" then lab0=312
if trim(laboratory)="13" then lab0=313
if trim(laboratory)="14" then lab0=314
if trim(laboratory)="15" then lab0=315
if trim(laboratory)="16" then lab0=316
if trim(laboratory)="17" then lab0=317
if trim(laboratory)="18" then lab0=318
if trim(laboratory)="19" then lab0=319
if trim(laboratory)="20" then lab0=320
if trim(laboratory)="21" then lab0=321
if trim(laboratory)="22" then lab0=322
if trim(laboratory)="23" then lab0=323
if trim(laboratory)="24" then lab0=324
if trim(laboratory)="25" then lab0=325
if trim(laboratory)="26" then lab0=326
if trim(laboratory)="27" then lab0=327
if trim(laboratory)="28" then lab0=328
if trim(laboratory)="29" then lab0=329
if trim(laboratory)="30" then lab0=330

cls
* Zimbabwe laboratories
if laboratory="BY_A" then lab0=401
if laboratory="MC_A" then lab0=402
if laboratory="MC_B" then lab0=403
if laboratory="MC_C" then lab0=404
if laboratory="MC_G" then lab0=405
if laboratory="MC_I" then lab0=406
if laboratory="MC_J" then lab0=407
if laboratory="MD_G" then lab0=408
if laboratory="ME_A" then lab0=409
if laboratory="ME_C" then lab0=410
if laboratory="ME_L" then lab0=411
if laboratory="ME_O" then lab0=412
if laboratory="ML_E" then lab0=413
if laboratory="ML_G" then lab0=414
if laboratory="ML_I" then lab0=415
if laboratory="ML_L" then lab0=416
if laboratory="MN_G" then lab0=417
if laboratory="MV_A" then lab0=418
if laboratory="MV_C" then lab0=419
if laboratory="MV_E" then lab0=420
if laboratory="MW_B" then lab0=421
if laboratory="MW_E" then lab0=422
if laboratory="MW_L" then lab0=423

drop sequence
drop res1 res2 res3
drop reason
drop sex
drop laboratory

rename reason0 to reason
rename sex0 to sex
rename lab0 to laboratory

savedata "temp0.rec" /replace

********************************
cls
close
```

```
read "temp0.rec"

define regyear0 ####
regyear0=year(regdate)

define regyear ####
regyear=regyear0

* correct errors in year of recording
if regyear0=1990 and laboratory=301 then regyear=1999
if regyear0=1990 and laboratory=306 then regyear=1999
if regyear0=1990 and laboratory=319 then regyear=1999
if regyear0=1990 and laboratory=320 then regyear=2000
if regyear0=1990 and laboratory=410 then regyear=2002

if regyear0=2000 and laboratory=408 then regyear=2002
if regyear0=2000 and laboratory=416 then regyear=2002
if regyear0=2000 and laboratory=419 then regyear=2002

if regyear0=2004 and laboratory=211 then regyear=2003
if regyear0=2004 and laboratory=223 then regyear=2003
if regyear0=2004 and laboratory=401 then regyear=2002
if regyear0=2004 and laboratory=408 then regyear=2002
if regyear0=2004 and laboratory=412 then regyear=2003
if regyear0=2004 and laboratory=413 then regyear=2002

if regyear0=2005 and laboratory=207 then regyear=2003
if regyear0=2033 and laboratory=207 then regyear=2003

label regyear "Year of registration"
labelvalue sex /1="Female" /2="Male" /9="Missing"
label sex "Sex of examinee"
labelvalue reason /0="Diagnosis"
labelvalue reason /1="Follow-up at 1 month"
labelvalue reason /2="Follow-up at 2 months"
labelvalue reason /3="Follow-up at 3 months"
labelvalue reason /4="Follow-up at 4 months"
labelvalue reason /5="Follow-up at 5 months"
labelvalue reason /6="Follow-up at 6 months"
labelvalue reason /7="Follow-up at 7 months"
labelvalue reason /8="Follow-up at 8 months or later"
labelvalue reason /10="Follow-up, month not stated"
labelvalue reason /99="Reason not stated"
label reason "Reason for examination"

labelvalue result1 /0.0="Negative"
labelvalue result1 /4.0="Positive, not quantified"
labelvalue result1 /5.0="Scanty, not quantified"
labelvalue result1 /0.1="Scanty, 1 AFB per 100 fields"
labelvalue result1 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result1 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result1 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result1 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result1 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result1 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result1 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result1 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result1 /1.0="1+ positive"
labelvalue result1 /2.0="2+ positive"
labelvalue result1 /3.0="3+ positive"
labelvalue result1 /9.0="No result recorded"
label result1 "Result of 1st examination"

labelvalue result2 /0.0="Negative"
labelvalue result2 /4.0="Positive, not quantified"
labelvalue result2 /5.0="Scanty, not quantified"
labelvalue result2 /0.1="Scanty, 1 AFB per 100 fields"
```

```
labelvalue result2 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result2 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result2 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result2 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result2 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result2 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result2 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result2 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result2 /1.0="1+ positive"
labelvalue result2 /2.0="2+ positive"
labelvalue result2 /3.0="3+ positive"
labelvalue result2 /9.0="No result recorded"
label result2 "Result of 2nd examination"

labelvalue result3 /0.0="Negative"
labelvalue result3 /4.0="Positive, not quantified"
labelvalue result3 /5.0="Scanty, not quantified"
labelvalue result3 /0.1="Scanty, 1 AFB per 100 fields"
labelvalue result3 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result3 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result3 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result3 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result3 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result3 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result3 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result3 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result3 /1.0="1+ positive"
labelvalue result3 /2.0="2+ positive"
labelvalue result3 /3.0="3+ positive"
labelvalue result3 /9.0="No result recorded"
label result3 "Result of 3rd examination"
label regdate "Date of registration"
label laboratory "Laboratory code"
keep country laboratory regdate regyear age sex reason result1 result2 result3
savedata "c_ex01.rec" /replace

********************************
* Test labels, sorting, and count

cls
close
read "c_ex01.rec"

tables country result1 /SLA /vl
tables country regyear

********************************
* Clean up and erase temporary session files

set echo=off
close
define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  cls
  erase "temp0.rec"
  erase "temp0.chk"
  cls
  type "All temporary files erased" /h2
 else
  type "All temporary files retained" /h2
endif
set echo=on
```

Exercise 2: Variability in serial smear results

At the end of this exercise you should be able to:
a. Create a subset of 'suspects' from the working dataset
b. Create a string variable that combines the three results for each examinee
c. Test the given hypothesis on variation in the serial pattern of the results
d. Reject or accept a study hypothesis for each country

The diligence of technicians may suffer if they are over-burdened with work. Decreasing diligence in sputum smear examinations may result in copying a first result.

This exercise examines a bit more closely the variability of serial smear grading among those with at least one positive result (it cannot be ascertained among the majority without any positive result).

In a given laboratory A we might find among suspects the following patterns:

Laboratory A Register

Examinee	Other variables	Res 1	Res 2	Res 3
Examinee 1		1+	1+	1+
Examinee 2		neg	neg	neg
Examinee 3		2+	2+	
Examinee 4		neg	neg	neg
Examinee 5		2+	2+	
Examinee 6		neg	1+	1+
Examinee 7		3+	3+	
Examinee 8		neg	neg	neg
Etc				

In a given laboratory B we might find among suspects the following patterns:

Laboratory B Register

Examinee	Other variables	Res 1	Res 2	Res 3
Examinee 1		1+	neg	1+
Examinee 2		neg		
Examinee 3		2+	1+	
Examinee 4		neg	neg	neg
Examinee 5		2+	3+	
Examinee 6		neg	1+	1+
Examinee 7		3+	1+	2+
Examinee 8		neg		
Etc				

If we compare the patterns found in laboratory A with those in laboratory B, we notice that there is much more variation in laboratory B than in laboratory A. In fact, there is virtually no variation in laboratory A for the series of smears for a given suspect.

The amount of tubercle bacilli is, however, not constant in a series of specimens. Most conspicuously, we see this phenomenon when we compare the number of bacilli found in an early morning specimen with an on-the-spot specimen from the same patient. But even if we took a series of 5 on-the-spot specimens from a patient, e.g., in two-hour intervals (as done in "front-loading"), it is likely that the grading of each of the smears made from these specimens will vary to some extent. This may be because the number of bacilli in the secretions varies and / or because the quality of the produced specimen varies and / or the laboratory technician takes by chance particles that differ in content: fresh sputum is not homogenous.

It is thus highly unlikely that all the results from a given examinee recorded in laboratory A reflect the true content of the series of smears. One becomes suspicious that once the technician in laboratory A found a slide to be positive with grade 2+, the subsequent specimen was not properly examined or perhaps even not examined at all, and the result of the first positive specimen was simply copied into the next column. Such observations can be made in seriously overworked laboratories which are forced to examine three smears until they can declare an examinee not to be a case, and if one specimen is positive, to examine additional specimens until the first positive is confirmed by a second positive smear.

By definition, we cannot examine variation among suspects with a series of three negative smears, which is regrettable because this is precisely the group in which this type of problem is most likely to occur. To assess the quality of examination among negative slides, a system of external quality assessment is required. Nevertheless, the results among suspects with at least one positive result may show the extent of variability between such results that might nevertheless be a useful indicator.

We do not know how much variation there must be to make the results look credible (and even if they vary, the technician could in fact have recorded a fictitious variation). What we can do, however, is to compare the extent of variation between laboratories, or in the data set available here, between the four countries, but we can only assess variations among suspects who are cases in the definition of this course.

In other words, the differences in variation are a crude tool to identify laboratories which pay more and which pay less attention to careful and recommended procedures for the examination of serial smears. This exercise should accomplish this.

Tasks:

Exercise hypothesis:

H_0:   In each study country, at least 60% of cases found among suspects with a complete diagnostic series show a variation in the serial pattern

---

- *Determine with a program C_EX02.PGM the proportions of smears with and without variation in serial smears by country*
- *Interpret the findings*

Solution to Exercise 2: Variability in serial smear results

Key Learning Points

When you have a hypothesis to test, remember that it may be logical to:

    a. Create and use a subset of the working dataset

    b. Create new variable(s)

Tasks:

Exercise hypothesis:

H_0:   In each study country, at least 60% of cases found among suspects with a complete diagnostic series show a variation in the serial pattern

- *Determine with a program C_EX02.PGM the proportions of smears with and without variation in serial smears by country*

- *Interpret the findings*

Solution

Determine with a program C_EX02.PGM the proportions of smears with and without variation in serial smears by country

The following summary output was created:

Crude: Proportion of Grading variation = With variation among all.

variable	stratum	Total N	n$_{Grading\ variation=With\ variation}$	%	(95% CI)
Grading variation	Total	7900	3466	43.9	(42.8-45.0)
Study country	Moldova	870	552	63.4	(60.2-66.6)
	Mongolia	1499	542	36.2	(33.8-38.6)
	Uganda	3465	1608	46.4	(44.8-48.1)
	Zimbabwe	2066	764	37.0	(34.9-39.1)

Crude: Proportion of Grading variation = With variation among all.

Interpret the findings

Conclusion: Except for Moldova, the hypothesis has to be refuted for each country. Of course, there is no accepted standard what constitutes an "acceptable" minimum level of variation that should be found. Nevertheless, it would appear that the level of variation particularly in Mongolia and Zimbabwe is unexpectedly low, that is the serial results raise some questions on the diligence of reading and reporting sputum smear examination results.

The program C_EX02.PGM that produced the above output is the following:

```
* Part C, Exercise 2

* Identifying patterns of serial smear results with identical individual results
```

```
* Objective of the exercise
* Identify series of identical result patterns in the four countries
* The reason for this exercise is that we hypothesize
*   that too regular patterns indicate that the laboratory
*   simply copies a positive result once found to (a) subsequent
*   result(s) rather than properly examining the individual smear
* Thus, this is analysis may be an indirect quality assurance program

* First decision: denominator:
* Define the denominator with the choice of the appropriate dataset
*    Data set must be suspects
*    Assessing variability among persons with only negative results
*       is biased as the proportion of these varies widely, thus excluding
*       such examinees
*    Assessing variability among patients with only two results provides
*       too little insight in variability, selecting thus those with three
*       results of which at least one is positive
*       Furthermore, those with unquantified positive results will also
*       bias the result

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:     Achilles Katamba, OR Paris 2003
* Zimbabwe:    Biggie Mabaera, OR Paris 2004

* Written by:    Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 28 Apr 2013

cls
close
logclose

*********************************************
* Selection process

cls
close

read "c_ex01.rec"

* All records in dataset:
*  128,808 records

* Include only suspects for analysis
select reason=0
* 89,362 records retained

* Select only examinees with three quantified smear results
select result1<4
select result2<4
select result3<4
* 61,064 records retained

define include #
                include=0
if result1>0 then include=1
if result2>0 then include=1
if result3>0 then include=1
select include=1
* 7,900 records retained

savedata "temp_01.rec" /replace

*********************************************
* Variable definition for analysis
```

```
cls
close

read "temp_01.rec"

define variation #
variation=1
if (result1=result2) and (result1=result3) then variation=0
label variation "Grading variation"
labelvalue variation /0="No variation"
labelvalue variation /1="With variation"

*********************************************
* Analysis: Hypothesis testing that
*   at least 60% have variation

cls
ciplot variation country /ng

*********************************************
* Clean up

define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  cls
  close
  erasepng /all /noconfirm
  erase "temp_01.chk"
  erase "temp_01.rec"
  cls
  type "All temporary files erased" /h2
 else
  type "All temporary files retained" /h2
endif
set echo=on
```
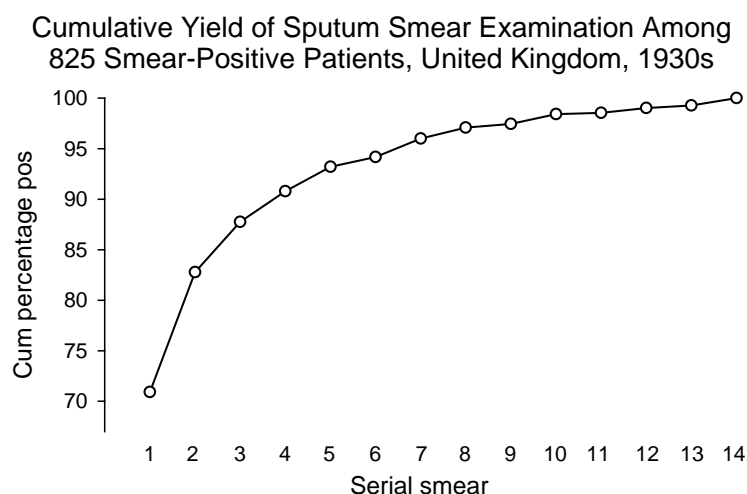
Exercise 3: Incremental yield from serial smears

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

  a. Create a subset of 'suspects' from the working dataset

  b. Create a string variable that combines the three results for each examinee

  c. Make calculations using a spreadsheet

  d. Test the given hypothesis on the incremental yield from the third smear

  e. Reject or accept a study hypothesis for each country

The diminishing return of serial smears is known from studies that have examined multiple serial specimens, as for example the following study from the 1930s:

Cumulative Yield of Sputum Smear Examination Among
825 Smear-Positive Patients, United Kingdom, 1930s

Hunter RA. Tubercle 1940;21:341-59

This study suggests that each serial smear adds an additional increment in case yield, but the incremental yield gets smaller with each additional examination. Program managers must thus arrive at some optimum that requires the least amount of work (number of smear examinations) to yield a large proportion of cases. The "three smear policy" is such a compromise that has been reached internationally and became reflected in the above mentioned guidelines.

The Union and WHO recommended in the past that each suspect should have three sputum smear examinations before being declared to be "sputum smear-negative". Some countries recommended only two examinations. The reason for this difference is that The Union and the WHO thought that making a third examination after two smears are negative would offer a sufficiently rewarding incremental yield (how much is rewarding – has anybody ever defined it?) from this third smear as to justify the additional work load for laboratories. Some microscopy laboratories are, however, so burdened with work (particularly in Africa) that a reduction in the required number of examinations would come as a great relief. It is also very possible that over-burdened laboratories may become less meticulous in the examination of a third smear after a first and second smear have been negative, which may reduce the potential incremental gain. Most of the studies determining the incremental yield from the third examination were done under relatively controlled conditions, but there was not much information around on the yield under routine conditions in low- and middle-income

countries. The primary hypothesis for the operations research study of the course cohorts of 2003 and 2004 was precisely to test the effectiveness under routine conditions from a representative sample of laboratories in four countries. As these were the only studies of this extent and representativeness, the published findings of these studies greatly contributed to the change in policy of WHO in June 2008 to recommend that routine screening of tuberculosis suspects should be limited to two serial examinations to exclude sputum smear-positive tuberculosis. This demonstrates how powerful a relatively simple study design can be in public health policy shaping, if carried out in a representative manner with diligent adherence to quality assurance.

In this exercise, the approach to this issue will be reproduced. We will analyze the data in two parts:

Part 1: incremental yield: we ask what proportion of all cases (being a case in any of up to three serial examinations) is found already on the first smear, which proportion is found on the second smear if the first is negative, and which proportion is found on the third smear after the first two have been negative.

Part 2: program efficiency of the third smear: in addition to the fraction of all ultimately positives, we also take the proportion of cases among all diagnostic examinees into account. The product of these two fractions is the proportion of cases found only with a third examination out of the total number of examinees with a diagnostic examination. We can express the efficiency by taking the reciprocal of this product and obtain the number of smears that need to be examined with a third smear to find one additional case among diagnostic examinees solely when doing a third serial smear (negative on the first two smears).

For *Part 1* we look only at cases and just determine frequencies of "essential patterns" if at least one of the three smears is positive. EpiData provides the possibility to display the point estimate and the 95% confidence interval around the three proportions.

For Part 2 we go a bit more into detail because we also need to know the total of all examinees who came for a diagnostic examination. We do thus not only need the three essential patterns among those positive on at least one of the three, but all six essential patterns. With our case definition that a suspect becomes a case once acid-fast bacilli are found implies that an additional third examination has to be done only if the two preceding examinations have been negative:

NNP / (NNP + NNN)

This is a fraction, but the hypothesis was about the number of smears. In analogy, you may consider the situation where you know that 20 out of 100 people have a characteristic and you now ask how many you have to examine to find the characteristic.

Confidence intervals

Our fraction might be very small despite the large number of suspects in the database. As the implication of refuting the hypothesis has serious programmatic consequences it is advisable to calculate confidence intervals around the number of smears and decide only to refute if the lower interval is in excess of the hypothesis number X.

The classic approach to estimating 95% confidence intervals is used when the population from which the cases arise is defined (observable) and a subset of this population is examined.

We define **P** as the proportion of cases found on the third smear only among those with three examinations:

P　　　=　　(NNP/(NNN+NNP)

The standard error of P [(SE(P)] is calculated from the square root of a function derived from P:

SE(P)　　=　　SQRT(P*(1-P)/(NNN+NNP))

And the 95% confidence intervals are:

$95\%_{low}$　=　P – 1.96*SE(P)

$95\%_{upper}$ =　P + 1.96*SE(P)

However, for the number of slides we will need the reciprocals of these values.

Tasks:

- *Determine with a program C_EX03.PGM the incremental yield of cases with serial smears with the denominator being all cases (Part 1)*

- *Determine in the same program for Part 2 the number of suspects with the six essential patterns listed in the table below. Use the following hypothesis to guide you in this part:*

<div style="border:1px solid">

Exercise hypothesis:

H_0:　Not more than 125 third smear examinations have to be made to find one additional case of tuberculosis in each of the four study countries

</div>

- *Create a table in spreadsheet by country as follows:*

	Moldova	Mongolia	Uganda	Zimbabwe	Total
Total					
Pattern					
N99					
NN9					
NNN					
NNP					
Npx					
Px					
Prop positive					
Yield					
First					
Second					
Third					

X

P
SE(P)
95% low
95% high

Smears
95% low
95% high

Hypothesis:

- ***Interpret the findings***

Solution to Exercise 3: Incremental yield from serial smears

Key Learning Points

When you have a hypothesis to test, remember that it may be logical to:

   a. Create and use a subset of the working dataset
   b. Create new variable(s)

Tasks:

- *Determine with a program C_EX03.PGM the incremental yield of cases with serial smears with the denominator being all cases (Part 1)*

- *Determine in the same program for Part 2 the number of suspects with the six essential patterns listed in the table below. Use the following hypothesis to guide you in this part:*

Exercise hypothesis:

H_0:   Not more than 125 third smear examinations have to be made to find one additional case of tuberculosis in each of the four study countries

- *Create a table in spreadsheet by country as follows:*

	Moldova	Mongolia	Uganda	Zimbabwe	Total
Total					
Pattern					
N99					
NN9					
NNN					
NNP					
Npx					
Px					
Prop positive					
Yield					
First					
Second					
Third					
X					
P					
SE(P)					
95% low					

95% high

Smears
    95% low
    95% high

Hypothesis:

- *Interpret the findings*

Solution:

Part 1: After creating the essential patterns and selecting only the cases among examinees with a diagnostic examination, we created the following output:

Incremental yield

Moldova

Essential patterns			
	N	%	(95% CI)
NNP	34	3.0	(2.2–4.2)
NPx	84	7.4	(6.0–9.1)
Px	1013	89.6	(87.6–91.2)
Total	1131	100.0	

Mongolia

Essential patterns			
	N	%	(95% CI)
NNP	12	0.7	(0.4–1.2)
NPx	42	2.4	(1.8–3.3)
Px	1663	96.9	(95.9–97.6)
Total	1717	100.0	

Uganda

Essential patterns			
	N	%	(95% CI)
NNP	107	1.5	(1.2–1.8)
NPx	487	6.7	(6.1–7.3)
Px	6686	91.8	(91.2–92.4)
Total	7280	100.0	

Zimbabwe

Essential patterns

	N	%	(95% CI)
NNP	155	4.5	(3.9-5.2)
NPx	325	9.4	(8.5-10.4)
Px	2969	86.1	(84.9-87.2)
Total	3449	100.0	

In Part 2, the following output was created in EpiData Analysis:

	country	nnp	npx	px	nnn	nn9	tot	sm95low	smpoint	sm95high	hypothesis
1	Moldova	34	84	1013	8424	1579	12713	186.3	248.8	374.3	Refute
2	Mongolia	12	42	1663	12264	708	15397	653.5	1023.0	2354.0	Refute
3	Uganda	107	487	6686	14736	3325	28666	116.7	138.7	171.0	Accept
4	Zimbabwe	155	325	2969	17740	2706	26601	99.8	115.5	136.9	Accept

Interpretation:

The recorded results show that the number of smears that need to be examined to find one additional case on a third serial smear examination that had not been found already on the first two exceeded 125 (one week's work) in Moldova and Mongolia, indicating the inefficiency of the requirement for three smears before declaring a patient smear-negative at least in these two countries.

We used the following program C_EX03.PGM to get these outputs:

```
* Part C, Exercise 3
* This is b_ex03 EpiData Analysis program
* to determine the incremental yield from serial smears

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:     Achilles Katamba, OR Paris 2003
* Zimbabwe:     Biggie Mabaera, OR Paris 2004

* Written by:    Hans L Rieder
* First version: 12 Feb 2009
* Last revision: 26 Apr 2018

cls
close
logclose

*********************************************
* Prepare data set

cls
close
logclose

read  "c_ex01.rec"

* Definition positive: any AFB in any of three results
* Values: "P" (positive) or "N" (negative)
* or "9" (unknown)

cls
```

```
* Case definition
gen i case=0
if result1>0 and result1<9 then case=1
if result2>0 and result2<9 then case=1
if result3>0 and result3<9 then case=1
label case "Microscopy-defined case"

cls
* Define essential patterns from
* all possible patterns
gen s(3) pattern="NNN"
if result3=9 then pattern="NN9"
if result2=9 then pattern="N99"
if result3>0 and result3<9 then pattern="NNP"
if result2>0 and result2<9 then pattern="NPx"
if result1>0 and result1<9 then pattern="Px"
label pattern "Essential patterns"

keep pattern reason case country
savedata "temp_01.rec" /replace


*************************************
* Analysis part 1
* Look for yield from 1st, 2nd, 3rd smear by country

cls
close
logclose

read  "temp_01.rec"

cls
set echo=off
select reason=0 and case=1
cls
type "Incremental yield" /h2
type "Moldova" /h2
freq pattern /c /ci if country=1
type "Mongolia" /h2
freq pattern /c /ci if country=2
type "Uganda" /h2
freq pattern /c /ci if country=3
type "Zimbabwe" /h2
freq pattern /c /ci if country=4
select
set echo=on


*************************************
* Analysis part 2
* Efficiency of the third smear in finding
*  an additional case: number of smears that
*  need to be examined among all diagnostic
*  examinees to find one additional case

cls
close
logclose

read "temp_01.rec"

aggregate pattern country /save="yield.rec" /replace /close

* Note: a more efficient way to do the following will be shown \
*  in Part D
cls
close
read "yield.rec"
select pattern="NNP"
gen i nnp=n
savedata "nnp.rec" /replace

cls
close
read "yield.rec"
```

```
select pattern="NPx"
gen i npx=n
savedata "npx.rec" /replace

cls
close
read "yield.rec"
select pattern="Px"
gen i px=n
savedata "px.rec" /replace

cls
close
read "yield.rec"
select pattern="NNN"
gen i nnn=n
savedata "nnn.rec" /replace

cls
close
read "yield.rec"
select pattern="NN9"
gen i nn9=n
savedata "nn9.rec" /replace

cls
close
read "yield.rec"
select pattern="N99"
gen i n99=n
savedata "n99.rec" /replace

cls
close
read "nnp.rec"
merge country /file="npx.rec"
merge country /file="px.rec"
merge country /file="nnn.rec"
merge country /file="nn9.rec"
merge country /file="n99.rec"

define tot #####
tot=nnp+npx++px+nnn+nn9+n99

define totpos #####
totpos=nnp+npx++px

drop n pattern mergevar
savedata "pattern.rec" /replace

cls
close
read "pattern.rec"

cls
define p #.#######
p=nnp/(nnp+nnn)

define sep #.#######
sep=sqrt(p*(1-p)/(nnp+nnn))

define cilow #.#######
cilow=p-1.96*sep

define cihigh #.#######
cihigh=p+1.96*sep

define smpoint ###.#
smpoint=1/p

define sm95low ###.#
sm95low=1/cihigh

define sm95high ###.#
sm95high=1/cilow
```

```
cls
          define hypothesis  _____
                    hypothesis="Accept"
if sm95low>125 then hypothesis="Refute"

cls
set display databrowser=on
browse country nnp npx px nnn nn9 n99 tot sm95low smpoint sm95high hypothesis
set display databrowser=off


***************************
* Clean up and erase temporary session files

set echo=off
define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  close
  cls
  type "Be patient, you will be alerted to completion" /h2
  erase "n99.chk    "
  erase "n99.rec    "
  erase "nn9.chk    "
  erase "nn9.rec    "
  erase "nnn.chk    "
  erase "nnn.rec    "
  erase "nnp.chk    "
  erase "nnp.rec    "
  erase "npx.chk    "
  erase "npx.rec    "
  erase "pattern.chk"
  erase "pattern.rec"
  erase "px.chk     "
  erase "px.rec     "
  erase "temp_01.chk"
  erase "temp_01.rec"
  erase "yield.chk  "
  erase "yield.rec  "
  cls
  type "All temporary files erased" /h2
 else
  type "All temporary files retained" /h2
endif
set echo=on
```

Exercise 4: Confirmatory results in serial smears

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

  a. Create a subset of 'suspects' from the working dataset, with the required number of examinations to test the hypotheses

  b. Make a distinction between scanty and positive smear results

  c. Create string variables that combines the three results for each examinee

  d. Recode some string variables to numeric variables

  e. Make calculations using a spreadsheet

  f. Test the given hypotheses on confirmatory results in serial smears

  g. Reject or accept a study hypothesis for each country

  h. Interpret your findings

The bacteriological definition by microscopy of a sputum smear-positive tuberculosis case following WHO required that a positive smear examination had to be confirmed by a second positive result.

This study:

Mabaera B, Lauritsen J M, Katamba A, Laticevschi D, Naranbat N, Rieder H L. Sputum smear-positive tuberculosis: empiric evidence challenges the need for confirmatory smears. Int J Tuberc Lung Dis 2007;11:959-64.

contributed to a policy change in WHO recommendations that were decided in June 2007 following the publication of these findings.

In this exercise, the approach to the problem is reproduced.

The dataset provided here allows the determination of how frequent a scanty positive or a positive smear result is actually confirmed in daily practice in these four countries. It allows further to determine how frequent such a confirmation can be made among suspects who actually had a complete set of examinations.

Exercise hypotheses

H_{01}: At least 80 per cent of suspects with at least one scanty or positive smear result have a confirmatory scanty or positive result

H_{02}: At least 90 per cent of suspects with three serial examination among which there is at least one scanty or positive smear result have a confirmatory scanty or positive result in another examination

Tasks:

- *Write a program C_EX04.PGM that determines the proportion of suspects who have a confirmatory examination, making a distinction between scanty and positive smears. Produce a table by country.*

- *Produce a second table in the same program to determine the proportion of suspects who have a confirmatory examination and who had a complete series of smears, making a distinction between scanty and positive smears.*

- *Interpret the findings.*

Solution to Exercise 4: Confirmatory results in serial smears

---

Key Learning Points

When you have a hypothesis to test, remember that it may be logical to:

    a. Create and use a subset of the working dataset

    b. Create new variable(s)

    c. Produce multiple frequencies of results with different selection criteria

    d. Make use of other software applications e.g. a spreadsheet to make calculations.

Exercise hypotheses

H_{01}:   At least 80 per cent of suspects with at least one scanty or positive smear result have a confirmatory scanty or positive result

H_{02}:   At least 90 per cent of suspects with three serial examination among which there is at least one scanty or positive smear result have a confirmatory scanty or positive result in another examination

Tasks:

- *Write a program C_EX04.PGM that determines the proportion of suspects who have a confirmatory examination, making a distinction between scanty and positive smears. Produce a table by country.*

- *Produce a second table in the same program to determine the proportion of suspects who have a confirmatory examination and who had a complete series of smears, making a distinction between scanty and positive smears.*

- *Interpret the findings.*

Solution

Producing the required results requires multiple frequencies with different selection criteria. The program C_EX04.PGM producing these is shown afterwards, followed by a summary table that is best made in a spreadsheet C_EX04.XLS.

Interpretation:

Moldova had the highest frequency of confirmatory results, in fact more than 95 per cent. As suggested in previous exercises, there might be considerable copying of results, thus it is doubtful to what extent the recorded confirmations correspond to actual results. The opposite is the case in Uganda, where fewer than 65 per cent had a confirmatory result (Table 1).

As shown in table 2, the absence of confirmatory results is simply attributable to the fact that once a smear is positive (or scanty), no further examination is being made. If such an examination is being made, then a confirmation was obtained in 90 per cent of more, with the exception of Zimbabwe, where it was just slightly below the critical proportion.

In summary, this exercise showed that confirmatory smears can generally be made, but in some countries, they are simply not sought. The more general question then is whether it is sensible to require such confirmatory smears, particular in the light that the treatment decision is not greatly affected by it, only the surveillance definition.

The program C_EX04.PGM:

```
* Part C, Exercise 4
* This is b_ex04 EpiData Analysis program
* to determine the frequency of confirmatory smears

* Moldova, Mongolia, Uganda, Zimbabwe
* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:     Achilles Katamba, OR Paris 2003
* Zimbabwe:    Biggie Mabaera, OR Paris 2004

* Written by:    Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 29 Apr 2013

**********************************************
* Selection
cls
close
logclose

read "c_ex01.rec"

* Selection criteria:
* - At least 1 smear must be positive
* - Diagnostic examinees only

* Definitions:
* - Case: at least 1 AFB in at least 1 smear
* - Scanty: 1-9 AFB or "Scanty not quantified"
* - Positive: any non-scanty positive result
* - Scanty series: at least 1 smear is scanty
* - Positive series: at least 1 is positive, none is scanty

gen i case=0
if result1>0 and result1<9 then case=1
if result2>0 and result2<9 then case=1
if result3>0 and result3<9 then case=1

select case=1
select reason=0

keep country result1 result2 result3
savedata "temp_01.rec" /replace

**********************************************
* New variable definition
cls
close
logclose

read "temp_01.rec"

* code for scanty results in series
                        define scanty1 <A>
if result1=0            then scanty1="N"
if result1>0  and result1<1 then scanty1="S"
if result1>=1 and result1<5 then scanty1="P"
```

```
if result1=5              then scanty1="S"
if result1=4              then scanty1="P"
if result1=9              then scanty1="9"

                     define scanty2 <A>
if result2=0              then scanty2="N"
if result2>0  and result2<1 then scanty2="S"
if result2>=1 and result2<5 then scanty2="P"
if result2=5              then scanty2="S"
if result2=4              then scanty2="P"
if result2=9              then scanty2="9"

                     define scanty3 <A>
if result3=0              then scanty3="N"
if result3>0  and result3<1 then scanty3="S"
if result3>=1 and result3<5 then scanty3="P"
if result3=5              then scanty3="S"
if result3=4              then scanty3="P"
if result3=9              then scanty3="9"

define scanty ___
scanty=scanty1+scanty2+scanty3

cls
gen i confirm=0
if substr(scanty,1,1)="N" and substr(scanty,2,1)="N" and substr(scanty,3,1)="P" then confirm=1
if substr(scanty,1,1)="N" and substr(scanty,2,1)="N" and substr(scanty,3,1)="S" then confirm=3
if substr(scanty,1,1)="N" and substr(scanty,2,1)="P" and substr(scanty,3,1)="9" then confirm=1
if substr(scanty,1,1)="N" and substr(scanty,2,1)="P" and substr(scanty,3,1)="N" then confirm=1
if substr(scanty,1,1)="N" and substr(scanty,2,1)="P" and substr(scanty,3,1)="P" then confirm=2
if substr(scanty,1,1)="N" and substr(scanty,2,1)="P" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="N" and substr(scanty,2,1)="S" and substr(scanty,3,1)="9" then confirm=3
if substr(scanty,1,1)="N" and substr(scanty,2,1)="S" and substr(scanty,3,1)="N" then confirm=3
if substr(scanty,1,1)="N" and substr(scanty,2,1)="S" and substr(scanty,3,1)="P" then confirm=4
if substr(scanty,1,1)="N" and substr(scanty,2,1)="S" and substr(scanty,3,1)="S" then confirm=4

cls
if substr(scanty,1,1)="P" and substr(scanty,2,1)="N" and substr(scanty,3,1)="9" then confirm=1
if substr(scanty,1,1)="P" and substr(scanty,2,1)="N" and substr(scanty,3,1)="N" then confirm=1
if substr(scanty,1,1)="P" and substr(scanty,2,1)="N" and substr(scanty,3,1)="P" then confirm=2
if substr(scanty,1,1)="P" and substr(scanty,2,1)="N" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="P" and substr(scanty,3,1)="9" then confirm=2
if substr(scanty,1,1)="P" and substr(scanty,2,1)="P" and substr(scanty,3,1)="N" then confirm=2
if substr(scanty,1,1)="P" and substr(scanty,2,1)="P" and substr(scanty,3,1)="P" then confirm=2
if substr(scanty,1,1)="P" and substr(scanty,2,1)="P" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="S" and substr(scanty,3,1)="9" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="S" and substr(scanty,3,1)="N" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="S" and substr(scanty,3,1)="P" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="S" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="P" and substr(scanty,2,1)="9" and substr(scanty,3,1)="9" then confirm=1

cls
if substr(scanty,1,1)="S" and substr(scanty,2,1)="N" and substr(scanty,3,1)="9" then confirm=3
if substr(scanty,1,1)="S" and substr(scanty,2,1)="N" and substr(scanty,3,1)="N" then confirm=3
if substr(scanty,1,1)="S" and substr(scanty,2,1)="N" and substr(scanty,3,1)="P" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="N" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="P" and substr(scanty,3,1)="9" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="P" and substr(scanty,3,1)="N" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="P" and substr(scanty,3,1)="P" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="P" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="S" and substr(scanty,3,1)="9" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="S" and substr(scanty,3,1)="N" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="S" and substr(scanty,3,1)="P" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="S" and substr(scanty,3,1)="S" then confirm=4
if substr(scanty,1,1)="S" and substr(scanty,2,1)="9" and substr(scanty,3,1)="9" then confirm=3

cls
label confirm "Confirmed by another positive"
labelvalue confirm /0="All negative"
labelvalue confirm /1="Pos not confirmed"
labelvalue confirm /2="Pos confirmed"
labelvalue confirm /3="Scanty not confirmed"
labelvalue confirm /4="Scanty confirmed"

cls
gen i scantpos=0
```

```
* Scanty, not confirmed
if scanty="NNS" then scantpos=1
if scanty="NS9" then scantpos=1
if scanty="NSN" then scantpos=1
if scanty="S99" then scantpos=1
if scanty="SN9" then scantpos=1
if scanty="SNN" then scantpos=1

cls
* Positive not confirmed
if scanty="NNP" then scantpos=2
if scanty="NP9" then scantpos=2
if scanty="NPN" then scantpos=2
if scanty="P99" then scantpos=2
if scanty="PN9" then scantpos=2
if scanty="PNN" then scantpos=2

cls
* Positive, confirmed, no Scanty in series
if scanty="NPP" then scantpos=3
if scanty="PNP" then scantpos=3
if scanty="PP9" then scantpos=3
if scanty="PPN" then scantpos=3
if scanty="PPP" then scantpos=3

cls
* Scanty, confirmed, no Positive in series
if scanty="NSS" then scantpos=4
if scanty="SNS" then scantpos=4
if scanty="SSN" then scantpos=4
if scanty="SS9" then scantpos=4
if scanty="SSS" then scantpos=4

cls
* Scanty-Positive, mixed scanty and positive in series
if scanty="NPS" then scantpos=5
if scanty="NSP" then scantpos=5
if scanty="PNS" then scantpos=5
if scanty="PPS" then scantpos=5
if scanty="PS9" then scantpos=5
if scanty="PSN" then scantpos=5
if scanty="PSP" then scantpos=5
if scanty="PSS" then scantpos=5
if scanty="SNP" then scantpos=5
if scanty="SP9" then scantpos=5
if scanty="SPN" then scantpos=5
if scanty="SPP" then scantpos=5
if scanty="SPS" then scantpos=5
if scanty="SSP" then scantpos=5

cls
label scantpos "Confirmation of smears"
labelvalue scantpos /1="Single Scanty"
labelvalue scantpos /2="Single Positive"
labelvalue scantpos /3="Positive confirmed by Positive"
labelvalue scantpos /4="Scanty confirmed by Scanty"
labelvalue scantpos /5="Scanty confirmed by Positive"

cls
define confres #
if confirm=1 or confirm=3 then confres=0
if confirm=2 or confirm=4 then confres=1
label confres "Confirmed by another positive"
labelvalue confres /0="Not confirmed"
labelvalue confres /1="Confirmed"

savedata "temp_02.rec" /replace

*******************************************
* Output for C_EX04

cls
close

read "temp_02.rec"
```

```
* Table 1. Confirmatory smears among all cases
cls
logclose
set echo=off
logopen "c_ex04_1.txt" /replace
ciplot confres country /ng if confirm<>0
ciplot scantpos country /ng if confirm<>0

title "Confirmation in all countries"
freq confres /c /ci if confirm<>0
freq scantpos /c /ci if confirm<>0

title "Confirmation in Moldova"
freq confres /c /ci if confirm<>0 and country=1
freq scantpos /c /ci if confirm<>0 and country=1

title "Confirmation in Mongolia"
freq confres /c /ci if confirm<>0 and country=2
freq scantpos /c /ci if confirm<>0 and country=2

title "Confirmation in Uganda"
freq confres /c /ci if confirm<>0 and country=3
freq scantpos /c /ci if confirm<>0 and country=3

title "Confirmation in Zimbabwe"
freq confres /c /ci if confirm<>0 and country=4
freq scantpos /c /ci if confirm<>0 and country=4
logclose
set echo=on

* Table 2. Confirmatory smears among all cases with three examinations
cls
logclose
logopen "c_ex04_2.txt" /replace
select
select confirm<>0
select substr(scanty,2,1)<>"9"
select substr(scanty,3,1)<>"9"

set echo=off
cls

title "Confirmation in all countries"
freq confres /c /ci
freq scantpos /c /ci

title "Confirmation in Moldova"
freq confres /c /ci  if country=1
freq scantpos /c /ci if country=1

title "Confirmation in Mongolia"
freq confres /c /ci  if country=2
freq scantpos /c /ci if country=2

title "Confirmation in Uganda"
freq confres /c /ci  if country=3
freq scantpos /c /ci if country=3

title "Confirmation in Zimbabwe"
freq confres /c /ci  if country=4
freq scantpos /c /ci if country=4
logclose
```

Exercise 4. Table 1. Confirmatory smears among all cases

	Moldova			Mongolia			Uganda			Zimbabwe			Total		
	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)
Total	1,131			1,717			7,280			3,449			13,577		
Not confirmed	151	13.4	(11.5-15.5)	89	5.2	(4.2-6.3)	2,804	38.5	(37.4-39.6)	672	19.5	(18.2-20.8)	3,716	27.4	(26.6-28.1)
Single scanty	27	2.4	(1.6-3.5)	24	1.4	(0.9-2.1)	43	0.6	(0.4-0.8)	98	2.8	(2.3-3.5)	192	1.4	(1.2-1.6)
Single positive	124	11.0	(9.3-12.9)	65	3.8	(3.0-4.8)	2,761	37.9	(36.8-39.0)	574	16.6	(15.4-17.9)	3,524	26.0	(25.2-26.7)
Confirmed	980	86.6	(84.5-88.5)	1,628	94.8	(93.7-95.8)	4,476	61.5	(60.4-62.6)	2,777	80.5	(79.2-81.8)	9,861	72.6	(71.9-73.4)
Positive+positive	843	74.5	(71.9-77.0)	1,502	87.5	(85.8-89.0)	4,342	59.6	(58.5-60.8)	2,563	74.3	(72.8-75.7)	9,250	68.1	(67.3-68.9)
Scanty+scanty	24	2.1	(1.4-3.1)	30	1.7	(1.2-2.5)	23	0.3	(0.2-0.5)	107	3.1	(2.6-3.7)	184	1.4	(1.2-1.6)
Scanty+positive	113	10.0	(8.4-11.9)	96	5.6	(4.6-6.8)	111	1.5	(1.3-1.8)	107	3.1	(2.6-3.7)	427	3.1	(2.9-3.5)

Exercise 4. Table 2. Confirmatory smears among all cases with three examinations

	Moldova			Mongolia			Uganda			Zimbabwe			Total		
	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)	Number	%	(95% CI)
Total	904			1,503			3,778			2,829			9,014		
Not confirmed	92	10.2	(8.4-12.3)	55	3.7	(2.8-4.7)	184	4.9	(4.2-5.6)	358	12.7	(11.5-13.9)	689	7.6	(7.1-8.2)
Single scanty	19	2.1	(1.3-3.3)	19	1.3	(0.8-2.0)	17	0.4	(0.3-0.7)	44	1.6	(1.2-2.1)	99	1.1	(0.9-1.3)
Single positive	73	8.1	(6.5-10.0)	36	2.4	(1.7-3.3)	167	4.4	(3.8-5.1)	314	11.1	(10.0-12.3)	590	6.5	(6.1-7.1)
Confirmed	812	89.8	(87.7-91.6)	1,448	96.3	(95.3-97.2)	3,594	95.1	(94.4-95.8)	2,471	87.3	(86.1-88.5)	8,325	92.4	(91.8-92.9)
Positive+positive	688	76.1	(73.2-78.8)	1,337	89.0	(87.3-90.4)	3,470	91.8	(90.9-92.7)	2,308	81.6	(80.1-83.0)	7,803	86.6	(85.8-87.3)
Scanty+scanty	20	2.2	(1.4-3.4)	21	1.4	(0.9-2.1)	22	0.6	(0.4-0.9)	76	2.7	(2.2-3.3)	139	1.5	(1.3-1.8)
Scanty+positive	104	11.5	(9.6-13.7)	90	6.0	(4.9-7.3)	102	2.7	(2.2-3.3)	87	3.1	(2.5-3.8)	383	4.2	(3.9-4.7)