Exercise: Quality assured quantitative data entry: Basic features of definition, control and verification.
J.Lauritsen, EpiData Association, Ortopedic Dept., Odense University Hospital & University of Southern Denmark.
February 2014. v1.0 . jl@epidata.dk

Content: Principle of definining data structures, entering data with control and comparison of data entry.
Duration: 5 hours including explanation of principles on blackboard. Task 10 will be homework assignment

Background:
For all studies regardless of design (Quantitative or Qualitative) we need to:
  • Collect existing litterature on the topic, study this and clarify the following: How do we understand the topic under study ? What is known and not known ? Which elements are important, how are they related etc. - in short: ***The conceptual understanding of the topic leading to a purpose for our study.***
  • Based on the purpose we decide on overall study design(quantitative/qualitative) and define a sampling frame, a setting. Which time period are we studying, which geographical reference, demographic characteristics such as age and sex, future or past information, specific design and size of the study. All of this leads into a strict description of our study - a protocol.

Understanding all of these aspects of study planning, inluding considerations on selection and information bias is crucial regardless of type of research. This understanding and our ability to inform readers about it sets the limits of possibilities for generalisation of our study, and thereby the scope of influence from our work on the theme in question.

Computer skills
Apart from ability to define precise research principles and conceptual models you also need to know basics of defining and handling empirical data. The following exercises will train your basic skills in creating and working with empirical quantitative data.

Preparation (if you have not already done so).
1. Install on you computer: EpiData Manager, EpiData EntryClient[1]
2. If you can run windows software on your pc, then install as well EpiData Analysis.
3. If you have access to Stata software, then install as well Stata

Task 1: **Work by yourself and create the following data structure.**

Variables in the study, which will be fields during data entry:
  Two basic variables: ***Sex (1:male, 2:female), Age***
  Text variable: ***The first and last name of a person you know***
  Date variable: ***Date of today***
  Identifying variable: ***Id (must be autoincrement, look under ..)***

Start Manager and create a dataform with fields for the variables mentioned.
A.Define (new) project.
B.Choose variable type by point and click on the toolbar
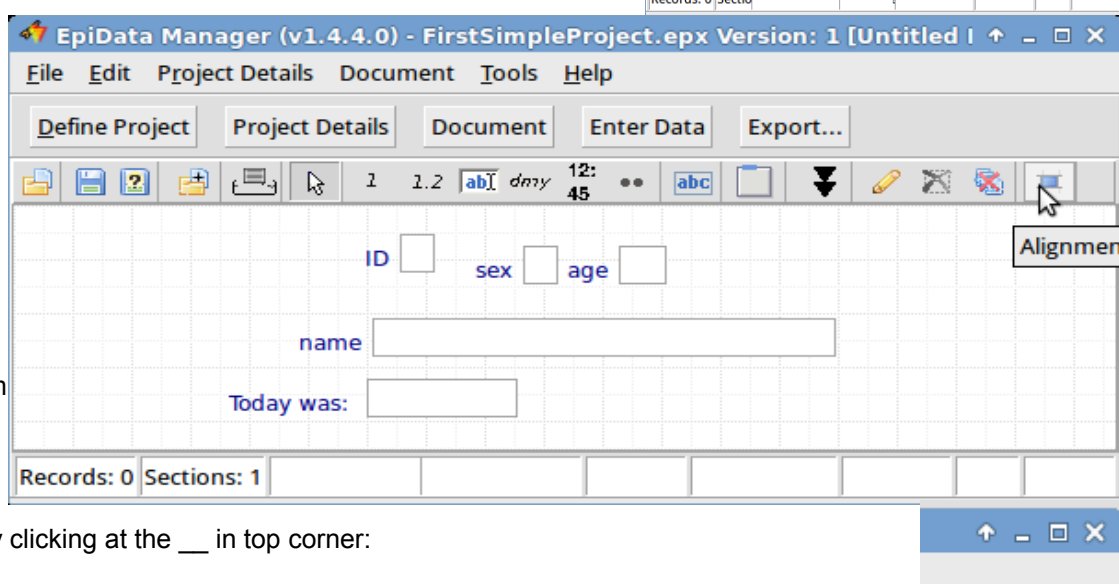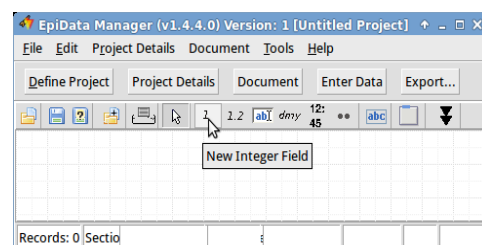**C.**Add next variable
D.Create an empty folder and save the project file there. "save as ...."
   (file menu)
   Your screen could look similar to this:
E. Notice ualigned top 3 fields. You can align them with the alignment tool, see arrow on right.
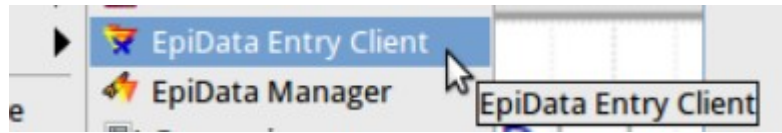**F.** When you like the alignment "save again".
   Minimise the Manager by clicking at the __ in top corner:

1   Software can be retrieved from: Http://www.epidata.dk/download.php

Task 2: Start EpiData EntryClient

Start it as you would always start software on
      your Computer.

Once you have started the Entry Client,
      then open the file you saved above from "recent files" :

Notice that you are warned that the file is already in use,
      since you did not close it in Manager
      - only minimised manager with
      the file.
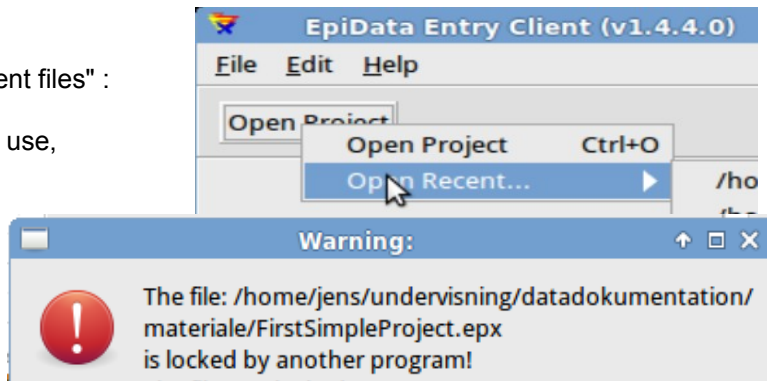
You are warned since it does not make sense to
      enter data and define structure at the same
   time. The remedy is to:

    a) Answer "NO"
    b) Shift to EpiData Manager and close the file in Manager.
    c) Re-open EntryClient and then find the file in "recent files".
    d) Could you now open without the warning ?   _____ (answer)
    e) Now you can enter five records (just make up some numbers).
         Notice how the counter in the bottom left corner changes for every record.
    f) What happens if you hit the + button, while you are in the date field ?   _____
    g) Look at the data you have entered under "Browse Data" in the menu
    h) Which keyboard shortcut will show you the data directly ?   _____
    i) Close the file and Entryclient
    j) Look with your file manager (explorer/my computer/....) in the folder
         where you saved the file. What is the name of the automatic backup file ?   _____
    k) Open on purpose the backup file with EntryClient - which warning is given ?

Task 3: Extend data definition with range and value labels. .

Many different controls can be added to a given data entry
      field, which becomes a variable during analysis. The
      basic principle is to highlight that field and set some
      conditions or attributes to it. Here we exemplify this with
      range and value labels. A value label is a text
      description of a value entered. In the value label you
      indicate which value is used for "missing" or "irrelevant".

Start EpiData Manager again and re-open the project.
    1. Highlight "sex" with the mouse
    2. Hit "enter" which will show "Field Properties".
    3. Notice the "Edit" to the right of "Valuelabel". Click
        on that and a small editor for value labels will
        show up. Add the values 1, 2 and 9, which is
        marked as missing. Press "ok".
        This will close the "Valuelabel Editor".
    4. Notice at the top of the "Field Property" form you
        see tab pages "Basic Extended Jumps etc". On
        the "Extended" tab set the "show value label set
        as note" for "sex". Press "Apply".
    5. For the field "age" define age range to 10-45.
    6. Close the file in Manager and open the project in
        EntryClient and add more records. Can you now
        enter the value 3 in sex ?
   7. Export the data to another system, e.g. Stata or DDI.
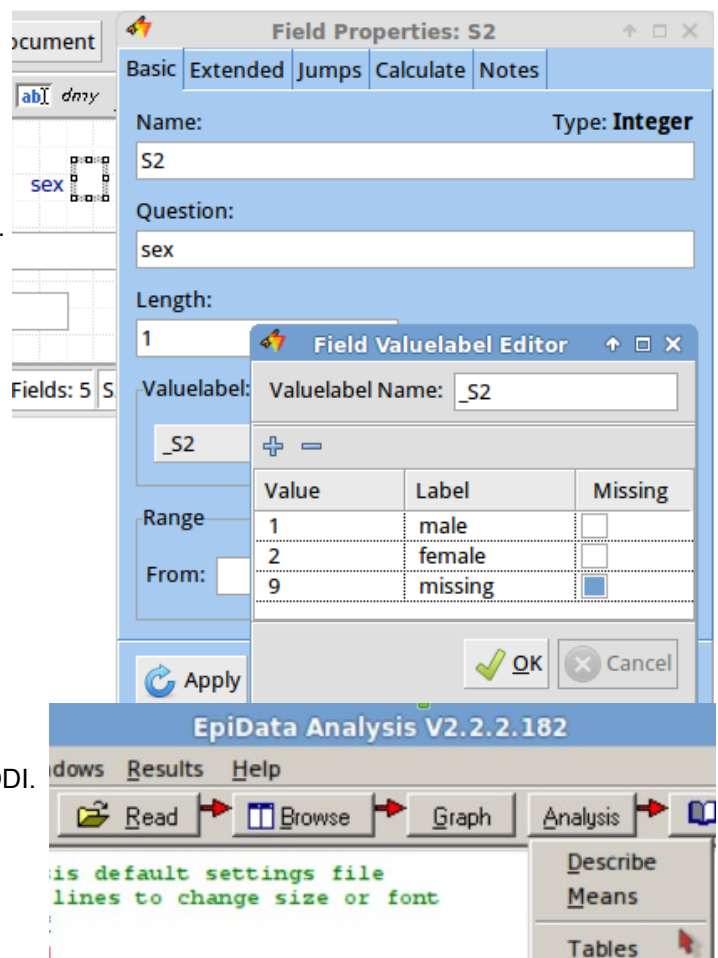
Task 4: First Quantitative Analysis

Open EpiData Analysis, see toolbar on right. With "Read",
followed by "Analysis"-"Describe"-"Tables" find mean of "age"
and a frequency tabulation of "Sex".

Task 5: Prepare NEW project with more variables.

In task 1-4 you have completed a very basic data definition and analysis with the questions mentioned. All other projects work the same way, but admittedly with a much more complex data structure, more controls of entry and more complicated statistical analysis.
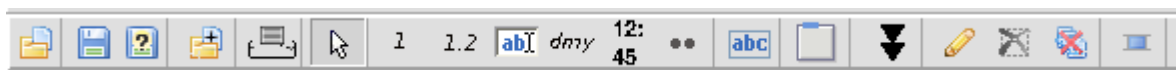
a. Start Manager and create a **new** project , followed by "Importing" structure from the **FirstSimpleProject** created in task 1-3. **Do not** import data from your first project.
   Which of the icons in the toolbar can do import ? Which shortcut will activate import ?

b. Add fields to accomplish this **final structure: (**Some of the lines will require more than one field !)
   0   One descriptive "**header**" at the top e.g. "Second Project exercise NN " (replace NN with your **name**)
   1   Identity variable, which must be integer length 6 digits and name **ID** !! **(not autoincrement)**
   2   Text variable: *The first and last name of the respondent*
   3   Date fields to calculate age: *Date of birth, Date of Today, Age*
   4   Four basic variables: *Sex (male,female)*
   5   Sleep duration: *Time went to bed, Time Got up this morning*, *Duration of sleep*
   6   Number of :   *Children*
   7   Interrupted sleep: *Count of sleep interruptions last saturday night (saturday-sunday)*
   *8   Header: **Backup and data issues:***
   *8a   **How much time are you willing to spend on making your computer work again***
       after a serious breakdown ? (one hour, one day, a few days, a week, a month, longer).

   8b   It turns out that your office computer or laptop actually broke down during last coffee break, therefore:
               How long time will it take you to buy a new computer, install software, restore files
               from backup and get ready to continue work on a replacement computer ?
                           (1-2 hours, about one day, less than a week, 1-4 weeks, Longer, Cannot my data are lost)
   9   Save the file as: **SecondProjectXX** where **xx** is replaced with your initials.

   10 Test that the file works as intended in EntryClient. **ONLY** enter **ONE** record with your own answers.

Task 6: Work in groups on the following:

a. Meet in the group and agree on a unique ID number that each of you have. (e.g. groups of 4-5 persons)
       The number must contain <u>six different digits</u>, the digits <u>cannot be sequential </u>(e.g. not 123456 or 765432)
b. Each of you now edits the entered record in your own computer to have the correct ID.
c. Each of you enter your own answers with your ID on the computer for all other group members, such that by the end all of your computers have observations from all group members.
d. Share the files, such that each of you have all **SecondProjectXX** from the group

Task 7: Compare content of the files:

Three evaluations of data content must now be made - use Manager Menu "Document":
Questions:
       a. Is your dataset following all definitions of your project ? "Data Content Validation"
       b. How many times occurs each ID in each of the XX files ? Use: "Count by ID" on all files in one run.
       c. Did you give the same answer on all computers ?   "Compare/validate Duplicate Files"
               Compare your XX project file to all other xx projects (one other at a time).
       c. Did you enter the same id number on all computers.

Task 8: Add Study Information and consider version of exported data

Extend the definition of the project with Study Information. In practice this forms a short protocol describing the study. In particular if you archive your data or share btw various collaborators this can be an efficient way of safeguarding against misunderstandings. Study Information is based on the "Dublin Core" standard.

       a. Find an extended explanation of "Dublin Core" elements on the Internet?
       b. Was "Dublin Core" elements included when you created a "Data Content Validation" report (task 7a)
       c. Export the data before and after one change to datastructure and entering one record in addition to previous content.Compare the export reports and answer this question together with the person sitting next to you:
               Which indications of version can we find in EpiData Project files ?

Task 9: Setup and control of Manager and EntryClient:

The ruling design choice of EpiData Software is to define reasonable default function and choices based on "good data quality" enhancement principles, but also to let the user be in control. That is the user can adapt the as well Manager and EntryClient according to own preferences.

Behaviour of the software is defined in preferences. Some preferences for e.g. screen font will follow the setup on a given computer, whereas other aspects are controlled via the project settings, which can only be changed by the project manager, since we wish to have the same restriction/control implemented on all data entered, regardless of who entered the data. Therefore project properties can only be changed when the project is open in EpiData Manager. Whereas if a specific user needs a larger screen font would not change the content of the data.

Therfore some settings are at the software level "Preferences" , whereas others are at the project level.

Project Property Specification
Open your project again and find the answer to these questions:

a. Can you decide to show or hide the names of the Fields on the screen ?
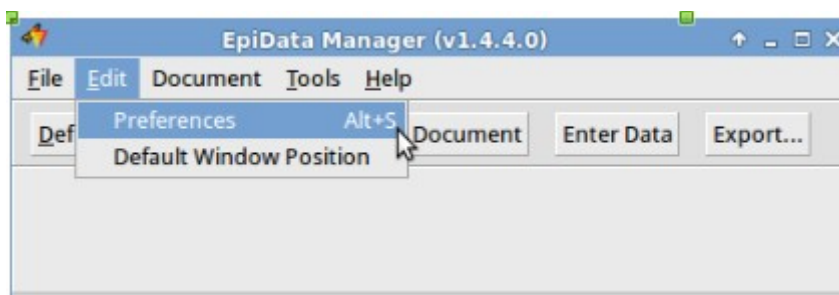b. Can you change the automatic backup interval ?
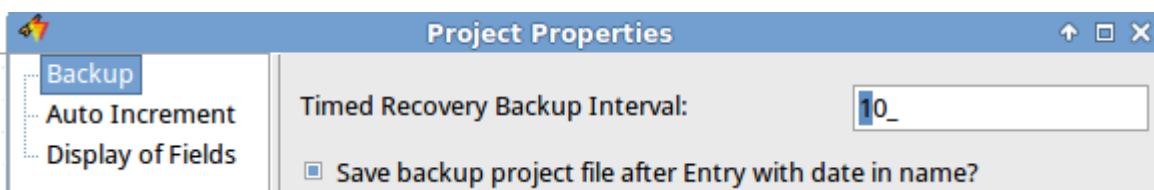
Preference setting
If you look in the menu of Manager and EntryClient under "Edit" you will see "Preferences" Here you can define a lot of parameters according to your own needs.

The figure shows EntryClient preferences:

a. Change some colours, backup folder or intervals.

b. Verify that it works after closing the software.

Position of software on screen:
Experiment with changing the position and size of Manager and EntryClient on your screen.

This will always be where you closed it last time. And it works for all forms as well, e.g. field properties.

c. Change positions, close the software and reopen to verify this.

d. Verify that clicking on "Default Window Position" restores all positions.
This can for example be helpful if you work on two screens at your office from your laptop and then suddenly a view is lost when you only use the laptop.
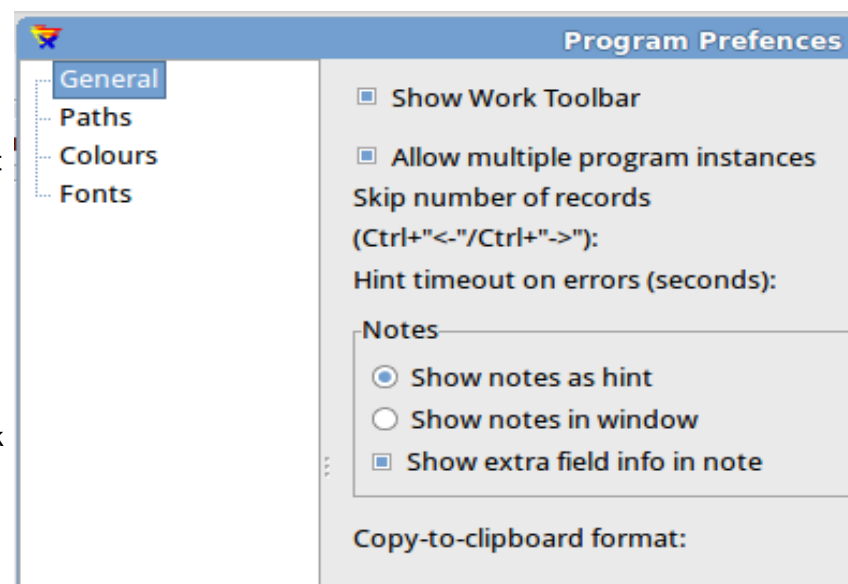
Task 10: Extended definitions of control and add study information:

Extend the definition of the fields to accomplish the following and verify in Entryclient when relevant, e.g. calculations, unique entry or jumps.

a. Calculate: age based on (today and date of birth) Notice the "tab page" "Calculate"

b. Insert jump such that "interrupted sleep" will be bypassed for those having no children.

c. Define field "id" a unique index key.

d. Encrypt the whole project with a Password.

e. What happens during Dataentry if you extend the definition for **sex** to "write label to string field" **name** ?

f. Calculate "Sleep duration" (time to bed and time to get up) Is there an extended challenge here in comparison to task 10a ?