# Health data collection and analysis

## Biostatistics, Epidemiology, Health Informatics

## Practical summary

## Student note

Edited by: Dr. habil. Attila Csaba Nagy, György Kirilla

Author: Dr. habil. Attila Csaba Nagy

2023

UNIVERSITY OF DEBRECEN

FACULTY OF HEALTH SCIENCES

# Health data collection and analysis

## Biostatistics, Epidemiology, Health Informatics

## Practical summary

## Student note

Edited by:

Dr. habil. Attila Csaba Nagy

Kirilla György

Author:

Dr. habil. Attila Csaba Nagy

Translated by:

Dr. habil. Attila Csaba Nagy

Debreceni Egyetemi Kiadó
Debrecen University Press

Proofread by:

Dr. habil. Móré Marianna

ISBN: 978-963-615-081-5

Published by: Debrecen University Press, member of the Association of Hungarian Book Publishers and Book Distributors founded in 1795

www.dupress.hu
Publisher in charge: Gyöngyi Karácsony

# Table of contents

# Foreword

This note is a practical summary designed to review the entire process of a short questionnaire study from its design to execution to analysis. The summary covers epidemiology, health informatics and biostatistics. The detailed theoretical background of each area is dealt with in separate subjects, but the three areas concerned are organically linked. Practical implementation can help to master the theory more easily, as well as to complete the learned theory. We strive to define the most important concepts according to aspects of practical application. The individual chapters follow each other in a logical order, but they are also intended to cover a specific area on their own. For ease of learning, key information is presented in outline form, and screenshots make it easier to understand/follow, whether it is related to the installation of the programs used or to illustrate a resultant product (output). We go through the installation of the programs, the selection and implementation of the appropriate method, and the interpretation of the results obtained. We close the material with useful links and the literature used.

Dr. habil. Attila Csaba Nagy

*"Non scholae sed vitae discimus."*

Epidemiology is a broader, more general science than just the epidemiology of communicable diseases. It practically covers the description of the characteristics of infectious and non-communicable diseases, the study of factors affecting their formation and course. Different relationships and characteristics are described by so-called epidemiological indicators. Epidemiology is an integral part of the entire spectrum of prevention from primary prevention to tertiary prevention.

Prevention can be divided into three main parts:

- Primary: health maintenance in healthy individuals free of disease (the aim is to maintain existing health), disease prevention, elimination of risk factors (e.g., smoking) e.g., through health education or vaccinations.
- Secondary: screening, the goal is to recognize the already developed, but still asymptomatic disease, to start therapy as soon as possible
- Tertiary: adequate care/care if possible complete rehabilitation

## Frequency measures

Frequency inferences cannot be drawn from absolute numbers because denominators are not available. If there were 10 deaths in County *A* and 20 in County B, it is not possible to compare the level of mortality without knowing the entire county population. In turn, frequency measures are already suitable for drawing conclusions. In this case, we have a benchmark that eliminates the effect of different denominators.

- Prevalence: P=n/N, i.e., dividing cases (disease or condition) by the total population. Synonyms: frequency, occurrence, constant cases. The prevalence of type 2 diabetes in Hungary is around 8% (800,000/1,000,000).

- Incidence: measures the occurrence of new cases, there are two types:
  - Cumulative incidence: CI=n/N, where the main difference from prevalence is that CI refers to a follow-up period, i.e., dividing the number of new cases occurring during that period by the population initially at risk. For example, if 10 out of 100 people get sick in a year, the absolute risk (cumulative incidence) is 10%.
  - Incidence density: ID=n/PT, refers more to a dynamic study population, the number of new cases over a given period, divided by total person time at risk (as long as someone lives, is in the study, is able to get sick, but is not sick). For example, if there will be 5 new cases in 10 person years, then 5/10=0.5 person-year$^{-1}$ is the absolute hazard (incidence density).

## Association measures

Knowing the frequency measures, we can calculate association measures. Association measures can be calculated for groups of exposed (risk/protective) and unexposed individuals.

- Difference measures:
  - Cumulative incidence difference: CID=$CI_1$-$CI_0$ The additional risk expresses the net excess risk due to exposure.
  - Incidence density difference: IDD= $ID_1$-$ID_0$
- Ratio measures:
  - The relative risk formula is RR= $CI_1/CI_0$, where CI1 is the incidence rate in the exposed group and $CI_0$ is the incidence rate in the non-exposed group. The interpretation of the relative risk is that the risk of the exposed group is X times greater than the risk of the non-exposed group (effectively the ratio of the absolute risks of the two groups (exposed/unexposed)). The neutral effect is indicated by a value of one: one time higher, is the same risk in the two groups, meaning it is not a real influencing factor.
  - Relative hazard: RH= $ID_1/ID_0$

## Epidemiological studies

Of the various epidemiological studies (diagnostic, prognostic, etiological), etiological studies are used most often. Etiological studies are used to identify factors (risk or protective) that potentially account for outcomes (disease, death, etc.). They are grouped into:

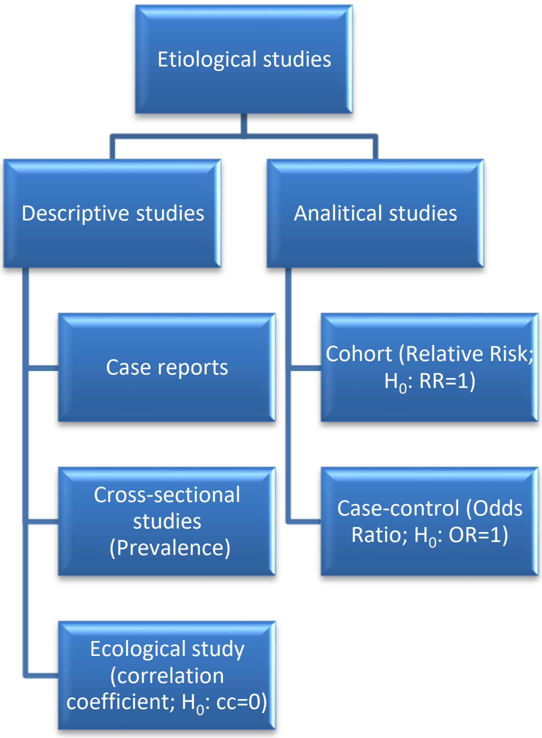

Etiological studies
Descriptive studies
Analitical studies
Case reports
Cross-sectional studies (Prevalence)
Ecological study (correlation coefficient; H$_0$: cc=0)
Cohort (Relative Risk; H$_0$: RR=1)
Case-control (Odds Ratio; H$_0$: OR=1)

Figure 1: Grouping of etiological studies

- Descriptive studies: due to their name, it is possible to describe phenomena, generate hypotheses with their help.
    - Case reports: describes individual/rare cases with a medical history.
    - Cross-sectional study: Prevalence can be measured with their help, such as disease frequencies measured through health surveys.
    - Ecological study: the units of observation are the groups; correlation coefficients can be calculated with their help. Ecological fallacy is a bias, when we try to interpret results at individual level (incorrectly) instead of group level.

- Analytical studies: we can test hypotheses generated by descriptive studies, as well as quantify the strength of the association.
    - Cohort: closed group (according to the Roman origin of the word, which meant closed military group), at the beginning of the study, everyone is healthy, and then during the follow-up, disease develops in both the exposed and non-exposed groups. The computed association measure is relative risk. These types of studies are time- and resource-consuming and therefore not suitable for rare diseases. We usually talk about prospective (forward-looking) studies, but it is possible to reduce the cost through retrospective studies. It can also be classified according to the nature of the sample; it can be dynamic (open cohort), variable (move/new entry), non-variable (closed cohort), or (often classified under the latter) fixed cohort (exposure category does not change either).
    - Case-control: in short, the opposite of the cohort, so it is fast, cheap, and suitable for rare diseases but not for rare exposure testing. For the already sick case group, we randomly select controls (from the same source population to represent them, regardless of their exposure status, since this is exactly what we want to get an idea of) who can catch the disease but are not sick. The association measure that can be calculated is the odds ratio.

## Validity

Validity is an important requirement for results. It means the degree to which the information collected accurately answers the research question. It can be divided into two groups:

- **Internal**: means that the result is valid for study participants; assumption: the study is free of systematic errors.
- **External**: means that the test result is also valid for the entire source population; assumptions: internal validity and representativeness.

Systematic errors can be broken down into 3 groups:

- **Information bias**: the observer (recorder/filler) records incorrect information, either due to recall bias or due to other factors, thus distorting, for example, the inclusion in the exposed or unexposed group.
- **Selection bias: the group selection is flawed, for example, the control group selection is not random,** but rather we put non-smokers in the control group, which thus will not be representative of the source population (for whom we want to draw a conclusion based on the study on my random sample).
- **Confounding factors:** three criteria need to be fulfilled:
  - Independent risk factor of the output variable
  - Related to suspected exposure (which we are investigating)
  - Not in the causal chain

  A fictitious example is a study analyzing the potential association between coffee consumption (exposure) and pancreatic cancer, in which smoking (meeting all three criteria) is a confounding factor. The latter is responsible for the real effect. (Common confounders are gender and age.)

If possible, it is necessary to prevent its occurrence:

- Randomization (a randomly selected sample, this is the best option)
- Restriction (the exclusion of a subgroup, e.g.: we exclude male participants, but we lose information)
- Matching (by age/gender matched sample)

If it has not been prevented, it should be corrected/adjusted for its effect:

- Stratified analysis (stratified along a confounder)
- Regression (faster and better solution, especially for many confounders/strata)

## Characteristics of Screening tests

Secondary prevention is screening. Screening tests have four main characteristics:

- Sensitivity: sensitivity, shows the true positivity rate, i.e., how good our test is at finding patients (a/a+c)
- Specificity: shows the true negative rate, i.e., how good our test is in identifying those free of disease (d/b+d)
- Positive predictive value (PPV): shows how likely an individual identified as positive by a screening test is to be ill (a/a+b)
- Negative predictive value (NPV): shows how likely an individual found negative on a screening test is to be free of the disease. (d/c+d)
- Prevalence: disease occurrence (a+c)/(a+b+c+d)

Table 1: New screening test and proven diagnosis

| | | **Diagnosis** | | |
|---|---|---|---|---|
| | | sick | not sick | |
| **Test** | positive | a | b | a+b |
| | negative | c | d | c+d |
| | | a+c | b+d | a+b+c+d |

The markings in each cell are:

- A: true positive: ill and aware
- B: false positive: not ill, may have false disease consciousness
- C: false negative: ill, but not aware of it, not receiving treatment in time
- D: true negative: not ill and aware

# Data collection

## About questionnaires

Questionnaires are a well-structured method of data collection consisting of a series of questions.

## Design questionnaires

For questionnaire surveys, we can use an existing, validated questionnaire (with the appropriate citation, of course) or we can create a new questionnaire. For newly created questionnaires to be validated, open-ended questions that can be completed in free text are essential. For final questionnaires, closed-ended questions are more common, also for ease of analysis. The order of questions is important, both in relation to and within appropriate sets of questions. Referring to our fictitious example: when exploring risk factors, it is worth starting with coffee, then continuing with smoking, and finally asking questions about alcohol consumption. When assessing these important health behavioral factors, it is easier for the respondent to answer a less sensitive question (coffee) first. In addition to the questions, the answers should follow each other in a logical order. The question number situation is contradictory, as more questions provide more information, but the respondent tends to get tired over time and give less accurate answers. In the past, the optimal number of questions was set at around 20-30, and the theoretical maximum was around 70. For modern online questionnaires, it is better to think about filling time, the optimal length is 10-20 minutes.

## Edit questionnaires

Previously, paper-based questionnaires were printed, nowadays, they are replaced by online questionnaires, which can be placed on your own website, bought from a survey company, or provided free of charge. Among the latter, Google Forms is popular, but it is better to use Microsoft Forms for tighter data management and security.

## Using Microsoft Forms

The following is a screenshot of how to access and use Microsoft Forms.

Microsoft Forms is part of the Microsoft Office/365 suite.

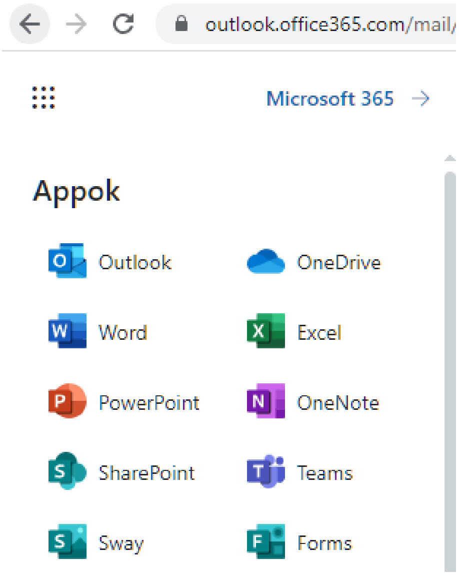


Figure 2: Microsoft Forms icon location

A new quiz or new form can be selected after launch:

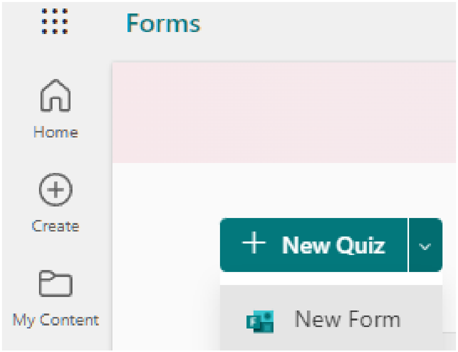


Figure 3: Creating a new form

As a first step, it's a good idea to name your new form/questionnaire:

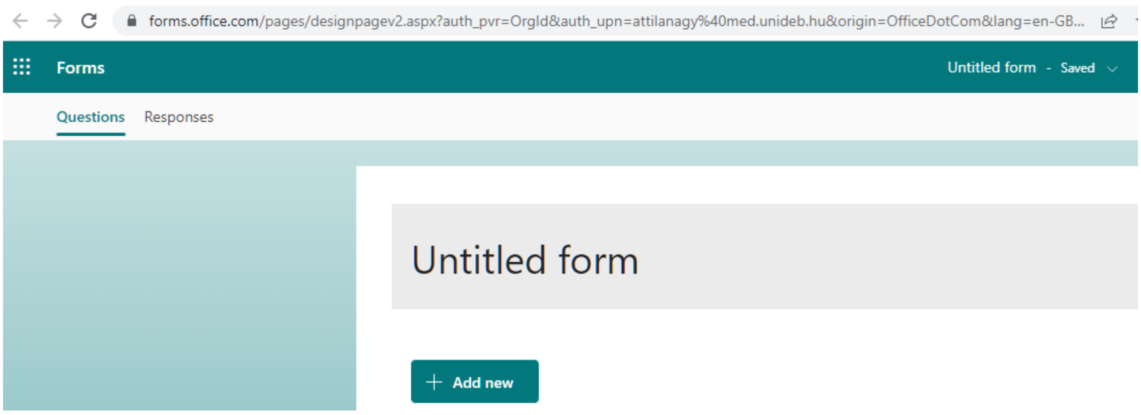

forms.office.com/pages/designpagev2.aspx?auth_pvr=OrgId&auth_upn=attilanagy%40med.unideb.hu&origin=OfficeDotCom&lang=en-GB...
Forms
Untitled form - Saved
Questions
Responses
Untitled form
Add new

Figure 4: Form naming

After naming the questionnaire, we can also provide a brief overview of the survey, after which we can choose from different types of questions:

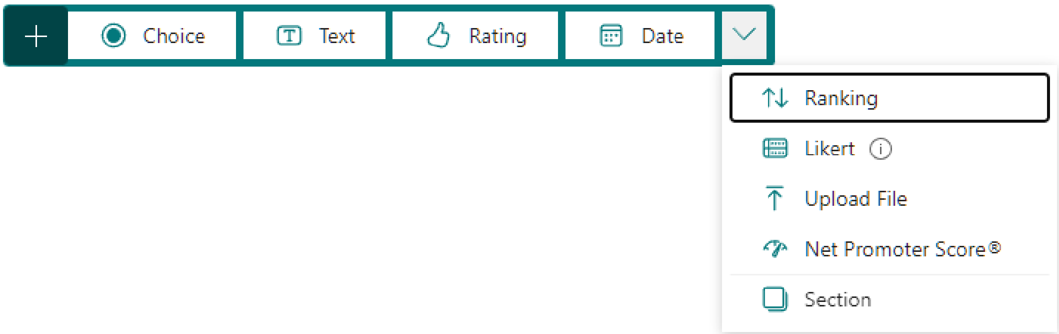
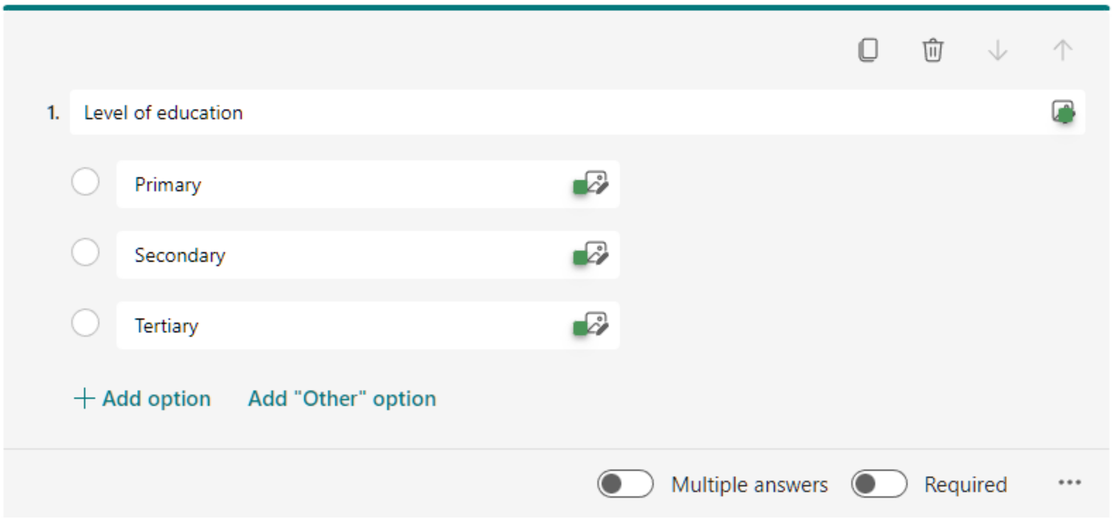


Figure 5: Question types

We can also select single- and multiple-choice questions.Questions can be edited, deleted, or the order can be changed, both in terms of questions and answers:



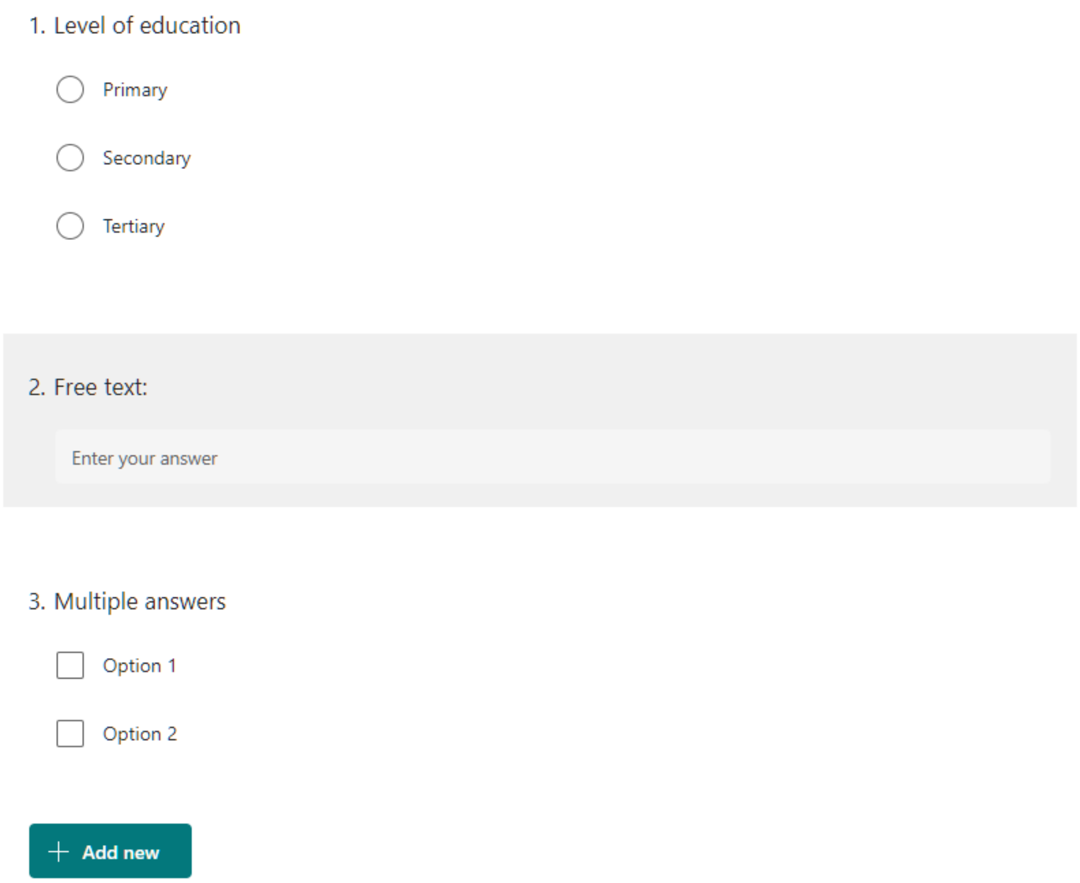
1. Level of education

○ Primary

○ Secondary

○ Tertiary

Figure 6: Add choices



1. Level of education

○ Primary

○ Secondary

○ Tertiary

2. Free text:

Enter your answer

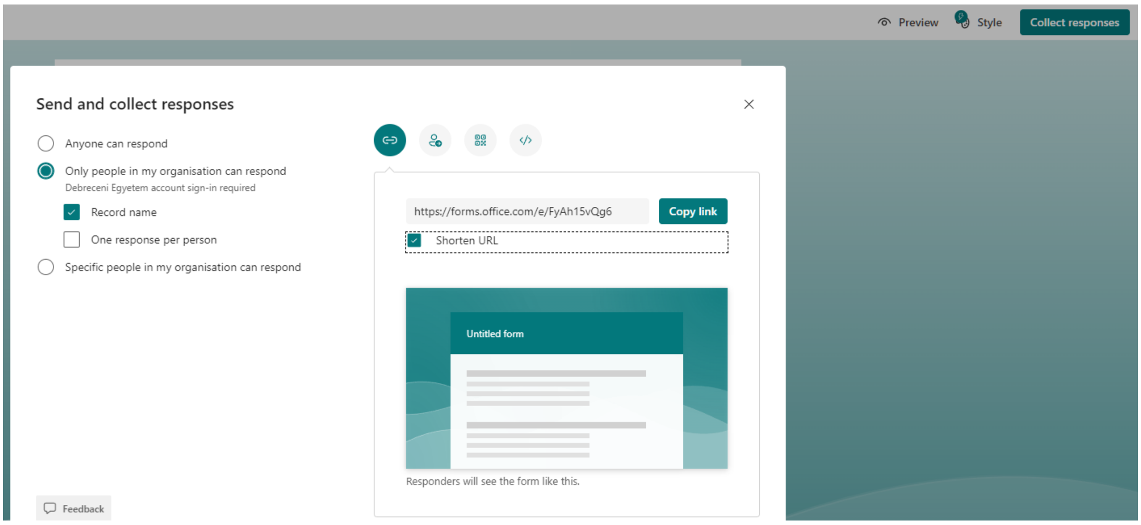
3. Multiple answers

□ Option 1

□ Option 2

Figure 7: Main question types

Once the questionnaire is ready, it can be used/distributed through the "Collect responses" option.



Send and collect responses

○ Anyone can respond

● Only people in my organisation can respond
Debreceni Egyetem account sign-in required

☑ Record name

□ One response per person

○ Specific people in my organisation can respond

https://forms.office.com/e/FyAh15vQg6

☑ Shorten URL

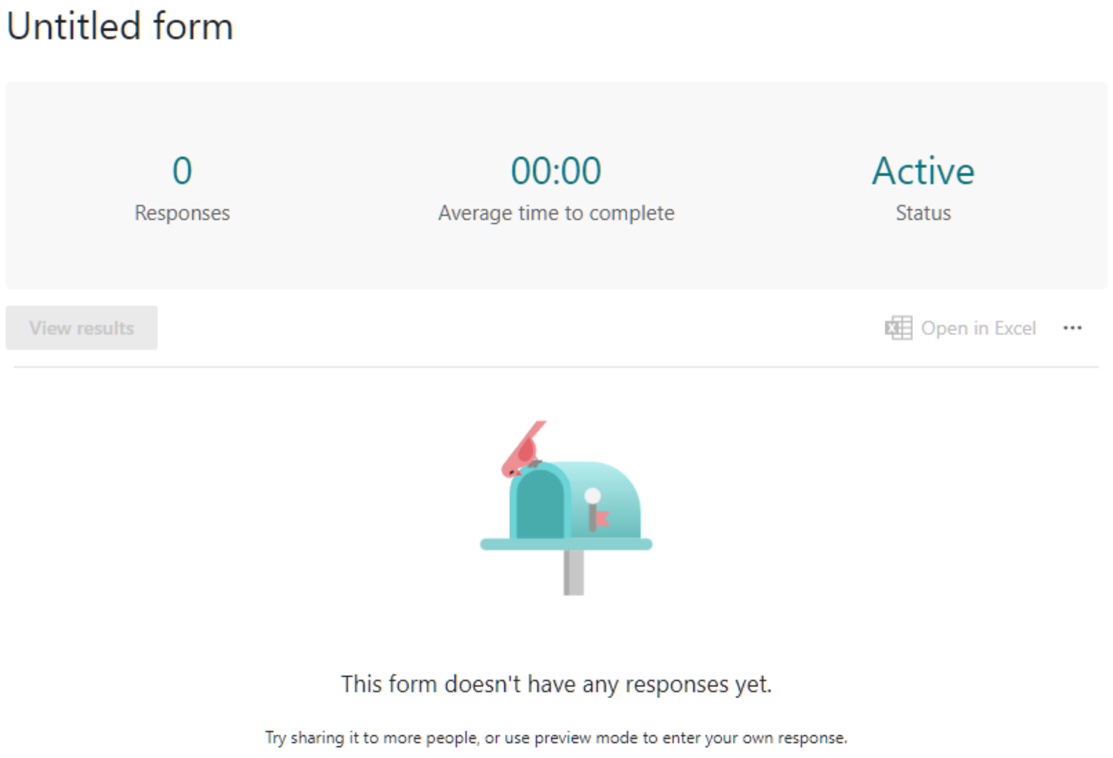
Responders will see the form like this.

Figure 8: Distribution of a form

It is worth setting the options "Anyone can respond" and "Shorten URL" for easier distribution of the questionnaire.

In the "Responses" menu item, there is detailed information on responses:

- Number of responses
- Average time to complete
- Results with charts
- Downloadable database in MS-Excel format


Untitled form

0
Responses

00:00
Average time to complete

Active
Status

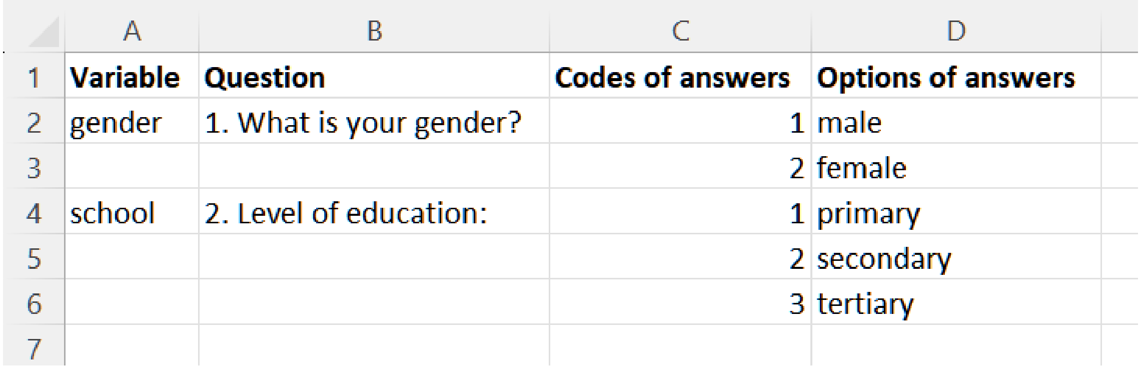
This form doesn't have any responses yet.

Try sharing it to more people, or use preview mode to enter your own response.

Figure 9: Response characteristics

# Data collection/data entry

An important principle is "garbage in, garbage out", that is, from "worthless" data it is possible to conduct "worthless" analyses. The same principle applies to the proper recording of data as well. The final database must be cleaned. We have already talked about the number and length of the questions. Questions should be grouped. A reference sample can be provided by the European Health Interview Survey (EHIS) questionnaire.
(https://ec.europa.eu/eurostat/documents/203647/203710/EHIS_wave_1_guidelines.pdf/ffbeb62c-8f64-4151-938c-9ef171d148e0)
Paper-based questionnaires require separate data recording. During recording, paper-based information is converted into digital characters (numbers and letters). During data entry, the missing values are encoded with "9" isolate the cause of the missing value (not written there by the responder or not recorded by the recorder). So many "9"-s are needed that it is not a realistic value for the given question (e.g., "999" for age). For analysis, we need to create a so-called codebook, which describes in detail the questions, possible answer options, and the variable names and coding used in the database. Statistical programs can easily cope with short variable names without special characters. In the codebook, we indicate the name of the variable, the corresponding question, and potential answers with the corresponding codes.



| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Variable** | **Question** | **Codes of answers** | **Options of answers** |
| 2 | gender | 1. What is your gender? | 1 | male |
| 3 | | | 2 | female |
| 4 | school | 2. Level of education: | 1 | primary |
| 5 | | | 2 | secondary |
| 6 | | | 3 | tertiary |
| 7 | | | | |

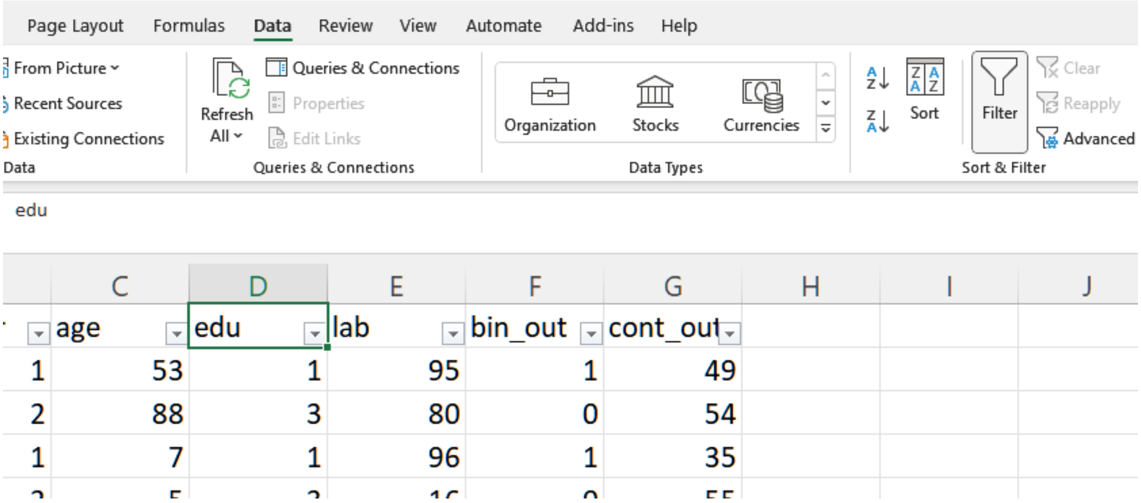
Figure 10: Codebook

## Double data entry

If you want to record a paper-based questionnaire, it is worth considering double data entry. This simply means that each questionnaire is recorded twice. The disadvantages of this are extra time and extra cost. The great advantage is that the database is free of recording errors, since the two databases can be compared, and the discrepancies can be retrieved.

## Data Cleansing

Before analysis, the raw database (recorded or received) must be cleaned. This process can be carried out even in MS-Excel. Very useful is the "Filter" function.

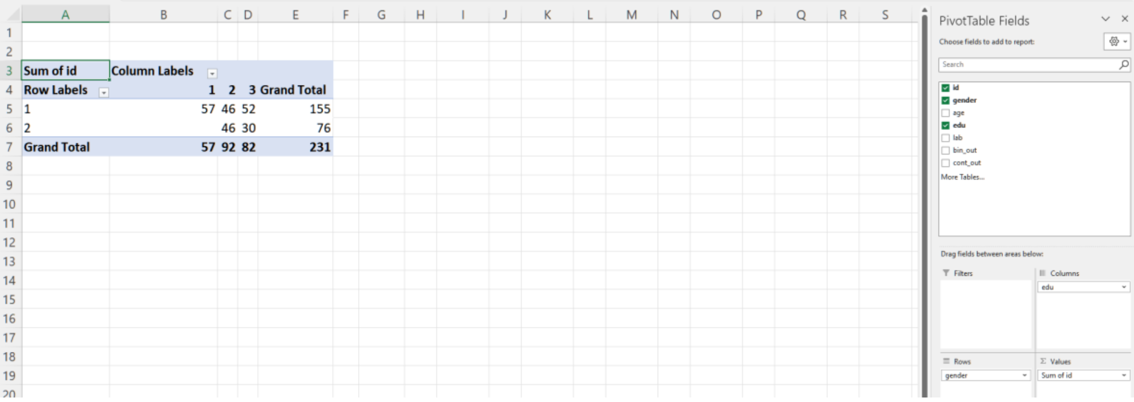


edu

|  | C | D | E | F | G |
|---|---|---|---|---|---|
|  | age | edu | lab | bin_out | cont_out |
| 1 | 53 | 1 | 95 | 1 | 49 |
| 2 | 88 | 3 | 80 | 0 | 54 |
| 1 | 7 | 1 | 96 | 1 | 35 |

Figure 11: Show filter

The filter shows the potential values in a column, so you can immediately see the wrong value. Erroneous, unrealistic, and 9-encoded values can be replaced with missing values. Among other functions (e.g., aggregation), the "PivotTable" can be found under the "Insert" menu item also helps with data cleaning.



| Sum of id | Column Labels |  |  |  |
|---|---|---|---|---|
| Row Labels | 1 | 2 | 3 | Grand Total |
| 1 | 57 | 46 | 52 | 155 |
| 2 |  | 46 | 30 | 76 |
| Grand Total | 57 | 92 | 82 | 231 |

PivotTable Fields

Choose fields to add to report:

- ☑ id
- ☑ gender
- ☐ age
- ☑ edu
- ☐ lab
- ☐ bin_out
- ☐ cont_out

More Tables...

Drag fields between areas below:

Filters

Columns: edu

Rows: gender

Values: Sum of id

Figure 12: Create a PivotTable

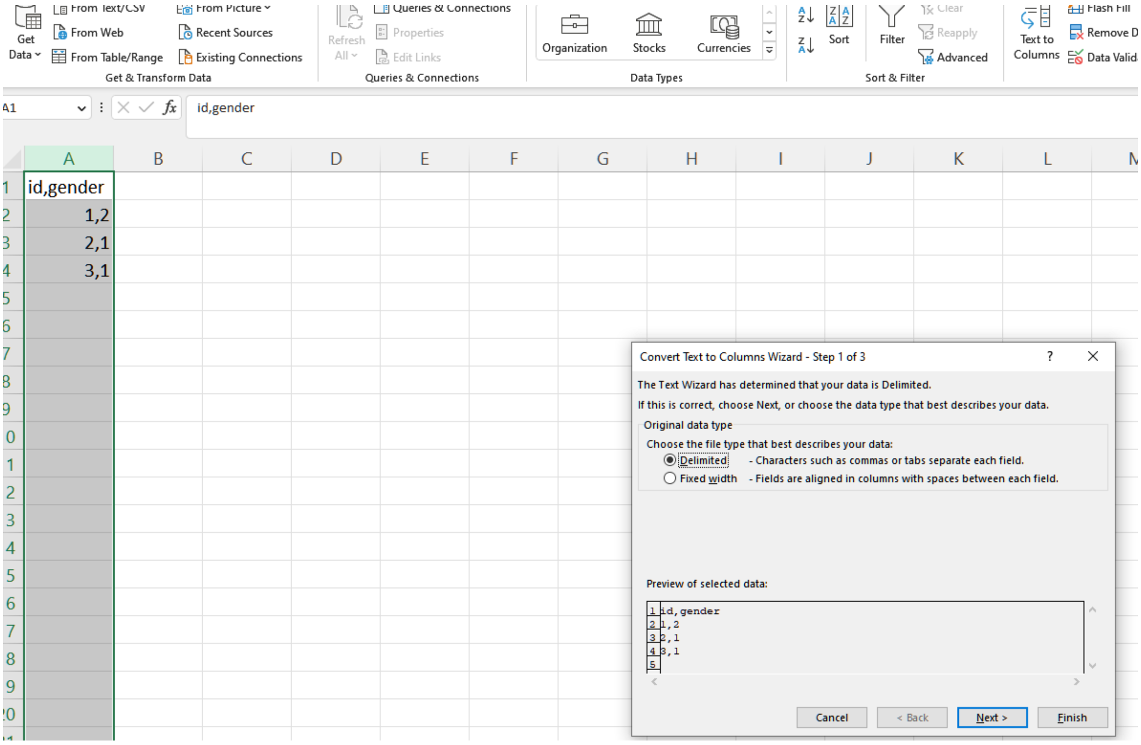
Using a simple drag-and-drop method, you can drag and drop individual variables onto a row or column section. You can even arrange multiple layers/subgroups along one axis. There is also a filter here to help you narrow down your database. And with "Values", not only the amounts can be calculated, but even the item numbers.

## Text in MS-Excel

One of the most common database formats is a text file. Most often, values are comma-separated (.csv), so comma-separated values are included in the file. A data file (.txt) separated by any other character (space, tab, etc.) is more common. In English and Hungarian data management, the decimal point is the main difference, which can be the source of additional data errors (e.g., numbers converted to dates). Hungarian .csv is a bit misleading, since the separator character here is not a comma, but a semicolon.

The "Text Cutting Wizard" can be brought up in two ways:

- copy text data to Excel
- Within the "Data" tab, select "Text to Columns"



| | A |
|---|---|
| 1 | id,gender |
| 2 | 1,2 |
| 3 | 2,1 |
| 4 | 3,1 |

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

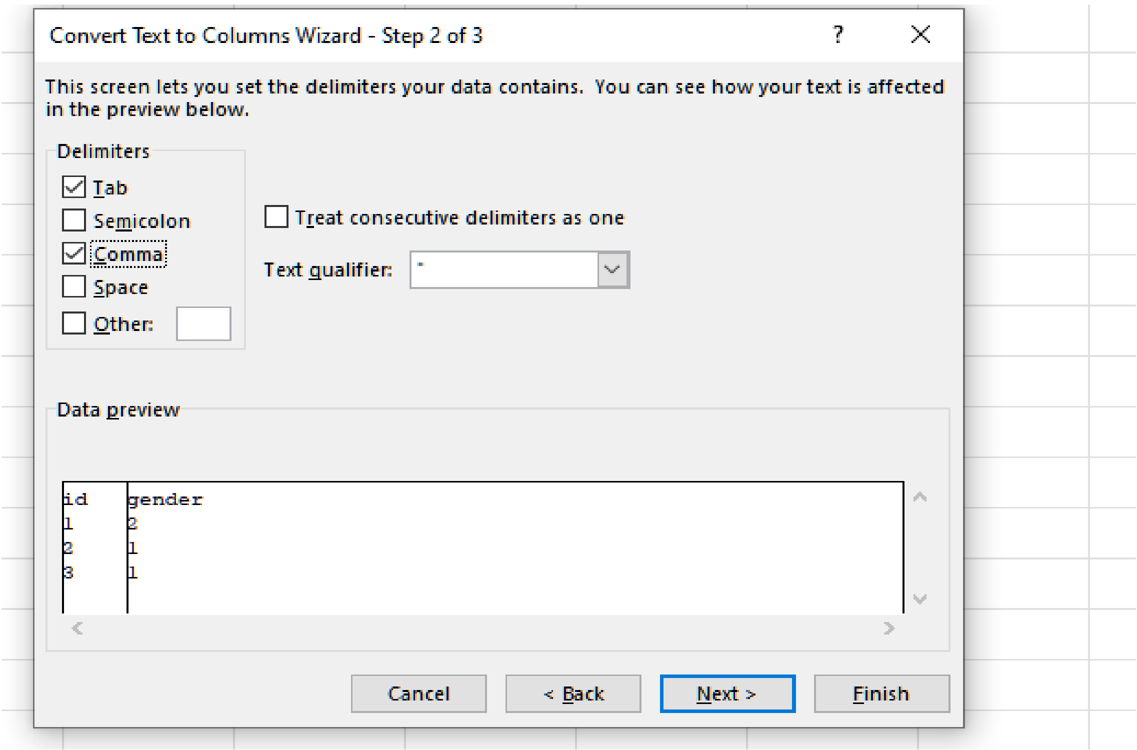
○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 id,gender
2 1,2
3 2,1
4 3,1
5
```

Figure 13: Text Splitting Wizard

It is important to choose a delimiter (also called a field separator), which can be any character besides the ones on the list.



Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab

☐ Semicolon

☑ Comma

☐ Space

☐ Other:

☐ Treat consecutive delimiters as one

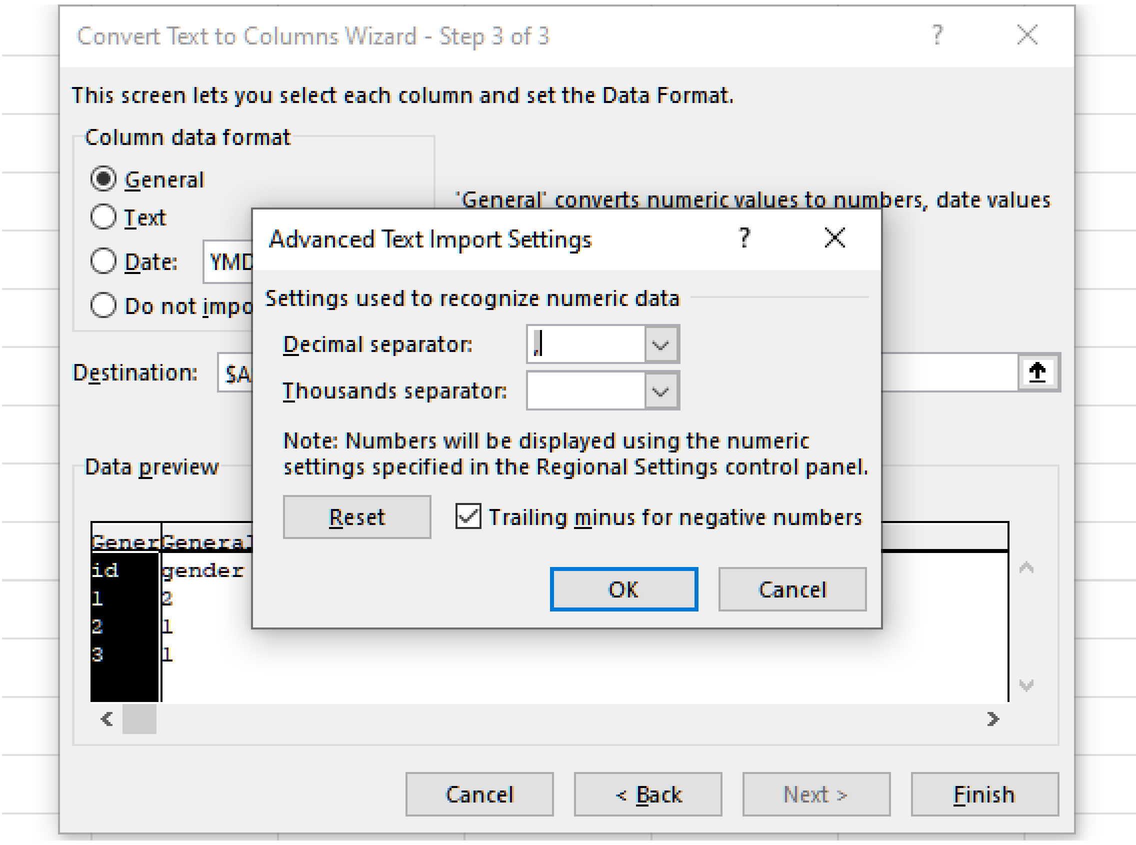
Text qualifier: "

Data preview

```
id  gender
1   2
2   1
3   1
```

Figure 14: Field separator

If you are inserting an English database, the decimal point and thousands separator on the next page are important.



Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

● General
○ Text
○ Date: YMD
○ Do not impo

'General' converts numeric values to numbers, date values

Destination: $A

Data preview

```
GenerGeneral
id   gender
1    2
2    1
3    1
```

Advanced Text Import Settings

Settings used to recognize numeric data

Decimal separator: ,

Thousands separator:

Note: Numbers will be displayed using the numeric settings specified in the Regional Settings control panel.

☑ Trailing minus for negative numbers

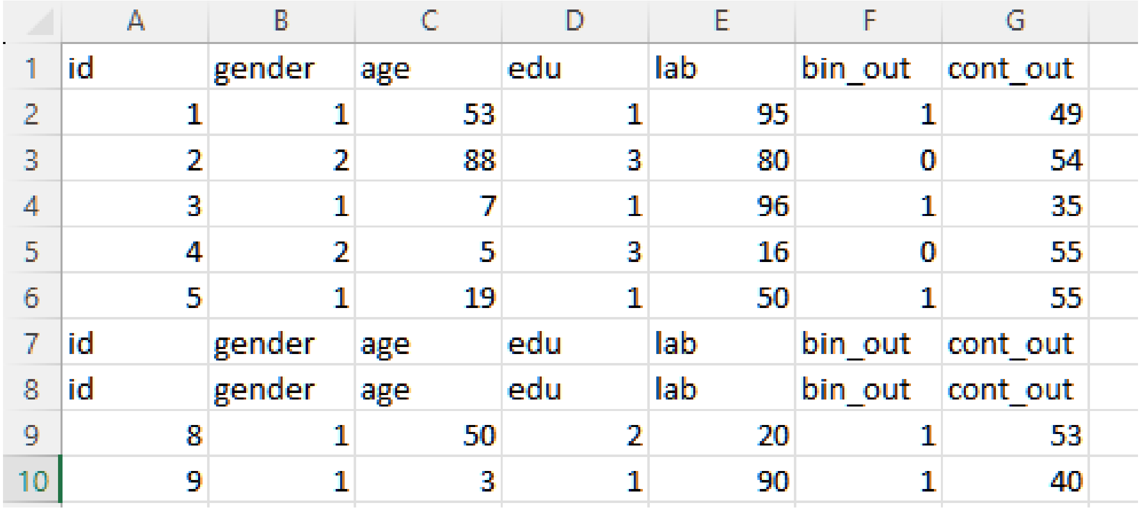
Figure 15: Set decimal point

In Hungarian Excel, the decimal point should be a point and the thousand separator should be a comma. This way, automatic conversion won't happen.

## Connect data / VLOOKUP function

There are two main types of database connectivity.

- *Append:* Concatenation of databases with the same structure can occur when multiple people enter questionnaires in the same system and end up wanting to create a large database. It's a simple matter of copying to each other, you can copy the heading to the bottom to be on the safe side, and after checking back, you can delete the copy and the new heading.



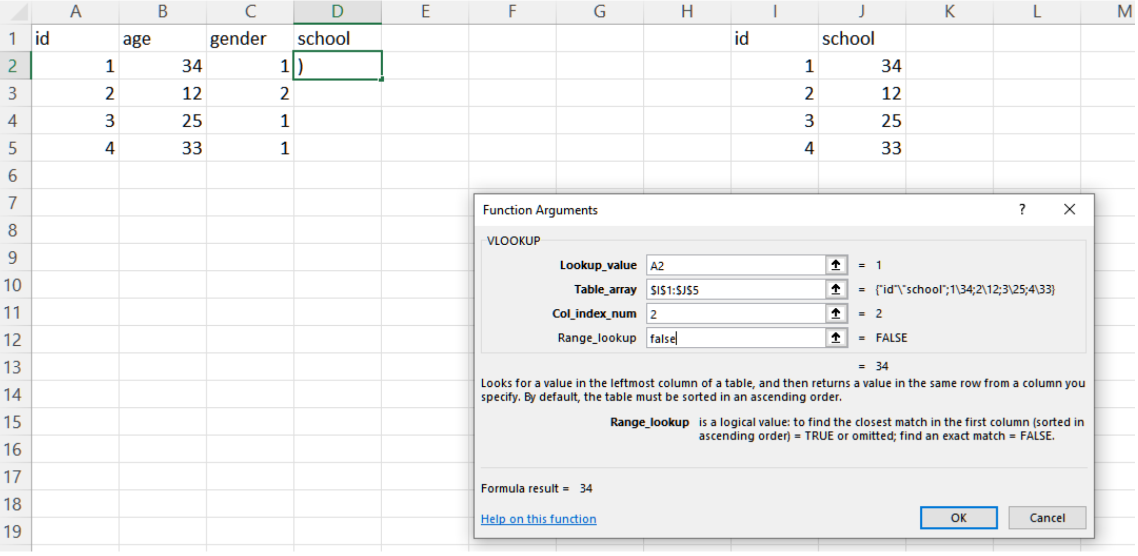
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | id | gender | age | edu | lab | bin_out | cont_out |
| 2 | 1 | 1 | 53 | 1 | 95 | 1 | 49 |
| 3 | 2 | 2 | 88 | 3 | 80 | 0 | 54 |
| 4 | 3 | 1 | 7 | 1 | 96 | 1 | 35 |
| 5 | 4 | 2 | 5 | 3 | 16 | 0 | 55 |
| 6 | 5 | 1 | 19 | 1 | 50 | 1 | 55 |
| 7 | id | gender | age | edu | lab | bin_out | cont_out |
| 8 | id | gender | age | edu | lab | bin_out | cont_out |
| 9 | 8 | 1 | 50 | 2 | 20 | 1 | 53 |
| 10 | 9 | 1 | 3 | 1 | 90 | 1 | 40 |

Figure 16: Appending databases

- *Merge:* used more frequently, in which case the existing database is concatenated with another. In simple terms, values are added to a variable that clearly identifies the database.

The two main functions of VLOOKUP function are exact match and categorization (closest match).

In an exact match, we look for the cutting point from another database along with a clear identifier. In this example, we bring education from another database:



| | A | B | C | D | I | J |
|---|---|---|---|---|---|---|
| 1 | id | age | gender | school | id | school |
| 2 | 1 | 34 | 1 | ) | 1 | 34 |
| 3 | 2 | 12 | 2 | | 2 | 12 |
| 4 | 3 | 25 | 1 | | 3 | 25 |
| 5 | 4 | 33 | 1 | | 4 | 33 |

Function Arguments

VLOOKUP

Lookup_value: A2 = 1

Table_array: $I$1:$J$5 = {"id"\"school";1\34;2\12;3\25;4\33}

Col_index_num: 2 = 2

Range_lookup: false = FALSE

= 34

Looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order.

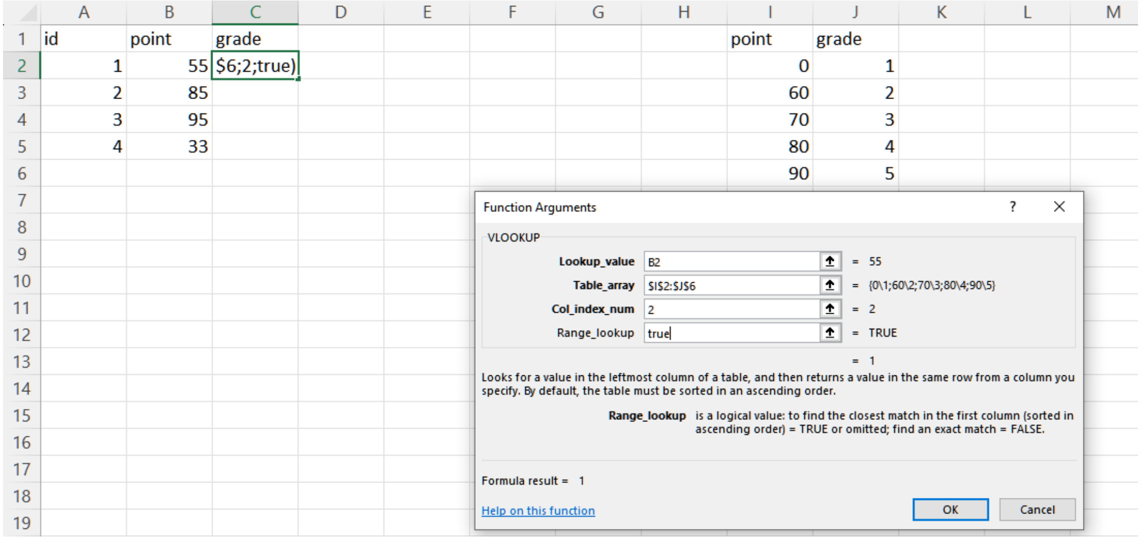
**Range_lookup** is a logical value: to find the closest match in the first column (sorted in ascending order) = TRUE or omitted; find an exact match = FALSE.

Formula result = 34

Figure 17: VLOOKUP setting for exact match

It is important to use an absolute reference for the value of the "Table" when selecting the secondary range ($ sign), so that when copying the formula, the range that represents the secondary data table will not slip.

During categorization, we convert a continuous variable to a categorical one (e.g., score-grade; age-age group):



| | A | B | C | I | J |
|---|---|---|---|---|---|
| 1 | id | point | grade | point | grade |
| 2 | 1 | 55 | $6;2;true) | 0 | 1 |
| 3 | 2 | 85 | | 60 | 2 |
| 4 | 3 | 95 | | 70 | 3 |
| 5 | 4 | 33 | | 80 | 4 |
| 6 | | | | 90 | 5 |

Function Arguments

VLOOKUP

| | | |
|---|---|---|
| **Lookup_value** | B2 | = 55 |
| **Table_array** | $I$2:$J$6 | = {0\1;60\2;70\3;80\4;90\5} |
| **Col_index_num** | 2 | = 2 |
| Range_lookup | true | = TRUE |

= 1

Looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order.

**Range_lookup** is a logical value: to find the closest match in the first column (sorted in ascending order) = TRUE or omitted; find an exact match = FALSE.

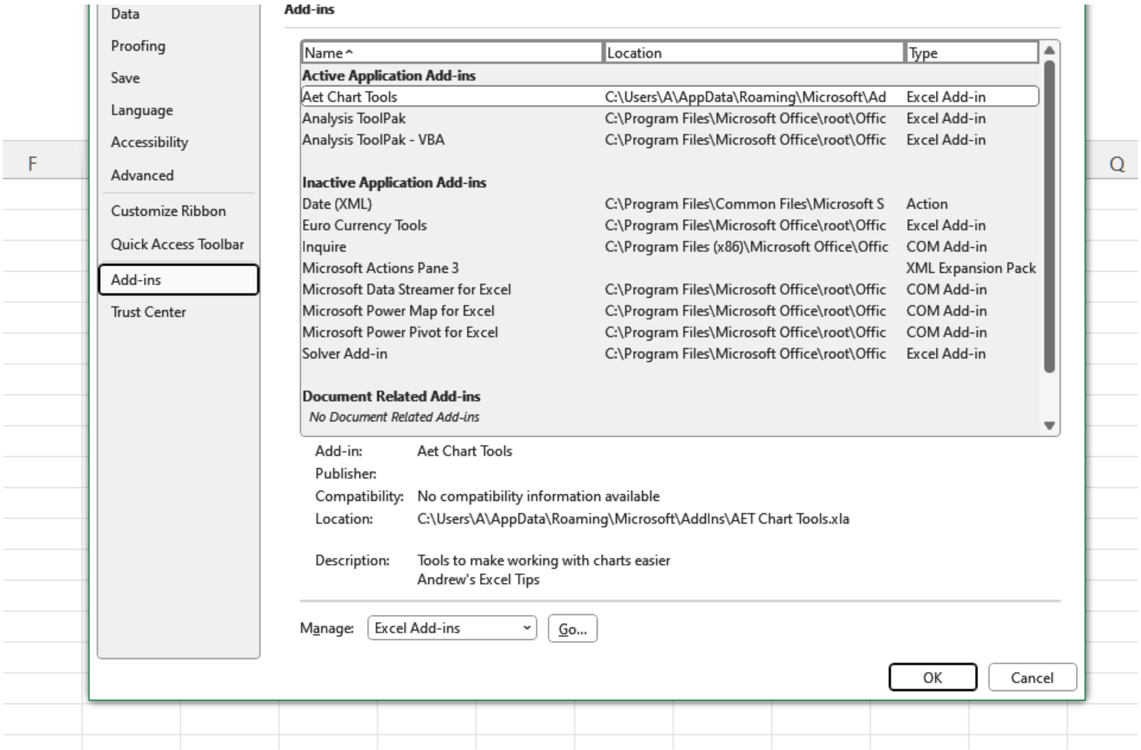
Formula result = 1

Figure 18: VLOOKUP setting for approximate match

In this case, you should still use an absolute reference, but "Range_lookup" is "TRUE", that is, unlike the previous example (where it was "FALSE"), it does not return a value if there is an exact match, but works along ranges of values.

## Data analysis in MS-Excel

MS-Excel has a surprising versatility in displaying and managing data. With built-in functions, a wide variety of statistical methods are available. There are also additional plugins that can be downloaded for free or for money. However, there is a built-in plugin that is not available by default, which is "Data Analysis". To bring it up, you need to go to "Excel Add-ins" after clicking "File", "Options", "Add-ins":



Add-ins

| Name ^ | Location | Type |
|---|---|---|
| **Active Application Add-ins** | | |
| Aet Chart Tools | C:\Users\A\AppData\Roaming\Microsoft\Ad | Excel Add-in |
| Analysis ToolPak | C:\Program Files\Microsoft Office\root\Offic | Excel Add-in |
| Analysis ToolPak - VBA | C:\Program Files\Microsoft Office\root\Offic | Excel Add-in |
| **Inactive Application Add-ins** | | |
| Date (XML) | C:\Program Files\Common Files\Microsoft S | Action |
| Euro Currency Tools | C:\Program Files\Microsoft Office\root\Offic | Excel Add-in |
| Inquire | C:\Program Files (x86)\Microsoft Office\Offic | COM Add-in |
| Microsoft Actions Pane 3 | | XML Expansion Pack |
| Microsoft Data Streamer for Excel | C:\Program Files\Microsoft Office\root\Offic | COM Add-in |
| Microsoft Power Map for Excel | C:\Program Files\Microsoft Office\root\Offic | COM Add-in |
| Microsoft Power Pivot for Excel | C:\Program Files\Microsoft Office\root\Offic | COM Add-in |
| Solver Add-in | C:\Program Files\Microsoft Office\root\Offic | Excel Add-in |
| **Document Related Add-ins** | | |
| *No Document Related Add-ins* | | |

Add-in: Aet Chart Tools
Publisher:
Compatibility: No compatibility information available
Location: C:\Users\A\AppData\Roaming\Microsoft\AddIns\AET Chart Tools.xla

Description: Tools to make working with charts easier
Andrew's Excel Tips

Manage: Excel Add-ins

Figure 19: Access Excel Add-ins

Then, by ticking the "Analysis ToolPak" options, "Data Analysis" will become available at the end of the "Data" menu item.

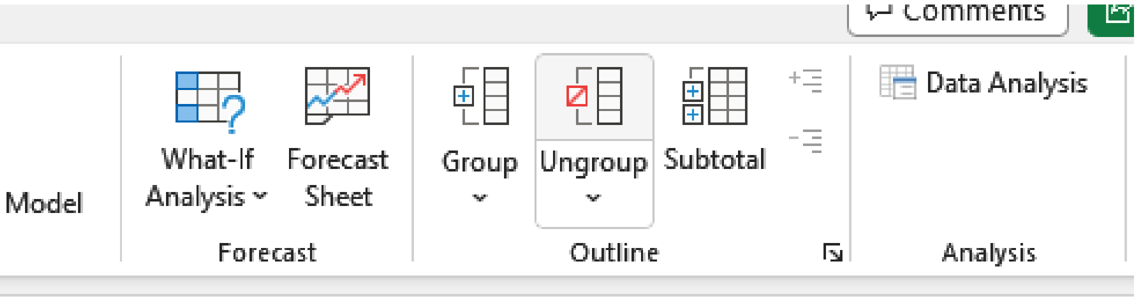
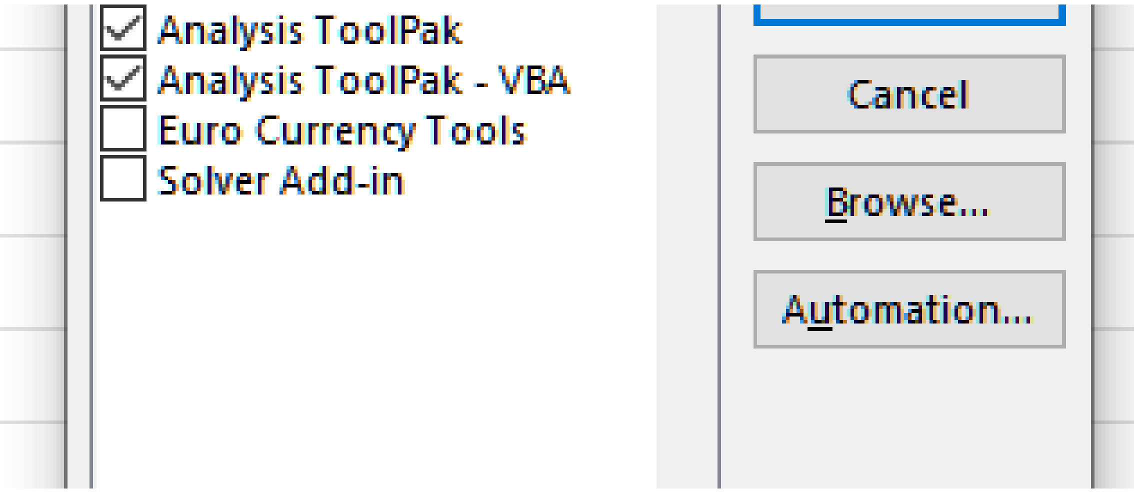


Figure 20: Data Analysis button location



☑ Analysis ToolPak
☑ Analysis ToolPak - VBA
□ Euro Currency Tools
□ Solver Add-in

Figure 21: Extensions required to visualize data analysis

Here, in addition to Descriptive Statistics, and to Singel Factor ANOVA, multiple linear regressions can also be performed (see later).

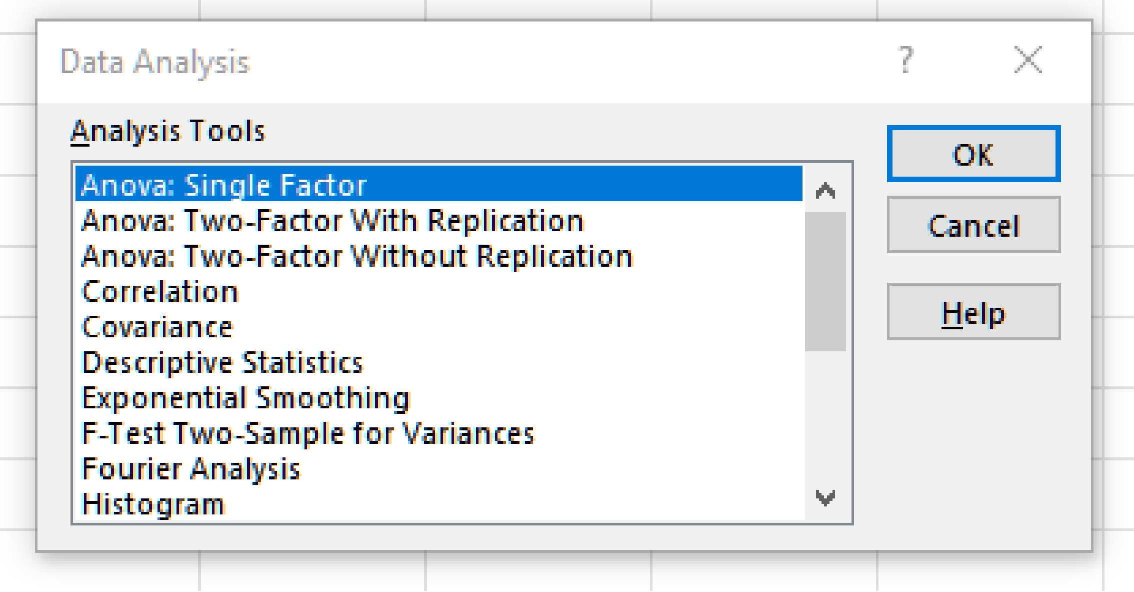


Figure 22: Analytical methods available in data analysis

# About statistical programs

Programs that enable statistical analysis, vary widely. For some, analysis is a secondary function (MS-Excel) and only limited methods are available. Software developed specifically for statistical analysis includes free (e.g.: R, PSPP), paid (e.g.: Stata, SPSS, SAS) versions, as well as programming languages (e.g.: Python) with statistical modules.

## Large Database Management (KNIME)

Nowadays, more, and larger databases are being created, in the so-called "Big Data" size. In addition to the huge size (volume) (e.g.: gigabyte, terabyte and petabyte), dynamic change is also a characteristic. MS-Excel has a limit of 1,048,576 rows and 16,384 columns. Excel is not a database manager, but a Spreadsheet Software. A database larger than this should be opened in a special database management software or statistical program. Relatively "smaller" databases are also worth converting or even opening with the help of special programs. It is worth saying a few words about Knime, which is also free. It is also ahead of statistical programs in terms of the speed of file operations. The program operates with so-called “nodes”. You can open files (CSV Reader), filter (Row/Column Filter) data, merge (Joiner) databases, and dump them into a (CSV Writer) file.

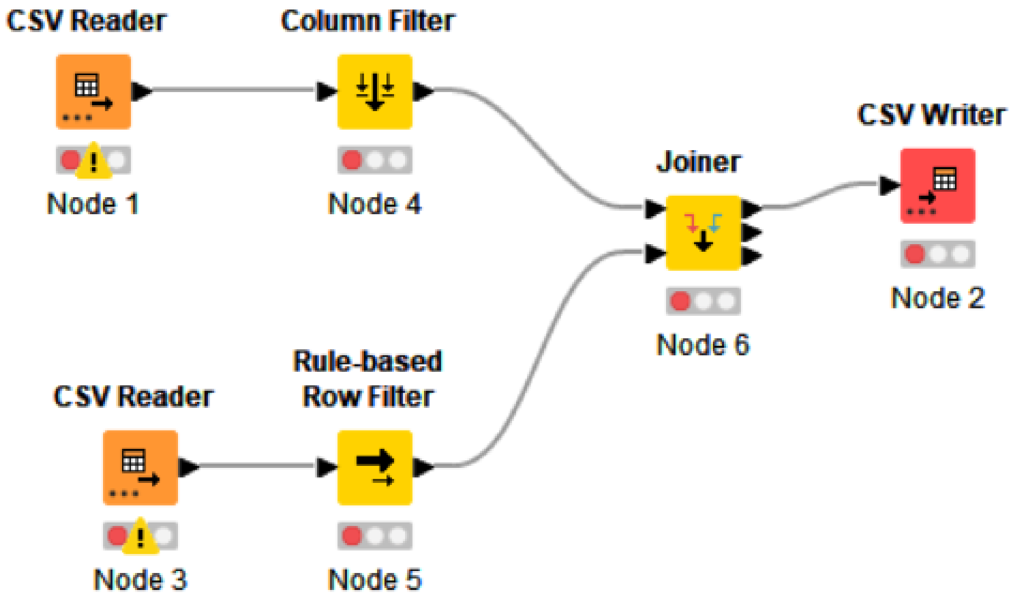
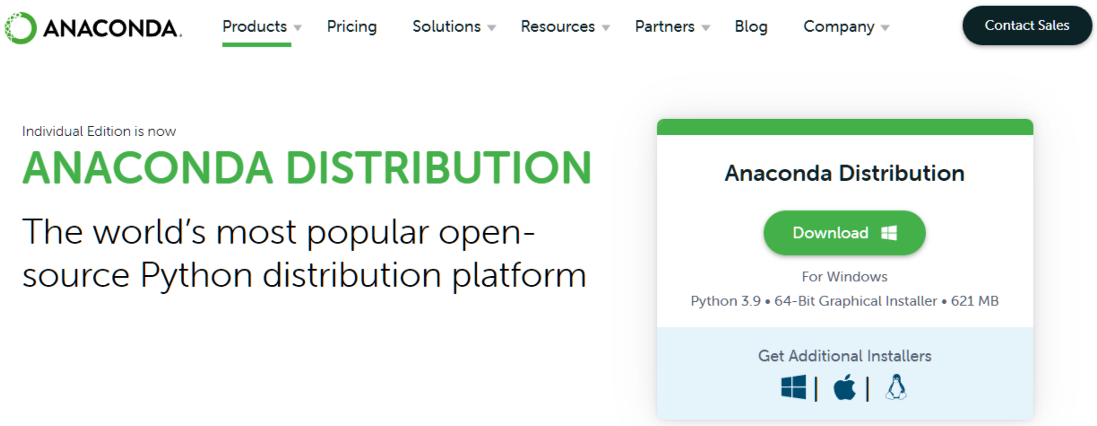


Figure 23: KNIME nodes

## Python

Python is one of the most widely used programming languages. As it has more and more data science packages, it is becoming increasingly popular among data scientists. It supports nearly all statistical methods by default or through free add-on packages. Its great advantage, besides being free, is its speed and wide range of applications. It can be downloaded for all operating systems, one of its most popular distributions is Anaconda (https://www.anaconda.com/products/distribution ).



Individual Edition is now

ANACONDA DISTRIBUTION

The world's most popular open-source Python distribution platform

Anaconda Distribution

For Windows

Python 3.9 • 64-Bit Graphical Installer • 621 MB

Figure 24: Anaconda Python distribution website

After downloading and installing, the "Jupyter Notebook (anaconda3)" icon is worth running.

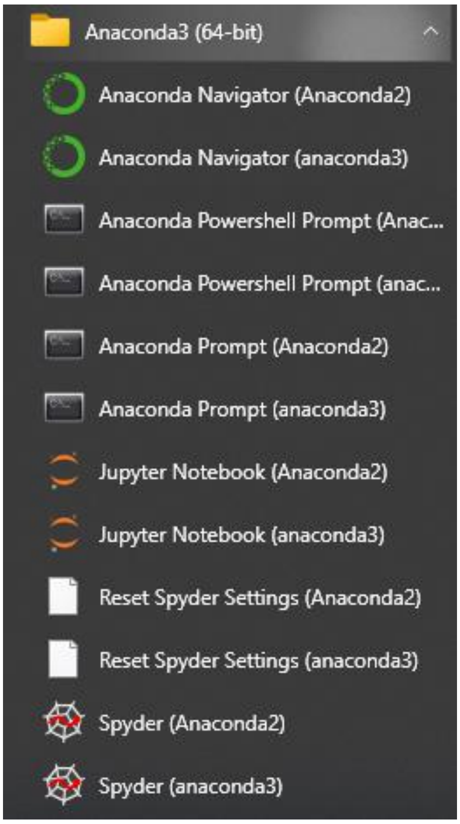

Anaconda3 (64-bit)
Anaconda Navigator (Anaconda2)
Anaconda Navigator (anaconda3)
Anaconda Powershell Prompt (Anac...
Anaconda Powershell Prompt (anac...
Anaconda Prompt (Anaconda2)
Anaconda Prompt (anaconda3)
Jupyter Notebook (Anaconda2)
Jupyter Notebook (anaconda3)
Reset Spyder Settings (Anaconda2)
Reset Spyder Settings (anaconda3)
Spyder (Anaconda2)
Spyder (anaconda3)

Figure 25: Anaconda3 in the Start menu

Once run, its interface will open in your default web browser.

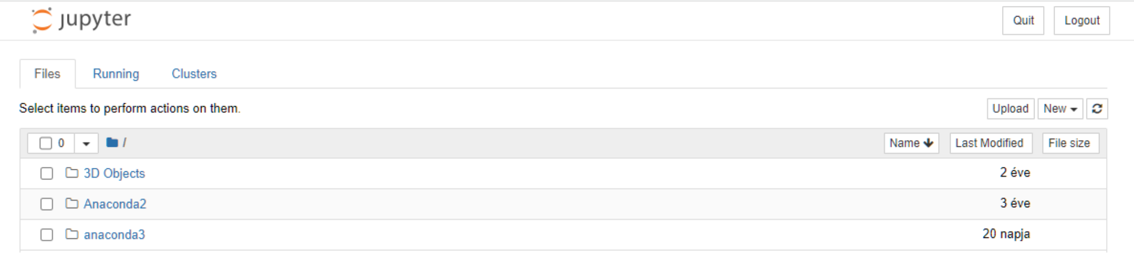


Figure 26: Anaconda interface in default browser (http://localhost:8888/tree)

To create a new file, you need to click on the "New" and then "Python 3" button.

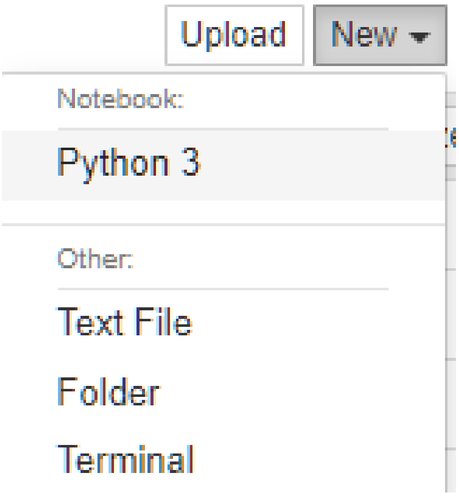


Figure 27: Creating a new Python file

The default file containing code and results an .ipynb extension and is located in the user's directory (C:\Users\username\).
There are many tutorials for programming Python available on the internet, and in this note we will get acquainted with just a few basic commands.

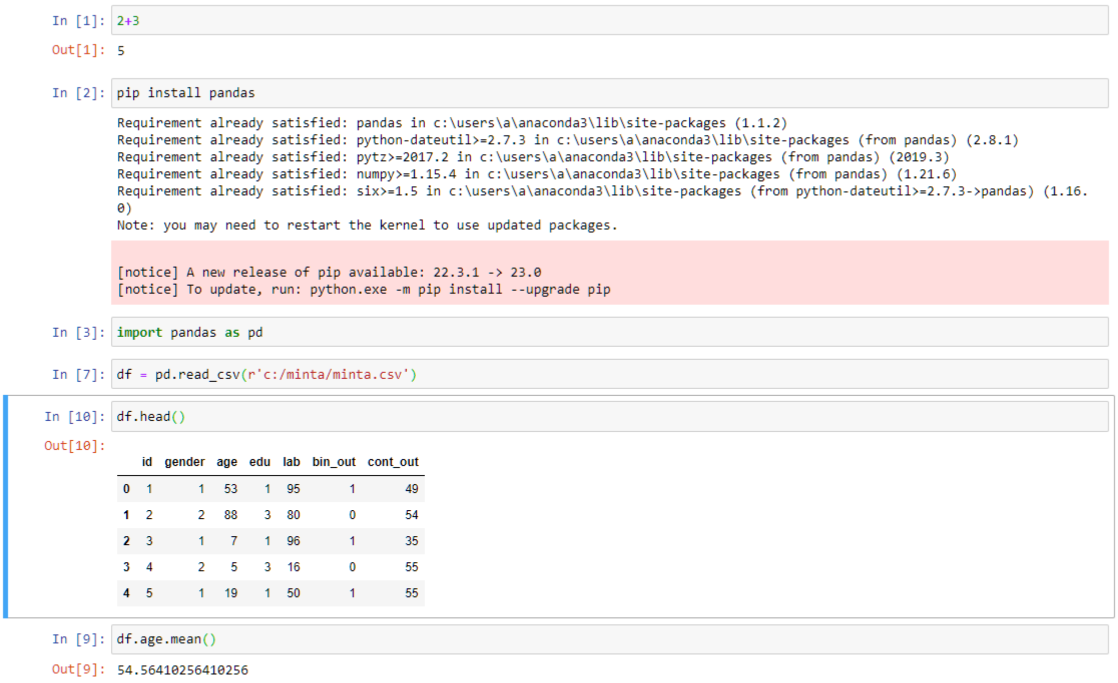
The commands in Figure 28, and the results obtained when they are run, are as follows:

In[1]: first command (2+3)
Out[1]: result of command (5)

You can run the command you typed by pressing Shift+Enter.
The "pip install pandas" command installs one of the most commonly used packages (pandas).

The command "import panda as pd" loads the "pandas" package and can be referred to as "pd" in the future.
The command "df = pd.read_csv(r'c:/sample/sample.csv')" loads our sample.csv file into a dataframe variable named df. If the separator is not the default comma, it must be specified separately or changed in the csv file first.
The command "df.head()" lists the contents of the df variable.
The command "df.age.mean()" returns the variable mean of age.

You can click on a specific row to select the row, currently In[10] is selected. The "x" key deletes (cuts) that line, and the "b" key opens a new command line.



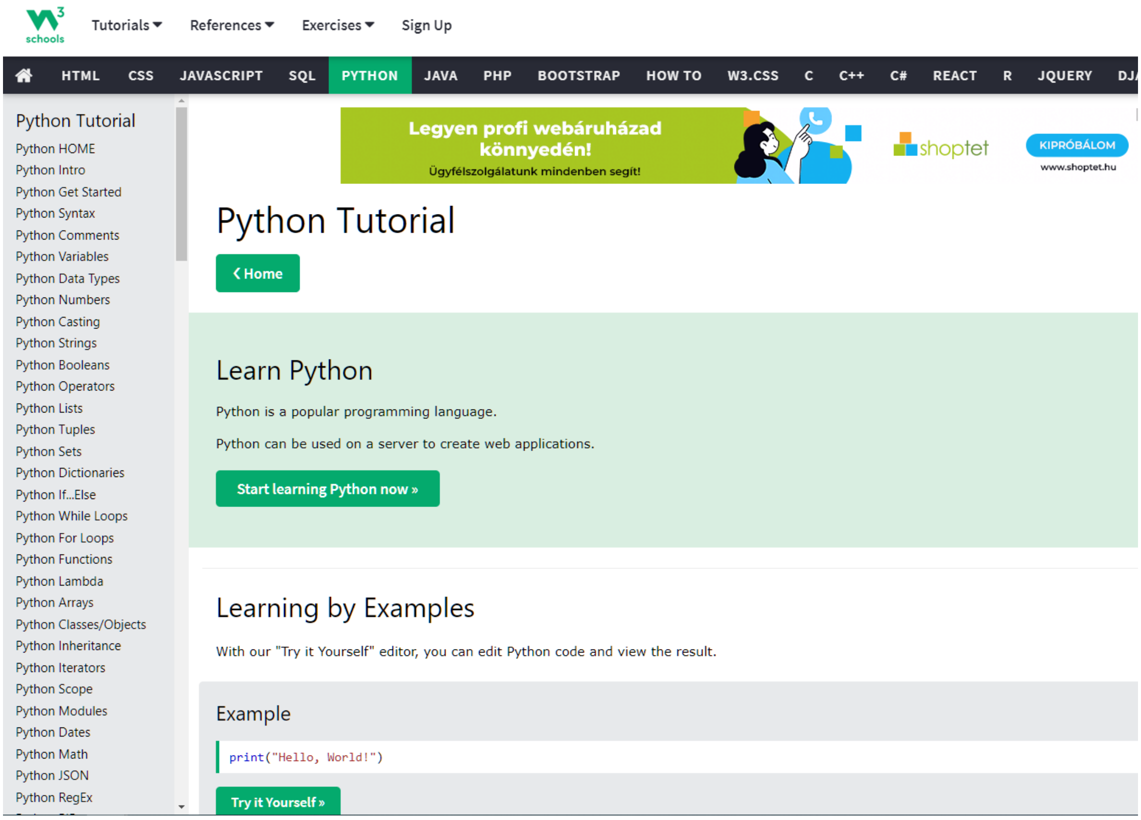
```
In [1]: 2+3
Out[1]: 5

In [2]: pip install pandas

        Requirement already satisfied: pandas in c:\users\a\anaconda3\lib\site-packages (1.1.2)
        Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\a\anaconda3\lib\site-packages (from pandas) (2.8.1)
        Requirement already satisfied: pytz>=2017.2 in c:\users\a\anaconda3\lib\site-packages (from pandas) (2019.3)
        Requirement already satisfied: numpy>=1.15.4 in c:\users\a\anaconda3\lib\site-packages (from pandas) (1.21.6)
        Requirement already satisfied: six>=1.5 in c:\users\a\anaconda3\lib\site-packages (from python-dateutil>=2.7.3->pandas) (1.16.
        0)
        Note: you may need to restart the kernel to use updated packages.

        [notice] A new release of pip available: 22.3.1 -> 23.0
        [notice] To update, run: python.exe -m pip install --upgrade pip

In [3]: import pandas as pd

In [7]: df = pd.read_csv(r'c:/minta/minta.csv')

In [10]: df.head()
Out[10]:
```

| | id | gender | age | edu | lab | bin_out | cont_out |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 53 | 1 | 95 | 1 | 49 |
| 1 | 2 | 2 | 88 | 3 | 80 | 0 | 54 |
| 2 | 3 | 1 | 7 | 1 | 96 | 1 | 35 |
| 3 | 4 | 2 | 5 | 3 | 16 | 0 | 55 |
| 4 | 5 | 1 | 19 | 1 | 50 | 1 | 55 |

```
In [9]: df.age.mean()
Out[9]: 54.56410256410256
```

Figure 28: Python commands and results

The W3 Schools page (https://www.w3schools.com/python/default.asp) helps you learn the programming language from the basics to the advanced level.



Python Tutorial

Learn Python

Python is a popular programming language.

Python can be used on a server to create web applications.

Learning by Examples

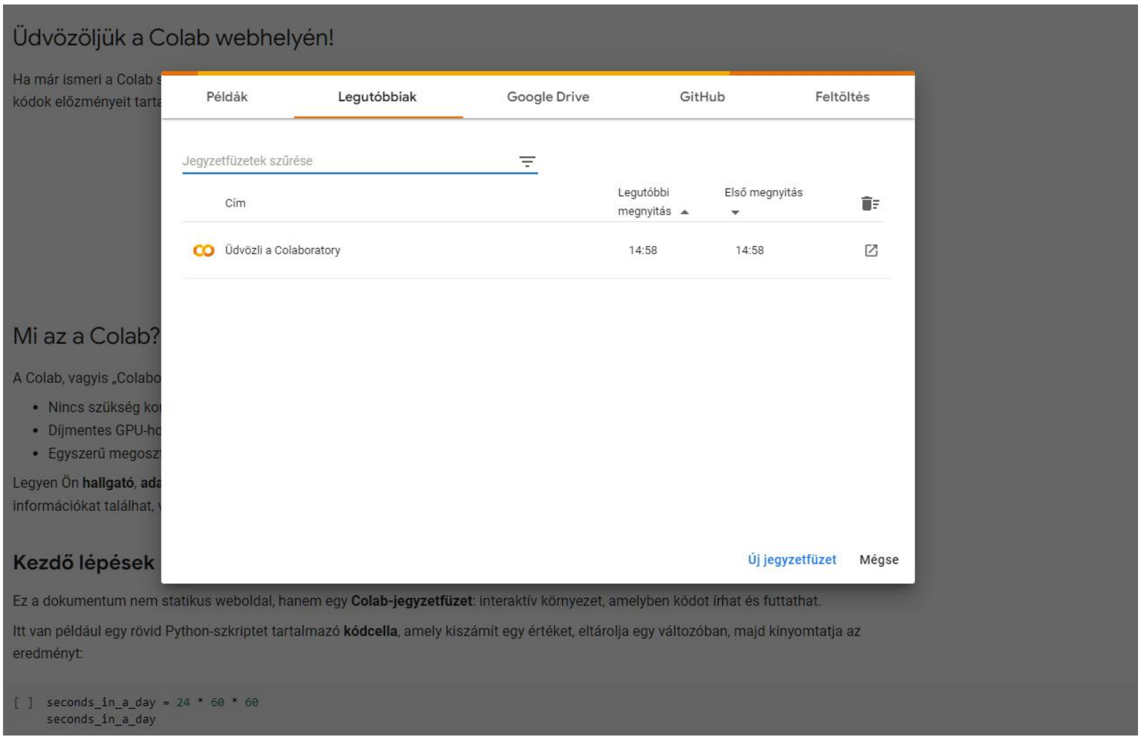
With our "Try it Yourself" editor, you can edit Python code and view the result.

Example

```
print("Hello, World!")
```

Figure 29: W3Schools website Python tutorial

Those who do not want to download Anaconda will also have the opportunity to use an online interface, as long as they have a Google account.



Üdvözöljük a Colab webhelyén!

Ha már ismeri a Colab s

kódok előzményeit tarta

| Cím | Legutóbbi megnyitás ▲ | Első megnyitás ▼ |
|---|---|---|
| Üdvözli a Colaboratory | 14:58 | 14:58 |

Mi az a Colab?

A Colab, vagyis „Colabo

- Nincs szükség ko
- Díjmentes GPU-ho
- Egyszerű megoszt

Legyen Ön **hallgató**, **ada**

információkat találhat, v

Kezdő lépések

Ez a dokumentum nem statikus weboldal, hanem egy **Colab-jegyzetfüzet**: interaktív környezet, amelyben kódot írhat és futtathat.

Itt van például egy rövid Python-szkriptet tartalmazó **kódcella**, amely kiszámít egy értéket, eltárolja egy változóban, majd kinyomtatja az eredményt:

```
[ ]  seconds_in_a_day = 24 * 60 * 60
     seconds_in_a_day
```

Figure 30: Google Colaboratory (https://colab.research.google.com/)

# Data analysis

After cleaning the data, we come to the data analysis. Data analysis can be divided into three main parts:

- descriptive analysis to obtain a picture of the study population, with particular reference to the main characteristics (including the outcome studied)
- simple analyses give an idea of potential assumptions (hypothesis) (e.g., t-test, chi-squared test)
- and multiple analyses, let us see the clear effect (adjusted for possible confounders)

Before discussing the details, it is worth briefly reviving some concepts. The p-value is derived from the English word “probability”. Most generally, we leave 5% to the role of chance (p=0.05) in analyses. If the role of chance is greater than 5%, we do not talk about an actual association, we do not reject the so-called null hypothesis ($H_0$, which indicates the absence of difference/association). The research hypothesis ($H_1$, is the opposite of the null hypothesis), is the assumption of the association between the potential influencing factor and the outcome. This is generated through descriptive etiological studies, and in connection with the analysis, descriptive analyses give the picture, which is crystallized by simple analysis, and finalized by multiple analyses. The most common p-value (0.05) has a 95% confidence interval. The latter illustrates the uncertainty of the estimate, the narrower it is, the more accurate the estimate, the smaller the differences that can be detected. That is, if we repeat the study 100 times (on different random samples), the results will be within this range 95 times. The other meaning is that the population parameter we are looking for (which we are interested in and why we are performing the study on a representative sample) is 95% likely to be in the range. The general formula for a 95% confidence interval is:

95%CI = point estimation ± 1.96xstandard error

The standard error varies from one calculated indicator to another (point estimation). The 1.96 is derived from the standard normal distribution (which has a 0 expected value and a standard deviation of 1). Most population parameters (body weight, blood sugar, etc.) follow a normal distribution.

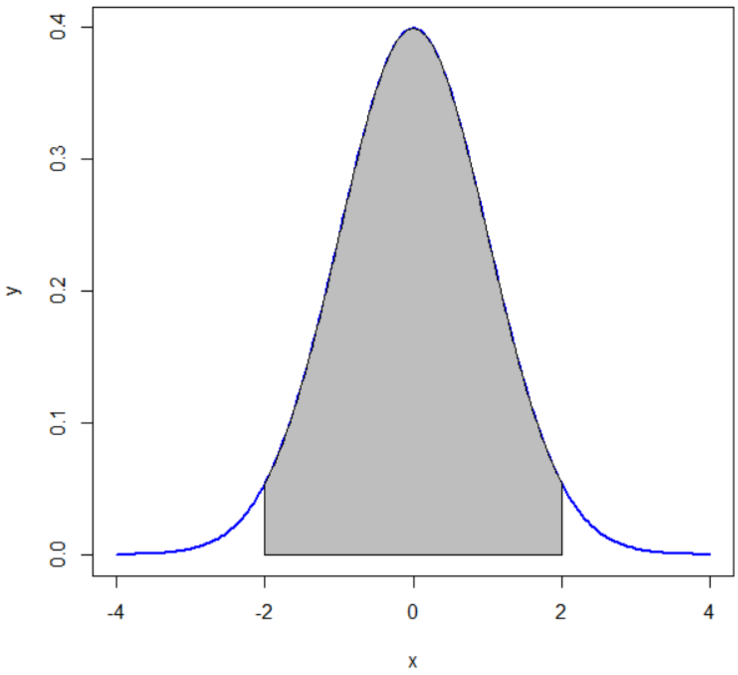


Figure 31: Standard normal distribution

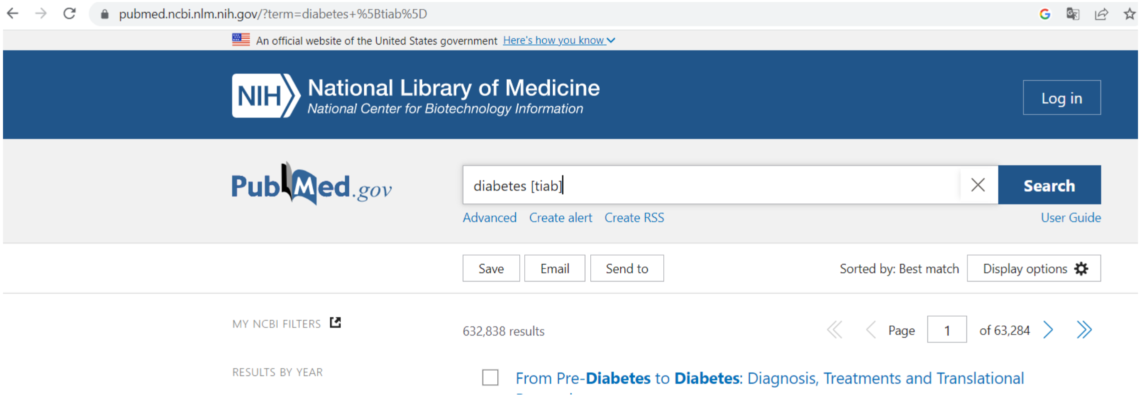
The area under the curve is 100%, that is, a probability of 1. When calculating the 95% confidence interval for the point estimation, we adapt this curve to our position (own standard deviation, our own sample size, and our own expected value (average)).

# Literature research

We cannot measure the entire population due to limited resources (time and money). Therefore, we take a representative sample (similar to the entire population (source population) in as many characteristics as possible) (usually by randomly sampling) and perform the analysis among them. In addition to representativeness, it is also important to have the right number of samples. The minimum number of samples needed to achieve the desired effect can be estimated by the statistical programs.

For this (and for further analysis) we will use the easily available and free R program.

The first step is to research the relevant literature. The most commonly used TAG on PubMed is TIAB TAG, which narrows results down to occurrences in the title and abstract, e.g.:



pubmed.ncbi.nlm.nih.gov/?term=diabetes+%5Btiab%5D

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

diabetes [tiab]

632,838 results

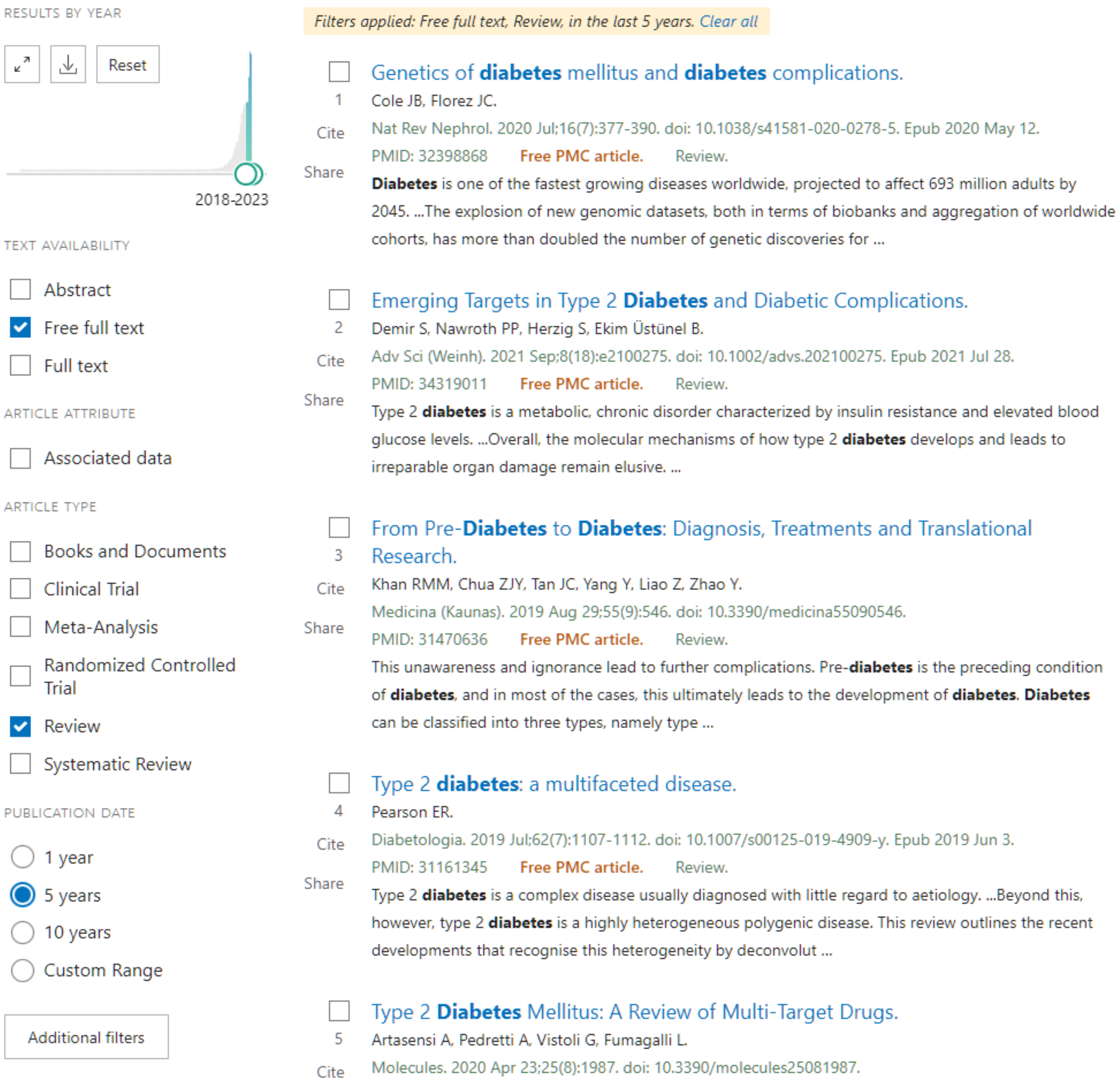
Page 1 of 63,284

Sorted by: Best match

□ From Pre-Diabetes to Diabetes: Diagnosis, Treatments and Translational

Figure 32: PubMed search using [tiab] TAG

Articles may be further narrowed down by publication date, language, etc. From the relevant articles, we can extract the results that are important to us. For a major topic review, it is worth turning on the "Review" or "Systematic Review" filter, as well as looking at the relevant results of the last 5 years. In the conditions shown in Figure 33, we filtered for the free available results ("Free full text").



TEXT AVAILABILITY

☐ Abstract
☑ Free full text
☐ Full text

ARTICLE ATTRIBUTE

☐ Associated data

ARTICLE TYPE

☐ Books and Documents
☐ Clinical Trial
☐ Meta-Analysis
☐ Randomized Controlled Trial
☑ Review
☐ Systematic Review

PUBLICATION DATE

○ 1 year
● 5 years
○ 10 years
○ Custom Range

Additional filters

*Filters applied: Free full text, Review, in the last 5 years. Clear all*

☐ 1 Cite Share

Genetics of **diabetes** mellitus and **diabetes** complications.
Cole JB, Florez JC.
Nat Rev Nephrol. 2020 Jul;16(7):377-390. doi: 10.1038/s41581-020-0278-5. Epub 2020 May 12.
PMID: 32398868 **Free PMC article.** Review.
**Diabetes** is one of the fastest growing diseases worldwide, projected to affect 693 million adults by 2045. ...The explosion of new genomic datasets, both in terms of biobanks and aggregation of worldwide cohorts, has more than doubled the number of genetic discoveries for ...

☐ 2 Cite Share

Emerging Targets in Type 2 **Diabetes** and Diabetic Complications.
Demir S, Nawroth PP, Herzig S, Ekim Üstünel B.
Adv Sci (Weinh). 2021 Sep;8(18):e2100275. doi: 10.1002/advs.202100275. Epub 2021 Jul 28.
PMID: 34319011 **Free PMC article.** Review.
Type 2 **diabetes** is a metabolic, chronic disorder characterized by insulin resistance and elevated blood glucose levels. ...Overall, the molecular mechanisms of how type 2 **diabetes** develops and leads to irreparable organ damage remain elusive. ...

☐ 3 Cite Share

From Pre-**Diabetes** to **Diabetes**: Diagnosis, Treatments and Translational Research.
Khan RMM, Chua ZJY, Tan JC, Yang Y, Liao Z, Zhao Y.
Medicina (Kaunas). 2019 Aug 29;55(9):546. doi: 10.3390/medicina55090546.
PMID: 31470636 **Free PMC article.** Review.
This unawareness and ignorance lead to further complications. Pre-**diabetes** is the preceding condition of **diabetes**, and in most of the cases, this ultimately leads to the development of **diabetes**. **Diabetes** can be classified into three types, namely type ...

☐ 4 Cite Share

Type 2 **diabetes**: a multifaceted disease.
Pearson ER.
Diabetologia. 2019 Jul;62(7):1107-1112. doi: 10.1007/s00125-019-4909-y. Epub 2019 Jun 3.
PMID: 31161345 **Free PMC article.** Review.
Type 2 **diabetes** is a complex disease usually diagnosed with little regard to aetiology. ...Beyond this, however, type 2 **diabetes** is a highly heterogeneous polygenic disease. This review outlines the recent developments that recognise this heterogeneity by deconvolut ...

☐ 5 Cite

Type 2 **Diabetes** Mellitus: A Review of Multi-Target Drugs.
Artasensi A, Pedretti A, Vistoli G, Fumagalli L.
Molecules. 2020 Apr 23;25(8):1987. doi: 10.3390/molecules25081987.

Figure 33: Set PubMed filter criteria

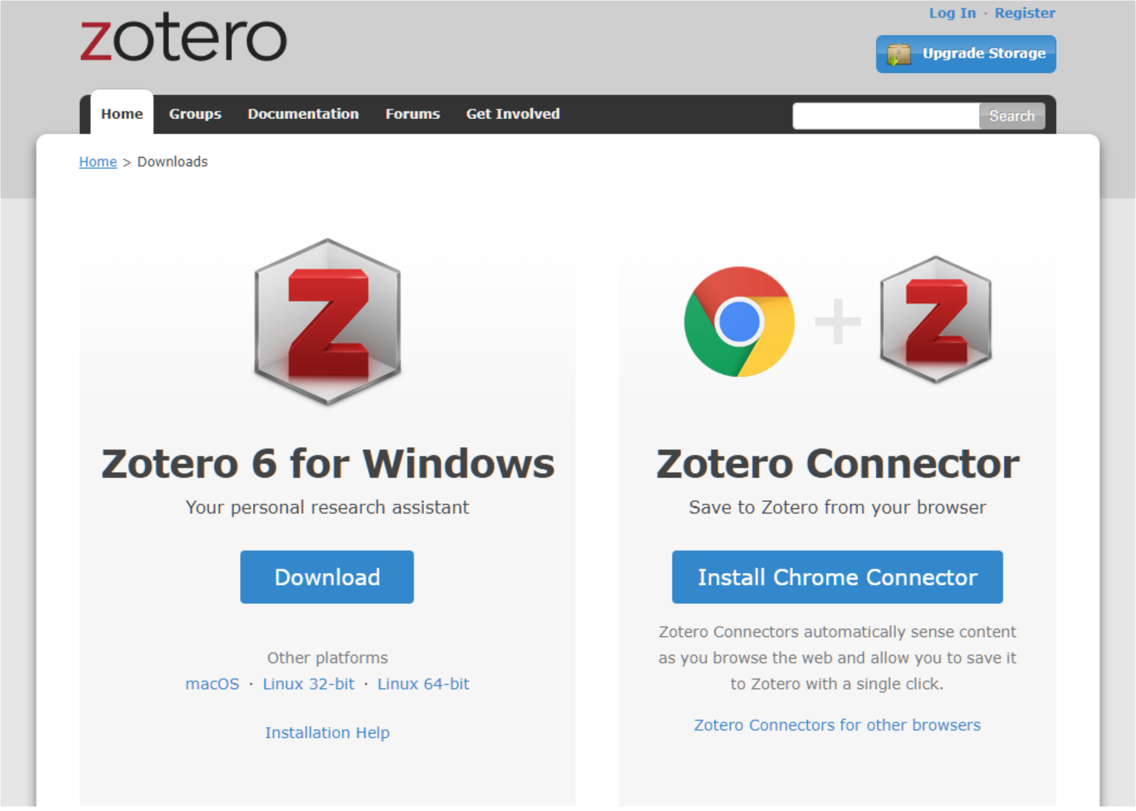
## Reference editor

References in the right format are essential for everything from project work to  theses and articles.

Most journals have their own rules for how to cite sources, but Harvard and Vancouver are the two most common styles. For the Vancouver type, we number the citations and indicate them in numerical order at the end. In Harvard style, the name(s) of the author(s) and the year of publication are given, and the cited references are listed in alphabetical order at the end of the work.

The form of reference for books, articles and websites is different in terms of the number of authors, publisher and date of access.

One of the most commonly used free reference editors is Zotero.



Zotero Connectors automatically sense content as you browse the web and allow you to save it to Zotero with a single click.

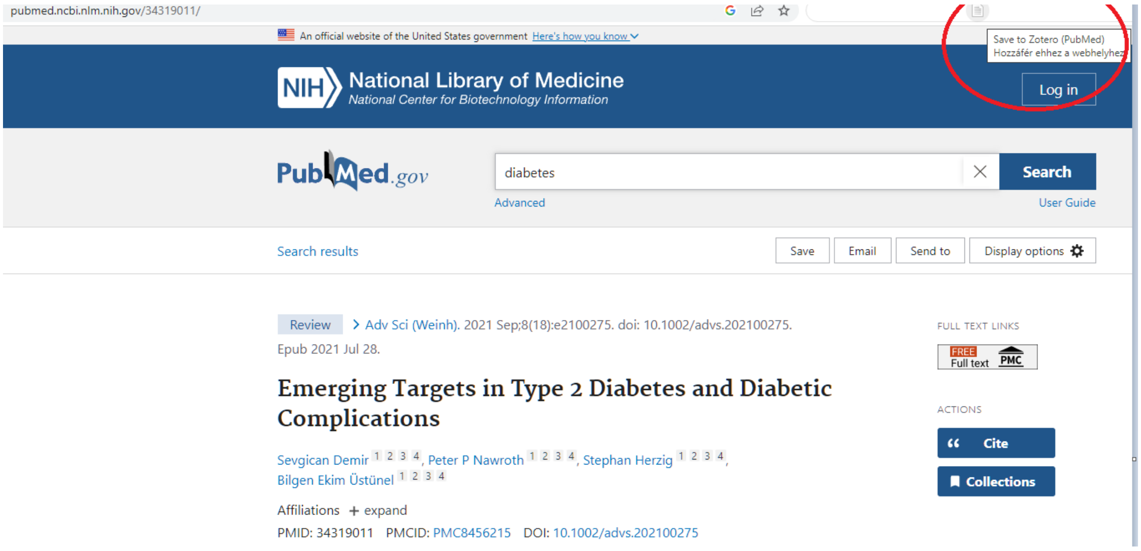
Figure 34: Download Zotero (https://www.zotero.org/download/)

We should install the main program and the Chrome Connector. The program integrates with both the browser and MS-Word.

Usage steps:

- Launch Zotero from the Start menu
- Pin browser extensions

For a useful article/page, we can save the link by clicking on the plugin, which is shown in Figure 35.



pubmed.ncbi.nlm.nih.gov/34319011/

An official website of the United States government Here's how you know

Save to Zotero (PubMed) Hozzáfér ehhez a webhelyhez

NIH National Library of Medicine National Center for Biotechnology Information

Log in

PubMed.gov

diabetes

Search

Advanced

User Guide

Search results

Save Email Send to Display options

Review > Adv Sci (Weinh). 2021 Sep;8(18):e2100275. doi: 10.1002/advs.202100275. Epub 2021 Jul 28.

Emerging Targets in Type 2 Diabetes and Diabetic Complications

Sevgican Demir 1 2 3 4, Peter P Nawroth 1 2 3 4, Stephan Herzig 1 2 3 4, Bilgen Ekim Üstünel 1 2 3 4

Affiliations + expand

PMID: 34319011 PMCID: PMC8456215 DOI: 10.1002/advs.202100275

FULL TEXT LINKS

FREE Full text PMC

ACTIONS

Cite

Collections

Figure 35: Save link to Zotero

After collecting the appropriate links, we can add them to Word (Add Citation) and create the Bibliography (Add Bibliography).

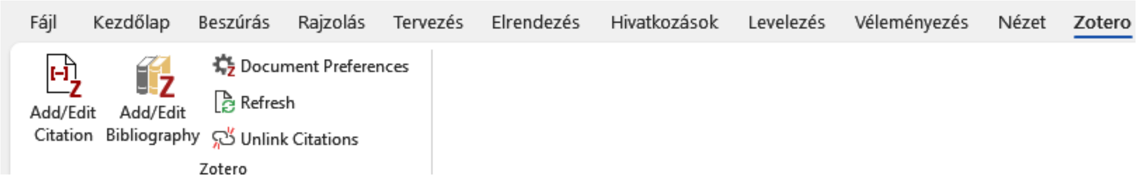
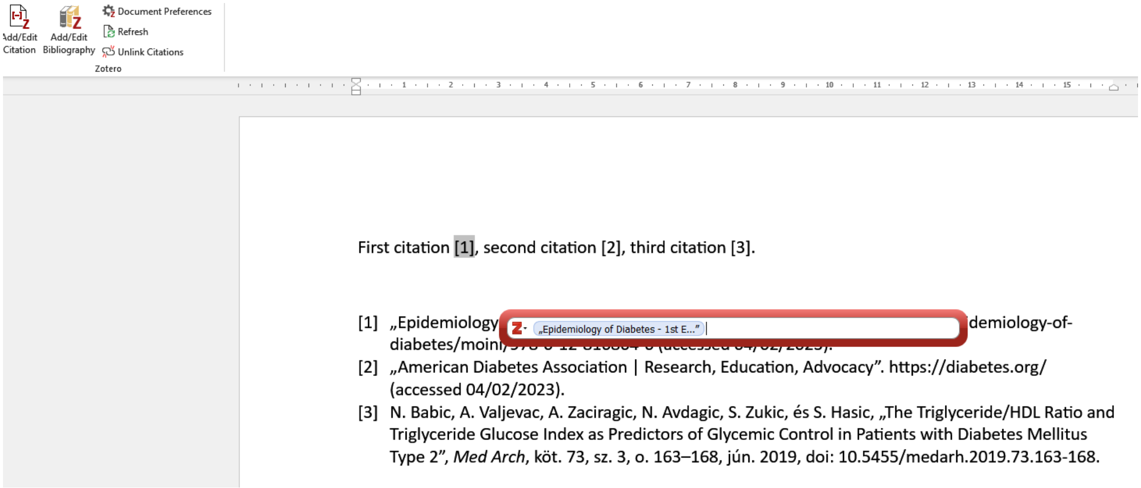


Figure 36: Zotero in Word

We can create a new link as well as edit the existing one by clicking on the "Add/Edit Citation" button, as shown in Figure 37.



First citation [1], second citation [2], third citation [3].

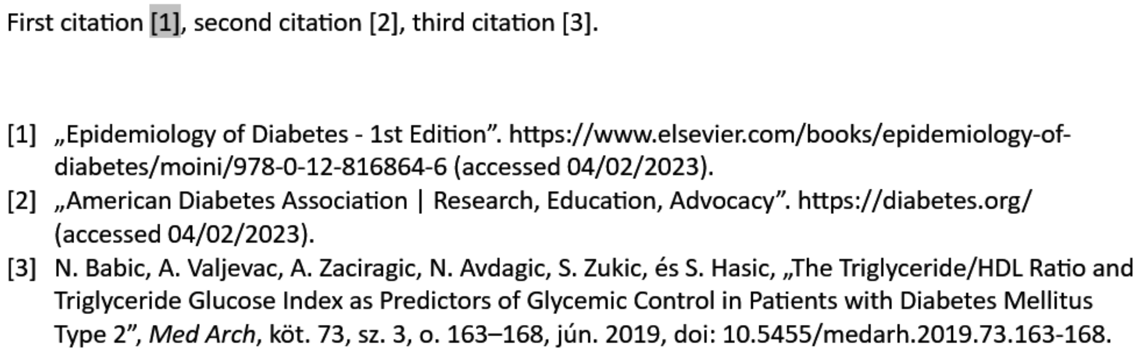
[1] „Epidemiology ... demiology-of-diabetes/moin...

[2] „American Diabetes Association | Research, Education, Advocacy". https://diabetes.org/ (accessed 04/02/2023).

[3] N. Babic, A. Valjevac, A. Zaciragic, N. Avdagic, S. Zukic, és S. Hasic, „The Triglyceride/HDL Ratio and Triglyceride Glucose Index as Predictors of Glycemic Control in Patients with Diabetes Mellitus Type 2", *Med Arch*, köt. 73, sz. 3, o. 163–168, jún. 2019, doi: 10.5455/medarh.2019.73.163-168.

Figure 37: Edit link with Zotero

The Bibliography is created using the "Add/Edit Bibliography" button.



First citation [1], second citation [2], third citation [3].

[1] „Epidemiology of Diabetes - 1st Edition". https://www.elsevier.com/books/epidemiology-of-diabetes/moini/978-0-12-816864-6 (accessed 04/02/2023).

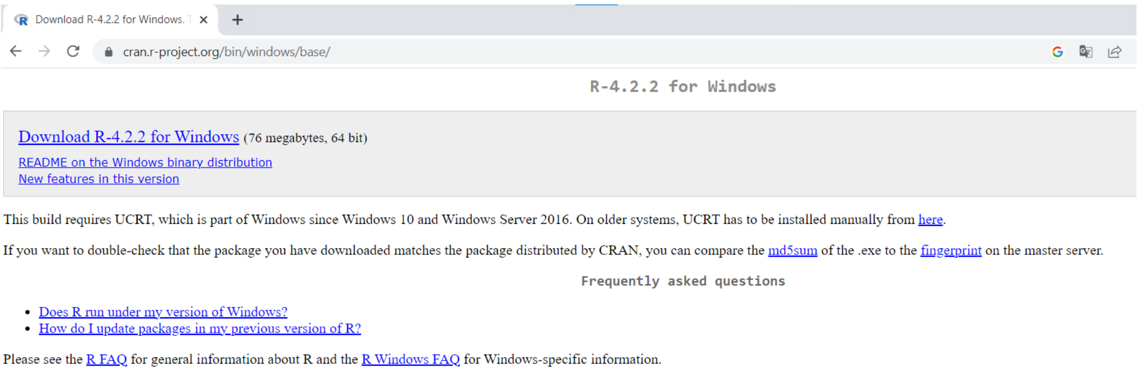
[2] „American Diabetes Association | Research, Education, Advocacy". https://diabetes.org/ (accessed 04/02/2023).

[3] N. Babic, A. Valjevac, A. Zaciragic, N. Avdagic, S. Zukic, és S. Hasic, „The Triglyceride/HDL Ratio and Triglyceride Glucose Index as Predictors of Glycemic Control in Patients with Diabetes Mellitus Type 2", *Med Arch*, köt. 73, sz. 3, o. 163–168, jún. 2019, doi: 10.5455/medarh.2019.73.163-168.

Figure 38: Bibliography by Zotero

## R program

The R program can be downloaded from the R-project website:


Download R-4.2.2 for Windows. T
cran.r-project.org/bin/windows/base/

R-4.2.2 for Windows

Download R-4.2.2 for Windows (76 megabytes, 64 bit)
README on the Windows binary distribution
New features in this version

This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server.

Frequently asked questions

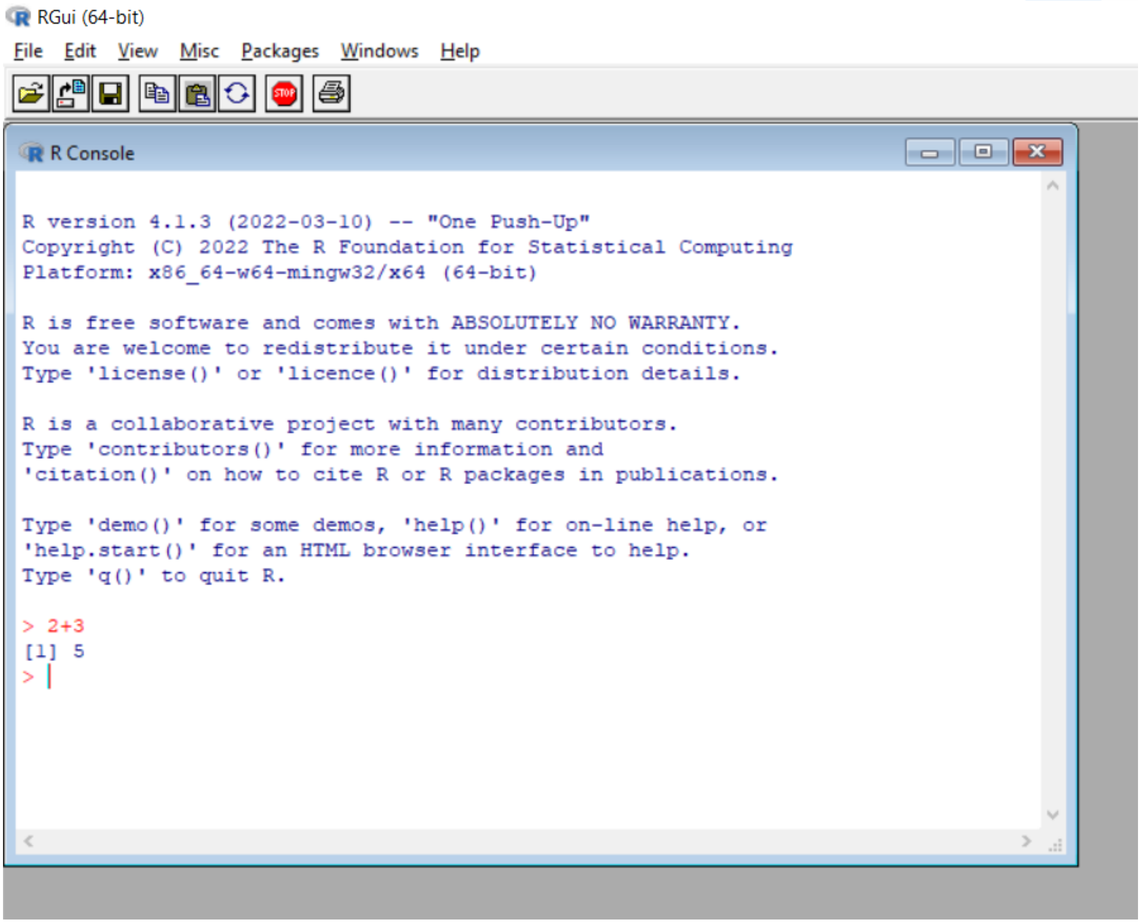
- Does R run under my version of Windows?
- How do I update packages in my previous version of R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Figure 39: Download R program

It is worth using the 64-bit version (x64) for better memory management.

After installation (preferably the default way), you can start the program. After entering commands, the result is immediately visible.

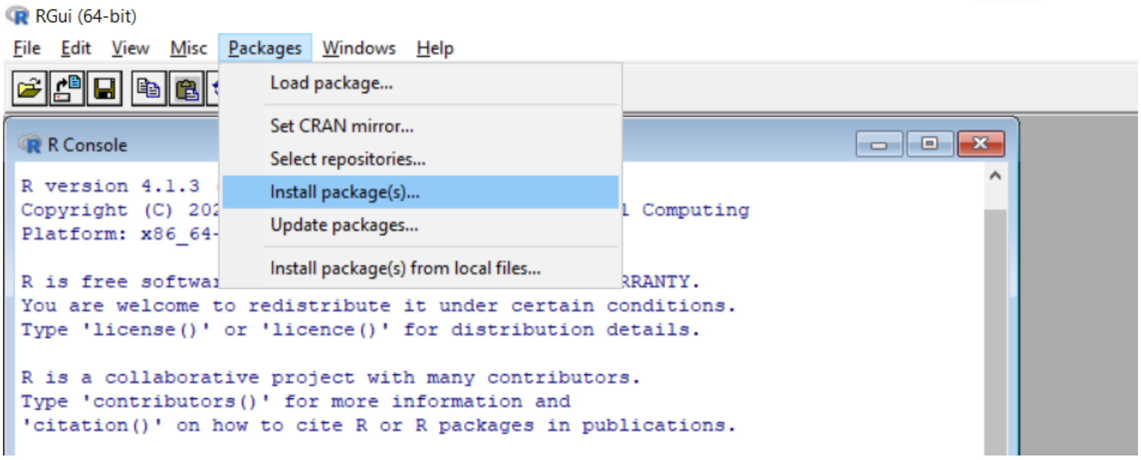


```
R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 2+3
[1] 5
> 
```

Figure 40: R program interface

Each package must be installed (install.package) and then loaded (library) to be available. The package installation can be done by command line or by clicking from the menu. The menu and the command could be seen below:

*install.packages("pwr")*



```
R version 4.1.3 (
Copyright (C) 202                           l Computing
Platform: x86_64-

R is free softwa                            RRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

Figure 41: Installing R package

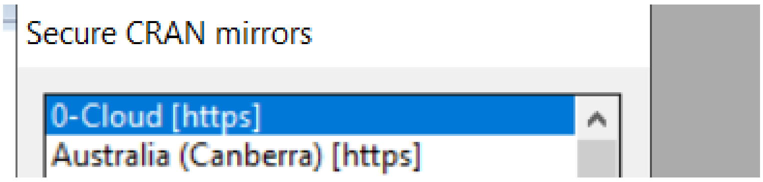


Figure 42: Choosing an R mirror server for installation

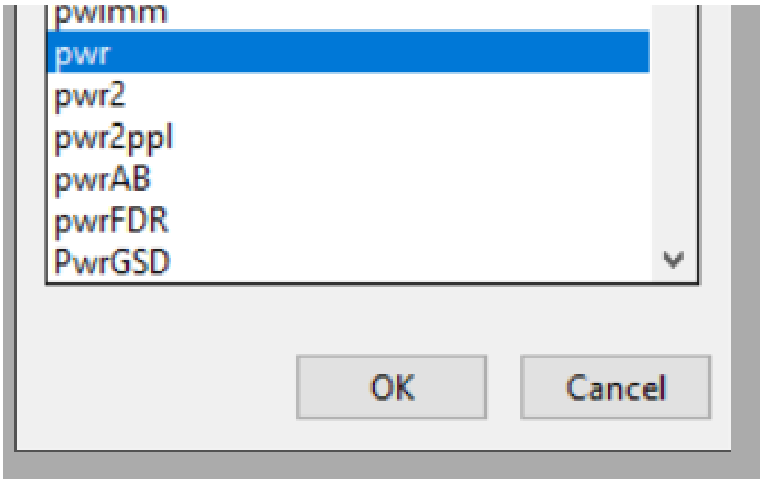


Figure 43: Installing the "PWR" package

To load the installed package:
*library(pwr)*

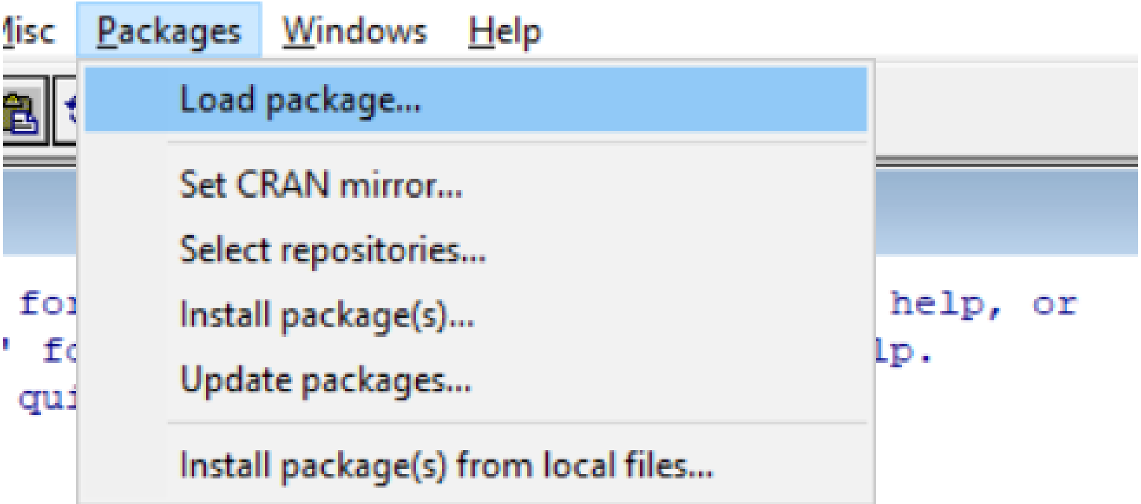


Figure 44: Loading R package

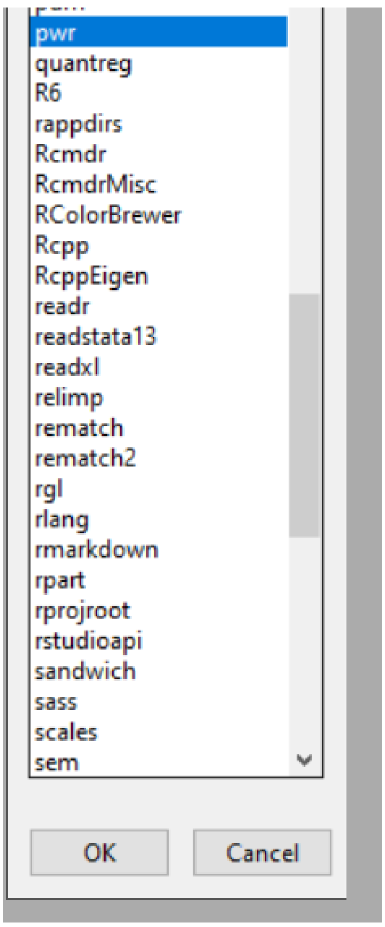
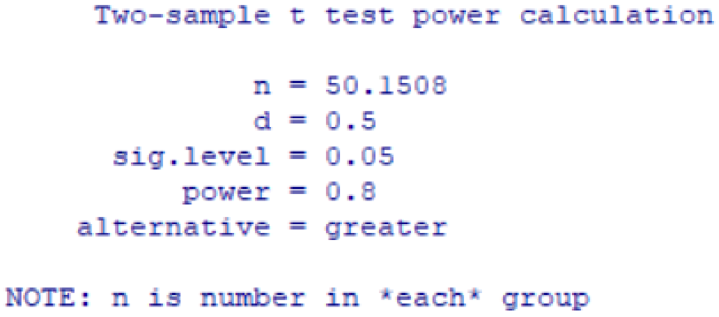


Figure 45: Loading package "PWR"

The install.packages command requires the quotation mark, but the library command prohibits it.

## Sample size estimation

After installing and loading the package, we can see, for example, the number of sample size needed to detect a difference of 0.5 with a p-value of 0.05 and a statistical power of 80% (complement of the type II error, the larger the better, at least 80%) for a two-sample t-test.
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="greater")



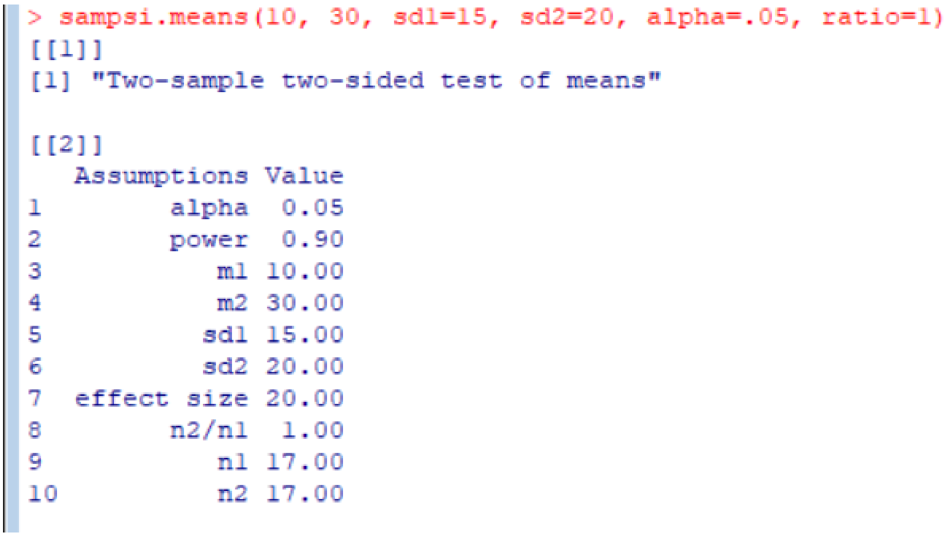
```
     Two-sample t test power calculation

              n = 50.1508
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = greater

NOTE: n is number in *each* group
```

Figure 46: Sample size estimation

Based on the estimation, at least 50 people per group are needed. You can also create your own functions. After typing the code below, a new command becomes available, that allows for a more accurate estimation of the number of sample size by using averages and standard deviations (which we know about, for example from PubMed).

```
sampsi.means<-function(m1, m2, sd1, sd2=NA, ratio=1, power=.90, alpha=.05, two.sided=TRUE, one.sample=FALSE){
effect.size<-abs(m2-m1)
sd2<-ifelse(!is.na(sd2), sd2, sd1)
z.pow<-qt(1-power, df=Inf, lower.tail=FALSE)
z.alph<-ifelse(two.sided==TRUE, qt(alpha/2, df=Inf, lower.tail=FALSE), qt(alpha, df=Inf, lower.tail=FALSE))
ct<-(z.pow+z.alph)
n1<-(sd1^2+(sd2^2)/ratio)*(ct)^2/(effect.size^2)
n<-(ct*sd1/effect.size)^2
if(one.sample==FALSE){
col1<-c("alpha", "power", "m1", "m2", "sd1", "sd2", "effect size", "n2/n1", "n1", "n2")
col2<-c(alpha,  power, m1, m2, sd1, sd2, effect.size, ratio, ceiling(n1), ceiling(n1*ratio))
}
else{
col1<-c("alpha", "power", "null", "alternative", "n")
col2<-c(alpha, power, m1, m2, ceiling(n))
}
ret<-as.data.frame(cbind(col1, col2))
ret$col2<-as.numeric(as.character(ret$col2))
colnames(ret)<-c("Assumptions", "Value")
description<-paste(ifelse(one.sample==FALSE, "Two-sample", "One-sample"), ifelse(two.sided==TRUE, "two-sided", "one-sided"), "test of means")
retlist<-list(description, ret)
return(retlist)
}
```

Then you can type the following command:
sampsi.means(10, 30, sd1=15, sd2=20, alpha=.05, ratio=1)



```
> sampsi.means(10, 30, sd1=15, sd2=20, alpha=.05, ratio=1)
[[1]]
[1] "Two-sample two-sided test of means"

[[2]]
   Assumptions Value
1        alpha  0.05
2        power  0.90
3           m1 10.00
4           m2 30.00
5          sd1 15.00
6          sd2 20.00
7  effect size 20.00
8        n2/n1  1.00
9           n1 17.00
10          n2 17.00
```

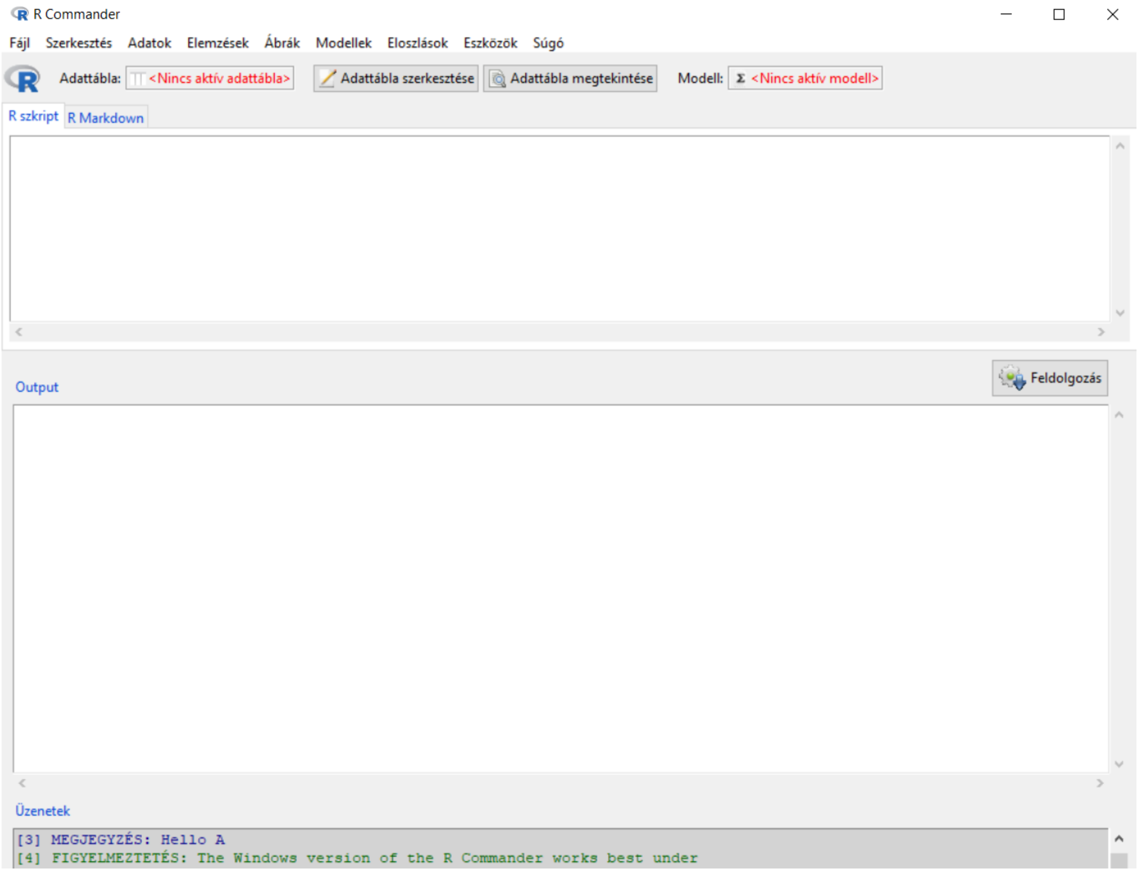
Figure 47: Result of sample size estimation

That is, for groups of 10 ± 15 (mean ± standard deviation) and 30 ± 20, the difference will be significant (p=0.05) if we work with a minimum of 17 people per group.

## Descriptive statistics

Categorical variables (e.g., gender or level of education) are characterized by percentage distribution. Continuous variables are characterized (for normal distribution) using the mean ± standard deviation, or median and interquartile range (for non-normal distribution).
To create descriptive statistics, the first step is to install and load the Rcommander (Rcmdr) package with a graphical interface.

```
install.packages("Rcmdr")
library(Rcmdr)
```



```
[3] MEGJEGYZÉS: Hello A
[4] FIGYELMEZTETÉS: The Windows version of the R Commander works best under
```

Figure 48: R commander

Once opened, we have the option to import a database. At the end of our note, you can find (and copy) the database used, which we will read from a *sample.csv*.

Other languages can also be downloaded, but due to the use of technical terms, it is worth using the English version.

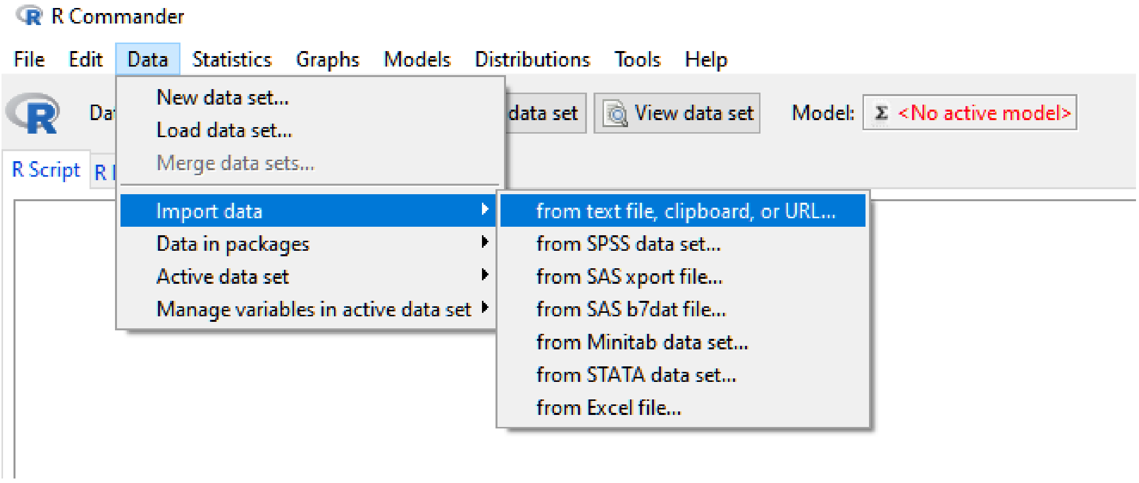
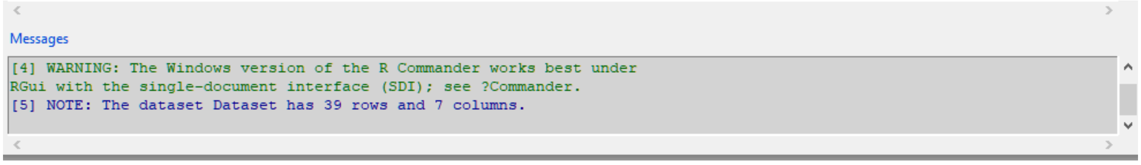


Figure 49: Importing file using R commander

We can leave everything as it is, especially the field separator, which in this case is the semicolon.

If loaded successfully, it prints the row and column numbers of the database.



```
[4] WARNING: The Windows version of the R Commander works best under
RGui with the single-document interface (SDI); see ?Commander.
[5] NOTE: The dataset Dataset has 39 rows and 7 columns.
```

Figure 50: Successful data retrieval

We have the option to edit or simply view the database.

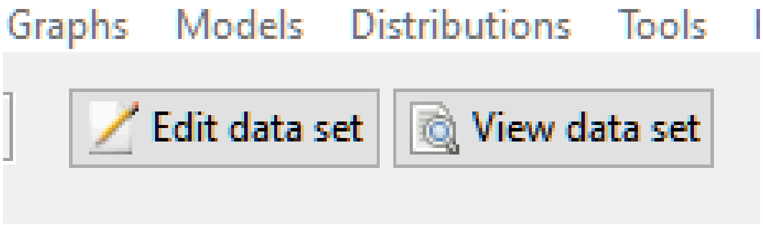
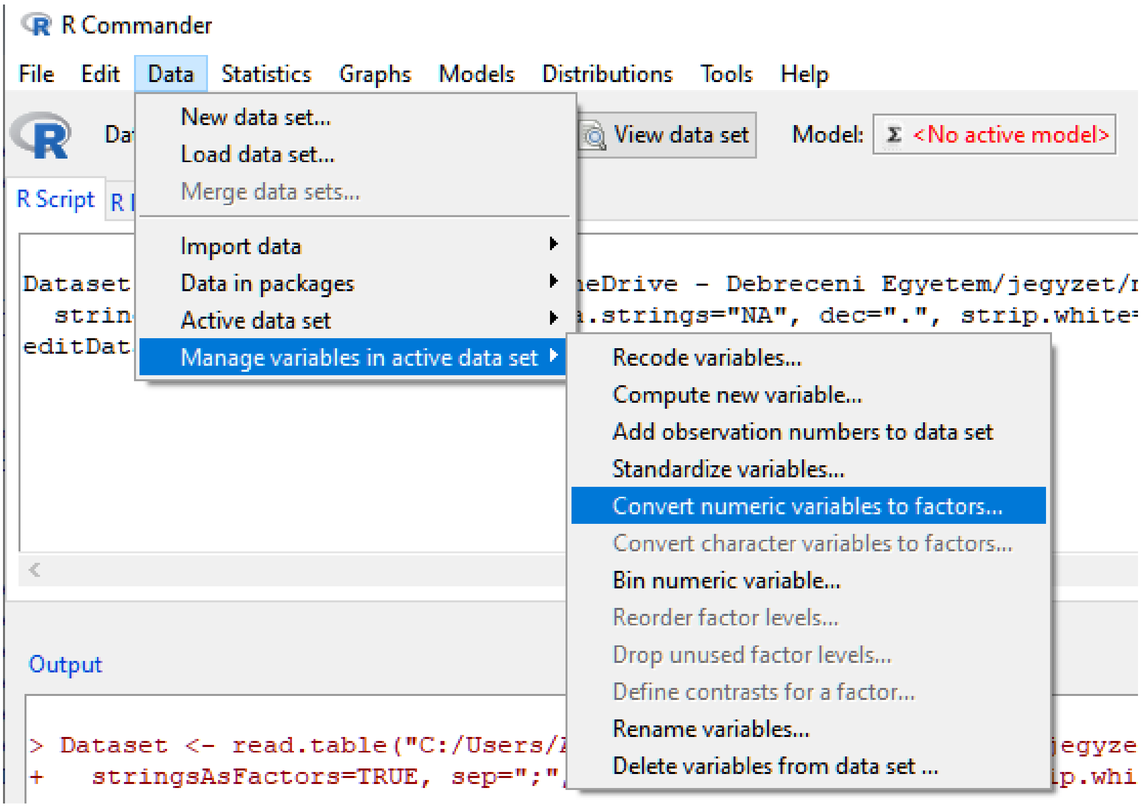


Figure 51: View/edit a database

As a first step, it is worth converting our categorical variables into factors.

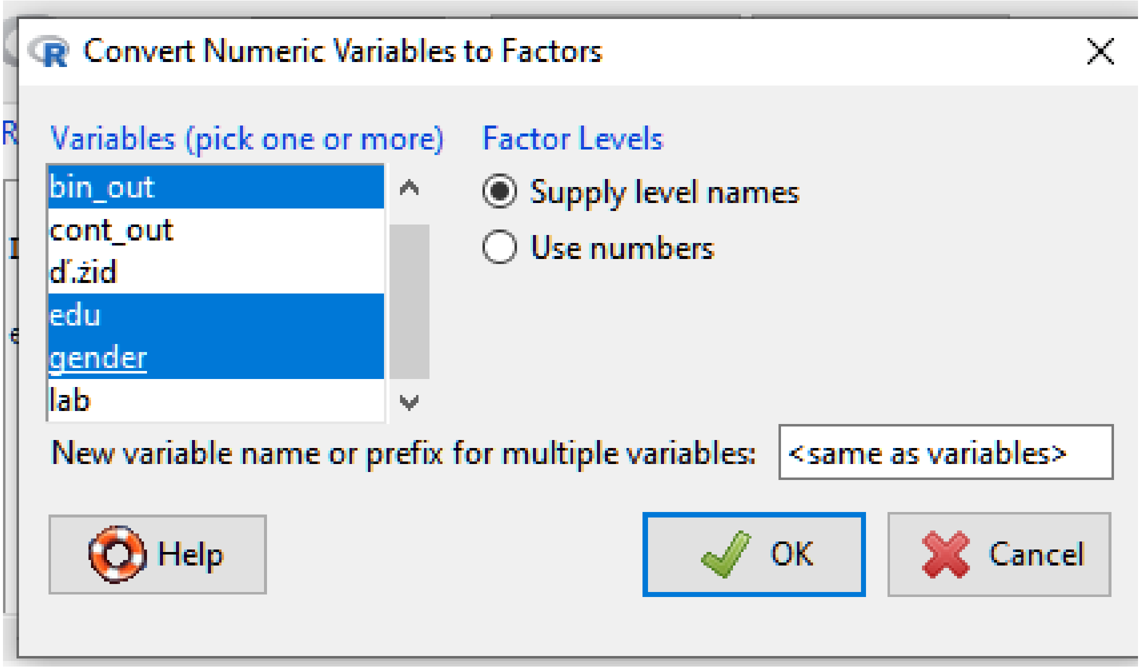


```
Dataset                                    eDrive - Debreceni Egyetem/jegyzet/r
  strin                                    .strings="NA", dec=".", strip.white=
editDat
```

```
> Dataset <- read.table("C:/Users/                                  jegyze
+   stringsAsFactors=TRUE, sep=";",                                 ip.whi
```

Figure 52: Converting variables to factors

Ctrl-click to select multiple variables at once.



Convert Numeric Variables to Factors

Variables (pick one or more)

- bin_out
- cont_out
- d'.žid
- edu
- gender
- lab

Factor Levels

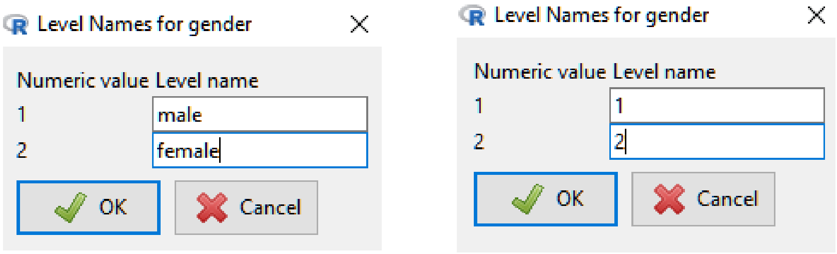
● Supply level names

○ Use numbers

New variable name or prefix for multiple variables: <same as variables>

Figure 53: Selecting multiple variables at once

Existing variables can be overwritten. After pressing the "OK" button, we must enter the labels for the values (label) or simply the value, without relabeling.



Level Names for gender

| Numeric value | Level name |
|---|---|
| 1 | male |
| 2 | female |

Level Names for gender

| Numeric value | Level name |
|---|---|
| 1 | 1 |
| 2 | 2 |

Figure 54: Relabeling variables

We can ask for a general descriptive characterization of all variables.

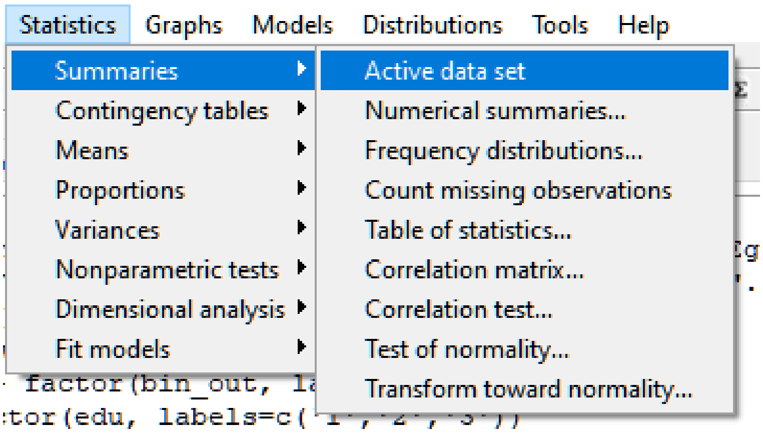
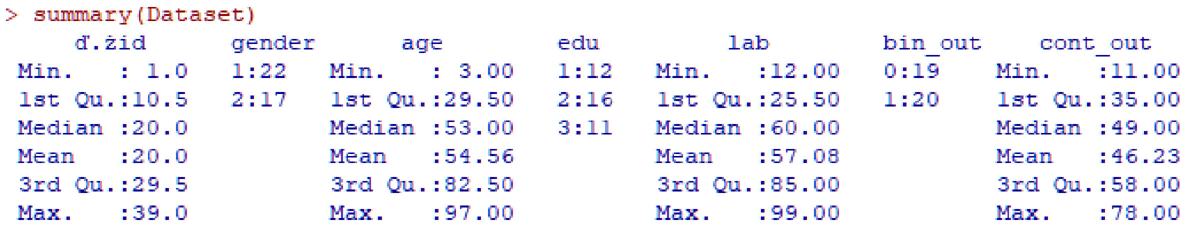


Figure 55: Query descriptive statistics

In this case, everything is treated as a continuous variable and the following parameters of the variables are given:

- Minimum
- First quartile (Q1), which is the 25% percentile value
- Median (Q2), or 50% percentile
- Third quartile (Q3), or 75% percentile value
- Maximum



```
> summary(Dataset)
     ď.žid       gender      age            edu        lab          bin_out     cont_out    
 Min.   : 1.0   1:22    Min.   : 3.00   1:12   Min.   :12.00   0:19    Min.   :11.00  
 1st Qu.:10.5   2:17    1st Qu.:29.50   2:16   1st Qu.:25.50   1:20    1st Qu.:35.00  
 Median :20.0           Median :53.00   3:11   Median :60.00           Median :49.00  
 Mean   :20.0           Mean   :54.56          Mean   :57.08           Mean   :46.23  
 3rd Qu.:29.5           3rd Qu.:82.50          3rd Qu.:85.00           3rd Qu.:58.00  
 Max.   :39.0           Max.   :97.00          Max.   :99.00           Max.   :78.00  
```

Figure 56: Descriptive statistics

We can also use commands (*summary*), which allows instructions to be executed without the graphical interface (GUI), i.e., Rcommander.

Continuous variables can be characterized separately by "Numerical Summaries":

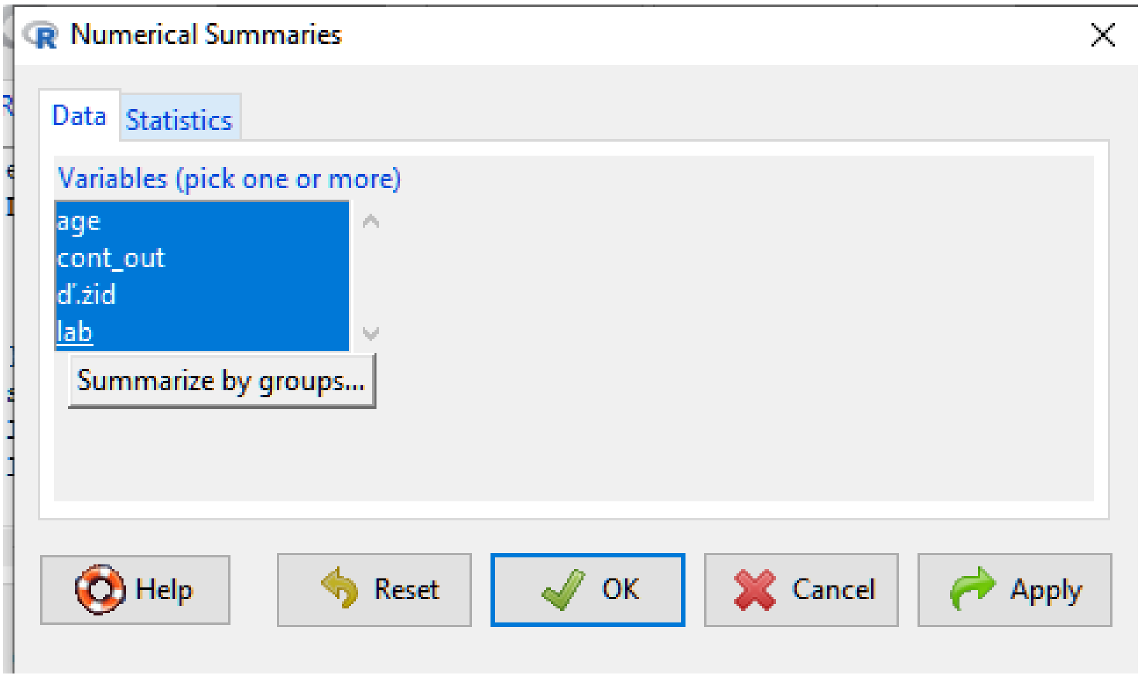
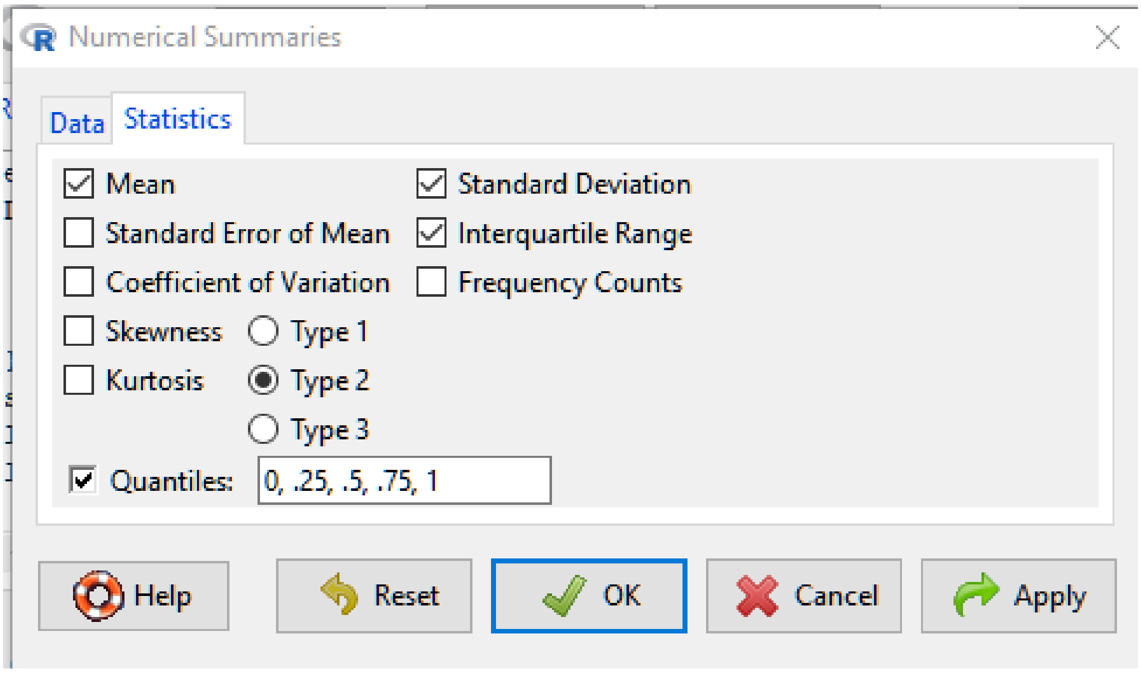


Figure 57: Selecting continuous variables for characterization

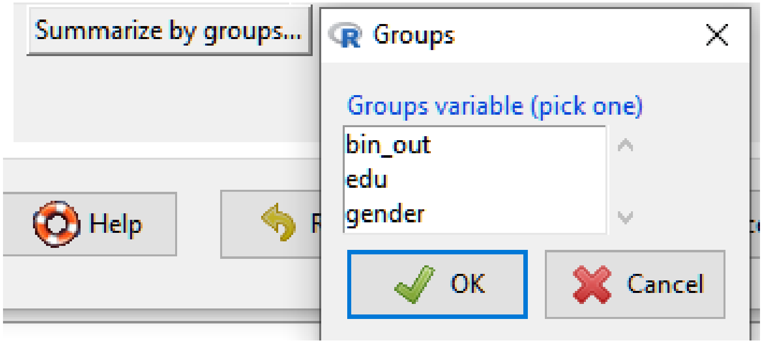
It is also possible to make multiple designations, as well as to select statistics that are interesting to us.



☑ Mean ☑ Standard Deviation
□ Standard Error of Mean ☑ Interquartile Range
□ Coefficient of Variation □ Frequency Counts
□ Skewness ○ Type 1
□ Kurtosis ● Type 2
○ Type 3
☑ Quantiles: 0, .25, .5, .75, 1

Figure 58: Setting up statistics

It is also possible to display layer by layer (as shown in Figure 59) along a given grouping variable (Summarize by groups).



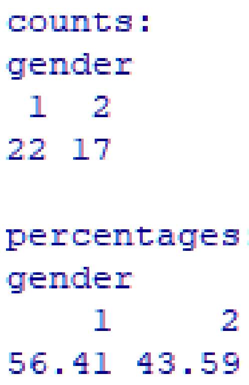
Groups variable (pick one)
bin_out
edu
gender

Figure 59: Stratified analysis by group

The characteristics shown are as follows:

- Mean
- Standard deviation (SD)
- Interquartile range (IQR)
- Percentiles
  - 0% (minimum)
  - 25% (Q1)
  - 50% (Q2), or median
  - 75% (Q3)
  - 100% (maximum)
  - sample size (N)

Categorical variables (after conversion to factors) are characterized by percentage distribution:



```
counts:
gender
 1  2 
22 17 

percentages:
gender
    1     2 
56.41 43.59 
```

Figure 60: Percentage distribution of categorical variables

The output table contains the absolute numbers and percentages.

For continuous variables, the distribution of data (normality) is also important. One of the main reasons is the description of the variable. If the distribution is normal (or does not deviate significantly from it, p>0.05) then we use the mean ± standard deviation. We usually round the numbers to two decimal places. The p-value is rounded to three decimal places. For non-normal distributions, median and IQR (interquartile range; [Q1-Q3]) characterizations are used.

One of the most commonly used methods for testing normality is the Shapiro-Wilk test.

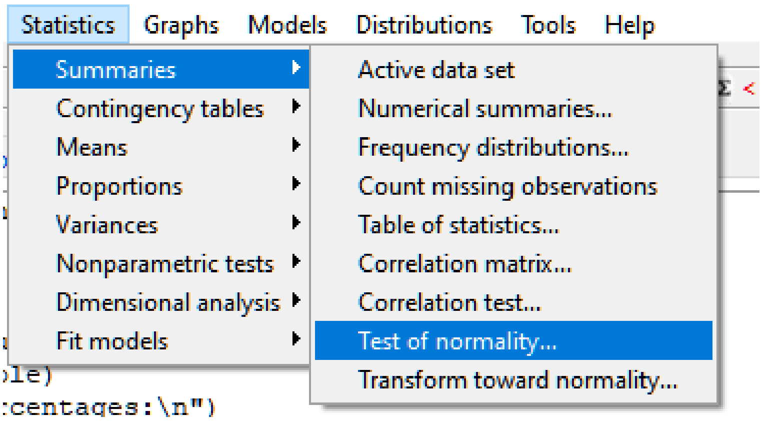
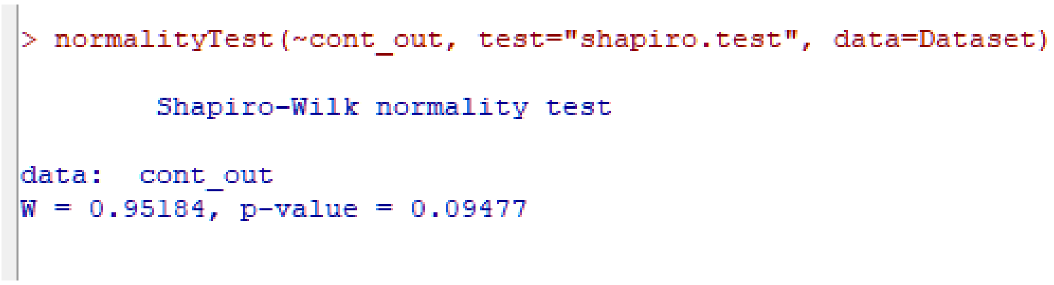


Figure 61: Normality testing



```
> normalityTest(~cont_out, test="shapiro.test", data=Dataset)

        Shapiro-Wilk normality test

data:  cont_out
W = 0.95184, p-value = 0.09477
```

Figure 62: Normality test result

Based on our results, it can be stated that the continuous outcome (cont_out) follows a normal distribution and does not differ significantly from it (p=0.095).

## Simple analysis

Descriptive analysis is followed by simple analysis. We study the association between an outcome (outcome, or dependent variable) and a potential influencing factor (explanatory or independent variable). With the help of crude analyses, a hypothesis can be generated, which will be evaluated using adjusted (multiple) analyses.

### Parametric tests

Parametric tests have greater statistical power, but certain conditions are required for their application. The most common condition is the normal (or near-normal) distribution of a continuous variable. The other most common condition is a sufficiently large number of sample size.

### *T-test*

The Student's t-test is suitable for comparing group averages. It is possible to compare the group mean to a constant number (constant) (Single-sample t-test; One-Sample t-test), and two average values of a group can be compared (before-after; Paired t-test), while the average of two independent groups (not necessarily with identical sample sizes) can also be compared (Two-sample t-test, Independent samples t-test).

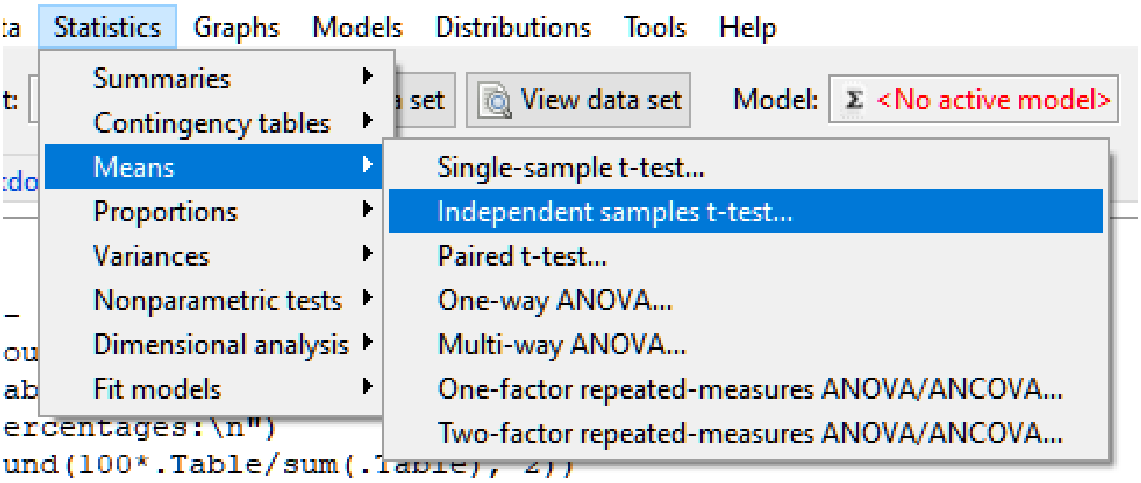


Figure 63: Two-sample t-test

For example, we can compare whether there is a difference in the continuous outcome (*cont_out*) by gender (*gender*).

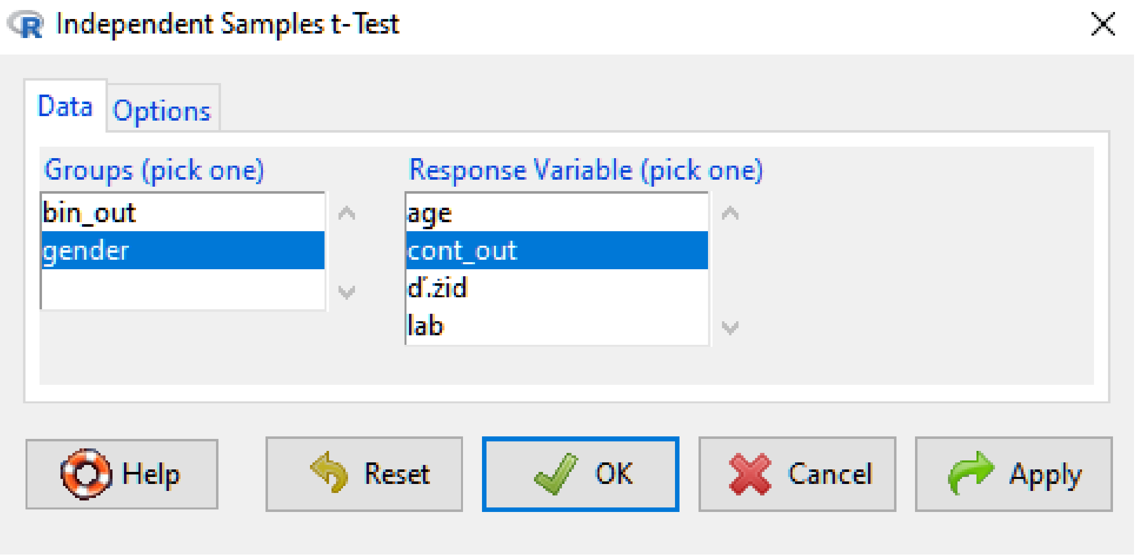
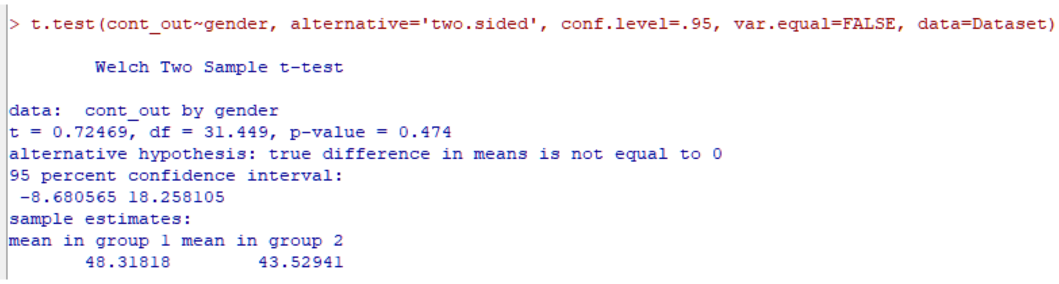


Figure 64: Two-sample t-test setting

Based on the result, we can say that there is no significant difference between the two group averages (48.32 vs. 43.53) (p=0.474) (the null hypothesis that there is no difference between the group averages remains in force).

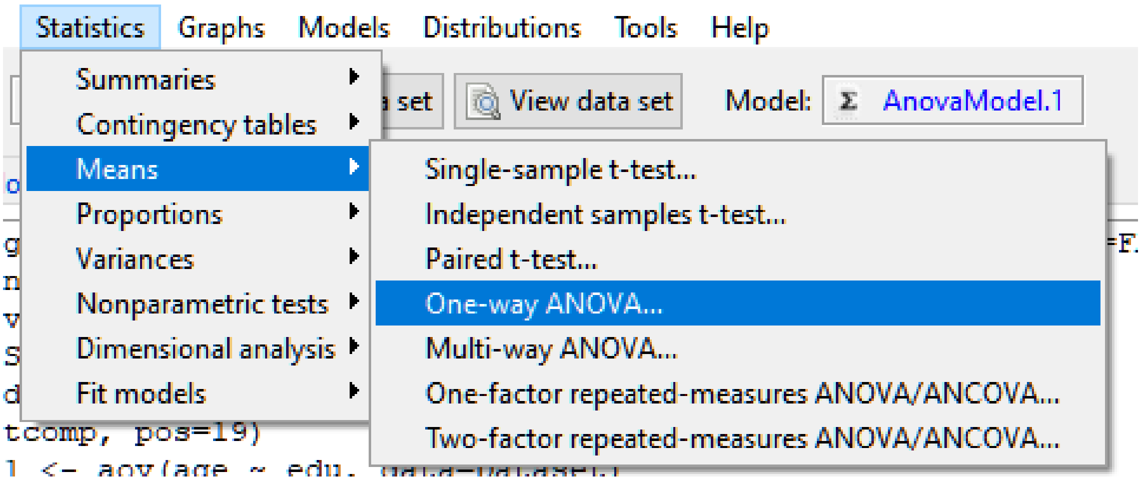


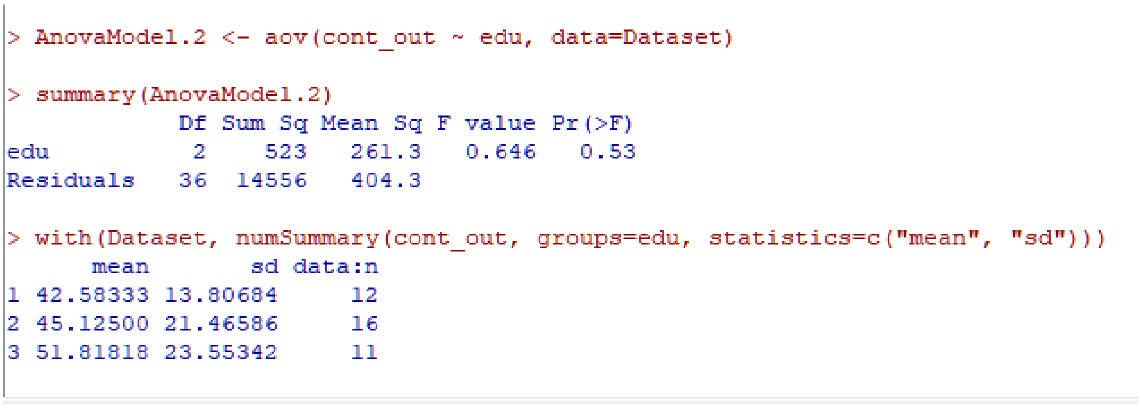
```
> t.test(cont_out~gender, alternative='two.sided', conf.level=.95, var.equal=FALSE, data=Dataset)

	Welch Two Sample t-test

data:  cont_out by gender
t = 0.72469, df = 31.449, p-value = 0.474
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.680565 18.258105
sample estimates:
mean in group 1 mean in group 2 
       48.31818        43.52941 
```

Figure 65: Two-sample t-test result

### *ANOVA*

ANOVA (Analysis of variance) is used when comparing more than two group averages. The name is a bit misleading, but the variances are only counted in the background by the test, in fact it is suitable for comparing averages.





```
> AnovaModel.2 <- aov(cont_out ~ edu, data=Dataset)

> summary(AnovaModel.2)
            Df Sum Sq Mean Sq F value Pr(>F)
edu          2    523   261.3   0.646   0.53
Residuals   36  14556   404.3

> with(Dataset, numSummary(cont_out, groups=edu, statistics=c("mean", "sd")))
      mean       sd data:n
1 42.58333 13.80684     12
2 45.12500 21.46586     16
3 51.81818 23.55342     11
```

Figure 66: ANOVA setup

Based on the result, it can be stated that there is no difference between the average values of continuous outcomes in terms of educational attainment (p=0.530).

### *Chi-squared test*

Associations between categorical variables can be detected by Pearson's chi-squared test. Also, the sample size is an important criterion for applicability. If there are not at least 5 observations in at least 80 % of the cells, it cannot be applied (Fisher's exact test is used instead).

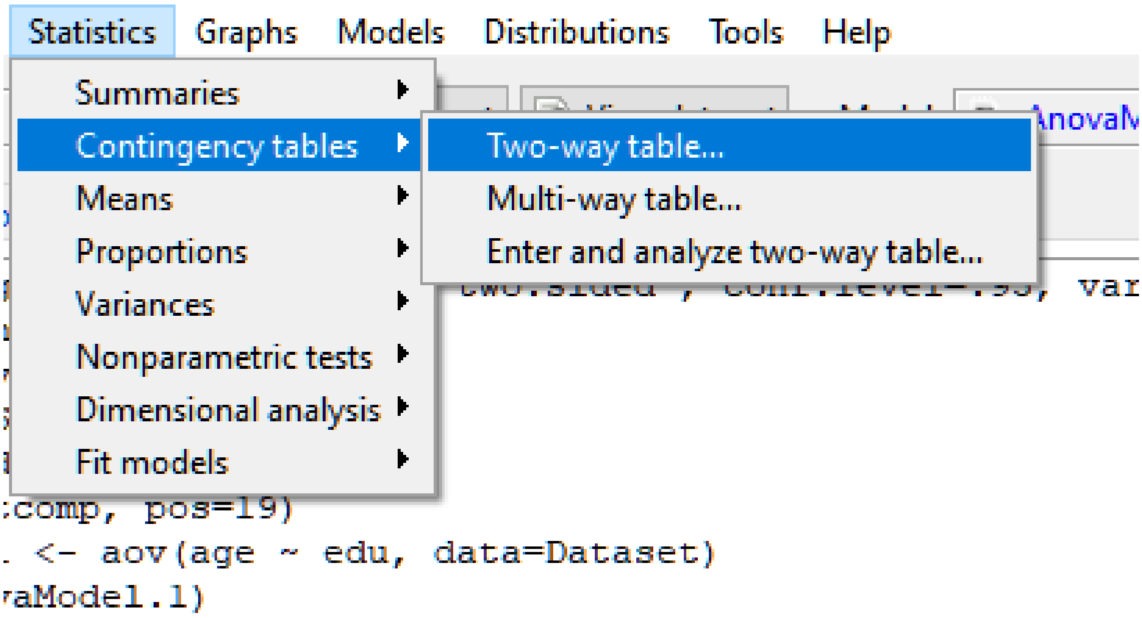
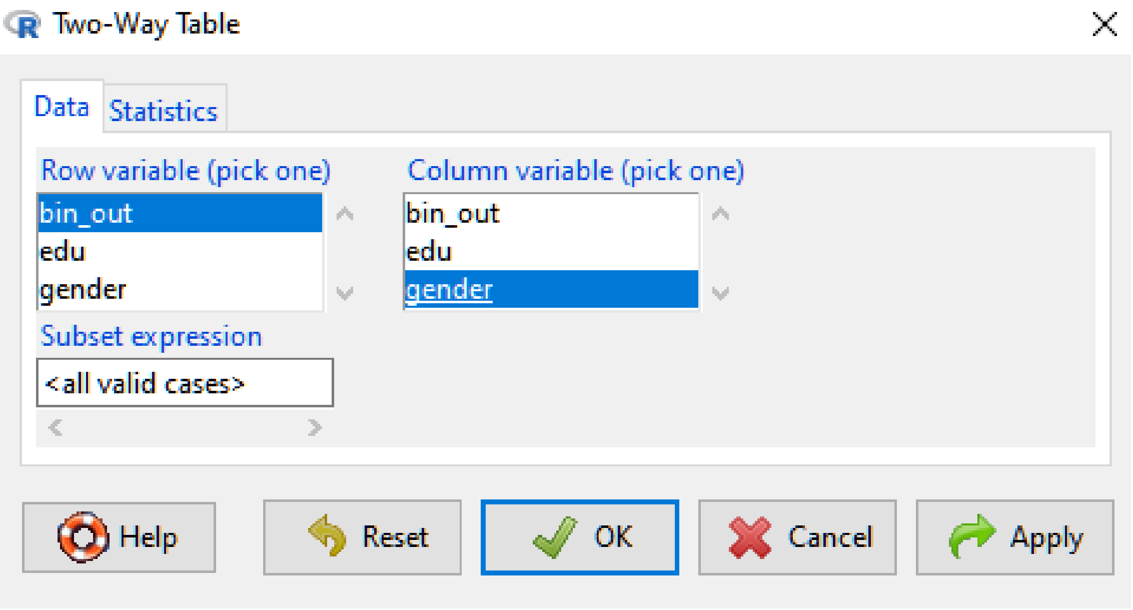


Figure 67: Choosing a Chi-squared test

Let's examine whether our binary (two categories, yes-no) outcome (*bin_out*) is related to gender.



Two-Way Table

Data | Statistics

Row variable (pick one)

- bin_out
- edu
- gender

Column variable (pick one)

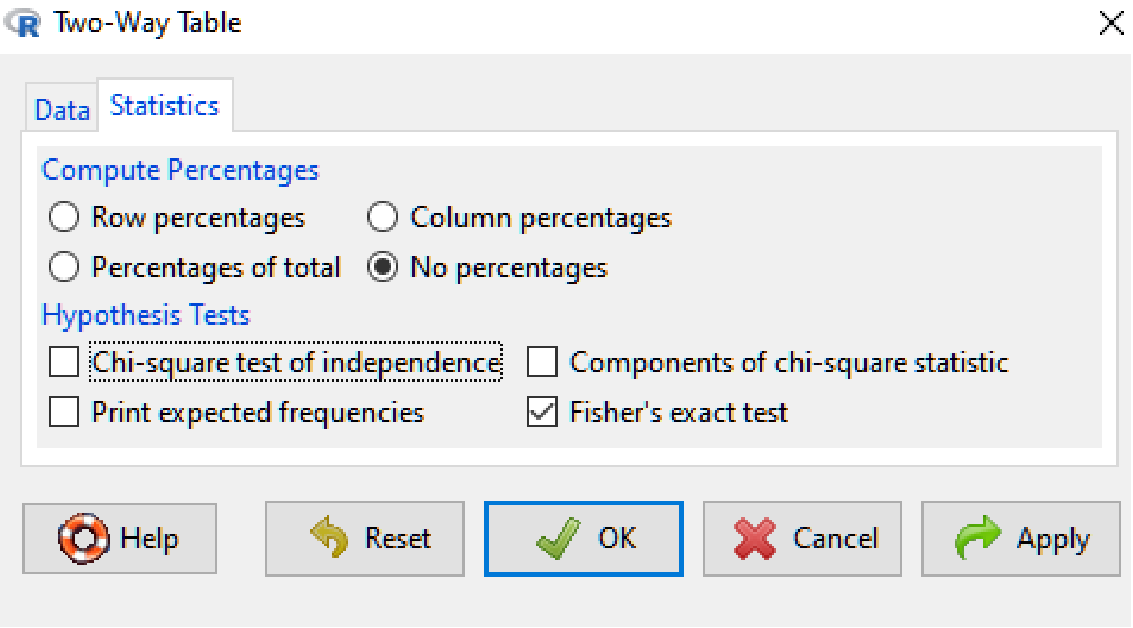
- bin_out
- edu
- gender

Subset expression

<all valid cases>

Figure 68: Chi-squared test setup

In the Statistics tab, you can also request Fisher's exact test (now stick to the default Chi-squared) or e.g. row percentages.



Two-Way Table

Data | Statistics

Compute Percentages

○ Row percentages ○ Column percentages

○ Percentages of total ● No percentages

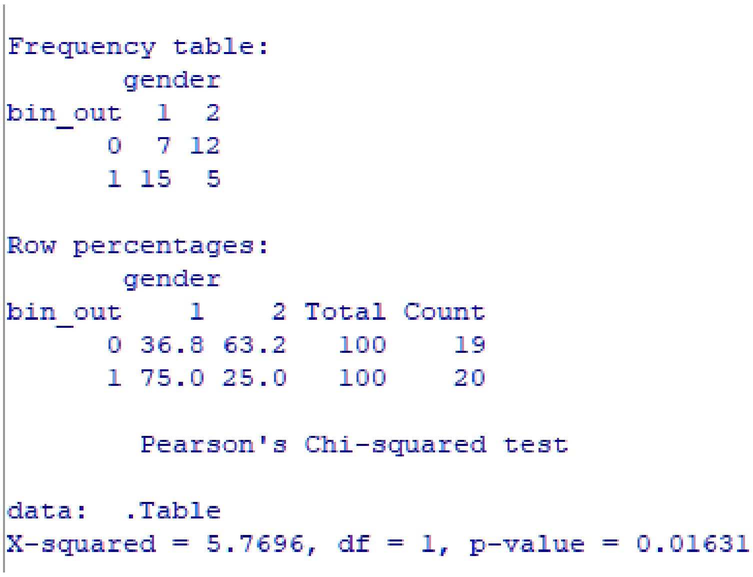
Hypothesis Tests

□ Chi-square test of independence □ Components of chi-square statistic

□ Print expected frequencies ☑ Fisher's exact test

Figure 69: Selection of Fisher's exact test

Based on our results, it can be said that there is a significant relationship (p=0.016) between gender and our binary output. On the basis of row percentages, we can see that there are more men among those with outcomes, while women are more numerous among those without outcomes.



```
Frequency table:
       gender
bin_out  1  2
      0  7 12
      1 15  5

Row percentages:
       gender
bin_out    1    2 Total Count
      0 36.8 63.2   100    19
      1 75.0 25.0   100    20

        Pearson's Chi-squared test

data:  .Table
X-squared = 5.7696, df = 1, p-value = 0.01631
```

Figure 70: Chi-squared test result

## Nonparametric tests

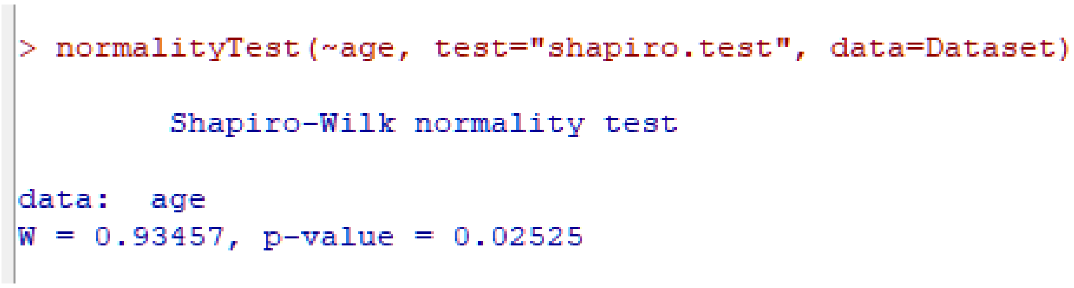
In the absence of assumptions, we have to use the statistically weaker non-parametric tests.

### *Wilcoxon test*

An alternative to t-tests. Instead of an average, it compares medians (more precisely, distributions, even with the same medians there may be significant differences).

- Single-sample t-test alternative
  - Single sample Wilcoxon test
- Paired t-test alternative
  - Wilcoxon signed-rank test/Paired samples Wilcoxon test
- Two-sample t-test alternative
  - Wilcoxon signed rank sum test/Mann–Whitney U test/Wilcoxon rank sum test

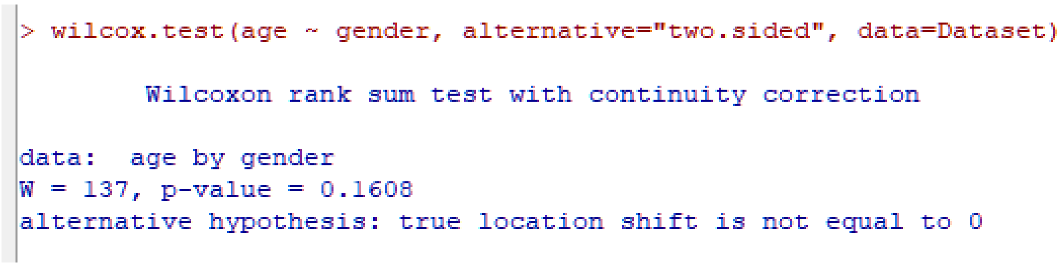
In our sample, the age *variable does* not follow a normal distribution (p=0.025), which justifies the use of a nonparametric test.



```
> normalityTest(~age, test="shapiro.test", data=Dataset)

        Shapiro-Wilk normality test

data:  age
W = 0.93457, p-value = 0.02525
```

Figure 71: Normality testing

Let's see if there is a difference between the medians (distributions) in terms of gender.



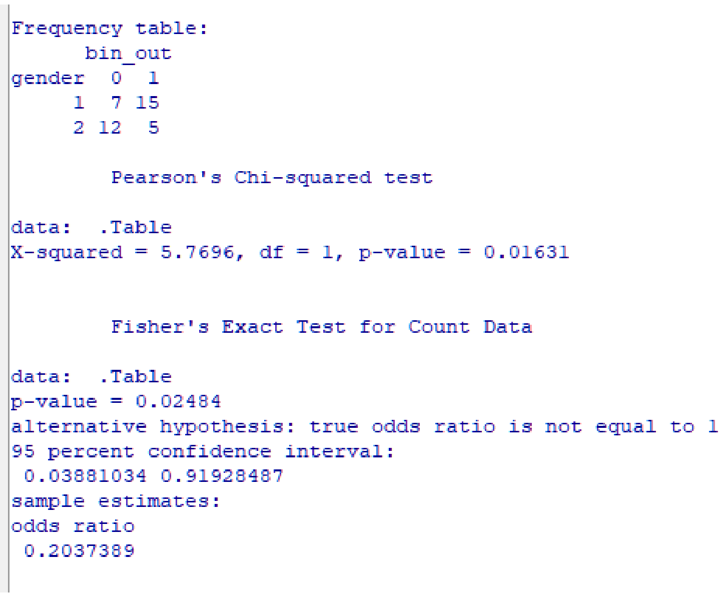
```
> wilcox.test(age ~ gender, alternative="two.sided", data=Dataset)

        Wilcoxon rank sum test with continuity correction

data:  age by gender
W = 137, p-value = 0.1608
alternative hypothesis: true location shift is not equal to 0
```

Figure 72: Mann-Whitney-Wilcoxon test

Based on the test result, there is no difference (p=0.161) between the two medians (distributions).

### *Fisher's exact test*

Suitable for comparing categorical variables with a smaller number of sample sizes. Sticking to the former example (row and column interchangeable, the p-value remains), the sample size was large enough for the Chi-squares test, each cell had at least 5 observations. If there had been e.g., 4 people in at least one of the cells, then 25% of all cells would not have met the condition, so there would not have been the minimum of 5 people in 80% of the cells (only 75%).



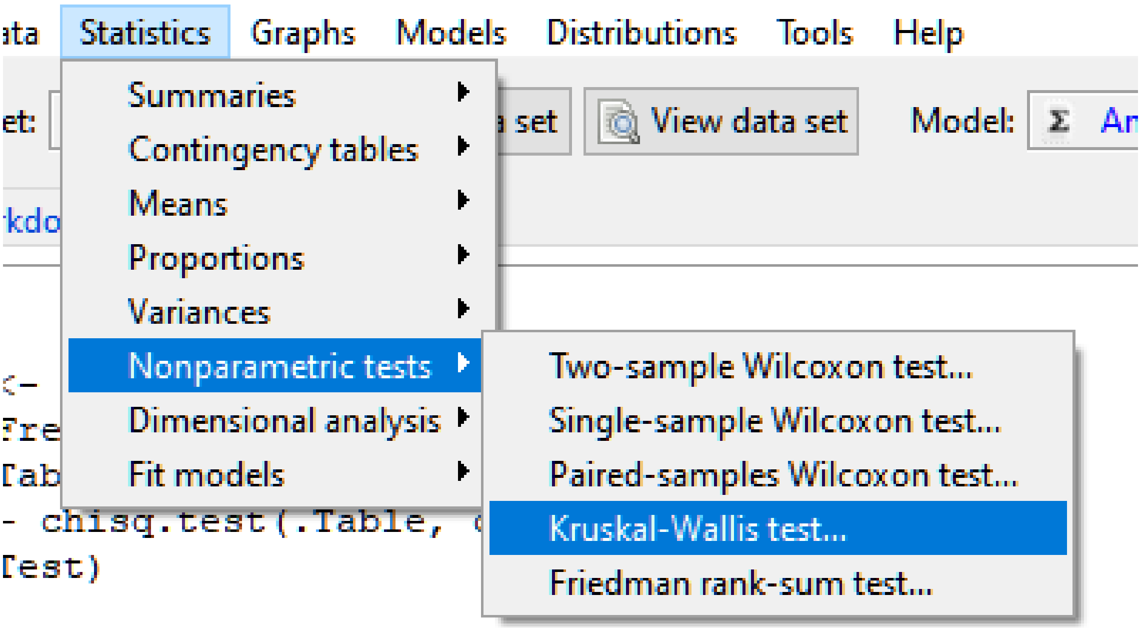
```
Frequency table:
      bin_out
gender  0  1
     1  7 15
     2 12  5

        Pearson's Chi-squared test

data:  .Table
X-squared = 5.7696, df = 1, p-value = 0.01631


        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.02484
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.03881034 0.91928487
sample estimates:
odds ratio
 0.2037389
```

Figure 73: Fisher's exact test result

Calculating the p-value of Fisher’s exact test, we obtain a similar value (0.016 vs 0.025). However, a significant p-value (close to 0.05) on the borderline can tip over, so let's monitor the fulfillment of this condition!

### *Kruskal-Wallis ANOVA*

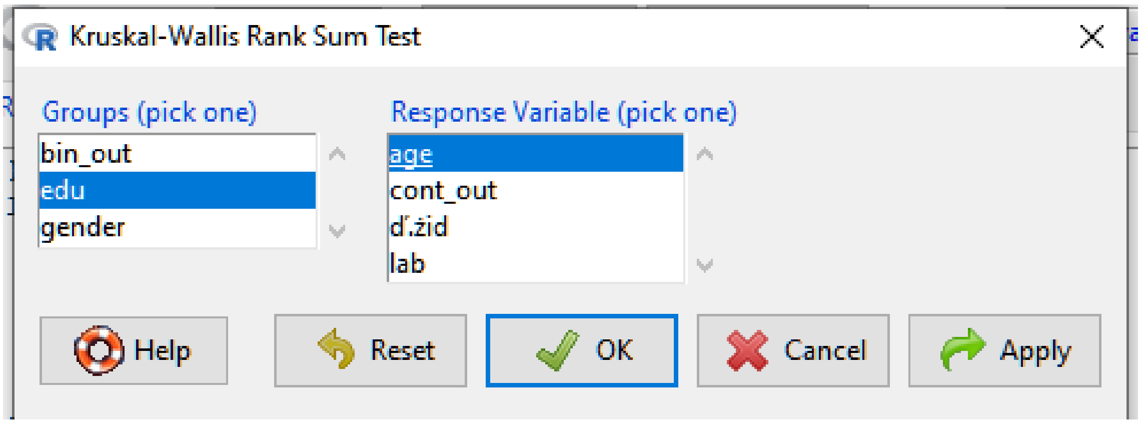
It is a nonparametric counterpart of ANOVA.



```
<-
Fre
Tab
- chisq.test(.Table,
Test)
```

Figure 74: Selection of Kruskal-Wallis ANOVA

Let us examine whether there is a difference in median ages across levels of education.


Kruskal-Wallis Rank Sum Test

Groups (pick one): bin_out, edu (selected), gender

Response Variable (pick one): age (selected), cont_out, d'.žid, lab

Figure 75: Kruskal-Wallis ANOVA

Based on the result of the test, there is a difference (between at least 2 groups), since p=0.042. Exactly which two groups have a difference is shown by the so-called post hoc tests. (Examples of such post-hoc tests are Bonferroni for ANOVA and Dunn for KW ANOVA.)

## Correlation

Correlation analysis can be used to analyze the association between two continuous variables. Along the two axes of the scattergram (x and y), the two variables are indicated (swapping the axes can only change the interpretation, not the p-value orcorrelation coefficient). A good example of overlapping statistical methods is correlation. For simple linear regression, the p-value of the regression coefficient is perfectly equal to the p-value of the correlation coefficient. The only difference is that the correlation coefficient shows the percentage movement of the two continuous variables together within the range from -1 to +1, while the regression coefficient shows the change in the outcome (in the independent variable) for one unit change (in the independent variable) on a scale of -∞ and +∞.

For normal distributions, the Pearson correlation is used, while for nonparametric distributions, the Spearman correlation is used.

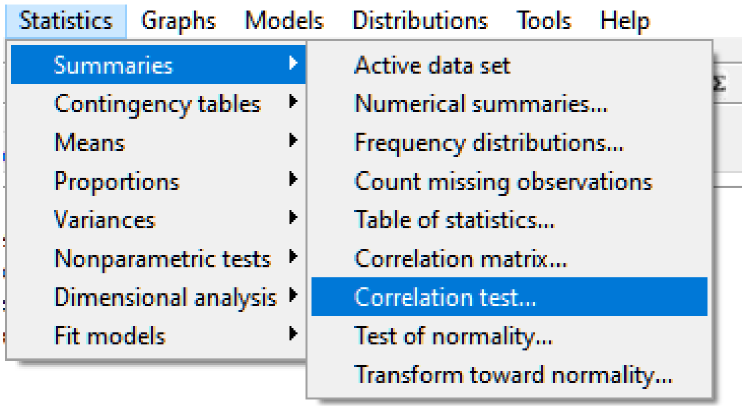
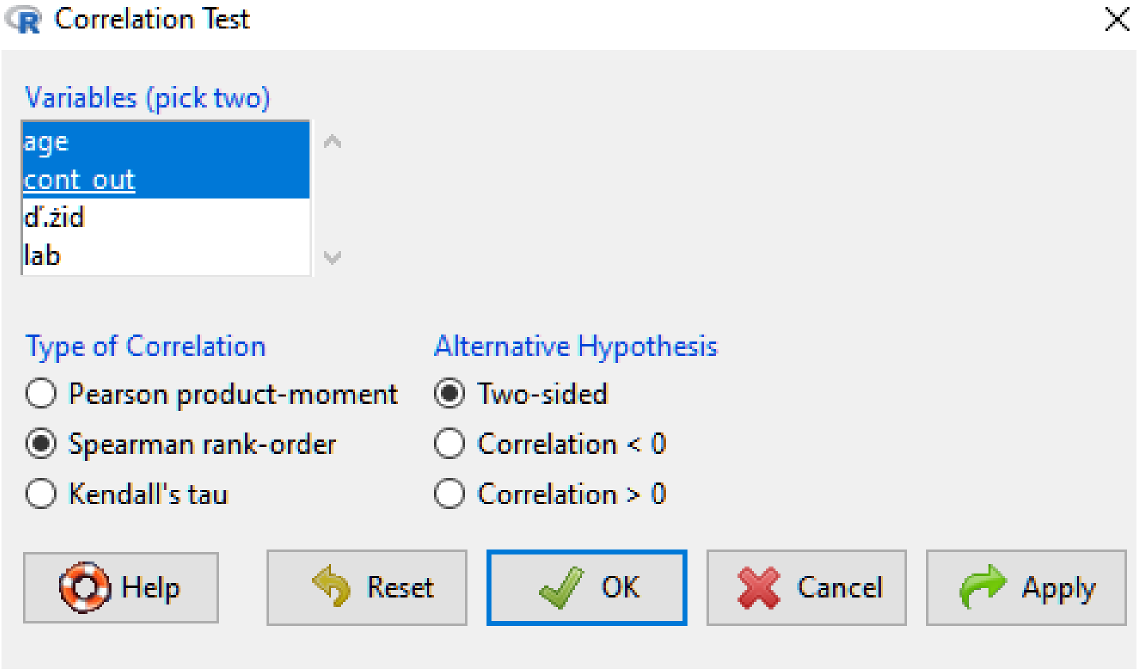


Figure 76: Correlation selection

Looking at the movement (correlation) of age (*age*, which does not show a normal distribution based on the results of the previous normality test) and our continuous outcomes together:



Correlation Test

Variables (pick two)

age
cont_out
ď.zid
lab

Type of Correlation

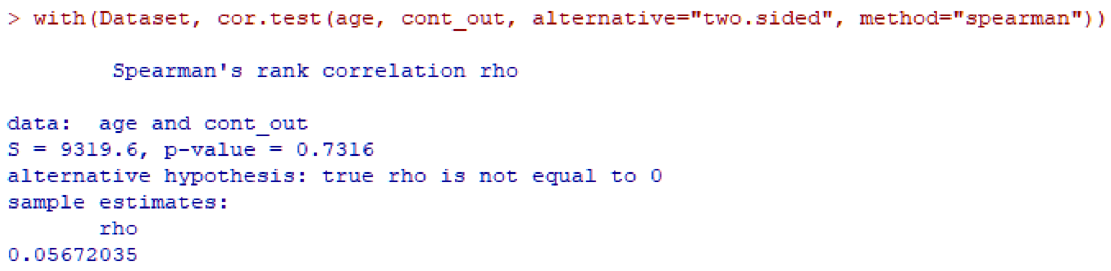
○ Pearson product-moment
● Spearman rank-order
○ Kendall's tau

Alternative Hypothesis

● Two-sided
○ Correlation < 0
○ Correlation > 0

Figure 77: Correlation setup

It can be said that based on the result of Spearman's correlation, a very weak (5.67%) co-movement is observed, which is also not significant (p=0.732).



```
> with(Dataset, cor.test(age, cont_out, alternative="two.sided", method="spearman"))

        Spearman's rank correlation rho

data:  age and cont_out
S = 9319.6, p-value = 0.7316
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.05672035
```

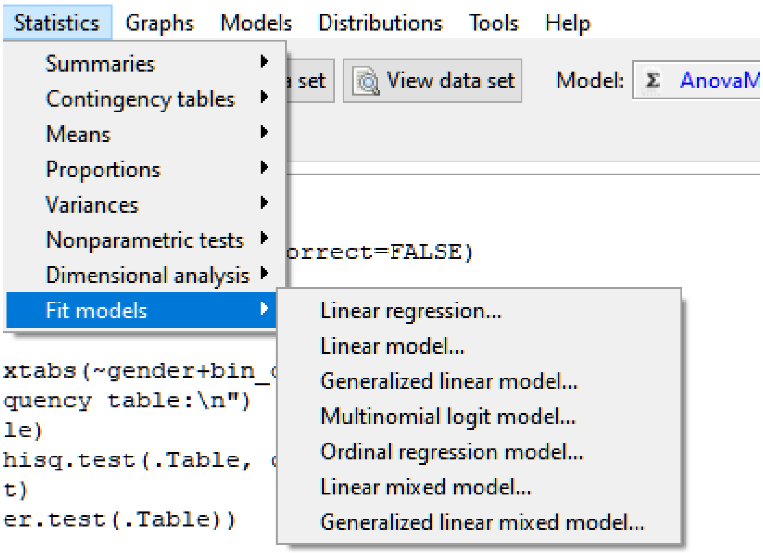
Figure 78: Spearman correlation result

## Multiple analysis

Based on the descriptive statistics, we got a picture of the database. For simple analyses, we looked at which variables significantly influenced our outcome. In connection with multiple analyses, adjusted measures are obtained, which are adjusted for potential confounders. The terms multiple and multivariate are often (incorrectly) interchanged. Multivariate means that we do not have one study outcome, but several (e.g., we want to test the effect of explanatory variables in diabetes and hypertension). The multiple model means that we have not only one explanatory (influencing) factor (as in simple analyses), but several.

### *Linear regression*

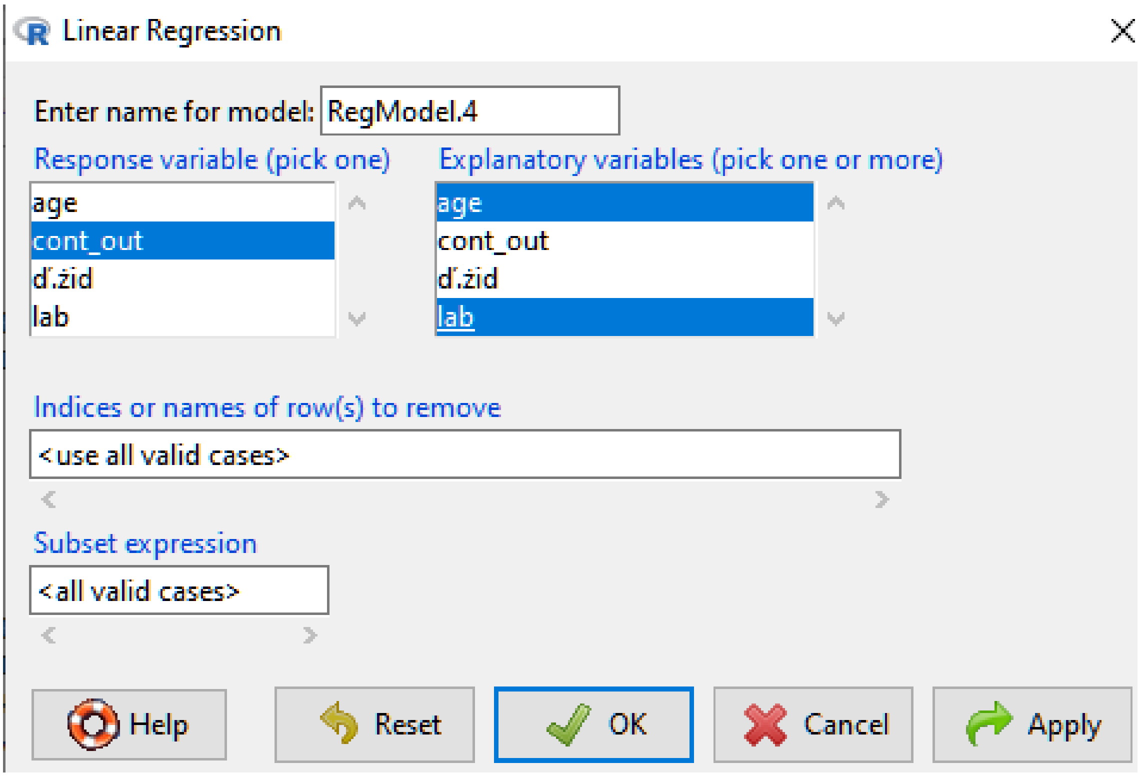
Among multiple models, we choose linear regression if our outcome is a continuous variable. Here, the distribution of the outcome is often examined (in fact, for OLS regression, the distribution of residues would have to be investigated in the case of a straight line more or less fitted to residues), if it is not nearly normal (i.e. the p-value does not have to be above 0.05, it is enough to have the distribution close to it, even if it is a little below it), then either we need to convert our variable (transform) or choose a statistically slightly weaker method (e.g. robust regression).



```
orrect=FALSE)
xtabs(~gender+bin_
quency table:\n")
le)
hisq.test(.Table,
t)
er.test(.Table))
```

Figure 79: Selecting a linear regression

Let's see if our continuous outcome is influenced by age and/or lab value.


Linear Regression

Enter name for model: RegModel.4

Response variable (pick one): cont_out

Explanatory variables (pick one or more): age, lab

Indices or names of row(s) to remove: <use all valid cases>

Subset expression: <all valid cases>

Figure 80: Linear regression setting

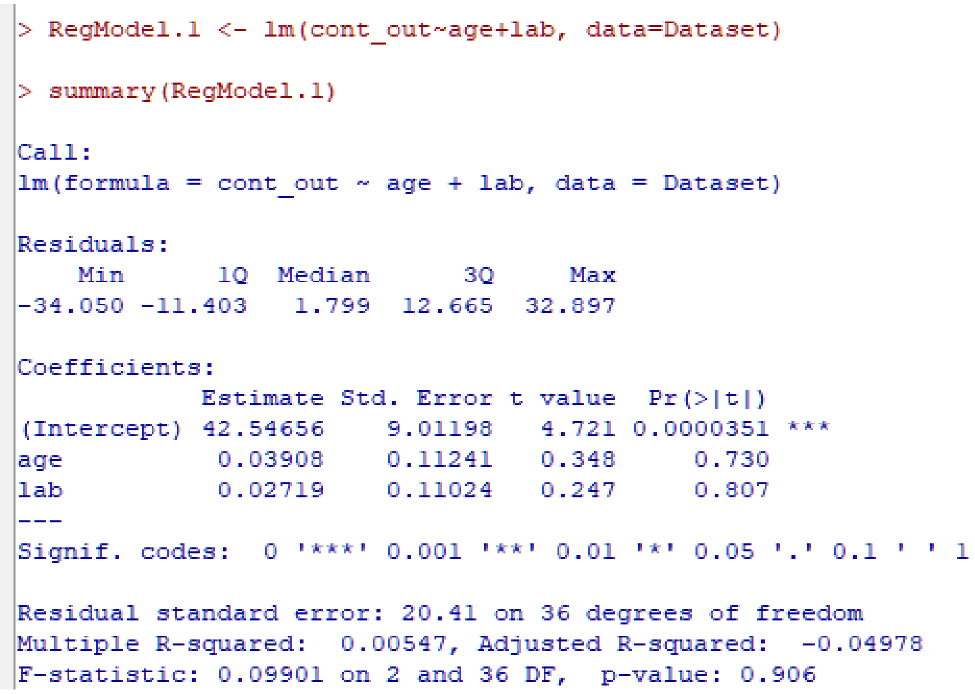

```
> RegModel.1 <- lm(cont_out~age+lab, data=Dataset)

> summary(RegModel.1)

Call:
lm(formula = cont_out ~ age + lab, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.050 -11.403   1.799  12.665  32.897 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 42.54656    9.01198   4.721 0.0000351 ***
age          0.03908    0.11241   0.348     0.730    
lab          0.02719    0.11024   0.247     0.807    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.41 on 36 degrees of freedom
Multiple R-squared:  0.00547, Adjusted R-squared:  -0.04978 
F-statistic: 0.09901 on 2 and 36 DF,  p-value: 0.906
```

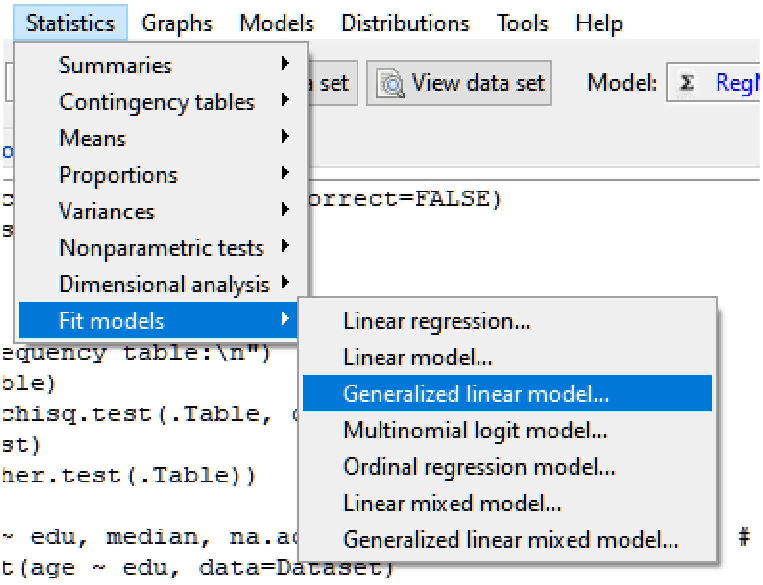
Figure 81: Linear regression result

None of these factors (age, lab) seem to have a significant influence on the outcome (p=0.348; p=0.807).

*Logistic regression*

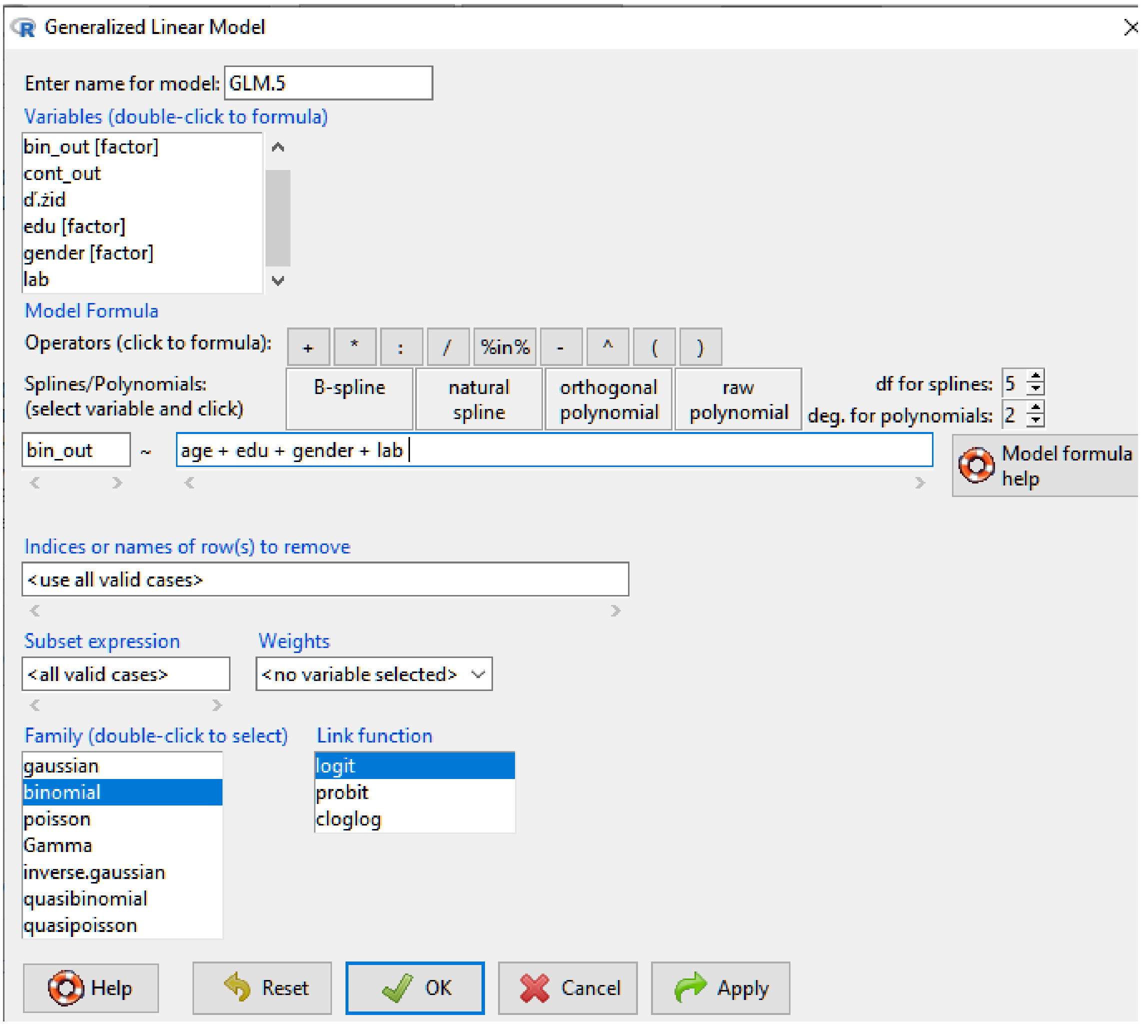
In logistical regression, our outcome is binary (yes-no; 1-0). Although the model gives the coefficients in the same way, it is more common to interpret the odds ratio.

Let's examine the potential influencing factors of our binary outcome.



```
orrect=FALSE)
equency table:\n")
ble)
chisq.test(.Table,
st)
her.test(.Table))
~ edu, median, na.ac
t(age ~ edu, data=Dataset)
```

Figure 82: Selecting logistic regression



Enter name for model: GLM.5

bin_out [factor]
cont_out
d.žid
edu [factor]
gender [factor]
lab

df for splines: 5
deg. for polynomials: 2

bin_out ~ age + edu + gender + lab

<use all valid cases>

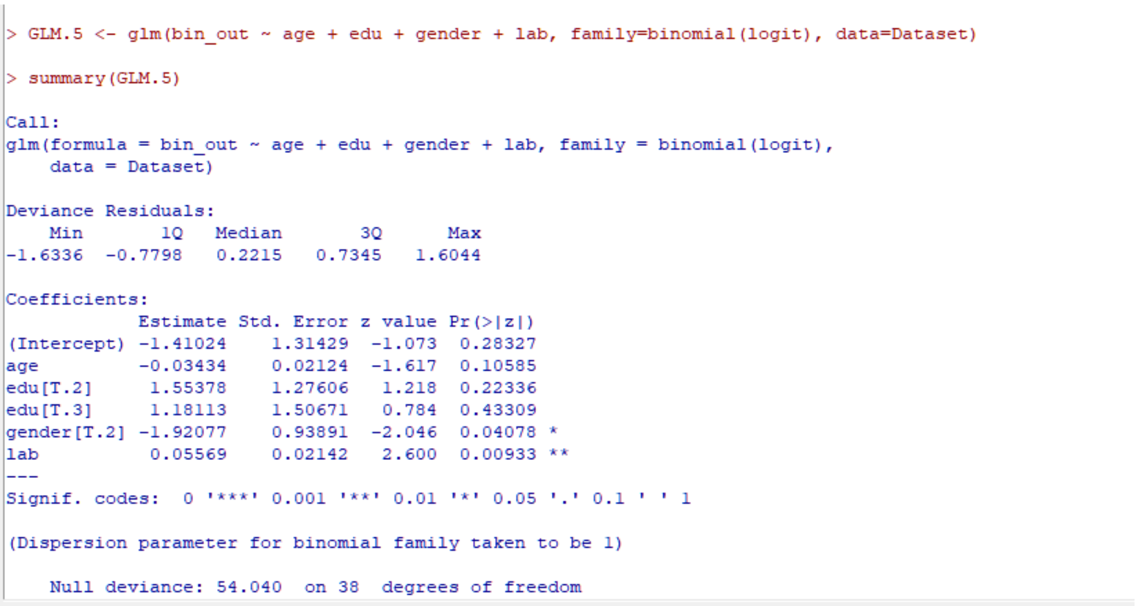
<all valid cases>

<no variable selected>

gaussian
binomial
poisson
Gamma
inverse.gaussian
quasibinomial
quasipoisson

logit
probit
cloglog

Figure 83: Setting logistic regression

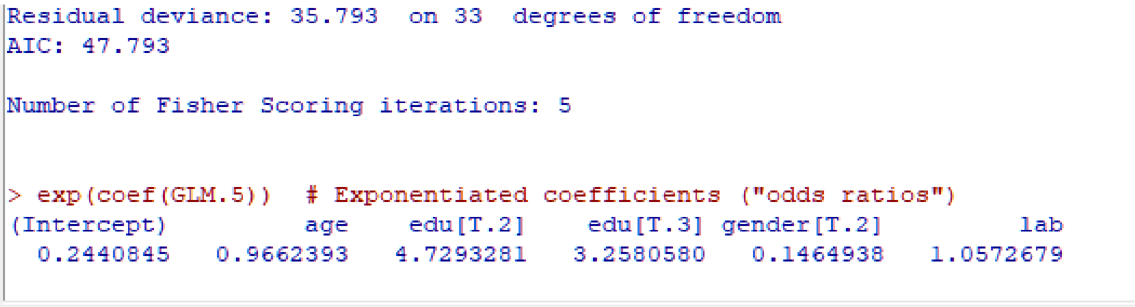


```
> GLM.5 <- glm(bin_out ~ age + edu + gender + lab, family=binomial(logit), data=Dataset)

> summary(GLM.5)

Call:
glm(formula = bin_out ~ age + edu + gender + lab, family = binomial(logit),
    data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6336  -0.7798   0.2215   0.7345   1.6044

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41024    1.31429  -1.073  0.28327
age         -0.03434    0.02124  -1.617  0.10585
edu[T.2]     1.55378    1.27606   1.218  0.22336
edu[T.3]     1.18113    1.50671   0.784  0.43309
gender[T.2] -1.92077    0.93891  -2.046  0.04078 *
lab          0.05569    0.02142   2.600  0.00933 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.040  on 38  degrees of freedom
```

Figure 84: Logistic regression result

Gender (*gender*; p=0.041) and lab value (lab; p=0.009) appear to be significant influencing factors.

Scrolling down, we also get the Odds Ratios.



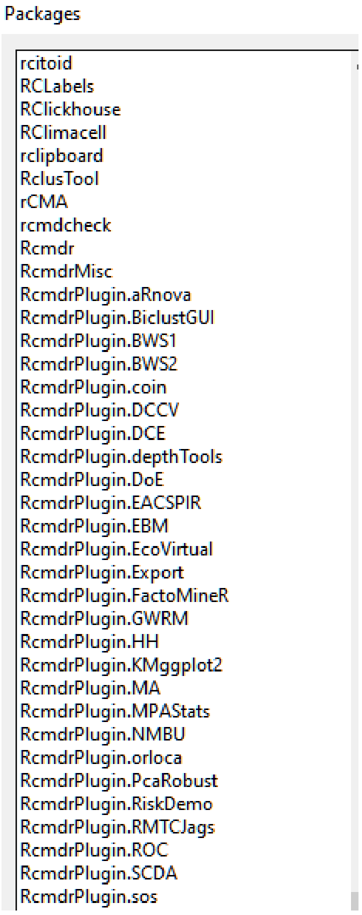
```
Residual deviance: 35.793  on 33  degrees of freedom
AIC: 47.793

Number of Fisher Scoring iterations: 5


> exp(coef(GLM.5))  # Exponentiated coefficients ("odds ratios")
(Intercept)         age    edu[T.2]    edu[T.3] gender[T.2]         lab
  0.2440845   0.9662393   4.7293281   3.2580580   0.1464938   1.0572679
```

Figure 85: Logistic regression results

The *protective* factor, as a forward-looking interpretation, makes the woman 0.15 times less likely to have the outcome compared to the man. The lab value appears to be a 1.05-fold risk factor.

### *Cox regression*

When Cox regression is used for survival analysis, the outcome is also binary (alive, dead). It's so much more than a logistic regression as time is also taken into consideration. As a first step, we install the plugin "*RcmdrPlugin.survival*" from the R packages:



Packages

rcitoid
RCLabels
RClickhouse
RClimacell
rclipboard
RclusTool
rCMA
rcmdcheck
Rcmdr
RcmdrMisc
RcmdrPlugin.aRnova
RcmdrPlugin.BiclustGUI
RcmdrPlugin.BWS1
RcmdrPlugin.BWS2
RcmdrPlugin.coin
RcmdrPlugin.DCCV
RcmdrPlugin.DCE
RcmdrPlugin.depthTools
RcmdrPlugin.DoE
RcmdrPlugin.EACSPIR
RcmdrPlugin.EBM
RcmdrPlugin.EcoVirtual
RcmdrPlugin.Export
RcmdrPlugin.FactoMineR
RcmdrPlugin.GWRM
RcmdrPlugin.HH
RcmdrPlugin.KMggplot2
RcmdrPlugin.MA
RcmdrPlugin.MPAStats
RcmdrPlugin.NMBU
RcmdrPlugin.orloca
RcmdrPlugin.PcaRobust
RcmdrPlugin.RiskDemo
RcmdrPlugin.RMTCJags
RcmdrPlugin.ROC
RcmdrPlugin.SCDA
RcmdrPlugin.sos

Figure 86: Install a package for Cox regression

After that, we can load this from Rcommander.

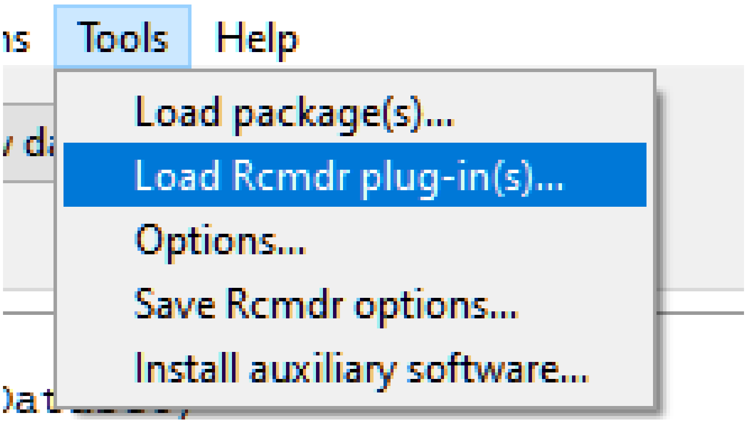


Figure 87: Load R package

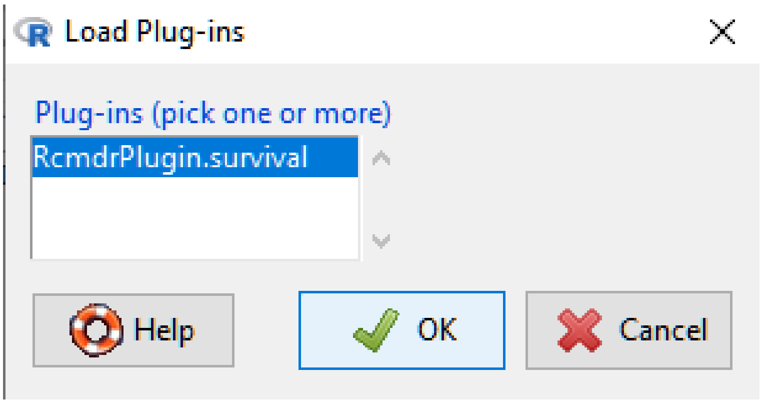


Figure 88: Choosing a package to use for survival analysis

Loading requires a restart of R commander (the existing database is deleted from memory).

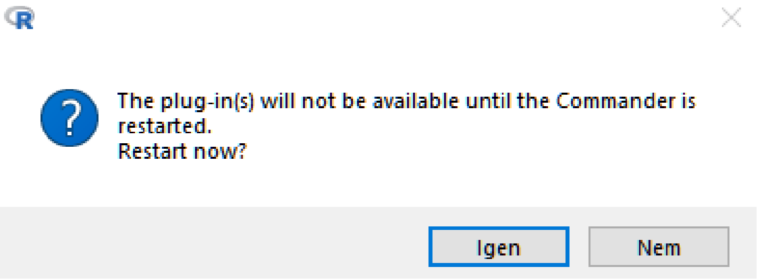

The plug-in(s) will not be available until the Commander is restarted.
Restart now?

Figure 89: R commander restart

You will get the new menu item.

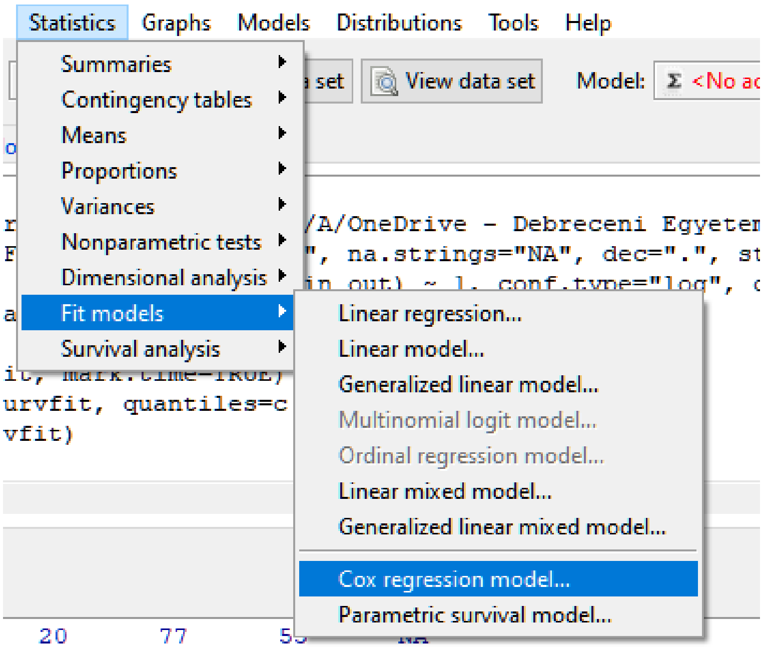


Figure 90: Cox regression selection

In the case of Cox regression, we can set the time (time), the event (event), for which the code of death is usually 1. It returns Hazard Ratios that are interpreted similarly to the Odds Ratio.

# Summary of methods of analysis

A summary diagram of the use of statistical methods depending on the research questions and the available data can be seen. The most common assumptions are normality and the number of sample size.

The summary diagram below starts from the research question, and then proceeds from the data to the final appropriate method based on the fulfillment of the conditions.

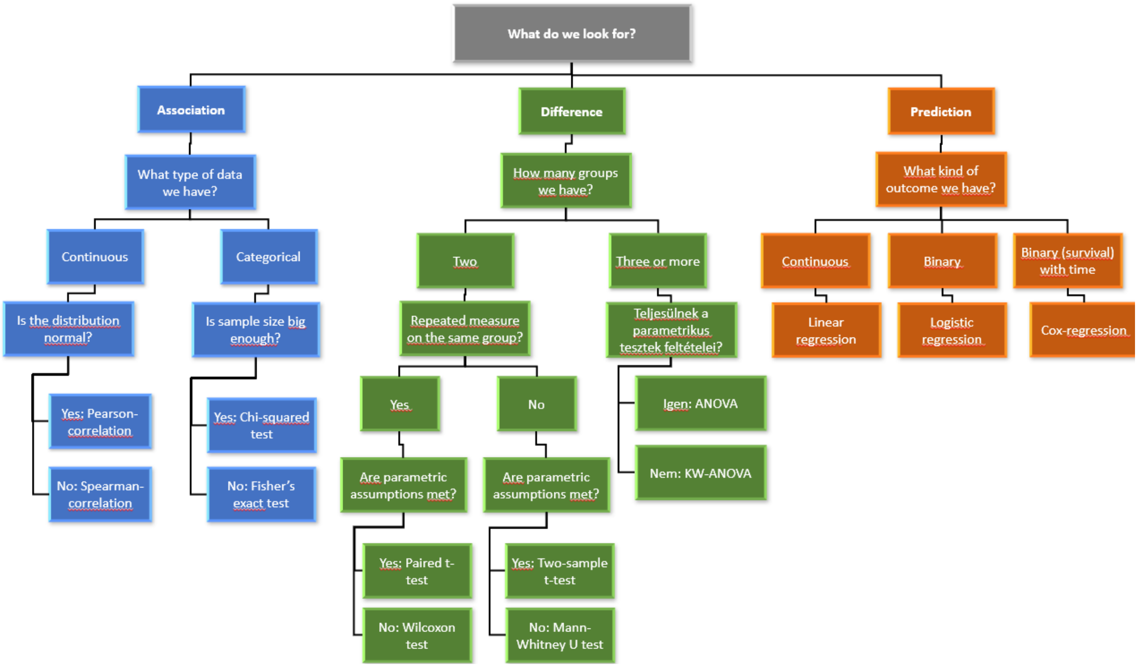


Figure 91: Summary of analysis methods

# Visualization of results

Proper visualization is essential when presenting our results. Due to their nature, different types of data are easier to review after selecting the appropriate chart type. To mention the main things, without being exhaustive:

- Time trends should be plotted on a line chart (e.g., prevalence over a longer time horizon)
- In general percentage distributions are illustrated on pie charts (e.g., gender distribution)
- Continuous variable is illustrated on a histogram
- Non-normal continuous variable shown in box plots
- The association between two continuous variables is described by a scattergram
- For categorical variables (e.g., educational attainment), a bar chart can also be used

# Epilogue

The literature used also contains useful links with brief descriptions/explanations. The database used is available with a codebook and allows you to reproduce the results obtained by the presented methods. The saying "Practice makes perfect" is especially true of statistics. If you are unsure about the conditions of application of a chosen method or its usability in the event of a particular problem, you should look for it. PubMed will help you find/check back on the right methods for your research question. YouTube often takes us through execution and even the interpretation in detail. Google Is Your Friend and we can use the right keywords to find valuable pages. It is a good idea to put key word combinations in quotation marks, or use the appropriate option (e.g., filetype:pdf) if you are searching for a specific file type. Let's practice as much as possible!

# Bibliography

1. The **relevant lecture materials of the University of Debrecen, Faculty of Health Sciences** have a detailed addition to the mentioned summary practical aspects.
2. The **European Health Interview Survey** (https://ec.europa.eu/eurostat/documents/203647/203710/EHIS_wave_1_guidelines.pdf/ffbeb62c-8f64-4151-938c-9ef171d148e0) can provide a validated sample for questionnaires
3. **Microsoft Forms** (https://www.office.com/launch/forms?auth=2) is a trusted data entry software.
4. **Deepl** is a useful translator beside Google Translate (https://www.deepl.com/translator)
5. **Zotero** is a free and easy-to-use reference editor (https://www. zotero.org/), which integrates with both the browser and Microsoft Word.
6. There are many online **databases** with data for international comparison, e.g. HFA explorer (https://gateway.euro.who.int/en/hfa-explorer/), Global Burden of Disease (GBD) (https://vizhub.healthdata.org/gbd-compare/)
7. **Laerd Statistics** (https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php) describes in detail the assumptions and also presents the details of the execution in SPSS.
8. There is also a detailed presentation with interpretation on the UCLA: **Statistical Consulting Group** page (https://stats.oarc.ucla.edu/other/dae/ ) for Stata, SAS, SPSS, Mplus and R programs.

# Database (semicolon separated values)

The contents of the database (sample.csv) used for the examples are with a Hungarian separator (; separator).

```
id;gender;age;edu;lab;bin_out;cont_out
1;1;53;1;95;1;49
2;2;88;3;80;0;54
3;1;7;1;96;1;35
4;2;5;3;16;0;55
5;1;19;1;50;1;55
6;2;92;3;12;0;58
7;2;97;3;63;0;14
8;1;50;2;20;1;53
9;1;3;1;90;1;40
10;1;29;1;88;1;38
11;2;94;3;39;0;75
12;2;82;2;74;0;14
13;1;77;3;89;1;78
14;1;21;1;46;0;65
15;1;29;1;21;0;57
16;2;37;2;42;0;36
17;1;86;2;60;0;15
18;2;15;2;26;0;38
19;1;54;3;99;1;52
20;1;28;3;17;0;77
21;1;77;2;95;1;48
22;2;70;1;77;0;35
23;2;45;3;22;0;11
24;2;44;2;59;1;73
25;1;43;2;39;1;50
26;1;72;1;12;0;18
27;1;45;2;41;0;58
28;1;30;1;99;1;23
29;2;62;2;88;0;72
30;1;65;2;25;1;39
31;2;9;2;23;1;13
32;2;97;1;16;0;47
33;1;49;2;70;1;26
34;1;89;2;35;1;39
35;1;84;2;77;1;75
36;2;90;3;97;1;64
37;1;78;2;64;0;73
38;2;30;1;82;1;49
39;2;83;3;82;1;32
```

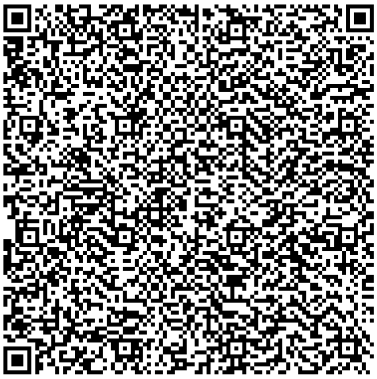
## Database (comma-separated values)

Database used for examples (sample.csv) in comma-separated form (which is the original English comma-separated values).

```
id,gender,age,edu,lab,bin_out,cont_out
1,1,53,1,95,1,49
2,2,88,3,80,0,54
3,1,7,1,96,1,35
4,2,5,3,16,0,55
5,1,19,1,50,1,55
6,2,92,3,12,0,58
7,2,97,3,63,0,14
8,1,50,2,20,1,53
9,1,3,1,90,1,40
10,1,29,1,88,1,38
11,2,94,3,39,0,75
12,2,82,2,74,0,14
13,1,77,3,89,1,78
14,1,21,1,46,0,65
15,1,29,1,21,0,57
16,2,37,2,42,0,36
17,1,86,2,60,0,15
18,2,15,2,26,0,38
19,1,54,3,99,1,52
20,1,28,3,17,0,77
21,1,77,2,95,1,48
22,2,70,1,77,0,35
23,2,45,3,22,0,11
24,2,44,2,59,1,73
25,1,43,2,39,1,50
26,1,72,1,12,0,18
27,1,45,2,41,0,58
28,1,30,1,99,1,23
29,2,62,2,88,0,72
30,1,65,2,25,1,39
31,2,9,2,23,1,13
32,2,97,1,16,0,47
33,1,49,2,70,1,26
34,1,89,2,35,1,39
35,1,84,2,77,1,75
36,2,90,3,97,1,64
37,1,78,2,64,0,73
38,2,30,1,82,1,49
39,2,83,3,82,1,32
```

## Database (QR code)

The QR code below contains the database. QR code generation was done with the free QR Code Generator. (https://goqr.me/)



The above image can be scanned off-camera with online QR code readers. (https://blog.qr4.nl/Online-QR-Code-Decoder.aspx )

# Codebook

id – unique identifier

gender – gender

- 1 - male
- 2 – female

age – age (years)

edu – educational attainment

- 1 – primary
- 2 – secondary
- 3 – tertiary

lab –laboratory parameters (continuous variable)

bin_out – binary outcome

- 0 – no
- 1 – yes

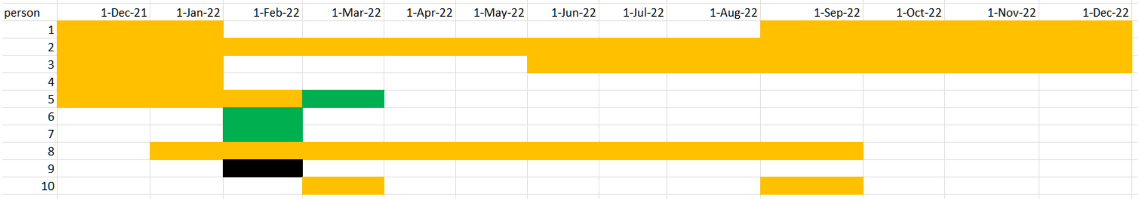
cont_out – continuous outcome

# Tasks

1. Calculate the mean of the age variable in the sample.csv and the 95% Confidence Interval of this point estimation in Excel, and based on this, declare whether the average age of the sample differs significantly from 35. Plot the average and the confidence interval.

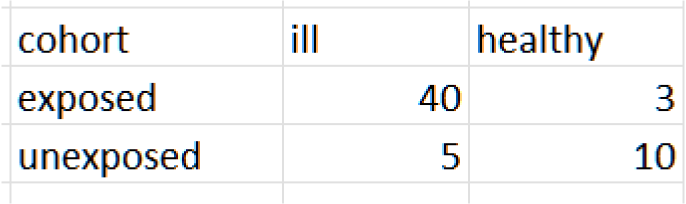
2. Given 50 men with a normal BMI, 80 obese men, 90 women with a normal BMI, and 20 obese women. Is there an association between gender and obesity?

3. In a follow-up study, one hundred people were followed for one year (01.01.2022-31.12.2022). Ten people have some kind of event (illness/death/move), the remaining ninety people have none of these events. The yellow cell represents illness, green represents movement, and black represents death. The studied disease is the flu (you can become infected repeatedly) but is not limited to the winter months. For colored cells, the given event is formed on the first day of the given month and lasts until the last day of the last colored month. (a) What is the prevalence on the first of July? (b) What is the one-year cumulative incidence and (c) what is the incidence density?

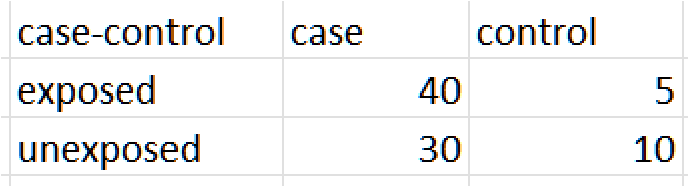


| person | 1-Dec-21 | 1-Jan-22 | 1-Feb-22 | 1-Mar-22 | 1-Apr-22 | 1-May-22 | 1-Jun-22 | 1-Jul-22 | 1-Aug-22 | 1-Sep-22 | 1-Oct-22 | 1-Nov-22 | 1-Dec-22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 🟨 | 🟨 |  |  |  |  |  |  |  | 🟨 | 🟨 | 🟨 | 🟨 |
| 2 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |
| 3 | 🟨 | 🟨 |  |  |  |  | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |
| 4 | 🟨 | 🟨 |  |  |  |  |  |  |  |  |  |  |  |
| 5 | 🟨 | 🟨 | 🟨 | 🟩 |  |  |  |  |  |  |  |  |  |
| 6 |  |  | 🟩 |  |  |  |  |  |  |  |  |  |  |
| 7 |  |  | 🟩 |  |  |  |  |  |  |  |  |  |  |
| 8 |  | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |  |  |  |
| 9 |  |  | ⬛ |  |  |  |  |  |  |  |  |  |  |
| 10 |  |  |  | 🟨 |  |  |  |  |  | 🟨 |  |  |  |

4. For the 2x2 contingency table below, calculate the association measures for the following (a) cohort and (b) case-control study!

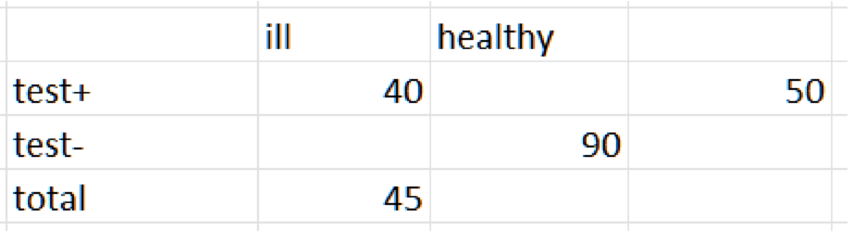


| cohort | ill | healthy |
|---|---|---|
| exposed | 40 | 3 |
| unexposed | 5 | 10 |



| case-control | case | control |
|---|---|---|
| exposed | 40 | 5 |
| unexposed | 30 | 10 |

5. Calculate the missing values, prevalence and test characteristics.



| | ill | healthy | |
|---|---|---|---|
| test+ | 40 | | 50 |
| test- | | 90 | |
| total | 45 | | |

# Solutions

1. In Excel, calculate the mean of the age variables in the sample.csv and the 95% confidence interval of this point estimation, and based on this, declare whether  the average age of the sample differs significantly from 35. Plot the average and confidence interval.

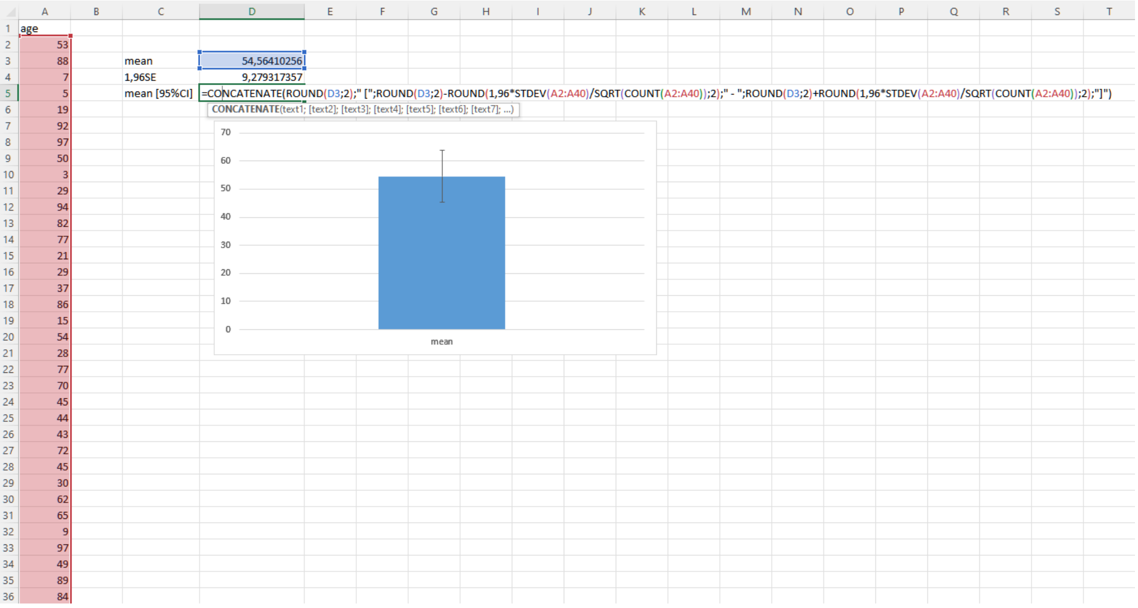
   The mean is [95% confidence interval]: 54.56 [45.28 - 63.84]

   The point estimation for my sample (54.56) is significantly higher than the hypothetical value of 35, based on the confidence interval, since the value according to the null hypothesis (35) is not in the 95% confidence interval.

   The result can be obtained in one cell using the following functions:

   = CONCATENATE(ROUND(D3;2);" [";ROUND(D3;2)-ROUND(1,96*STDEV(A2:A40)/SQRT(COUNT(A2:A40));2);" - ";ROUND(D3;2)+ROUND(1,96*STDEV(A2:A40)/SQRT(COUNT(A2:A40));2);"]")

   Functions used:
   CONCATENATE: concatenates the specified numbers/characters
   ROUND: rounds the number to a specific decimal place
   STDEV: returns the standard deviation of a selected range
   SQRT: calculates square root
   COUNT: counts the cells in the given range, in this case it matches the sample size



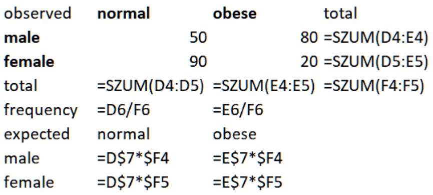
| | A | B | C | D |
|---|---|---|---|---|
| 1 | age | | | |
| 2 | 53 | | | |
| 3 | 88 | | mean | 54,56410256 |
| 4 | 7 | | 1,96SE | 9,279317357 |
| 5 | 5 | | mean [95%CI] | =CONCATENATE(ROUND(D3;2);" [";ROUND(D3;2)-ROUND(1,96*STDEV(A2:A40)/SQRT(COUNT(A2:A40));2);" - ";ROUND(D3;2)+ROUND(1,96*STDEV(A2:A40)/SQRT(COUNT(A2:A40));2);"]") |
| 6 | 19 | | | |
| 7 | 92 | | | |
| 8 | 97 | | | |
| 9 | 50 | | | |
| 10 | 3 | | | |
| 11 | 29 | | | |
| 12 | 94 | | | |
| 13 | 82 | | | |
| 14 | 77 | | | |
| 15 | 21 | | | |
| 16 | 29 | | | |
| 17 | 37 | | | |
| 18 | 86 | | | |
| 19 | 15 | | | |
| 20 | 54 | | | |
| 21 | 28 | | | |
| 22 | 77 | | | |
| 23 | 70 | | | |
| 24 | 45 | | | |
| 25 | 44 | | | |
| 26 | 43 | | | |
| 27 | 72 | | | |
| 28 | 45 | | | |
| 29 | 30 | | | |
| 30 | 62 | | | |
| 31 | 65 | | | |
| 32 | 9 | | | |
| 33 | 97 | | | |
| 34 | 49 | | | |
| 35 | 89 | | | |
| 36 | 84 | | | |

CONCATENATE(text1; [text2]; [text3]; [text4]; [text5]; [text6]; [text7]; ...)
70
60
50
40
30
20
10
0
mean

2. Given 50 men with a normal BMI, 80 obese men, 90 women with a normal BMI, and 20 obese women. Is there an association between gender and obesity?
   The chi-squared test is suitable for answering the question:

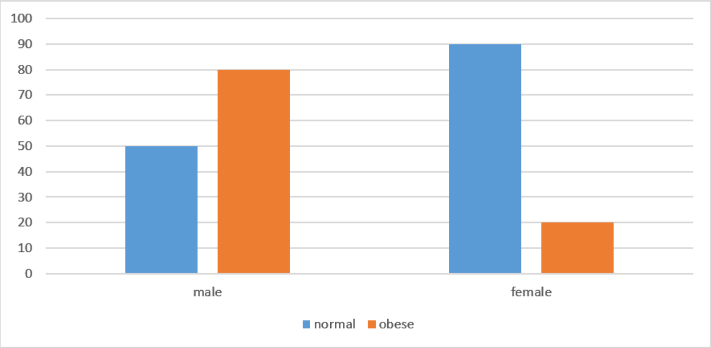
- We can indicate the aggregated case numbers in a contingency table.
- After that, we need to calculate the sum of the rows and columns.
- Next come the frequencies that occur in the columns.
- The expected case numbers can be calculated using the results of the previous calculations.

The association is significant, p<0.001, and based on the figure, there are more obese men.

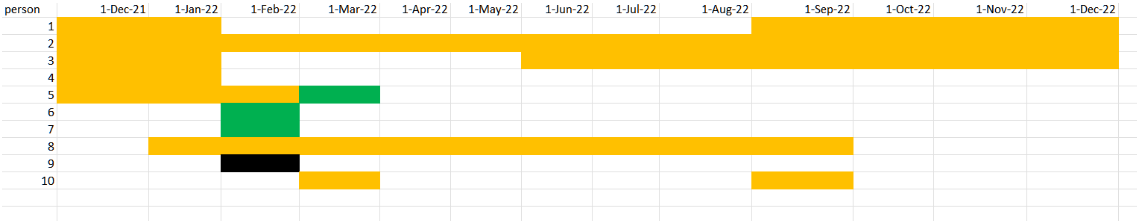


| observed | **normal** | **obese** | total |
|---|---|---|---|
| **male** | 50 | 80 | =SZUM(D4:E4) |
| **female** | 90 | 20 | =SZUM(D5:E5) |
| total | =SZUM(D4:D5) | =SZUM(E4:E5) | =SZUM(F4:F5) |
| frequency | =D6/F6 | =E6/F6 | |
| expected | normal | obese | |
| male | =D$7*$F4 | =E$7*$F4 | |
| female | =D$7*$F5 | =E$7*$F5 | |

=CHISQ.TEST(D4:E5;D9:E10) 1,13E-11



3. In a follow-up study, we follow one hundred people for a year (01.01.2022-31.12.2022). Ten people have some kind of event (illness/death/move), the remaining ninety people don't have any of these events. The yellow cell indicates the disease, the green cell indicates the movement, the black cell indicates the death. The studied disease is the flu (can infect repeatedly) but is not limited to the winter months. For colored cells, the given event is formed on the first day of the given month and lasts until the last day of the last colored month. (a) What is the prevalence on the first of July? (b) What is the one-year cumulative incidence and (c) what is the incidence density?



(a) P=n/N=3/96=0.0312

The prevalence is 3.12%, as there are three patients on the first of July and the total population is 100-4 (three individuals have moved away and one has died.)

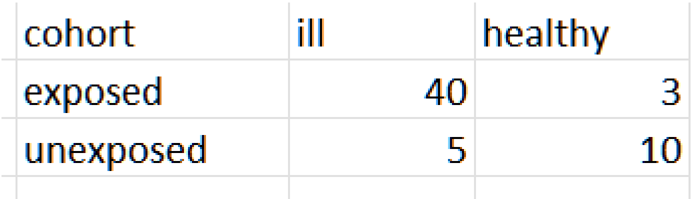
(b) CI=n/N=5/95=0.0526

The cumulative incidence of 1-year is 5.26%, as 5 new events (occurred more than once in some individuals) occurred and initially 100-5 people (5 people were already sick) were at risk.

(c) ID=n/total person-time=5/1115=0.0045/person-month=0.0538/person-year

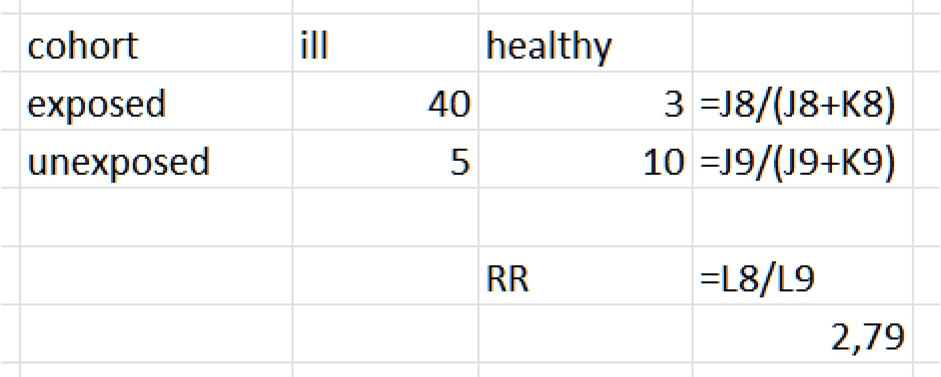
A total of five new events occurred. Ninety people had no events, their person time was 90*12=1080 months, as long as they were at risk (not sick and in the study). The remaining 10 people have a person time of 7+0+4+11+0+1+1+3+1+7=35 months, for a total of 1080+35=1115 months. This person-months 1115/12=92.92 person-years.

4. For the 2x2 contingency tables below, calculate the association measures for the following cohort and case-control studies!
(a)



| cohort | ill | healthy |
|---|---|---|
| exposed | 40 | 3 |
| unexposed | 5 | 10 |

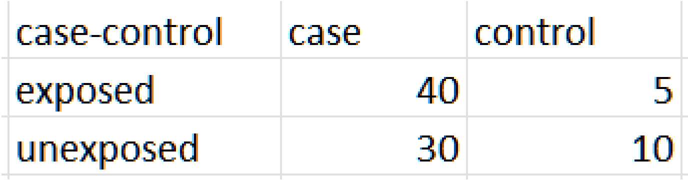
In the case of a cohort study, the relationship indicator is the relative risk when the two absolute risks (cumulative incidence) are divided by each other.



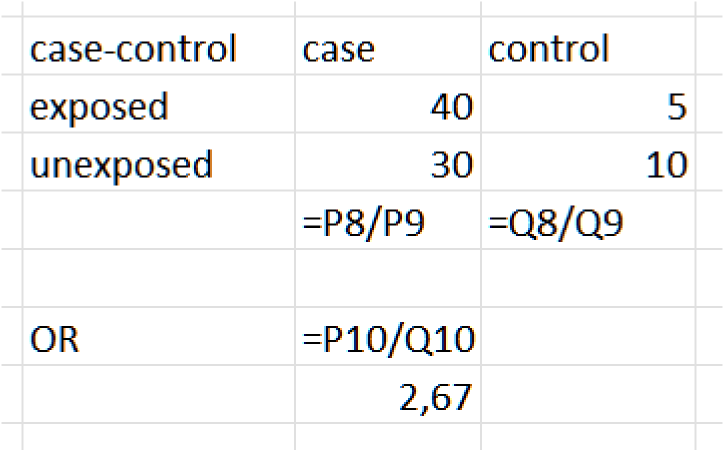
| cohort | ill | healthy | |
|---|---|---|---|
| exposed | 40 | 3 | =J8/(J8+K8) |
| unexposed | 5 | 10 | =J9/(J9+K9) |
| | | | |
| | | RR | =L8/L9 |
| | | | 2,79 |

Based on relative risk, exposed people have a 2.79x higher risk of having the disease compared to non-exposed people.

(b)



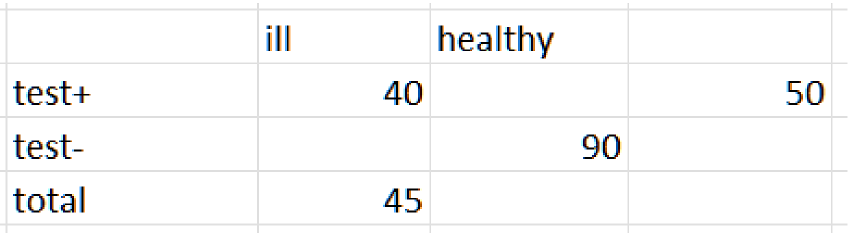
| case-control | case | control |
|---|---|---|
| exposed | 40 | 5 |
| unexposed | 30 | 10 |



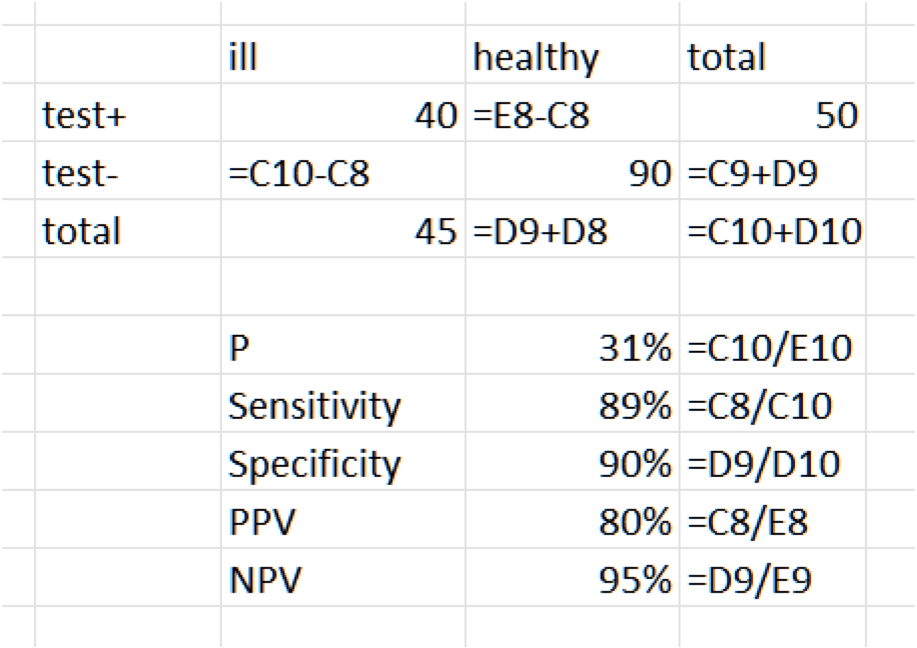
| case-control | case | control |
|---|---|---|
| exposed | 40 | 5 |
| unexposed | 30 | 10 |
|  | =P8/P9 | =Q8/Q9 |
|  |  |  |
| OR | =P10/Q10 |  |
|  | 2,67 |  |

The odds of exposure quotient are the odds ratio. In this case, there is a 2.67x higher chance of exposure among cases.

5. Calculate the missing values, prevalence and test characteristics!



| | ill | healthy | |
|---|---|---|---|
| test+ | 40 | | 50 |
| test- | | 90 | |
| total | 45 | | |



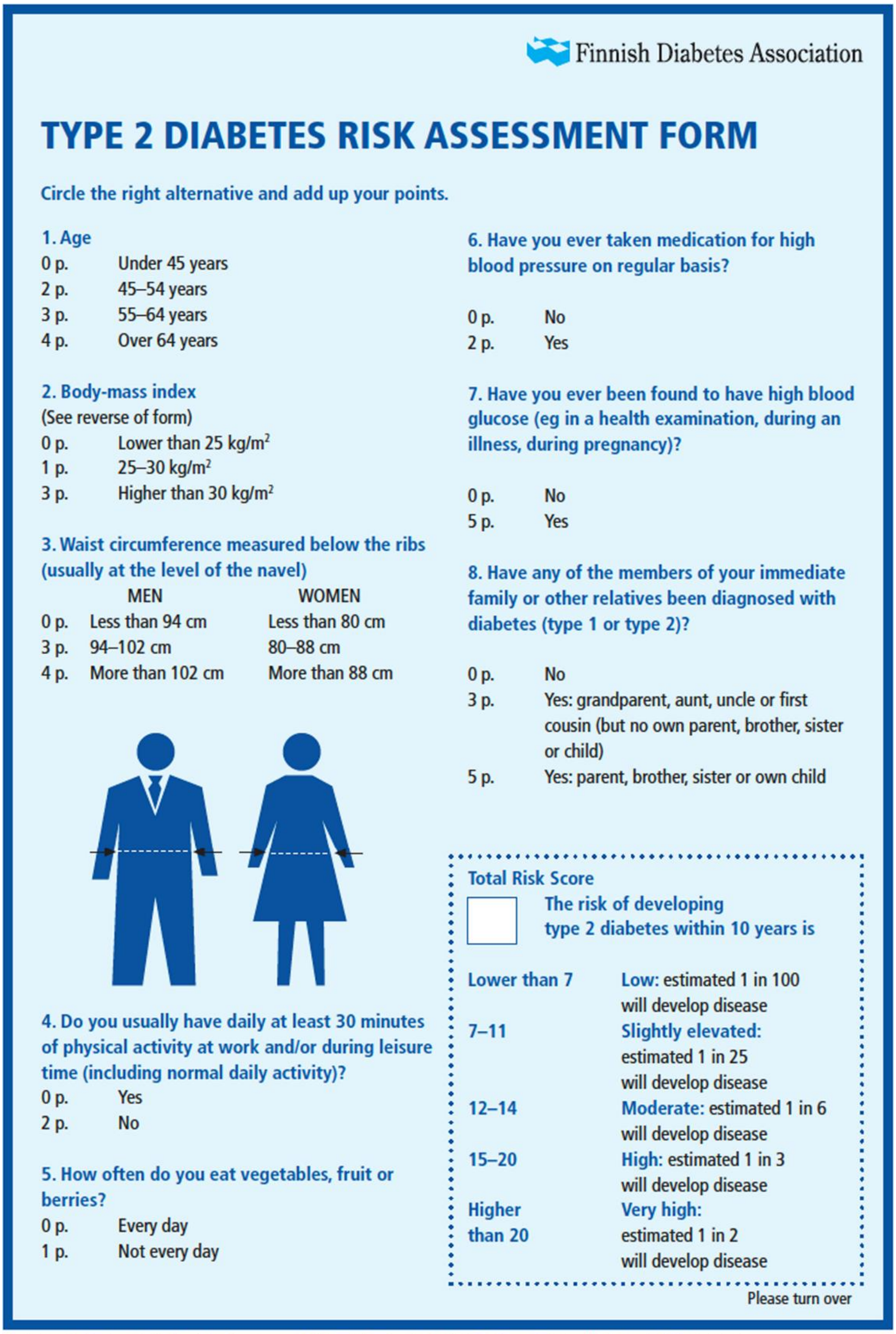
| | ill | healthy | total |
|---|---|---|---|
| test+ | 40 | =E8-C8 | 50 |
| test- | =C10-C8 | 90 | =C9+D9 |
| total | 45 | =D9+D8 | =C10+D10 |
| | | | |
| | P | 31% | =C10/E10 |
| | Sensitivity | 89% | =C8/C10 |
| | Specificity | 90% | =D9/D10 |
| | PPV | 80% | =C8/E8 |
| | NPV | 95% | =D9/E9 |

The frequency of the disease (prevalence) is 31%.
The test correctly identifies 89% of patients. (sensitivity)
The test correctly identifies 90% of healthy people. (specificity)
Individuals found positive are 80% likely to be sick. (PPV)
Individuals found negative have a 95% probability of being healthy. (PPV)

# Annex

Some important questionnaires excerpts with attribution are attached below.

## FINDRISC Questionnaire[1]



Finnish Diabetes Association

### TYPE 2 DIABETES RISK ASSESSMENT FORM

**Circle the right alternative and add up your points.**

**1. Age**

| | |
|---|---|
| 0 p. | Under 45 years |
| 2 p. | 45–54 years |
| 3 p. | 55–64 years |
| 4 p. | Over 64 years |

**2. Body-mass index**
(See reverse of form)

| | |
|---|---|
| 0 p. | Lower than 25 kg/m$^2$ |
| 1 p. | 25–30 kg/m$^2$ |
| 3 p. | Higher than 30 kg/m$^2$ |

**3. Waist circumference measured below the ribs (usually at the level of the navel)**

| | MEN | WOMEN |
|---|---|---|
| 0 p. | Less than 94 cm | Less than 80 cm |
| 3 p. | 94–102 cm | 80–88 cm |
| 4 p. | More than 102 cm | More than 88 cm |

**4. Do you usually have daily at least 30 minutes of physical activity at work and/or during leisure time (including normal daily activity)?**

| | |
|---|---|
| 0 p. | Yes |
| 2 p. | No |

**5. How often do you eat vegetables, fruit or berries?**

| | |
|---|---|
| 0 p. | Every day |
| 1 p. | Not every day |

**6. Have you ever taken medication for high blood pressure on regular basis?**

| | |
|---|---|
| 0 p. | No |
| 2 p. | Yes |

**7. Have you ever been found to have high blood glucose (eg in a health examination, during an illness, during pregnancy)?**

| | |
|---|---|
| 0 p. | No |
| 5 p. | Yes |

**8. Have any of the members of your immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)?**

| | |
|---|---|
| 0 p. | No |
| 3 p. | Yes: grandparent, aunt, uncle or first cousin (but no own parent, brother, sister or child) |
| 5 p. | Yes: parent, brother, sister or own child |

**Total Risk Score**

□ **The risk of developing type 2 diabetes within 10 years is**

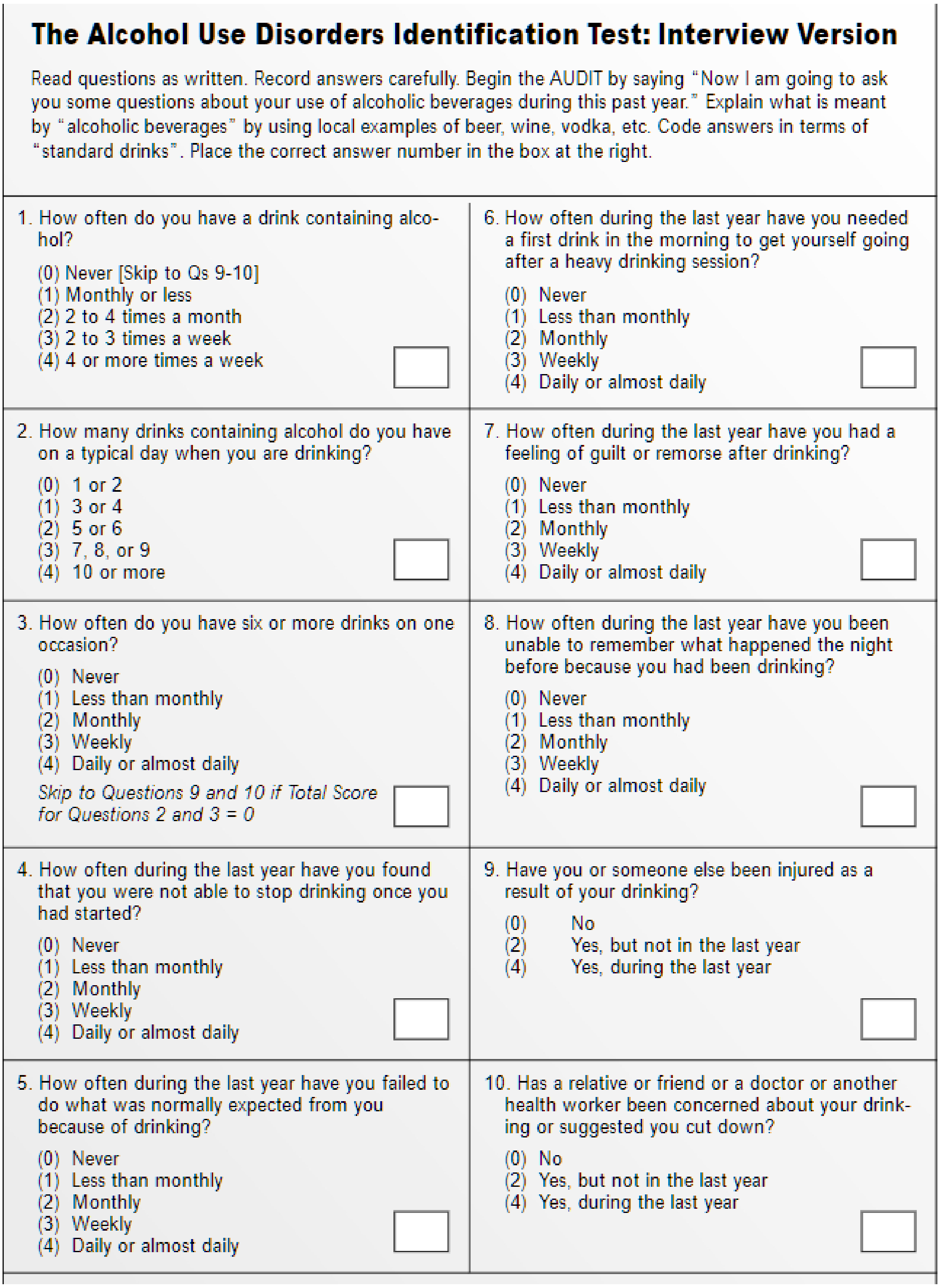
| | |
|---|---|
| Lower than 7 | Low: estimated 1 in 100 will develop disease |
| 7–11 | Slightly elevated: estimated 1 in 25 will develop disease |
| 12–14 | Moderate: estimated 1 in 6 will develop disease |
| 15–20 | High: estimated 1 in 3 will develop disease |
| Higher than 20 | Very high: estimated 1 in 2 will develop disease |

Please turn over

Test designed by Professor Jaakko Tuomilehto, Department of Public Health, University of Helsinki, and Jaana Lindström, MFS, National Public Health Institute.

[1] P. Böhme, A. Luc, P. Gillet, és N. Thilly, „Effectiveness of a type 2 diabetes prevention program combining FINDRISC scoring and telephone-based coaching in the French population of bakery/pastry employees", Eur J Clin Nutr, köt. 74, sz. 3, Art. sz. 3, márc. 2020, doi: 10.1038/s41430-019-0472-3.

## Alcohol Use Disorders Identification Test[2]



### The Alcohol Use Disorders Identification Test: Interview Version

Read questions as written. Record answers carefully. Begin the AUDIT by saying "Now I am going to ask you some questions about your use of alcoholic beverages during this past year." Explain what is meant by "alcoholic beverages" by using local examples of beer, wine, vodka, etc. Code answers in terms of "standard drinks". Place the correct answer number in the box at the right.

1. How often do you have a drink containing alcohol?

(0) Never [Skip to Qs 9-10]
(1) Monthly or less
(2) 2 to 4 times a month
(3) 2 to 3 times a week
(4) 4 or more times a week □

2. How many drinks containing alcohol do you have on a typical day when you are drinking?

(0) 1 or 2
(1) 3 or 4
(2) 5 or 6
(3) 7, 8, or 9
(4) 10 or more □

3. How often do you have six or more drinks on one occasion?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

*Skip to Questions 9 and 10 if Total Score for Questions 2 and 3 = 0* □

4. How often during the last year have you found that you were not able to stop drinking once you had started?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily □

5. How often during the last year have you failed to do what was normally expected from you because of drinking?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily □

6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily □

7. How often during the last year have you had a feeling of guilt or remorse after drinking?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily □

8. How often during the last year have you been unable to remember what happened the night before because you had been drinking?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily □

9. Have you or someone else been injured as a result of your drinking?

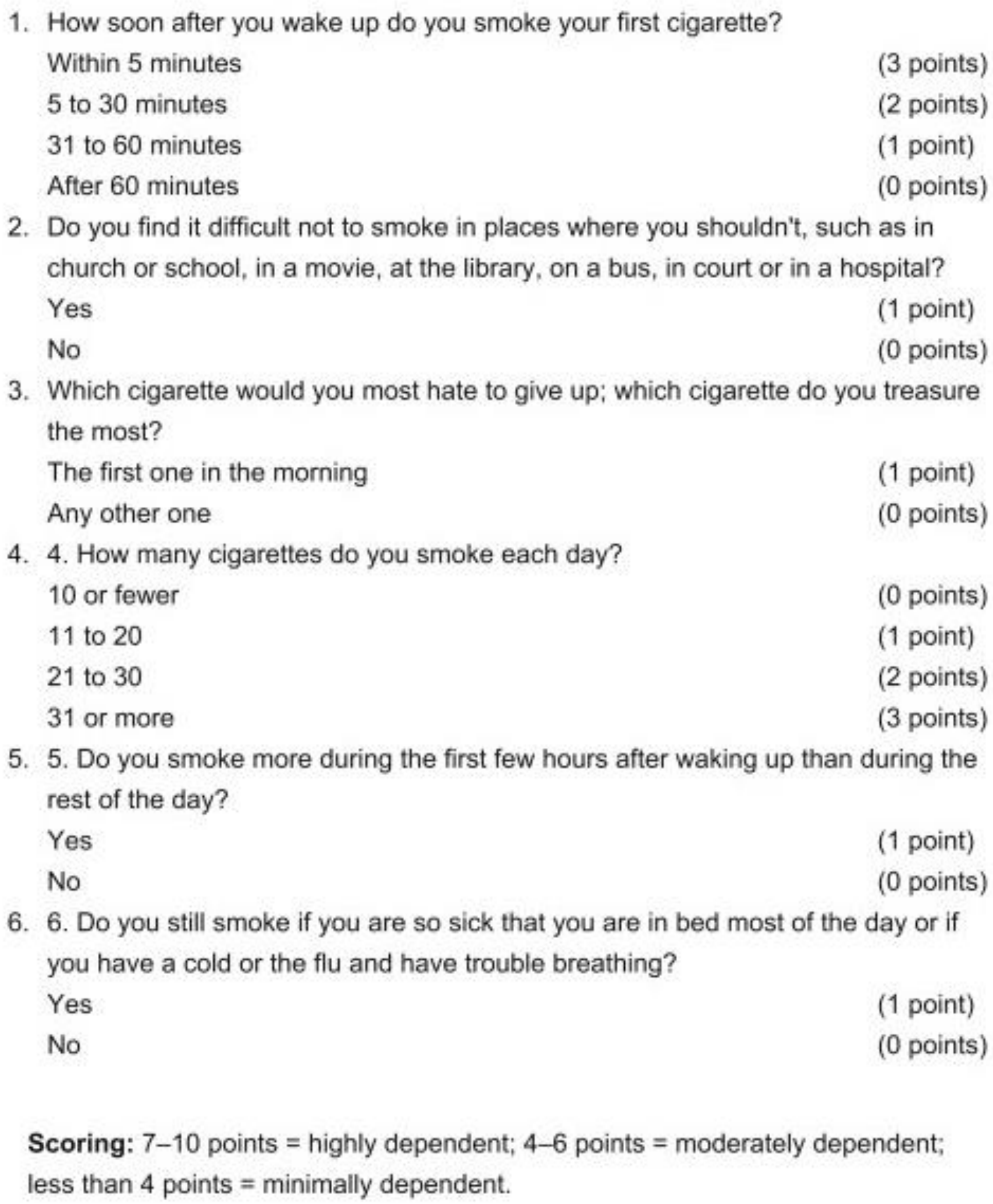
(0) No
(2) Yes, but not in the last year
(4) Yes, during the last year □

10. Has a relative or friend or a doctor or another health worker been concerned about your drinking or suggested you cut down?

(0) No
(2) Yes, but not in the last year
(4) Yes, during the last year □

[2] AUDIT : the Alcohol Use Disorders Identification Test : guidelines for use in primary health care [Internet]. [cited Feb 16, 2023]. Available from: https://www.who.int/publications-detail-redirect/WHO-MSD-MSB-01.6a

## Fagerström test for nicotine dependence[3]



1. How soon after you wake up do you smoke your first cigarette?

| | |
|---|---|
| Within 5 minutes | (3 points) |
| 5 to 30 minutes | (2 points) |
| 31 to 60 minutes | (1 point) |
| After 60 minutes | (0 points) |

2. Do you find it difficult not to smoke in places where you shouldn't, such as in church or school, in a movie, at the library, on a bus, in court or in a hospital?

| | |
|---|---|
| Yes | (1 point) |
| No | (0 points) |

3. Which cigarette would you most hate to give up; which cigarette do you treasure the most?

| | |
|---|---|
| The first one in the morning | (1 point) |
| Any other one | (0 points) |

4. 4. How many cigarettes do you smoke each day?

| | |
|---|---|
| 10 or fewer | (0 points) |
| 11 to 20 | (1 point) |
| 21 to 30 | (2 points) |
| 31 or more | (3 points) |

5. 5. Do you smoke more during the first few hours after waking up than during the rest of the day?

| | |
|---|---|
| Yes | (1 point) |
| No | (0 points) |

6. 6. Do you still smoke if you are so sick that you are in bed most of the day or if you have a cold or the flu and have trouble breathing?

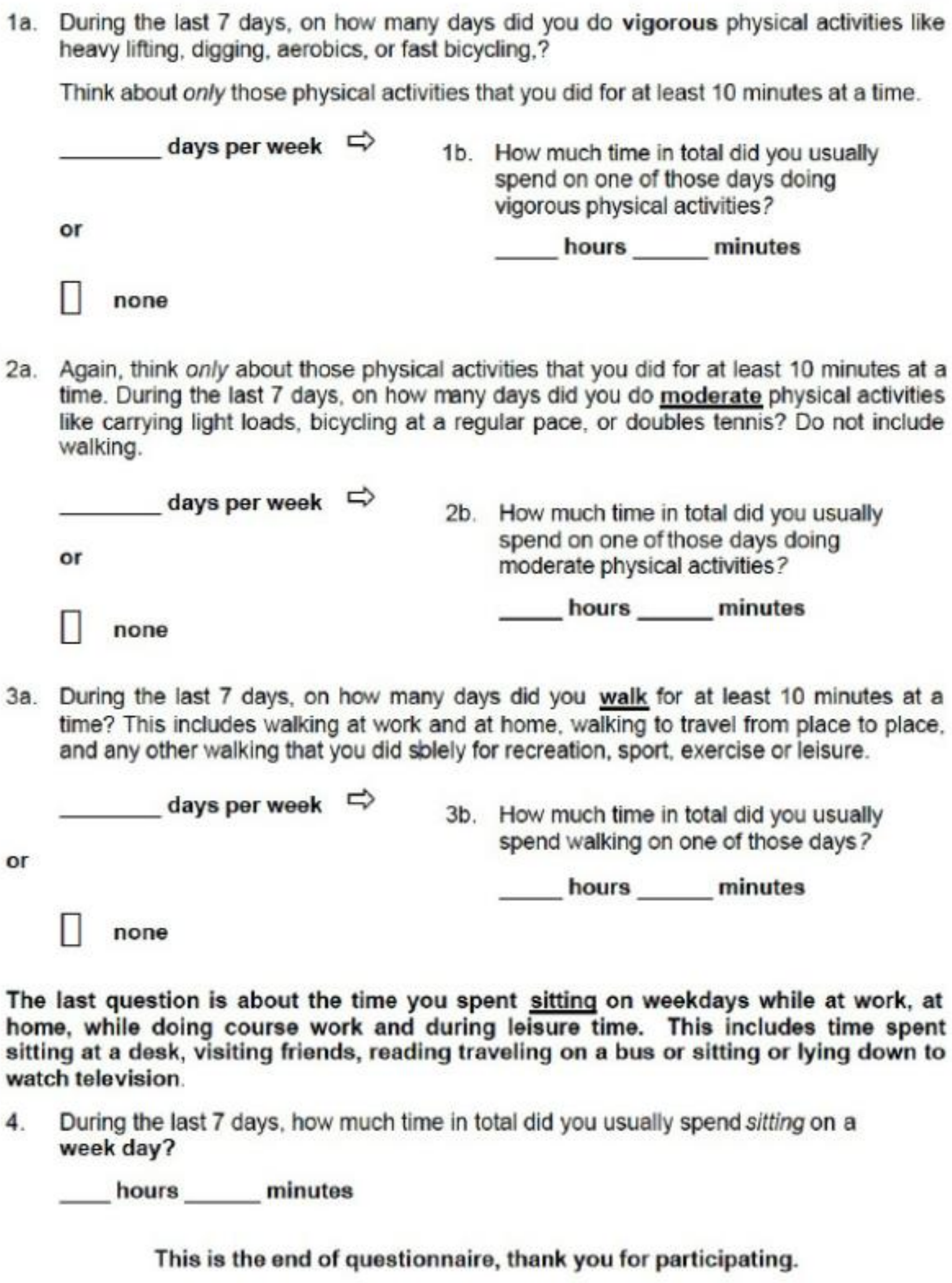
| | |
|---|---|
| Yes | (1 point) |
| No | (0 points) |

**Scoring:** 7–10 points = highly dependent; 4–6 points = moderately dependent; less than 4 points = minimally dependent.

[3] Fagerstrom Test for Nicotine Dependence - an overview | ScienceDirect Topics [Internet]. [cited Feb 16, 2023]. Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/fagerstrom-test-for-nicotine-dependence

## Abbreviated version of the International Physical Activity Questionnaire[4]

IPAQ (International Physival Activity Questionnaire Short Form)



1a. During the last 7 days, on how many days did you do **vigorous** physical activities like heavy lifting, digging, aerobics, or fast bicycling,?

Think about *only* those physical activities that you did for at least 10 minutes at a time.

_________ **days per week** ⇨

**or**

□ **none**

1b. How much time in total did you usually spend on one of those days doing vigorous physical activities?

_____ **hours** ______ **minutes**

2a. Again, think *only* about those physical activities that you did for at least 10 minutes at a time. During the last 7 days, on how many days did you do **moderate** physical activities like carrying light loads, bicycling at a regular pace, or doubles tennis? Do not include walking.

_________ **days per week** ⇨

**or**

□ **none**

2b. How much time in total did you usually spend on one of those days doing moderate physical activities?

_____ **hours** ______ **minutes**

3a. During the last 7 days, on how many days did you **walk** for at least 10 minutes at a time? This includes walking at work and at home, walking to travel from place to place, and any other walking that you did sblely for recreation, sport, exercise or leisure.

_________ **days per week** ⇨

**or**

□ **none**

3b. How much time in total did you usually spend walking on one of those days?

_____ **hours** ______ **minutes**

**The last question is about the time you spent sitting on weekdays while at work, at home, while doing course work and during leisure time. This includes time spent sitting at a desk, visiting friends, reading traveling on a bus or sitting or lying down to watch television.**

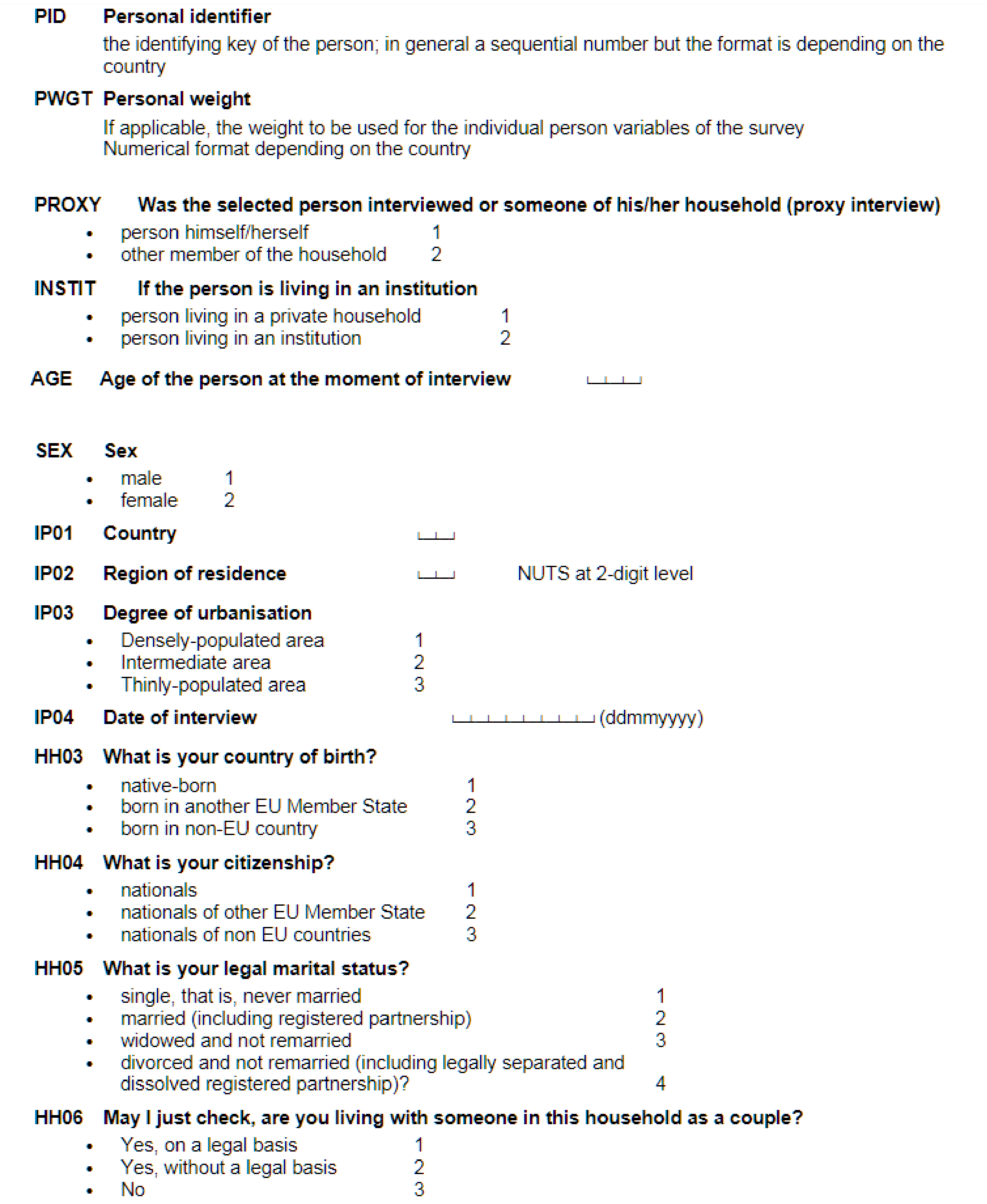
4. During the last 7 days, how much time in total did you usually spend *sitting* on a **week day?**

____ **hours** ______ **minutes**

**This is the end of questionnaire, thank you for participating.**

[4] Cardol M, de Haan RJ, de Jong BA, van den Bos GA, de Groot IJ. Psychometric properties of the Impact on Participation and Autonomy Questionnaire. Arch Phys Med Rehabil. 2001 Feb;82(2):210–6. https://www.researchgate.net/figure/International-Physical-Activity-Questionnaire-short-form-IPAQ-SF_fig1_275642934

## European Health Interview Survey 2019 – Questionnaire (relevant parts)[5]



**PID** **Personal identifier**

the identifying key of the person; in general a sequential number but the format is depending on the country

**PWGT** **Personal weight**

If applicable, the weight to be used for the individual person variables of the survey
Numerical format depending on the country

**PROXY** **Was the selected person interviewed or someone of his/her household (proxy interview)**

- person himself/herself 1
- other member of the household 2

**INSTIT** **If the person is living in an institution**

- person living in a private household 1
- person living in an institution 2

**AGE** **Age of the person at the moment of interview**

**SEX** **Sex**

- male 1
- female 2

**IP01** **Country**

**IP02** **Region of residence** NUTS at 2-digit level

**IP03** **Degree of urbanisation**

- Densely-populated area 1
- Intermediate area 2
- Thinly-populated area 3

**IP04** **Date of interview** (ddmmyyyy)

**HH03** **What is your country of birth?**

- native-born 1
- born in another EU Member State 2
- born in non-EU country 3

**HH04** **What is your citizenship?**

- nationals 1
- nationals of other EU Member State 2
- nationals of non EU countries 3

**HH05** **What is your legal marital status?**

- single, that is, never married 1
- married (including registered partnership) 2
- widowed and not remarried 3
- divorced and not remarried (including legally separated and dissolved registered partnership)? 4

**HH06** **May I just check, are you living with someone in this household as a couple?**

- Yes, on a legal basis 1
- Yes, without a legal basis 2
- No 3

[5] European Health Interview Survey - Microdata - Eurostat [Internet]. [cited Feb 16, 2023]. Available from: https://ec.europa.eu/eurostat/web/microdata/european-health-interview-survey

**HH07 What is the highest education leaving certificate, diploma or education degree you have obtained? Please include any vocational training.**

1

| | |
|---|---|
| • no formal education or below ISCED 1 | 1 |
| • primary education (ISCED 1) | 2 |
| • lower secondary education (ISCED 2) | 3 |
| • upper secondary education (ISCED 3) | 4 |
| • post-secondary but non-tertiary education (ISCED 4) | 5 |
| • first stage of tertiary education (ISCED 5) | 6 |
| • second stage of tertiary education (ISCED 6) | 7 |

**HH08 How would you define your current labour status?**

| | |
|---|---|
| • working for pay or profit (including unpaid work for a family business or holding, including an apprenticeship or paid traineeship, including currently not at work due to maternity, parental, sick leave or holidays) | 1 |
| • unemployed | 2 |
| • pupil, student, further training, unpaid work experience | 3 |
| • in retirement or early retirement or has given up business | 4 |
| • permanently disabled | 5 |
| • in compulsory military or community service | 6 |
| • fulfilling domestic tasks | 7 |
| • other | 8 |

**HH09 Have you ever worked for pay or profit?**

| | |
|---|---|
| • Yes | 1 |
| • No | 2 |

**HH10 Are (Were) you an employee, self-employed or working without payment as a family worker?**

| | |
|---|---|
| • employee | 1 |
| • self-employed | 2 |
| • family worker | 3 |

**HH11 What type of work contract do (did) you have?**

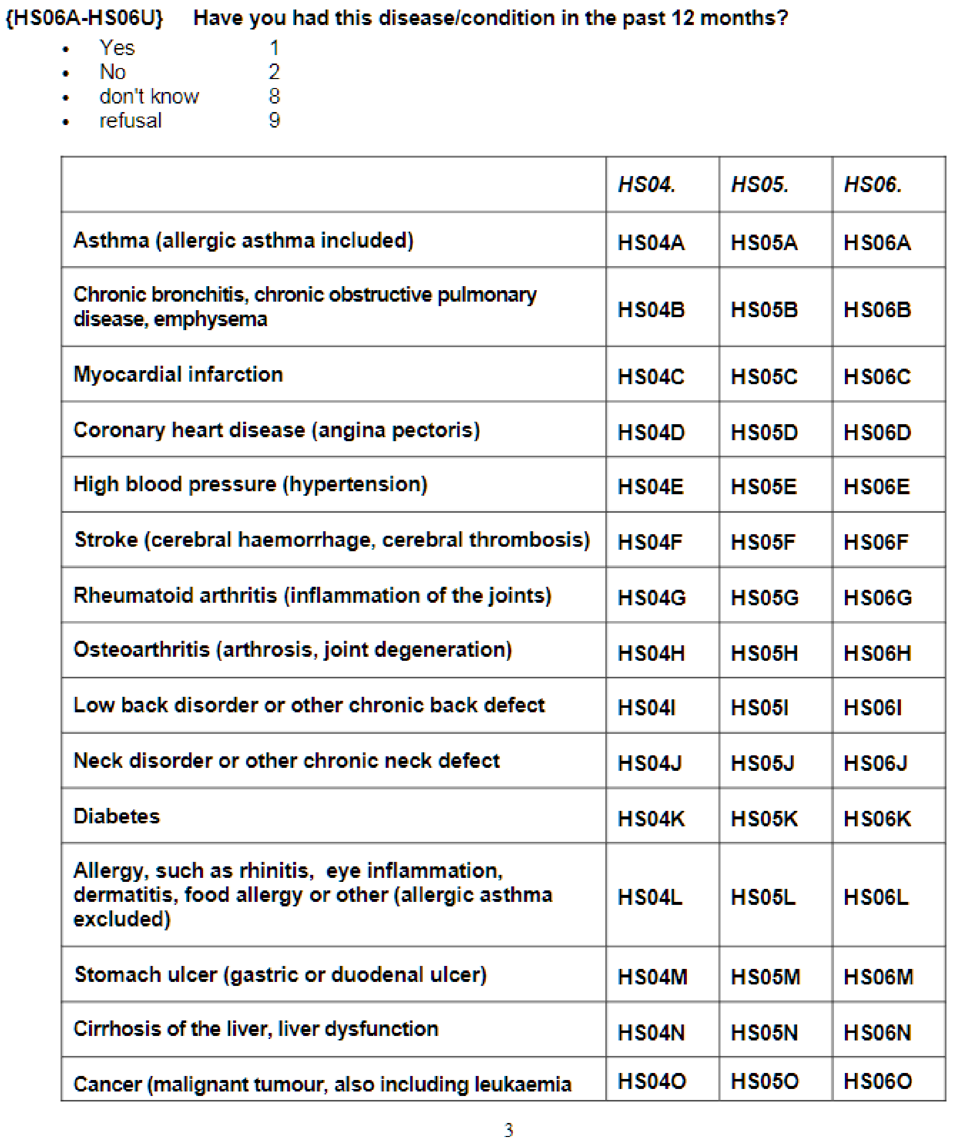
| | |
|---|---|
| • permanent job/work contract of unlimited duration | 1 |
| • temporary job/work contract of limited duration | 2 |

**HH12 In your (main) job do (did) you work full-time or part-time?**

| | |
|---|---|
| • full-time | 1 |
| • part-time | 2 |

**HH13 What is (was) your occupation in this job?**

ı ı ı ISCO-88 COM, 2 digits



**{HS06A-HS06U} Have you had this disease/condition in the past 12 months?**

| | |
|---|---|
| Yes | 1 |
| No | 2 |
| don't know | 8 |
| refusal | 9 |

| | *HS04.* | *HS05.* | *HS06.* |
|---|---|---|---|
| Asthma (allergic asthma included) | HS04A | HS05A | HS06A |
| Chronic bronchitis, chronic obstructive pulmonary disease, emphysema | HS04B | HS05B | HS06B |
| Myocardial infarction | HS04C | HS05C | HS06C |
| Coronary heart disease (angina pectoris) | HS04D | HS05D | HS06D |
| High blood pressure (hypertension) | HS04E | HS05E | HS06E |
| Stroke (cerebral haemorrhage, cerebral thrombosis) | HS04F | HS05F | HS06F |
| Rheumatoid arthritis (inflammation of the joints) | HS04G | HS05G | HS06G |
| Osteoarthritis (arthrosis, joint degeneration) | HS04H | HS05H | HS06H |
| Low back disorder or other chronic back defect | HS04I | HS05I | HS06I |
| Neck disorder or other chronic neck defect | HS04J | HS05J | HS06J |
| Diabetes | HS04K | HS05K | HS06K |
| Allergy, such as rhinitis, eye inflammation, dermatitis, food allergy or other (allergic asthma excluded) | HS04L | HS05L | HS06L |
| Stomach ulcer (gastric or duodenal ulcer) | HS04M | HS05M | HS06M |
| Cirrhosis of the liver, liver dysfunction | HS04N | HS05N | HS06N |
| Cancer (malignant tumour, also including leukaemia | HS04O | HS05O | HS06O |

**{SF02-SF10} How much of the time, during the past 4 weeks…**

| | | All of the time | Most of the time | Some of the time | A little of the time | None of the time | Don't know | Refus al |
|---|---|---|---|---|---|---|---|---|
| **SF02** | **Did you feel full of life?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF03** | **Have you been very nervous?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF04** | **Have you felt so down in the dumps that nothing could cheer you up?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF05** | **Have you felt calm and peaceful?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF06** | **Did you have a lot of energy?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF07** | **Have you felt down-hearted and depressed?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF08** | **Did you feel worn out?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF09** | **Have you been happy?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
| **SF10** | **Did you feel tired?** | 1 | 2 | 3 | 4 | 5 | 8 | 9 |

**HC01 During the past 12 months, that is since (date one year ago), have you been in hospital as an inpatient, that is overnight or longer?**

| | |
|---|---|
| • Yes | 1 |
| • No | 2 |
| • don't know | 8 |
| • refusal | 9 |

**HC02 How many separate stays in hospital as an inpatient have you had since (date one year ago)? Count all the stays that ended in this period.**

| | |
|---|---|
| • number of stays | └┴┘ |
| • don't know | 98 |
| • refusal | 99 |

**HC03 Thinking of this/these inpatient stay(s), how many nights in total did you spend in hospital?**

| | |
|---|---|
| • number of nights | └┴┴┘ |
| • don't know | 998 |
| • refusal | 999 |

**HC04 During the past 12 months, that is since (date one year ago), have you been admitted to hospital as a day patient, that is admitted to a hospital bed, but not required to remain overnight?**

| | |
|---|---|
| • Yes | 1 |
| • No | 2 |
| • don't know | 8 |
| • refusal | 9 |

**HC05 How many days have you been admitted as a day patient since (date one year ago)?**

| | |
|---|---|
| • number of days | └┴┴┘ |
| • don't know | 998 |
| • refusal | 999 |

**HC06 During the past 12 months, was there any time when you really needed to be hospitalised following a recommendation from a doctor, either as an inpatient or a day patient, but did not?**

| | |
|---|---|
| • Yes, there was at least one occasion | 1 |
| • No, there was no occasion | 2 |
| • don't know | 8 |
| • refusal | 9 |

**HC07 What was the main reason for not being hospitalised?**

| | |
|---|---|
| • Could not afford to (too expensive or not covered by the insurance fund) | 1 |
| • Waiting list, other reasons due to the hospital | 2 |
| • Could not take time because of work, care for children or for others | 3 |
| • Too far to travel / no means of transportation | 4 |
| • Fear of surgery / treatment | 5 |
| • Other reason | 6 |
| • don't know | 8 |
| • refusal | 9 |

**PA01 Have you ever been vaccinated against flu?**

- Yes 1
- No 2
- don't know 8
- refusal 9

**PA02 When were you last time vaccinated against flu?**

- Since the beginning of this year 1
- Last year 2
- Before last year 3
- don't know 8
- refusal 9

**PA03 Can I just check, what month was that?**

- ⌴⌴ Month (01 …12;
- Don't know 99

**PA04 Has your blood pressure ever been measured by a health professional?**

- Yes 1
- No 2
- don't know 8
- refusal 9

**PA05 When was the last time that your blood pressure was measured by a health professional?**

- Within the past 12 months 1
- 1-5 years ago 2
- More than 5 years ago 3
- don't know 8
- refusal 9

**PA06 Has your blood cholesterol ever been measured?**

- Yes 1
- No 2
- don't know 8
- refusal 9

**PA07 When was the last time that your blood cholesterol was measured?**

- Within the past 12 months 1
- 1-5 years ago 2
- More than 5 years ago 3
- don't know 8
- refusal 9

**PA08 Has your blood sugar ever been measured?**

- Yes 1
- No 2
- don't know 8
- refusal 9

**PA12 and {PA12A-PA12E} What was the reason for this last mammography?**

| | |
|---|---|
| Reasons specified | 1 |
| Don't know | 8 |
| Refusal | 9 |

**if PA12=1 ("reasons specified") then {PA12A-PA12E}**

**PA12A Myself or my GP/family doctor or a specialist noticed something not quite right in my breast (e.g. a lump)**

| | |
|---|---|
| Yes | 1 |
| no | 2 |

**PA12B My GP/family doctor or a specialist advised me to have it without there being something wrong**

| | |
|---|---|
| Yes | 1 |
| no | 2 |

**PA12C Because of breast cancer in my family**

| | |
|---|---|
| Yes | 1 |
| no | 2 |

**PA12D Invitation from a national or local screening programme**

| | |
|---|---|
| Yes | 1 |
| no | 2 |

**PA12E Other reason**

| | |
|---|---|
| Yes | 1 |
| no | 2 |

If PA12 equals 1 (reasons specified) and {PA12A-PA12I} is not ticked, we consider the answer for {PA12A-PA12I} as a "No".

**PA13 Have you ever had a cervical smear test?**

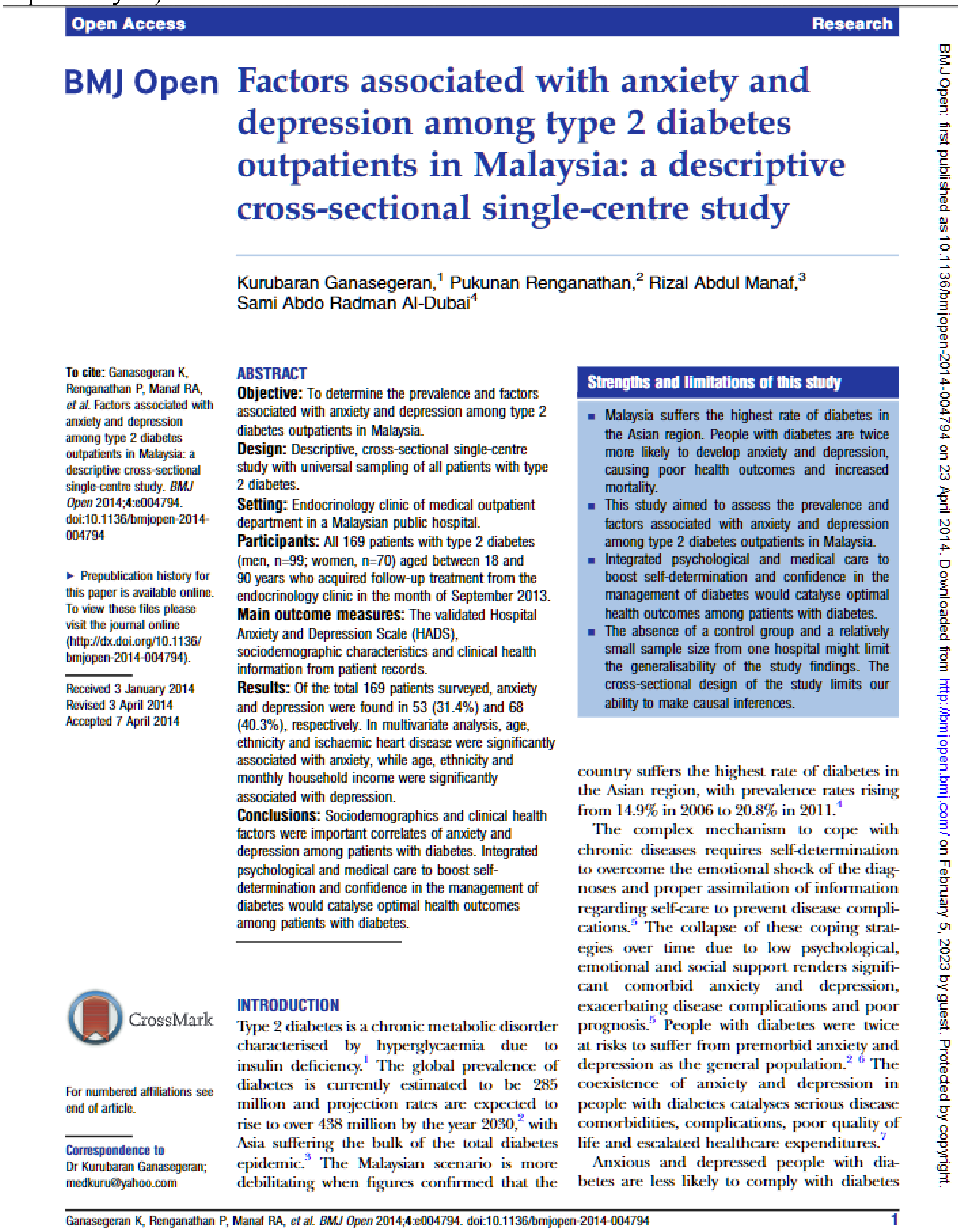
| | |
|---|---|
| Yes | 1 |
| No | 2 |
| don't know | 8 |
| refusal | 9 |

**PA14 When was the last time you had a cervical smear test?**

| | |
|---|---|
| Within the past 12 months | 1 |
| More than 1 year, but not more than 2 years | 2 |
| More than 2 years, but not more than 3 years | 3 |
| Not within the past 3 years | 4 |
| don't know | 8 |
| refusal | 9 |

## Scientific publication

The results of correct construction are demonstrated in the article below (descriptive, simple, multiple analysis).[6]



Open Access Research

BMJ Open

# Factors associated with anxiety and depression among type 2 diabetes outpatients in Malaysia: a descriptive cross-sectional single-centre study

Kurubaran Ganasegeran,[1] Pukunan Renganathan,[2] Rizal Abdul Manaf,[3] Sami Abdo Radman Al-Dubai[4]

**To cite:** Ganasegeran K, Renganathan P, Manaf RA, *et al*. Factors associated with anxiety and depression among type 2 diabetes outpatients in Malaysia: a descriptive cross-sectional single-centre study. *BMJ Open* 2014;4:e004794. doi:10.1136/bmjopen-2014-004794

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2014-004794).

Received 3 January 2014
Revised 3 April 2014
Accepted 7 April 2014

### ABSTRACT

**Objective:** To determine the prevalence and factors associated with anxiety and depression among type 2 diabetes outpatients in Malaysia.

**Design:** Descriptive, cross-sectional single-centre study with universal sampling of all patients with type 2 diabetes.

**Setting:** Endocrinology clinic of medical outpatient department in a Malaysian public hospital.

**Participants:** All 169 patients with type 2 diabetes (men, n=99; women, n=70) aged between 18 and 90 years who acquired follow-up treatment from the endocrinology clinic in the month of September 2013.

**Main outcome measures:** The validated Hospital Anxiety and Depression Scale (HADS), sociodemographic characteristics and clinical health information from patient records.

**Results:** Of the total 169 patients surveyed, anxiety and depression were found in 53 (31.4%) and 68 (40.3%), respectively. In multivariate analysis, age, ethnicity and ischaemic heart disease were significantly associated with anxiety, while age, ethnicity and monthly household income were significantly associated with depression.

**Conclusions:** Sociodemographics and clinical health factors were important correlates of anxiety and depression among patients with diabetes. Integrated psychological and medical care to boost self-determination and confidence in the management of diabetes would catalyse optimal health outcomes among patients with diabetes.

### Strengths and limitations of this study

- Malaysia suffers the highest rate of diabetes in the Asian region. People with diabetes are twice more likely to develop anxiety and depression, causing poor health outcomes and increased mortality.
- This study aimed to assess the prevalence and factors associated with anxiety and depression among type 2 diabetes outpatients in Malaysia.
- Integrated psychological and medical care to boost self-determination and confidence in the management of diabetes would catalyse optimal health outcomes among patients with diabetes.
- The absence of a control group and a relatively small sample size from one hospital might limit the generalisability of the study findings. The cross-sectional design of the study limits our ability to make causal inferences.

### INTRODUCTION

Type 2 diabetes is a chronic metabolic disorder characterised by hyperglycaemia due to insulin deficiency.[1] The global prevalence of diabetes is currently estimated to be 285 million and projection rates are expected to rise to over 438 million by the year 2030,[2] with Asia suffering the bulk of the total diabetes epidemic.[3] The Malaysian scenario is more debilitating when figures confirmed that the country suffers the highest rate of diabetes in the Asian region, with prevalence rates rising from 14.9% in 2006 to 20.8% in 2011.[4]

The complex mechanism to cope with chronic diseases requires self-determination to overcome the emotional shock of the diagnoses and proper assimilation of information regarding self-care to prevent disease complications.[5] The collapse of these coping strategies over time due to low psychological, emotional and social support renders significant comorbid anxiety and depression, exacerbating disease complications and poor prognosis.[5] People with diabetes were twice at risks to suffer from premorbid anxiety and depression as the general population.[2 6] The coexistence of anxiety and depression in people with diabetes catalyses serious disease comorbidities, complications, poor quality of life and escalated healthcare expenditures.[7]

Anxious and depressed people with diabetes are less likely to comply with diabetes

For numbered affiliations see end of article.

**Correspondence to**
Dr Kurubaran Ganasegeran; medkuru@yahoo.com

[6] Ganasegeran K, Renganathan P, Manaf RA, Al-Dubai SAR. Factors associated with anxiety and depression among type 2 diabetes outpatients in Malaysia: a descriptive cross-sectional single-centre study. BMJ Open. 2014 Apr 23;4(4):e004794.

self-care recommendations.[6] The diagnosis of diabetes is a life-threatening stressor that demands high mental and physical accommodations due to elevated feelings of fear.[8] Depression among people with diabetes adds an increased burden to patient adherence, compliance and poor prognosis for quality health outcomes.[9] Depression in the diabetes population has been associated with potential sociodemographic and clinical factors.[7] Ageing,[2] ethnicity,[8] socioeconomic status,[10] education level[11] and unemployment[12] were important correlates for depression among people with diabetes.

Common diabetes vascular complications like ischaemic heart disease (IHD), cerebrovascular accidents (CVAs) and diabetic nephropathy had caused significant rates of mortality and poor quality of life.[2,11] Malaysia topped the world in diabetic nephropathy, with almost 15 000 patients requiring dialysis and 2000 acquiring kidney transplants.[13] Diabetes-related complications and associated comorbidities have been proven to amplify psychiatric conditions.[2]

Numerous studies from developed and developing countries assessed factors affecting anxiety and depression among people with diabetes.[2,6,14] Irish and Mexican studies concluded that the prevalence of anxiety and depression was considerably higher among people with diabetes in comparison to the general population.[6,9] A Malaysian study recommended that early psychiatric screening was required owing to elevated risks for anxiety and depression among people with diabetes.[8] This study aimed to determine the prevalence and factors associated with anxiety and depression among outpatients with diabetes in a Malaysian public hospital.

## METHODS

### Study setting and population

This cross-sectional single-centre study was conducted in the month of September 2013 among all 169 patients with type 2 diabetes aged between 18 and 90 years who acquired follow-up treatment from the Endocrinology Clinic at the Medical Outpatient Department of Tengku Ampuan Rahimah Hospital (HTAR), Selangor, Malaysia. Objectives and benefits of the study were explained in verbal and written form attached to the questionnaires. Patients were assured that their participation was confidential and would not affect their medical treatment outcomes. A written consent was obtained from those who agreed to participate. Patients with type 1 and gestational diabetes were excluded from the study.

### Ethical issues

This study complied with the guidelines convened in the Declaration of Helsinki. The study was conducted as part of a larger study that explored anxiety and depression among outpatients in Malaysia.

### Study instruments

A self-administered questionnaire consisting of three parts was used in this study:

The first part included items on sociodemographics (gender, age, ethnicity, marital status, education level, residence, monthly household income and employment status).

The second part assessed anxiety and depression among patients with diabetes. Anxiety is defined as subjective experience of fear and its' physical manifestations while depression is defined as anhedonia (reduced positive affect).[15] To explore anxiety and depression among patients with diabetes, we used the Hospital Anxiety and Depression Scale (HADS), originally developed by Zigmond and Snaith,[16] and validated among the Malaysian population.[17] This widely used self-assessment tool measures the level of emotional distress (anxiety and depression) in various clinical settings, including diabetes population.[2,6,18] HADS is comprised of 14 items, 7 of which measures anxiety (HADS-A) and another 7 measures depression (HADS-D). These items are scored on a four-point Likert scale ranging from 0 (not present) to 3 (considerable). Item scores were summed to provide subscaled scores of anxiety and depression, ranged between 0 and 21, and total summed score ranged from 0 to 42. A higher score represents higher anxiety or depression.[18] The scores are categorised as follows: normal (0–7) and caseness which includes mild distress (8–10), moderate distress (11–14) and severe distress (15–21).[18] The questionnaire was administered in English.

The third part included clinical health information of the patients derived from medical records.

### Baseline data definitions

#### Type 2 diabetes

The presence of diabetes diagnosed by a physician with a fasting plasma glucose value of 7 mmol/L (126 mg/dL) or higher,[19] and patients currently being administered with oral hypoglycaemic drugs or insulin therapy as documented in medical records were included in this study.

#### Diabetes vascular complications

Vascular complications in diabetes were considered when patients were diagnosed with CVA, IHD or nephropathies. Patients diagnosed with either one vascular complication over the past year were included in this study. CVA was defined as hemiparesis cases proven by medical and CT scan records.[20] IHD was considered to exist with a history of angina or acute coronary syndromes elicited among patients with diabetes and documented in medical records.[2] Nephropathy is defined by proteinuria >500 mg in 24 h among patients with diabetes from medical records.[1]

#### Diabetes comorbid conditions

Patients with diabetes were classified as hypertensive if they were previously diagnosed and were currently on antihypertensive medications[2] as documented in medical records. Dyslipidaemia was defined as high plasma triglyceride concentration, low high-density lipoprotein cholesterol concentration and increased concentration of low-density lipoprotein cholesterol[21] with patients currently on statin medications as documented in medical records.

### The Malaysian healthcare system

Public healthcare providers across the nation are mainly entrusted by the Ministry of Health Malaysia with the commitment of 'healthcare access to all'.[4] The public healthcare is highly subsidised through general revenue and taxation collected by the federal government, and with a minimal registration fee of US$0.33 or MYR1. Malaysians are granted free access to clinical consultations, treatment and medications both as outpatients or inpatients in all public health facilities within the country.[4] HTAR is the second busiest public health facility in terms of patient admissions and outpatient services in Malaysia at the time of this study.[4]

### Statistical analysis

Analysis was performed using Statistical Package for Social Sciences (SPSS) (V.16.0, IBM, Armonk, New York, USA). Descriptive analysis was performed for all variables in this study. To check for the validity of the HADS among Malaysian population, an exploratory factor analysis was performed using principal component method with varimax rotation and Cronbach's α was used to test the internal consistency of the scale. Anxiety and depression scores were expressed as mean and SDs. Test of normality was performed for total anxiety and depression subscale scores. T test and analysis of variance (ANOVA) test were applied to compare anxiety and depression across sociodemographic and clinical health variables. In case of ANOVA, post hoc test was used to determine where the significant difference was. Multiple linear regression analysis using 'Enter' technique was employed to obtain significant factors associated with anxiety and depression among patients with diabetes. The accepted level of significance was set below 0.05 (p<0.05). Multicollinearity was checked between independent variables.

## RESULTS

### Sociodemographic characteristics and clinical health information of the respondents

One hundred sixty-nine patients were included in this study. Of the total, 99 (58.6%) were men and 70 (41.4%) were women. The mean (±SD) age of the patients was 36.9 (±15.9) years and the majority aged less than 50 years, 137 (81.1%; table 1).

Baseline clinical data of the patients are summarised in table 2. Of the total patients, 53 (31.4%) were diagnosed for diabetes vascular complications. Twelve patients (7.1%) were diagnosed for CVA, 24 (14.2%) were diagnosed for IHD and 17 (10.1%) developed nephropathy. Forty-four (26.0%) patients developed at least one comorbid condition, while 21 (12.4%) had two comorbid conditions. Cronbach's α coefficient for HADS-A subscale was 0.83, while Cronbach's α coefficient for HADS-D subscale was 0.71. Mild anxiety and depression were found in 33 (19.5%) and 49 (29.0%) of the patients, respectively. Moderate anxiety and depression were found in 16 (9.5%) patients respectively. Severe anxiety and depressive symptoms were detected in four (2.4%) and three (1.8%) of the patients, respectively.

Table 1 Sociodemographic characteristics of the respondents (n=169)

| Characteristics | N | Percentage |
|---|---|---|
| Gender | | |
| Male | 99 | 58.6 |
| Female | 70 | 41.4 |
| Age (years) | | |
| <50 | 137 | 81.1 |
| ≥50 | 32 | 18.9 |
| Ethnicity | | |
| Malay | 53 | 31.3 |
| Chinese | 88 | 52.1 |
| Indian | 28 | 16.6 |
| Marital status | | |
| Single | 63 | 37.3 |
| Married | 106 | 62.7 |
| Highest education level | | |
| High school | 75 | 44.4 |
| Tertiary education | 94 | 55.6 |
| Residence | | |
| Urban | 132 | 78.1 |
| Rural | 37 | 21.9 |
| Monthly household income (MYR) | | |
| <3000 | 56 | 33.1 |
| ≥3000 | 113 | 66.9 |
| Current employment status | | |
| Employed | 119 | 70.4 |
| Unemployed | 50 | 29.6 |

MYR1 is equivalent to US$0.33 at the time of study.

### Association between anxiety and depression and sociodemographics of the respondents

Table 3 shows the association between anxiety and depression with sociodemographic characteristics. Patients aged 50 years or older had higher anxiety score (9.1±4.6) compared with those aged less than 50 years (6.4±2.7, p<0.001). Significant associations were observed between anxiety and ethnicity (p<0.001); post hoc tests showed that Indians exhibited higher anxiety score (8.4±4.2) in comparison to Chinese (6.6±2.4, p=0.044). Patients graduated from high school exhibited higher anxiety score (7.5±4.0) in comparison to those with a tertiary degree (6.4±2.5, p=0.037). In addition, patients aged 50 years or older were more depressed (9.2±4.0) in comparison to those aged less than 50 years (6.3±2.9, p<0.001). Significant associations were observed between depression and ethnicity (p<0.001);

**Table 2** Clinical health information of the respondents (n=169)

| Characteristics | N | Percentage |
|---|---|---|
| *Diabetes vascular complications* | | |
| Cerebrovascular accident | | |
| Yes | 12 | 7.1 |
| No | 157 | 92.9 |
| Ischaemic heart disease | | |
| Yes | 24 | 14.2 |
| No | 145 | 85.8 |
| Diabetic nephropathy | | |
| Yes | 17 | 10.1 |
| No | 152 | 89.9 |
| Comorbidities | | |
| Diabetes alone | 104 | 61.6 |
| Diabetes with hypertension or dyslipidaemia | 44 | 26.0 |
| Diabetes with hypertension and dyslipidaemia | 21 | 12.4 |
| Anxiety | | |
| Normal | 116 | 68.6 |
| Mild | 33 | 19.5 |
| Moderate | 16 | 9.5 |
| Severe | 4 | 2.4 |
| Depression | | |
| Normal | 101 | 59.7 |
| Mild | 49 | 29.0 |
| Moderate | 16 | 9.5 |
| Severe | 3 | 1.8 |

**Table 3** Association between anxiety and depression with sociodemographic characteristics of the respondents (n=169)

| | Anxiety | | Depression | |
|---|---|---|---|---|
| Characteristics | Mean (SD) | p Value | Mean (SD) | p Value |
| Gender | | | | |
| Male | 7.0 (3.5) | | 7.1 (3.5) | |
| Female | 6.8 (3.0) | 0.737 | 6.6 (3.1) | 0.345 |
| Age (years) | | | | |
| <50 | 6.4 (2.7) | | 6.3 (2.9) | |
| ≥50 | 9.1 (4.6) | <0.001 | 9.2 (4.0) | <0.001 |
| Ethnicity | | | | |
| Malay | 6.5 (3.8) | | 6.9 (3.1) | |
| Chinese | 6.6 (2.4) | | 5.9 (2.9) | |
| Indian | 8.4 (4.2) | 0.035 | 9.8 (3.5) | <0.001 |
| Marital status | | | | |
| Single | 6.7 (2.7) | | 6.9 (2.9) | |
| Married | 7.0 (3.6) | 0.601 | 6.8 (3.6) | 0.894 |
| Highest education level | | | | |
| High school | 7.5 (4.0) | | 7.7 (3.7) | |
| Tertiary education | 6.4 (2.5) | 0.037 | 6.2 (2.9) | 0.006 |
| Residence | | | | |
| Urban | 6.8 (2.9) | | 6.7 (3.1) | |
| Rural | 7.2 (4.5) | 0.569 | 7.6 (3.9) | 0.125 |
| Monthly household income (MYR) | | | | |
| <3000 | 7.5 (4.4) | | 8.7 (3.6) | |
| ≥3000 | 6.6 (2.6) | 0.090 | 6.0 (2.8) | <0.001 |
| Current employment status | | | | |
| Employed | 6.6 (3.2) | | 6.4 (3.3) | |
| Unemployed | 7.6 (3.4) | 0.078 | 7.9 (3.2) | 0.007 |

post hoc tests revealed that Indians exhibited higher depression (9.8±3.5) in comparison to Malays (6.9±3.1) and Chinese (5.9±2.9, p<0.001, respectively). Similarly, patients who graduated from high school exhibited greater depression (7.7±3.7) in comparison to tertiary graduates (6.2±2.9, p=0.006). Patients with a monthly household income of less than MYR3000 have higher depression score (8.7±3.6) compared to those with higher income (6.0±2.8, p<0.001). Similarly, unemployed patients portrayed higher depression score (7.9±3.2) in comparison to those employed (6.4±3.3, p=0.007).

### Association between anxiety and depression and clinical health information of the respondents

Patients diagnosed for IHD exhibited higher anxiety score (8.7±4.2) in comparison to those without such complication (6.6±3.1, p=0.004). In addition, significant associations were observed between depression and disease comorbidities (p=0.010); post hoc tests showed that patients with associated hypertension or dyslipidaemia had higher depression score (7.5±3.2) in comparison to those without comorbid conditions (6.3±3.4, p=0.009; table 4).

### Factors associated with anxiety among patients with diabetes by multiple linear regression

Table 5 exhibits the factors associated with anxiety among patients with diabetes. Patients aged ≥50 years had on the average 2.3 (95% CI 0.9 to 3.6) higher anxiety score in comparison to those aged less than 50 years (p=0.001). Indians had on an average 1.7 (95% CI 0.3 to 3.2) higher anxiety score compared with Malays (p=0.018). Patients diagnosed with IHD had on an average 1.5 (95% CI 0.1 to 2.9) higher anxiety score in comparison to those without such a condition (p=0.039).

### Factors associated with depression among patients with diabetes by multiple linear regression

Table 6 shows the factors associated with depression among patients with diabetes. Patients aged ≥50 years had on the average 1.4 (95% CI 0.2 to 2.7) higher depression score in comparison to those aged less than 50 years (p=0.027). Indians had on the average 2.7 (95% CI 1.4 to 4.0) higher depression score compared with Chinese (p<0.001). Patients with a monthly household income of less than MYR3000 had on an average 1.9 (95% CI 0.8 to 3.0) higher depression score compared to those with a higher income (p=0.001).

## DISCUSSION

This study aimed to determine the prevalence and factors associated with anxiety and depression among

**Table 4** Association between anxiety and depression with clinical health information of the respondents (n=169)

| | Anxiety | | Depression | |
|---|---|---|---|---|
| Characteristics | Mean (SD) | p Value | Mean (SD) | p Value |
| *Diabetes vascular complications* | | | | |
| Cerebrovascular accident | | | | |
| Yes | 6.6 (4.1) | | 6.7 (4.8) | |
| No | 6.9 (3.3) | 0.742 | 6.9 (3.2) | 0.823 |
| Ischaemic heart disease | | | | |
| Yes | 8.7 (4.2) | | 7.8 (4.1) | |
| No | 6.6 (3.1) | 0.004 | 6.7 (3.2) | 0.131 |
| Diabetic nephropathy | | | | |
| Yes | 7.4 (2.2) | | 6.4 (3.0) | |
| No | 6.8 (3.4) | 0.492 | 6.9 (3.4) | 0.548 |
| Comorbidities | | | | |
| Diabetes alone | 6.6 (3.3) | | 6.3 (3.4) | |
| Diabetes with hypertension or dyslipidaemia | 7.2 (3.4) | | 7.5 (3.2) | |
| Diabetes with hypertension and dyslipidaemia | 7.7 (3.0) | 0.289 | 8.4 (2.9) | 0.010 |

diabetes outpatients in Malaysia. Of the 169 patients with diabetes surveyed, 31.4% perceived anxious states while 40.3% exhibited depressive symptoms. The estimated rate of anxiety reported in this study was similar to an Irish sample (32%),[6] but relatively lower than that found in Mexican (52.9%)[9] and Pakistani (57.9%) participants.[2] In contrary, self-reported depression rates reported in this study were similar than that found in Mexican (47.7%)[9] and Pakistani (43.5%) samples,[2] but comparatively higher than that found in Irish participants (22.4%).[6] In the final model, age, ethnicity and history of IHD were significantly associated with anxiety, while factors significantly associated with depression were age, ethnicity and monthly household income.

Ageing appears to accelerate diabetes vascular complications and hyperglycaemic crisis, causing poor functional status and high mortality rates.[22] Dysregulation of the hypothalamic-pituitary-adrenal axis and overactivation of the sympathetic nervous system due to fear of hypoglycaemia, complications or mortality are immediate physiological processes that prompt higher anxiety states among older population.[5] This study found a significantly higher anxious state among older patients compared with younger ones. Collins *et al*[6] reported otherwise.

The development of vascular complications is a predictive factor for psychological morbidity among people with diabetes.[23] This study found a significantly higher anxiety level among patients with IHD. Khuwaja *et al*[2] reported similar associations.

The increased susceptibility to various diseases, disabilities and social isolation among older population causes serious psychological repercussions.[24] This study found a significantly higher depression score among older patients in comparison to younger ones. Similar findings were found among patients with diabetes in other countries.[2 25]

Latest statistics revealed by the Ministry of Health Malaysia reported that the prevalence of diabetes was the highest among Indian ethnic (24.9%), followed by Malay ethnic (17%) and Chinese ethnic (13.9%).[4] Minority ethnic groups have been known to experience higher anxiety and depression rates.[26 27] This study found a

**Table 5** Factors associated with anxiety among patients with diabetes by multiple linear regression (n=169)

| | | | | | 95% CI | |
|---|---|---|---|---|---|---|
| Predictors | B | SE | β | p Value | Lower | Upper |
| Age (years) | | | | | | |
| <50 | Ref | Ref | Ref | Ref | Ref | Ref |
| ≥50 | 2.3 | 0.7 | 0.3 | 0.001 | 0.9 | 3.6 |
| Ethnicity | | | | | | |
| Malay | Ref | Ref | Ref | Ref | Ref | Ref |
| Chinese | 0.7 | 0.6 | 0.1 | 0.194 | −0.4 | 1.8 |
| Indian | 1.7 | 0.7 | 0.2 | 0.018 | 0.3 | 3.2 |
| Highest education level | | | | | | |
| High school | 0.1 | 0.5 | 0.0 | 0.871 | −1.0 | 1.1 |
| Tertiary educated | Ref | Ref | Ref | Ref | Ref | Ref |
| Having ischaemic heart disease | | | | | | |
| Yes | 1.5 | 0.7 | 0.2 | 0.039 | 0.1 | 2.9 |
| No | Ref | Ref | Ref | Ref | Ref | Ref |

**Table 6** Factors associated with depression among patients with diabetes by multiple linear regression (n=169)

| | | | | | 95% CI | |
|---|---|---|---|---|---|---|
| **Predictors** | **B** | **SE** | **β** | **p Value** | **Lower** | **Upper** |
| Age (years) | | | | | | |
| <50 | Ref | Ref | Ref | Ref | Ref | Ref |
| ≥50 | 1.4 | 0.6 | 0.2 | 0.027 | 0.2 | 2.7 |
| Ethnicity | | | | | | |
| Malay | 0.4 | 0.5 | 0.1 | 0.458 | −0.7 | 1.4 |
| Indian | 2.7 | 0.7 | 0.3 | <0.001 | 1.4 | 4.0 |
| Chinese | Ref | Ref | Ref | Ref | Ref | Ref |
| Highest education level | | | | | | |
| High school | −0.3 | 0.5 | −0.1 | 0.548 | −1.4 | 0.7 |
| Tertiary educated | Ref | Ref | Ref | Ref | Ref | Ref |
| Monthly household income (MYR) | | | | | | |
| <3000 | 1.9 | 0.6 | 0.3 | 0.001 | 0.8 | 3.0 |
| ≥3000 | Ref | Ref | Ref | Ref | Ref | Ref |
| Current employment status | | | | | | |
| Employed | Ref | Ref | Ref | Ref | Ref | Ref |
| Unemployed | −1.7 | 1.5 | −0.2 | 0.265 | −4.7 | 1.3 |
| Comorbidities | | | | | | |
| Diabetes alone | Ref | Ref | Ref | Ref | Ref | Ref |
| Diabetes with hypertension or dyslipidaemia | −2.6 | 1.5 | −0.4 | 0.080 | −5.5 | 0.3 |
| Diabetes with hypertension and dyslipidaemia | −2.3 | 1.7 | −0.2 | 0.189 | −5.7 | 1.1 |

significantly higher anxiety and depression level among Indian patients in comparison to other ethnicities. A recent Malaysian study which reported similar associations postulated that minority ethnic Indians experienced extensive psychological comorbidities due to triadic stressors of socioeconomic constraints, poor education level and perceived discrimination.[8]

Higher depression states in unemployment is caused by reduced sociological functions such as status identity, social contacts, participation in collective purposes and regular activities.[12] This study found a significantly higher depression status among unemployed patients in comparison to those being employed. Kaur *et al*[8] reported similar consistencies. In addition, this study found a significantly higher depression level in lower income patients. Similar findings were reported in a Malaysian study.[8] Reduced confidence due to economic instability and increased healthcare expenditures for routine diabetes screening complications, comorbid conditions and adherence to treatment pose substantial psychological illness among people with diabetes.[9]

Diabetes comorbid conditions like hypertension and dyslipidaemia has been known to amplify disease complications and poor treatment outcomes.[21 28] Increased rates of depression have been found in diabetes patients with hypertension.[28] An exponential rise of mortality rates due to serious cardiovascular disease complications in dyslipidaemia would contribute to high depression rates among patients with diabetes due to reduced quality of life and poor prognosis.[7 21] This study found a significantly higher depression score among patients with diabetes and hypertension or dyslipidaemia. Khuwaja *et al*[2] found similar findings.

Higher education attainment has been linked to be a protective factor against anxiety and depression among people with diabetes.[6 11 29] Education drives individuals towards proper understanding of disease mechanisms and complications, prompting increased compliance and adherence towards disease treatment for better health outcomes. Our study found a significantly lower anxiety and depression level among tertiary educated patients in comparison to high school graduates.

## LIMITATIONS

The absence of a control group and a small sample size from a single hospital might limit the generalisability of the study findings. In addition, the heterogeneity of the sample in this study caused by the wide range of age affects the prevalence rates and may limit the exploration of anxiety and depression in the youngest age groups. The cross-sectional design of the study limits our ability to make causal inferences. Further research is needed to address these limitations.

## CONCLUSION

Sociodemographics and clinical factors were important correlates of anxiety and depression among patients with diabetes. Age, ethnicity and IHD were significantly associated with anxiety. Factors significantly associated with depression were age, ethnicity and monthly household income.

## RECOMMENDATIONS

Early recognition of vulnerable factors associated with anxiety and depression in people with diabetes is necessary to promote patient adherence and compliance to diabetes control. Collaborative teamwork between healthcare providers and patients through a compassionate and holistic approach in recognising early neurotic features is essential to prevent disease comorbidities and mortalities. Rejuvenating primary healthcare policies from an essentially 'reactive-based system' (responding only when individuals are sick) to a 'proactive-based system' (focus on overall mental and physical health well-being) needs to be drafted immediately and amalgamated in all public health facilities within Malaysia. Increasing patient awareness to boost self-determination and confidence through integrated psychological and medical care in the management of diabetes would catalyse optimal health outcomes, as mused Osler (1961):

> Care more for the individual patient than for the special features of the disease…Put yourself in his place…The kindly word, the cheerful greeting, the sympathetic look—these the patient understands. Sir William Osler (Aphorisms from his bedside teachings and writings, 1961)

**Author affiliations**
[1]International Medical School, Management and Science University (MSU), Shah Alam, Selangor, Malaysia
[2]Clinical Research Center, Tengku Ampuan Rahimah Hospital (HTAR), Klang, Selangor, Malaysia
[3]Community Health Department, Faculty of Medicine, Universiti Kebangsaan Malaysia (UKM), Kuala Lumpur, Malaysia
[4]Department of Community Medicine, International Medical University (IMU), Kuala Lumpur, Malaysia

**Contributors** KG had the original idea, designed the study, drafted the first version of the manuscript and is responsible for the final version. PR and RAM assisted with the study design, data collection, the literature research and editing of the manuscript. KG and SARA-D contributed to the study coordination, data analysis and preparation of the manuscript. SARA-D revised the final draft critically for important intellectual content. All authors have contributed to and approved the final version of the manuscript.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None.

**Patient consent** Obtained.

**Ethics approval** Study protocol was approved by the Medical Research Ethics Committee (MREC), Ministry of Health Malaysia (NMRR-13-643-14711).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

## REFERENCES

1. Fowler MJ. Microvascular and macrovascular complications of diabetes. *Clin Diabetes* 2008;26:77–82.
2. Khuwaja AK, Lalani S, Dhanani R, *et al*. Anxiety and depression among outpatients with type 2 diabetes: a multi-centre study of prevalence and associated factors. *Diabetol Metab Syndr* 2010;2:72.
3. International Diabetes Federation. *Diabetes Atlas.*Vol 4. 2010. http://www.worlddiabetesfoundation.org/composite-35.htm (accessed 23 Dec 2013).
4. Statistics Malaysia and health facts. *Ministry of Health Malaysia*. 2014. http://www.moh.gov.my (accessed 31 Mar 2014).
5. Gonzalez JS, Sabrina A, Havah E, *et al*. Psychological issues in adults with type 2 diabetes. In: Pagoto S, eds. *Psychological co-morbidities of physical illness: a behavioral medicine perspective, Chapter II*, Springer Science Business Media LLC, 2011:73–121. doi:10.1007/978-1-4419-0029-6_2
6. Collins MM, Corcoran P, Perry IJ. Anxiety and depression symptoms in patients with diabetes. *Diabet Med* 2009;26:153–61.
7. Engum A, Mykletun A, Midthjell K, *et al*. Depression and diabetes—a large population-based study of sociodemographic, lifestyle, and clinical factors associated with depression in type 1 and type 2 diabetes. *Diabetes Care* 2005;28:1904–9.
8. Kaur G, Tee GH, Ariaratnam S, *et al*. Depression, anxiety and stress symptoms among diabetics in Malaysia: a cross sectional study in an urban primary care setting. *BMC Fam Pract* 2013;14:69.
9. Tovilla-Zarate C, Juarez-Rojop I, Jimenez YP, *et al*. Prevalence of anxiety and depression among outpatients with type 2 diabetes in the Mexican population. *PLoS ONE* 2012;7:e36887.
10. Everson SA, Maty SC, Lynch JW, *et al*. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *J Psychosom Res* 2002;53:891–5.
11. Peyrot M, Rubin R. Levels and risks of depression and anxiety symptomatology among diabetic adults. *Diabetes Care* 1997;20:585–90.
12. Palizgir M, Bakhtiari M, Esteghamati A. Association of depression and anxiety with diabetes mellitus type 2 concerning some sociological factors. *Iran Red Crescent Med J* 2013;15:644–8.
13. National Renal Registry Malaysia. *Clinical Research Center Ministry of Health Malaysia*. 2006. http://www.crc.gov.my (accessed 15 Nov 2013).
14. Katon W, Unutzer J, Russo J. Major depression: the importance of clinical characteristics and treatment response to prognosis. *Depress Anxiety* 2010;27:19–26.
15. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)*. Vol 4. American Psychiatric Association, 2000.
16. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–70.
17. Fatt QK, Atiya AS, Heng NGC, *et al*. Validation of the Hospital Anxiety and Depression Scale and the psychological disorder among premature ejaculation subjects. *Int J Impot Res* 2007;19:321–5.
18. Whelan-Goodinson R, Ponsford J. Validity of the Hospital Anxiety and Depression Scale to assess depression and anxiety following traumatic brain injury as compared with the Structured Clinical Interview for DSM-IV. *J Affect Disord* 2009;114:94–102.
19. American Diabetes Association. Standards of medical care in diabetes. *Diabetes Care* 2006;29:4–42.
20. Vaz NC, Ferreira AM, Kulkarni MS, *et al*. Prevalence of diabetic complications in rural Goa, India. *Indian J Community Med* 2011;36:283–6.
21. Mooradian AD. Dyslipidemia in type 2 diabetes mellitus. *Nat Clin Pract Endocrinol Metab* 2009;5:150–9.
22. Morley JE. The elderly type 2 diabetic patient: special considerations. *Diabet Med* 1998;15:41–6.
23. Almawi W, Tamim H, Al-Sayed N, *et al*. Association of comorbid depression, anxiety, and stress disorders with type 2 diabetes in Bahrain, a country with a very high prevalence of type 2 diabetes. *J Endocrinol Invest* 2008;31:1020–4.
24. Ganatra HA, Zafar SN, Qidwai W, *et al*. Prevalence and predictors of depression among an elderly population of Pakistan. *Aging Ment Health* 2008;12:349–56.
25. Mosaku K, Kolawole B, Mume C, *et al*. Depression, anxiety and quality of life among diabetic patients: a comparative study. *J Natl Med Assoc* 2008;100:73–8.
26. Dunlop DD, Song J, Lyons JS, *et al*. Racial or ethnic differences in rates of depression among preretirement adults. *Am J Public Health* 2003;93:1945–52.
27. Fisher L, Laurencin G, Chesla CA, *et al*. Depressive affect among four ethnic groups of male patients with type 2 diabetes. *Diabetes Spectr* 2004;17:215–19.
28. Thomas J, Jones G, Scarinci I, *et al*. A descriptive and comparative study of the prevalence of depressive and anxiety disorders in low-income adults with type 2 diabetes and other chronic illnesses. *Diabetes Care* 2003;26:2311–17.
29. Bener A, Al-Hamaq AO, Dafeeah EE. High prevalence of depression, anxiety and stress symptoms among diabetes mellitus patients. *Open Psychiatry J* 2011;5:5–12.