Appendix A1

Excel and SPSS

by

Gert Nieuwenhuis
Tilburg University
The Netherlands

August 2009

This reader treats some statistical techniques concerning the computer packages Excel and SPSS. It follows the statistical approach in the book **Statistical Methods for Business and Economics**, 2009, by Gert Nieuwenhuis. The notes are based on the versions Excel 2007 and SPSS 16.0. Below, Section A1.j belongs to Chapter *j* of the book.

The choice has been made to mainly use Excel for the first chapters of the book (roughly, Chapters 1 – 15), although occasionally SPSS is used too. The reason is that Excel is rather accessible and many students have used it during their secondary school education. Furthermore, in the opinion of the author Excel is preferred since for many Excel-commands statistical knowledge is needed. (Many SPSS-commands work like a black box and statistical knowledge is hardly necessary to conduct them.) In Part I of this reader many Excel techniques are explained, at the moment they are needed when reading the text book.

But for later sections (roughly, Sections A1.16 – A1.25), when techniques for Inferential Statistics are demonstrated, SPSS is mostly used since then this package is better equipped. Still, Excel is also frequently used for the more arithmetic or probabilistic results. Part II of these notes gives a first, guided introduction to SPSS, It is situated between the Sections A1.15 and A1.16, but it is completely self-supporting and might also be located otherwise. Part III is mainly about SPSS-options concerning Chapters 16 – 25 of the book (although occasionally Excel is used too). The approach is rather ad hoc, but the author believes (and has the experience) that – after the first guided SPSS introduction of Part II – it works well.

The author wants to thank Prof. Dr. Arthur van Soest for permitting me to use the Statistica datasets and the basic version of Part II.

Part I: Sections A1.1 – A1.15

A1.1 Introduction and basic concepts

Some elementary **Excel** procedures are considered; they simplify operations on data.

***Sum* Σ**
This button can be found under `Home`; the underlying function calculates the sum of data and puts the result in a previously selected cell:

- Select the cell;
- click **Σ**;
- select the data you want to add up;
- press `Enter`.

Easy Excel-functions
You may want to have some easy functions available immediately. These functions can be chosen by checking them after clicking, with the right-hand mouse button, on the `Status Bar` (the lowest bar of the Excel screen). Some (frequently chosen) easy functions that can be checked are:

`Average`	(to determine the mean of data)
`Count`	(to count the number of data points, including the non-numerical ones)
`Numerical Count`	(to count the number of numeric data points)
`Max`	(to determine the maximum of a dataset)
`Min`	(to determine the minimum of a dataset)
`Sum`	(to calculate the summation of a dataset)

Having checked these functions, their values will always appear on the `Status Bar` as soon as you select data.

The Fill Handle, to fill data into adjacent cells
This is the small black square in the lower-right corner of selected cells. It will be illustrated with an example.

Suppose you have two data columns, each with 25 data points; or to say it otherwise: you have 25 pairs of data, where a pair is the combination of two data points of a row in the dataset. You want to create a new column that contains the 25 differences of the first and the second data point in the pairs. Then you do not have to conduct the subtract-operation 25 times. Excel offers the possibility to do it in a cell for the first pair and to use the lower-right corner of that cell to do the same for the other pairs of data.

To be more precise, suppose that the positions A1-A25 and B1-B25 contain the data and that you want to create the differences of the two data points of the pairs in the positions C1-C25. Then do the following steps:

- Type = in cell C1;
- click A1, type - (that is, "minus") in C1, and click B1;
- press `Enter` (the difference of the first pair appears in C1);
- select C1 and place the cursor at the lower-right corner of cell C1;
- press the left mouse-button and keep on doing so while drawing the corner to position C25 (the 25 differences of the pairs appear in the cells C1-C25).

Exercise A1.1
Column A contains the data 1, 2, 3, 4; column B contains the data 5, 6, 7, 8. Create in column C the data of the operation 2A – 3B by using the fill handle.

The use of the fill handle to transform data into percentage growth data
Suppose that the GDPs for the most recent ten years are in A1-A10; call them $x_1, \cdots, x_{10}$ where the subscript 1 refers to the first year in order (so, to ten years ago) and 10 refers to last year. However, you want the **percentage growth** data of the GDPs; that is: the data that express the percentage change of the years when compared to the year before. So, you want:

$$y_t = 100 \times (x_t - x_{t-1}) / x_{t-1} \quad \text{for } t = \mathbf{2}, ..., 10$$

Below, this formula is used to create these growth (%) data; it is illustrated for $t = 2$.

- Type =100*(in cell B**2**;
- click A2, type - ("minus") in B2, click A1, type)/ in B2, click A1;
- press `Enter` (the percentage growth of year 2 when compared to year 1 appears in B2);
- select B2 and place the cursor at the lower-right corner of cell B2;
- press the left mouse-button and keep on doing so while drawing the corner to cell B10 (the **nine** growth data appear in the cells B2-B10).

A1.2 Tables and Graphs

Some **Excel techniques** will be considered that can be used to create tables and/or graphs. Important statistical concepts of Chapter 2 of the book are: (cumulative, relative) frequency distribution, distribution function, bar chart, pie chart and histogram.

But first: Is Data Analysis installed?
First check whether `Data Analysis` is installed (under `Data`). If not, do the following:

- Click the **Microsoft Office Button** in the upper-left corner of the Excel screen, click **Excel Options**, and then click **Add-Ins**.
- In the **Add-ins** box, identify the add-in that you want to enable (in this case: the **Analysis Toolpak**) and note the Add-in type located in the **Type** column.
- Select the Add-in type in the **Manage** box and then click **Go**.

Creation of a frequency distribution from the original data

Start from the original data. Create a new column with heading `Bin`. In case of a discrete variable, type the **different** values of the dataset under `Bin`; in case of a continuous variable, type under `Bin` the *endpoints* of the classes of the classification that you want.

- Choose `Data` / `Data Analysis` / `Histogram`;
- import the original data into `Input Range`;
- import the Bin-values into `Bin Range`;
- choose a first position for your `Output Range`;
- if you also want a graph, check `Chart Output`;
- `OK`.

As a result, the frequency distribution appears (with a graph if `Chart Output` was checked). In case of data of a continuous variable, the resulting frequency for an endpoint is the frequency of the interval that is formed by that endpoint and the previous endpoint.

Excel does not have a direct command to obtain the ***relative frequency distribution***; this more specific distribution can be determined from the frequency distribution by using the fill handle (see Section A1.1).

By checking `Cumulative Percentage` in the creation procedure above, the ***cumulative relative frequency distribution*** (in percentages) can be obtained. However, the accompanying chart is not very useful.

Exercise A2.1

Dataset (of a discrete variable): 6, 6, 8, 6, 10, 10, 8, 9, 9, 6.
Determine the (cumulative, relative) frequency distribution of this dataset.
[Remarks: although 7 is not in the dataset, it is wise to include it under `Bin`. Why?]

Creation of bar chart, pie chart, histogram

In the book, a bar chart presents the (relative) frequency distribution of data of a discrete or coded-qualitative variable, and it has its bars separated. A histogram is for classified data of a continuous variable; its bars are adjacent. However, the Excel-options `Histogram` and `Bar Chart` are less strictly defined. Excel has the options `Charts` (under `Insert`) and `Histogram` (under `Data` / `Data Analysis`) to create graphical presentations of frequency distributions.

The Excel-option `Histogram` can be used to create a histogram or bar chart directly **from the original data** (see above for its construction), but only if these data are numerical (quantitative or qualitative but coded as numbers).

Bar charts and pie charts can be created from `Charts`, but only **if the data are already summarized** in terms of a (relative) frequency distribution.

So, to create a histogram or bar chart from original, quantitative data, you need the Excel-option `Histogram`. If the data are already summarised in a frequency distribution, the Excel option `Charts` has to be used.

a) Creation of bar chart or histogram from original, numerical (quantitative) data

- Create the frequency distribution while checking `Chart Output`, as described above under *Creation of a frequency distribution from the original data*;
- for a bar chart (so: with separated bars, for data of a discrete or coded-qualitative variable), the chart output can be used (and refined if wanted);
- for a histogram (so, with adjacent bars, for data of a continuous variable), right-click one of the bars, choose `Format Data Series` and put `Gap Width` on `No Gap` (0 %); click `OK`.

b) Creation of bar chart or pie chart from un-coded, qualitative data

- Select the data;
- use `Data` / `Sort` to order the data alphabetically;
- create the frequency distribution (that is, the different non-numeric values of the dataset combined with their frequencies) by counting the frequencies of the different values; hence, the different values are listed jointly with the corresponding frequencies;
- select the overview of the frequency distribution (so, values and frequencies), click the wanted bar chart type in `Insert`, `Charts`; for instance: the first 2D-one of `Column` (or `Pie`);
- if you want to change (in the bar chart) the name of the series: right click a bar, choose `Select Data` and `Edit` (left-hand), and type the `Series name` you want.

c) Creation of bar chart, histogram or pie chart for a given (classified, relative) frequency distribution

Starting point is a given (classified, relative) frequency distribution. That is: for a continuous variable, we already have a column with endpoints of the classes and a column with the accompanying (relative) frequencies; for a discrete variable, we already have a column with the different values and a column with the accompanying (relative) frequencies. See above for the creation of such a distribution from the original data.

- Select the (relative) frequencies of the distribution;
- click the wanted bar chart type in `Insert`, `Charts`; for instance: the first 2D-one of `Column` (or `Pie`); the chart follows;

- to change the numbers 1, 2, 3, ... on the horizontal axis into the set of different values (or the endpoints of the classes), right-click a bar and choose `Select Data`;
- click `Edit` under `Horizontal Axis Labels` and insert the different values or endpoints into `Axis label range`; `OK`;
- if a histogram (for a continuous variable) is wanted, right-click one of the bars, choose `Format Data Series` and put `Gap Width` on `No Gap` (0 %); click `OK`.

Exercise A2.2

a. For the data of the former exercise, construct a bar chart of the relative frequency distribution.
b. Consider the following classified relative frequency distribution of a continuous variable:

Class	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	**Total**
Rel frequency	0.1	0.2	0.1	0.4	0.2	1

Create a nice histogram with Excel.

Scatter plots

See also Section A1.5. These graphical plots can be used to show the relationship between two sets of numbers; one set is put horizontally, the other is put vertically. This scatter plot option is also used frequently as an Excel-detour to overcome the omission of direct creations of graphs. For instance, it can be used to create a graph of a distribution function.

Creation of a graph of the distribution function for a classification of a continuous dataset

- Firstly, create the **cumulative** relative frequency distribution (that is: a column of endpoints of the classes accompanied by a column of the corresponding cumulative relative frequencies). Choose a suitable starting point *s* of the first class, in accordance with the common width of the classification. Extend the cumulative relative frequency distribution by letting it start at (*s*, 0).
- Select the data of the cumulative relative frequency distribution (endpoints + cumulative relative frequencies);
- under `Insert, Charts` click `Scatter` and choose the type: `Scatter with Straight Lines and Markers`;
- if wanted, the graph can be refined.

Recall that the distribution function of a dataset of a **discrete** variable is a step-function. This cdf can also be created with the option `Scatter`, but it takes more experience. Its construction will not be considered here.

The scatter plot option is also used to give a graphical presentation of **time series**.

Exercise A2.3

a. Create a graph of the distribution function of the distribution in part **b.** of the previous exercise.
b. The table below contains the GDPs of Belgium for the period 2000 – 2005:

Year	2000	2001	2002	2003	2004	2005
GDP (billions of euro)	251.7	258.9	267.7	274.7	289.5	298.5

Source: Statistics Belgium (2006)

Create a nice graph to present these data.

A1.3 Measures of location

Some **Excel techniques** will be considered. The most common measures of location are under the button `Function` f_x (under `Formulas`)

Under this button, Excel houses several functions. The category `All` lists them alphabetically. The Excel-functions `Median` and `Average` can be used to calculate the median and the mean of a dataset. The Excel-function `Mode` is used to calculate the mode of a dataset, but it is unreliable: the answer depends on the ordering of the data; see the columns in Example A3.2 below.

The use of `Function` f_x

Things are illustrated for the function `Average`; `Median` and other functions (for instance `ABS`, absolute value) go similarly.

- Put the cursor in the cell where you want your answer to appear;
- click `Insert Function` (under `Formulas`);
- choose the category and the function you want (for instance `Average`); press `OK`;
- import the data into `Number 1`;
- press `OK` (the answer appears in the cell that was chosen before).

Exercise A3.1

Determine the mode, median and mean of the dataset: 1, 2, 3, 4, 5, 6.

Exercise A3.2

Determine the mode of each of the columns in the table below. Guess what the mode will be before calculating it.

1	1	1	2	2	2
1	2	2	1	1	2
2	1	2	1	2	1
2	2	1	2	1	1

The Excel-function Sumproduct

This function is easy when a **weighted mean** has to be calculated since it generates the sum $\sum w_i x_i$ for the weights *w* and the observations *x*.

- Put the cursor in the cell where you want your answer to appear;
- click `Insert Function`;
- choose the category `All` and the function `Sumproduct`; press `OK`;
- import the *w*-data into `Array 1`;
- import the *x*-data into `Array 2`;
- press `OK` (the answer appears in the cell that was chosen before).

The use of Excel-functions directly in the data sheet

Things are illustrated with `Average` and `Sumproduct`.

For `Average`:

- Put the cursor in the cell where you want your answer to appear;
- type =average(in that cell, select the data, and type) in that cell;
- press `Enter` (the answer appears in the cell that was chosen).

For `Sumproduct`:

- Put the cursor in the cell where you want your answer to appear;
- Type =sumproduct(in that cell, select the *w*-data, type , in that cell, select the *x*-data, and type) in that cell;
- press `Enter` (the answer appears in the cell that was chosen).

The Excel-function Geomean

With this function the **geometric mean** r_g of a sequence $r_1,\cdots r_n$ of growth variables (proportions, numbers usually between 0 and 1) can be calculated. However, it wants the $1+r_t$ as input and returns $1+r_g$.

- First, create a column that contains the $1+r_t$;
- put the cursor in the cell where you want your answer to appear;
- click `Insert Function`;
- double-click the function `Geomean`, for instance in the category `Statistical`;
- import the $1+r_t$ into `Number 1`;
- press `OK`;

- $1 + r_g$ appears in the cell that was chosen; subtract 1.

[[Or: type =geomean(in a cell, select the data $1 + r_t$, type) in that cell, and press `Enter`. Don't forget to subtract 1.]]

Exercise A3.3
The following proportions are the interest rates for three successive years: 0.2, 0.1, -0.1. Show that the geometric mean equals 0.0591, which corresponds to 5.91%.

A1.4 Measures of variation

Some **Excel and SPSS techniques** will be considered. In Chapter 4 of the book, several new statistics were introduced: quartile, percentile, variance, standard deviation. When the underlying dataset is given, they can be calculated with Excel and SPSS. Note, however, that Excel and SPSS may give (slightly) different answers for quartiles and percentiles.

Excel-functions
The required Excel functions can be found under `Insert Function` and `Statistical` (or `All`).

Quartile:	Choose `quartile`; import the data into `Array` and type 1 or 2 or 3 in `Quart`.
Percentile:	Choose `percentile`; import the data into `Array` and type the wanted **proportion** (number between 0 and 1) in `K`.
Variance:	Use `var` for a sample dataset and `varp` for a population dataset.
Standard deviation:	Use `stdev` for a sample dataset and `stdevp` for a population dataset.

Excel-calculation of statistics in one go
Mean, median, mode, variance and standard deviation can also – in one go – be obtained by taking the following actions:

`Data / Data Analysis / Descriptive Statistics`

Import the data into `Input Range`, choose your `Output Range`, and check `Summary Statistics`; click `OK`.

Notice, however, that the obtained values for variance and standard deviation are **sample** results. That is: when following this road to calculate variance and standard deviation, Excel automatically assumes that the dataset is a sample dataset. When using this method to obtain the corresponding **population** statistics, multiply the sample variance and sample standard deviation respectively by $(N-1)/N$ and $\sqrt{(N-1)/N}$.

SPSS-options

Mean and standard deviation can be obtained in one go:

`Analyze / Descriptive Statistics / Descriptives`

Place the name of the variable(s) of interest under `Variable(s)` and click `OK`.

Notice that the **sample** standard deviation is calculated. If a **population** standard deviation is wanted, multiply the result by $\sqrt{(N-1)/N}$.

Mean, median, mode, standard deviation, variance, quartiles, percentiles can be obtained in one go:

`Analyze / Descriptive Statistics / Frequencies`

Under `Statistics`, check the functions you want; for `Percentiles`, fill in the percentages that are wanted and use the `Add`-button. Click `Continue` and `OK`.

SPSS box plot options

A. Using `Graphs / Box plot`:

1) Box plot(s) for groups of data in **different** columns (as in Example 4.3 of the book)
   - Choose `Simple`, check `Summaries of separate variables` and click `Define`;
   - import the variable(s) for which you want box plots (in one figure) into `Boxes Represent`;
   - if wanted, import the variable (if present) that contains the numbers or names (labels) of the objects (cases) into `Label Cases by`;
   - click `OK`.

2) Box plots for groups of data that are all in **one** column (as in Example 4.4)

   - Choose `Simple`, check `Summaries for groups of cases` and click `Define`;
   - import the data-column into `Variable`;
   - import the column with the group-indication into `Category Axis`;
   - if wanted, import the variable (if present) that contains the numbers or names (labels) of the objects (cases) into `Label Cases by`;
   - click `OK`.

B. Using `Analyze / Descriptive Statistics / Explore`, if also Tukey's hinges and outliers are wanted:

- Import the data-columns into `Dependent List`;

- check Both;
- under Statistics, check Percentiles and Outliers;
- click Continue and OK.

Notice that a part of the printout is about the 25%, 50% and 75% percentiles, and that Tukey's quartiles (which correspond to the locations of the hinges) and alternative ones (that often are different) are given. The box plots are also created, in different pictures.

SPSS interactive histogram-option (used for the bar charts of Example 4.4)

- Graphs / Interactive / Histogram;
- drag the data-column into the box in the middle;
- drag the column with the group-indication into the box Panel Variables;
- click OK.

A1.5 Pairs of variables

Some **Excel and SPSS techniques** will be considered. Important concepts of Chapter 5 are **scatter plot** (with **regression line**), **covariance**, **correlation**, **regression** and **contingency table**. It will be described how Excel and SPSS can be used; only the basic steps are mentioned. The independent variable (or its data column) is denoted by *x*, the dependent variable (or its data column) by *y*.

Scatter plot with regression line in Excel

- Select the *x*- and *y*-data;
- click Scatter under Insert;
- choose the type you want.

To add the regression line:

- Right-click on one of the dots;
- choose Add Trendline;
- choose Linear and check Display Equation on chart (if wanted);

Scatter plot with regression line in SPSS

- Graphs / Scatter;
- choose Simple and click Define;
- put *y* in Y Axis and *x* in X Axis;
- OK.

To add the regression line:

- Double-click the plot;
- right-click one of the dots;
- choose Add Fit Line at Total;
- choose Linear.

To add the data labels (as in the scatter plot of Example 5.6):

- Double-click the plot;
- right-click one of the dots;
- choose `Show Data Labels`.

Covariance and correlation in Excel

- Both functions `covar` and `correl` can be found under the `Function` f_x: in the list in the category `Statistical`;
- double-click the function; import the *x*- and *y*-data into `Array 1` and `Array 2`, respectively;
- notice that `covar` considers all data to be **population** data. If your dataset is a sample dataset, multiply the `covar`-result by n and divide by $n-1$.

Covariance and correlation in SPSS

- `Analyze` / `Correlate` / `Bivariate`;
- put *x* and *y* in the `Variables` window;
- check `Pearson`;
- under `Options`, check `Cross-product deviations and covariances`;
- `Continue` and `OK`.

Regression-printout in Excel

- `Data` / `Data Analysis` / `Regression`; `OK`;
- put *y* in `Input Y Range` and *x* in `Input X Range`;
- check `Output Range` and choose your position;
- (check `Residuals` if wanted);
- `OK`.

Regression-printout in SPSS

- `Analyze` / `Regression` / `Linear`;
- put *y* in `Dependent` and *x* in `Independent(s)`;
- (if wanted, under `Save` and `Residuals`, check `Unstandardized` and click `Continue`);
- `OK`.

Contingency tables with Excel

The dataset always has to have three columns; one of them must identify the respective elements (as 'year' in Example 5.13 and 'ID' in Example 5.14). Below, this variable will be called ID.

- From `Insert`, click `Pivot Table`;
- insert all data (including the headings) into `Table/Range`;
- choose your `Location`; `OK`;

- from the Pivot Table Field List, drag one of the two non-ID variables into Column Labels and the other into Row Labels; drag ID into Σ Values;
- click Sum of ID and Value Field Settings; choose Count and click OK.

To obtain charts for the frequency distributions of variable I by the other variable (II):

- In the resulting table above, select all cells except the totals of the rows and the columns;
- In Insert, Charts, click the first sub-type of Column;
- Take care that the values of variable II are on the horizontal axis; otherwise, swap the two variables by dragging.

To change counts into percentages:

- Double-click Count of ID in the chart;
- Under Show values as, choose Percentage of row.

Contingency tables with SPSS

- Analyze / Descriptive Statistics / Crosstabs;
- Put variable I in Row(s) and variable II in Column(s);
- Check Display clustered bar charts;
- Click Cells and check Counts (observed) and/or Percentages (row);
- Continue; OK.

A1.6 Definitions of probability

The use of the Excel-function Sampling in Example 6.7

The Excel function Sampling chooses arbitrarily from some input range. That is why the values of the twenty coins have to be typed in the first column and are taken as input range.

Type the values of the twenty coins in positions A1 – A20 of the Excel data sheet. That is: type three times 2, five times 5, nine times 10 and three times 20. Choose the option Sampling under Data / Data Analysis and press OK. Select A1 – A20 as Input Range, check Random and take 20000 as Number of Samples, take B1 as Output Range and press OK. The 20000 simulated drawings from the wallet appear in column B. Of course, your data in column B will differ from the data found by the author, in Xmp06-07.xls.

Type BIN in C1, and the numbers 2, 5, 10 and 20 in C2 – C5. To obtain $n(A)$, $n(B)$, $n(C)$ and $n(D)$ for the case that $n = 100$, choose Data / Data Analysis / Histogram / OK. Select the data in B1 – B100 as Input Range, take C2 – C5 as Bin-Range, and choose your Output Range (for instance, E1) and press OK. (Do not check Chart Output; only the frequency distribution is wanted.) The frequencies

$n(A)$, $n(B)$, $n(C)$ and $n(D)$ can be read from the resulting frequency distribution; the relative frequencies for the table follow by dividing the frequencies by 100. Of course, your answers will differ from the numbers in the table of Example 6.7.

To find $n(A)$, $n(B)$, $n(C)$ and $n(D)$ for the cases that n = 200, 500 etc., just repeat the above procedure while 100 is replaced by 200, 500 etc.

A1.7 Calculation of probabilities

In Excel, the function `randbetween` – with `Bottom` 1 and `Top` N – randomly generates one of the numbers of $\{1, 2, \cdot\cdot\cdot, N\}$; the option `sampling` generates arbitrarily chosen numbers from some list of numbers. The following example is a simulation application.

The use of the Excel-function Sampling in the Table and Figure of Example 7.6

- Type 1, 2, · · ·, 25 in the positions A1 – A25.
- Choose `Data/Data Analysis/Sampling`; take A1 – A25 as `Input Range`, 25000 as `Number of Samples` and B1 as `Output Range`; `OK`.
- Use `Data/Data Analysis/Histogram` to create the frequency distribution of the simulated data in column B; take B1 – B25000 as `Input Range`, A1 – A25 as `Bin Range`, D1 as `Output Range` and do **not** check `Chart Output`.
- Divide the resulting frequencies by 25000 to obtain the relative frequencies; of course, your relative frequencies will differ from those in the table of Example 7.6.
- Use `Charts` (under `Insert`) to create the bar chart of your relative frequency distribution.

The use of the Excel-function `Randbetween` in Exercise 7.38ab

Use the Excel-function `randbetween(1,6)` to simulate one realisation of X = 'number of eyes with a fair die' in position A1 of the Excel data sheet. Draw A1 (with the fill handle) to position A10000, thus generating a sample of 10000 simulated realisations of X. Fix these data by using `copy`, `paste special`, `values`.

It also is asked to determine the relative frequency of the outcome k (for k = 1, 2, · · ·, 6). Again, the option `Histogram` can be used. But also the following approach is possible. For example for k = 2, the relative frequency can be obtained as follows:

- Type `=IF(A1=2,1,0)` in C1;
- draw this field to C10000;
- count the number of ones in column C by using `sum;`
- divide the result by 10000

A1.8 Probability distribution, expectation and variance

The use of the Excel-function `Randbetween` in Exercise 8.28bcd

Use the Excel-function `randbetween(1,6)` to simulate one realisation of X = 'number of eyes with a fair die' in position A1 of the Excel data sheet. Draw A1 (with the fill handle) to position A5000, thus generating a sample of 5000 simulated realisations of X. Fix these data by using `copy`, `paste special`, `values`.

It also is asked to determine the relative frequencies of the outcome k (for $k = 1, 2, \cdots, 6$). For example for $k = 2$, the relative frequency can be obtained as follows:

- Type `=IF(A1=2,1,0)` in C1;
- draw this field to C5000;
- count the number of ones in column C by using `sum;`
- divide the result by 5000

A1.9 Families of discrete distributions

In Excel, probabilities for binomial and hypergeometric distributed random variables can respectively be calculated with `binomdist` and `hypgeomdist`

Calculation in Excel of probabilities about Y if Y ~ Bin(n, k)

- $P(Y = k)$ can be calculated with `binomdist(`*k,n,p,* **`0`**`)`
- $P(Y \le k)$ can be calculated with `binomdist(`*k,n,p,* **`1`**`)`

For instance:

if $Y \sim Bin(9,0.4)$, then $P(Y = 3)$ follows with `binomdist(3,9,0.4,0)`

if $Y \sim Bin(9,0.4)$, then $P(Y \le 4)$ follows with `binomdist(4,9,0.4,1)`

Calculation in Excel of probabilities about Y if Y ~ H(n; M, N)

- $P(Y = k)$ can be calculated with `hypgeomdist(k,n,M,N)`

For instance:

if $Y \sim H(9;50,100)$, then $P(Y = 4)$ follows with `hypgeomdist(4,9,50,100)`

Calculation in Excel of probabilities about Y if Y ~ Po(μ)

- $P(Y = k)$ can be calculated with `poisson(k,μ,` **`0`**`)`
- $P(Y \le k)$ can be calculated with `poisson(k,μ,` **`1`**`)`

For instance:

if $Y \sim Po(8.7)$, then $P(Y = 7)$ follows with `poisson(7,8.7,0)`

if $Y \sim Po(15.1)$, then $P(Y \le 14)$ follows with `poisson(14,15.1,1)`

Creation of pdf and bar chart of Bin(100, 0.05)

- Type "x" in A1 and "$P(X = x)$" in B1 of an empty Excel sheet;
- type 0 in A2 and 1 in A3;
- select A2 and A3 simultaneously and use the fill handle (see Section A1.1) to draw the lower-right angle to position A102, thus creating the numbers 0, 1, $\cdots$, 100 in the field A2-A102;
- type `=binomdist(A2,100,0.05,0)` in position B2 and press enter;

- select B2 and use the fill handle to draw the lower-right angle of B2 to position B102. Thus, the desired binomial pdf appears.
- For the bar chart, see Section A1.2.

A1.10 Families of continuous distributions

In Excel, probabilities for normal distributions can be calculated with `normdist`. The inverse operation (when a probability is given and a quantile has to be calculated) can be done with `norminv`.

Calculation in Excel of probabilities about Y if Y ~ N(μ, σ²)

- $f(y)$, the value of the **pdf** f at y, follows with `normdist(y,μ,σ,0)`
- $P(Y \le y)$ can be calculated with `normdist(y,μ,σ,1)`
- For given $A = P(Y \le b)$, the constant b follows with `norminv(A,μ,σ)`

For instance for $Y \sim N(10.8, 4)$:

$f(9.2)$ follows with `normdist(9.2,10.8,2,0)`

$P(Y \le 7.3)$ follows with `normdist(7.3,10.8,2,1)`

The constant b with $P(Y \le b) = 0.72$, follows with `norminv(0.72,10.8,2)`

Calculation in Excel of probabilities about Z if Z ~ N(0, 1)

- $P(Z \le z)$ can be calculated with `normSdist(z)`
- For given $A = P(Z \le b)$, the constant b follows with `normSinv(A)`

For instance:

$P(Z \le 0.8)$ follows with `normSdist(0.8)`

The constant b with $P(Z \le b) = 0.9$ follows with `normSinv(0.9)`

Remarks:

- All Excel functions `norm(s)dist` and `norm(s)inv` look to the **left**
- `normSdist(z)` gives the same result as `normdist(z,0,1,1)`
- `normSinv(A)` gives the same result as `norminv(A,0,1)`
- since probabilities for $Y \sim N(\mu, \sigma^2)$ can be transformed into probabilities for $Z \sim N(0, 1)$, the Excel-function `normdist` is not indispensable.

Calculation of z_α

- since, under the standard normal pdf, z_α has an area $1 - \alpha$ at its **left** side, z_α follows with `normSinv(1-α)`

For instance:

$z_{0.05}$ follows with `normSinv(0.95)`

Creation of the graph of the N(10, 16)-pdf

We use `normdist(x,10,4,0)` in a clever way. Notice that the larger part of the total probability 1 lies between $10 - 3.5\times4 = -4$ and $10 + 3.5\times4 = 24$.

- Type –4 and –3.8 in A1 and A2, respectively.
- Select the field A1 – A2 and use the fill handle to draw the lower-right angle to position A141, thus creating the values –4, –3.8, –3.6, · · ·, 23.8, 24 in the field A1 – A141.
- Type `=normdist(A1,10,4,0)` in B1 and press Enter.
- Select B1 and use the fill handle to draw the lower-right angle to position B141, thus creating the corresponding values $f(x)$ in column B.
- Create the graph of *f* by constructing the scatter plot of the B-data on the A-data; see Section A1.5.

The use of the Excel-function `Rand` in Example 10.3

- `Rand()` is used to obtain simulated observations from $U(0, 1)$.
- `Sqrt` is used to find the simulations from *f*.
- `data/data analysis/histogram` is used to obtain the frequency distribution, taking 0.1, 0.2, · · ·, 1 as `Bin Range`.
- Next, the relative frequency distribution is created by dividing by 10000.
- All relative frequencies are divided by the common width 0.1 of the ten classes, to obtain the relative frequency **density**.
- This relative frequency density can be graphed (see Section A1.2).

A1.12 Random samples

Exercise 12.12 (with SPSS)

b. Determine estimates of μ_X, σ_X^2, σ_X, μ_Y, σ_Y^2 and *p*.
   - Open the file.
   - Choose `Analyze/Descriptive Statistics /Descriptives`.
   - Select the three variable names and transport them with the arrow to the right-hand window.
   - Choose `OK`.

d. Find estimates for $\sigma_{X,Y}$ and $\rho_{X,Y}$.
   - Choose `Analyze/Correlate/Bivariate`,
   - select the variables *X* and *Y* and transport them with the arrow to the right-hand window.
   - Under `Options`, check `cross-products and covariances`, and `continue` and `OK`.

A1.13 The sample mean

Exercise 13.17d (jointly with Excel and SPSS)

- In Excel, use `Rand()` and the fill handle to simulate 1000 rows and 12 columns with $U(0, 1)$ observations.

- Use the fill handle to create a column with 1000 observations of $S - 6$.
- Copy-paste this column into an empty SPSS data sheet.
- Use `Graphs/Histogram` to create a figure of a histogram of these 1000 observations jointly with the best fitting normal curve.

A1.15 Interval estimation and hypothesis testing: a general introduction

Example 15.10 (with SPSS)

- From the SPSS data sheet, choose `Analyze / Descriptive Statistics / Descriptives`.
- Place (with the arrow) the variable `Ass_Time` under the heading `Variable(s)`; press `OK`

Exercise 15.26 (with SPSS)

Use the function `MEAN` via `Transform / Compute Variable` to find the average-variable. Calculate its mean and standard deviation via `Descriptive statistics`. Use them to determine the standardised observations (of X), again via `Transform / Compute Variable`.

To find the means and stdevs per country, use `Analyze/Compare means/ Means`.

Part II: A first guided introduction to SPSS

*Part II of this reader offers a first introduction to SPSS. The so-called **Statistica** dataset is used to illustrate the principle ideas of SPSS. You are supposed to read this chapter behind your PC, and simultaneously make the (small) exercises.*

1. Introduction

SPSS is the market leader in the area of data analysis. The package was originally conceived as a statistical package for the social sciences, but nowadays it has a much wider application. There are versions of SPSS for many types of computers. The version that was used when writing the present notes, is called SPSS for Windows version 16.0. Like many other packages for a Windows environment, this version is meant to be very user-friendly; it is mainly menu-driven. This, incidentally, does not mean that every feature of the program is easily found and self-explanatory. However, after a first guided introduction, it is certainly possible to explore and use the possibilities of the package without technical help.

The standard format of the data that are input to SPSS is a rectangular data matrix. The rows are ***cases***, for instance individuals or households; the columns are variables, like age or education. We will demonstrate the basic principles of SPSS by using the Statistica dataset.

Statistica is an imaginary European county with more than 10000 adult inhabitants. Among the records of the counties public administration there is a database called `statistica`, which keeps track of data on the adult population of the county. Besides personal information the database also contains information on the family (= household) the adult belongs to. The data matrix `statistica` is about the following variables:

ds	dummy single (= 1 for an adult who runs a household on his/her own, the respondent. The respondent might be heading a household with one or more minors. All other adults are coded ds = 0)
dh	dummy head of household (= 1 if the adult is not the only adult in the household but is head of the household. By definition, this household consists of two or more adults. Observe that *ds*dh* = 0)
dp	dummy partner (= 1 if the adult is the partner of the head of the household. By definition, we have *dp*ds* = 0. "Partner" is defined as being married to the head of the household. Persons cohabiting are not considered partners)

By these definitions there are four exclusive situations an adult can have with respect to his or her position within the household:

(1) single (ds = 1)

(2) *head of a household with two or more adults (dh = 1)*
(3) *partner of the head of the household (dp = 1)*
(4) *none of the above possibilities (ds = dh = dp = 0)*

df dummy female (= 1 if the adult is female)
weight weight in kg
length height (= length) in cm
age age in years (not rounded off)
edu education level (the lowest level is coded 1, the highest is coded 5)
wage hourly wage (in euro) during the last year
hours number of hours worked per week
nkids number of children (<18 years)
fs family size (= number of family members; i.e., number of adults + number of children under 18)
finc annual net family income * 1000 euro (including children's allowance)
foodexp annual family expenditures on food * 1000 euro
housexp annual family expenditures on housing * 1000 euro
clotexp annual family expenditures on clothes * 1000 euro
recrexp annual family expenditures on recreation * 1000 euro
findex registration number of the family (i.e., the identification number of the family to which the adult belongs according to the registry)

In SPSS, a distinction is made between ***categorical*** and ***numerical*** variables. Categorical variables are also called ***qualitative***. They have no natural numerical value, but give a partition into categories or classes. Numerical variables correspond to what are called "***quantitative*** data". The categorical variables in the data set of Statistica have the following categories

ds 0 in a more-person household
1 in a one-person household

dh 0 not a head of a more-person household
1 head of a more-person household

dp 0 no partner of head of a more-person household
1 partner head of a more person household

df 0 male
1 female

edu 1 low
5 high

The first nine cases of the data matrix look as follows:

1	0	0	1	59.5	168.0	102.43	1	0.00	0	0	1	17.67	4.88	5.81	1.50	1.53	1
1	0	0	0	86.8	190.0	65.82	2	0.00	0	0	1	11.61	3.26	4.51	0.93	0.84	2
1	0	0	1	56.5	168.2	29.54	4	0.00	0	1	2	20.29	5.98	7.05	1.71	1.80	3
1	0	0	0	76.1	172.1	34.91	2	9.17	23	0	1	30.91	7.48	14.10	2.53	2.59	4
1	0	0	0	74.4	171.3	36.66	2	6.60	31	0	1	16.00	4.58	4.69	1.24	1.32	5
1	0	0	1	72.8	163.6	35.21	3	14.27	27	0	1	121.57	23.58	46.94	10.54	10.80	6
1	0	0	1	71.7	171.2	45.32	3	8.59	23	0	1	19.72	5.61	7.46	1.62	1.41	7
1	0	0	0	74.3	175.8	24.10	3	0.00	0	0	1	25.58	6.50	10.08	2.06	2.19	8
1	0	0	1	55.3	158.7	22.75	1	6.18	31	0	1	19.20	5.70	7.98	1.69	1.54	9

There are 18 columns that correspond to the 18 variables as they are described above. Every row corresponds to the data of one adult individual.

2. Into SPSS

To be able to analyze the data of Statistica with SPSS, as a first step we have to start up SPSS. A spreadsheet-like screen appears (comparable to Excel). To analyze data, these data first have to be read into SPSS. They are put into a data sheet, like in Excel. How to import the data, depends on how the data are stored. The most important types of data we will read into SPSS are:

1) SPSS files: these are files in SPSS format. Usually they are created with SPSS. They can be imported directly. Usually they have the extension `.sav`. The advantage of using this type of files is that not only the values of the variables are stored, but also the names of the variables, and, if required, the meaning of certain values of the variables (***value labels***, see below).

2) ASCII files. These are data in text format, which can be read by all word processors and statistical software packages. For example, the data in the `Statisti.dat` file are stored in this format. ASCII files only contain the values of the variables. The names of the variables (and value labels, if required) have to be added by the user before the data can be analyzed.

Below, it will be assumed that all data are in de dot (.) mode and not in the comma (,) mode. That is, it is assumed that your computer and SPSS want to have the data in such a way that the decimal dot is used and not the decimal comma.

In order to be able to do the exercises, we first consider reading data of type 2). Again, we use the Statistica data to illustrate.

3. How to read ASCII-data

The ASCII file `Statisti.dat` contains the data matrix of which the first few rows are printed above. From the data matrix, it is not clear what the data mean and what the variables should be called. We have to add this information ourselves.

For reading a data matrix, choose from the SPSS-menu the option

```
File|Read Text Data
```

Go to the correct directory to find the desired file. Then, 6 steps are needed: In step 2, choose `Delimited`. Choose '`Finish`' when you arrive at step 6. The (heading) names of the variables can be filled in (in the sheet `Variable View`) after having opened the dataset.

Exercise: Read the data into SPSS and (in the sheet `Variable View`) add the names of the variables: respectively, ds dh dp df weight length age edu wage hours nkids fs finc foodexp housexp clotexp recrexp findex

We can improve the documentation of the data by adding ***variable labels*** and ***value labels.*** The variable labels are the long descriptions of the variables. Value labels are the descriptions of the categories of the variables. As an illustration, we will consider Statisti.dat. Carry out the following actions.

1. Choose the sheet `Variable View`
2. Click in the screen that has appeared on `Label`
3. Enter for the first variable the text "dummy single"
4. Enter in `Values` the `Value` "0" and `Label` "in a more-person household" and click on `Add`
5. Also enter in `Values` the `Value` "1" and `Label` "in a one-person household" and click on `Add`
6. Click on `OK`; now the description of the variable ds is complete

Exercise: Document also the variables df and weight; why does it not make sense to make value labels for the variable weight?

4. Storing in SPSS format

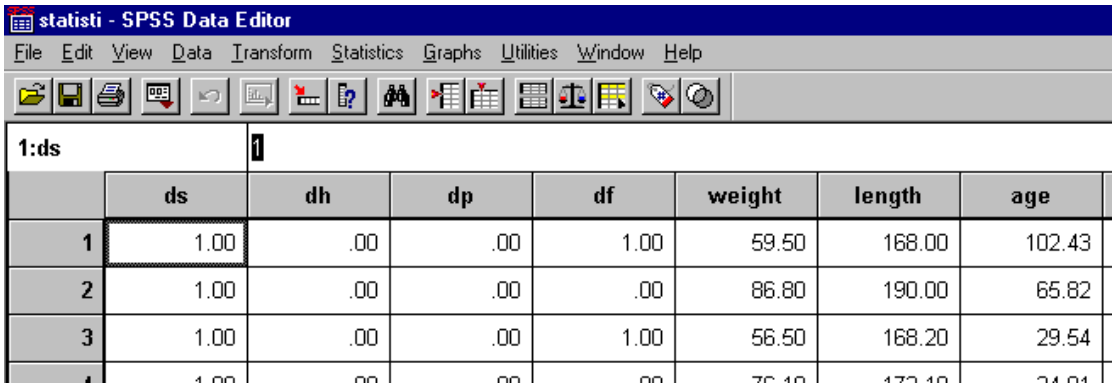
You can store the thus documented file in such a way that this information about variable names, labels and value labels is saved (with extension `.sav`) by clicking the save-icon.

5. Importing Data in SPSS format

When we have a data file in SPSS format, including names of the variables and possibly value labels, etc., it is very easy to load it into SPSS for (further) analysis. We have to fetch the data that will fill the spreadsheet. The name of the Statistica file in SPSS format is `statisti_all.sav`. By using

```
File|Open|Data
```

this file can be found and read into SPSS. As also appears from the picture below, SPSS already 'knows' what the data stand for: the names of the variables are written above the columns. The case-numbers are given in front of the rows. You may obtain more information by clicking on `Utilities|Variables`. By clicking on the variables, their descriptions are made visible. A text string which describes a variable is called a ***variable label***. Text strings that describe the categories of the categorical variables are called ***value labels***.



	ds	dh	dp	df	weight	length	age
1	1.00	.00	.00	1.00	59.50	168.00	102.43
2	1.00	.00	.00	.00	86.80	190.00	65.82
3	1.00	.00	.00	1.00	56.50	168.20	29.54

6. Describing one variable: univariate exploratory analysis

We are ready now to analyze the data. We start with the well-known methods to describe data. This can be done in two ways: numerically and graphically. First we examine every variable separately. The numerical way to look at data is to request frequency distributions (categorical variables) or means and variances (numerical variables).

Frequencies
All numerical analyses (and also some graphical analyses) in the menu fall under the header `Analyze`. Frequency distributions can be generated by

```
Analyze|Descriptive Statistics|Frequencies
```

There appears a list of variables that can be brought to an empty window by using the arrow. Now carry out the following actions.

1. Click on *df* and then on the arrow
2. Click on *edu* and then on the arrow
3. Click on `OK`

Now an output screen appears with results. Interpret these results.

Next, we want to illustrate the outcomes with a picture. Go back to

```
Analyze|Descriptive Statistics| Frequencies
```

and note that the previous setup is still available. Now carry out the following actions:

1. Click on `charts`
2. Click on `bar charts`
3. Click on `Continue`
4. Click on `OK`

Now you see, apart from the numerical outcomes, also the bar chart depicted in the output window.

One final question: you can apply the procedure `Frequencies` also to the variable *weight*. Why is that not a good idea?

Descriptives
The most important procedure to examine numerical data is `Descriptives`. This procedure is reached from the SPSS main menu with

```
Analyze|Descriptive Statistics|Descriptives
```

Now there appear, just as with `Frequencies`, two windows: one in which the variables from the dataset are displayed and an empty window that can be filled with variables by clicking the arrow. Now carry out the following actions:

1. Click on *housexp*
2. Click on the arrow
3. Do the same for *recrexp*, *foodexp* and *clotexp*
4. Click on `OK`

Again, an output window appears. Interpret its contents.

Of course, there are again ways to adapt the output to our wishes. Click (in `Descriptives`) on `Options` and request apart from the already indicated measures also the standard error of the mean and the variance (what, again, were the differences with standard deviation?).

Graphs
The graphical module allows for a multitude of pictures that can be made from the data. We do this using the expense-variables that were already analyzed with `Descriptives`. Click on:

```
Graphs|Bar
```

Now a window appears in which the bar chart can be specified. Click on `Simple` and on `Summaries of separate variables`. Next, click on `Define`. Now a screen appears in which we can select variables in the usual way. Select the variables *housexp*, *recrexp*, *foodexp* and *clotexp*, transport them to `Bars Represent`, and click on `OK`. In the output window now the bar chart appears. Interpret it!

Exercise: Find out - starting from `Graph` - how you can make a line-chart, area chart and pie chart. Which one is the most attractive?

Exercise: Make from Graph a histogram of *foodexp*, and also show the best fitting normal distribution in the graph.

Exercise: Create a bar chart for *edu* from Analyze|Descriptive Statistics |Frequencies.

7. Describing two variables: bivariate exploratory analysis

More exciting than examining the distributions of single variables is the analysis of relations **between** variables. This makes it clear how different variables within a subject of interest are connected. Here we restrict ourselves to relations between pairs of variables and make the distinction between categorical and numerical variables.

Two categorical variables
Categorical variables have the property that it is illegal to make calculations with the category numbers. As a consequence, relations between categorical variables can primarily only be represented by tables. SPSS contains a number of procedures to generate tables. We restrict ourselves here to the most commonly used procedure:

Analyze|Descriptive Statistics|Crosstabs

After clicking Crosstabs, a window with variables appears. Besides, there appear, amongst others, a window with the title Row(s) and a window with the title Column(s). By using the arrows, the variables can be transported to these windows. Now carry out the following actions

1. Click on *edu*
2. Click on the arrow in front of Row(s)
3. Click on *dp*
4. Click on the arrow in front of Column(s)
5. Click on OK

In the output window there appears a cross-table of 'education' by 'having a partner'. The numbers in the table do not give a clear analysis. The interpretation of the table becomes easier when percentages are calculated. To achieve this, carry out the following actions:

1. Go back to Crosstabs
2. Click on Cells
3. Click Observed off
4. Click under the header Percentages on Column
5. Click on Continue and next on OK

Now the table of 'education' by 'partner' looks different, and it is easier to make a statement about the education level of partners in more-person households. What do you think of this level?

Exercise: Use the table to compare education levels of heads and partners.

Exercise: Make a two-way table of *dh* by *dp*; explain why one of the cells is empty.

A categorical and a numerical variable
The simplest way to relate a categorical and a numerical variable to each other is to compare the means of the numerical variable within the categories of the categorical variable. Also for this analysis SPSS has a simple procedure, which can be called from the main menu by;

```
Analyze|Compare Means|Means
```

Then a screen appears with the familiar list of variables and two windows with titles `Dependent List` en `Independent List`, respectively. In the dependent list the dependent variables are selected; these are the numerical variables. In the independent list the independent, categorical variables are selected. Now carry out the following actions:

1. Put in `Dependent List` the variables *foodexp*, *clotexp*, *recrexp* and *housexp*.
2. Put in the `Independent List` the variable *fs*
3. Click on `OK`

The output screen now shows the different average expenditures by size of the household.

Of course it is attractive to show these averages in a graphical way. We can do this again with a bar chart. Click on:

```
Graphs|Bar
```

and make a `Simple` bar chart with the option `summaries for groups of cases`. Next, carry out the following actions:

1. Click on `Define`
2. Move the variable *fs* to `Category Axis` using the arrow
3. Check `Other statistic (e.g., mean)`
4. Put the variable *foodexp* in `Variable`
5. Click on `OK`

The output window then shows the bar chart we were looking for.

Exercise: Construct in the same way a line chart, an area chart and a pie chart. In which respect is the pie chart different? What could be the reason for that?

Two numerical variables

Numerical variables often do not lend themselves for representation by tables. It is, however, possible to express the degree of relationship between two numerical variables by a number. This number is the correlation coefficient (see Section 5.1 of the book *Statistical Methods for Business and Economic*). It can be requested in SPSS by:

`Analyze|Correlate|Bivariate`

After clicking these options in the menu, there appears a screen with variables and an empty window, in which the variables can be placed in the usual way. Now carry out the following actions

1. Move the variables *weight* and *length* to the `Variables`-window
2. Click on `OK`

From the output it appears that SPSS writes the correlations in a matrix. As a consequence, not only the correlation between *weight* and *length* is given, but also between *weight* and *weight* and between *length* and *length* (which, of course, are equal to 1).

Exercise: Calculate with SPSS the correlation matrix of the variables *finc*, *foodexp*, *clotexp*, *housexp* en *recrexp*.

A graphical way to represent the relations between numerical variables is the ***scatterplot***. This plot can be requested from the SPSS menu by:

`Graphs|Scatter`

Next, click `Simple`, and then `Define`. After that, variables can again be selected. Carry out the following actions:

Put *length* in `X-Axis`
Put *weight* in `Y-Axis`
Click on `OK`

Now a scatterplot of *weight* by *length* has appeared. Because of the large number of cases, many individual points are not visible, but it is clear that, generally speaking, weight increases with length.

In an ***overlay-plot*** various scatterplots are displayed in one graph. In order to define scatterplots for overlay plots, pairs of variables have to be chosen together.

Exercise: Construct an overlay-plot of *foodexp*, *clotexp*, *housexp* and *recrexp* on the *Y*-axis and *finc* on the *X*-axis.

From the pictures of the above exercise it appears that the relations between *family income* and the various expenditure variables can be well described by straight lines (this also was clear from the correlation coefficients), but that the slopes (angles) of the lines are different. This, again, can be numerically confirmed by linear regression.

Simple Linear Regression
For this, SPSS also has a procedure. From the SPSS menu it can be reached by:

```
Analyze|Regression|Linear
```

There appears a screen (the so-called ***Linear Regression Window***) in which variables can be selected. Now carry out the following actions:

Put the variable *foodexp* in the window `Dependent`
Put the variable *finc* in the window `Independent(s)`
Click on `OK`

In the output screen a comprehensive description of the regression analysis appears. It is important that in the `Coefficients` part, in the column with heading `B` at `net family income * 1000`, the value 0.235 appears. This is the value of b_1 in the regression equation $y = b_0 + b_1 x$, where y is 'expenditure on food' and x is 'family income'.

Exercise: Carry out the same procedure for *clotexp*, *recrexp* and *housexp*, and verify that the output is in accordance with the overlay plot.

Part III: Sections A1.16 – A1.25

A1.16 Confidence intervals and tests for μ and p

Calculation in Excel of probabilities about Y if $Y \sim t_{n-1}$

Recall that the Excel-functions `normsdist` and `normsinv` for the standard normal distribution were considered in Section A1.10; they "look to the left". The corresponding Excel-functions for t_{n-1} look to the **right**.

- $P(Y > b)$ can be calculated with `tdist(b,n-1,1)`.
  (For technical reasons, the **1** has to be added in this command.)
- So, $P(Y \leq b)$ can be calculated with **1-**`tdist(b,n-1,1)`
- For given $A = P(Z > b)$, the constant b follows with `tinv(2A,n-1)`

For instance for $Y \sim t_{13}$:

- $P(Y > 0.2) = 0.4223$, which follows from `tdist(0.2,13,1)`
- The constant b with $P(Y \leq b) = 0.95$ satisfies: $P(Y > b) = 0.05$. So, it follows from `tinv(0.1,13)`; the answer is $b = 1.7709$.

Remarks:

- All Excel functions `tdist` and `tinv` look to the **right**.
- In the command `tinv`, the right-hand area has to be **doubled**.

Calculation of $t_{\alpha;n-1}$ with Excel

- Since $t_{\alpha;n-1}$ wants an area α at its **right**-hand side, it can be calculated with `tinv(2α,n-1)`

For instance:

$t_{0.05;13}$ follows with `tinv(0.10,13)`

Intermezzo: how to open an Excel-file in SPSS?

From the `SPSS Data Editor` onwards, use `Open` and `Data`. To find the Excel-dataset, don't forget to choose `All files` in `Files of type`. If necessary, check `Read variable names`. Be careful; don't loose observations!

Tests and CI's with respect to p (SPSS)

The only sample statistic that is needed for such tests and CI's, is the sample proportion $\hat{p}$. With SPSS, it can be obtained with:

`Analyze / Descriptive Statistics / Frequencies`

Next, the value of the test statistics and/or the CI can be calculated easily.

Tests and CI's with respect to μ (SPSS)

Two possible methods are mentioned below. The first only calculates the values $\bar{x}$ and s, so that the value of the test statistic and/or the CI can be calculated easily. The second is

the standard SPSS procedure that immediately yields the value of the test statistic and a **95%-**confidence interval.

1) Compute the sample statistics 'sample mean' and 'sample standard deviation' (and substitute them into the formulas of the test statistic and the interval estimator). With SPSS, these statistics can be calculated with:
   `Analyze / Descriptive Statistics / Descriptives`
2) Use the **one-sample *t*-test** button of SPSS, under:
   `Analyze / Compare Means / One-Sample T Test`
   (SPSS uses `Test Value` to denote the **hinge**.)

A1.17 Statistical inference about σ^2

It will be explained how to use Excel for the calculation of probabilities and quantiles regarding a χ^2_ν-distributed random variable. Furthermore, it is explained how SPSS can be used to create a histogram with best-fitting normal curve included and to calculate the sample mean and sample standard deviation of a dataset.

Below, it is assumed that the rv *W* has the probability distribution χ^2_ν.

The Excel-function `chidist`

The function `chidist` can be used to calculate probabilities with respect to *W*. The function calculates areas under the χ^2_ν-density and looks **to the right**. Here is the general rule:

$$P(W > b) = \texttt{chidist(b; } \nu\texttt{)}$$

Hence, probabilities like $P(W < b)$ and $P(c < W < d)$ cannot be determined directly; you'll first have to calculate the probabilities $P(W > b)$, and $P(W > c)$ and $P(W > d)$. Always make rough graphs of the density and indicate the area that is asked for.

The Excel-function `chiinv`

The function `chiinv` works the other way round. It calculates quantiles of the distribution χ^2_ν. To determine $\chi^2_{\alpha;\nu}$ (the $(1-\alpha)$-quantile that cuts of an area α at the right-hand side), the following rule can be used:

$$\chi^2_{\alpha;\nu} = \texttt{chiinv(}\alpha\texttt{; } \nu\texttt{)}$$

Notice that the position under the χ^2_ν-density that cuts off an area *p* at the **left**-hand side, cuts of an area $1 - p$ at the right-hand side; so it is equal to $\chi^2_{1-p;\nu}$.

Calculation of sample mean and sample standard deviation with SPSS

- Open your dataset;
- `Analyze` / `Descriptive Statistics` / `Descriptives`;
- Import your variable(s) into the window `Variable(s)`;
- `OK`.

Histogram with SPSS

- `Graphs` / `Histogram`;
- Import your variable into `Variable`;
- Check `Display normal curve` (if wanted);
- `OK`.

The computation of the difference D of variables X and Y in different columns (SPSS)
`Transform` / `Compute Variable`; type D in `Target Variable` and $X–Y$ in `Numeric Expression`; `OK`.

A1.18 Confidence intervals and tests to compare two parameters

SPSS is case-oriented and Excel is cell-oriented
Recall from the first part of Section 18.2 that SPSS is **row-oriented** (also called **case-oriented**) and Excel is **cell-oriented**; see also Example 2.2. The standard procedures of SPSS consider the information along one row as belonging to one element (case). That is why SPSS prefers to have the data of two independent samples (1 and 2) below each other, in one column; a second column informs about the number of the sample (1 or 2) where the observations come from.

When the data are paired as in the **paired observations design** b), SPSS prefers to have them organised in pairs along the rows. Hence, the first column contains the data of sample 1 and the second column contains the data of sample 2.

In Excel it is usually best to put the data of the two samples in different columns, for both designs a) and b).

The Excel-function `Fdist`

The function `Fdist` can be used to calculate probabilities with respect to an F_{ν_1,ν_2}-distributed random variable F. The function calculates areas under the F_{ν_1,ν_2}-density and looks **to the right**. Here is the general rule:

$$P(F > b) = \texttt{Fdist(b;}\nu_1,\nu_2\texttt{)}$$

Hence, probabilities like $P(F < b)$ and $P(c < F < d)$ cannot be determined directly; you'll first have to calculate the probabilities $P(F > b)$, and $P(F > c)$ and $P(F > d)$. Always make rough graphs of the density and indicate the area that is asked for.

The Excel-function `Finv`

The function `Finv` works the other way round. It calculates quantiles of the distribution F_{ν_1,ν_2}. To determine $F_{\alpha;\nu_1,\nu_2}$ (the $(1-\alpha)$-quantile that cuts of an area α at the **right**-hand side), the following rule can be used:

$$F_{\alpha;\nu_1,\nu_2} = \texttt{Finv}(\alpha;\nu_1,\nu_2)$$

Notice that the position under the F_{ν_1,ν_2}-density that cuts off an area *p* at the **left**-hand side, cuts of an area $1-p$ at the right-hand side; so it is equal to $F_{1-p;\nu_1,\nu_2}$.

The standard F-test in SPSS is different from the one in Section 18.3

SPSS offers a standard test to find out whether two population variances are equal. However, the test statistic that is used in that test is **not** the same as the test statistic presented in the Section 18.3 of the book.

Non-standard SPSS approach that is useful for the calculation of many CIs and tests

The ingredients of many confidence intervals and test statistics can be obtained just by using the SPSS command:

`Analyze` / `Descriptive Statistics` / `Descriptives`

This approach also offers alternatives to the two standard tests below.

The standard independent samples t-test of SPSS

This standard test can only be used correctly in case of the **independent samples** experimental design. Furthermore, the data have to be organised in the way SPSS prefers it; row-oriented, see Section 18.2.

- `Analyze` / `Compare Means` / `Independent-Samples T Test`;
- import the sample data in `Test Variable(s)`;
- import the column with the sample-numbers into `Grouping Variable`; click `Define Groups` and fill in `Group` 1 and `Group` 2;
- `Continue` and `OK`.

The standard paired-samples t-test of SPSS

This standard test can only be used correctly in case of a paired-samples experiment. The data have to be organised in the way SPSS prefers it, in different columns.

- `Analyze` / `Compare Means` / `Paired-Samples T Test`;
- import the pair of variables into `Paired Variables`;
- `OK`.

Non-standard way to do the paired-samples t-test with SPSS

This is only possible if the two sets of observations are in two separate columns of equal length.

- Use Transform / Compute Variable to create the column of the differences;
- use Analyze / Descriptive Statistics / Descriptives to determine the sample mean and sample standard deviation of the differences;
- use these results to calculate the ***val*** yourself.

Histograms (SPSS) for separate groups when the data are in one column

- Graphs / Interactive / Histogram;
- drag the variable with the data into the upper empty box;
- drag the variable (often string (categorical), but not always) that distinguishes the subgroups into the lower empty box;
- (if wanted, under the tab Histogram, check Normal curve);
- OK.

About z-tests to compare population proportions with SPSS

The necessary ingredients of interval estimators and test statistics can often easily be found by way of the SPSS option:

Analyze / Descriptive Statistics / Frequencies

A1.19 Simple linear regression

It is explained how SPSS can be used to run the regression of *y* on *x*. Furthermore, the creation of confidence and prediction intervals is considered, and it is explained how residuals, standardized residuals and predicted values can be calculated with SPSS.

(Simple and overlay) scatter plots with regression line (SPSS)

- Graphs / Scatter;
- If you want to compare two variables: choose Simple; check Define; import the dependent variable into Y-Axis and the independent variable into X-Axis; OK.
- If you want to have more than one scatter plot in the same figure (hence, with the same horizontal variable): choose Overlay; select one pair of variables and import them into Y-X Pairs; if necessary, use the Swap Pair button ↔ to change the order. Do the same for another pair; etc. Click OK.
- To add the regression line, double left-click the scatter plot to initiate the Chart Editor; right click one of the dots; choose Add Fit Line at Total.

Running the regression of y on x with SPSS

Firstly, open the dataset of interest.

- Analyze / Regression / Linear; the Linear Regression Window appears;

- use the arrows to import your dependent variable into `Dependent` and your independent variable(s) into `Independent(s)`;
- press `OK`.

Determination of a 95%-CI for β_1

In the Linear Regression Window, click the `Statistics` button and check `Confidence Intervals`, and `Continue` and `OK`. Notice that SPSS can only calculate CI's with the confidence level 95%.

Determination (SPSS) of a CI for $E(Y_p)$ ***and a PI for*** Y_p ***when*** x_p ***is given***

Firstly, type the value of x_p into a new row (case) of the SPSS data sheet. Since Y_p is not (yet) observed, the accompanying observation of Y cannot be typed. We start from the Linear Regression Window onwards.

- Click the `Save` button;
- under `Prediction Intervals`, check `Mean` (if you want a CI for $E(Y_p)$) and/or `Individual` (if you want a PI for Y_p);
- indicate the confidence level that you want;
- click `Continue` and `OK`.

The interval(s) for x_p appear(s) in the data sheet (not as `Output`). The lower and upper bounds of the CI for $E(Y_p)$ respectively arise in the columns called `LMCI` and `UMCI` (Lower / Upper CI for the Mean). The lower and upper bounds of the PI for Y_p respectively arise in the columns called `LICI` and `UICI` (Lower / Upper CI for the Individual value of Y_p). Notice that SPSS is not very consistent: the PI is also denoted as CI.

Creation of all (standardised) residuals and all predictions $\hat{y}_i$

Start from the Linear Regression Window.

- Click the `Save` button;
- under `Predicted Values`, check `Unstandardized` to create all $\hat{y}_i$;
- under `Residuals`, check `Unstandardized` (for the e_i) and/or `Standardized` (for all standardised residuals);
- `Continue` and `OK`.

The $\hat{y}_i$, e_i and standardised residuals appear in the SPSS data sheet under the respective headings `PRE`, `RES`, `ZRE`.

More decimals

To obtain more decimals for a number in the printout, left-click the number a few times.

A1.20 Multiple linear regression: introduction

The SPSS-techniques are similar to Chapter 19; see Appendix A1.19.

Running the regression of y on $x_1, \cdots, x_k$ with SPSS

Firstly, open the dataset of interest.

- `Analyze / Regression / Linear`; the Linear Regression Window appears;
- use the arrows to import your dependent variable into `Dependent` and your independent variables into `Independent(s)`;
- press `OK`.

Determination (SPSS) of a CI for $E(Y)$ and a PI for Y when $x_1, \cdots, x_k$ are given

In SPSS, this construction is started up by adding an extra case to the *n* cases that are already present in the data sheet. For this extra case the observation of *Y* cannot be filled in (since it is not observed yet) but the observations of the independent variables are included. Next: the regression is run and in the `Regression Window`, under the `Save` button (Prediction interval), the options `Mean` (duty (i), see the book) and/or `Individual` (duty (ii)) are checked and the confidence level is indicated.

Creation (SPSS) of all (standardised) residuals e_i and predictions $\hat{y}_i$

See Appendix A1.19.

A1.21 Multiple linear regression: extension

It will be explained how (in SPSS) the column with observations of a **function** $h(X)$ of *X* can be created when the column with observations of *X* are already given. Especially, X^2, the natural logarithm log(*X*) and **dummy variables** based on *X* will be considered. Also the observations of **interaction terms** can be obtained easily. To detect collinearity, the **correlation matrix** is important.

To see the developments when adding new variables to a linear regression model, the **block-approach** is a useful tool. It easily generates the value of the test statistic in a partial *F*-test. **Stepwise regression** is often used to get a first idea about useful independent variables within a list of potential predictors. However, it is better to use common sense than the biased conclusions of a statistical package.

Creation (SPSS) of the observations of $h(X)$

- `Transform / Compute Variable`;
- type the name (that you want to assign to $h(X)$) in `Target Variable` (for instance, a suitable naming for $h(X) = X^2$ and $h(X) = \ln(X)$ is respectively XX and LOGX);
- in `Numeric Expression`, type the mathematical expression of $h(X)$ (for instance, if $h(X) = X^2$, type X*X; if $h(X) = \ln(X)$, type LN(X) or import `LN(numexpr)` from the `Functions` listing);

- OK.

The creation (SPSS) of one dummy variable based on a <u>string</u> variable V

Starting point is a string (= qualitative) variable *V*; assume that one of the outcomes is *A*. The data of *V* are in one of the columns.

The procedure creates a column with 1 if *V* equals *A* and 0 if *V* is not *A*.

- Transform / Compute Variable;
- in Target Variable, type the name that you want to assign to the dummy (for instance, D);
- in Numeric Expression, type *V*='*A*' and click OK

The creation (SPSS) of one dummy variable based on a <u>numeric</u> variable V

Starting point is a numeric (= quantitative) variable *V*; one of the outcomes is (say) 18. The data of *V* are in one of the columns.

The procedure to create a column with 1 if *V* equals 18 and 0 otherwise, is similar to the procedure for a qualitative variable, but now it has to be typed: *V*=18 (without quotes).

The creation (SPSS) of m-1 dummies if the variable V takes m levels

First find out whether the variable is string or numeric and decide which level will be base level.

- Carry out the above procedure to create the first dummy;
- do the procedure again for the second dummy, but give this dummy a different name under Target Variable
- do the procedure again for the third button (if necessary), etc.

The creation (SPSS) of the interaction term X_1X_2 from the variables X_1 and X_2

- Transform / Compute Variable;
- in Target Variable, type the name that you want to assign to the interaction term (for instance, X1X2);
- in Numeric Expression, type X1*X2 and click OK

Correlation matrix (SPSS)

- Analyze / Correlate / Bivariate;
- import the variables for which you want pair-wise correlations;
- check Pearson; OK.

Importing the independent variables in blocks (SPSS)

This **block-approach** is suitable when you want to see the SSE decreasing and the r^2 increasing when new variables are added to the regression model. It is also useful for partial *F*-tests.

- Analyze / Regression / Linear;
- import the independent variables you want to include firstly (for instance, for the reduced model);

- click Next and import the independent variables you want to include secondly (for instance, the additional variables that complete the complete model);
- if wanted, repeat the last step;
- under the Statistics button, check R squared change;
- Continue; OK.
  (For the partial *F*-test, the ***val*** can be found in the Model Summary part of the printout: the number under the heading F change that belongs to Model 2.)

Stepwise regression (SPSS)

- Analyze / Regression / Linear;
- import the dependent and independent variables;
- under Method, choose Stepwise

A1.22 Multiple linear regression: model violations

Weighted least-squares and SPSS

- Analyze / Regression / Linear;
- in the linear regression window, import *y* and *x* in the usual way;
- import the weights into WLS Weight

The value of the DW-statistic D (with SPSS)

- Analyze / Regression / Linear;
- in the linear regression window, click the Statistics button and check Durbin-Watson;
- the value appears in the printout.

Note that the DW-bounds (d_L and d_U) cannot be obtained with SPSS (or Excel). You will need a table.

Two-stage least-squares and SPSS

- Analyze / Regression / 2-Stage Least Squares;
- import *y* into Dependent, *x* into Explanatory, the observations of the instrumental variable into Instrumental.

Logistic regression and SPSS

- Analyze / Regression / Binary Logistic;
- import *y* into Dependent and the *x*-variable(s) into Covariates

Creation (SPSS) of lags of a variable Y:

- Transform / Compute Variable;
- for lag 1: type (for instance) Y_1 in Target Variable and LAG(Y) in Numeric Expression;

- for lag 2: type (for instance) Y_2 in `Target Variable` and LAG(Y,2) in `Numeric Expression`.

Lilliefors test for normality (SPSS)

- Use the *z*-data (the data after having standardised your original data);
- `Analyze` / `Descriptive Statistics` / `Explore`; check `Both`, and under `Plots` check `Normality Plots with Tests`.

Lilliefors test for normality of residuals (SPSS)

Conduct the above analysis on the sequence of residuals.

A1.23 Time series and forecasting

Many statistical packages offer special procedures to create a moving average series or an exponentially smoothed series from a time series $y_1,\cdots,y_n$. However, it always takes some time to learn how these procedures have to be applied. Moreover, the researcher has to be careful since the procedures can be different from the procedures sketched in this book.

(But essentially there hardly is a need to generate such moving average and exponentially smoothed series with a statistical package since they easily can be created with Excel and, if wanted, copy-pasted into the data sheet that is used. Below, it will be explained how moving average and exponentially smoothed series can be generated with Excel.)

The creation of a moving average series with Excel

Suppose that the underlying time series is in field A1-An, and that a 3-period moving average series is wanted.

- Type in B2: =(A1+A2+A3)/3;
- press `Enter`;
- use the lower-right angle of B2 (the fill-handle) to draw the formula to position B(n-1);
- the series arises in the positions B2-B(n-1).

The creation of an exponentially smoothed series with Excel

Suppose that the underlying time series is in field A1-An, and that an exponentially smoothed series with smoothing constant 0.3 is wanted.

- Type in B1: =A1 and press `Enter`;
- type in B2: =0.3*A2+0.7*B1 and press `Enter`;
- select B2 and use the lower-right angle (the fill-handle) of B2 to draw the formula to position Bn;
- the smoothed series arises in the positions B1-Bn

The creation of a moving average smoothed series with SPSS

We only consider the procedure for the *m*-period case for **odd** values of *m*; we demonstrate it for $m = 3$. The time series $y_1,...,y_n$ comes from the variable *Y*.

- Under `Transform/Create Time Series/Function`, choose `Centered Moving Average` and take `Span` $m = 3$;
- import *Y* into `New Variable(s)`; `OK`.
  (The new variable Y_1 in the data sheet is the MA-series.)

The creation of an exponentially smoothed series with SPSS

We consider exponential smoothing of a time series $y_1,...,y_n$ (of a variable *Y*) with smoothing constant $w = 0.8$. The procedure is based on the LAG function.

- First, create the column $T = 1, \ldots, n$ via `Transform/Compute Variable` by typing `$casenum` (one of the functions in the `All`-list) in `Numeric Expression`;
- next, use `Transform/Compute Variable` to create the variable *S* that has y_1 (the first value of *Y*) on **all** positions; so, if $y_1 = 8$ type 8 in `Numeric Expression`; `OK`;
- again, go to `Transform/Compute Variable` and replace for *S* the numeric expression by 0.8*Y+0.2*LAG(S,1); click the `If` button, check `Include if case satisfies condition` and type T>1 in the box; `Continue`; `OK`; `OK`.

The creation of trend- and dummy columns in time series regression models (SPSS)

Things are illustrated for quarterly data, in presence of a column called Q = 'Quarter' with values 1, 2, 3 and 4.

- Create the column T = 'time' with values $1, 2, \cdots, n$ via `Transform/Compute Variable` by typing `$casenum` (one of the functions in the `All`-list) in `Numeric Expression`;
- to generate T^2 or T^3, the SPSS option `Transform` / `Compute` can again be used;
- the quarter dummies can in SPSS be created (via `Transform` / `Compute`) in the way you learned before, by making use of the variable *Q* that refers to the successive quarters within each year. For instance: for the dummy D1 of quarter 1 type Q = 1 in `Numeric Expression`.

A1.24 Chi-square tests

Chi-square goodness of fit tests (SPSS)

As input dataset, SPSS does not use the original data points. Instead the corresponding (integer-valued) numbers of the classes are used. So, the (transformed) dataset of size *n* has to consists of the numbers of the classes (usually the numbers $1, 2, \ldots, k$).

- `Analyze/Nonparametric Tests/Chi Square`;

- import the variable into `Test Variable List`;
- in `Expected Values`, check `Values` and fill in (and `Add`) the corresponding values of e_l (or E_l) for all classes l. `Values` lets you enter a list of values that are proportional to your expectations, so it is also possible to fill in (and `Add`) the corresponding values of p_i as specified in H_0 for each class. Note that the values must appear in order, so the first value corresponds to class 1 and the last value corresponds to class k. If you assume in H_0 that the variable is equally distributed among all categories, then you can just check the box `All categories equal`.
- In the printout, the ***val*** is called Chi-Square.

Test for association (independence) in SPSS
- `Analyze / Descriptive Statistics / Crosstabs`;
- put one variable in `Row(s)`, the other in `Column(s)`;
- check `Chi-Square` under `Statistics`;
- check `Observed` and `Expected` under `Cells`
  (In the printout, the ***val*** is called Pearson Chi-Square.)

A1.25 Non-parametric statistics

Wilcoxon rank sum test (for two independent samples) with SPSS
The dataset has to be organized as SPSS prefers it: case oriented. That is, all data have to be in one column (say X) while another column (say S) gives the sample (1 or 2).
- `Analyze/Nonparametric Tests/2 Independent Samples`;
- import the data column X into `Test Variable List`;
- import S into `Grouping Variable` and `Define` your `Groups`, `Continue`;
- check `Mann-Whitney U`;
- `OK`.
  (The printout gives the ***val***s of the Wilcoxon rank sum test and its large-samples version respectively under `Wilcoxon W` and `Z`.)

Sign test (for two matched samples) with SPSS
The two (paired) samples have to be in different columns.
- `Analyze/Nonparametric Tests/2 Related Samples`;
- import the pair of variables into `Test Pairs`;
- check `Sign`; `OK`.
  (The ***val*** of the sign test can easily be calculated from the `Frequencies` part of the printout.)

Wilcoxon's signed rank sum test (for two matched samples) with SPSS
The two (paired) samples have to be in different columns.
- `Analyze/Nonparametric Tests/2 Related Samples`;
- import the pair of variables into `Test Pairs`;

- check `Wilcoxon`; `OK`.
  (The ***val*** of Wilcoxon's signed rank sum test is called *Z*, in `Test Statistics`.)

Kruskal-Wallis test (for two or more independent variables) with SPSS

The dataset has to be organized as SPSS prefers it: case oriented. That is, all data have to be in one column (say *X*) while another column (say *S*) gives the sample (1, 2, ..., *k*).

- `Analyze/Nonparametric Tests/k Independent Samples`;
- import the data column *X* into `Test Variable List`;
- import *S* into `Grouping Variable` and `Define` your `Groups`, `Continue`;
- check `Kruskal-Wallis H`;
- `OK`.
  (The printout gives the ***val*** of the test after `Chi-Square`.)