AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 2, 195–206

Teaching Statistics with Excel
A Big Challenge for Students and Lecturers

Christine Duller
Johannes Kepler University Linz

Abstract: This paper has a look on the implementation of basic statistical methods in Excel, in particular in describing data. Often a frequency distribution is the starting point for describing a dataset, but its calculation is not that easy in Excel. Therefore the first section gives some alternatives to the common way. The next part presents some interesting solutions for measures of central tendency and variability provided by Excel. The last important topic is about creating graphs, where Excel has an advantage: creating histograms with unequal interval width is possible.

Zusammenfassung: Dieser Beitrag diskutiert einige der Probleme, die bei der deskriptiven Statistik mit Excel auftreten. Häufigkeitsverteilungen sind meistens der Ausgangspunkt für eine umfassendere Datenanalyse, aber deren Berechnung in Excel ist nicht so leicht wie man vermuten könnte. Daher werden im ersten Abschnitt Alternativen zum üblichen Weg aufgezeigt. Der folgende Abschnitt zeigt interessante Ergebnisse bei der Berechnung von Lage- und Streuungsmaßzahlen. Der letzte wichtige Bereich ist der Erstellung von Grafiken gewidmet, bei denen Excel einen wesentlichen Vorteil bietet: Histogramme sind auch mit unterschiedlichen Intervallbreiten darstellbar.

Keywords: Spreadsheet, Frequency Distribution, Descriptive Statistics, Histogram.

1 Introduction

Teaching statistics is a big challenge, teaching statistics with Excel is an even bigger challenge. But you have to accept it, because special software for statistical analysis is either very expensive (e.g. SAS) or is provided with a command line interface (e.g. R) and therefore not easy to use for somebody without any experience.

Excel is used in many different ways for teaching statistics: Simulations are used to teach distributions (Doane, 2004) or for demonstrating experimental power and variability (Horgan, 1999; Martin, 2007; Mills, 2002). Combinatorial ideas can be illustrated (Kühleitner, 2007; Borovcnik, 2007) and calculation and illustration of probabilities (Bartz, 2007) or frequencies (Hunt and Mashhoudy, 2004) is another important part. Excel is a powerful tool for creating almost all kinds of graphs (Hunt, 1996, 2003; Hunt and Mashhoudy, 2004). Moreover Excel is implemented to arise statistical thinking (Nash and Quon, 1996) or for understanding important ideas in statistics (Price and Zhang, 2007). Last but not least lecturers could use Excel to generate individualized tasks for students (Hunt, 2005, 2007).

This paper has a look in the problems occurring by using Excel in a basic course of statistics. Therefore it has a look on the implementation of basic statistical methods in

Excel, in particular in describing data. The topics are creating frequency distributions, calculating measures of central tendency and variability, and creating histograms. The various examples show some snares set by Excel and give some hints how to avoid embarrassments.

2 Frequency Distributions

A frequency distribution is a table used to organize data. This definition indicates that a frequency table is something very basic in the large field of statistics. So calculating a frequency table should not be too hard. But in Excel its one of the most complicated things. Usually the functions implemented in Excel are easy to use, you have just to paste the function and set the references or values. This does not work for the Excel function *Frequency*. So the next step would be to read the help about *Frequency*, where you will find the following:

> ***Frequency***
> Calculates how often values occur within a range of values, and then returns a vertical array of numbers. For example, use *Frequency* to count the number of test scores that fall within ranges of scores. Because *Frequency* returns an array, it must be entered as an array formula.

Most students are irritated by hearing formula, but what does array formula mean? Another hint from the help:

> ***Array Formulas***
> An array formula can perform multiple calculations and then return either a single result or multiple results. Array formulas act on two or more sets of values known as array arguments. Each array argument must have the same number of rows and columns. You create array formulas in the same way that you create other formulas, except you press *Ctrl* + *Shift* + *Enter* to enter the formula. Array constants can be used in place of references when you don't want to enter each constant value in a separate cell on the worksheet. Some of the built-in functions are array formulas, and must be entered as arrays to get the correct results.

This information will not be very helpful for beginners. If you want to calculate frequencies with Excel you have to do the following steps:

- Prepare a worksheet for the results (Figure 1).
- Mark the area for the results (here *B2:B5*).
- Call the function *Frequency*.
- Mark the area of the dataset.
- Mark the area for the values (here *A2:A5*).
- Press the combination *Ctrl* + *Shift* + *Enter* to get all results.

It seems to be really hard work to get a simple frequency table in Excel. Moreover the remark in the help text, that each array argument must have the same number of rows and columns, is not very helpful in this case, because it is not true.

However, apart from this the example shows even more interesting details. The Microsoft Excel help file gives some remarks in addition to an explanation and the syntax:

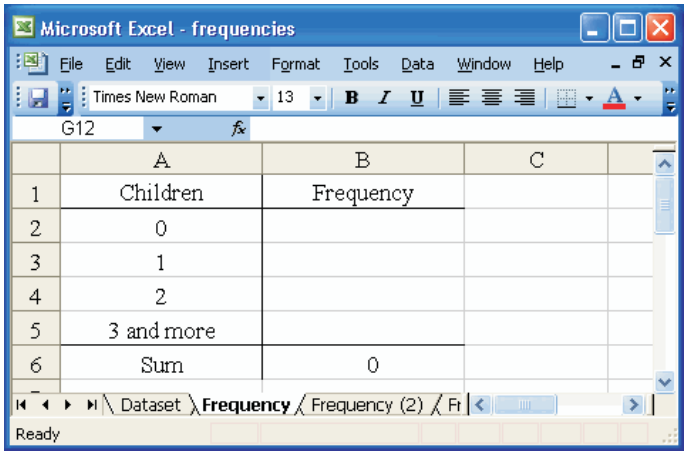


Figure 1: Prepare a table for the results

Remarks

- FREQUENCY is entered as an array formula after you select a range of adjacent cells into which you want the returned distribution to appear.
- The number of elements in the returned array is one more than the number of elements in bins_array. The extra element in the returned array returns the count of any values above the highest interval. For example, when counting three ranges of values (intervals) that are entered into three cells, be sure to enter FREQUENCY into four cells for the results. The extra cell returns the number of values in data_array that are greater than the third interval value.
- FREQUENCY ignores blank cells and text.
- Formulas that return arrays must be entered as array formulas.

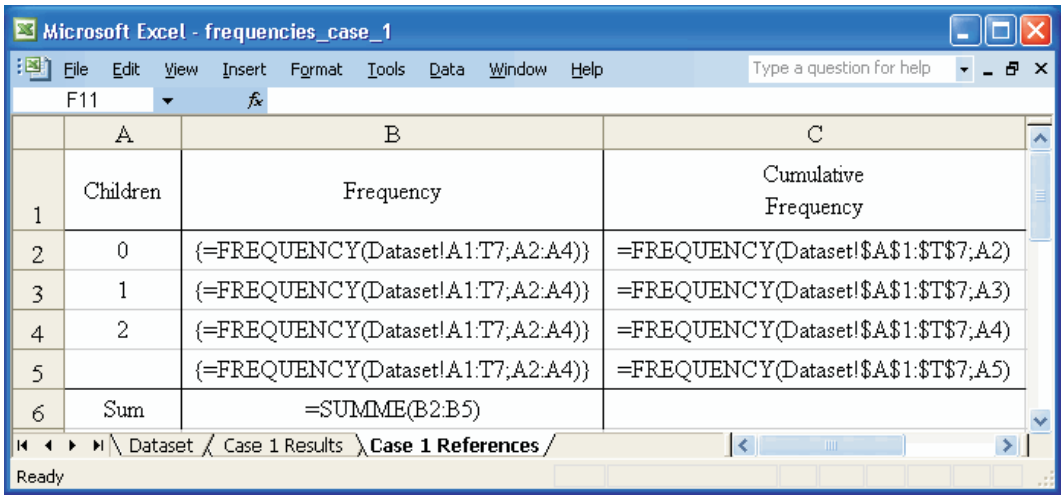
The second remark implies, that you would get the same results with setting the reference for the categories to A2:A4. So there are three different possibilities to handle this formula:

1. The number of elements in the reference for the values is one less than the number of results (Figure 2, column B).
2. The number of elements in the reference is the same as the number of results, but the last reference for the value contains no numeric value (i.e. “3 and more”).
3. The number of elements is the same, but the last reference for the value contains the maximum value of the dataset (i.e. 7).

In the first two cases the last cell of the results returns the frequency for values above the last numeric value (here above 2). In the third case the final result counts all values, which are above 2 but at least 7.

But the last case also provides another possibility to calculate frequencies: In this case you get the cumulative frequencies by using the formula not as an array formula but as a common formula instead (Figure 2, column C). This method doesn’t work for the other two cases, as is shown in Figure 4.

Another possibility to get frequencies is by using the function Countif.



Microsoft Excel - frequencies_case_1

File Edit View Insert Format Tools Data Window Help Type a question for help

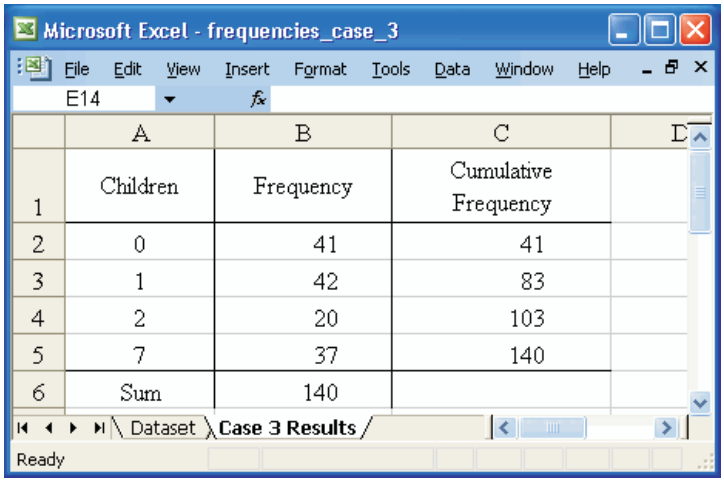
F11

	A	B	C
1	Children	Frequency	Cumulative Frequency
2	0	{=FREQUENCY(Dataset!A1:T7;A2:A4)}	=FREQUENCY(Dataset!$A$1:$T$7;A2)
3	1	{=FREQUENCY(Dataset!A1:T7;A2:A4)}	=FREQUENCY(Dataset!$A$1:$T$7;A3)
4	2	{=FREQUENCY(Dataset!A1:T7;A2:A4)}	=FREQUENCY(Dataset!$A$1:$T$7;A4)
5		{=FREQUENCY(Dataset!A1:T7;A2:A4)}	=FREQUENCY(Dataset!$A$1:$T$7;A5)
6	Sum	=SUMME(B2:B5)	

Dataset / Case 1 Results \ Case 1 References /

Ready

Figure 2: Handling of Frequency, case 1



Microsoft Excel - frequencies_case_3

File Edit View Insert Format Tools Data Window Help

E14

	A	B	C	D
1	Children	Frequency	Cumulative Frequency	
2	0	41	41	
3	1	42	83	
4	2	20	103	
5	7	37	140	
6	Sum	140		

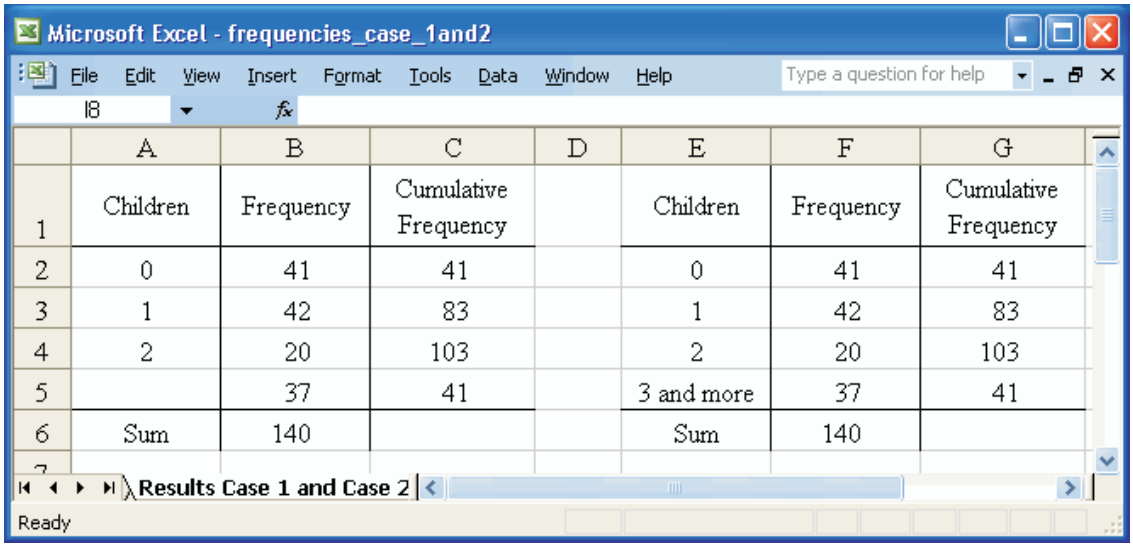
Dataset \ Case 3 Results /

Ready

Figure 3: Results of Frequency, case 3

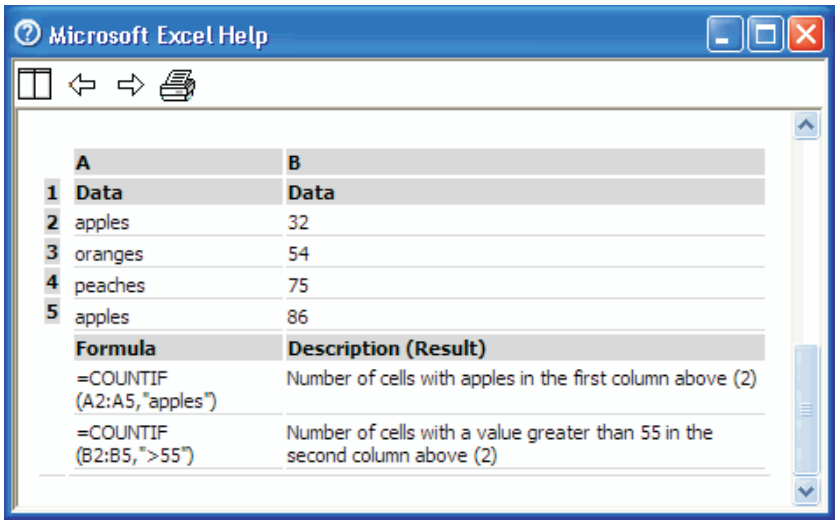
Countif
Counts the number of cells within a range that meet the given criteria.
COUNTIF(range, criteria)
Range is one or more cells to count, including numbers or names, arrays, or references that contain numbers. Blank and text values are ignored.
Criteria is the criteria in the form of a number, expression, cell reference, or text that defines which cells will be counted.

But also this function is a little bit tricky, because it is not easy to find out, how to use this formula for interval-scaled variables. For our example it is easy to calculate the frequencies for 0, 1 or 2 children, but how to create the criteria for the last category (3 and more)? The example in the helpfile (Figure 5) is not very useful for this question, because using “>A4” is impossible and using “>2” not very comfortable, since using references instead of values is the better way (especially when changing the data later on).



	A	B	C	D	E	F	G
1	Children	Frequency	Cumulative Frequency		Children	Frequency	Cumulative Frequency
2	0	41	41		0	41	41
3	1	42	83		1	42	83
4	2	20	103		2	20	103
5		37	41		3 and more	37	41
6	Sum	140			Sum	140	

Figure 4: Results of Frequency, case 1 and 2



	A	B
1	**Data**	**Data**
2	apples	32
3	oranges	54
4	peaches	75
5	apples	86

Formula	Description (Result)
=COUNTIF (A2:A5,"apples")	Number of cells with apples in the first column above (2)
=COUNTIF (B2:B5,">55")	Number of cells with a value greater than 55 in the second column above (2)

Figure 5: Example in Countif

By pure chance – and meanwhile (since Office 2007) also on the internet homepage of Microsoft – you can find the solution: Using ">&A4" as criteria works well.

3 Measures of Central Tendency and Variability

Lets have a look on other topics in elementary statistics, e.g. measures of central tendency, more precisely mean, mode, median and percentiles. The mean can be calculated with the statistical function *Average* and using this function is no problem. *But* there exists a function *Averagea* too, which also calculates the mean – with the difference, that the arguments could be numbers, text *and* logical values. For example the average of "wahr" (German word for true) and "true" is 0.5, whatever this means.

The function *Mode* returns the most common value in a data set. But how does Excel deal with the problem when there is more than one most common value. Statistical software packages usually return the smallest of all possible modes, but Excel has created an own - not very useful - way to solve this problem. Excel takes the first occurring possible mode, this leads to different results if you reorder the dataset.

Table 1: Different modes for the same, but reordered dataset

4	1
4	1
1	4
1	4
0	0
mode=4	mode=1

It is possible to avoid this problem by sorting the data before using the function *Mode*, if there is only one variable to be analyzed. In case of more than one variable the best solution is to calculate the frequencies and evaluate the mode by yourself.

Median
The median of a distribution of values is a number, such that at least 50% of the values are less than or equal to the median and at least 50% of the values are greater than or equal to the median.

Percentile
The p-percentile of a distribution of values is a number x_p, such that at least a percentage p of the values are less than or equal to x_p and at least a percentage $1-p$ of the values are greater than or equal to x_p.

This definition for a percentile is the generalization of the definition for the median. Many statistical books mention percentiles only for probability distributions, but not for samples. For arbitrary $0 < p < 1$ the following different definitions were found. We first define some necessary notation. Let $\lfloor x \rfloor$ be the integer part of x, $\{x\}$ denotes the fractional part of x, i.e. $x - \lfloor x \rfloor$, $\lceil x \rceil$ is the smallest integer $\geq x$, $r(x)$ is x rounded to the nearest integer, and finally $x_{(i)}$ represents the order statistic, where $i = 1, \ldots, n$.

In Casella and Berger (2002) the sample percentile is given by

$$x_p = \begin{cases} x_{(r(np))} & \text{for } 1/2n < p < 1/2, \\ x_{(n+1-r(n(1-p)))} & \text{for } 1/2 < p < 1 - 1/2n, \end{cases} \tag{1}$$

and for $p = 1/2$ the common sample median is used. The verbal definition for a percentile given in Casella and Berger (2002) differs a little bit from the one given above: A percentile is the observation such that *approximately* np of the observation are less than this observation.

In Hogg and Tanis (2005) the sample percentile is calculated as

$$x_p = x_{(\lfloor (n+1)p \rfloor)} + \{(n+1)p\} \left(x_{(\lceil (n+1)p \rceil)} - x_{(\lfloor (n+1)p \rfloor)} \right) . \tag{2}$$

The verbal definition for the percentile is the same as in Casella and Berger (2002), the results are sometimes different.

In Johnson (2004) the calculation of a sample percentile is made by

$$x_p = \begin{cases} (x_{(np)} + x_{(np+1)})/2 & \text{for } np \text{ integer,} \\ x_{(\lceil np \rceil)} & \text{otherwise.} \end{cases} \tag{3}$$

In Excel the sample percentile is calculated by (Hafner and Waldl, 2000)

$$x_p = x_{(\lfloor p(n-1)+1 \rfloor)} + \{p(n-1)+1\} \left(x_{(\lceil p(n-1)+1 \rceil)} - x_{(\lfloor p(n-1)+1 \rfloor)} \right) . \tag{4}$$

The (verbal) definition for a sample percentile seems to be the same as in Casella and Berger (2002), but the result differs from result (1) as well as from result (2). An explicit definition for a percentile is not given in Excel.

For the dataset

Value	Frequency
0	32
1	43
2	46
3	59
4	70
Sum	250

the 30th-percentile is given by

$$x_{0.3} = \begin{cases} 1 & \text{using (1),} \\ 1.3 & \text{using (2),} \\ 1.5 & \text{using (3),} \\ 1.7 & \text{using (4).} \end{cases}$$

The question arises, how to teach (different) results for the same question to newcomers in statistics, particularly with regard to different results offered by different statistical software packages (Hyndman and Fan, 1996; Langford, 2006).

Measures of Variability

Functions for calculating or estimating the variance in Excel have the possibility to do this including numbers, text and logical values. Again this does not make sense. Moreover the algorithm for computing or estimating the variance was very poor until version 2002, e.g. calculating the variance for a dataset insisting of 10 objects, each with the same value 123456789 led to a variance of 5.12. This bug was corrected in Excel 2003, now the variance is calculated correctly for the mentioned dataset. Nevertheless, there are other data (consisting of more objects with larger values) with wrong results for the variance. (The variance is calculated correctly in SPSS 12.0 or in R). This is only one point out

of many about the accuracy of statistical procedures in Excel (Heiser, 2006; Keeling and Pavur, 2007; Knüsel, 1998, 2005; McCullough and Wilson, 1999, 2002, 2005).

A new highlight is offered by Excel 2007, which has a surprising calculation for multiplications with result 65535: In some versions the multiplication "77.1*850" (or other multiplications with result 65535) will give 100000 as result! An update with an Excel 2007 Hotfix package (Beard, 2007; Microsoft, 2007) will solve this problem.

4 Graphs

Good graphs should illustrate numerical information without distortion (Cleveland, 1994; Oliver, 1998). Excel offers an impressive variety of graphs, but creating a histogram for distributions with unequal group intervals or creating a c.d.f. for discrete distributions is not very common in Excel.

Histogram
(*http://www.stats.gla.ac.uk/steps/glossary/presenting_data.html#hist*, 2008-01-28)
A histogram is a way of summarizing data that are measured on an interval scale. It is often used in exploratory data analysis to illustrate the major features of the distribution of the data in a convenient form. It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles might be drawn of non-uniform height.

Statistical software usually offer histograms, but not for the special case of unequal intervals. However sometimes there is the need for that special kind, because of given definitions for interesting variables, i.e. the SME-Definition of the European Commission (2003) for medium-sized (headcount < 250), small (headcount < 50) and micro (headcount < 10) enterprises. For illustration the difference between incorrect (but easy to create) and correct graphical information, the following example is chosen. It contains a (relative) frequency table for body heights of 65 students using intervals with unequal widths.

Table 2: Body Height

Interval	Body Height	Rel. Frequency	Width	Density
1	$155 < x \leq 160$	0.062	5	0.012
2	$160 < x \leq 165$	0.246	5	0.049
3	$165 < x \leq 170$	0.215	5	0.043
4	$170 < x \leq 175$	0.200	5	0.040
5	$175 < x \leq 185$	0.215	10	0.022
6	$185 < x \leq 195$	0.062	10	0.006

The benefit of Excel is the possibility to create a histogram with unequal intervals, the drawback is that you need some experience to do so (Hunt, 1996, 2003). Comparing

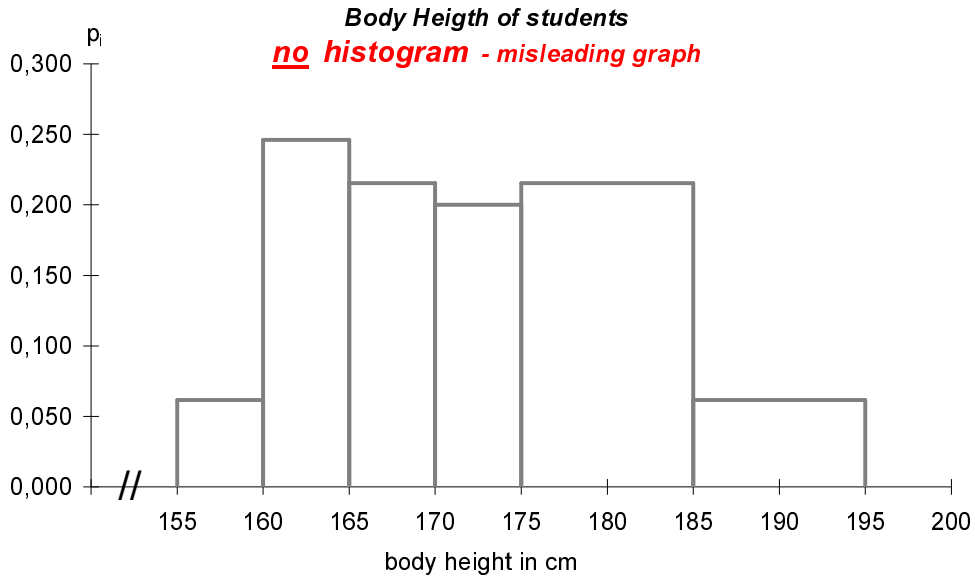


Figure 6: Misleading graph

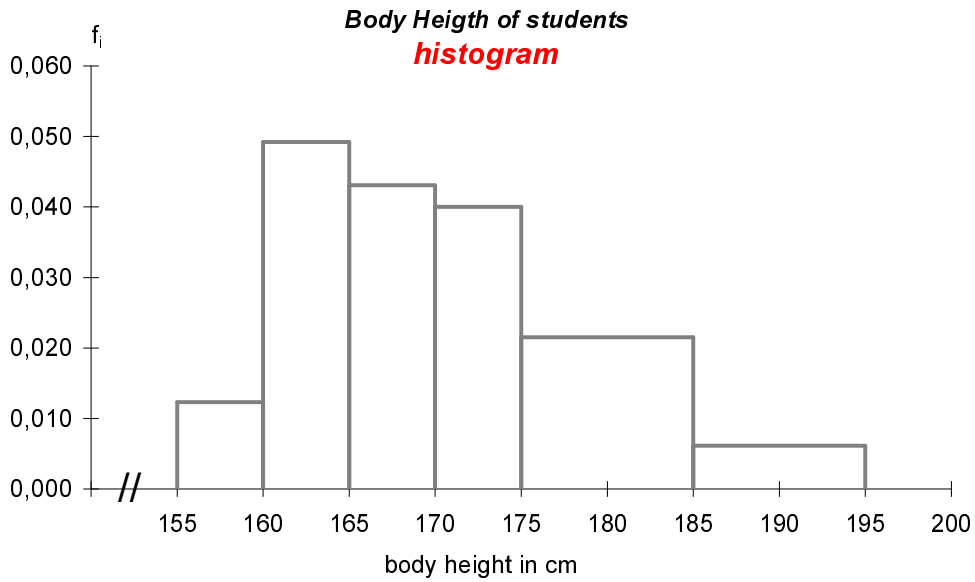


Figure 7: Correct histogram

the two graphs in Figure 6 and 7, the necessity of a histogram can easily be seen. The first graph (Figure 6) is misleading, because it seems that there are twice as much people between 175cm and 185cm as between 165cm and 170cm. The histogram in Figure 7 shows the proportion correctly.

To create a histogram with Excel, you have to organize the data in a very special way (Table 3), because each point of each rectangle has to be an entry in the data array. Now the histogram can be created with the Chart Wizard of Excel, choosing *XY (Scatter)* with option *lines only*. The same chart is useful for creating cumulative distribution functions.

Table 3: Data for histogram

Body height	Density f_i
155	0.0000
155	0.0123
160	0.0123
160	0.0000
160	0.0492
165	0.0492
165	0.0000
165	0.0431
⋮	⋮
190	0.0062
190	0.0000
190	0.0062
195	0.0062
195	0.0000

Teaching histograms with unequal widths of intervals is a valuable preliminary work for continuous probability densities: Histograms show relative frequencies as (rectangular) areas, the area below a continuous probability density shows the probability.

5 Summary

This paper had a look on the implementation of basic statistical methods in Excel, in particular for describing data. Excel is not produced to analyze data in a statistical sense, but to a certain extent it can be utilized for it. The implemented functions are sometimes a little bit tricky, some functions are not very meaningful and – last but not least – users should keep their eye on accuracy. On the other hand Excel is a well known and frequently used software, so the best way to deal with it is to teach statistics with Excel with a critical point of view.

Teaching statistics was a challenge in the past and will be one in the future. In the past the challenge was often the calculation itself. Now the calculation is done by the computer and the today challenge is to work with the various software packages in the right way, understanding the results and staying distrustful to the results.

References

Bartz, S. (2007). Excelblatt vereinfacht Stochastik. *Stochastik in der Schule*, *27*(2), 25-29.

Beard, J. (2007). *LAWTECH GURU BLOG.* http://www.lawtechguru.com/archives/2007/09/25_excel_2007_multiplication_math_bug.html, 2008-01-28.

Borovcnik, M. (2007). Das Sammelbildproblem - Rosinen und Semmeln und Verwandtes: Eine rekursive Lösung mit Irrfahren. *Stochastik in der Schule*, *27*(2), 19-24.

Casella, G., and Berger, R. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, California: Duxbury.

Cleveland, W. (1994). *The Elements of Graphing Data*. Summit, New Jersey: Hobart Press.

Doane, D. (2004). Using Simulation to Teach Distributions. *Journal of Statistics Education*, *12*(1).

European Commission. (2003). *Commission Recommendation of 6 May 2003 concerning the definition of small and medium-sized enterprises.* http://europa.eu/eur-lex/pri/en/oj/dat/2003/l_124/l_12420030520en00360041.pdf, 2008-01-28.

Hafner, R., and Waldl, H. (2000). *Statistik für Sozial- und Wirtschaftswissenschaftler, Band 2*. Wien: Springer.

Heiser, D. (2006). Microsoft Excel 2000 and 2003 faults, problems, workarounds and fixes. *Computational Statistics and Data Analysis*, *51*, 1442-1443.

Hogg, R., and Tanis, E. (2005). *Probability and Statistical Inference* (Seventh ed.). New York: Macmillan.

Horgan, G. (1999). Use of Spreadsheets for Demonstrating Experimental Power and Variability. *Journal of Statistics Education*, *7*(1).

http://www.stats.gla.ac.uk/steps/glossary/presenting_data.html#hist. (2008-01-28).

Hunt, N. (1996). Spreadsheet Histograms. *Teaching Statistics*, *18*(1), 10-11.

Hunt, N. (2003). Handling Continuous Data in Excel. *Teaching Statistics*, *25*(2), 42-45.

Hunt, N. (2005). Using Microsoft Office to Generate Individualized Tasks for Students. *Teaching Statistics*, *27*(2), 45-48.

Hunt, N. (2007). Individualized Statistics Coursework Using Spreadsheets. *Teaching Statistics*, *29*(2), 38-43.

Hunt, N., and Mashhoudy, H. (2004). Charts in Excel – A Series Matter. *Teaching Statistics*, *26*(2), 49-53.

Hyndman, R., and Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*(4), 361-365.

Johnson, R. (2004). *Miller and Freund's Probability and Statistics for Engineers*. Upper Saddle River, New Jersey: Prentice Hall.

Keeling, K., and Pavur, R. (2007). A comparative study of the reliability of nine statistical software packages. *Computational Statistics and Data Analysis*, *51*, 3811-3831.

Knüsel, L. (1998). On the Accuracy of Statistical Distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis*, *26*, 375-377.

Knüsel, L. (2005). On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, *48*, 445-449.

Kühleitner, M. (2007). Sammelbildproblem: Eine Simulation mit Excel. *Stochastik in der Schule*, *27*(1), 24-26.

Langford, E. (2006). Quartiles in Elementary Statisitcs. *Journal of Statistics Education*, *14*.

Martin, C. (2007). A Simulation Based on Goldratt's Matchstick/Die Game. *Decision Sciences Journal of Innovative Education*, *5*(2), 423-429.

McCullough, B., and Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis*, *31*, 27-37.

McCullough, B., and Wilson, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and XP. *Computational Statistics and Data Analysis*, *40*, 713-721.

McCullough, B., and Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, *49*, 1244-1252.

Microsoft. (2007). http://support.microsoft.com/kb/943075.

Mills, J. (2002). Computer Simulation Methods to teach Statistics: A Review of the Literature. *Journal of Statistics Education*, *10*(1).

Nash, J., and Quon, T. (1996). Issues in Teaching Statistical Thinking with Spreadsheets. *Journal of Statistics Education*, *4*(1).

Oliver, F. (1998). *How to present information in graphs and diagrams.* Notes on behalf of the Examinations Board of the Royal Statistical Society, http://www.therss.org.uk/exams/docs/diagrams.pdf, 2008-01-28.

Price, B., and Zhang, X. (2007). The Power of Doing: A Learning Exercise that Brings the Central Limit Theorem to Life. *Decision Sciences Journal of Innovative Education*, *5*(2), 405-411.

Authors' Address:

Christine Duller
Institut für Angewandte Statistik
Johannes Kepler Universität Linz
Altenberger Str. 69
A-4040 Linz

Tel. +43 732 2468-9128
Fax +43 732 2468-9846
E-mail: christine.duller@jku.at
Homepage: http://www.ifas.jku.at/