Elaborated Paper on the Use of LFS Micro Data

Present paper in brief In conjunction with preparing the paper on *Guidelines for external data users: Guidelines for publishing LFS data* the need emerged for an elaborated description of the background to the guidelines set out. This paper gives an outline of how to use micro data in the most expedient way – the paper is thus aimed at specialized users.

Guidelines set out First, the guidelines set out and their background is discussed. Consequently, quarterly and annual estimates are outlined as well as the method of weighting focussing on the two elements: design weight and correction for non-response.

The guidelines are described in *Guidelines for publishing LFS data*:

	Result	Instruction
Quarterly figures	Under 4,000 persons (weighted figures)	Figures are not published
	4,000-7,000 persons (weighted figures)	Figures can be published with the caution that figures are subject to some uncertainty*
Annual figures	Under 2,000 persons (weighted figures)	Figures are not published
	2,000 – 4,000 persons (weighted figures)	Figures can be published with the caution that figures are subject to some uncertainty*

* Figures in the group that can be published subject to caution should only be published in special cases. Statistics Denmark recommends that this is discussed before figures are published with employees from the Labour Force Survey.
.

There are, to a very large extent, practical considerations related to the specific break downs. Subsequently, the lower limit for the quarterly figures was originally 3,500, but this did not harmonize well with the instruction to publish only figures in thousands. The variability for the weighted estimates varies, implying that two figures of 4,000 are not necessarily subject to an equal certainty.

The threshold limits are determined by the number of interviews on which they are based – the greater the number of interviews, the greater the certainty. As the number of interviews is higher for annual figures, the weighted threshold limit value is lower. The estimates are based on persons employed in different occupational categories, as the statistical uncertainty is here typically relatively high.

By means of figures 1 and 2 in the annex, attempts have been made to illustrate the reasons for choosing the selected threshold limits

Quarterly estimates First, we take a look at the quarterly estimates. Generally speaking, the estimates indicate that the coefficient of variation reaches 20 pct. for a size estimation of 7,000, while they exceed 25 pct. for a size estimation of 4,000. With 95-percent certainty, a result of 7,000 ranges between 4,200 and 9,800, whereas a result of 4,000 ranges between 2,000 and 6,000. Depending on the size estimation, the results can potentially vary between, respectively, factor 2.3 (for 7,000) and factor 3 (for 4,000).

Annual estimates For the annual estimates, the variability is slightly lower and this is also the case for the threshold limits used for publication. For the same reason, these results can better be used for purposes of compiling structural statistics, although there is, of course, still sampling errors. The results reaching about 4,000 have a coefficient of variation around 12 pct., while the coefficient of variation for results around 2,000 reaches 18 pct. With 95-percent certainty, estimates of 4,000 range between 3,000 and 5,000, while estimates of 2,000 range between 1,200 and 2,800. For the upper as well as the lower threshold limits, the certainty of the figures is considerably higher than is the case for the quarterly estimates, which is, of course, also the purpose of the annual data.

Design weight   Each interview in the LFS has, on average, a final value of 200 for quarterly data and a final value of 50 for annual data. The weight of each interview is primarily formed by the design weight, which depends on the size of each individual group drawn for the survey. For the quarterly dataset from the 4th quarter 2011, the design weights vary between 108 and 600, the average was 199, while the median was 220. A quarterly weighted estimate of 7,000 thus comprises (disregarding the correction for non-response) between 12 and 65 interviews.

Statistics Denmark's Office for Survey and Method recommends that statistical statements should be based on at least 20-25 observations. It is also very rarely the case that 7,000 weighted persons are based on less than 20 observations – consequently, 7,000 is a very appropriate rule of thumb.
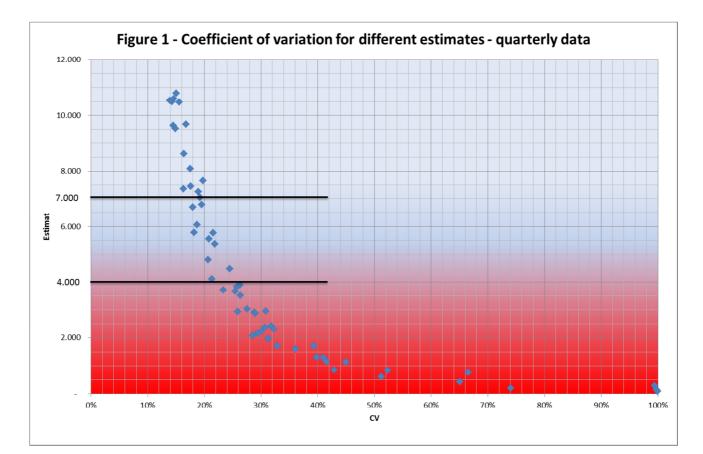
Conversely, an estimate of 4,000 can be based on between 7 and 37 observations. This will be the case where estimates under 4,000 contain an appropriate number of observations, implying that the variability of the result should thus be sufficiently high for publication. Here, expert users with access to micro data can test the number of observations of a specific estimate for the purpose of getting an indication of the statistical variability.

Non-respons   We have, so far, only looked at the clear-cut sampling errors. But there may also be a problem with regard to the bias as a result of non-response. Generally slightly more than every second person in the sample participates in the survey. However, the non-response can fluctuate considerably between different groups in society. The non-response can be analysed by looking at the second element in the weight of figures – the calibration weight or the correction for non-response. This weight of correction depends partly on the model of weighting, partly by the specific non-response for different groups of respondents. It has an average of 1, but varies from between 0.1 to 4 and illustrates, e.g. the scale of the non-response in a specific group of respondents in relation to the average non-response among all groups of respondents.

The table below shows an example of 3 estimates, which are apparently very similar to each other, whether weighted or unweighted, and when we look at the coefficient of variation. However, this picture differs when the correction for non-response is taken into consideration. It can be seen that the bottom group has an estimated non-response, which is 11 pct. higher than the average non-response in the survey – this indicates that there is a higher risk of biased responses. The top group has an estimated non-response, which is 11 pct. lower than the average, and the risk of biased responses is thus lower for this group.

3-digit ISCO	Weighted estimate	Unweighted estimate	CV	Correction for non-response
Management of the main activity in manufacturing, mining and quarrying, construction, supply, distribution, etc.	10,539	56	14%	-11%
Activities in the domain of food, beverages and tobacco	10,493	56	14%	0%
Activities in the domain of vegetation	10,796	52	15%	11%

It is particularly important to keep an eye on the correction for non-response, when we look at small groups, which are very close to the threshold limits used for publication, in order to examine the risk of a systematic bias in the existing responses. One example illustrating the problem involved in the correction for non-response can be seen in the group of immigrants. Here, the correction for non-response for Danish citizens is -5.3 pct., i.e. they participate to a slightly higher degree than the average, while the correction for non-response for non-western foreign citizens is 46.2 pct. The risk of systematically biased results is thereby considerably higher than for non-western foreign citizens than for Danish citizens. This is important to keep in mind

when the robustness of one's results is considered. If you have any questions as to the robustness, you are always welcome to contact one of Statistics Denmark's employees responsible for with the Labour Force Survey.

Annex: Relationship between the estimated size and coefficient of variation for the different occupational categories

Figure 1 - Coefficient of variation for different estimates - quarterly data

Figure 2 - Coefficient of variation for different estimates - annual data