# Dissemination of wage data from the Spanish LFS: anonymized microdata files

Miguel Ángel García Martínez. INE-SPAIN

## Background

Unlike what happens in other countries of the European Union, the Spanish LFS doesn't collect any quantitative information on income in the survey. Only qualitative information about receiving unemployment subsidies or allowances (to provide the LFS variable REGISTER), or the self-assessment of the respondent as 'retired' (to provide LFS variable MAINSTAT) are collected.

The questions on income administered through personal or telephone interview are always problematic. Different tests carried out in Spain including questions on income within the LFS showed important problems of validity of the information obtained and confirmed the sensitivity of this kind of information.

On the other hand, the analytical power of labour income in the household/population surveys, and particularly in LFS, is out of discussion and consequently such information was already required in the Council Regulation 577/1998 information as part of the variables to be provided, although in a voluntary basis.

The Regulation (EC) Nº 1372/2007 of the European Parliament and of the Council stated as compulsory in the LFS the information on 'wages from the main job'. The variable can be obtained from administrative sources and, in that case, the transmission of the data will be carried out within twenty one months of the end of the reference month. The Commission Regulation (EC) Nº 377/2008 describe the variable INDECIL as annual and it will be code in deciles.

## Main problems associated to interview on income in population/household surveys

The problems related to the capture of income information from computed assisted personal or telephonic interview (CAPI/CATI interview) are diverse.

On one hand we have the intrinsic difficulty of the concept itself (gross vs. net wages, the components to be included, the lapse of time to be considered in the remuneration, how to deal with 'small' contracts, etc.) that are conditioned by the specific objectives of the survey. In our case (LFS) the wages variable is claimed to be considered as a classification variable and focused on monthly remuneration. Currently, the deciles should be calculated taking into account the 'take-home' pay but other consistent measures are allowed. In the next future there seems that the gross definition will be preferred.

Other factor that influence the quality of the information on wages directly requested by questions in household's surveys is the possibility of 'proxy' respondents who can answer

instead of the person directly concerned. This issue can be especially relevant in the case of income from self-employment.

Finally, there is an underlying problem of sensitiveness of the information requested that is frequently present in the interview and produces an underestimation bias of income (very often observed in many countries, and well-known in income surveys in Spain). Qualitative studies show the risk that this 'sensitivity' factor of the income variables is transmitted to other variables of the survey changing the 'cognitive framework' of the interview and conditioning the response differently (for the worse) than when we just ask for the labour market situation.

All these issues made us to decided focusing the efforts on obtaining the information from the available administrative sources (namely, social security records and income-tax registers)

## Accessing to the administrative sources to obtain the wage deciles

The European legal basis was crucial for accessing to the income-tax sources as our national law excluded specifically this administrative sources from the access for statistical purposes (that is in fact granted in general for other administrative sources).

Thus, the Regulation (CE) 1372/2007 of the European Parliament and the Council stating as compulsory the information on wages of main job and allowing a deadline of twenty-one months after the reference period for deliver the data, and the Regulation (CE) 377/2008 of the European Commission coding the variable INCDECL as deciles of the monthly wages of the main job, determined the way to proceed in the methodology to produce the information in Spain.

More concrete aspects of the methodology can be consulted in the INE web site or in previous LFS Workshop papers. In this one we are focusing on the evolution of the dissemination of the statistical information of the wages from the LFS in Spain since we could obtain the information under the European Regulations.

## Initial dissemination of data on deciles (November 2010 publishing data for 2006-2009)

The first dissemination of LFS-wage information was in November 2010, publishing the whole series of data available at that moment (2006-2009). The tables were designed to provide the number of employees crossing by deciles and a set of variables considered to have a significant influence in wage levels. The full-time/part-time distinction was included in all the tables in order to 'separate' the effect of this crucial variable from the others.

The following variables were selected to cross tabulate deciles and full-time/part-time distinction: sex, age group (five ten-years groups), citizenship, region, educational attainment (seven groups), field of education (wide field), occupation (1-digit ISCO level), activity (A21 NACE), type of contract (permanent vs. Temporal), people working in the workplace (five groups), supervisory responsibilities, time working in the company, public of private employee distinction and underemployment situation (national definition)

An example of this kind of tables is shown as table 1 (cross tabulation by sex).

**Table 1. Employees by full-time/part-time distinction, sex and decile.**

**Wage deciles**

**Wage decile of main job**

**Employees by full-time/part-time distinction, sex and decile. Number of employees and percentages by sex.**

**Unit: Thousands of persons**

| | 2015 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Number of employees** | | | | | | | | | | | |
| **Total** | | | | | | | | | | | |
| **Total** | 14.756,7 | 1.474,8 | 1.476,1 | 1.476,0 | 1.475,5 | 1.475,9 | 1.475,5 | 1.475,5 | 1.475,8 | 1.475,8 | 1.475,8 |
| **Males** | 7.704,2 | 378,3 | 519,2 | 628,9 | 759,8 | 867,0 | 901,0 | 907,6 | 873,8 | 897,1 | 971,4 |
| **Females** | 7.052,5 | 1.096,6 | 956,9 | 847,1 | 715,7 | 608,9 | 574,5 | 567,9 | 602,0 | 578,7 | 504,4 |
| **Full-time** | | | | | | | | | | | |
| **Total** | 12.220,3 | 94,0 | 819,7 | 1.256,7 | 1.342,0 | 1.423,7 | 1.439,9 | 1.443,3 | 1.457,2 | 1.470,5 | 1.473,4 |
| **Males** | 7.057,7 | 47,0 | 351,5 | 572,9 | 708,8 | 855,9 | 890,1 | 897,6 | 867,9 | 894,6 | 971,4 |
| **Females** | 5.162,6 | 47,1 | 468,2 | 683,8 | 633,1 | 567,8 | 549,8 | 545,7 | 589,3 | 575,9 | 502,0 |
| **Part-time** | | | | | | | | | | | |
| **Total** | 2.536,4 | 1.380,8 | 656,4 | 219,3 | 133,6 | 52,2 | 35,6 | 32,2 | 18,6 | 5,3 | 2,4 |
| **Males** | 646,5 | 331,3 | 167,8 | 56,0 | 51,0 | 11,1 | 10,9 | 10,0 | 5,9 | 2,5 | .. |
| **Females** | 1.889,9 | 1.049,5 | 488,7 | 163,3 | 82,6 | 41,1 | 24,7 | 22,2 | 12,7 | 2,8 | 2,4 |
| **Percentage** | | | | | | | | | | | |
| **Total** | | | | | | | | | | | |
| **Total** | 100,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 | 10,0 |
| **Males** | 100,0 | 4,9 | 6,7 | 8,2 | 9,9 | 11,3 | 11,7 | 11,8 | 11,3 | 11,6 | 12,6 |
| **Females** | 100,0 | 15,5 | 13,6 | 12,0 | 10,1 | 8,6 | 8,1 | 8,1 | 8,5 | 8,2 | 7,2 |
| **Full-time** | | | | | | | | | | | |
| **Total** | 100,0 | 0,8 | 6,7 | 10,3 | 11,0 | 11,7 | 11,8 | 11,8 | 11,9 | 12,0 | 12,1 |
| **Males** | 100,0 | 0,7 | 5,0 | 8,1 | 10,0 | 12,1 | 12,6 | 12,7 | 12,3 | 12,7 | 13,8 |
| **Females** | 100,0 | 0,9 | 9,1 | 13,2 | 12,3 | 11,0 | 10,6 | 10,6 | 11,4 | 11,2 | 9,7 |
| **Part-time** | | | | | | | | | | | |
| **Total** | 100,0 | 54,4 | 25,9 | 8,6 | 5,3 | 2,1 | 1,4 | 1,3 | 0,7 | 0,2 | 0,1 |
| **Males** | 100,0 | 51,2 | 25,9 | 8,7 | 7,9 | 1,7 | 1,7 | 1,5 | 0,9 | 0,4 | .. |
| **Females** | 100,0 | 55,5 | 25,9 | 8,6 | 4,4 | 2,2 | 1,3 | 1,2 | 0,7 | 0,1 | 0,1 |

**Source: LFS-Spain**

Instituto Nacional de Estadística

To complement the information about the distribution of wages, we published one table including wage data on the lower limits and averages by deciles. The idea was that this information, in combination to the other tables on distribution by deciles of the employees were used to make approximate estimates of average wages by the above-mentioned variables.

**Table 2. Average wage and lower limit by decil**

**Average wages**

**Gross monthly average wage of main job**

**Average wage and lower limit by decil**

**Unit: Euros**

| | Gross monthly average wage of main job | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2015 | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Lower limit | .. | 680,00 | 979,52 | 1.215,71 | 1.402,85 | 1.596,79 | 1.814,04 | 2.136,72 | 2.607,24 | 3.424,75 |
| Average | 420,05 | 828,96 | 1.102,14 | 1.310,96 | 1.497,67 | 1.703,73 | 1.960,42 | 2.363,55 | 2.964,23 | 4.784,50 |

**Source: LFS-Spain**

Instituto Nacional de Estadística

## Extension of the statistical information to average wages by decile (November 2014 publishing data 2006-2013)

Even users welcomed this new information about decile distribution of employees, they still demanded additional results on average wages.

After consulting the agencies involved in the delivery of the basic wage data to produce the deciles, we reached an agreement to publish average wages additionally to the information on distribution of employees by decile of each category. These results were published in November 2014 for the whole series 2006-2014.

As an example of the kind of data added, the cross tabulation by sex is provided in table 3.

**Table 3. Average wage by full-time/part-time distinction, sex and decile.**

**Average wages**

**Gross monthly average wages of the main job**

**Average wages by full-time/part-time distinction, sex and decile**

**Unit: Euros**

| | 2015 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Total** | | | | | | | | | | | |
| Total | 1.893,70 | 420,05 | 828,96 | 1.102,14 | 1.310,96 | 1.497,67 | 1.703,73 | 1.960,42 | 2.363,55 | 2.964,23 | 4.784,50 |
| Males | 2.122,47 | 454,86 | 833,73 | 1.101,65 | 1.312,87 | 1.497,04 | 1.705,75 | 1.961,57 | 2.360,62 | 2.963,63 | 4.858,79 |
| Females | 1.643,79 | 408,04 | 826,37 | 1.102,51 | 1.308,93 | 1.498,57 | 1.700,57 | 1.958,58 | 2.367,81 | 2.965,15 | 4.641,41 |
| **Full-time** | | | | | | | | | | | |
| Total | 2.142,03 | 615,79 | 850,46 | 1.103,99 | 1.312,19 | 1.497,28 | 1.704,22 | 1.961,12 | 2.364,19 | 2.964,56 | 4.786,61 |
| Males | 2.248,91 | 616,29 | 849,69 | 1.104,03 | 1.314,08 | 1.496,88 | 1.706,01 | 1.961,70 | 2.360,87 | 2.963,46 | 4.858,79 |
| Females | 1.995,93 | 615,29 | 851,05 | 1.103,95 | 1.310,06 | 1.497,87 | 1.701,31 | 1.960,17 | 2.369,08 | 2.966,27 | 4.646,94 |
| **Part-time** | | | | | | | | | | | |
| Total | 697,24 | 406,72 | 802,10 | 1.091,56 | 1.298,61 | 1.508,55 | 1.684,22 | 1.928,74 | 2.313,37 | 2.871,55 | 3.487,48 |
| Males | 742,16 | 431,98 | 800,29 | 1.077,30 | 1.295,97 | 1.509,48 | 1.684,87 | 1.949,70 | 2.323,48 | 3.022,91 | .. |
| Females | 681,87 | 398,74 | 802,73 | 1.096,45 | 1.300,24 | 1.508,30 | 1.683,93 | 1.919,29 | 2.308,68 | 2.732,65 | 3.487,48 |

**Source: LFS-Spain**

Instituto Nacional de Estadística

## The anonymized microdata file on wages from LFS in Spain (October 2016 for the period 2006-2015)

Users acknowledged the efforts for providing LFS data based on wages, but they still claimed for having microdata including individual salaries to analyze in depth the conformation of salaries.

Once again, after consulting to the suppliers agencies of the basic administrative data, we were allowed to build up a anonymized microdata file of wages from LFS whenever the confidentiality were respected (business as usual), but particularly concerning the higher salaries which were considered as more prone to potential disclosure.

To include the 'exact' monthly wage in a microdata file really increases radically the risk of identification comparing to categorical variables (e.g. the decile) in combination with other variables. To balance this risk, these other variables had to be aggregated of even suppressed.

For this reason, the approach followed to create the file was, first, to select just the variables that were judged as highly relevant when it comes to influence the wage, and even for these variables, to aggregate categories. Secondly, to ensure that the file is not linkable to other anonymized LFS files. In other words, to make the file 'unique'[1].

Concerning the 'scope' of the registers to be included in the anonymized microdata file of wages from LFS, we opted for include the whole registers surveyed in the sample, also including other persons in employment, unemployed and out of the labour force and those aged less than 16. This allows to incorporate the 'household composition' dimension in the wage analysis and to look the number of wage earners and other income receivers (qualitative information on retirement or other pensions and unemployment allowances or subsidies).

The characteristics of the microdata file and the grouping of variables included in it are the following:

1. It is a national file. No information about NUTS is included. The only 'territorial' information is a distinction based directly from the stratification of the Spanish LFS in three categories according to the size (inhabitants) of the municipality of residence: less than 10.000 inhabitants, 10.000-99.999, and 100.000 or more or being a municipality of special relevance within the province (NUTS3)
2. Age is grouped in five years intervals (65+ the last one)
3. Activity (NACE) and occupation (ISCO) variables are aggregated into ten significant groups each one. This seems pretty hard, but as an advantage, it makes almost straightforward to compare even when the NACE/ISCO classification changes and helps to prevent the 'over-use' of a rather small sample of employees.
4. Usual hours are provided in ten hours intervals
5. Educational attainment is grouped in seven categories

[1] In contrast, the decile variable is linked to the anonymized subsample variables in the Spanish LFS. Other example of 'linkable' anonymized microdata files in Spain is the 'flow' files that permit to link the common part of the sample among the quarters that the household remain in the sample.

6. Field of the educational attainment is limited to the first level
7. Size of the firm is presented in five specific groups plus the different 'don't know' categories.
8. Time working in the enterprise, as well as time since last worked, are grouped in six categories
9. All 'atypical work' variables are summarized in just one synthetic and dichotomous variable
10. A dichotomous variable on receiver of income is also included.
11. Self-assessed labor status is included (MAINSTAT)
12. And, of course, the 'exact' monthly wage for those earning less than 5.000 euros and the average wage for the corresponding sex for those earning 5.000 or more.

The graphs included in ANNEX I illustrate the file content. All the data refer to 2015, the last reference period available by now (data published in November 2016)

## Coherence issues

One drawback of grouping the wages of 5.000 euros or higher in the same average value is that the estimates, in general, are distorted whenever we try to segment by other variables. In the graphs of Annex I, the employees earning 5,000 euros or more per month are clearly identified grouped in the right (graph 1) and upper part (graph 2-5).

Thus, for example, comparing to published results, the estimates calculated from the anonymized microdata file of wages LFS in Spain will be the same only if the whole group (monthly earners >= 5,000) is included in just one of the categories involved. To illustrate this fact, see table 4.

The full time/part time distinction is fully equal as all persons earning 5,000 euros or more are all full time employees and the average has been segmented by sex, by definition.

On the other hand, when we classified by any other variable, the results differ. The example in the table is provided using the distinction between permanent or temporal job.

Other issue related with coherence between this file and other LFS data (but shared with other subsample variables) is the fact that the main collective of analysis are the employees and the total estimate for this group is not coherent to the annual averages of the four quarter in the Spanish LFS. Remember that INCDECIL is an annual variable, and the coherence criteria for the subsample file don't deal with professional status.

**Table 4. Comparison between published data and data calculated from anonymized microdata files**

| Example of coincidences | | | |
|---|---|---|---|
| **Average wages** | | | |
| **Average gross monthly wages from main job** | | | |
| **Average wages by full time / part time distinction and sex** | | | |
| **Unit: Euros** | | | |
| | **Published data** | | **Results from anonimised microdata** |
| | **2015** | | **2015** |
| **Total** | | | **Total** |
| **Total** | 1.893,70 | | 1.893,70 |
| **Males** | 2.122,47 | | 2.122,47 |
| **Females** | 1.643,79 | | 1.643,79 |
| **Full time** | | | **Full time** |
| **Total** | 2.142,03 | | 2.142,03 |
| **Males** | 2.248,91 | | 2.248,91 |
| **Females** | 1.995,93 | | 1.995,93 |
| **Part time** | | | **Part time** |
| **Total** | 697,24 | | 697,24 |
| **Males** | 742,16 | | 742,16 |
| **Females** | 681,87 | | 681,87 |

| Example of differences | | | |
|---|---|---|---|
| **Average wages** | | | |
| **Average gross monthly wages from main job** | | | |
| **Average wages by duration of the contract** | | | |
| **Unit: Euros** | | | |
| | **Published data** | | **Results from anonimised microdata** |
| | **2015** | | **2015** |
| **Total** | | | **Total** |
| **Total** | 1.893,70 | | 1.893,70 |
| **Permanent** | 2.090,16 | | 2.089,37 |
| **Temporal** | 1.314,51 | | 1.316,83 |

## Summary remarks

The legal basis provided by EU-Regulations allowed us to develop a suitable system of exploitation of administrative wage data. Without this legal basis, the procedures would have been limited to only part of the relevant sources (Social Security records, in the case).

The added wage information in the LFS was really welcomed by our national users who didn't stop pushing for more detailed information.

The anonymized microdata of wages LFS in Spain allow a very approximate reproduction of published results. The differences are provoked by the aggregation of the wages >=5.000 in just the average for each sex. This is a tradeoff between analytical power and the special risk of disclosure of the higher wages.

ANNEX I

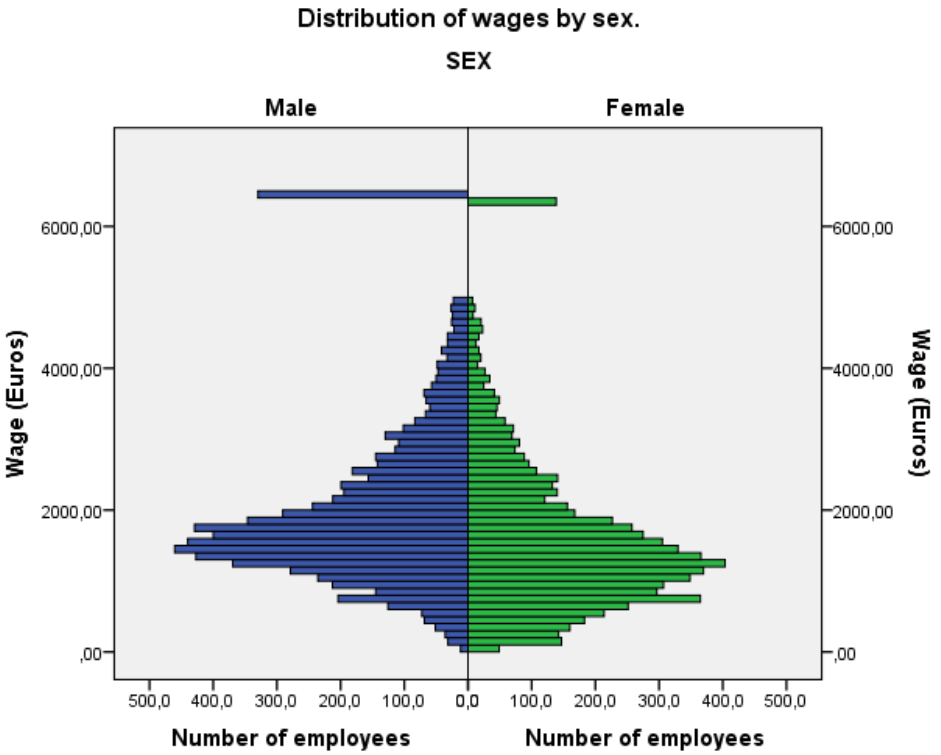
Selected graphs to illustrate the content of the anonymized microdata file of wages from LFS-Spain.

The reference period is 2015. Last available data by May 2017.

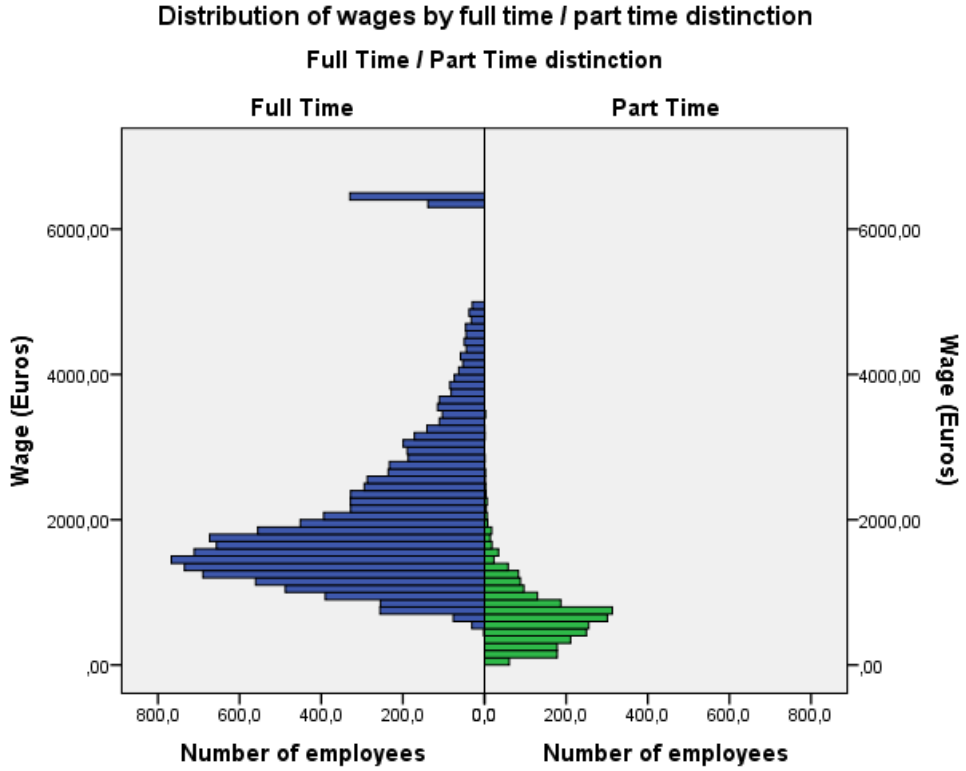
The graphs have been produced using SPSS version 22 software.

**Graph 1**

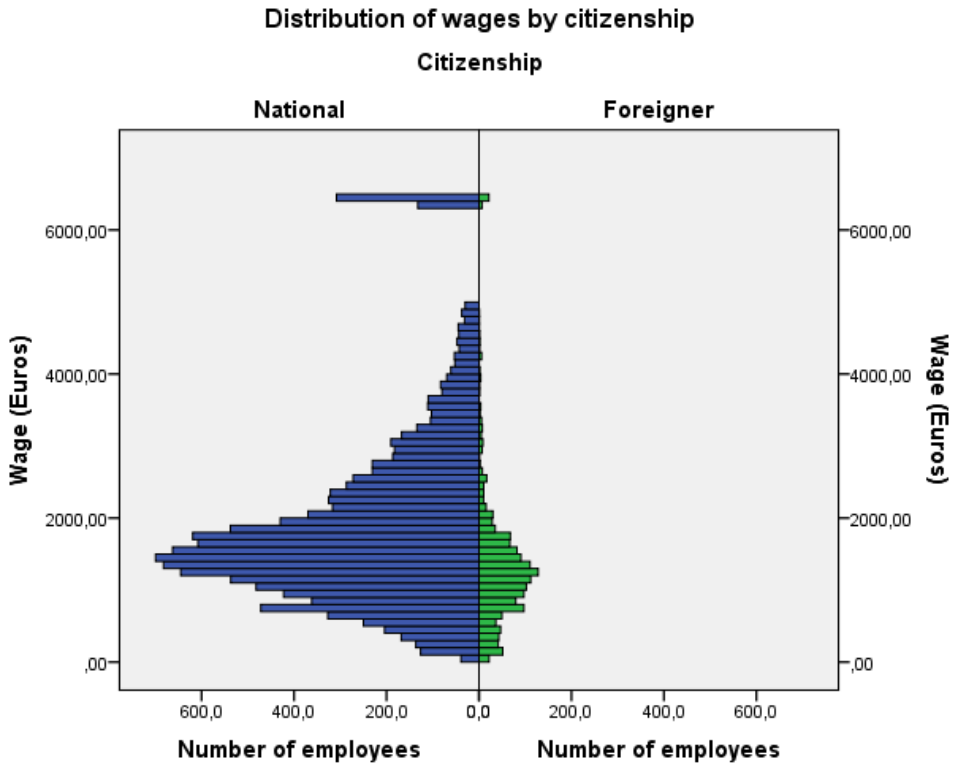
**Graph 2**




**Graph 3**



**Graph 4**




**Graph 5**