ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LIX 29 Number 4, 2011

SAMPLE REPRESENTATIVENESS VERIFICATION OF THE FADN CZ FARM BUSINESS SAMPLE

M. Prášilová, R. Zeipelt

Received: March 8, 2011

Abstract

PRÁŠILOVÁ, M., ZEIPELT, R.: *Sample representativeness verification of the FADN CZ farm business sample.* Acta univ. agric. et silvic. Mendel. Brun., 2011, LIX, No. 4, pp. 251–256

Sample representativeness verification is one of the key stages of statistical work. After having joined the European Union the Czech Republic joined also the Farm Accountancy Data Network system of the Union. This is a sample of bodies and companies doing business in agriculture. Detailed production and economic data on the results of farming business are collected from that sample annually and results for the entire population of the country´s farms are then estimated and assessed. It is important hence, that the sample be representative. Representativeness is to be assessed as to the number of farms included in the survey and also as to the degree of accordance of the measures and indices as related to the population. The paper deals with the special statistical techniques and methods of the FADN CZ sample representativeness verification including the necessary sample size statement procedure. The Czech farm population data have been obtained from the Czech Statistical Office data bank.

represenatativeness, sample, Farm Accountancy Data Network (FADN), agriculture, statistical methods, results of farming

In 1995 the project of test farms network started being introduced into practice in the Czech Republic. At that time a sample survey was designed and carried out by the Research Institute of Farm Economics (VÚZE) for the first time concerning the economic outcomes of farming enterprises in the CR in 1994. At the time of the CR admission to EU being prepared the efficiency of Czech agriculture had been assessed using different categories and procedures of economic outcomes assessment as against the EU practices. Therefore, already in 2001 it was decided to pay attention to the EU standards of legislation and to gradually accept the EU Member States economic outcomes indicators system in agriculture. In the same year the VÚZE in Prague started testing the application of the FADN farm accountancy data system. Currently, the FADN farm sample offers an important source of information on the actual economic situation of the CR farming enterprises to the EU.

MATERIAL AND METHODS

The aim of the paper is to verify the representativeness of selected data collected by means of the sample survey of the CR farming bodies economic outcomes in 2008 as based on the FADN network. The farm population is represented by Czech Statistical Office data.

The estimates from sample surveys are not as exact as the full census results since every sample survey is affected by a sampling error (ČERMÁK, VRABEC, 1999). In order to carry out more surveys based on samples it is useful to obtain a picture on numbers and sizes of the sample units (farms) as concerns their areas, production volumes, receipts and other volume indicators as well as it concerns hectare yields, costs per hectare of farm area and receipts per hectare, i.e., intensity figures. Exploratory analysis is aimed at these results.

Exploratory analysis helps to a fast recognition of the variation series types, i.a., to detection of the so-called outliers. Processing of the basic data into the so-called box-and-whisker plot is the primary

device of the analysis. The graph presents very simply and intuitively the location of the five most important values of the variation series, viz., the maximum and minimum, the median and both the two quartiles. It is accepted in general that, there are more outliers in larger populations, both absolutely and relatively. Outliers can be identified by means of the method, and they can be removed without risk of the population information power reduction. Homogeneity of the population can be increased this way and variation of the variables reduced. Data processed this way have been operated further on only.

First it is good to verify whether the same structure of bodies and companies in the size classes holds both in the population and the sample. The χ^2 test procedure can be applied to this end by means of which it can be verified whether the same frequency distribution holds for the population and the sample (ANDĚL, 2003). The following test criterion is employed:

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} \quad ,$$

where:
k ..... No. of classes,
n_j ..... actual frequencies,
np_j ..... theoretical frequencies (in the population).

If the χ^2 test criterion does not exceed χ^2_α, coincidence of the frequency distributions by regions both in the population and the sample can be accepted.

Further on, the differences in percentages of subjects in the sample and the population can be verified by means of the relative frequency test:

$$u = \frac{f_i - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}},$$

where:
u ..... test criterion value,
f_i ..... sample relative frequency,
π_0 ..... expected relative frequency value in the population,
n ..... sample size.

In case, the test criterion value exceeds the normal distribution critical value, a statistically significant difference is accepted. In such a case, the population can again be subjected to exploration analysis, and after one or two outliers having been removed, variation of the variable can be decreased and insignificance of the difference can be supported.

The sample data (the relative frequency of the variable) from the reference year can be tested against the sample data from the year before. The test procedure on the difference between two relative frequencies can be employed here. The test criterion is as follows:

$$u = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{\bar{p} \times \bar{q}}{n}}},$$

where

$$\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}, \quad \bar{q} = 1 - \bar{p}, \; n = \frac{n_1 \times n_2}{n_1 + n_2}$$

and where:
u ..... test criterion value,
n_1 ..... first sample size,
n_2 ..... second sample size,
m_1 ..... No. sample units with the certain attribute in the first sample,
m_2 ..... No. sample units with the certain attribute in the second sample.

In case, the test criterion value does not exceed the normal distribution critical value, then the difference between the two relative frequencies proves insignificant.

Besides the representativeness of the numbers of subjects sampled, the representativeness of indicators obtained had to be tested. For further testing of representativeness of sample surveys, tests were carried out in cases where data for the population were available, concerning the differences between sample values and actual population values (data from the official exhaustive statistical reporting). The test procedure concerning the μ_0 parameter value is applicable here:

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}},$$

where:
t ..... test criterion value,
$\bar{x}$ ..... arithmetic average of the variable studied,
μ_0 ..... average from the exhaustive reporting,
s^2 ..... sample variance,
n ..... sample size.

In case, the t-test criterion value exceeds the Student t-distribution critical value for the α significance level and the (n–1) degrees of freedom, then the difference between the average values in the sample and in the population is statistically significant.

The test procedures applied so far, should support the assumption that, based on the sample survey carried out, proper quality estimates on the Czech Republic agriculture can be constructed.

RESULTS

Sample survey representativeness

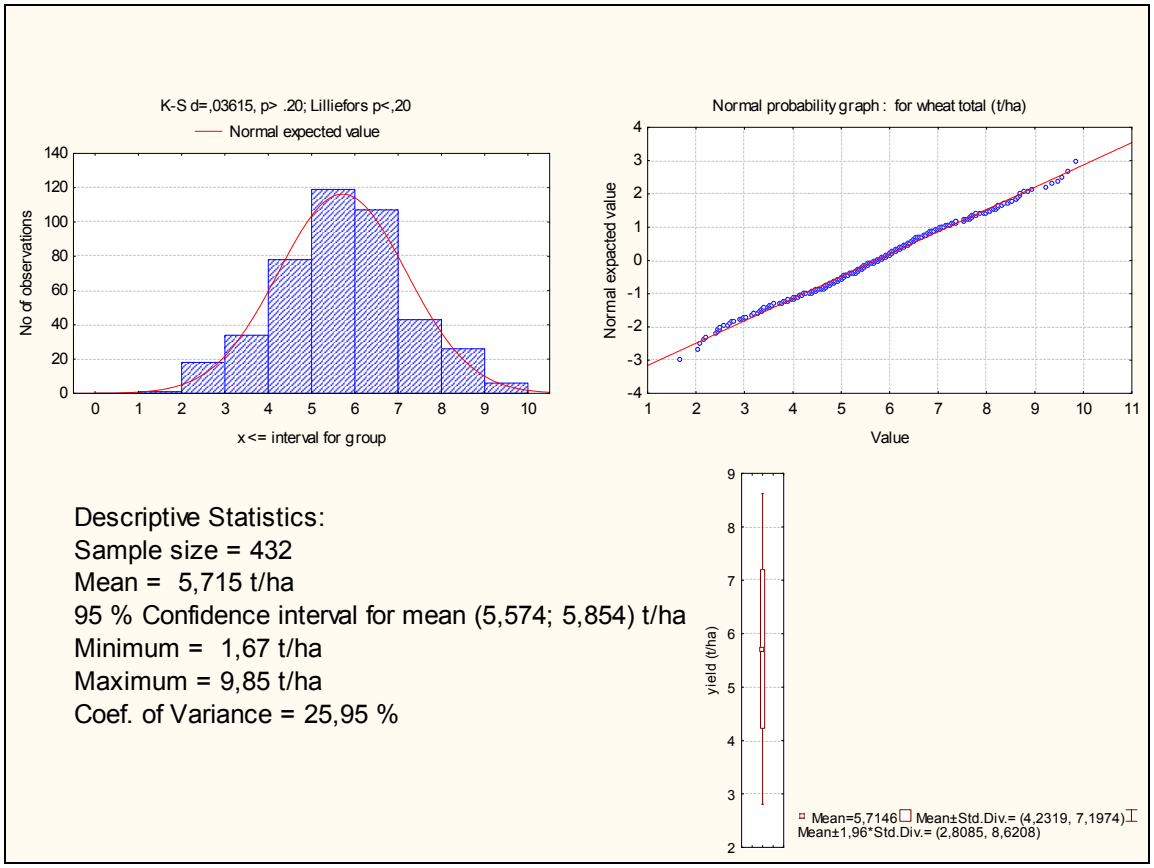
Testing of sample survey representativeness has been carried out concerning the test farms sample as to the selected crop yields and livestock efficiency. The authors first preserved (on purpose) all the values obtained from all samples and at the second step only, the data were subjected to exploratory

I: *Test criterion values for separate crops 2008*

Crop	Yields t/ha					Test criterion (t-test)	
	CR Total	Single-entry bookkeeping farms	No. subjects	Double-entry bookkeeping farms	No. subjects	Single-entry bookkeeping farms	Double-entry bookkeeping farms
Wheat	5.77	4.9391	610	5.715	432	–16.9871	–0.7763
Barley	4.65	4.3453	508	4.63	419	–6.8727	–0.3217
Potato	25.83	22.774	127	23.609	132	–4.8376	–2.9873
Rape	2.94	3.1306	305	2.932	362	4.1732	–0.2123

II: *Test criterion values for dairycows in 2008 (Farms with single-entry bookkeeping, Farms with double-entry bookkeeping, Total) in liters per day*

Indicator	Livestock efficiency					Test criterion (t-test)	
	CR Total	Single-entry bookkeeping farms	No. subjects	Double-entry bookkeeping farms	No. subjects	Single-entry bookkeeping farms	Double-entry bookkeeping farms
Milk yield (liters/day)	18.51	15.76	167	17.3	284	–6.9915	–5.0675



1: *Exploratory analysis, Wheat Total, FADN CZ 2008 (Double-entry bookkeeping farms) in t/ha*

analysis (Figure 1). The sample units have been grouped into two groups, the group of farms with single-entry bookkeeping and the group of farms with double-entry bookkeeping. The following Tables I and II are based on the original unadjusted FADN CZ data bank and they contain the test computation results for the entire Republic.

As concerns the CR Total, concordance has been obtained in the unadjusted data for 2008 on the hectare yields data from test farms network with double-entry bookkeeping and the assumed hectare yields by Czech Statistical Office. Potatoes show a significant difference in both the types of farms, that can be caused by the hectare yields high

variation level due to the year. The coefficient of variation is high in potatoes (31.26% for double-entry bookkeeping farms, 36.18% for single-entry bookkeeping farms).

In livestock production, the test farms sample representativeness verification was carried out for average livestock efficiency. The question was asked whether there was a difference between the value from the test farms sample and the actual population value (from the official statistical reporting). In case, the obtained test criterion value exceeds the normal distribution value, then the difference between the sample and the population values can be considered statistically significant. As the critical normal distribution value, the $u_{0.05} = 1.96$ value has been chosen. The following Table II contains the test results obtained.

The high variation of the unadjusted data caused the results of the testing procedure. As soon as the data were subjected to exploratory analysis, all the test exercises mentioned above demonstrated an insignificant difference between the sample results and the official statistics results, at 95% probability level. It may be considered a topic for further discussion, whether it is acceptable to eliminate the outliers in primary data (assuming they are correct).

Representativeness of surveys in the regions

Average hectare yields of selected crops and yields of livestock production in separate regions were tested in a similar way. In most cases, a difference between the population average (the CZSO data) and the sample average was not found significant which means that, the test criterion value is below 1.96 and the p value is above the 0.05 significance level. It is possible to state, hence, that the sample is well representative of the population and the sample survey results can well be generalized. A difference between the sample and the population averages has been found at 95% confidence in some of the regions. After a detailed analysis it has been discovered that, it concerned the regions where there is a low number of sample units and/or high variation of hectare yields. After raising the confidence level up to 99%, the differences showed themselves insignificant.

DISCUSSION AND CONSLUSION

The Farm Accountancy Data Network covered 1,677 farming subjects in 2008. This comparatively large sample was tested as to representativeness. Special statistical procedures demonstrated that, some of the data contained many outlying (extreme) values. These can cause a problem when verifying representativeness. Therefore, the outlying values were removed in such cases and as a result of the effort by the authors, verification of representativity of the indicators of the FADN CZ network in 2008 was accomplished.

As it has already been mentioned, the sample survey of economic subjects in the Czech agriculture shifted to the system of indicators used for agriculture of the EU member States. Currently, the FADN sample serves the most important data source on the situation in the CR agriculture. It is being employed by State bodies, research institutions and Universities. Also the farmers themselves use it for comparison and assessment of the positions of their own farms. The same sample offers basic data for economic development estimation over the next years, too. All these circumstances show that the sample can only be employed on the condition of its own representativeness, i.e., if it well represents the entire population of agricultural enterprises in the Czech Republic. The present paper demonstrates the representativeness verification procedures concerning the population of agricultural enterprises in 2008. The acceptance of the results is positive. Most of the indicators under study were representative, both for the CR total and for separate regions. It seems, too, that representativity is violated in case, there are outliers in the population. Exclusion of the outliers then facilitated representativity. A problem remains then, if the sample size becomes reduced following the exclusion of outliers. The actual 2008 population of farms did not contain many outliers and the exclusion of these in principle did not affect the sample size. The 2008 sample of 1,677 units can be considered sufficiently large as it concerns representativeness. The regular annual partial replacement of sample units remains a problem, anyway. The replacement is caused both by further non-existence of some of the sample units and by disinterest of some of the farmers in continued co-operation. New sample units have to be recruited, then. Recently, the replacement covered 5% of farms, which is admissible. The sample units replacement together with other conditions make it necessary to perform the sample representativeness verification annually following the method presented. It is understood, the method has to be developed and improved gradually. In 2009, the sample was increased and the questionnaire was expanded by new indicators. This has been a positive fact as it concerns the representativeness verification, but the operation has become more demanding from the viewpoint of the reporting units. The general wish is to make the sample of agricultural subjects in the FADN CZ stable, with minimal necessary replacement. Then only the sample will represent the population of farms properly, even after possible exclusion of outliers.

SUMMARY

Economic outcomes of Czech agriculture have since 1994 been assessed based on the surveys carried out in the network of the agricultural primary production test farms (FADN CZ). The FADN farms report their results annually, hence, currently the sample of farms is comparatively stable with a long series of results. The network has been organized and administered by the Research Institute of Agricultural Economics in Prague (ÚZEI, currently) under the title of „Sample survey of economic outcomes by enterprising subjects". Already during the period of the CR admission to EU being prepared, it was decided to pay attention to EU legislation as concerns the surveys and to accept gradually the system of economic results indicators in agriculture of EU member States. The new system employs the agricultural bookkeeping data network with some other indicators added. Currently, the FADN sample serves an important information resource to the EU, concerning the actual economic situation of Czech agriculture, and it also can be applied for international comparison. The FADN data have to be faultless and of high quality and the sample of reporting units has to be representative. The paper deals with the sample representativeness verification and with representativeness of some selected indicators of the agricultural primary production. The basic data have been obtained from the last survey finished in 2008. The entire population is being described by the ČSÚ data. The agricultural bookkeeping data network covered 1,677 farm subjects in 2008, grouped into farms with single-entry and with double-entry bookkeeping. Outliers were discovered and excluded by means of exploratory analysis. Representativeness by regions and by production branches was verified by means of statistical test procedures. Using one-sample and two-sample tests, average levels of the intensity indicators in crop production and livestock production were compared. The original data from the farms with double-entry bookkeeping in 2008 showed in most cases full conformity with the data from the entire population. Once the outliers were excluded from the data, the indicators under study showed conformity with the ČSÚ data from the both types of farms.

REFERENCES

ANDĚL, J., 2003: *Statistické metody*. Praha: Matfyz-press. ISBN 80-86732-08-8.

ČERMÁK, V., VRABEC, M., 1999: *Teorie výběrových šetření – Část I*. Praha: VŠE. ISBN 80-7079-191-8.

ČERMÁK, V., VRABEC, M., 1998: *Teorie výběrových šetření – Část II*. Praha: VŠE. ISBN 80-7079-609-X.

Collective of authors, 2010: *Výběrové šetření hospodářských výsledků zemědělských podniků v síti FADN CZ za rok 2008*. Praha: ÚZEI.

PACÁKOVÁ, I., 2008: *Statistika v terénních průzkumech.* Praha: Professional Publishing. ISBN 978-80-86946-74-0.

Address

doc. Ing. Marie Prášilová, CSc., doc. Ing. Rudolf Zeipelt, CSc., Česká zemědělská univerzita v Praze, Katedra statistiky, Kamýcká 129, 165 21 Praha 6, e-mail: prasilova@pef.czu.cz, zeipelt@pef.czu.cz