Interactive Data Analysis of OECD – Data: Introduction to Permutation Tests

Grzegorz Kończak[1], Karsten Lübke[2]

Abstract

For a long time, there has been a call for the use of real, multivariate data within statistical education (e.g. GAISE, 2016) and to integrate civic data (e.g. ProCivicStat, 2018).

On the other hand, there is some evidence that Simulation Based Inference, i.e. Bootstrapping and Permutation Tests can help to improve conceptual understanding of inferential statistics (see e.g. Chance *et al.*, 2016; Maurer and Lock, 2016; Tintle *et al.*, 2018; Hildreth *et al.*, 2018; Lübke *et al.*, 2019). In general, the permutation methods are considered to be more powerful than bootstrap (Berry *et al.*, 2014) for calculating e.g. p-values. In permutation tests a test statistic is calculated for the observed data. Next the distribution of this test statistic is estimated using permutations of the observations over all possible arrangements (Kończak, 2016) or by Monte-Carlo Methods (Ernst, 2004).

By the help of an interactive Shiny Apps (Doi *et al.*, 2016) we combine these ideas, real, civic data and permutation tests: From the OECD Regional Well-Being (OECD Regions and Cities at a Glance 2018) data-set users can choose the indicators and the two countries they want to compare. By using ggplot2 graphics (Wickham 2016) we hope to ease the transition from learning to doing statistics (see e.g. McNamara, 2019). Currently the app is available in Polish, German and English.

Keywords: OECD Well-Being, Permutation Tests, Shiny Application, Statistical Education

JEL Classification: C12, C15, I31

1. Introduction

Statistical methods allow comparison of population characteristics. Parametric tests such as *t*-test and non-parametric tests such as U-Mann-Whitney's test are used to compare e.g. expected values or other aspects of the distribution. In recent years, computer simulation methods such as bootstrap and permutation tests have been increasingly used to compare population characteristics. The main goal of this paper is to present the idea of permutation tests with use of the real civic data with an interactive Shiny Apps.

Permutation methods precede many traditional parametric statistical methods, but only recently permutation methods have become part of the mainstream discussion of statistical

[1] University of Economics in Katowice, grzegorz.konczak@ue.katowice.pl.
[2] FOM University of Applied Sciences, Lissaboner Allee 7, 44269 Dortmund, Germany, karsten.luebke@fom.de.

tests. Permutation tests were proposed by Fisher (1925) and further developed by Pitman (1937). Berry *et al.* (2014) indicate that in 1923 Neyman (Spława-Neyman, 1923) introduced a permutation model for analyzing field experiments in agriculture, but the article was not translated into English until 1990.

Shiny Apps are web-based applications that enable easy interactive analysis and can be programmed in R. Numerous examples for different purposes exist (Doi *et al.*, 2016) and used in a consistent way they may also help to narrow the gap between learning and doing statistics (see e.g. McNamara, 2019) within the learning curve.

To engage students and to show the value of statistics as well as the concept of permutation tests we use the OECD Regional Well-Being (OECD Regions and Cities at a Glance 2018) data-set in which users can choose the indicators and the two countries they want to compare.

2. The idea of permutation tests

Permutation tests were proposed by Fisher (1925) and Pitman (1937). These tests are computer-intensive statistical methods. Instead of comparing the observed value of the test statistic to a known standard distribution, the reference distribution is generated from the sample data in permutation tests. These tests can give results that are as accurate as than those obtained with the use of traditional statistical methods.

The concept of these tests is simpler than of the tests based on normal distribution. The idea of permutation tests is like the bootstrap method but in bootstrap samples are taken with replacement. The main application of these tests is a two-sample problem (Efron and Tibshirani, 1983).

The permutation testing is based on the 5 following steps (Good, 2006):

1. Identify the null hypothesis and the formulate the alternative hypotheses.
2. Choose a test statistic (T).
3. Compute the test statistic (T_0).
4. Determine the frequency distribution of the statistic under the null hypothesis.
5. Make a decision using this distribution as a guide.

We will concentrate on testing equality of means of two distributions. The form of the null hypothesis is that the two distribution will be following

$$H_0: F_1 = F_2 \ (H_0: \mu_1 = \mu_2)$$

against the alternative

$$H_1: \mu_1 \neq \mu_2.$$

The form of the alternative hypothesis is not the simple negation of the null hypothesis. Due to the form of the alternative hypothesis the test statistics that discriminates between the null hypothesis and the alternative should be chosen, so instead of sample mean also sample proportion for categorical data can be chosen.

Let us consider two samples S_1 and S_2 where $S_1=\{x_1, x_2, \ldots, x_{n1}\}$ and $S_2 = \{y_1, y_2, \ldots, y_{n2}\}$. We will consider the statistic

$$T = \bar{X} - \bar{Y} \tag{1}$$

where $\bar{X}$ and $\bar{Y}$ are respectively means of the first and the second sample. Lest us denote by T_0 the value of this test statistic for the obtained data. It is easy to notice that big absolute values of the test statistic T_0 will lead to the rejection of the null hypothesis. To obtain a distribution of the test statistic under the null hypothesis data labels should be permuted N times. Let us consider the joined sample $S = S_1 \cup S_2$. The joined sample S should be N times randomly divided into two samples of sizes n_1 and n_2. For each case the value of the test statistic T is calculated. We obtain values T_1, T_2, ..., T_N. The empirical distribution of these values is an estimate of the distribution of the statistic T under the null hypothesis of identical distributions.

The achieved significance level (ASL) should be determined (Efron and Tibshirani, 1983). ASL is the probability of observing at least that a value as large as T_0 when the null hypothesis is true. The value of ASL is determined based on a series of data permutations.

$$ASL = P_{Ho}\{|T| \geq |T_0|\}. \tag{2}$$

The smaller the value of ASL the stronger is evidence against the null hypothesis. For a given significance level α (like 0.05, 0.1 or 0.01) we reject the null hypothesis if ASL is less or equal to α. Otherwise we do not reject H_0. Usually the value of ASL is unknown. It can be estimated as follows:

$$\widehat{ASL} = \frac{card\{i:|T_i| \geq |T_0|\}}{N}. \tag{3}$$

3. OECD Well-Being Data

OECD Regional Well-Being data (2019) is used in the paper. The data shows how regions perform when it comes to education, environment, safety, and other topics important to your well-being. This interactive site allows you to measure well-being in your region and compare it with 402 other OECD regions based on eleven topics central to the quality of our lives.

The data allows to measure well-being of 402 regions in 36 countries. There are included 25 EU countries (without Bulgaria and Romania) and United Kingdom, Australia, Canada, Chile, Israel, Japan, Korea, Mexico, Switzerland, Turkey, United States. The number of regions

is differing for various countries from 1 in Luxembourg and 2 in Slovenia up to 51 in United States. Each region is measured in eleven following topics: income, jobs, housing, health, access to services, environment, education, safety, civic engagement and governance, community, and life satisfaction (for details see Table 1). A score has been calculated for each topic so that places and topics within and across countries could be easily compared.

Table 1. Variables and quartiles for 402 regions in OECD Well-Being Data

No.	Variable	Unit	Region		
			1st Qu.	Median	3rd Qu.
1	Labour force with at least secondary education	%	60.70	81.60	90.10
2	Employment rate	%	61.40	67.60	72.97
3	Unemployment rate	%	4.100	5.500	8.575
4	Household disposable income per capita	constant USD PPP	11362	17725	23136
5	Homicide rate	per 100000 people	0.700	1.200	3.700
6	Mortality rate	per 100000 people	7.100	8.050	9.400
7	Life expectancy	number of years	78.00	80.50	82.20
8	Air pollution (level of PM2.5)	$\mu g/m^3$	8.1	12.4	17.6
9	Voter turnout	%	56.17	71.00	83.30
10	Broadband access	% of households	67.00	78.00	86.22
11	Number of rooms per person	rooms per person	1.300	1.800	2.100
12	Perceived social network support	%	85.90	91.45	93.90
13	Self assessment of life satisfaction	index 0 to 10	6.025	6.800	7.375

4. Permutation Test for OECD Well-Being Data

The idea of the use of permutation tests is presented for the 1st variable Labour force with at least secondary education (Table 1) for two countries: Germany and Poland. There are 16 regions in Germany (land) and 16 regions in Poland (voivodeship). The values for the considered variable for regions of Germany and Poland are the following (%):
Germany: 85.1, 87.2, 87.6, 92.2, 84.1, 86.2, 85.6, 91.8, 85.5, 84.3, 83.8, 85.2, 94.0, 91.4, 87.0, 94.5.
Poland: 93.5, 95.3, 96.1, 96.2, 93.2, 95.3, 93.2, 92.9, 95.1, 92.3, 93.3, 95.3, 93.9, 93.7, 90.3, 93.9.

The mean values of the variable Labour force with at least secondary education are for Germany 87.844% and for Poland 93.969%.

Note that the inference here is not from sample to population but about form sample to data generating process.

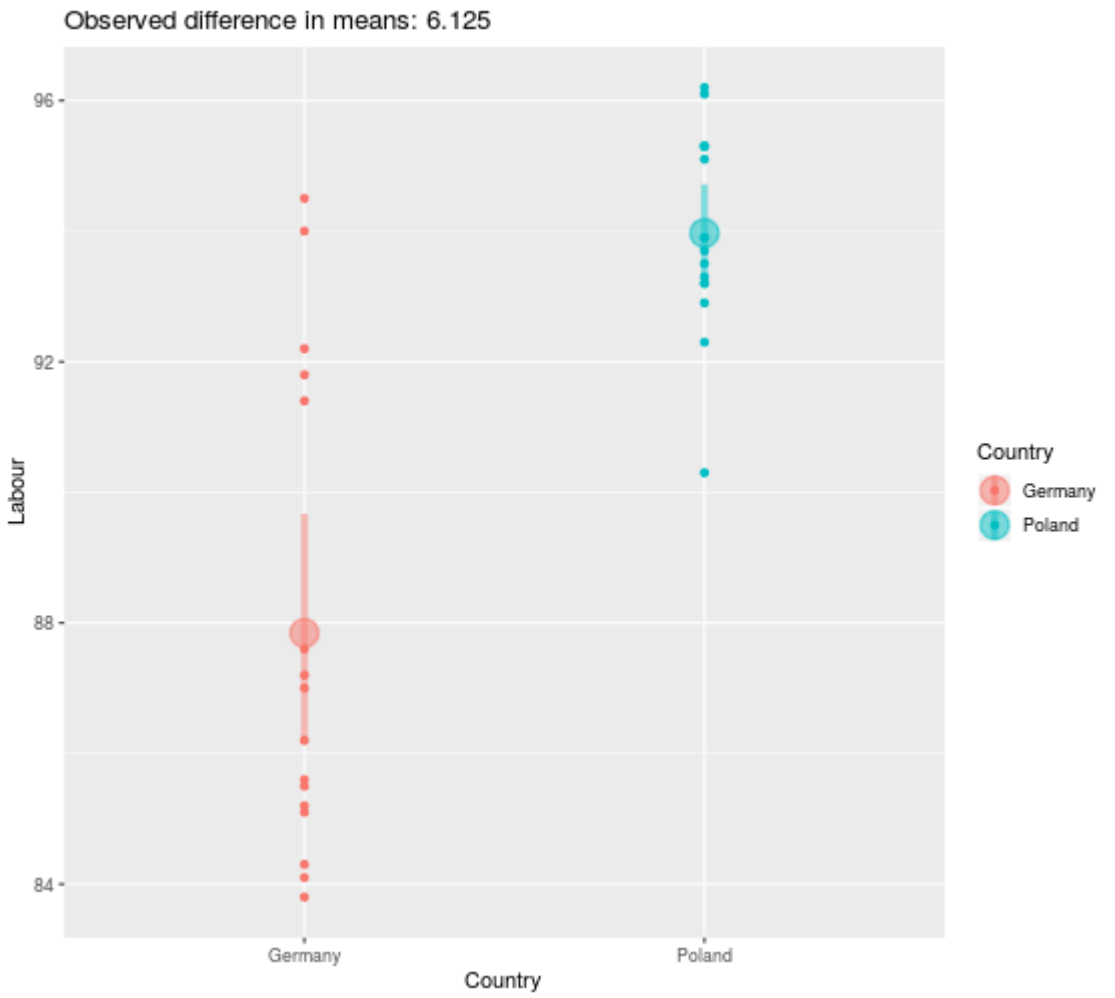

Figure 1. Observed difference for the variable *Labour force with at least secondary education* for Germany and Poland

The values for considered regions and results of permutation testing are presented in Fig. 1. It could be seen that the mean in percentage of workers with at least secondary education in Poland is greater than in Germany. Fig. 2 show the empirical distribution of the test statistic T under H_0 and the value T_0 obtained for the original data. The difference of means for this variable for German and Poland regions data is equal to -6,125. The values of the test statistic T obtained for the $N = 100$ of data permutation are between -4 and 4. This leads to the rejection of the null hypothesis: it is very unlikely – if the model of no difference in distribution would be true - to observe such an extreme value in a sample as we actually did. So the data provides evidence that there is a difference in the distribution of Labour force with at least secondary education for Germany and Poland.

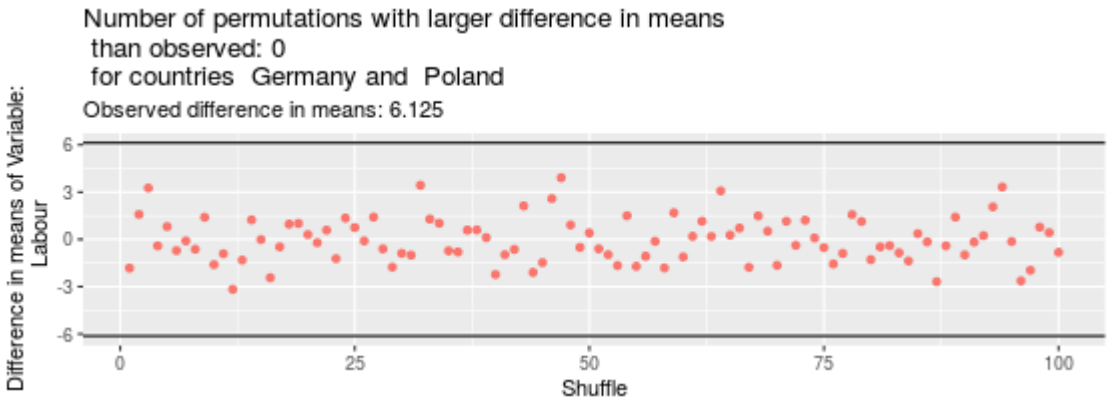

Figure 2. Permuted difference in means for the variable *Labour force with at least secondary education* for Germany and Poland

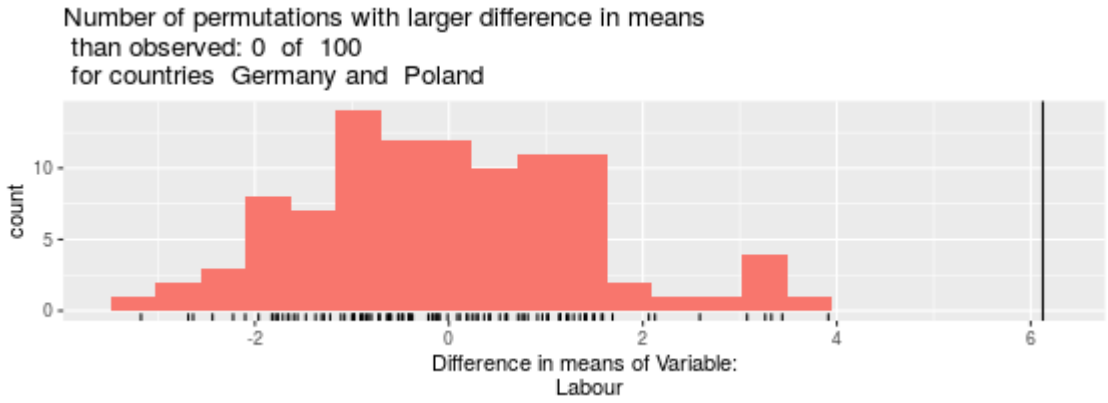

Figure 3. Permutation distribution of difference in means for the variable *Labour force with at least secondary education* for Germany and Poland

Fig. 4 shows a screenshot from the app (https://fomshinyapps.shinyapps.io/OECD-Permutation) [access 2020.01.24]).

Figure 4. Screenshot of the shiny app

5. Conclusions

With real, civic data, available in an interactive way, students can be motivated to investigate data, learn about society for variables and countries they may be interested in – and not just of Labour force with at least secondary education for Germany and Poland as in our example.

By using permutation tests, statistical thinking in terms of models, distributions, hypothesis, and randomness can be promoted: How would the sampled data look like if there would be no difference in the distributions?

We think that such apps offer a nice opportunity for statistic teachers to ask good questions and to show why statistical literacy and thinking is such an important skill in times of big data.

References

Berry, K.J., Johnston, J.E., Mielke Jr., P.W. (2014). *A Chronicle of Permutation Statistical Methods*, Springer International Publishing, New York.

Chance, B., Wong, J., Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, 24(3), 114-126.

Doi, J., Potter, G., Wong, J., Alcaraz, I., Chi, P. (2016). Web application teaching tools for statistics using R and shiny. *Technology Innovations in Statistics Education*, 9(1).

Efron, B., Tibshirani, R. (1983). *An Introduction to the Bootstrap*. Science Business Media, Inc.

Ernst, M.D. (2004) Permutation methods: a basis for exact inference. *Statistical Science*, 19(4), 676-685.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

GAISE (2016). GAISE College Report ASA Revision Committee: Guidelines for Assessment and Instruction in Statistics Education College Report 2016.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer Science Business Media, Inc., New York.

Good, P. (2006). *Resampling Methods. A Practical Guide to Data Analysis*. Birkhauser. Boston-Basel-Berlin, 2006.

Hildreth, L.A., Robison-Cox, J., Schmidt, J. (2018) Comparing Student Success and Understanding in Introductory Statistics under Consensus and Simulation-Based Curricula. *Statistics Education Research Journal*, 17(1).

Maurer, K., Lock, D. (2016). Comparison of learning outcomes for simulation-based and traditional inference curricula in a designed educational experiment. *Technology Innovations in Statistics Education*, 9(1).

Kończak, G. (2014). On the modification of the non-parametric test for comparing locations of two populations, *Proceedings of COMPSTAT 2014*, ed.: M. Gilli, G. Gonzalez-Rodriguez, A. Nieto-Reyes, 35-44.

Kończak, G. (2016). Testy permutacyjne. Teoria i zastosowania. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Lübke, K., Gehrke, M., Markgraf, N. (2019). Statistical Computing and Data Science in Introductory Statistics. In *Applications in Statistical Computing* (pp. 139-150). Springer, Cham.

Lehmann, E.L., Romano, J.P. (2005). *Testing Statistical Hypothesis*, Springer Science+Business, Inc., New York.

McNamara, A. (2019). Key attributes of a modern statistical computing tool. *The American Statistician*, 73(4), 375-384.

OECD Regional Well-Being, https://www.oecdregionalwellbeing.org [access: 2019.12.22].

ProCivicStat (2018). Engaging civic statistics: A call for action and recommendations, http://iase-web.org/islp/pcs [access: 2020.01.22].

Pitman, E.J.C. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society* 4, 119-130.

Romano, J. (1989). Bootstrap and Randomization Tests of Some Nonparametric Hypotheses, *The Annals of Statistics*, vol. 17, No. 1, 141-159.

Spława-Neyman, J. (1923). Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych. Rocznik Nauk Rolniczych, v. 10, s. 1-51.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation-based inference. *Journal of Statistics Education*, 26(2), 103-109.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.