Döhrn, Roland

**Working Paper**

# Comparing forecast accuracy in small samples

Ruhr Economic Papers, No. 833

**Provided in Cooperation with:**
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

*Suggested Citation:* Döhrn, Roland (2019) : Comparing forecast accuracy in small samples, Ruhr Economic Papers, No. 833, ISBN 978-3-86788-966-7, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen, https://doi.org/10.4419/86788966

This Version is available at:
https://hdl.handle.net/10419/209589

RUHR
ECONOMIC PAPERS

Roland Döhrn

# Comparing Forecast Accuracy in Small Samples

#833

# Imprint


## Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany



## Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Roland Döhrn, Prof. Dr. Manuel Frondel, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

## Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de



## Ruhr Economic Papers #833

Responsible Editor: Roland Döhrn





ISSN 1864-4872 (online) – ISBN 978-3-86788-966-7

Ruhr Economic Papers #833

Roland Döhrn

# Comparing Forecast Accuracy in Small Samples

UNIVERSITÄT DUISBURG ESSEN

RWI

## Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/86788966
ISSN 1864-4872 (online)
ISBN 978-3-86788-966-7

Roland Döhrn[1]

# Comparing Forecast Accuracy in Small Samples

## Abstract

*The Diebold-Mariano-Test has become a common tool to compare the accuracy of macroeconomic forecasts. Since these are typically model-free forecasts, distribution free tests might be a good alternative to the Diebold-Mariano-Test. This paper suggests a permutation test. Stochastic simulations show that permutation tests outperform the Diebold-Mariano-Test. Furthermore, a test statistic based on absolute errors seems to be more sensitive to differences in forecast accuracy than a statistic based on squared errors.*

*JEL-Code: C14, C15, C53*

*Keywords: Macroeconomic forecast; forecast accuracy; Diebold-Mariano test; permutation test*

*December 2019*

*1 Roland Döhrn, UDE and RWI.* *- The author thanks Boris Blagov, Angela Fuest, and Martin Micheli for valuable comments to a earlier version of this paper.* *- All correspondence to: Roland Döhrn, RWI, Hohenzollernstr. 1-3, 45128 Essen, Germany, e-mail: roland.doehrn@rwi-essen.de*

## 1. Introduction

Comparing the accuracy of macroeconomic forecasts is difficult for various reasons: The task is, in Diebold's (2015: 1) words, to compare "model-free forecasts". Furthermore, the samples are often quite small, the properties of the forecast errors are unknown, and, finally, forecasts are not independent from earlier forecasts of the same institution and of other institutions (Gallo et al. 2002). A widely used and simple test for comparing predictive accuracy is the Diebold Mariano-(DM)-Test (Diebold and Mariano 1995). However, the test statistic tends to be oversized when samples are small and when forecast errors have heavy tails. Furthermore, non-stationary errors may cause problems. However, the strength of the test is its simplicity, and a simple approach to overcome these problems, at least partially, is the modification proposed by Harvey, Leybourne and Newbold (1997).[1]

As an alternative to the DM-Test, forecast comparisons in a model-free world suggest the use of a non-parametric test. However, different from other research-fields with typically small samples such as meteorology (Peissendorfer, Barnett 1983) or medicine (Ludbrook, Dudley 1998), these approaches have not become popular in the forecast comparison literature. This paper tries to fill this gap by proposing permutation tests.[2]

In many applications, permutation tests have turned out to be a powerful alternative to parametric tests. Commonly, they are applied to two independent samples. A test statistic is chosen (e.g. the difference of means) and calculated for the original sample. After that, the observations are resampled without repetition, and the test statistic is calculated for each resampling. By that, an empirical distribution of the test statistic is generated. Finally, the test statistic of the original data is compared to the empirical distribution.

---

[1] *Originally meant just for comparing forecasts (Diebold 2015), in the follow-up a rich literature emerged that pushed ahead the DM-test in various directions, in particular to make it applicable to model-based forecasts (for an overview see Diebold 2015, Coroneo and Iacome 2015).*

[2] *A simple alternative in the case of matched pairs would be a sign test, or, as a more powerful one, a signed rank test (Butar Butar, Bandularesi 2009).*

Adopting this approach to forecasts has to take into account that forecasts relate to a specific year. Thus, resampling all data would not be helpful. Therefore, a permutation test for matched pairs must be used. This also allows for a simple interpretation of the test as a treatment problem: Two forecasters know the same raw data, but they 'treat' them differently, e.g. by using different models. The test evaluates, whose treatment is more 'successful' in terms of forecast accuracy.

Table 1 illustrates a permutation test for matched pairs. For simplicity, only five years are considered. The numbers are absolute forecast errors of two forecaster (FC1 and FC2). The difference in the mean absolute error is -0.32. Next, all possible permutations are calculated, in the present case 32 ($=2^5$). The first is identical to the original data. The second is a combination of four values taken from FC1 and one from FC2 and so forth. Of particular interest are permutations 5 and 6, since they are the only ones showing a difference in forecast means exceeding that of the original data. In total, three of 32 permutations, i.e. 9.4% of all cases, show a difference in the mean absolute error that is equal to or smaller than the original. Thus, the null hypothesis of equal mean forecast errors cannot be rejected at the usual level of significance.

Table 1
**Forecast permutation an differences in forecast errors: an example**

| | | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 | Mean | Diff |
|---|---|---|---|---|---|---|---|---|
| Original | FC1 | 0.4 | 3 | 2 | 0.4 | 0.1 | 1.18 | |
| | FC2 | 0.8 | 3.9 | 1.7 | 0.8 | 0.3 | 1.5 | -0.32 |
| Perm 1 | FC1 | 0.4 | 3 | 2 | 0.4 | 0.1 | 1.18 | |
| | FC2 | 0.8 | 3.9 | 1.7 | 0.8 | 0.3 | 1.5 | -0.32 |
| Perm 2 | FC1 | 0.4 | 3 | 2 | 0.4 | **0.3** | 1.22 | |
| | FC2 | 0.8 | 3.9 | 1.7 | 0.8 | **0.1** | 1.46 | -0.24 |
| Perm 3 | FC1 | 0.4 | 3 | 2 | **0.8** | **0.3** | 1.26 | |
| | FC2 | 0.8 | 3.9 | 1.7 | **0.4** | **0.1** | 1.42 | -0.16 |
| … | | | | | | | | |
| Perm 5 | FC1 | 0.4 | 3 | **1.7** | 0.4 | 0.1 | 1.12 | |
| | FC2 | 0.8 | 3.9 | **2** | 0.8 | 0.3 | 1.56 | -0.44 |
| Perm 6 | FC1 | 0.4 | 3 | **1.7** | 0.4 | **0.3** | 1.16 | |
| | FC2 | 0.8 | 3.9 | **2** | 0.8 | **0.1** | 1.52 | -0.36 |
| … | | | | | | | | |
| Perm 31 | FC1 | **0.8** | **3.9** | **1.7** | **0.8** | 0.1 | 1.46 | |
| | FC2 | **0.4** | **3** | **2** | **0.4** | 0.3 | 1.22 | 0.24 |
| Perm 32 | FC1 | **0.8** | **3.9** | **1.7** | **0.8** | **0.3** | 1.5 | |
| | FC2 | **0.4** | **3** | **2** | **0.4** | **0.1** | 1.18 | 0.32 |

*Author's calculations.*

This paper analyses the use of permutation test for forecast comparisons in more depth. It starts with a short description of the test (Section 2). In Section 3, an application is presented, in which the outcome of the permutation test is also compared to the traditional DM-test. Section 4 provides a more detailed analysis and evaluates the power of the permutation test in comparison to the DM-test. Section 5 concludes.

## 2. Permutation test for matched pairs

The setup of the permutation test has already been outlined briefly in the introduction. For a more general description, we consider a bivariate sample of size T.

(1) $$(e_{A,1}\ e_{B,1})\ (e_{A,2}\ e_{B,2})\ (e_{A,3}\ e_{B,3})\ \ldots.\ (e_{A,T}\ e_{B,T})$$

Each element of the sample relates to a specific year and it contains a forecast error of forecaster A and of forecaster B. The permutations are calculated for each of these pairs, i.e. the elements will not be changed between the pairs.

The null hypothesis to be tested is that Forecasters A and B have the same accuracy. Depending on the loss function, one can use the difference of mean absolute errors as a test statistic ($\theta_{MAE}$) or the difference in the mean squared errors ($\theta_{MSE}$); the latter is more in line with the DM-test.

(2) $$\theta_{MAE} = \frac{1}{T}\sum_t |e_A| - \frac{1}{T}\sum_t |e_B|$$

(3) $$\theta_{MSE} = \frac{1}{T}\sum_t (e_A)^2 - \frac{1}{T}\sum_t (e_B)^2$$

To describe the test, it is helpful to re-write pairs in (1) in a way that accounts for the order of the two elements in each pair. The first element $E_1$ of each pair t is one of the two members of the matched pairs. The second element $E_2$ is the member not chosen as the first element.

(4) $$(E_{1,t} = (e_{A,t} \vee e_{B,t})\ \ E_{2,t} \in (e_{A,t}, e_{B,t}) \neg E_{1,t})$$

Now all pairs are resampled in this way P times to get P permutations. The total number of permutations is $2^T$. If T is not too large, the full set of possible permutations can be calculated. However, with T rising the

number of permutations grows exponentially, and accordingly the execution time of the test. Therefore, it is advisable to use bootstrapping for large Ts. Whether full set or bootstrapped: For each permutation p the test statistics $\theta_{L,p}$ are calculated according to (2) and (3); where p denotes the permutation and L the loss function (L∈(MAE,MSE)). After each permutation, the test statistic of the original data $\theta_{L,o}$ is compared to $\theta_{L,p}$. The result of this comparison is stored in a count-vector c.

(5) $$c_{L,p} = \begin{cases} 1 & if\ \theta_{L,p} \geq \theta_{L,o} \\ 0 & else \end{cases}$$

After having completed the permutations, the test statistic s of the permutation test is:

(6) $$s_L = \frac{\sum c_L}{P} \quad for\ l = MAE, MSE$$

Thus, the test statistic s is the share of permutations showing a higher difference in mean absolute or squared errors than the original data. A different interpretation is rendered by looking at the distribution of the $\theta_{L,p}$: It is tested, whether the original $\theta$ is located at the upper or at the lower end of this distribution.

From (4) it becomes evident, that the distribution of the $\theta_{L,p}$ will be symmetric in the case the full set of permutations is calculated. Every combination of forecasts appearing as $E_1$ will also appear once as $E_2$, and the $E_2$ resp. $E_1$ will be the same complement. If the permutations are bootstrapped, symmetry is not warranted, but it will be reached asymptotically.

With $\alpha$ being the level of significance, the null should be rejected if $s<\alpha/2$ or $s>(1-\alpha/2$. In the first case, an overwhelming share of permutations shows a $\theta$ which is smaller than the observed one. This indicates that the mean error of forecast A is larger than the mean error of forecast B at a level of significance of $\alpha$. In the second case, most permutations show a larger error, meaning that forecast A tends to be the more accurate one.

## 3. An example

Subsequently, the test procedure will be applied to two samples of forecasts. The first example relates to one year ahead forecasts for the

US economy in the period 2008 to 2018. They have been contributed by the forecasters to the Consensus Economics survey of December.[3] Thus they can be assumed to be released in the second half of November or in early December. Table 2 presents some descriptive statistics of these forecasts. It shows that MAE as well as MSE do not deviate too much except for the forecast of the Eaton Corporation which shows lower values.

Table 2

**Accuracy measures of US-December GDP forecasts for the following year**

| Forecaster | Acronym | MAE | MSE |
|---|---|---|---|
| Eaton Corporation | EATON | 0.45 | 0.35 |
| Moody's Analytics | MOODY | 0.68 | 0.63 |
| The Conference Board | CONF | 0.60 | 0.49 |
| Fannie Mae | FANNIE | 0.59 | 0.53 |
| Nat Assn of Home Builders | NAH | 0.55 | 0.50 |
| Swiss Re | SWISS | 0.72 | 0.62 |
| Univ of Michigan – RSQE | UoM | 0.58 | 0.53 |
| Wells Fargo | WELLS | 0.62 | 0.60 |

*Source: Consensus economics; author's calculations.*

Table 3 shows the results of three tests comparing the accuracy of these forecast pairwise: The DM-Test, the permutation test for MSE (PERMSE), and the permutation test for MAE (PERMAE). In neither case does the DM-Test reject the null of equal forecast accuracy. The PERMSE-Test shows in one case that mixing the forecast of EATON with another forecasts generates a higher MSE in more than 95% of all permutations, i.e. the null of equal forecast accuracy can be rejected with an error margin of 10%. The PERMAE-test generates an even clearer result: The EATON forecast shows a lower MAE than four of the competing forecasts, in three cases at an error margin of 10%, in one case it is significant even at the 5%-level. Furthermore, the NAH-forecast outperforms one of its competitors.

[3] Consensus Economics collects the forecasts in the first week of a month. However, it is unclear when the forecasts were produced. In three cases, missing data were taken from the January survey.

Table 3

**Pairwise tests for equal forecast accuracy of eight December forecasts of U.S. GDP, 2008-2018**

| | Diebold-Mariano-Test, p-values[1] | | | | | | |
|---|---|---|---|---|---|---|---|
| | MOODY | CONF | FANNIE | NAH | SWISS | UoM | WELLS |
| EATON | 0.231 | 0.137 | 0.270 | 0.457 | 0.241 | 0.189 | 0.150 |
| MOODY | | 0.555 | 0.423 | 0.081 | 0.947 | 0.573 | 0.937 |
| CONF | | | 0.811 | 0.974 | 0.556 | 0.809 | 0.450 |
| FANNIE | | | | 0.800 | 0.394 | 0.996 | 0.778 |
| NAH | | | | | 0.474 | 0.845 | 0.735 |
| SWISS | | | | | | 0.703 | 0.961 |
| UoM | | | | | | | 0.668 |
| | **Permutation test, MSE[2]** | | | | | | |
| | MOODY | CONF | FANNIE | NAH | SWISS | UoM | WELLS |
| EATON | 0.894 | 0.949 | 0.854 | 0.816 | 0.890 | 0.833 | **0.951** |
| MOODY | | 0.246 | 0.301 | 0.104 | 0.472 | 0.268 | 0.457 |
| CONF | | | 0.594 | 0.520 | 0.729 | 0.580 | 0.617 |
| FANNIE | | | | 0.359 | 0.740 | 0.500 | 0.614 |
| NAH | | | | | 0.813 | 0.582 | 0.668 |
| SWISS | | | | | | 0.311 | 0.473 |
| UoM | | | | | | | 0.739 |
| | **Permutation text, MAE[2]** | | | | | | |
| | MOODY | CONF | FANNIE | NAH | SWISS | UoM | WELLS |
| EATON | **0.950** | **0.986** | 0.869 | 0.777 | **0.965** | 0.905 | **0.970** |
| MOODY | | 0.259 | 0.222 | 0.102 | 0.610 | 0.172 | 0.298 |
| CONF | | | 0.500 | 0.355 | 0.823 | 0.463 | 0.551 |
| FANNIE | | | | 0.297 | 0.882 | 0.492 | 0.563 |
| NAH | | | | | **0.954** | 0.691 | 0.707 |
| SWISS | | | | | | 0.161 | 0.231 |
| UoM | | | | | | | 0.633 |

*Author's calculations. Abbreviations of the institutions see table 2. - [1]Modified Diebold-Mariano-Statistics according to Harvey, Leybourne, Newbold (1993). – [2]Share of forecast permutations of forecast in row and column showing a higher difference in mean errors than the forecast in the row.*

The second example looks at forecasts of the German economy. Five spring forecasts for the current year are considered, mostly published in April. The sample is 2011 to 2018; its size is delimited by the forecast of the German government, which was made public for the first time in 2011.[4] Thus, the second example approaches the limits of testing, since only 8 observations are available.

Table 4 exhibits some characteristics of these forecasts. Since – different from the U.S. case – the publication date of these forecasts is known, it also shows the average length of the forecast horizon. It is quite similar; the difference between the forecast published earliest and the one

[4] *The Bundesbank as well as the Government have a longer record of forecasts. However, earlier forecasts were only for internal use.*

published latest is only 28 days. Therefore, the information the forecasts are built on should not differ substantially. Again, the mean errors are within a narrow range.

Table 4

**Accuracy measures of the forecasts analyzed**

| Forecaster | Acronym | MAE | MSE | Average Forecast horizon (days) |
|---|---|---|---|---|
| IMF | IMF | 0.47 | 0.35 | 276 |
| Joint Forecast | GD | 0.41 | 0.21 | 263 |
| Bundesregierung | BR | 0.32 | 0.20 | 253 |
| European Commission | EU | 0.32 | 0.19 | 251 |
| Institut der deutschen Wirtschaft | IW | 0.47 | 0.25 | 248 |

*Source: RWI Forecast database; author's calculations.*

Table 5 shows results similar to those in the U.S. case. The DM-test rejects the null only in one case, and only at an error margin of 10%. Both permutation tests reject the null at an error margin of 5% for the IMF forecast relative to the EU forecast as well as to the government's forecast. In both cases the IMF provides the less accurate forecast.

Table 5

**Pairwise tests for equal forecast accuracy of five spring forecasts of German GDP, 2011-2018**

| | | | | |
|---|---|---|---|---|
| | | **Diebold-Mariano-Test, p-values[1]** | | |
| | IMF | EU | IW | BR |
| GD | 0.308 | 0.528 | 0.588 | 0.700 |
| IMF | | 0.121 | 0.573 | **0.096** |
| EU | | | 0.465 | 0.617 |
| IW | | | | 0.557 |
| | | **Permutation test, MAE[2]** | | |
| | IMF | EU | IW | BR |
| GD | 0.910 | 0.242 | 0.727 | 0.313 |
| IMF | | **0.023** | 0.289 | **0.023** |
| EU | | | 0.773 | 0.625 |
| IW | | | | 0.266 |
| | | **Permutation test, MSE[2]** | | |
| | IMF | EU | IW | BR |
| GD | 0.758 | 0.102 | 0.820 | 0.102 |
| IMF | | **0.023** | 0.523 | **0.023** |
| EU | | | 0.938 | 0.688 |
| IW | | | | 0.086 |

*Author's calculations. Abbreviations of the institutions see table 4. - [1]Modified Diebold-Mariano-Statistics according to Harvey, Leybourne, Newbold (1993). – [2]Share of forecast permutations of forecast in row and column showing a higher difference in mean errors than the forecast in the row.*

## 4. Power of the test

As the examples show, the permutation test seems to be more sensitive to differences in forecast accuracy than the DM-test. In the following, it will be analyzed whether this is just an outcome that is specific to the dataset considered, or the observation can be generalized.

A tool to compare the power of statistical tests are power functions as proposed by Butar Butar/Park (2008) and Butar Butar/Bandularesi (2009). The underlying idea is straightforward: two data sets to be compared are randomly drawn from the same distribution, and thereafter one of the datasets is "shocked" by adding a constant. It is then tested whether both datasets have the same mean. Since it is known that this should not be the case, we expect the test to reject this hypothesis.

The simulation design is as follows. In step one, two random datasets are drawn. In step two, a constant $\mu$ is added to all elements in the first dataset to ensure that means differ. In step three, it is tested whether the mean error of the first dataset is larger than that of the second. Steps one to three are repeated m times, counting the share of the m repetitions rejecting the null of equal means at an error margin of $\alpha$. After that, the simulations start again with a larger $\mu$ in the second step. The entire procedure is repeated until $\mu$ is large enough that 100% of the tests reject the null. For positive $\mu$ the power functions relate the complement of the share of simulations making a type II error (i.e. do not reject the null) to the imputed difference $\mu$ in the means of two datasets. The slope of the functions should be positive, and the functions should converge to one for large values of $\mu$.

In Butar Butar/Park (2008) and Butar Butar/Bandularesi (2009) the datasets are drawn randomly from different statistical distributions. In the present case, a normal distribution with zero mean and a standard deviation of one will be used. Both vectors of length T are interpreted as forecast errors. To be able to analyze the impact of different volatilities of the forecast errors the standard deviation of both vectors is rescaled

using the factors 0.5, 1, 1.5 and 2.[5] To evaluate, how the power of the permutation test is influenced by the sample size T, the power functions

are calculated for sample sizes of 8, 10, 12, and 14 years. $\mu$ is increased stepwise by 0.05, and the number of repetitions m is set to 1000. Since the shocked data indeed have a larger mean than the non-shocked, a one-sided test is conducted with $\alpha$ = 5%.

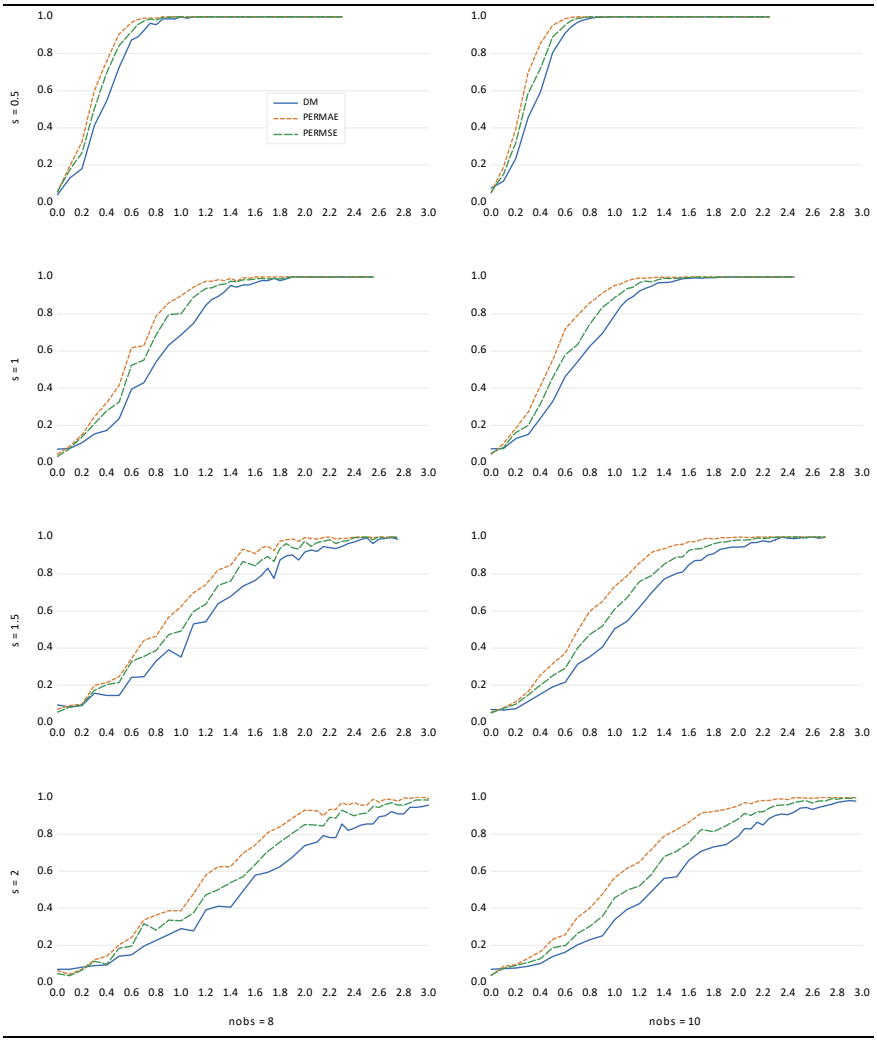
Figure 1 exhibits the power functions of the DM-test as well as of the permutation tests PERMAE and PERMSE. Three aspects become evident. First, having more observations shifts the power function generally to the left, i.e. the power of all test increases. Second, having more volatile errors shifts all power functions to the right in general, i.e. they reduce the power of all three test. Third, the order of the three tests is the same in all simulation experiments and for all shocks $\mu$. The power of PERMAE is highest, PERMSE comes second and the DM-test is the least powerful among these tests.

Table 6 summarizes the results shown in Chart 1. It displays the $\mu$ (rounded to 0.05) at which the margin of 95% of the tests rejecting the null is hit. The figures become smaller in general from the top to the bottom in each block, and they increase from left to right. Comparing the three tests, they are lowest for PERMAE, a bit larger for PERMSE, and largest for DM-test for all combinations of sample size and volatility of the forecast errors.

[5] *Using the root of MSE as an approximation of the standard deviation of forecast errors, 0.5 is about the value observed for spring forecasts of GDP in the current year. A standard deviation of 1 is approximately found in the winter forecast of GDP for the next year. Standard deviations of 1.5 and 2 are typically for forecasts of more volatile variables such as gross fixed capital formation or GDP forecasts over a longer horizon.*

Figure 1

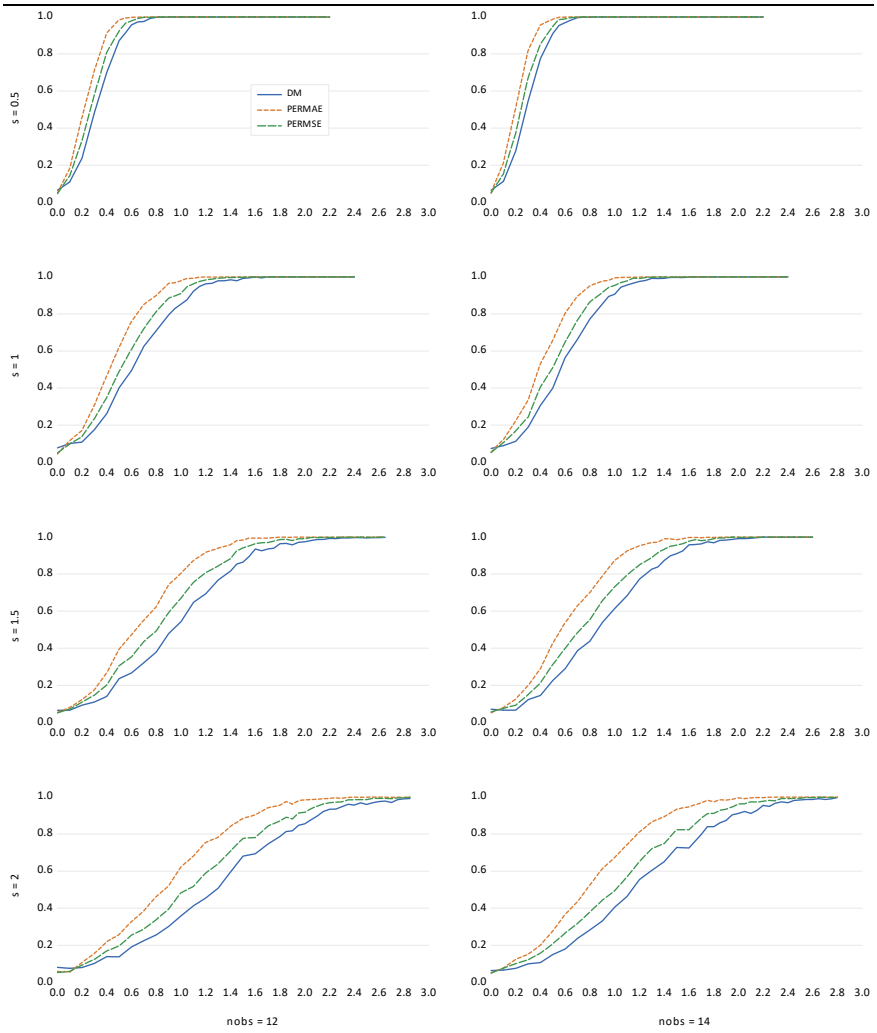
**Power functions for the permutation and the DM test for different standard deviations of errors and sample sizes**




*Author's computations*

Figure 1 (continued)

**Power functions for the permutation and the DM test for different standard deviations of errors and sample sizes**




*Author's computations*

Table 6

**Difference in mean forecast errors at which 95% of tests reject to null[1]**

| Number of observations | Standard deviation of errors | | | |
|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 |
| | Permutation test MAE | | | |
| 8 | 0.60 | 1.20 | 1.80 | 2.30 |
| 10 | 0.50 | 1.00 | 1.50 | 2.00 |
| 12 | 0.50 | 0.90 | 1.40 | 1.80 |
| 14 | 0.40 | 0.80 | 1.20 | 1.70 |
| | Permutation test MSE | | | |
| 8 | 0.65 | 1.30 | 1.85 | 2.65 |
| 10 | 0.60 | 1.20 | 1.75 | 2.30 |
| 12 | 0.55 | 1.10 | 1.55 | 2.10 |
| 14 | 0.55 | 1.00 | 1.45 | 2.00 |
| | DM-test | | | |
| 8 | 0.75 | 1.40 | 2.35 | 3.00 |
| 10 | 0.70 | 1.30 | 2.10 | 2.70 |
| 12 | 0.60 | 1.20 | 1.80 | 2.35 |
| 14 | 0.55 | 1.10 | 1.60 | 2.20 |

*Autor's computations. – [1]at a 5% level (one-sided test)*

Figure 1 also includes the results for $\mu = 0$, i.e. for the case in which the means indeed should be equal. Since all values are different from zero and positive, although small, they demonstrate that there also is a type I error associated with the tests. To assess the power of the tests fairly, the type I error should be considered, too. Table 7 shows the type I error of the simulation experiments. As a general result, type I errors seem to be somewhat smaller for PERMAE and PERMSE than for the DM-test. Concerning the relation to sample size and standard deviations of the errors, however, the simulations do not exhibit clear-cut results.

## 5. Conclusions

For deciding whether one forecast shows smaller errors than another the Diebold-Mariano (DM)-Test has become a widely used tool. It is – as Diebold (2015) clarifies – designed for model-free forecasts, for which macroeconomic forecasts are typical examples. This paper proposes permutation tests for paired observations as an alternative approach to test for equal mean forecast errors. It shows in a series of simulation experiments that permutation tests are more powerful than the DM-test, and that a test statistic based on absolute errors detects differences in forecast accuracy earlier than a test statistic based on squared errors. Furthermore, it is shown that high volatility of the forecast errors and

Table 7

**Type I errors of the tests**

Share of cases with $\mu$=0 in which the null of equal forecast errors is rejected at $\alpha$ =0.05 in per cent

| Number of observations | Standard deviation of errors | | | |
|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 |
| | Permutation test MAE | | | |
| 8 | 5,5 | 4.3 | 7.0 | 6.3 |
| 10 | 5.0 | 4.3 | 5.4 | 3,7 |
| 12 | 4.8 | 4.4 | 5.3 | 5.9 |
| 14 | 5.1 | 5.4 | 5.2 | 5.1 |
| | Permutation test MSE | | | |
| 8 | 5.5 | 3.1 | 5.5 | 4.7 |
| 10 | 5.2 | 4.9 | 5.0 | 3.8 |
| 12 | 4.6 | 4.9 | 5.1 | 5.3 |
| 14 | 4.9 | 5.2 | 5.6 | 4.9 |
| | DM-test | | | |
| 8 | 3.9 | 7.0 | 9.3 | 7.1 |
| 10 | 7.5 | 7.2 | 6.9 | 7.1 |
| 12 | 6.5 | 7.7 | 6.5 | 8.1 |
| 14 | 6.5 | 7.3 | 7.1 | 6.9 |

*Autor's computations.*

small samples reduce the power of all tests. Finally, the type I error margin of all tests is about 5%, with the DM-test showing somewhat larger errors.

These results are derived from simulation experiments with random data. Applied to real forecasts, all tests can be expected to be more powerful, since errors of forecasts for the same economic entity for the same year are intercorrelated empirically. Therefore, the differences in errors of two forecast observed empirically should be smaller than the differences between two random error vectors, and a deviating forecast accuracy might be easier to detect.

Notwithstanding these advantages, the permutation test faces a technical problem: The number of possible permutations increases quadratic to the number of observations. For 20 observations, e.g., more than one million permutations must be calculated, and for 21 observations already more than two million. Therefore, the computation time will increase quadratically to the number of observations. For comparing forecasts over longer periods, Monte Carlo methods may be employed to draw a random sample of n permutation replications (Efron, Tibshirani 1993: 207).

## Literature


Butar Butar, F. and A. Bandularesi (2009), Comparison of the power of paired samples using permutation tests. *Journal of mathematical sciences & mathematical education* 3 (2): 19-30.

Butar Butar, F. and J-W. Park (2008), Permutation tests for comparing two populations. *Journal of mathematical sciences & mathematical education* 4 (2): 19-31.

Coroneo, L., and F. Iacone (2015). *Comparing predictive accuracy in small samples*: Department of Economics and Related Studies, University of York.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Test. *Journal of Business & Economic Statistics*, 33(1), 1-9. doi:10.1080/350015.2014.983236

Diebold, F. X. and R.S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253-263.

Efron, B. and R. J. Tibshirani (1993), An Introcution to the Bootstrap. London: Chapman & Hall.

Gallo, G. M., Granger, C. J., and Yongil, J. (2002). Copycats and Common Swings: The Impact of the Use of Forecasts in Information Sets. *IMF Staff Paper*, 49(1), 4-21.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13, 281-291.

Ludbrook. J and H. Dudley (1998), Why Permutation Tests are Superior to t and F-Test in Biomedical Research. *The American Statistician* 52 (2): 127-132.

Peissendorfer, R.W and T. Barnett (1983), Numerical Model-Reality Intercomparison Test Using Small-Sample Studies. *Journal of the Atmosphere Sciences* 40: 184-1896