The Stata Journal (2009)
9, Number 3, pp. 454–465

Nonparametric testing of distributions—the Epps–Singleton two-sample test using the empirical characteristic function

Sebastian J. Goerg
Max Planck Institute for Research on Collective Goods
Kurt-Schumacher-Straße 10
53113 Bonn, Germany
goerg@coll.mpg.de

Johannes Kaiser
Deutsche Bundesbank[1]
Wilhelm-Epstein-Straße 14
60431 Frankfurt, Germany
johannes.kaiser@bundesbank.de

Abstract. In statistics, two-sample tests are used to determine whether two samples have been drawn from the same population. An example of such a test is the widely used Kolmogorov–Smirnov two-sample test. There are other distribution-free tests that might be applied in similar occasions. In this article, we describe a two-sample omnibus test introduced by Epps and Singleton, which usually has a greater power than the Kolmogorov–Smirnov test although it is distribution free. The superiority of the Epps–Singleton characteristic function test is illustrated in two examples. We compare the two tests and supplement this contribution with a Stata implementation of the omnibus test.

Keywords: st0174, escftest, nonparametric tests, Kolomogorov–Smirnov, Epps–Singleton, two-sample case

1 Introduction

In many empirical scientific fields, statistical tests are used to determine whether two samples have been drawn from the same population. The commonly used procedure is to test the data in question against the null hypothesis, H_0, that the underlying distributions of the two samples are equal. The Kolmogorov–Smirnov two-sample (KS) test, the Wilcoxon–Mann–Whitney rank-sum (MW) test, and the Epps–Singleton (ES) test are examples of this approach. Implementations of the KS and MW tests are included in Stata. In this article, we introduce a Stata implementation of the ES test. The KS and ES tests are able to detect differences in distributions—be it by location, scale, or family. The MW test detects only locational shifts. The reason for this is its directional

---

[1] Any opinions expressed in this paper are those of the authors and do not reflect the opinion of the Bundesbank.

alternative hypothesis, H_1, which states that the underlying distribution of one sample is stochastically larger than the underlying distribution of the other sample.

It has been shown by Epps and Singleton (1986) that the ES test is usually more powerful than the KS test. There exists one more advantage of the ES test over the KS test: An assumption of the KS test is that the data are drawn from a continuous distribution. Contrary to that, both continuous and discrete data may be used for the ES test. This also holds true for the MW test.

In the following, the rationale of the ES test is described. We next explain the syntax of the Stata implementation. Then we apply the tests to two examples and compare the results. Finally, we close with some short conclusions.

2 The ES test

In this section, we give a brief outline of the ES test and concentrate on the important relations and functions. Hereby, we limit our remarks to a description of the procedure and leave out details on proofs and derivations. The interested reader can find these details in the original paper by Epps and Singleton (1986).

The *p*-value of the ES test gives the probability of falsely rejecting H_0 that both samples have been drawn from the same population. It tests for dissimilarities by comparing the empirical characteristic functions, $\phi_1(t)$ and $\phi_2(t)$, of the two samples instead of the observed distributions, F_1 and F_2.

The empirical characteristic function is the Fourier transform of the observed distribution function. The characteristic function of a distribution can be used to conveniently derive its moments and thus contains more information than just one measure, like the mean, the median, or the variance. However, this also holds true for the probability density. Additionally, the use of the probability density is more intuitive than the use of the characteristic function. Epps (1993) describes the geometrical representation of the characteristic function as the center of mass of a distribution wrapped around the unit circle in the complex plane. These caveats raise doubts on the necessity of applying them: Why should one use the empirical characteristic function for statistical tests?

One advantage of the characteristic function is that it can be used as a representation of distributions whose probability densities cannot be specified. One example is the family of alpha-stable distributions introduced by Paul Lévy, where only three distributions (Gaussian, Cauchy, and Lévy) in closed form for densities are known. Typical applications for distributions whose forms are not closed are models with returns from stock markets (Epps 1993; Borak, Härdle, and Weron 2005).

Another advantage, and more relevant here, is that the characteristic function is completely defined for discrete and continuous data, while the distribution function is completely defined only for continuous data. For discrete data, the distribution function is defined only in certain points.

An important prerequisite for the application of the test is that all observations are independent, both within and across samples. The null hypothesis of the test states

$$H_0 : \phi_1(t) = \phi_2(t), \text{with} -\infty < t < \infty$$

The characteristic function is defined as $\phi_k(t) = \int_{-\infty}^{\infty} e^{itx} dF_{n_k}(x)$, where t is a real number and $i = \sqrt{-1}$. For a sample, k, with a size of n_k, with X_{km} denoting the mth observation in sample k, and a distribution function $F_{n_k}(x)$, the empirical characteristic function is defined as

$$\phi_{n_k}(t) = \int_{-\infty}^{\infty} e^{itx} dF_{n_k}(x) = n_k^{-1} \sum_{m=1}^{n_k} e^{itX_{km}}$$

To make use of the characteristic function for the ES test, a set of parameters $t_1, t_2, \ldots, t_J$ has to be chosen. For the sake of applicability, these parameters need to be calibrated to provide the test with a sufficient power against a broad class of alternatives. Epps and Singleton (1986) did simulations with nine different families of distributions[2] in 30 samples altogether. They found that with $t_1 = 0.4$ and $t_2 = 0.8$ ($J = 2$), the test performed optimally, conditional on their sample of 30 comparisons. In the following, we will briefly summarize the proceedings as described by Epps and Singleton (1986). For a more exhaustive description of the calibration, refer to their work.

The t_j need to be standardized with an estimate of scale $\widehat{\sigma}$—Epps and Singleton (1986) claim that a sufficiently good scale measure for $\widehat{\sigma}$ is the semi-interquartile range. As a consequence, the test is carried out with $\widetilde{t_j} = t_j/\widehat{\sigma}, j = 1, 2$.

For each X_{km}, a 4×1 vector $g(X_{km})$ is created:

$$g(X_{km}) = (\cos\, t_1 X_{km}, \sin\, t_1 X_{km}, \cos\, t_2 X_{km}, \sin\, t_2 X_{km})'$$

Let g_k contain the real and imaginary parts of the characteristic function of the sample for both t_1 and t_2:

$$g_k = n_k^{-1} \sum_{m=1}^{n_k} g(X_{km})$$

Let $G_2 = g_1 - g_2$ be the difference between both vectors. If H_0 was true, $\sqrt{n_1 + n_2} G_2$ would be distributed asymptotically as multivariate $N(\vec{0}, \Omega)$. Epps and Singleton derive an estimator for the covariance matrix Ω. Let $\nu_k = n_k/(n_1 + n_2)$ be the share of sample k in the combined sample and

$$\widehat{S}_k = \frac{n_k - 1}{n_k} \text{cov}\{g(X_{km})\}$$

be the sample covariance matrix of sample k. A sufficient estimator for Ω would then be

$$\widehat{\Omega} = \frac{1}{\nu_1} \widehat{S}_1 + \frac{1}{\nu_2} \widehat{S}_2$$

---

[2] They chose normal, uniform, Cauchy, Laplace, symmetric stable, gamma, Poisson, binomial, and negative binomial distributions.

The test statistic of the ES test is defined as $W_2 = (n_1 + n_2) \cdot G_2' \cdot \widehat{\Omega}^+ \cdot G_2$ with $\widehat{\Omega}^+$ being the generalized inverse of $\widehat{\Omega}$. W_2 is distributed asymptotically as chi-squared with r degrees of freedom, where r denotes the rank of $\widehat{\Omega}^+$. This is how the p-level of the test can be computed. Roughly spoken, W_2 is a measure for the statistical distance between the empirical characteristic functions of both samples standardized by the variance–covariance matrices, with the characteristic functions being descriptors for the distributions underlying the two samples in question.

If the sample size of both observations is small, Epps and Singleton suggest to use a small-sample correction factor, $\widehat{C}(n_1, n_2)$. They conducted simulations and concluded that W_2 can be excessive for small n_k. Hence, if each one of the two samples includes less than 25 observations, a factor of

$$\widehat{C}(n_1, n_2) = \left\{1 + (n_1 + n_2)^{-0.45} + 10.1\left(n_1^{-1.7} + n_2^{-1.7}\right)\right\}^{-1}$$

should be applied on the test statistic W_2. The idea behind $\widehat{C}$ was to find a transformation $T(W_2; n_1, n_2) = C(n_1, n_2) \cdot W_2$ fulfilling $\sup P\{T(W_2; n_1, n_2) \geq \chi^2_\alpha\} \leq \alpha$, with χ^2_α being the $1 - \alpha$ percentile of the χ^2 distribution with four degrees of freedom. Epps and Singleton estimated the highest value of $C(n_1, n_2)$ in 1,000-trial simulations with different α's and sample sizes. The parameters of the correction factor $\widehat{C}$ were estimated to minimize the error $C(n_1, n_2) - \widehat{C}(n_1, n_2)$.

Epps and Singleton compared their test with the Anderson–Darling, the Cramér–von Mises, and the KS tests by means of computational simulations and came to the following conclusions:

- If discrete data are used, apply the ES test.
- If continuous data are used, the KS test usually has a lower power than the ES test.
- Sometimes, the Anderson–Darling and the Cramér–von Mises tests can have a higher power than the ES test.

3 The escftest command

3.1 Description

We include with this article a Stata implementation of the ES test in the program `escftest`. After installation, the new commands `escftest` and `help escftest` are available. In the algorithm described above, both matrix and vector operations are used. We used a Mata function in the code to accomplish these calculations. The reader should be aware that Mata was introduced to the Stata software package in version 9, so the command will refuse to work in versions earlier than 9.

3.2 Syntax

The syntax of the command to execute the ES characteristic function test is

`escftest` *varname* [*if*] [*in*], `group(`*groupvar*`)` [`t1(`#`)` `t2(`#`)`]

varname specifies the variable to test.

3.3 Options

`group(`*groupvar*`)` is required. It specifies the grouping variable. There must be exactly two different groups in the specified sample.

`t1(`#`)` specifies the parameter t_1 as defined by Epps and Singleton (1986). In this paper, details on this parameter are given in section 2. If omitted, the default is `t1(0.4)`. It should not be necessary to specify `t1()`.

`t2(`#`)` specifies the parameter t_2 as defined by Epps and Singleton (1986). In this paper, details on this parameter are given in section 2. If omitted, the default is `t2(0.8)`. It should not be necessary to specify `t2()`.

3.4 Saved results

Normally, it should not be necessary to modify `t1()` or `t2()`. These parameters should be modified only if one wants to calibrate the test for a specific task. `escftest` saves some of the results of the performed test in `r()`:

Scalars	
`r(crit_val_1)`	the critical value for the test statistic W_2 at a significance level of 0.01
`r(crit_val_5)`	the critical value for the test statistic W_2 at a significance level of 0.05
`r(crit_val_10)`	the critical value for the test statistic W_2 at a significance level of 0.1
`r(p_val)`	the p-value associated with the actual test statistic W_2
`r(correction)`	the small-sample correction factor, C (if applied)
`r(t1)`	the value used for t_1 in the empirical characteristic function
`r(t2)`	the value used for t_2 in the empirical characteristic function
Macros	
`r(group1)`	value of the grouping variable for the first group
`r(group2)`	value of the grouping variable for the second group

4 Some applications

In this section, we compute two examples with the tests mentioned above. The first application refers to the numerical example from Epps and Singleton (1986); the data are taken from a study by Delse and Feather (1968). In this study, the ability of two groups to control salivation is compared; one group receives a biofeedback stimulus and the other group does not. The second example is taken from the field of experimental economics and applies an intercultural methodology introduced by Goerg and Walkowitz (2008) on Chinese and Germans.

First, we take a glance at the example described by Delse and Feather (1968). They investigate the effect of letting subjects hear a salivation signal and try to control their salivation. For the study, 20 subjects were equally distributed in two groups. Each subject was told to try to increase his salivation rate when observing a light signal on the left side and to decrease it when observing a light signal on the right side. In the experiment, one of the two groups received a biofeedback stimulus in terms of a tone (1,000 cycles per second, 0.2 seconds) for each saliva drop collected by a special apparatus. The other group did not receive such feedback. The data collected are shown in the table in figure 1. Each observation represents the difference between the mean number of saliva drops over 13 increase signals and the mean number of drops over 13 decrease signals. The data are taken from Hollander and Wolfe (1999, 180).[3] The quantile–quantile plot in figure 1 already reveals that the data of the two groups are not identically distributed.

Biofeedback group	Control group
-1.15	-0.35
-0.15	2.55
2.48	1.73
3.25	0.73
3.71	0.35
4.29	2.69
5.00	0.46
7.74	-0.94
8.38	-0.37
8.60	12.07

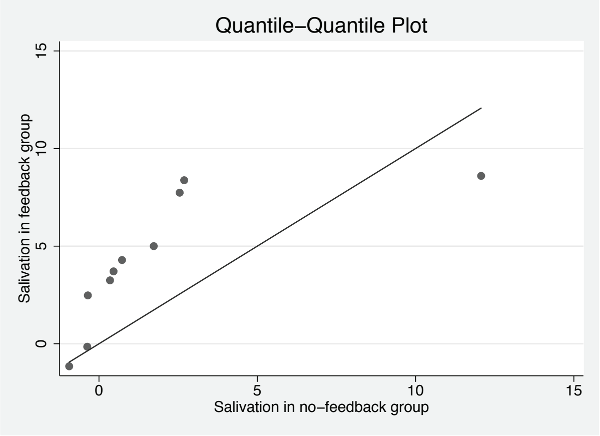

Figure 1. Values and quantile–quantile plot of the study by (Delse and Feather 1968)

Before taking a look at a comparison of the results of the ES test with the results of the KS test, we would like to mention that the numerical example from section 5 of Epps and Singleton (1986) contains an error that is either a simple typing error or a programming error: On page 202, the scale measure $\widehat{\sigma}$ for standardizing t_j is stated to be 1.95. This is not correct. If one calculates $\widehat{\sigma}$ by hand, it becomes clear that this value should be 2.05. Christian Rojas,[4] who did some research on the ES test, arrived at the same conclusion. Nevertheless, the result of the numerical example is correct.

The variable `salivationDF` gives the participant's mean change rate of salivation from the Delse and Feather study. The variable `groupDF` defines the two subject groups in the study: group one with the biofeedback stimulus and group two without it. Both groups consist of 10 participants. Let's take a look at the test results:

---

[3] Epps and Singleton take the data from an earlier edition of the same book.

[4] See http://www.umass.edu/resec/faculty/rojas/index.shtml.

```
. escftest salivationDF, group(groupDF)
Epps-Singleton Two-Sample Empirical Characteristic Function test
Sample sizes: groupDF = 1           10
              groupDF = 2           10
              total                 20
t1                               0.400
t2                               0.800
Critical value for W2 at 10%     7.779
                          5%     9.488
                          1%    13.277
Test statistic W2               15.141
Ho: distributions are identical
P-value                        0.00442
Note: a small sample correction factor of C(10,10) = 0.60140 has been applied
to W2.
```

The ES test gives the necessary values of the test statistic W_2 for significance at 10%, 5%, and 1%. In this example, the test statistic $W_2 = 15.141$ totals to a value much higher than the necessary 13.277 for significance at the 1% level: the p-level is at 0.44%. A small-sample correction factor is applied because both observations are smaller than 25.

```
. ksmirnov salivationDF, by(groupDF) exact
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group       D       P-value      Exact
 ----------------------------------------------
 1:                0.1000    0.905
 2:               -0.6000    0.027
 Combined K-S:     0.6000    0.055       0.035
```

Because of the small sample size, we apply `ksmirnov, exact`. The KS test gives the p-value for the one-sided comparison, once with a smaller group 1 and once with a smaller group 2. The combined value gives the exact p-value for the two-sided comparison. H_0 is rejected at a level of 5.5%. This is a much weaker significance level than the one for the ES test.

The second example is from the field of experimental economics.[5] A popular research question in this field is the comparison of economic behavior across different populations and decision conditions. Typical characteristics of data obtained by economic experiments are relatively small sample sizes and often the discreteness of attributes. The last point forbids the application of the KS test. Thus the question of whether behavior between subject groups differs and by what means is normally determined by the MW test. In contrast to the ES test, the MW test has a directional alternative hypothesis, H_1, which is that one sample is stochastically larger than the other. On one hand, if significant results are obtained by the MW test, they include more information than results from the ES test. On the other hand, if no sample is stochastically larger, the MW test finds no differences. The following example, where the KS test is not applicable,

---

[5] In contrast to experiments in psychology, participants in experiments by economists receive a payoff that is determined by the decisions made in the experiment. This is done to ensure monetary incentives, which economists are interested in.

illustrates this limit of the MW test and the advantage of the ES test. The features of data gathered by economic experiments, described above, make the ES test a valuable tool for this research area where it is casually applied (for example, Henrich [2000], Eckel and Grossman [1998], and Hoffman, McCabe, and Smith [1996]).

In the experiment by Goerg and Walkowitz (2008), the cooperative behavior of participants from different countries is compared. Participants received an initial endowment of 10 Talers.[6] Two matched participants had to decide simultaneously and anonymously whether to send a part of their initial endowments to the matched player. The transfer amount had to be an integer between 0 and 10. This transferred amount reached the matched player doubled. The total payoff for the participant was his or her initial endowment minus the amount sent to the other player plus the doubled amount sent from the other player.

A participant who tries to maximize his own payoff would transfer nothing and hope that the matched player would send something to him. A player who wants to maximize the collective payoff would send everything and expect the matched player to transfer everything, too. Thus transferring nothing is understood as no cooperation, transferring something is understood as gradual cooperation, and transferring everything is understood as full cooperation. The method is introduced in more detail in Goerg and Walkowitz (2008), where it is applied on participants from Israel and Palestine.

The new and yet unpublished data that is discussed here contains the choices of 20 participants in China and 20 participants in Germany. The variable `cooperation` contains the transferred amount between 0 and 10, and the variable `country` defines the two groups.

---

[6] A fictional currency used in the experiment, with a fixed exchange rate to Euros.

Cooperation in Germany	Cooperation in China
0	10
1	4
2	0
2	5
2	10
2	10
3	0
3	5
3	6
3	7
4	10
5	3
5	1
5	7
5	0
6	8
10	1
10	5
10	8
10	10

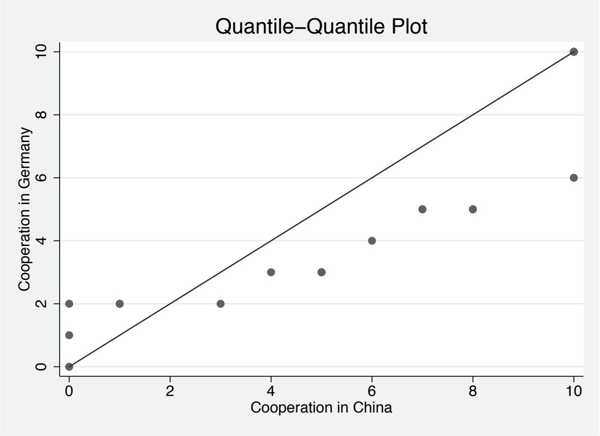

Figure 2. Cooperation in China and Germany

The quantile–quantile plot in figure 2 reveals differences between the two samples. Recall that the participants could choose only integer numbers as transfer amounts. The discreteness of the observed attribute rules out the application of the KS test to the data. We will search for quantitative support of this qualitative result by applying the MW test and the ES test. Let's start with the MW test:

```
. ranksum cooperation, by(country)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
     country |      obs    rank sum    expected
-------------+---------------------------------
           C |       20       441.5         410
           G |       20       378.5         410
-------------+---------------------------------
    combined |       40         820         820
unadjusted variance     1366.67
adjustment for ties      -28.72
                      ----------
adjusted variance       1337.95
Ho: cooper~n(country==C) = cooper~n(country==G)
             z =   0.861
    Prob > |z| =   0.3891
```

The two-sided rank-sum test reveals no significant difference ($p = 0.3891$) between the behavior in the two countries. The differences revealed by the quantile–quantile plot are of a kind that the MW test is not capable of showing. In contrast to this, the ES test detects more types of deviations than does the MW test. Thus the ES test leads to a different result:

```
. escftest cooperation, group(country)
Epps-Singleton Two-Sample Empirical Characteristic Function test
Sample sizes: country = C           20
              country = G           20
              total                 40
t1                               0.400
t2                               0.800
Critical value for W2 at 10%     7.779
                          5%     9.488
                          1%    13.277
Test statistic W2                8.900
Ho: distributions are identical
P-value                        0.06364
Note: a small sample correction factor of C(20,20) = 0.76092 has been applied
to W2.
```

The ES test finds a significant difference between the distributions of behavior in the two countries, with a *p*-value of 0.0636. Obviously, the distribution of cooperative behavior in the two populations (participants in Germany and participants in China) differs. In both countries, the experimental conditions were kept identical regarding stakes, incentives, and distributions of demographic attributes among the participants. Thus the observed differences are most likely implied by the different cultural backgrounds.

The rank-sum test could not detect differences between participants from the two countries. This example impressively demonstrates the importance of the ES test for situations where discrete data are investigated, and these situations frequently occur in the field of experimental economics. While the MW test captures only central tendencies, the ES test can capture distributional characteristics.

5 Conclusions

In this article, we briefly described a powerful alternative to the Kolmogorov–Smirnov two-sample test and a complement to the Wilcoxon–Mann–Whitney rank-sum test, namely, the Epps–Singleton characteristic function test. We explained the use of the Stata implementation and applied the tests on two examples. The first example compared the *p*-levels of the KS test with those of the ES test and showed that the *p*-level of the ES test is far better. The second example showed a situation where the KS test cannot be applied and the MW test does not lead to significant results.

We provide the community with a Stata implementation of the ES test and hope that it might be of use. There is still room for future work; neither the Cramér–von Mises nor the Anderson–Darling two-sample test has been introduced to Stata so far (the Anderson–Darling goodness-of-fit test has already been adopted to Stata by Royston [1996]).

6 Acknowledgments

We thank the editor and the anonymous referee for speed improvements in the supplemented program code and for very valuable comments and suggestions. Furthermore, we thank Christian Rojas for discussions with us and for providing a Matlab implementation of the ES test.

7 References

Borak, S., W. Härdle, and R. Weron. 2005. Stable distributions. SFB 649 Discussion Paper 2005-008, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin. http://141.20.100.9/papers/pdf/SFB649DP2005-008.pdf.

Delse, F. C., and B. W. Feather. 1968. The effect of augmented sensory feedback on the control of salivation. *Psychophysiology* 5: 15–21.

Eckel, C. C., and P. J. Grossman. 1998. Are women less selfish than men?: Evidence from dictator experiments. *Economic Journal* 108: 726–735.

Epps, T. W. 1993. Characteristic functions and their empirical counterparts: Geometrical interpretations and applications to statistical inference. *American Statistician* 47: 33–38.

Epps, T. W., and K. J. Singleton. 1986. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation* 26: 177–203.

Goerg, S. J., and G. Walkowitz. 2008. Presentation effects in cross-cultural experiments: An experimental framework for comparisons. Discussion Paper No. 4/2008, Bonn Econ Discussion Papers, Bonn Graduate School of Economics, University of Bonn. http://econpapers.repec.org/paper/bonbonedp/bgse4_5f2008.htm.

Henrich, J. 2000. Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* 90: 973–979.

Hoffman, E., K. A. McCabe, and V. L. Smith. 1996. On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* 25: 289–301.

Hollander, M., and D. A. Wolfe. 1999. *Nonparametric Statistical Methods*. 2nd ed. New York: Wiley.

Royston, P. 1996. sg47: A plot and a test for the χ^2 distribution. *Stata Technical Bulletin* 29: 26–27. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 142–144. College Station, TX: Stata Press.

About the authors

Sebastian Goerg holds a master's degree in economics. His main areas of research include experimental economics, algorithmic learning rules, and intercultural behavioral economics.

Johannes Kaiser holds a master's degree in business and engineering from the University of Karlsruhe and obtained a doctorate in economics from the University of Bonn. His research interests include behavioral and experimental economics, behavioral finance, and nonparametric statistics.