Goodness of fit to the family of distribution.

Abstract of Ph.D. Theses

By Éva Osztényiné Krauczi
Supervisor:
Professor Sándor Csörgő
Consultants:
Professor Gyula Pap and Professor Gábor Szűcs

Doctoral School in Mathematics and Computer Science
Bolyai Institute, University of Szeged
Szeged, 2016

1 Introduction

The thesis is devoted to the study of goodness of fit in the case of various distributions. Let $X_1, \ldots, X_n$ be a sample (independent identically distriduted random variables) from an unknown distribution with distribution function F. The simple hypothesis is

$$\mathcal{H}_0 : F = F_0,$$

where F_0 is a given distribution function, and the composite hypothesis is

$$\mathcal{H}_0 : F \in \mathcal{F},$$

where $\mathcal{F}$ denotes the family of probability distributions.

In Chapter 2 we collect the historical preliminaries using the comprehensive summary provided by del Barrio, Cuesta-Albertos and Matrán [9]. For the overview we recall the first tests which are suitable for goodness of fit to a fixed distribution paying special attention to the development of the asymptotic theory of goodness of fit tests. The goodness of fit to the family of distributions and their asymptotic theories are considered focusing to two classes of this procedure: tests of fit based on the empirical distribution function (EDF), and the regression and correlation tests of fit.

In Chapter 3 we suggest a goodness of fit procedure to the uniform distribution on $[0, 1]$ and to the uniform family. The idea is the following: consider a random uniform sample on $[0, 1]$, let the simple size be n. Moreover, there is a given deterministic distance level $d_n \in (0, 1)$ for all n. We push through this distance level on $[0, 1]$ and we observe how many nonempty disjoint classes breaks up the elements of the order statistics into. The elements of the order statistics belong to the same class, where the distance between any two neighbouring elements is not greater than d_n. The classes belong to a given sample at a given distance level is called the number of clusters. S. Csörgő and Wu showed that the number of clusters is asymptotically normal for three different distance level sequences. We extend the results of S. Csörgő and Wu [6] to multivariate limit theorems for uniform distributions on different intervals. These theorems are applied for testing uniformity on a known and an unknown interval. We prove that the joint cluster count vector is asymptotically normal in all cases. Thus, these tests define asymptotically χ^2 tests for a uniform distribution or for the uniform family. We simulated powers of the new tests as well.

In Chapter 4 we investigate a goodness of fit test to the normal family, based on the L^2-Wasserstein distance, proposed by del Barrio, Cuesta-Albertos, Matrán and Rodrí- guez-Rodríguez [10]. They obtained the location- and scale-free test statistic for the null hypothesis $\mathcal{H}_0 : F \in \mathbf{N}$, where $\mathbf{N}$ denotes the normal family. A simulation study was performed to evaluate the power of the BCMR-test and to make comparisons with other tests of normality.

In Chapter 5 we present the weighted version of the quantile correlation test for goodness of fit to the logistic family, introduced by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [10], and del Barrio, Cuesta-Albertos and Matrán [9]. The use of weight functions in the test statistics were suggested independently from each

other by de Wet in [7] and [8] and by S. Csörgő [4] and [5]. We prove the results of S. Csörgő [5] for location and scale logistic family with the weight function for location family suggested by de Wet. Del Barrio, Cuesta-Albertos and Matrán [9] obtained the asymptotic distribution as the Karhunen–Loève expansion of the weighted Brownian-bridge. With the same technique we determine the infinite series representation of our limiting distribution. Similarly to the previous results a simulation study was performed to evaluate the power of the tests.

The author has written three papers on the subject of the thesis. Joint cluster counts from uniform distribution is published in Krauczi [16]. Krauczi [14] contains the study of the quantile correlation test for normality. Finally, the results of the weighted quantile correlation test for the logistic family are from Balogh and Krauczi [2].

All convergence relations are understood throughout the thesis as $n \to \infty$, and let $\to_{\mathcal{D}}$ denote convergence in distribution and let $\to_{\mathbf{P}}$ denote convergence in probability.

2 Historical preliminaries

As an overview we recall the first tests which are suitable for goodness of fit to a fixed distribution paying special attention to the development of the asymptotic theory of goodness of fit tests. The first goodness of fit procedure is the χ^2-test proposed by Pearson [17]. Under the null-hypothesis, this test has asymptotic distribution χ^2. The EDF-tests and the recovery of their asymptotic distribution have received special attention. These tests use different functional distances to measure the discrepancy between the hypothesized distribution function and the empirical distribution function. A section is devoted to the problem of the goodness of fit to the family of distributions and their asymptotic theories. The first studies are occurred in the most interesting case, for the Gaussian family. Then we adapt all the procedures considered in the first subsection for the case of the parametric family. The simple idea is choosing an adequate estimator of the parameter and replacing the fixed distribution by the distribution with the estimated parameter. Finally we recall the regression and correlation tests, the very popular Wilk–Shapiro-test of normality [19] and it's further modifications. The asymptotic result are also considered.

3 Goodness of fit to the uniform family

Introduction and preliminary results

We suggest a goodness of fit procedure to the uniform distribution on $[0, 1]$ and to the uniform family. The idea is the following: let $U_1, \ldots, U_n$ be a random uniform sample (independent uniformly distributed on $[0, 1]$ random variables). Moreover, there is a given deterministic distance level $d_n \in (0, 1)$ for all n. We push through this distance level on $[0, 1]$ and we observe how many nonempty disjoint classes breaks up the elements of the order statistics into. The elements of the order statistics belong to the same class, where the distance between any two neighbouring elements is not

greater than d_n. The classes belong to a given sample at a given distance level is called the number of clusters and denoted $K_n(d_n)$.

We recall that S. Csörgő and Wu [6] showed that the number of clusters is asymptotically normal for three different distance level sequences.

Theorem 1 (Csörgő and Wu [6]) *(i) If $nd_n \to 0$ and $n^2 d_n \to \infty$, then*

$$\begin{aligned}\Delta_n &= \sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}(1-e^{-nd_n})}} \le x \right) - \Phi(x) \right| \\ &= O\left(\sqrt{(nd_n + \varepsilon_n) \log \frac{1}{nd_n}} + \frac{\log(n\sqrt{d_n})}{n\sqrt{d_n}} \right),\end{aligned}$$

where $\varepsilon_n = \sqrt{(4\log n)/n}$, and so $(K_n - ne^{-nd_n})/(n\sqrt{d_n}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

(ii) If $0 < \liminf_n nd_n \le \limsup_n nd_n < \infty$, then

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-2nd_n}(e^{nd_n} - 1 - n^2 d_n^2)}} \le x \right) - \Phi(x) \right| = O\left(\frac{\log^{3/4} n}{n^{1/4}} \right),$$

and hence if $nd_n \to c \in (0,\infty)$, then $(K_n - ne^{-nd_n})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma)$, where $\sigma = e^{-2c}[e^c - 1 - c^2]$.

(iii) If $nd_n \to \infty$ and $ne^{-nd_n} \to \infty$, then

$$\Delta_n = O\left(\frac{(nd_n)^{3/2}}{\sqrt{e^{nd_n}}} + \sqrt{\varepsilon_n nd_n \log(ne^{-nd_n})} + \sqrt{\frac{e^{nd_n}}{n}} \log(ne^{-nd_n}) \right),$$

where Δ_n is as in the case (i) and $\varepsilon_n = \sqrt{(4\log n)/n}$ again, and so

$$\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

We extend the results of S. Csörgő and Wu [6] to multivariate limit theorems for uniform distributions on different intervals. These theorems are applied for testing uniformity on a known and an unknown intervals.

Theoretical results

We investigate the joint behaviour of K_n's for sequences of different distance levels. Set $J \ge 1$ and let $d_{n1} \le d_{n2} \le \cdots \le d_{nJ}$, $n \in \mathbb{N}$, be distance levels. If the sample comes from the uniform distribution on the unit interval $[0,1]$, then $K_{nj}(d_{nj})$ denote the numbers of clusters corresponding to the distance levels d_{nj} for all n and j. Consider the random vector

$$\mathbf{K}_n = \frac{1}{\sqrt{n}} \left(\frac{K_{n1}(d_{n1}) - m_{n1}}{\sigma_{n1}}, ..., \frac{K_{nJ}(d_{nJ}) - m_{nJ}}{\sigma_{nJ}} \right)^\top, \qquad n \in \mathbb{N}, \tag{1}$$

with the sequences $m_{nj} = ne^{-nd_{nj}}$ and

$$\sigma_{nj}^2 = e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2), \qquad j = 1, \ldots, J.$$

Theorem 2 *Let $d_{n1} \le d_{n2} \le \cdots \le d_{nJ}$, $n \in \mathbb{N}$, be distance levels satisfying one of the following conditions:*

(T1) $nd_{nj} \to 0$, $n^2 d_{nj} \to \infty$;

(T2) $0 < \liminf_n nd_{nj} \le \limsup_n nd_{nj} < \infty$;

(T3) $nd_{nj} \to \infty$, $ne^{-nd_{nj}} \to \infty$.

Moreover, assuming additionally

$$s_{ij} := \lim_{n\to\infty} \frac{e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni} d_{nj})}{\sigma_{ni}\sigma_{nj}} \in \mathbb{R}, \qquad 1 \le i < j \le J, \tag{2}$$

and let $s_{jj} := 1$ and $s_{ji} := s_{ij}$. Then

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma) \tag{3}$$

with the covariance matrix $\Sigma = (s_{ij})_{i,j=1,\dots,n}$.

One of the corollary of this theorem is that the limiting distribution can be obtained with the block diagonal matrix Σ.

Corollary 3 *Suppose $J \ge 2$, and $0 \le J_1 \le J_2 \le J$ are such, that distance levels d_{nj} satisfy condition (T1) for $j \le J_1$ and condition (T3) for $j > J_2$. Moreover, assume additionally the following conditions:*

(i) For $i < j \le J_1$ it holds $s_{ij} := \lim_{n\to\infty} \sqrt{d_{ni}/d_{nj}} \in \mathbb{R}$.

(ii) For $J_1 < j \le J_2$ $c_j := \lim_{n\to\infty} nd_{nj} \in \mathbb{R}$ exists. Then for $J_1 < i < j \le J_2$ it holds

$$s_{ij} := \frac{(e^{c_i} - 1 - c_i c_j)}{\sqrt{(e^{c_i} - 1 - c_i^2)(e^{c_j} - 1 - c_j^2)}}.$$

(iii) For $J_2 < i < j$ $s_{ij} := \lim_{n\to\infty} e^{-n(d_{nj}-d_{ni})/2} \in \mathbb{R}$ also exists.

And let $s_{ji} := s_{ij}$ and $s_{jj} := 1$. Then (3) holds with the block diagonal matrix

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix},$$

where the dimensions of blocks Σ_1, Σ_2 and Σ_3 are the respectively $J_1 \times J_1$, $(J_2 - J_1) \times (J_2 - J_1)$ and $(J - J_2) \times (J - J_2)$. The components of blocks are defined with s_{ij}'s above.

We apply Corollary 3 to typical sequences suggested by S. Csörgő and Wu in [6], hereby we choose the parameters to obtain a diagonal covariance matrix. S. Csörgő and Wu give well-behaving examples called typical sequences. A typical sequence $(d_n)_{n=1,2\dots}$ for the case (T1) is $d_n = n^{-\alpha}$ for some $\alpha \in (1, 2)$. In particular, we take $d_{nj} = n^{-\alpha_j}$ for $j \le J_1$, with $\alpha_1 > \alpha_2 > \cdots > \alpha_{J_1}$ resulting in $s_{ij} = 0$ for $i < j \le J_1$. In the case (T2) the existence of the limit $c := \lim_{n\to\infty} nd_n \in \mathbb{R}$ gives the typical sequence $(d_n)_{n=1,2\dots}$. Here let $0 \le J_2 - J_1 \le 2$ which means that Σ_2 is at most a 2×2 matrix and take $c_{J_2} = (e^{c_{J_1+1}} - 1)/c_{J_1+1}$ in the case $J_2 - J_1 = 2$. A typical sequence $(d_n)_{n=1,2\dots}$ for the case (T3) is $d_n = \beta(\log n)/n$ for some $\beta \in (0, 1)$. So we take $d_{nj} = \beta_j(\log n)/n$ for $j > J_2$, with $\beta_i < \beta_j$ for $J_2 < i < j$ resulting again $s_{ij} = 0$. Under these special choices Corollary 3 reduces to the following one.

Corollary 4

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, E_J),$$

where E_J denotes the unit matrix with dimension J.

If the sample comes from the uniform distribution on the known interval $[a, b]$ with $a, b \in \mathbb{R},\ a < b$, then we prove with applying a linear transformation of the interval $[a, b]$ onto the interval $[0, 1]$, that the transformed cluster count vector is also asymptotically normal distributed under the correctly transformed assumptions.

Let $V_1, V_2, \ldots, V_n$ be independent random variables, each uniformly distributed on the interval $[a, b]$ with $a, b \in \mathbb{R},\ a < b$, known. Set $J \geq 1$ and let $d_{n1} \leq d_{n2} \leq \cdots \leq d_{nJ}$ be distance levels. Let $K_{nj}^{a,b}(d_{nj})$ be numbers of clusters corresponding to the distance levels d_{nj}, $j = 1, \ldots, J$. Set

$$m_{nj}^{a,b} = ne^{-\frac{nd_{nj}}{b-a}}, \qquad \sigma_{nj}^{a,b} = \sqrt{e^{-2\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{nj}}{b-a}} - 1 - \left(\frac{nd_{nj}}{b-a}\right)^2\right)},$$

and

$$\mathbf{K}_n^{a,b} = \frac{1}{\sqrt{n}}\left(\frac{K_{n1}^{a,b}(d_{n1}) - m_{n1}^{a,b}}{\sigma_{n1}^{a,b}}, \ldots, \frac{K_{nJ}^{a,b}(d_{nJ}) - m_{nJ}^{a,b}}{\sigma_{nJ}^{a,b}}\right)^\top.$$

Theorem 5 *Suppose each d_{nj} satisfies one of the conditions (T1), (T2) or (T3'), where*

(T3') $nd_{nj} \to \infty$, $ne^{-\frac{nd_{nj}}{b-a}} \to \infty$.

In addition, s_{ij}'s exist for which

$$e^{-\frac{nd_{ni}}{b-a}-\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{ni}}{b-a}} - 1 - \frac{nd_{ni}}{b-a}\frac{nd_{nj}}{b-a}\right)/\sigma_{ni}^{a,b}\sigma_{nj}^{a,b} \to s_{ij}, \qquad 1 \leq i < j \leq J, \tag{4}$$

and let $s_{ii} := 1$ and $s_{ji} := s_{ij}$. Then it holds

$$\mathbf{K}_n^{a,b} \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma) \tag{5}$$

with the covariance matrix $\Sigma = (s_{ij})_{i,j=1,\ldots,J}$.

Finally, the sample comes from the uniform distribution on the unknown interval. Let $V_1, \ldots, V_n$ be independent, uniformly distributed random variables on the interval $[a, b]$ with $a < b$ being unknown and let $V_{1,n}, \ldots, V_{n,n}$ be the ordered sample. We investigated a counterpart of Theorems 2 and 5 when the endpoints of the interval are estimated by $\hat{a}_n = V_{1,n}$ and $\hat{b}_n = V_{n,n}$. In an analogue to the previous notations, for given $J \geq 1$ and distance levels $d_{n1} < \cdots < d_{nJ}$ denote $\hat{K}_{nj}(d_{nj})$ numbers of clusters corresponding to the distance levels d_{nj}, $j = 1, \ldots, J$, set

$$\hat{m}_{nj} = ne^{-\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}}, \qquad \hat{\sigma}_{nj} = \sqrt{e^{-2\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}}\left(e^{\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}} - 1 - \left(\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}\right)^2\right)}$$

and

$$\widehat{\mathbf{K}}_n = \frac{1}{\sqrt{n}} \left(\frac{\hat{K}_{n1}(d_{n1}) - \hat{m}_{n1}}{\hat{\sigma}_{n1}}, \ldots, \frac{\hat{K}_{nJ}(d_{nJ}) - \hat{m}_{nJ}}{\hat{\sigma}_{nJ}} \right)^\top .$$

Theorem 6 *The assumptions of Theorem 5 are satisfied, and consider the covariance matrix Σ from there. Then*

$$\widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma). \tag{6}$$

Statistical results

It follows from theoretical results that we obtain asymptotically χ^2 test for goodness of fit under the simple and the composite null hypotheses.

First consider the simple null hypothesis asserting that a sample $X_1, \ldots, X_n$ has the uniform distribution on $[0, 1]$. Given $J \geq 1$ and distance levels $d_{n1} \leq \cdots \leq d_{nJ},\ n \in \mathbb{N}$, each satisfying one of the conditions (T1), (T2) or (T3) such that the condition (2) holds. The covariance matrix Σ as in Theorem 2 is nonsingular. Let $\mathbf{K}_n$ be the vector given by (1). Then from (3) it follows that under the simple null hypothesis

$$C_n := \mathbf{K}_n^\top \Sigma^{-1} \mathbf{K}_n \xrightarrow{\mathcal{D}} \chi^2_J,$$

where χ^2_J is a random variable with the chi-square distribution with J degrees of freedom. So, C_n defines a test for uniformity on $[0, 1]$, called here *the cluster test* and denoted by C_n.

Now, consider the composite null hypothesis asserting that a sample comes from the family of all uniform distributions on $\mathbb{R}$. Distance levels $d_{n1} \leq \cdots \leq d_{nJ},\ n \in \mathbb{N}$, each satisfy the conditions of Theorem 6, then under the simple null hypothesis we obtain

$$\widehat{C}_n := \widehat{\mathbf{K}}_n^\top \Sigma^{-1} \widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \chi^2_J.$$

Accordingly it may seemed, that the composite hypothesis may be tested like the previous paragraph. The problem is that as we don't know the explicit value a and b, so the component of the covariance matrix Σ can't be determined, hence the test statistics $\widehat{C}_n$ can't be counted based on a given sample. Therefore we test the composite null hypothesis with another procedure. Here, we propose a possible solution based on the random transformation of the sample $V_1, \ldots, V_n$ coming from an unknown interval into the unit interval as follows:

$$\left(\frac{V_{2,n} - V_{1,n}}{V_{n,n} - V_{1,n}}, \ldots, \frac{V_{n-1,n} - V_{1,n}}{V_{n,n} - V_{1,n}} \right).$$

Here $\tilde{K}_{n-2,j}(d_{nj})$ denote the numbers of clusters corresponding to the distance levels d_{nj} for the randomly transformed sample, $j = 1, \ldots, J$, and let

$$\tilde{\mathbf{K}}_{n-2} := \frac{1}{\sqrt{n}} \left(\frac{\tilde{K}_{n-2,1}(d_{n1}) - m_{n-2,1}}{\sigma_{n-2,1}}, \ldots, \frac{\tilde{K}_{n-2,J}(d_{nJ}) - m_{n-2,J}}{\sigma_{n-2,J}} \right)^\top$$

be a vector of normalized numbers of clusters of the randomly transformed sample. In addition let $\tilde{\Sigma}$ be the covariance matrix computed using the randomly transformed sample. Then

$$C_n^{\mathrm{mod}} := \tilde{\mathbf{K}}_{n-2}^{\top} \tilde{\Sigma}^{-1} \tilde{\mathbf{K}}_{n-2} \xrightarrow{\mathcal{D}} \chi_J^2.$$

So, C_n^{mod} defines a test for uniformity, called here *the modified cluster test* and denoted by C_n^{mod}. These tests define asymptotically χ^2 tests for a uniform distribution or for the uniform family. This means that asymptotic critical values of these tests are given by quantiles of the chi-square distribution with J degrees of freedom.

We simulated powers of the new tests against some continuous alternative distributions on $[0, 1]$ and we compared these tests with the data driven smooth test introduced in Inglot and Ledwina [12]. The conclusion is that the cluster tests perform less well than other procedures unless some highly oscillating alternatives.

4 Goodness of fit to the normal family

Introduction and preliminary results

We perform a simulation study of the goodness of fit test to the normal family based on the L^2-Wasserstein distance, proposed by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [10]. They obtained the location- and scale-free test statistic for the null hypothesis $\mathcal{H}_0 : F \in \mathbf{N}$, where $\mathbf{N}$ denotes the normal family. This testing procedure belongs to the class of minimum distance tests (using the distance of quantile functions); on the other hand it is asymptotically equivalent with a correlation test. The name of this test derives from these two different approaches: the quantile correlation test.

To describe their proposal, let $\mathcal{P}_2(\mathbb{R})$ be the set of probabilities on $\mathbb{R}$ with a finite second moment. For probabilities P_1 and P_2 in $\mathcal{P}_2(\mathbb{R})$ the L_2-Wasserstein distance between P_1 and P_2 is

$$\mathcal{W}(P_1, P_2) = \inf \left\{ [E(X_1 - X_2)^2]^{1/2}, \mathcal{L}(X_1) = P_1, \mathcal{L}(X_2) = P_2 \right\},$$

where $\mathcal{L}(X)$ denotes the probability distribution of the random variable X. It can be explicitly obtained in terms of quantile functions:

$$\mathcal{W}(P_1, P_2) = \left[\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 \mathrm{d}t \right]^{1/2},$$

where F_1^{-1} and F_2^{-1} are quantile function associated with the probabilities P_1 and P_2.

If P is a probability distribution in $\mathcal{P}_2(\mathbb{R})$ with distribution function F, mean μ_0 and standard deviation σ_0, then L_2-Wasserstein distance-square between F and the class of all normal laws $\mathbf{N}$ is

$$\mathcal{W}^2(P, \mathbf{N}) := \inf\{\mathcal{W}^2(P, N_\sigma^\mu), N_\sigma^\mu \in \mathbf{N}\} = \sigma_0^2 - \left(\int_0^1 F^{-1}(t) \Phi^{-1}(t) \mathrm{d}t \right)^2,$$

where Φ^{-1} is the standard normal quantile function. The ratio $\mathcal{W}^2(P,\mathbf{N})/\sigma_0^2$ is not affected by location or scale changes of F. Hence, it can be considered as a measure of dissimilarity between F and $\mathbf{N}$.

Given a random sample $X_1,\ldots,X_n$ from F, now the empirical version of the ratio $\mathcal{W}(P,\mathbf{N})/\sigma_0$ may be obtained. Then the location- and scale-free BCMR-test statistic for the null hypothesis $H_0 : F \in \mathbf{N}$ is

$$T_n := \frac{\mathcal{W}^2(F_n,\mathbf{N})}{S_n^2} = 1 - \frac{\left[\int_0^1 Q_n(t)\Phi^{-1}(t)dt\right]^2}{S_n^2} = 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi^{-1}(t)\,\mathrm{d}t\right]^2}{S_n^2},$$

where S_n^2 denotes the empirical variance.

Del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [10] investigated the asymptotic distribution of the test statistic under the null-hypothesis. They managed to produce the limit distribution in two different forms. The first form is functionals of the Brownian bridge, the second is a series of random variables. Let φ denote the standard normal density function and B denote the Brownian bridge, and let

$$a_n = \frac{1}{n}\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{\left[\varphi(\Phi^{-1}(t))\right]^2}dt.$$

Theorem 7 (del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez[10]) *If $F \in \mathbf{N}$, then*

$$\begin{aligned} n(T_n - a_n) &\xrightarrow{\mathcal{D}} \int_0^1 \frac{B^2(t) - E(B^2(t))}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t - \left[\int_0^1 \frac{B(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 - \left[\int_0^1 \frac{B(t)\Phi^{-1}(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 \\ &\stackrel{\mathcal{D}}{=} -\frac{3}{2} + \sum_{j=3}^{\infty}\frac{Z_j^2 - 1}{j}, \end{aligned}$$

where $(Z_j)_{j=3}^{\infty}$ is a sequence of independent standard normal random variables.

Simulation results

In the simulation study the distribution function of the limiting random variable above is computed numerically in two different ways. You can see the asymptotic distribution in Fig. 1. Next, using different sample sizes from $n = 10$ to $n = 100\,000$, we simulate the distribution function of the BCMR-test statistic $n(T_n - a_n)$. As shown in Fig. 2, we find that the convergence is overall very slow.

A simulation study was performed to evaluate the power of the BCMR test against many continuous alternative distributions and make comparisons with five other tests of normality. The first of these tests is Shapiro–Wilk's W test [19], for which there is a specific interest in the comparison for the small $n = 20$ and 50, while for $n = 100$ we use the Shapiro–Francia [18] extension of the W test, denoted by W'. Among the EDF tests we considered the Kolmogorov–Smirnov D test [13], with the modification

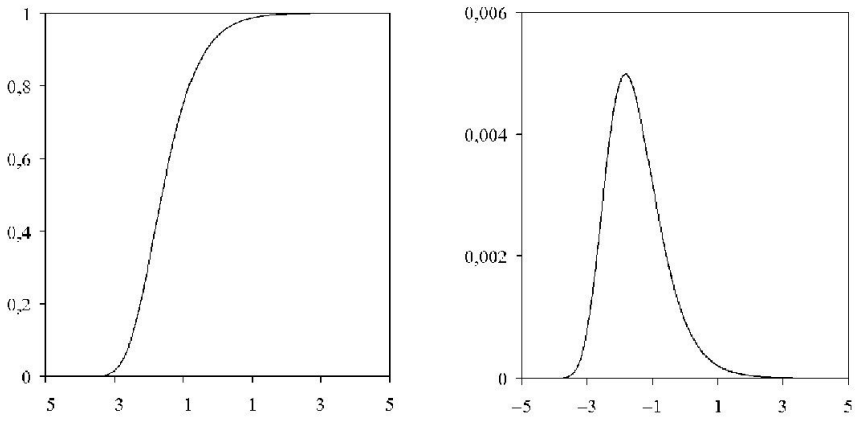

Figure 1: The asymptotic distribution function (left) and its density (right)

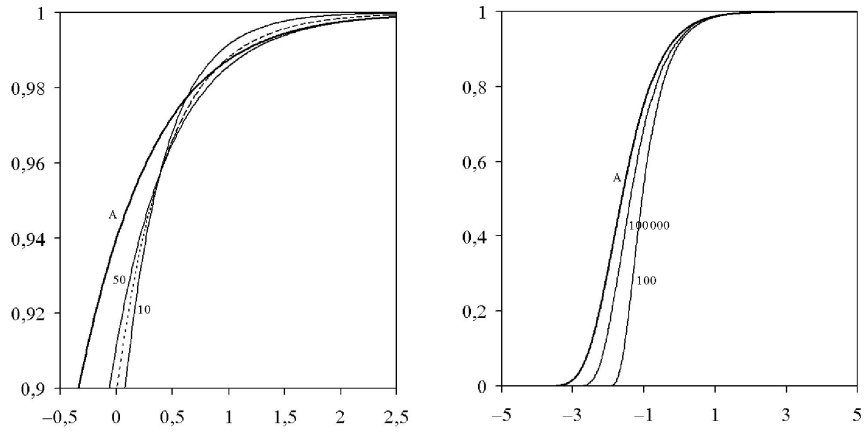

Figure 2: The distribution functions of the BCMR-test statistic $n(T_n - a_n)$ for $n = 10$, 20 (dotted line), 50 and the asymptotic distribution function the thicker line marked with A on the left (left). The same for $n = 100$ and 100 000 (right)

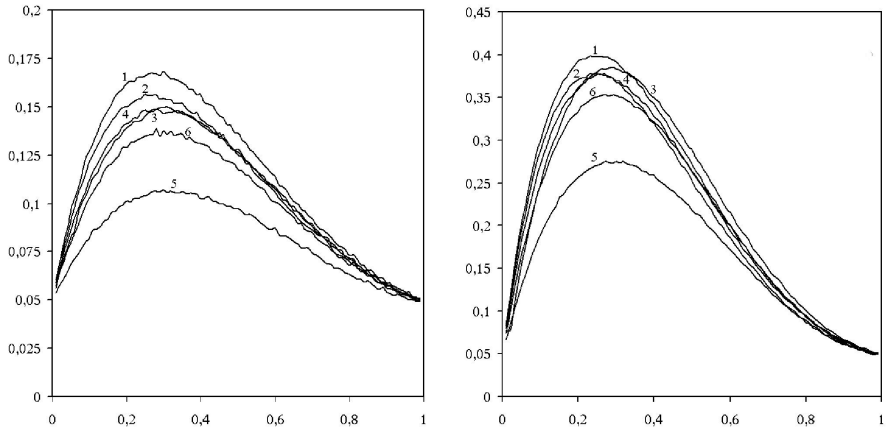

Figure 3: Power of the BCMR, W, ISE, BHEP, D and A^2 test as a function of λ for $CN(\lambda, 4)$ (left) and the same for $CN(\lambda, 9)$: 1=BCMR test; 2=W test; 3=ISE test; 4=BHEP test; 5=D; 6=A^2 test (right)

suggested by Stephens [20], and the Anderson-Darling A^2 test [1]. The fourth test we have chosen is based on density estimation, the integrated squared error (ISE) test of Bowman and Foster [3] with fixed kernels. The fifth test is based on the empirical characteristic function, the BHEP test from Epps and Pulley [11]. The two figures in Fig. 3 contain a comparison of the powers of five tests against contaminated normal alternatives for $n = 20$ and $\alpha = 0.05$ significance level. The symbol $CN(\lambda, \sigma^2)$ stands for the contaminated normal distribution $F(x) = (1 - \lambda)\Phi(x) + \lambda\Phi(x/\sigma)$, $x \in \mathbb{R}$, for all $0 < \lambda < 1$ and $\sigma > 0$.

A rough general conclusion of this study is that the BCMR-test usually performs better than the other tests, except for the combination of the Wilk–Shapiro- and the Shapiro–Francia-test. In most cases the properties of the latter combination and the properties of the BCMR quantile correlation test appear to be very similar to each other. Since under the null hypothesis the asymptotic distribution for Wilk–Shapiro-test is the same as for the BCMR-test, thus the result of the power study isn't surprising.

5 Goodness of fit to the logistic family

We present the weighted version of the quantile correlation test statistics for goodness of fit to the logistic family, introduced by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [10], and del Barrio, Cuesta-Albertos and Matrán [9]. The use of weight functions in the test statistics were suggested independently from each other by de Wet in [7] and [8] and by S. Csörgő in [4] and [5]. It is an interesting fact that there the authors' motivations were considerably different. S. Csörgő showed that the suitably weighted versions of the correlation tests have limiting distribution for more

family of probability distributions; de Wet expected "the loss of degrees of freedom" in the limiting null distribution (in the case of the normal family this means that the first two terms are missing in the infinite series representation of the asymptotic distribution). We prove the results of S. Csörgő [5] for location and scale logistic family with the weight function for location family suggested by de Wet.

For a given distribution function $G(x),\ x \in \mathbb{R}$, and for $\theta \in \mathbb{R}$ and $\sigma > 0$, let $G_\sigma^\theta(x) = G((x-\theta)/\sigma)$, $x \in \mathbb{R}$, and consider the location-scale family

$$\mathcal{G}_{l,s} = \{G_\sigma^\theta : \theta \in \mathbb{R}, \sigma > 0\}.$$

Denote by $Q_G(t) = G^{-1}(t), 0 < t < 1$, the quantile function of G. Consider a weight function $w : (0,1) \to [0,\infty)$ satisfying $\int_0^1 w(t)\,\mathrm{d}t = 1$, and define the weighted r-th moment

$$\mu_r(G,w) := \int_0^1 (Q_G(t))^r w(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^r w(G(x))\mathrm{d}G(x).$$

Assume that $\mu_1(G,w)$ and $\mu_2(G,w)$ are finite, and define also the weighted variance:

$$\nu(G,w) := \mu_2(G,w) - \mu_1^2(G,w) \geq 0.$$

The weighted L_2-Wasserstein distance with weight function w of two distributions F and G can be defined as

$$\mathcal{W}_w(F,G) := \left[\int_0^1 \left(Q_F(t) - Q_G(t)\right)^2 w(t)\,\mathrm{d}t\right]^{\frac{1}{2}}.$$

Therefore the weighted L_2-Wasserstein distance $\mathcal{W}_w(F,\mathcal{G}_{l,s}) = \inf\{\mathcal{W}_w(F,G) : G \in \mathcal{G}_{l,s}\}$ between F and location-scale family $\mathcal{G}_{l,s}$, scaled to F is

$$\frac{\mathcal{W}_w^2(F,\mathcal{G}_{l,s})}{\nu(F,w)} = 1 - \frac{\left[\int_0^1 Q_F(t)Q_G(t)w(t)\mathrm{d}t - \mu_1(F,w)\mu_1(G,w)\right]^2}{\nu(F,w)\nu(G,w)},$$

as derived in [5].

Consider a random sample $X_1,\ldots,X_n$ with common distribution function F, and let a fixed distribution function G. We would like to test the null hypothesis $\mathcal{H}_0 : F \in \mathcal{G}_{l,s}$. Letting Q_n be the sample quantile function, in order to define the following test statistics

$$\begin{aligned} V_n &:= 1 - \frac{\left[\int_0^1 Q_n(t)Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_0^1 Q_n(t)w(t)\mathrm{d}t\right]^2}{\nu(G,w)\left[\int_0^1 Q_n^2(t)w(t)\mathrm{d}t - \left(\int_0^1 Q_n(t)w(t)\mathrm{d}t\right)^2\right]} \\ &= 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\left\{\int_{\frac{k-1}{n}}^{\frac{k}{n}} Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right\}\right]^2}{\nu(G,w)\left[\sum_{k=1}^n X_{k,n}^2\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t - \left(\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right)^2\right]}, \end{aligned}$$

derived from the weighted L^2-Wasserstein distance between the empirical distribution of the sample and the location-scale family $\mathcal{G}_{l,s}$.

Csörgő determined the limiting behaviour of this statistics, below we use the following general result due to Csörgő [5].

Theorem 8 (Csörgő [5]) *Let w be a nonnegative integrable function on the interval $(0,1)$, for which $\int_0^1 w(t)\,\mathrm{d}t = 1$. Suppose that G has finite weighted second moment and that it is twice differentiable on the open interval (a_G, b_G) such that $g(x) = G'(x) > 0$ for all $x \in (a_G, b_G)$, and let B denote the Brownian bridge. If the conditions*

$$\sup_{0<t<1} \frac{t(1-t)|g'(Q_G(t))|}{g^2(Q_G(t))} < \infty, \qquad \int_0^1 \frac{t(1-t)}{g^2(Q_G(t))} w(t)\mathrm{d}t < \infty,$$

and

$$n\int_0^{\frac{1}{n+1}} \left[Y_{1,n} - Q_G(t)\right]^2 w(t)\mathrm{d}t \xrightarrow{\mathbf{P}} 0, \qquad n\int_{\frac{n}{n+1}}^1 \left[Y_{n,n} - Q_G(t)\right]^2 w(t)\mathrm{d}t \xrightarrow{\mathbf{P}} 0$$

are satisfied, the following asymptotic is valid:

If F belong to $\mathcal{G}_{l,s}$ generated by G, then

$$nV_n \xrightarrow{\mathcal{D}} V_g := \frac{1}{\nu(G,w)} \left\{ \int_0^1 \frac{B^2(t)}{g^2(Q_G(t))} w(t)\mathrm{d}t - \left[\int_0^1 \frac{B(t)}{g(Q_G(t))} w(t)\mathrm{d}t\right]^2 \right\}$$
$$- \left[\frac{1}{\nu(G,w)} \int_0^1 \frac{B(t)Q_G(t)}{g(Q_G(t))} w(t)\mathrm{d}t - \frac{\mu_1(G,w)}{\nu(G,w)} \int_0^1 \frac{B(t)}{g(Q_G(t))} w(t)\mathrm{d}t\right]^2.$$

This theorem was used to establish the asymptotic distributions of the test statistics specialized to the logistic family.

Results

Consider the logistic distribution function $G(x) = 1/(1+e^{-x}),\ x \in \mathbb{R}$, and $\mathcal{G}_{l,s}$ denotes the logistic location-scale family as defined above. With the weight function suggested by de Wet [8] for the logistic location family $w(t) = 6t(1-t),\ 0 < t < 1$, with the weighted first moment $\mu_1(G,w) = 0$ and with the weighted second moment $\mu_2(G,w) = \pi^2/3 - 2$ the location-scale-free test statistic specializes to

$$V_n = 1 - \frac{\left[\sum_{k=1}^n a_{k,n} X_{k,n}\right]^2}{\left(\frac{\pi^2}{3} - 2\right)\left[\sum_{k=1}^n b_{k,n} X_{k,n}^2 - \left(\sum_{k=1}^n b_{k,n} X_{k,n}\right)^2\right]},$$

where the coefficients are given explicitly by

$$\begin{aligned} a_{k,n} &= \int_{\frac{k-1}{n}}^{\frac{k}{n}} 6t(1-t)\ln\left(\frac{t}{1-t}\right)\mathrm{d}t \\ &= \frac{k^2(3n-2k)}{n^3}\ln\frac{k}{n-k} - \frac{(k-1)^2(3n-2k+2)}{n^3}\ln\frac{k-1}{n-k+1} \\ &\quad + \ln\frac{n-k}{n-k+1} + \frac{1-2k}{n^2} + \frac{1}{n}\,, \\ b_{k,n} &= \int_{\frac{k-1}{n}}^{\frac{k}{n}} 6t(1-t)\mathrm{d}t = \frac{3(2k-1)}{n^2} + \frac{2(-3k^2+3k-1)}{n^3}\,. \end{aligned}$$

As a corollary to the asymptotic results from [5] we obtain the following limiting distribution of the test statistics V_n.

Theorem 9 *If the distribution function F of the sample belongs to the logistic location-scale family $\mathcal{G}_{l,s}$ then the rescaled statistic nV_n has the asymptotic distribution*

$$\begin{aligned} nV_n \xrightarrow{\mathcal{D}} V :=& \frac{1}{\pi^2/3-2}\left\{\int_0^1 \frac{6B^2(t)}{t(1-t)}\,\mathrm{d}t - \left[\int_0^1 6B(t)\,\mathrm{d}t\right]^2\right\} \\ & - \left[\frac{1}{\pi^2/3-2}\int_0^1 6B(t)\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t\right]^2, \end{aligned}$$

where the integrals exists with probability 1.

Del Barrio, Cuesta-Albertos and Matrán [9] obtained the asymptotic distribution as the Karhunen–Loève expansion of the weighted Brownian-bridge. With the same technique we determine the infinite series representation of our limiting distribution.

Theorem 10 *The limiting distribution V can be represented alternatively as*

$$V \overset{\mathcal{D}}{=} \frac{1}{\frac{\pi^2}{3}-2}\sum_{k=2}^{\infty}\frac{6}{k(k+1)}Z_m^2 - \left[\frac{1}{\frac{\pi^2}{3}-2}\sum_{l=1}^{\infty}\frac{3\sqrt{4l+1}}{l(l+1)(2l-1)(2l+1)}Z_{2l}\right]^2,$$

where $(Z_m)_{m=1}^{\infty}$ is an infinite sequence of independent identically distributed standard normal random variables, and the series converges with probability 1.

Simulation results

Similarly to the previous section a simulation study was performed to evaluate the power of the tests. The numerical results is presented in Table 1. We compare the new test with Meintanis-tests based on the empirical characteristic function and the empirical momentum generating function from [15].

A rough general conclusion of this study is that a simply computable test statistic is obtained, the asymptotic critical values may be used and the test seems to be average powerful.

References

[1] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

[2] F. Balogh and É. Krauczi. Weighted quantile corelation test for the logistic family. *Acta Scientiarum Mathematicarum.(Szeged)*, 80(1-2):307–326, 2014.

[3] A. Bowman and P. Foster. Adaptive smoothing and density-based tests of multivariate normality. *JASA. Journal of the American Statistical Association*, 88:529–537, 1993.

[4] S. Csörgő. Weighted correlation tests for scale families. *Test*, 11(1):219–248, 2002.

[5] S. Csörgő. Weighted correlation tests for location-scale families. *Mathematical and Computer Modelling*, 38(7-9):753–762, 2003. Hungarian applied mathematics and computer applications.

[6] S. Csörgő and W. B. Wu. On the clustering of independent uniform random variables. *Random Structures Algorithms*, 25(4):396–420, 2004.

[7] T. de Wet. Discussion of "Contributions of empirical and quantile processes to the asymptotictheory of goodness-of-fit tests". *Test*, 9(1):74–79, 2000.

Table 1: Empirical powers (in %) for nV_n against some alternatives ($n = 20, 50$ and 100 sample sizes, $*$ 100% empirical power, α significance level

Alternatives	20	50	100	20	50	100
$N(0,1)$	5	6	8	2	2	4
Uniform	13	47	93	5	29	82
Cauchy	88	99	*	84	99	*
Laplace	26	39	55	17	29	43
Exp(1)	70	99	*	56	97	*
Triangle(I)	4	7	13	2	3	6
Triangle(II)	21	61	97	11	43	91
Beta(2;2)	6	15	40	2	7	24
Weibull(2)	12	25	54	5	15	38
Gamma(2,1)	40	81	99	27	69	98
Lognormal	86	*	*	79	*	*
Student(5)	16	19	21	10	12	13
χ^2_1	94	*	*	88	*	*
Negativ Exp	69	99	*	56	97	*
α		0,10			0,05	

[8] T. de Wet. Goodness-of-fit tests for location and scale families based on a weighted L_2-Wasserstein distance measure. *Test*, 11(1):89–107, 2002.

[9] E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, 9(1):1–96, 2000. With discussion.

[10] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[11] T. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70:723–726, 1983.

[12] T. Inglot and T. Ledwina. Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Applications*, 417(1):124–133, 2006.

[13] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale del Istituto Italiano degli Attuari*, 4:83–91, 1933.

[14] É. Krauczi. A study of the quantile correlation test of normality. *Test*, 18(1):156–165, 2009.

[15] S. G. Meintanis. Goodness-of-fit tests for the logistic distribution based on empirical transforms. *Sankhyā. The Indian Journal of Statistics*, 66(2):306–326, 2004.

[16] K. É. Osztényiné. Joint cluster counts from uniform distribution. *Probability and Mathematical Statistics*, 33(1):93–106, 2013.

[17] E. S. Pearson. A further development of tests for normality. *Biometrika*, 22:239–249, 1930.

[18] M. W. Shapiro, S.S. and H. Chen. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 63:1343–72, 1968.

[19] S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[20] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.