On Testing Goodness-of-fit for Cauchy Distribution

František Rublík

Institute of Measurement Science, Slovak Academy of Sciences
Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic
E-mail: umerrubl@savba.sk

Abstract. *A simulation comparison of the Henze test with the extreme quantile test for testing the hypothesis that the sample is drawn from the Cauchy distribution, is presented. The results suggest that the extreme quantile test is better for the sample sizes $n \leq 50$, while the Henze test is better for $n \geq 100$.*

Suppose that $x_1, \ldots, x_n$ is a random sample. By the null hypothesis H_0 it is understood throughout the paper that the sample is drawn from a Cauchy distribution with unknown parameter μ of location and unknown parameter σ of scale. As usual, by Cauchy distributions we understand the probabilities on real line, corresponding to the distribution functions

$$F(x, \mu, \sigma) = \frac{1}{2} + \frac{1}{\pi} \arctan(\frac{x-\mu}{\sigma}) , \quad \mu , \sigma \in R , \ \sigma > 0 . \tag{1}$$

Let

$$F_n(x) = \frac{1}{n} \text{card} \{ j; \ j \leq n, \ x_j \leq x \} \tag{2}$$

denote the empirical distribution function. The null hypothesis is in Section 4.14 of [2] recommended to be tested by means of the Anderson-Darling test statistic

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \ \log \left(Z^{(i)} (1 - Z^{(n+1-i)}) \right) , \tag{3}$$

where $Z^{(1)} \leq \ldots \leq Z^{(n)}$ are order statistics computed from

$$Z_j = F(x_j, \hat{\mu}, \hat{\sigma}) , \quad j = 1, \ldots, n . \tag{4}$$

Here F is the function (1) and the estimators

$$\hat{\mu} = \sum_{i=1}^{n} g_{ni} x_n^{(i)} , \quad g_{ni} = \frac{1}{n} G\Big(\frac{i}{n+1}\Big) , \tag{5}$$

$$G(u) = \frac{\sin(4\pi(u-0.5))}{\tan(\pi(u-0.5))} = -4 \sin^2(\pi u) \cos(2\pi u) ,$$

$$\hat{\sigma} = \sum_{i=1}^{n} c_{ni} x_n^{(i)} , \quad c_{ni} = \frac{1}{n} J\Big(\frac{i}{n+1}\Big) , \tag{6}$$

$$J(u) = \frac{8 \tan(\pi(u-0.5))}{\sec^4(\pi(u-0.5))} = -8 \cos(\pi u) \sin^3(\pi u)$$

are the ones proposed in [1], $x_n^{(i)}$ is the ith order statistic computed from $x_1, \ldots, x_n$. As has already been mentioned on pp. 72 - 73 of [5], for the constants from (5) the equality $\sum_i g_{ni} = 1$ does not hold and therefore (5) is not an equivariant estimate of the location parameter. Consequently, the distribution of (3) depends in this case on the parameter of the sampled Cauchy distribution (simulations show that it depends on the ratio μ/σ). When the constants $c(\alpha, n)$ from the Table 4.26 on p. 163 of [2] are used, then according to the simulation results from p. 73 of [5] for $\mu = 20$, $\sigma = 1$ and $n = 50$

$$P(A^2 > c(0.1, 50)) = 0.57\,,$$

and similarly by simulation one can find out that for $\mu = 20$, $\sigma = 1$ and $n = 100$

$$P(A^2 > c(0.1, 1000)) = 0.41\,,$$

(where $c(\alpha, n)$ is obtained for sampling from the Cauchy distribution with $\mu = 0$, $\sigma = 1$). These values are far above the nominal level $\alpha = 0.1$, the same effect can be observed when the asymptotic constants $c(\alpha, +\infty)$ are used.

Another test of the mentioned null hypothesis has been proposed in [3]. This test is based on the statistic

$$D_{n\lambda} = n \int_{-\infty}^{+\infty} \left| \Psi_n(t) - e^{-|t|} \right|^2 e^{-\lambda|t|}\, \mathrm{d}t\,, \qquad \Psi_n(t) = \frac{1}{n} \sum_{j=1}^{n} e^{\mathrm{i}tY_j}\,,$$

where

$$Y_j = \frac{X_j - \hat{\mu}}{\hat{\sigma}}\,. \tag{7}$$

According to the formula (1.4) on p. 268 of [3]

$$D_{n,\lambda} = \frac{2}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\lambda}{\lambda^2 + (Y_j - Y_k)^2} - 4 \sum_{j=1}^{n} \frac{1+\lambda}{(1+\lambda)^2 + Y_j^2} + \frac{2n}{2+\lambda}\,, \tag{8}$$

and the estimates

$$\hat{\mu} = \begin{cases} \dfrac{x_n^{(k)} + x_n^{(k+1)}}{2} & n = 2k\,, \\ \\ x_n^{(k+1)} & n = 2k+1\,, \end{cases} \tag{9}$$

$$\hat{\sigma} = \frac{\hat{\xi}_{0.75n} - \hat{\xi}_{0.25n}}{2} \tag{10}$$

are used, where (cf. (2))

$$\hat{\xi}_{pn} = \inf\{t;\ F_n(t) \geq p\,\} \tag{11}$$

denotes the sample pth quantile. The Gürtler-Henze test rejects the null hypothesis if $D_{n,\lambda} > d(n, \lambda, \alpha)$, where under validity of H_0

$$P(D_{n,\lambda} > d(n, \lambda, \alpha)) = \alpha\,. \tag{12}$$

Independently of [3] another test for testing H_0 was presented in [6]. This test is an inuitive modification of the quantile test from [5] and the results of [5] are derived by means of the general theory of the quantile test from [4].

Suppose that $\hat{\mu}$ is the median (9), $\hat{\sigma}$ is the trigonometric estimate(6) and

$$\boldsymbol{\Delta}_n = \left(F(x_n^{(1)}, \hat{\mu}, \hat{\sigma}) - \frac{1}{n+1}, F(x_n^{(n)}, \hat{\mu}, \hat{\sigma}) - \frac{n}{n+1} \right)'\,, \tag{13}$$

where F is the function (1). Further, put

$$\mathbf{\Sigma}_n = \mathbf{A}_n + \mathbf{G}_n \,, \tag{14}$$

where

$$\begin{aligned}
&\mathbf{A}_n = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \,, \quad a_{ij} = \min\{p_i, p_j\}\,(1 - \max\{p_i, p_j\}) \,,\ p_1 = \tfrac{1}{n+1} \,,\ p_2 = \tfrac{n}{n+1} \,, \\
&\mathbf{G}_n = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \,, \\
&g_{ij} \;=\; \frac{\sin^2(p_i\pi)\sin^2(p_j\pi)}{4} - \frac{sin^2(p_i\pi)}{2}\min\{p_j, 1-p_j\} - \\
&\quad -\; \frac{sin^2(p_j\pi)}{2}\min\{p_i, 1-p_i\} - \frac{\sin(2p_i\pi)\sin(2p_j\pi)}{2\pi^2} \,.
\end{aligned} \tag{15}$$

By means of Theorem 1.2 of [5] for $n > 2$ one easily finds out that for $n > 2$ the matrix $\mathbf{\Sigma}_n$ is regular. The test statistic presented in [6] is

$$Q_n = n\mathbf{\Delta}_n' \mathbf{\Sigma}_n^{-1} \mathbf{\Delta}_n \tag{16}$$

and the null hypothesis H_0 is rejected whenever

$$Q_n > q(\alpha, n) \,. \tag{17}$$

Here the constant $q(\alpha, n)$ fulfills under the validity of H_0 the equality

$$P(Q_n > q(\alpha, n)) = \alpha \,, \tag{18}$$

the values of $q(\alpha, n)$ for selected α and n can be found in Table 3 of [6].

The following table contains simulation estimates of the probabilities of rejection (cf. (8)–(11), (12) and (13)–(16), (18))

$$r_D = P(D_{n,\lambda} > d(n, \lambda, \alpha)) \,, \quad r_Q = P(Q_n > q(\alpha, n)) \,,$$

where $\lambda = 5$ and $\alpha = 0.1$ (according to the simulation estimates from [3] the value $\lambda = 5$ turns out to be an optimal choice of λ). All the simulations have been carried out by means of MATLAB, version 4.2c.1 with $N = 10000$ trials for each particular case, except for the values of r_D for $n = 100$ and $n = 200$, which are taken from [3] (and based also on $N = 10000$ trials). The alternatives considered in the following table are those defined on pp. 279–280 of [3].

n	20		50		100		200	
	r_D	r_Q	r_D	r_Q	r_D	r_Q	r_D	r_Q
N(0,1)	0.46	**0.52**	0.97	**1**	**1**	**1**	**1**	**1**
NC(0.1,0.9)	0.1	**0.11**	0.09	**0.12**	0.1	**0.11**	**0.11**	**0.11**
NC(0.3,0.7)	0.09	**0.13**	0.12	**0.17**	**0.19**	0.17	**0.32**	0.16
NC(0.5,0.5)	0.1	**0.17**	0.24	**0.28**	**0.48**	0.28	**0.79**	0.28
NC(0.7,0.3)	0.17	**0.25**	**0.53**	0.51	**0.86**	0.51	**0.99**	0.51
NC(0.9,0.1)	0.33	**0.40**	0.87	**0.88**	**1**	0.88	**1**	0.87
Student(2)	0.13	**0.17**	0.36	**0.53**	0.7	**0.8**	**0.96**	0.93
Student(3)	0.20	**0.23**	0.62	**0.81**	0.95	**0.98**	**1**	**1**
Student(4)	0.25	**0.29**	0.77	**0.91**	0.99	**1**	**1**	**1**
Student(5)	0.28	**0.33**	0.83	**0.95**	**1**	**1**	**1**	**1**
Student(7)	0.34	**0.38**	0.89	**0.98**	**1**	**1**	**1**	**1**
Student(10)	0.38	**0.42**	0.93	**0.99**	**1**	**1**	**1**	**1**
Tukey(1)	**0.13**	0.09	**0.14**	0.07	**0.18**	0.06	**0.23**	0.05
Tukey(0.2)	0.21	**0.25**	0.65	**0.84**	0.96	**0.99**	**1**	**1**
Tukey(0.1)	0.31	**0.35**	0.86	**0.97**	**1**	**1**	**1**	**1**
Tukey(0.05)	0.38	**0.43**	0.93	**0.99**	**1**	**1**	**1**	**1**
Uniform	0.84	**0.95**	**1**	**1**	**1**	**1**	**1**	**1**
Logistic	0.34	**0.38**	0.90	**0.99**	**1**	**1**	**1**	**1**
Laplace	0.18	**0.19**	0.55	**0.85**	0.93	**1**	**1**	**1**
Gumbel	0.39	**0.61**	0.97	**1**	**1**	**1**	**1**	**1**

The table shows that for sample sizes n not larger than 50 the extreme quantile test proposed in [6] gives for overwhelming majority of the considered alternatives better results than the Henze test and therefore for these n the extreme quantile test should be used. For n about 100 the number of alternatives for which the given test is better than the other is approximately the same both for the extreme quantile and for the Henze test, but since the performance of the later test in the cases when it is more powerfull turns out to be more striking than the performance of the extreme quantile test in the cases which are in his favour (for $n = 200$ the situation is in favour of the Henze test), for $n \geq 100$ the Henze test should be used.

References

[1] Chernoff, H., Gastwirth, J. and Johns, M. *Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation*. Ann. Math. Statist. 38(1967), 52 – 72.

[2] D'Agostino, R. and Stephens, M.(eds.) *Goodness-of-Fit Techniques*. Marcel Dekker Inc., New York, 1986.

[3] Gürtler, N. and Henze, N. *Goodness-of-fit tests for Cauchy distribution based on the empirical characteristic function*. Ann. Inst. Statist. Math. 52(2000), 267 – 286.

[4] Rublík, F. *A quantile goodness-of-fit test applicable to distributions with non-differentiable densities*. Kybernetika 33(1997), 505 – 524.

[5] Rublík, F. *A goodness-of-fit test for Cauchy distribution*. In: Probastat'98, Proceedings of the Third International Conference on Mathematical Statistics, Tatra Mountains Mathematical Publications 17(1999), Bratislava, pp. 71 – 81.

[6] Rublík, F. *A quantile goodness-of-fit for Cauchy distribution, based on extreme order statistics*. Applications of Mathematics 46(2001), 339 – 351.