RESEARCH ARTICLE

Testing for goodness rather than lack of fit of continuous probability distributions

Stefan Wellek

1 Department of Biostatistics, CIMH Mannheim, Mannheim Medical School of the University of Heidelberg, Mannheim, Germany, 2 Department of Medical Biostatistics, Epidemiology & Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

* stefan.wellek@zi-mannheim.de

Abstract

The vast majority of testing procedures presented in the literature as goodness-of-fit tests fail to accomplish what the term is promising. Actually, a significant result of such a test indicates that the true distribution underlying the data differs substantially from the assumed model, whereas the true objective is usually to establish that the model fits the data sufficiently well. Meeting that objective requires to carry out a testing procedure for a problem in which the statement that the deviations between model and true distribution are small, plays the role of the alternative hypothesis. Testing procedures of this kind, for which the term tests for equivalence has been coined in statistical usage, are available for establishing goodness-of-fit of discrete distributions. We show how this methodology can be extended to settings where interest is in establishing goodness-of-fit of distributions of the continuous type.

1 Introduction

Goodness-of-fit tests belong to the oldest and most frequently used methods of statistical inference. A chapter devoted to them can be found in almost every textbook for statisticians working in whatever area of application. Likewise, there are numerous authoritative expositions of the mathematical theory of these methods, beginning with Cramèr’s classical text [1], through all three editions of Lehmann’s “Testing Statistical Hypotheses” [2–4], to Volume 2A of “Kendall’s Advanced Theory of Statistics” [5], to mention just a few highly influential references of this category. Virtually all inferential procedures presented in the existing literature as tests of goodness-of-fit, share one crucial feature: The statement that the model to be fitted coincides with the true distribution from which the data are taken, plays the role of the null hypothesis, implying that a significant result actually indicates lack rather than goodness of fit of the model. This is clearly at variance with the fact that in the vast majority of applications, interest will be in proving rather than falsifying the model so that such a test typically fails to serve the purpose of its user. For instance, many of the most widely used methods of statistical analysis rely on the assumption that the data follow a specific distributional law (namely the Gaussian), and it is widespread practice to make sure of the adequacy of this assumption in a preliminary test. In the latter, the hypothesis one aims to establish states that the corresponding model
holds at least approximately true. In the existing literature, the term lack-of-fit test occurs quite infrequently. The main exception is research on methods to be used for detecting misspecifications of linear or generalized linear regression models (see, e.g., [6–9]).

Not surprisingly, the discrepancy between a test of goodness-of-fit and a procedure enabling one to establish the respective model, is addressed at least implicitly in some of the classical expositions of the topic (see, e.g., [5], §§25.6–7). The usual recommendation for finding a way around that difficulty is to take steps for increasing the power of the test, preferably, since increasing the order of magnitude of the sample size will rarely be feasible, by checking the \( p \)-value against an increased threshold, e.g. 10 instead of 5 percent and deciding in favor of goodness-of-fit if the test does not reject the null hypothesis even at that increased level of significance. However, it is a basic fact that “inverting” a test of a given null hypothesis \( H_0 \) against some alternative \( H_1 \) by declaring \( H_0 \) to be statistically proven if it cannot be rejected, fails to produce a test controlling the risk of taking an erroneous decision in favor of \( H_0 \). A procedure which is tailored for serving the latter purpose (“proof of the null hypothesis”), is what has become quite popular in biostatistics since the last few decades under the name equivalence test. Construction of a test of that kind requires to enlarge \( H_0 \) before defining it as the new alternative hypothesis, through introducing some indifference zone around the respective point in the parameter space consisting of models deviating from the model of interest by an amount considered still acceptable. The basic requisites for reformulating the testing problem in that way are a suitable measure of distance between true and hypothesized model (often called a metric in a mathematically not fully precise terminology) and a numerical specification of the maximum tolerable value of that distance (called equivalence margin in biostatistical contexts).

Up to now, equivalence tests for goodness-of-fit problems have been made available for problems of establishing models for discrete distributions (see [10], Ch. 9). In Section 2, we develop a framework for equivalence testing in settings where the objective is to establish goodness-of-fit of some hypothesized continuous distribution (like the standard normal law) to the true distribution underlying a dataset under analysis. In the proposed hypotheses formulation, the indifference zone around the model to be established, consists of all Lehmann alternatives [11] to the corresponding cumulative distribution function (cdf) for which the ratio \( \theta \), say, between the value of the true and the hypothetical cdf at any point in the sample space, falls in a sufficiently narrow interval around unity. Except for considering cdf’s rather than survivor functions which give the probabilities in the right-hand tail of a distribution, the Lehmann parameter \( \theta \) coincides with what plays, under the term ‘hazard ratio’, a prominent role in statistical survival analysis. In Section 3, a uniformly (in \( \theta \)) most powerful, exact test for a hypothesis of this form is derived and shown that its power does only depend on the Lehmann parameter \( \theta \), not on the cdf one aims to fit to the data. In Section 4. results of studying the UMP test by means of exact numerical methods are presented, focusing on comparisons with tests being available for grouped data taken from the distribution to be fitted. The question of how to extend the approach to settings where the model to be established involves unknown parameters (like location and scale), will be addressed in Section 5. An illustrating example is presented in Section 6.

2 Assumptions and hypotheses formulation

Throughout we assume that the assessment of goodness-of-fit of the distribution of interest can be based on a random sample \( X_1, \ldots, X_n \) of size \( n \in \mathbb{N} \) of mutually independent observations. The common distribution of these observations is assumed to be of the continuous type, with \( F \) as the cdf. The cdf of the distribution to be fitted will be denoted by \( F_0 \) and assumed to
have support on some (maybe unbounded) interval on the real line. A reasonable basis for constructing a region in the space of all continuous cdf’s on the real line which can be considered equivalent (i.e., sufficiently similar) to the cdf \( F_0 \) specified under the model of interest, consists of including all Lehmann alternatives \( F(\cdot) = [F_0(\cdot)]^{\theta} \) for which the maximum vertical distance between \( F \) and \( F_0 \) does not exceed some suitably chosen margin \( \delta > 0 \). Making this idea precise, we start with defining equivalence of the hypothesized to the true distribution by the condition

\[
\|F - F_0\| \equiv \sup_{-\infty < x < \infty} |F(x) - F_0(x)| < \delta \quad \text{for some (fixed) } \delta > 0. \tag{1}
\]

Modifying the proof given in [12] for survival rather than distribution functions in the straightforward way, it can be shown that for Lehmann alternatives, the condition (1) is equivalent to

\[
|\theta^{1/(1-\theta)} - \theta^0/(1-0)| < \delta. \tag{2}
\]

From basic properties of the expression on the left-hand side of (2) as a function of \( \theta \), we can eventually conclude that the goodness-of-fit criterion (1) is satisfied if and only if there holds

\[
(1 + \varepsilon)^{-1} < \theta < 1 + \varepsilon \quad \text{for some suitable } \varepsilon > 0. \tag{3}
\]

Given the equivalence margin \( \delta \) to \( \|F - F_0\| \), the corresponding value of \( \varepsilon \) is obtained by solving the equation:

\[
(1 + \varepsilon)^{-1/(1+\varepsilon)} - (1 + \varepsilon)^{-(1+\varepsilon)/\varepsilon} = \delta. \tag{4}
\]

Table 1 shows the values of \( \varepsilon \) determined from Eq (4) for a selection of customary specifications of \( \delta \), together with the right-hand limit of the equivalence range for \( \log \theta \) which one is used to consider as the parameter of interest in the proportional hazards model for survival distributions [13].

Fig 1 visualizes the equivalence region of the proposed form for the standard normal cdf \( \Phi \) in the space \( \mathcal{F} = \{F \mid F : \mathbb{R} \to [0, 1], F = \Phi^{\theta}, \theta \in (0, \infty)\} \) for \( \delta = 0.15 \).

3 A uniformly most powerful test for establishing goodness-of-fit in the continuous case

Adopting the conceptualization of the notion of goodness-of-fit of a specified distribution with cdf \( F_0 \) to a sample from an unknown continuous univariate distribution proposed in the Introduction and letting the indifference zone around the model consist of Lehmann alternatives to \( F_0 \), we need a test for the problem

\[
H : \theta \leq 1/(1 + \varepsilon) \quad \text{or} \quad \theta \geq 1 + \varepsilon \quad \text{vs.} \quad K : 1/(1 + \varepsilon) < \theta < 1 + \varepsilon. \tag{5}
\]

In deriving an optimum solution to this problem, we will make use of the following

| \( \delta \) | 0.05 | 0.075 | 0.10 | 0.15 | 0.20 | 0.25 |
|-----|-----|-----|-----|-----|-----|-----|
| \( \varepsilon \) | 0.1457 | 0.2266 | 0.3135 | 0.5077 | 0.7341 | 1.0000 |
| \log(1 + \varepsilon) | 0.1360 | 0.2042 | 0.2727 | 0.4106 | 0.5505 | 0.6931 |

https://doi.org/10.1371/journal.pone.0256499.t001
Lemma 1. Let the $X_i, i = 1, \ldots, n$ be i.i.d. with continuous cdf $[F_0(\cdot)]^\theta$ and let

$$\hat{\theta}_n(F_0) = -n/\sum_{i=1}^{n}\log[F_0(X_i)].$$

Then, the distribution of $\hat{\theta}_n(F_0)$ does not depend on $F_0$, and its cdf is, for any $\theta$, exactly given by $t \rightarrow 1 - F_{2\nu}(2\theta n/t)$, where $F_{2\nu}(\cdot)$ denotes the cdf of a central chi-square distribution with $\nu = 2n$ degrees of freedom.

Proof. Due to the well-known basic property of the probability-integral transform, we can write for any $i \in \{1, \ldots, n\}$: $-\log F_0(X_i) \overset{d}{=} -\log U_i^{1/\theta} = \theta^{-1}\log (1/U_i)$, with $U_i \sim \mathcal{U}(0, 1)$ and $\overset{d}{=}$ denoting equality in distribution. The distribution of $\log(1/U_i)$ has cdf

$$P[\log(1/U_i) \leq \bar{u}] = P[U_i \geq e^{-\bar{u}}] = 1 - e^{-\bar{u}} = G_{1,1}(\bar{u}) \quad \forall \bar{u} \geq 0,$$

with $G_{1,1}(\cdot)$ denoting the cdf of a gamma distribution with parameters $(1, 1)$.

$$\Rightarrow \; P_{\theta}^t[\hat{\theta}_n \leq t] = P_{\theta}^t\left[\sum_{i=1}^{n} -\log F_0(X_i) \geq n/t\right]$$

$$= P_{(0,1)}^t\left[\theta^{-1}\sum_{i=1}^{n}\log(1/U_i) \geq n/t\right]$$

$$= P_{(0,1)}^t\left[\sum_{i=1}^{n}\log(1/U_i) \geq \theta n/t\right]$$

$$\overset{(7)}{=} 1 - G_{n,1}(\theta n/t) = 1 - F_{2\nu}(2\theta n/t), \; t > 0.$$
The following proposition states that there is an optimum test for (5) and describes the computational steps to be taken in order to carry out that procedure.

**Proposition 1.** Let the common distribution of the $X_i$ under $\theta = 1$ be absolutely continuous with density $f_0(\cdot)$. Then, there exists a uniformly most powerful test for (5) which rejects the null hypothesis $H$ of relevant deviations of the true from the hypothesized distribution if and only if it turns out that

$$C_n(\alpha, \varepsilon) < -n/\sum_{i=1}^{\infty} \log[F_0(X_i)] < C_n^c(\alpha, \varepsilon).$$  \hspace{1cm} (8)

The critical constants $C_n(\alpha, \varepsilon)$, $C_n^c(\alpha, \varepsilon)$ have to be determined through solving the equations

$$F_{\frac{\alpha}{2n}}(2(1+\varepsilon)^{-1}n/c_1) - F_{\frac{\alpha}{2n}}(2(1+\varepsilon)^{-1}n/c_2) = \alpha$$

$$= F_{\frac{\alpha}{2n}}(2(1+\varepsilon)n/c_1) - F_{\frac{\alpha}{2n}}(2(1+\varepsilon)n/c_2), \quad 0 < c_1 < c_2 < \infty.$$ \hspace{1cm} (9)

**Proof.** For absolutely continuous $F_0, X_i \sim F_0^{(n)} \forall i$ implies that the joined density $g_0(n)$, say, of the sample $(X_1, \ldots, X_n)$ is given by

$$g_0(n)(x_1, \ldots, x_n) = \theta^n \prod_{i=1}^{n} (F_0(x_i))^{\theta-1} f_0(x_i) = \theta^n \cdot \exp\{\theta \sum_{i=1}^{n} \log F_0(x_i)\} \cdot \prod_{i=1}^{n} f_0(x_i).$$

Thus, $g_0(n)$ is an element of a 1-parameter exponential family, with $\sum_{i=1}^{n} \log F_0(X_i)$ as a sufficient statistic for $\theta$. Hence, a well-known theorem on the existence of UMP tests for equivalence hypotheses about parameters of families of that structure (cf. [10], Appendix A.1) applies, according to which a UMP level-$\alpha$ test for $(H, K)$ has rejection region

$$C = \{(x_1, \ldots, x_n) \in \mathbb{R}^n | (\sum_{i=1}^{n} \log F_0(x_i)) \in (\tilde{c}_1, \tilde{c}_2)\},$$

where $(\tilde{c}_1, \tilde{c}_2)$ solves the equations

$$P_{\theta_1}[(\sum_{i=1}^{n} \log F_0(X_i)) \in (\tilde{c}_1, \tilde{c}_2)] = \alpha = P_{\theta_2}[(\sum_{i=1}^{n} \log F_0(X_i)) \in (\tilde{c}_1, \tilde{c}_2)],$$ \hspace{1cm} (10)

with $\theta_1 = 1/(1+\varepsilon)$, $\theta_2 = 1 + \varepsilon$. In view of $\hat{\theta}_n \equiv -n/\sum_{i=1}^{n} \log F_0(X_i)$, Lemma 1 implies that a unique solution of (10) is obtained by setting $\tilde{c}_i = -n/C_n^{(i)}(\alpha, \varepsilon), \nu = 1, 2$, with $C_n^{(i)}(\alpha, \varepsilon), k = 1, 2$, being defined as in (8), (9).

**Remark 1.** The optimal critical constants $C_n(\alpha, \varepsilon)$, $C_n^c(\alpha, \varepsilon)$ are not given explicitly but have to be calculated from (9) by means of a suitable numerical algorithm. All results to be presented in the subsequent sections were obtained by means of the program provided as Supplementary Material under the name UMPTestForGoF both as SAS/IML- and R-script.

**Remark 2.** In terms of both the algorithm for determining the critical constants and its power, the UMP test given by (8) is completely distribution-free. The cdf $F_0$ defining the model whose goodness-of-fit one wants to establish, is only used for computing the test statistic.

**Remark 3.** Transforming the sufficient statistic $\sum_{i=1}^{n} \log F_0(X_i)$ to the equivalent statistic $\hat{\theta}_n(F_0) = -n/\sum_{i=1}^{n} \log F_0(X_i)$ in writing down the decision rule of the UMP test, is simply a matter of conceptual convenience: $\hat{\theta}_n(F_0)$ is easily shown to be the ordinary ML estimator of the Lehmann parameter $\theta$ and thus a quantity straightforward to interpret.
4 Numerical results on the UMP test for goodness-of-fit

Table 2 gives a tabulation of the critical values and the power against the null alternative \( \theta = 1 \) of the exact UMP test derived in Section 3 for three different choices of the equivalence margin \( \varepsilon \) and sample sizes \( n \) ranging from 10 to 200. As usual, the significance level \( \alpha \) is chosen to be 0.05 throughout. Comparing the entries in different lines of the same block of the table reveals the effect of increasing the sample size on basic characteristics of the test: The left- and right-hand limit of the critical interval which has to be checked for inclusion of the observed value of the ML estimator \( \hat{\theta}_n(F_0) \), is monotonically decreasing, and increasing, respectively, in \( n \), in a way making the corresponding intervals a nested sequence of sets. Furthermore, as is mandatory for any reasonable test for the problem put forward in (5), the power increases likewise monotonically in \( n \) to unity. The effect of increasing the equivalence margin \( \varepsilon \) becomes obvious from comparisons between homologous entries in the different blocks of the table. As has clearly to be expected, the critical interval is shrinking in length as \( \varepsilon \) decreases, and the power attainable with a given sample size declines fairly rapidly as the margin is tightened.

Recalling the standard approaches to testing for lack-of-fit of a fully specified distribution to that underlying a given dataset, it seems natural to compare the new test with a goodness-of-fit testing procedure which uses grouped data. In the context of lack-of-fit testing, it is often recommended (see, e.g., \[5\], §25.22) to choose for grouping classes of equal probability in terms of the distribution \( F_0 \) under assessment. Focusing on the best known and perhaps most interesting special case that the distribution one aims to fit is the standard normal \( \mathcal{N}(0, 1) \) so that there holds \( F_0 = \Phi \), and that the number \( k \), say, of classes to be formed equals 5, a partition of the range of \( X \) which is in line with that recommendation, is given by the intervals \((-\infty, -0.8416], (-0.8416, -0.2533], (-0.2533, 0.2533], (0.2533, 0.8416], (0.8416, \infty)\). If the equivalence margin for the Lehmann parameter is chosen to be \( \varepsilon = 0.5077 \) (corresponding to a maximal allowable vertical distance of \( \delta = 0.15 \) between \( \Phi^\theta \) and \( \Phi \)), the probability masses, say \( p_j(1+\varepsilon) \), of these intervals under \( \theta = 1 + \varepsilon \) are computed to be

\[
\begin{array}{ccccc}
 j & 1 & 2 & 3 & 4 & 5 \\
 p_j(1+\varepsilon) & 0.08834 & 0.16286 & 0.21173 & 0.25138 & 0.28569 \\
\end{array}
\]

Table 2. Critical constants and power of the UMP test at level \( \alpha = 0.05 \) to detect perfect fit of the model \[\theta = 1\] for various sample sizes and choices of the equivalence margin.

| \( \varepsilon \) | \( n \) | \( C_1^\varepsilon(x, \varepsilon) \) | \( C_2^\varepsilon(x, \varepsilon) \) | \( \text{POW}(1) \) |
|---|---|---|---|---|
| 0.3135 | 10 | 0.9836 | 1.0423 | 0.0723 |
| " | 20 | 0.9830 | 1.0427 | 0.1043 |
| " | 50 | 0.9593 | 1.0671 | 0.2919 |
| " | 100 | 0.9047 | 1.1227 | 0.7183 |
| " | 200 | 0.8587 | 1.1737 | 0.9727 |
| 0.5077 | 10 | 0.9821 | 1.0767 | 0.1142 |
| " | 20 | 0.9560 | 1.1039 | 0.2496 |
| " | 50 | 0.8510 | 1.1215 | 0.7869 |
| " | 100 | 0.7883 | 1.2887 | 0.9857 |
| " | 200 | 0.7481 | 1.3473 | 1.0000 |
| 0.7341 | 10 | 0.9629 | 1.1450 | 0.2118 |
| " | 20 | 0.8670 | 1.2444 | 0.5738 |
| " | 50 | 0.7400 | 1.3946 | 0.9743 |
| " | 100 | 0.6854 | 1.4822 | 0.9998 |
| " | 200 | 0.6504 | 1.5496 | 1.0000 |

https://doi.org/10.1371/journal.pone.0256499.t002
The goodness-of-fit test for multinomial distributions established in Ch. 9.1 of [10] defines equivalence between multinomial distributions of common dimension \( k \) in terms of the Euclidean distance between the corresponding parameter vectors \( \pi \) and \( \pi' \), where \( \pi \) refers to the unknown distribution underlying the data and \( \pi' \) to the model to be fitted. With the grouped-data problem under consideration, we have \( \pi_j = 1/5 \forall j = 1, \ldots, 5 \), and the Euclidean distance of the vector \( \pi(1+\varepsilon) \) with the above listed components from \( \pi' \) is computed to be \( d(\pi, \pi') = \varepsilon = 0.1548 \). On the other hand, \( \pi(1+\varepsilon) \) is the parameter vector of the multinomial distribution into which the uniform distribution on \( \{1, \ldots, 5\} \) generated from \( N(0, 1) \) by means of the chosen partition, is mapped through moving \( \theta \) to the (right-hand) boundary of its equivalence range. Thus, it seems reasonable to consider the problem

\[
H_0 : d(\pi, \pi') \geq \varepsilon = 0.1548 \text{ versus } H_1 : d(\pi, \pi') < 0.1548 \quad (11)
\]

concerning the parameter \( \pi \) of a multinomial distribution over \( \{1, \ldots, 5\} \), as the grouped-data analogue of the testing problem to which the results shown in the middle block of Table 2 relate. As shown in Wellek, loc. cit., an asymptotically valid test of (11) is given by the following decision rule:

\[
\text{Reject } H_0 \iff d^2(\hat{\pi}, \pi') < \varepsilon^2 - u_{1-\alpha} v_s(\pi, \pi')/\sqrt{n}, \quad (12)
\]

where

\[
v^t_s(\hat{\pi}, \pi') = 4 \left[ \sum_{j=1}^{k} (\hat{\pi}_j - \pi'_j)^2 \hat{\pi}_j - \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} (\hat{\pi}_{j_1} - \pi'_{j_1})(\hat{\pi}_{j_2} - \pi'_{j_2})\hat{\pi}_{j_1}\hat{\pi}_{j_2} \right], \quad (13)
\]

and \( \hat{\pi} \) denotes the vector of relative frequencies of observations falling in the different classes. For small to moderate sample sizes, the power of the grouped-data goodness-of-fit test given by (12) and (13) against the null alternative \( \pi = \pi' \) can be computed even exactly. With \( k = 5 \), \( \pi' = (1/5, \ldots, 1/5) \), \( \varepsilon = 0.1548 \), one finds that the power values computed for the exact UMP test with the non-grouped continuous data reduce to

| \( n \) | 10 | 20 | 50 | 100 | 200 |
|---|---|---|---|---|---|
| \( \text{POW}(1) \) | 0.0116 | 0.0032 | 0.2258 | 0.7015 | 0.9856 |

Thus, the fact being well known (see, e.g., [14], Ch.27) for the lack-of-fit case that grouping entails substantial losses in efficacy, has also to be stated for problems of testing for goodness-of-fit.

5 A glance at options for testing for goodness-of-fit of distributions involving nuisance parameters

Upon noticing the results presented in Sections 3 and 4, it seems natural to ask the question whether the approach admits generalization to settings where one aims to fit a distribution involving unknown nuisance parameters rather than being fully specified. The best known special case of such a problem is testing for normality, which is to say that the hypothesis of interest states that, except for differences one is willing to accept, the distribution underlying the data has cdf \( F = F_0\left(\frac{X}{\sigma}\right) \), with \( F_0 = \Phi \) and arbitrary \((\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\). Adapting the hypothesis formulation proposed in the case of a fully specified distributional model to this setting is
straightforward, leading to consider the testing problem

\[ H : (\theta, \mu, \sigma) \in ((-\infty, 1/(1+\varepsilon)] \cup [1 + \varepsilon, \infty)) \times \mathbb{R} \times \mathbb{R}_+ \]

versus \n\[ K : (\theta, \mu, \sigma) \in ((1/(1+\varepsilon), 1 + \varepsilon)) \times \mathbb{R} \times \mathbb{R}_+ . \]  

(14)

As before, \( \theta \) denotes the Lehmann parameter inducing potential deviations of the true distribution from the distribution of the form assumed under the model to be fitted.

A promising and often successful approach to the construction of equivalence tests in multi-parameter families of distributions uses the maximum likelihood estimator of the parameter of interest as pivotal quantity (for the theoretical basis of that approach see [10], Ch. 3.4). The log-likelihood function and its first-order derivatives associated with a sample (\( X_1, \ldots, X_n \)) of i.i.d. observations from \([F_0 \left( \frac{-\theta}{\sigma} \right)]^0\) are readily obtained to be

\[ l(x_1, \ldots, x_n; \theta, \mu, \sigma) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log F_0 \left( \frac{x_i - \mu}{\sigma} \right) + \sum_{i=1}^{n} \log \left( \frac{1}{\sigma} f_0 \left( \frac{x_i - \mu}{\sigma} \right) \right) . \]  

(15)

\[ \frac{\partial l}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \log \frac{x_i - \mu}{\sigma} \]  

(16.a)

\[ \frac{\partial l}{\partial \mu} = -\frac{1}{\sigma} \sum_{i=1}^{n} \left( \frac{f_0}{F_0} \right) \left( \frac{x_i - \mu}{\sigma} \right) - \frac{1}{\sigma} \sum_{i=1}^{n} \left( \frac{f_0}{F_0} \right) \left( \frac{x_i - \mu}{\sigma} \right) \]  

(16.b)

\[ \frac{\partial l}{\partial \sigma} = -\frac{1}{\sigma} \sum_{i=1}^{n} \left( \frac{f_0}{F_0} \right) \left( \frac{x_i - \mu}{\sigma} \right) \cdot \left( \frac{x_i - \mu}{\sigma} \right) - \frac{n}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} \left( \frac{f_0}{F_0} \right) \left( \frac{x_i - \mu}{\sigma} \right) \cdot \left( \frac{x_i - \mu}{\sigma} \right) . \]  

(16.c)

Solving the corresponding system of score equations by means of the Newton-Raphson algorithm or an alternative numerical technique is an easy exercise, and the roots almost surely exist. The same cannot be said of the maximum likelihood estimator: examining the function \((\theta, \mu, \sigma) \mapsto l(x_1, \ldots, x_n; \theta, \mu, \sigma)\) for various datasets revealed that it fails to attain almost surely a global maximum in the interior of the parameter space. Hence, carrying out a construction requiring that the MLE of \( \theta \) exists and is asymptotically normal, is not practicable here.

In contrast to the MLE, the score statistic \( U(\theta) \equiv l(x_1, \ldots, x_n; \hat{\theta}, \hat{\mu}(\theta), \hat{\sigma}(\theta)) \) with \((\hat{\mu}(\theta), \hat{\sigma}(\theta))\) as the solution to \( \frac{\partial U}{\partial \mu} = 0, \frac{\partial U}{\partial \sigma} = 0 \) for fixed \( \theta \), is almost surely well defined. However, no approach to basing tests for interval hypotheses on a statistic of this form is at hand. An option for making use of \( U(\theta) \) anyway for the purpose in mind, is to split up the equivalence testing problem (14) into the two one-sided testing problems

\[ H_{1l} : (\theta, \mu, \sigma) \in (\infty, \frac{1}{1+\varepsilon}) \times \mathbb{R} \times \mathbb{R}_+ \quad \text{vs.} \quad K_{1l} : (\theta, \mu, \sigma) \in (\frac{1}{1+\varepsilon}, \infty) \times \mathbb{R} \times \mathbb{R}_+ . \]  

(17.1)

and

\[ H_{1r} : (\theta, \mu, \sigma) \in (1+\varepsilon, \infty) \times \mathbb{R} \times \mathbb{R}_+ \quad \text{vs.} \quad K_{1r} : (\theta, \mu, \sigma) \in (-\infty, 1+\varepsilon) \times \mathbb{R} \times \mathbb{R}_+ . \]  

(17.2)

Rejecting \( H_{1l} \) and \( H_{1r} \), when it turns out that there holds \( U(1/(1+\varepsilon))/\hat{\mathcal{V}}_i > z_{\gamma} \) and \( U(1+\varepsilon)/\hat{\mathcal{V}}_i < z_{\gamma} \), respectively, yields asymptotically valid tests for (17.1) and (17.2), provided \( \hat{\mathcal{V}}_{i1}^2 \) consistently estimates the asymptotic variance of \( U(1/(1+\varepsilon)) \) under \( \theta = 1/(1+\varepsilon) \) and \( \hat{\mathcal{V}}_{i1}^2 \) that of \( U(1+\varepsilon) \) under \( \theta = 1 + \varepsilon \). As usual, \( z_\gamma \) stands for the \( \gamma \)-quantile of \( \mathcal{N}(0, 1) \) for arbitrary \( \gamma \in (0, 1) \).
In order to avoid the possible inconsistency of the score test pointed out by D.A. Freedman [15], the variance estimators \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \) should be based on the expected covariance matrix \( V(\theta, \mu, \sigma) \), say, of the score statistics \( \left( \frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \mu}, \frac{\partial}{\partial \sigma} \right) \) evaluated at \((\theta, \mu, \sigma) = (1/(1 + \varepsilon), \hat{\mu}(1/(1 + \varepsilon)), \hat{\sigma}(1/(1 + \varepsilon)))\) and \((\theta, \mu, \sigma) = (1 + \varepsilon, \hat{\mu}(1 + \varepsilon), \hat{\sigma}(1 + \varepsilon))\), respectively. No explicit expressions are available for the elements of \( V(\theta, \mu, \sigma) \). However, the expected value of the random variables corresponding to the expressions appearing on the right-hand side of Eq (16.a-c) can be computed sufficiently fast and with high degree of accuracy by means of standard numerical integration techniques (like Gauss-Legendre quadrature), and the same holds true for the squares of these variables and thus also for the entries in \( V(\theta, \mu, \sigma) \). Afterwards, \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \) is computed as the reciprocal (1,1)-element of \( V(1/(1 + \varepsilon), \hat{\mu}(1/(1 + \varepsilon)), \hat{\sigma}(1/(1 + \varepsilon)))^{-1} \) and \( V(1 + \varepsilon, \hat{\mu}(1 + \varepsilon), \hat{\sigma}(1 + \varepsilon))^{-1} \), respectively.

The last step of the construction consists in combining the two score tests for the one-sided problems (17.1) and (17.2) into a test for the two-sided equivalence problem (14) to be solved when interest is in establishing goodness-of-fit. The combined test rejects the null hypothesis \( H \) of (14) if and only if both of the critical inequalities \( U(1/(1 + \varepsilon))/\hat{v}_1 > z_{\alpha/2} \) and \( U(1 + \varepsilon)/\hat{v}_2 < z_{\alpha/2} \) are found to hold true. According to a well-known and frequently exploited principle from the theory of equivalence testing (cf. [10, 16]; Ch. 7.1), the asymptotic validity of the one-sided tests with rejection regions \{ \( U(1/(1 + \varepsilon))/\hat{v}_1 > z_{\alpha/2} \) \} and \{ \( U(1 + \varepsilon)/\hat{v}_2 < z_{\alpha/2} \) \} implies that the test for goodness-of-fit of a distribution from the family \( \{ F_0(\frac{\longrightarrow}{\mu, \sigma}) \}_{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+} \) carried-out in this way, is likewise asymptotically valid in terms of the significance level.

The facts stated so far admit the conclusion that in theory, testing for goodness-of-fit of distributional models involving unknown parameters is not an insurmountable challenge for statistical inference. However, upon studying the power of such a test, this judgement can hardly be maintained: Determining the rejection probability of the double one-sided score-test procedure described above by means of Monte Carlo simulation of normally distributed data reveals that even against the null alternative of perfect fit of the model \( \theta = 1 \) and for a choice of the equivalence margin of moderate strictness \( \varepsilon = 0.5077, \) corresponding to a maximum acceptable vertical distance of the cdf’s of \( \delta = 0.15 — \text{Recall Table 1} \), several thousands (!) of observations are required in order to attain a power of 80%. Thus, establishing goodness-of-fit of some distributional shape rather than a specific element of the corresponding family of distributions (like \( \mathcal{N}(\mu, \sigma^2) \)) by means of a testing procedure providing satisfactory control over both kinds of error-risk, is hardly an option for practice.

### 6 Illustrating example

Fig 2 shows the plot of the empirical cdf of a simulated random sample of size \( n = 100 \) from \( \mathcal{N}(0, 1) \), together with the theoretical cdf \( \Phi \) to be assessed for goodness-of-fit to these data.

In order to apply the UMP test of Section 3 with \( F_0 = \Phi \), the maximum likelihood estimate of the Lehmann parameter \( \theta \) has to be determined. With the data behind the empirical cdf plotted above and \( F_0 = \Phi \), evaluating Eq (6) yields \( \hat{\theta}_a(F_0) = 1.0793 \). If the equivalence margin for \( \theta \) is chosen to be \( \varepsilon = 0.5077 \), it is seen from Table 2 that the observed value of \( \hat{\theta}_a(F_0) \) has to be checked for inclusion in the interval \((0.7883, 1.2887)\). Since 1.0793 is an inner point of this interval, the conclusion is that the goodness-of-fit test for standard normality at level \( \alpha = 0.05 \) leads under the chosen specifications to rejecting the null hypothesis of lack-of-fit.

Assuming next that the distribution one wants to fit is \( \mathcal{N}(\mu, \sigma^2) \) with both parameters being unknown and keeping the equivalence limits for \( \theta \) set at \( 1/(1 + \varepsilon) \) and \( 1 + \varepsilon \) with \( \varepsilon = 0.5077 \), carrying out the double one-sided score test described in Section 5 has to start with...
computing the restricted MLE’s of \((\mu, \sigma)\) given \(\theta = 1/(1 + \epsilon)\) and \(\theta = 1 + \epsilon\). From the dataset under analysis, these estimates are obtained to be \((\hat{\mu}, \hat{\sigma}) (1/(1 + \epsilon)) = (0.3760, 0.8434)\) and \((\hat{\mu}, \hat{\sigma})(1 + \epsilon) = (-0.3308, 1.0636)\). Calculating the corresponding efficient scores of \(\theta\) gives \(U(1/(1 + \epsilon)) = 0.4444\) and \(U(1 + \epsilon) = 2.2406\), with estimated asymptotic variances \(\tilde{\nu}_l^2 = 1.636889\) and \(\tilde{\nu}_r^2 = 0.276968\). Finally, standardizing the two score statistics yields \(U(1/(1 + \epsilon))/\tilde{\nu}_l = 1.751258\) and \(U(1 + \epsilon)/\tilde{\nu}_r = 0.844329\) of which only the first one leads to rejecting the corresponding one-sided null hypothesis whereas \(U(1 + \epsilon)/\tilde{\nu}_r\) is much larger than \(z_\alpha = -1.64485\). Hence, testing for normal shape of the distribution underlying the dataset fails to lead to a decision in favor of the alternative hypothesis of the corresponding equivalence problem (14).

7 Discussion

As major strengths of the procedure obtained in the core of this paper for testing for goodness-of-fit of a fully specified distribution of the continuous type to the distribution underlying a given random sample, the following facts can be adduced:

(i). The alternative hypothesis which can be declared established upon a positive result, states that the model fits the data sufficiently well rather than meriting rejection because of marked discrepancies from the true distribution.

(ii). The method is fully exact and satisfies the strongest of the optimality criteria having been in use for hypothesis tests since the beginnings of classical frequentist inference.

(iii). Due to the availability of a fairly compact source code both in SAS/IML and R, the practical implementation of the test, as well as the algorithm for exact power and sample size computation, is fast and easy.
(iv). The primary metric, in terms of which the region of distribution functions defined equivalent to the distribution function specified by the model, is fully intuitive also for applied research workers without advanced statistical training. It is the same in terms of which alternatives to the null hypothesis of perfect fit of the model are expressed in the Kolmogorov test.

Admittedly, the last of these advantages comes into play only as long as one is willing to rely on the semiparametric model which assumes that the true distribution function underlying the data differs from that specified by the model through taking all values of the latter to the $\theta$-th power. Except for ignoring right-censoring and applying the transformation $u \mapsto u^\theta$ to cumulative distribution rather than survivor functions, this model is the same as Cox’s [13] well-known proportional hazards model. If the proposed test does not lead to a decision in favor of equivalence between actual and hypothesized distribution, one cannot rule out the possibility that the true distribution differs from $F_0$ nowhere by more than a given margin $\delta$ without satisfying the modified Cox model.

### Supporting information

S1 File.  
(R)

S2 File.  
(SAS)

S1 Dataset. Sample used for illustration.  
(XLSX)

### Author Contributions

Conceptualization: Stefan Wellek.

Formal analysis: Stefan Wellek.

Investigation: Stefan Wellek.

Methodology: Stefan Wellek.

Software: Stefan Wellek.

Supervision: Stefan Wellek.

Validation: Stefan Wellek.

Visualization: Stefan Wellek.

Writing – original draft: Stefan Wellek.

Writing – review & editing: Stefan Wellek.

### References

1. Cramér H. Mathematical Methods of Statistics. Princeton: Princeton University Press; 1946.
2. Lehmann EL. Testing Statistical Hypotheses. New York: John Wiley; 1959.
3. Lehmann EL. Testing Statistical Hypotheses. Second Edition. New York: John Wiley; 1986.
4. Lehmann EL, Romano JP. Testing Statistical Hypotheses. Third Edition. New York: Springer; 2005.
5. Kendall MG, Stuart A, Ord K, Arnold S. Kendall’s Advanced Theory of Statistics. Volume 2A—Classical Inference and Relationship. London: Edward Arnold; 1999.
6. Neill JW, Johnson DE. Testing for Lack of Fit in Regression—a Review. COMMUN STAT A-THEOR. 1984; 13(4):485–511. https://doi.org/10.1080/03610928408828696

7. Christensen R. Lack-of-Fit Tests Based on Near or Exact Replicates. ANN STAT. 1989; 17(2):673–683. https://doi.org/10.1214/aos/1176347133

8. Su JQ, Wei LJ. A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model. J AM STAT ASSOC. 1991; 86(414):420–426. https://doi.org/10.1080/01621459.1991.10475059

9. Miller FR, Neill JW. Lack of Fit Tests for Linear Regression Models with Many Predictor Variables Using Minimal Weighted Maximal Matchings. J MULTIVARIATE ANAL. 2016; 150:14–26. https://doi.org/10.1016/j.jmva.2016.05.005

10. Wellek S. Testing Statistical Hypotheses of equivalence and noninferiority. Second Edition. Boca Raton: CRC Press; 2010.

11. Lehmann EL. The power of rank tests. ANN MATH STAT. 1953; 14:35–44.

12. Wellek S. A log-rank test for equivalence of two survivor functions. Biometrics. 1993; 49:877–81. https://doi.org/10.2307/2532208

13. Cox DR. Regression Models and Life Tables (with Discussion). J ROY STAT SOC B. 1972; 34:187–220.

14. DasGupta A. Asymptotic Theory of Statistics and Probability. New York: Springer; 2008.

15. Freedman DA. How can the score test Be inconsistent? TAS. 2007; 61:291–95.

16. Berger RL. Multiparameter hypothesis testing and acceptance sampling. Technometrics. 1982; 24:295–300. https://doi.org/10.2307/1267823

---

Testing for goodnes s rather than lack of fit of continuous probability distributions

---

6. Neill JW, Johnson DE. Testing for Lack of Fit in Regression—a Review. COMMUN STAT A-THEOR. 1984; 13(4):485–511. https://doi.org/10.1080/03610928408828696

7. Christensen R. Lack-of-Fit Tests Based on Near or Exact Replicates. ANN STAT. 1989; 17(2):673–683. https://doi.org/10.1214/aos/1176347133

8. Su JQ, Wei LJ. A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model. J AM STAT ASSOC. 1991; 86(414):420–426. https://doi.org/10.1080/01621459.1991.10475059

9. Miller FR, Neill JW. Lack of Fit Tests for Linear Regression Models with Many Predictor Variables Using Minimal Weighted Maximal Matchings. J MULTIVARIATE ANAL. 2016; 150:14–26. https://doi.org/10.1016/j.jmva.2016.05.005

10. Wellek S. Testing Statistical Hypotheses of equivalence and noninferiority. Second Edition. Boca Raton: CRC Press; 2010.

11. Lehmann EL. The power of rank tests. ANN MATH STAT. 1953; 14:35–44.

12. Wellek S. A log-rank test for equivalence of two survivor functions. Biometrics. 1993; 49:877–81. https://doi.org/10.2307/2532208

13. Cox DR. Regression Models and Life Tables (with Discussion). J ROY STAT SOC B. 1972; 34:187–220.

14. DasGupta A. Asymptotic Theory of Statistics and Probability. New York: Springer; 2008.

15. Freedman DA. How can the score test Be inconsistent? TAS. 2007; 61:291–95.

16. Berger RL. Multiparameter hypothesis testing and acceptance sampling. Technometrics. 1982; 24:295–300. https://doi.org/10.2307/1267823