Statistica Sinica **28** (2018), 2823-2840
doi:https://doi.org/10.5705/ss.202016.0507

ON P-VALUES

Laurie Davies

University of Duisburg-Essen

Abstract: In statistics P-values are mostly used in the context of hypothesis testing. Software for linear regression assigns a P-value to every covariate which corresponds to testing the hypothesis that the 'true' value of the regression coefficient is zero. In this paper several different uses and interpretations of P-values will be discussed ranging from the use of P-values as measures of approximation for parametric models, for location-scale M-functionals to Jeffreys' criticism of P-values and to the choice of covariates in linear regression without an error term. The approach is neither frequentist nor Bayesian. It is not frequentist as the P-values are calculated and interpreted for the data at hand, and simply being a P-value makes it non-Bayesian.

Key words and phrases: Approximate models, approximation regions, choice of covariates, functionals, prediction, P-values and approximation.

1. Peter Hall

I met Peter Hall for the first time in 1996. Robert Staudte had invited me to La Trobe where we discussed my ideas on statistics (Davies (1995)). Staudte decided to arrange a one-day workshop and invited Peter Hall to give a talk. After that I met Peter several times at various conferences and workshops. He did not agree with my approach but nevertheless encouraged me to continue, one of the very few people who did so. In particular he invited me to contribute Davies (2008) to the Journal of the Korean Statistical Society and arranged for the discussants. In 2010 we were colleagues for a short time at the University of California, Davis, and in 2013 he invited me to Melbourne for a six-week stay. That was the last time I saw him. We have lost a wonderful colleague and a wonderful man.

2. Introduction

All branches of knowledge which require the analysis of data make use of P-values. Unfortunately in many cases 'make use of' could be replaced by 'abuse', the many reports of widespread abuse are convincing. In response The American

Statistician published a statement on P-values by the American Statistical Association together with supplementary material consisting of statements by several statisticians and philosophers (Wasserstein and Lazar (2016)).

The most detailed of the supplementary material is Greenland et al. (2016). The authors point out that there are many ways in which any usefulness of a P-value can be invalidated. One example is to perform several experiments and report only the one with the smallest P-value. Problems of this nature will not be discussed here. The paper will be concerned only with the relationship between the data and a posited model or posited parameter values.

All figures were produced using R R Core Team (2014). The calculations were done with FORTRAN 77 source code.

3. Probability Models and Approximation

There are two meaning of the word 'model' is statistics. The first meaning refers to a parametric family of distributions, for example the normal model as the family of all normal distributions. This meaning of the word 'model' is common in much of statistics, in particular in Bayesian statistics where such models are the objects of study.

The second meaning is that of a single probability measure. In this sense of the word the $N(0,1)$ probability measure is one model, the $N(0,2)$ probability measure is another model. This is the sense in which the word will be used in this paper. Models in this sense are the atoms so to speak of probability theory and hence the basic objects of stochastic modelling. The meaning of the word 'model' in the first sense is a parametric family of models in the second sense.

The two different meanings of the word 'model' are not just a question of notation or definition. They reflect different approaches to statistics. This may be seen in Birnbaum (1962) where a parametric family of probability measures has to be adequate without specifying the adequacy of any individual measure. This is necessary as the Likelihood Principle requires the proportionality of two different densities for all values of the parameter and not just for the adequate ones. A similar problem occurs when testing hypotheses. The parametric model is declared adequate without specifying the adequate values of the parameter. A hypothesis $H_0 : \mu = \mu_0$ is then tested to see whether $\mu = \mu_0$ is consistent with the data. It only makes sense to do this if the adequate parameter values have not been specified when declaring the whole family to be adequate as otherwise the test would be superfluous.

A model P is an adequate approximation to data $\mathbf{x}_n$ if typical data sets generated under P look like $\mathbf{x}_n$ (Davies (2014)). To make this susceptible to mathematical analysis the term 'look like' must be expressible in numerical quantities. This may not always be possible or easy; see Neyman, Scott and Shane (1953, 1954); Buja et al. (2009), Chapter 5.8 of Huber (2011) and Davies and Krämer (2016). In the following it will be assumed that 'look like' has a precise mathematical expression.

Given a probability measure P a sample of size n generated under P will be denoted by $\mathbf{X}_n(P) = (X_1(P), \ldots, X_n(P))$. Given further a number $\alpha, 0 < \alpha \leq 1$, denote by $E_n(P, \alpha)$ a subset of $\mathbb{R}^n$ such that

$$\mathbf{P}(\mathbf{X}_n(P) \in E_n(P, \alpha)) = \alpha\,. \tag{3.1}$$

The interpretation is that typical samples $\mathbf{X}_n(P)$ generated under P lie in $E_n(P, \alpha)$. A sample $\mathbf{x}_n$, not necessarily generated under P, will look like a typical sample $\mathbf{X}_n(P)$ if $\mathbf{x}_n \in E_n(P, \alpha)$. The choice of $E_n(P, \alpha)$ depends on those statistics of the data regarded as important for the analysis to be conducted.

Given a family $\mathcal{P}$ of probability measures an α-approximation region for the data $\mathbf{x}_n$ is defined by

$$\mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{P}) = \{P \in \mathcal{P} : \mathbf{x}_n \in E_n(P, \alpha)\}. \tag{3.2}$$

It follows that for each $P \in \mathcal{P}$

$$\mathbf{P}(P \in \mathcal{A}(\mathbf{X}_n(P), \alpha, \mathcal{P})) = \mathbf{P}(\mathbf{X}_n(P) \in E_n(P, \alpha)) = \alpha \tag{3.3}$$

for all $P \in \mathcal{P}$ so that $\mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{P})$ has the property of a confidence region.

The definition (3.2) makes no assumption that the data $\mathbf{x}_n$ were generated under some model $P \in \mathcal{P}$. The interpretation is that $\mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{P})$ specifies those models P for which $\mathbf{x}_n$ 'looks like' a 'typical sample' $\mathbf{X}_n(P)$ generated under P. It is possible for an approximation region to be empty which is not the case for confidence regions.

As an example consider the family of normal distributions $\mathcal{N} = \{(\mu, \sigma) : N(\mu, \sigma^2)\}$. An approximation region can be based on the mean, the variance, outliers and the distance of the empirical measure to the model $N(\mu, \sigma^2)$ as measured by the Kuiper metric d_{ku} (Kuiper (1962)). More precisely put $\mathbf{y}_n = (\mathbf{x}_n - \mu)/\sigma$ and

$$\begin{cases} S_1(\mathbf{y}_n) = \sqrt{n}\,|\mathrm{mean}(\mathbf{y}_n)|, & S_2(\mathbf{y}_n) = \sum_{i=1}^n y_i^2, \\ S_3(\mathbf{y}_n) = \max_i |y_i|, & S_4(\mathbf{y}_n) = d_{\mathrm{ku}}(\mathbb{P}(\mathbf{y}_n), N(0,1)), \end{cases} \tag{3.4}$$

where $\mathbb{P}(\mathbf{y}_n)$ is the empirical measure based on $\mathbf{y}_n$. For each of the four statistics

one can determine the P-values given by

$$\begin{cases} p_1(\mu,\sigma) = \mathbf{P}(S_1(\mathbf{Y}_n) \geq S_1(\mathbf{y}_n)) = 2(1 - \text{pnorm}(S_1(\mathbf{y}_n))), \\ p_2(\mu,\sigma) = 2\min(p, 1-p), p = \mathbf{P}(S_2(\mathbf{Y}_n) \leq S_2(\mathbf{y}_n)) = \text{pchisq}(S_2(\mathbf{y}_n), n), \\ p_3(\mu,\sigma) = \mathbf{P}(S_3(\mathbf{Y}_n) \geq S_3(\mathbf{y}_n)) = 1 - (2\text{pnorm}(S_3(\mathbf{y}_n)) - 1)^n, \\ p_4(\mu,\sigma) = \mathbf{P}(S_4(\mathbf{Y}_n) \geq S_4(\mathbf{y}_n)) = 1 - \text{pkuip}(S_4(\mathbf{y}_n)), \end{cases} \tag{3.5}$$

where $\mathbf{Y}_n$ are i.i.d. $N(0,1)$. The P-values $\text{pkuip}(q)$ for the Kuiper metric are the asymptotic P-values, see for example Proposition 12.3.6 of Dudley (1989). The approximation region is then defined by

$$\begin{aligned} \mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{N}) = \{(\mu,\sigma) : p_1(\mu,\sigma) \geq p_1, p_2(\mu,\sigma) \geq p_2, \\ p_3(\mu,\sigma) \geq p_3, p_4(\mu,\sigma) \geq p_4\}, \end{aligned} \tag{3.6}$$

where the $p_1, \ldots, p_4$ are chosen to give an α-approximation region. Choosing the p_i to satisfy $p_1 + \cdots + p_4 = 1 - \alpha$ gives an α^*-approximation region with $\alpha^* > \alpha$. The value of α^* can be determined by simulations. A better approximation to an α-approximation region can now be obtained by choosing the p_i to satisfy $p_1 + \cdots + p_4 = 1 + \alpha^* - 2\alpha$. For a normal sample of size $n = 27$, $\alpha = 0.9$ and with the p_i equal gives $\alpha^* = 0.925$. A more accurate 0.9-approximation region can be obtained by putting $p_i = (1 + 0.925 - 2 * 0.9)/4 = 0.125/4 = 0.03125$ instead of $p_i = 0.025$.

The following data give the quantity of copper in milligrams per litre in a sample of drinking water (Davies (2014)):

$$\left.\begin{array}{l} 2.16, 2.21, 2.15, 2.05, 2.06, 2.04, 1.90, 2.03, 2.06, \\ 2.02, 2.06, 1.92, 2.08, 2.05, 1.88, 1.99, 2.01, 1.86, \\ 1.70, 1.88, 1.99, 1.93, 2.20, 2.02, 1.92, 2.13, 2.13. \end{array}\right\} \tag{3.7}$$

The covering probability for (μ,σ) with $p_1 = \cdots = p_4 = 0.025$ is $\alpha^* = 0.9293$. The correction described above gives $p_1 = \cdots = p_4 = 0.0323$. The 0.9 approximation region $\mathcal{A}(\mathbf{x}_n, 0.9, \mathcal{N})$ for the data (3.7) based on these p_i data set is shown in the upper panel of Figure 1. The lower panel shows a surface plot of the minimum of the four P-values of (3.5). An approximation region for μ alone can be obtained by projecting $\mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{N})$ onto the μ-axis:

$$\mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{N}_\mu) = \{\mu : \text{there exists some } \sigma \text{ s.t. } (\mu,\sigma) \in \mathcal{A}(\mathbf{x}_n, \alpha, \mathcal{N})\} \tag{3.8}$$

The result is the interval $[1.938, 2.094]$. The standard 0.9 confidence interval for μ based on the t-statistic is the smaller interval $[1.978, 2.054]$. This is not always the case. It is possible for an approximation region to be much smaller than a corresponding confidence region, indeed, it may be empty.

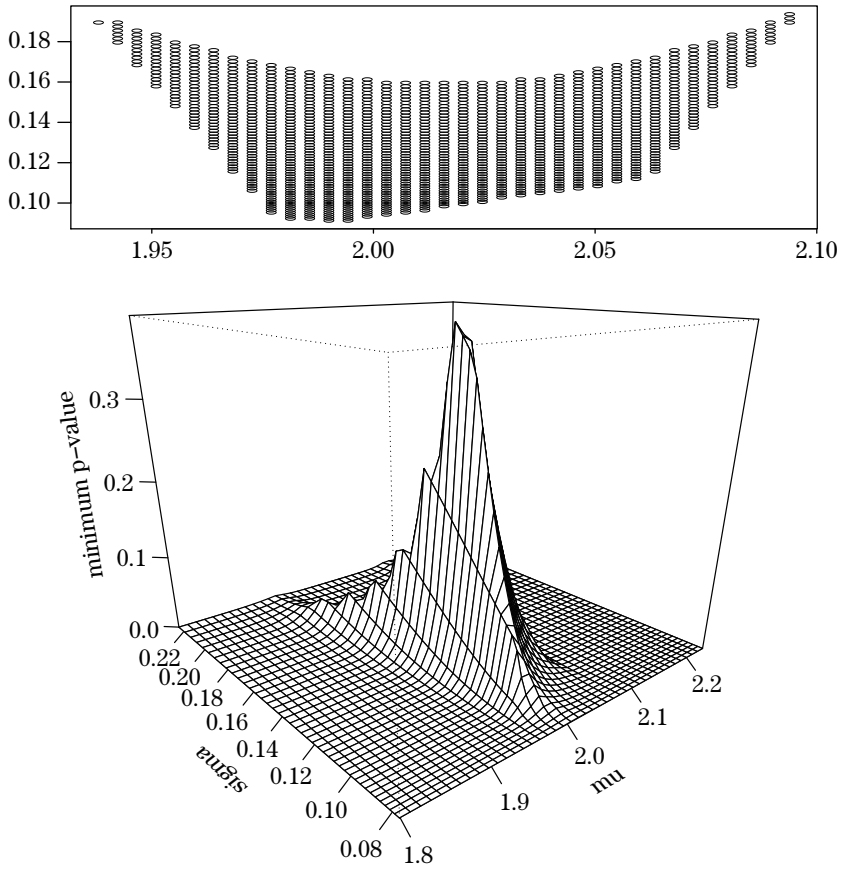

Figure 1. Upper panel: the approximation region $\mathcal{A}(\mathbf{x}_n, 0.9, \mathcal{N})$ for the data (3.7). Lower panel: a surface plot of the minimum P-values of (3.5).

The approximation region (3.6) is based on the P-values of the four statistics of (3.4) for each parameter pair (μ, σ). These values give a measure of how well the data can be approximated by an i.i.d. $N(\mu, \sigma^2)$ model. For example, the four P-values associated with the mean 2.016 and standard deviation 0.116 of the copper data are 1.000, 0.991, 0.157 and 0.935. The smallest value is 0.157 which is the outlier measure S_3 and is due to the observation 1.70.

For the copper data the values of μ and σ which give the best approximation in the sense of the largest minimum P-value are $(\mu, \sigma) = (1.999, 0.130)$ with minimum value 0.447 and P-values 0.489, 0.447, 0.448, 0.774.

Because of (3.3) an approximation region is sometimes interpreted as a confidence region. Such an interpretation however causes difficulties. Consider the family $\mathcal{N}$ of normal models as above. A standard confidence region for the 'true'

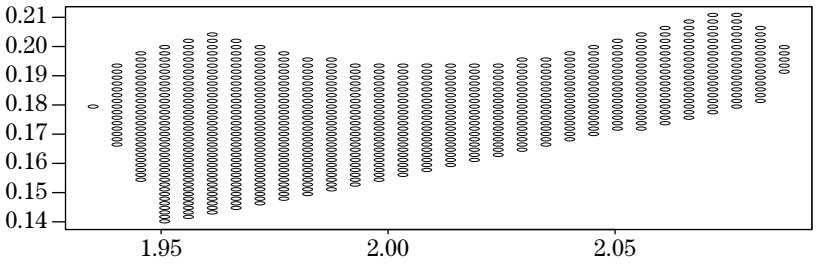

Figure 2. The approximation region $\mathcal{A}(\mathbf{x}_n, 0.9, \mathcal{N})$ for the data (3.7) but with 1.7 replaced by 1.4.

value μ_0 of μ is based on the assumption that there is indeed a 'true' value μ_0 of μ. That is, the data were generated under $N(\mu_0, \sigma^2)$ for some σ. This assumption is not checked in the formal inference phase and consequently a confidence region for μ_0 is never empty. The interpretation is that it is a measure of the precision with which μ_0 can be determined. Thus a small confidence region is 'a good thing'.

The approximation region (3.8) on the other hand is not based on the assumption that the data were indeed generated as i.i.d. $N(\mu, \sigma^2)$ for some (μ, σ). It specifies those μ-values, if any, for which $N(\mu, \sigma^2)$ is an adequate approximation to the data for some σ. Thus if the adequacy region (3.8) is small this simply means that there are few values of μ for which $N(\mu, \sigma^2)$ is an adequate approximation to the data for some σ. It is not a measure of precision. If one imagines the data gradually becoming less and less normal then the region (3.6) will become smaller and eventually will be the empty set. One way of doing this is to gradually increase one value of the sample until this value becomes incompatible with a Gaussian distribution. As an example Figure 2 gives the 0.9 approximation region for the copper data of (3.7) but with the smallest observation of 1.7 being replaced by 1.4. If the 1.7 is replaced by 1.2989 the approximation region as calculated has exactly one point $(1.9692, 0.2086)$ with P-values 0.428, 0.198, 0.0348 and 0.0324.

Interpreting (3.8) as a confidence region leads to complications. As the data become less and less like Gaussian data the region becomes smaller and smaller which is interpreted as an increase in precision. Thus on this interpretation replacing 1.7 by 1.2989 in (3.7) leads to exact values for (μ, σ) namely $(1.9692, 0.2086)$. When the region becomes empty this is as if one goes from infinite precision to no information at all. A discussion can be found in `http://andrewgelman.com/2011/08/25/` From the point of view of approximation

there is no problem of interpretation. The set of adequate parameter values becomes smaller and smaller and eventually becomes the empty set, that is, there are no adequate parameter values at all.

The approximation region (3.6) will pick up an outlier if it is sufficiently large in the sense that the approximation region is empty because of the outlier. Thus for the copper data if the 1.7 is replaced by 1.29 the 0.9-approximation interval is empty. The largest minimum P-value is 0.030 attained for $(\mu, \sigma) = (1.972, 0.211)$ with P-values 0.484, 0.167, 0.032 and 0.030. It is noteworthy that the smallest P-value, 0.030, is that of S_4 which checks the Kuiper distance and not that of S_3 which was explicitly included to check outliers. This can be traced back to the use of the statistics S_1 and S_2 which are based on the mean and variance which are notoriously poor for identifying outliers (see Davies (2014, p. 94)).

The obvious remedy is to replace the mean and standard deviation by M-functionals of location and scale (see Huber and Ronchetti (2009, Chaps 4-5)). Apart from a larger domain, functionals have the further advantage that they can be constructed to have certain properties such as boundedness, continuity and even differentiability (see Davies (2014, pp. 107-108)). This will be illustrated using M-functionals defined as follows. Given ψ- and χ-functions ψ and χ, respectively, and a probability measure P over $\mathbb{R}$, the M-functional T_M is defined by $T_M(P) = (T_L(P), T_S(P))$ where $T_L(P)$ and $T_S(P)$ solve

$$\begin{cases} \int \psi\left(\dfrac{x - T_L(P)}{T_S(P)}\right) dP(x) = 0, \\ \int \chi\left(\dfrac{x - T_L(P)}{T_S(P)}\right) dP(x) = 0. \end{cases} \tag{3.9}$$

The functions ψ and χ can be so chosen so that (i) (3.9) has a unique solution for all P with a largest atom of less than 0.5 and (ii) the functional $T_M(P)$ is locally uniformly Fréchet differentiable in a Kolmogorov neighbourhood of P see Davies (1998) and Hampel et al. (1986, p. 54). This gives stability of analysis with respect to P.

The ψ and χ functions used here are taken from Davies (2014, Chap. 5) and are

$$\psi(u) = \psi(u, c) = \frac{\exp(uc) - 1}{\exp(uc) + 1}, \quad \chi(u) = \frac{u^4 - 1}{u^4 + 1}, \tag{3.10}$$

where c is a tuning constant set here to 5.

Let $\mathbf{X}_n(P)$ denote a sample of size n of i.i.d. random variable with distribution P. Then

$$\psi\left(\frac{X_i(P)-T_L(P)}{T_S(P)}\right)$$

has mean zero and variance

$$\mathbf{E}\left\{\psi^2\left(\frac{X_i(P)-T_L(P)}{T_S(P)}\right)\right\}=\int\psi^2\left(\frac{u-T_L(P)}{T_S(P)}\right)\,dP(u)$$

which implies that

$$\frac{\sum_{i=1}^n\psi(X_i(P)-T_L(P))/T_S(P)}{\sqrt{\sum_{i=1}^n\psi^2(X_i(P)-T_L(P))/T_S(P)}}$$

is asymptotically $N(0,1)$. The same result holds for χ.

Based on this the P-values $p_1(\mu,\sigma)$ and $p_2(\mu,\sigma)$ are replaced by

$$p_1(\mu,\sigma)=2\{1-\text{prgau}(|S_\psi(\mu,\sigma)|)\},\quad p_2(\mu,\sigma)=2\{1-\text{prgau}(|S_\chi(\mu,\sigma)|)\}\tag{3.11}$$

where

$$S_\psi(\mu,\sigma)=\frac{\sum_{i=1}^n\psi(x_i-\mu)/\sigma}{\sqrt{\sum_{i=1}^n\psi^2(x_i-\mu)/\sigma}}\text{ and }S_\chi(\mu,\sigma)=\frac{\sum_{i=1}^n\chi(x_i-\mu)/\sigma}{\sqrt{\sum_{i=1}^n\chi^2(x_i-\mu)/\sigma}}\,.\tag{3.12}$$

The P-values (3.11) are asymptotic ones but as both ψ and χ are bounded they are quite accurate even for small values of n.

The outlier statistic S_3 and the Kuiper distance statistic S_4 can be incorporated into the definition of the approximation but this requires taking into account that for the scale functional $T_S(N(0,1))=0.647161$. To make it asymptotically Fisher consistent (see Huber and Ronchetti (2009, p. 9)) it must be multiplied by $1/0.647161=1.54521$. In fact we use a finite sample size correction f_n. The outlier and Kuiper statistics become

$$S_3(\mathbf{y}_n)=\max_i|y_i|\text{ and }S_4(\mathbf{y}_n)=\text{pkuip}(\mathbb{P}_n(\mathbf{y}_n,N(0,1)))\tag{3.13}$$

respectively where $y_i=(x_i-\mu)/(f_n\sigma)$. The resulting P-values are

$$\begin{cases}p_3(\mu,\sigma)=\mathbf{P}(S_3(\mathbf{Y}_n)\ge S_3(\mathbf{y}_n))=1-(2\text{pnorm}(S_3(\mathbf{y}_n))-1)^n,\\ p_4(\mu,\sigma)=\mathbf{P}(S_4(\mathbf{Y}_n)\ge S_4(\mathbf{y}_n))=1-\text{pkuip}(S_4(\mathbf{y}_n)).\end{cases}\tag{3.14}$$

The resulting approximation region is

$$\begin{aligned}\tilde{\mathcal{A}}(\mathbf{x}_n,\alpha,\mathcal{N})=\{(\mu,\sigma):\,&p_1(\mu,\sigma)\ge p_1,p_2(\mu,\sigma)\ge p_2,\\ &p_3(\mu,\sigma)\ge p_3,p_4(\mu,\sigma)\ge p_4\}\end{aligned}\tag{3.15}$$

with $p_1(\mu,\sigma$ and $p_2(\mu,\sigma)$ given by (3.11) and $p_3(\mu,\sigma)$ and $p_4(\mu,\sigma)$ given by (3.14).

The $p_1,\dots,p_4$ must be chosen to give an α-approximation region. This can be done in the same manner as described above and results in $p_i=0.0323$. The approximation region is shown in Figure 3. It can be compared with Figure 1. The 0.9-approximation interval for μ is $[1.963,2.069]$ compared to $[1.938,2.094]$

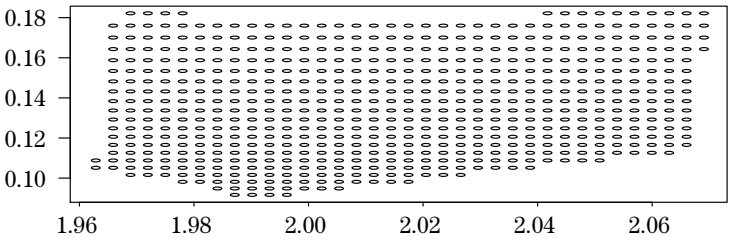

Figure 3. The 0.9-approximation region (3.15) for the copper data with $p_i = 0.03225$ ($\alpha = 0.9$) for the copper data using the psi- and chi- functions of (3.10) with $c = 5$.

for the 0.9-approximation region (3.6) and $[1.978, 2.054]$ for the standard 0.9-confidence interval based on the t-statistic. The largest minimum P-value is 0.442 attained for $(\mu, \sigma) = (2.014, 0.138)$ with P-values 0.467, 0.442, 0.468 and 0.711.

If the 1.7 of the copper data is replaced by 1.376 the 0.9-approximation region (3.15) is empty. This is larger than the corresponding value 1.2947 for the region (3.6). The largest minimum P-value is $0.0320 < 0.03225$ attained for $(\mu, \sigma) = (1.966, 0.182)$ with P-values 0.0323, 0.058, 0.0320 and 0.106. It is seen that the P-value of the outlier statistic is the smallest one.

The approximation region $\tilde{\mathcal{A}}(\mathbf{x}_n, \alpha, \mathcal{N})$ of (3.15) is for the family of Gaussian models and may well be empty. Indeed for many data sets on water quality the approximation region would be empty for this reason. See also the data sets of Stigler (1977). Nevertheless a point estimate and an interval of plausible values for the quantity of interest are still required. This can be done by simply dropping the two statistics S_3 and S_4 of (3.13) in the definition of the approximation region. The Gaussian family $\mathcal{N}$ is replaced by the family $\mathcal{P}$ of all probability measure on $\mathbb{R}$ and the quantity of interest, the copper content of the water sample, is now identified with the location functional $T_L(P)$ rather than with μ of the $N(\mu, \sigma^2)$ model. More precisely the approximation region becomes

$$\begin{aligned}\mathcal{A}_M(\mathbf{x}_n, \alpha, \mathcal{P}) = \{(T_L(P), T_S(P)) : p_1(T_L(P), T_S(P)) \geq p_1(P), \\ p_2(T_L(P), T_S(P)) \geq p_2(P), \mathrm{pkuip}(\mathbb{P}(\mathbf{x}_n), P) \leq 1 - p_3\},\end{aligned} \tag{3.16}$$

where

$$\begin{aligned}p_1(T_L(P), T_S(P)) &= \mathbf{P}(|S_{M,\psi}(\mathbf{X}_n(P), T_L(P), T_S(P))| \\ &\geq |S_{M,\psi}(\mathbf{x}_n, T_L(P), T_S(P))|),\end{aligned} \tag{3.17}$$

$$\begin{aligned}p_2(T_L(P), T_S(P)) &= \mathbf{P}(|S_{M,\chi}(\mathbf{X}_n(P), T_L(P), T_S(P))| \\ &\geq |S_{M,\chi}(\mathbf{x}_n, T_L(P), T_S(P))|)\end{aligned} \tag{3.18}$$

with

$$\begin{cases} S_{M,\psi}(\mathbf{x}_n, T_L(P), T_S(P)) = \dfrac{\sum_{i=1}^n \psi(x_i - T_L(P))/T_S(P)}{\sqrt{\sum_{i=1}^n \psi^2(x_i - T_L(P))/T_S(P)}}, \\ S_{M,\chi}(\mathbf{x}_n, T_L(P), T_S(P)) = \dfrac{\sum_{i=1}^n \chi(x_i - T_L(P))/T_S(P)}{\sqrt{\sum_{i=1}^n \chi^2(x_i - T_L(P))/T_S(P)}}. \end{cases} \tag{3.19}$$

The statistics $S_{M,\psi}(\mathbf{X}_n(P), T_L(P), T_S(P))$ and $S_{M,\chi}(\mathbf{X}_n(P), T_L(P), T_S(P))$ are asymptotically $N(0,1)$ distributed. The requirement $\text{pkuip}\,(\mathbb{P}(\mathbf{x}_n), P) \leq 1 - p_3$ forces P into a $O(1/\sqrt{n})$ Kolmogorov neighbourhood of $\mathbb{P}_n$. This together with the locally uniform Fréchet differentiability of $T_M = (T_L, T_S)$ (see pages 107-108 of Davies (2014)) implies that the convergence to the $N(0,1)$ distribution is uniform over the Kolmogorov neighbourhood. Thus asymptotically and uniformly in P,

$$\begin{cases} p_1(T_L(P), T_S(P)) \to 2\{1 - \text{pnorm}(S_{M,\psi}(\mathbf{x}_n, T_L(P), T_S(P)))\} \\ p_2(T_L(P), T_S(P)) \to 2\{1 - \text{pnorm}(S_{M,\chi}(\mathbf{x}_n, T_L(P), T_S(P)))\}. \end{cases} \tag{3.20}$$

With this asymptotic approximation the approximation region $\mathcal{A}_M(\mathbf{x}_n, \alpha, \mathcal{P})$ becomes

$$\begin{aligned} &\tilde{\mathcal{A}}_M(\mathbf{x}_n, \alpha, \mathcal{P}) = \{(T_L(P), T_S(P)) : \\ &2\{1 - \text{pnorm}(S_{M,\psi}(\mathbf{x}_n, T_L(P), T_S(P)))\} \geq p_1, \\ &2\{1 - \text{pnorm}(S_{M,\chi}(\mathbf{x}_n, T_L(P), T_S(P)))\} \geq p_2, \\ &\text{pkuip}(\mathbb{P}(\mathbf{x}_n), P) \leq 1 - p_3\}. \end{aligned} \tag{3.21}$$

It is shown in the top panel of Figure 4 for the copper data with $p_i = 0.0333$ corresponding to $\alpha = 0.9$ in a first approximation. It may be compared with the approximation region based on the Gaussian distribution as shown in Figure 1. The lower panel of Figure 4 shows a surface plot of the minimum of the two asymptotic P-values of (3.20).

The approximation region for $T_L(P)$ is obtained by projecting the approximation region onto the x-axis. For the copper data with $\alpha = 0.9$ it is $[1.966, 2.069]$ compared with the standard 0.9-confidence interval $[1.978, 2.054]$ based on the t-statistic.

The approximation region (3.21) remains unchanged if the smallest observation 1.7 is replaced by zero. This is in sharp contrast to the approximation region based on the Gaussian family of models which is empty in this case. The 0.9-confidence interval based on the t-statistic is $[1.821, 2.085]$. This stability is a strong argument for the use of the M-functional rather than basing the analysis

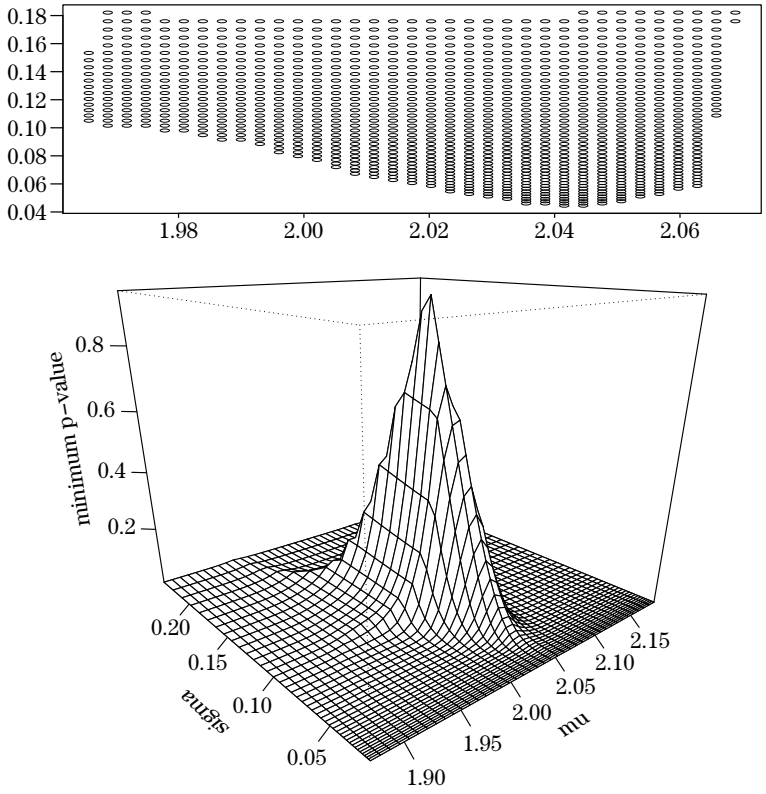

Figure 4. Upper panel: the 0.9 approximation region of the location and scale functionals (T_L, T_S) for the copper data using the psi- and chi- functions of (3.10) with $c = 5$. Lower panel: a surface plot of the $\min(p_1(T_l, T_S), p_2(T_L, T_S))$ of (3.20).

on the Gaussian model.

The use of (3.11) and (3.12) to calculate an approximation region seems to be new. Confidence regions receive scant attention in the robustness literature and are typically calculated, if at all, by appealing to the asymptotic normality of the functionals (T_L, T_S) themselves. Such asymptotic confidence regions are perforce elliptical whereas those based on (3.11) and (3.12) are not as can be seen in Figure 4.

Following the recommendation above, the following will be based on the M-functional of (3.9) with ψ- and χ-functions given by (3.10).

As an example consider again the copper data (3.7) Suppose the legal limit is 2.1 milligrams of copper per litre of water and we wish to test the hypothesis that this is exceeded. The standard analysis based on the Gaussian family is to identify the quantity of copper in the drinking water with μ leading to the null

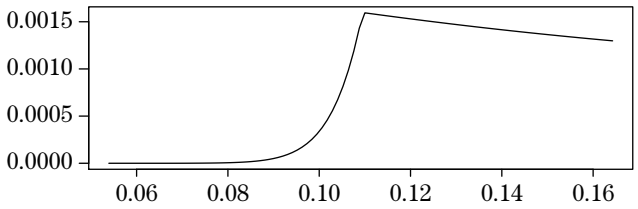

Figure 5. The $\min(p_1(2.1, T_S), p_2(2.1, T_S))$ of (3.20) for $H_0 : T_L \geq 2.1$ plotted against a 0.9-approximation interval for T_S derived from (3.21). The maximum minimum value is 0.0016.

hypothesis

$$H_0 : \mu \geq \mu_0 = 2.1. \tag{3.22}$$

The P-value based on the t-statistic is 0.00044.

If the analysis is based on the M-functional the quantity of copper in the drinking water is identified with T_L. The null hypothesis becomes

$$H_0 : T_L \geq 2.1. \tag{3.23}$$

Neither the hypothesis (3.22) nor the hypothesis (3.23) mentions scale. In the case of the t-statistic the scale σ used is the standard deviation of the data. Even if this is consistent with the underlying Gaussian assumption, which it may not be, it is not the only such value of σ. It would be possible to do the same for the hypothesis (3.23) and base the calculation of the P-value on the corresponding value of T_S. However in accordance with the philosophy of this paper we calculate a P-value for each value of T_S consistent with the data.

Given α the set of T_S values consistent with the data is the projection of the α-approximation region $\tilde{\mathcal{A}}_n$ of (3.21) onto the y-axis. For the copper data and $\alpha = 0.9$ it is $[0.054, 0.164]$. The $p_i(T_L, T_S), i = 1, 2$,-values of (3.20) are now calculated with $\mu = \mu_0$ and for each $\sigma \in [0.054, 0.164]$ with appropriate changes made to $p_1(T_L, T_S)$ depending on the form of H_0 (one-sided or two-sided) .

Figure 5 shows the $\min(p_1(T_L, T_S), p_2(T_L, T_S))$ for $\alpha = 0.9$ and $H_0 : T_L \geq 2.1$ implying $T_L = 2.1$. The largest minimum P-value is 0.0016.

4. Jeffreys, P-values and Prediction

The following taken from page 385 of Jeffreys (1961) is often cited as an argument against the use of P-values:

> .... gives the probability of departures, measured in a particular way, equal to or greater than the observed set, and the contribution

> from the actual value is nearly always negligible. What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems to be a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it.

What Jeffreys means by 'a hypothesis predicting' is not at all clear. He writes that a hypothesis has not predicted observable results that have not occurred but it also has not predicted observable results which have occurred. It seems necessary to clarify what is meant by a hypothesis predicting.

A non-vacuous prediction consists of specifying a probability α and a set $\mathcal{S}(\alpha)$ such that the prediction is correct if $X \in \mathcal{S}(\alpha)$ and the probability that it is correct is α. It is worthy of note that the larger α the more vacuous the prediction so to speak. The expression 'not predicted to occur' will be understood as 'predicted not to occur' rather than as 'forgetting to predict'. As a simple example put $\alpha = 0.95$ and $\mathcal{S}(\alpha) = (-1.96, 1.96)$. Then the prediction is $-1.96 \leq X \leq 1.96$. Suppose however that $X = 3.121$ is observed. This value was predicted not to occur. The P-value is $\mathbf{P}(|X| > 3.121) = 0.0018$ and for this to be a successful prediction would require $\alpha = 0.9982$ rather than the chosen $\alpha = 0.95$.

If it were agreed beforehand that a false prediction would lead to the null hypothesis being rejected, then this is done because a value predicted not to occur, namely 3.121, did in fact occur. This seems an unremarkable procedure.

The concept of adequate approximation can be looked at in terms of prediction. Given a number α and based on a model P a prediction has to be made about a sample $\mathbf{x}_n$. In making the prediction it has to be decided which aspects of the data are regarded as important. In the definition of the approximation region (3.6) the important aspects are given by the statistics $S_i, i = 1, \ldots, 4$ of (3.4) with $P = N(\mu, \sigma^2)$. The corresponding prediction is that all the inequalities of (3.5) will hold with $\mathbf{y}_n = (\mathbf{x}_n - \mu)/\sigma$ replacing $\mathbf{Y}_n$. If the prediction is correct then the model $N(\mu, \sigma^2)$ is accepted as an adequate approximation to the data.

5. P-values and Choice of Covariates in Stepwise Regression

The following is based on Davies (2017). Given a data set of size n consisting of a dependent variable $\mathbf{y}(n)$ and $p(n)$ covariates $\mathbf{x}(n)$, the problem is to decide which if any of the covariates to include. The discussion below will be restricted to the case where $p(n)$ is chosen by stepwise regression but the idea can be

extended to considering all subsets of the covariates as long as $p(n)$ is not too large, say $p(n) \le 20$ (see Davies (2016)).

It would seem that all procedures for choosing the covariates are based on the standard linear model

$$\mathbf{Y}(n) = \mathbf{X}(n)\beta(n) + \varepsilon(n). \tag{5.1}$$

The procedure to be described below is not based on this model. The basic idea is to compare the covariates $\mathbf{x}(n)$ with covariates which are simply standard Gaussian white noise. A covariate $\mathbf{x}_j$ is included only if it is significantly better than white noise.

Suppose that $p_0 \le n-2$ with indices $\mathcal{S}_0$ have already been been included in the regression and that the sum of squared residuals is ss_0. Denote by ss_j the sum of squared residuals if the covariate $\mathbf{x}_j$ with $j \notin \mathcal{S}_0$ is included. The next candidate for inclusion is that covariate for which ss_j is smallest. Including this covariate leads to a sum of squared residuals

$$ss_{01} = \min_{j \notin \mathcal{S}_0} ss_j.$$

Replace all the covariates not in $\mathcal{S}_0$ in their entirety by standard Gaussian white noise. Let SS_j denote the sum of squared residuals if the random covariate corresponding to $\mathbf{x}_j$ is included. The inclusion of the best of the random covariates leads to a sum of squared residuals

$$SS_{01} = \min_{j \notin \mathcal{S}_0} SS_j.$$

The probability that the best random covariate is better than the best of the actual covariates is

$$\begin{aligned}\mathbf{P}(SS_{01} < ss_{01}) &= 1 - \mathbf{P}(SS_{01} \ge ss_{01}) = 1 - \mathbf{P}\left(\min_{j \notin \mathcal{S}_0} SS_j \ge ss_{01}\right) \\ &= 1 - \prod_{j \notin \mathcal{S}_0} \mathbf{P}(SS_j \ge ss_{01}).\end{aligned}$$

It has been shown by Lutz Dümbgen that

$$SS_j \stackrel{D}{=} ss_0(1 - B_{1/2,(n-p_0-1)/2}), \tag{5.2}$$

where $B_{a,b}$ denotes a beta random variable with parameters a and b and distribution function $\text{pbeta}(\cdot, a, b)$. From this it follows that

$$\mathbf{P}(SS_j \ge ss_{01}) = \text{pbeta}\left(1 - \frac{ss_{01}}{ss_0}, \frac{1}{2}, \frac{(n-p_0-1)}{2}\right)$$

so that finally,

$$\mathbf{P}(SS_{01} \le ss_{01}) = 1 - \text{pbeta}\left(1 - \frac{ss_{01}}{ss_0}, \frac{1}{2}, \frac{(n - p_0 - 1)}{2}\right)^{p(n)-p_0}. \tag{5.3}$$

This is the P-value for the inclusion of the next covariate. The simplest procedure is to specify $\alpha < 1$ and to continue the stepwise selection until the first P-value exceeds α. Those covariates up to but excluding this last one are the selected ones. The stopping rule is

$$ss_{01} > ss_0\left(1 - \text{qbeta}\left((1-\alpha)^{1/(p(n)-p_0)}, \frac{1}{2}, \frac{(n - p_0 - 1)}{2}\right)\right) \tag{5.4}$$

where $\text{qbeta}(\cdot, a, b)$ is the quantile function of the beta distribution with parameters a and b.

The procedure is conceptually and algorithmically simple. It requires no regularization parameter or cross-validation or an estimate of the error term in (5.1). It is invariant with respect to affine changes of unit of the covariates and equivariant with respect to a permutation of the covariates. It can be extended to non-linear parametric regression, robust regression and the Kullback-Leibler discrepancy where appropriate.

As an example we take the leukemia data (Golub et al. (1999) `http://www-genome.wi.mit.edu/cancer/` which was analysed in Dettling and Bühlmann (2003). These consist of data on $n = 72$ samples of tissue with $p(n) = 3{,}571$ covariates. The dependent variable $\mathbf{y}(n)$ is either 0 or 1 depending on whether the patient suffers from acute lymphoblastic leukemia or acute myeloid leukemia. The first five genes in order of inclusion with their associated P-values as defined by (5.3) are as follows:

gene number	1,182	1,219	2,888	1,946	2,102
P-value	0.0000	8.57e-4	3.56e-3	2.54e-1	1.48e-1

(5.5)

According to this relevant genes are 1,182, 1,219 and 2,888 and given these the remaining 3,568 are no better than random noise. This applies to the gene 1,946 but if a simple linear regression is performed using this gene alone its P-value in the linear regression is 7.75e-9. This is much smaller than the 0.254 in (5.5). The P-value (5.3) takes into account the stepwise nature of the procedure, in particular that gene 1,946 is the best of the remaining genes once the genes 1,182, 1,219 and 2,888 have been included. A simple linear regression does not take this into account.

The data were gathered in the hope of using the gene expression data to classify the patients. If the classification is based on genes 1,182, 1,219 and

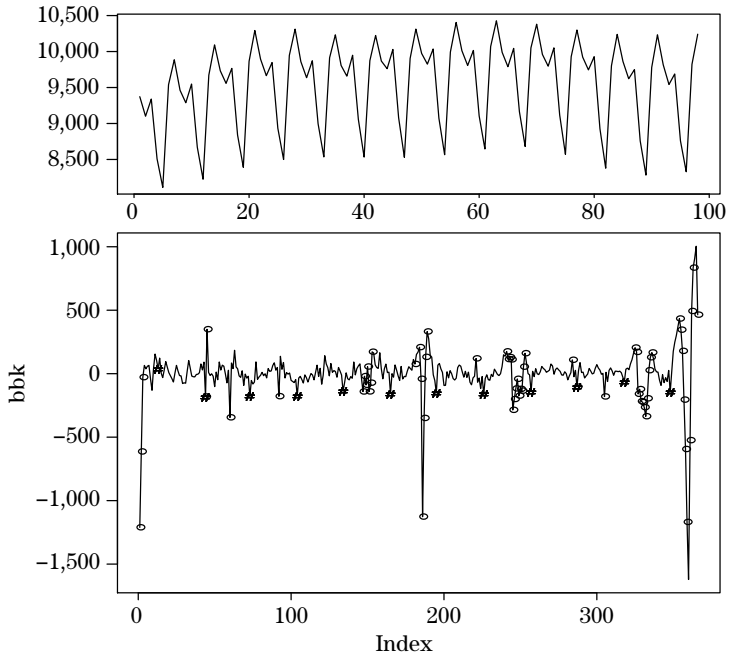

Figure 6. Upper panel: the first 98 values of the regression function of the birthdays data. Lower panel: the residuals for each day of the year averaged over the twenty years.

2,888, a simple linear regression results in one misclassification. In Dettling and Bühlmann (2003) the authors considered 42 different classification schemes. Only two of them resulted in a single misclassification. They used a 1-nearest-neighbour method based on 25 and 3,571 genes. For this particular data set the procedure described above attains the same result and moreover specifies the relevant genes.

A second example is the daily number of births in the United States from 1st January 1969 to the 31st December 1988. They are available as 'Birthdays' from the R-package 'mosaicData'. The data were analysed by removing a trend and then performing a linear regression using the covariates

$$xs_j(i) = \sin\left(\frac{\pi ji}{n}\right) \text{ and } xc_j(i) = \cos\left(\frac{\pi ji}{n}\right),\ i = 1, \ldots, n, j = 0, \ldots, n$$

with $n = 7{,}305$ giving 14,611 covariates in all. The constant term xc_0 was included by default. The remaining 14,610 were treated pairwise, that is, the sin and cos terms together on the basis that only the frequency was important. Setting the cut-off P-value to 0.01 and using the robustified version based on the Huber ψ-function resulted in 106 of the 7,305 pairs being included.

The top panel of Figure 6 shows the regression function for the first 98 days. The five strongest periods are in order 7 days, 3.5 days, 365 days (one year), 182.6 days (six months), and 2.33 days (third of a week) . The lower panel shows the residuals for each day of the year averaged over the twenty years. From left to right the "o" symbols show New Year, 14th February (Valentine), the 29th February (leap year), the 1st April, Memorial Day, 4th July (Independence Day), 8th August (no explanation), Labor Week, 10th October (no explanation), Halloween, Thanksgiving and Christmas. The hash marks "#" show the 13th of each month. A Bayesian analysis of the same data is to be found in Gelman et al. (2013).

References

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association* **57**, 269–326.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**, 4361–4383.

Davies, L. (2014). *Data Analysis and Approximate Models In Monographs on Statistics and Applied Probability* **133**. CRC Press.

Davies, L. (2017). Stepwise choice of covariates in high dimensional regression. arXiv:1610.05131 [math.ST].

Davies, P. L. (1995). Data features. *Statistica Neerlandica* **49**, 185–245.

Davies, P. L. (1998). On locally uniformly linearizable high breakdown location and scale functionals. *The Annals of Statistics* **26**, 1103–1125.

Davies, P. L. (2008). Approximating data (with discussion). *Journal of the Korean Statistical Society* **37**, 191–240.

Davies, P. L. (2016). Functional choice and non-significance regions in regression. arXiv: 1605.01936 [math.ST].

Davies, L. and W. Krämer (2016). Stylized facts and simulating long range financial data. arXiv: 1612.05229v1 [q-fin.ST].

Dettling, M. and P. Bühlmann (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–1069.

Dudley, R. M. (1989). *Real Analysis and Probability*. Wadsworth and Brooks/Cole, California.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis* (3rd Edition). CRC Texts in Statistical Science. CRC Press.

Golub, T., Slonim, D., Tamayo. P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and

Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician* **70**. Supplemental material to 'The ASA's statement on p-values: context, process and purpose'.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Huber, P. J. (2011). *Data Analysis*. Wiley, New Jersey.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics* (2nd Edition). Wiley, New Jersey.

Jeffreys, H. (1961). *Theory of Probability* (3rd Edition). Oxford Classic Texts in the Physical Sciences. Oxford University press.

Kuiper, N. H. (1962). On a metric in the space of random variables. *Statistica Neerlandica* **16**, 231–235.

Neyman, J., Scott, E. L. and Shane, C. D. (1953). On the spatial distribution of galaxies a specific model. *Astrophysical Journal* **117**, 92–133.

Neyman, J., Scott, E. L. and Shane C. D. (1954). The index of clumpiness of the distribution of images of galaxies. *The Astrophysical Journal Supplement Series* **8**, 269–294.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stigler, S. M. (1977). Do robust estimators work with real data? (with discussion). *The Annals of Statistics* **5**, 1055–1098.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on P-values: Context, process, and purpose. *The American Statistician* **70**.

Faculty of Mathematics, University of Duisburg-Essen, 45711 Essen, Germany

E-mail: laurie.davies@uni-due.de

(Received November 2016; accepted October 2017)