ISSN 1973-2201 *STATISTICA, anno LXXXIV, n. 2, 2024*
DOI: https://doi.org/10.60923/issn.1973-2201/22076

STATISTICS: TRUTH, ONTOLOGY, APPROXIMATION, HONESTY AND INDOCTRINATION

Laurie Davies[1]
University of Duisburg-Essen, Essen, Germany

SUMMARY

This additional material was presented by the author during the discussion meeting on the paper by Dümbgen and Davies (2024), held on October 24, 2024, at the Department of Statistical Sciences, University of Bologna.

1. TRUTH

Truth is an essential part of statistical inference, almost all concepts of statistical inference are truth dependent. Examples are hypothesis testing, p-values and confidence intervals. The most cited paper of the JRSS B is Benjamini and Hochberg (1995), where the false discovery rate is based on hypotheses being either 'true' or 'not-true' The Bayesians are not immune, the title of Fraser *et al.* (2016) is "Reputability and the quest for truth".

Although statistical models are much too simple to be true, it is common practice to behave as if they are true, that is, as if the data were generated as described by the model. This implies true parameter values and these should be estimated as precisely as possible leading to the use of optimal estimation procedures. This is standard practice and called by Tukey the "assumed (revealed?) truth" approach (Tukey (1993d)).

2. TRUTH AND ONTOLOGY

Truth depends on existence or being and ontology is the philosophical study of existence or being; see Merricks (2007) and the resulting discussions in various philosophical journals. The ontological question of the existence of true parameter values is never posed in statistics, not even in the ASA statement on *P*-values:

https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

[1] Corresponding Author. E-mail: laurie.davies@uni-due.de

If it were to be posed the answer would in general be negative. It could be argued that there is a true value of the gravitational constant. Gravitational attraction and Hooke's law are essentially linear in Cavendish's experiment giving an example of a linear regression with true values Falconer (1999).

In contrast the production of riboflavin using bacteria is a complicated biological process and the model of the logarithmic rate of increase of riboflavin as linear function of a sparse subset of the gene expressions is simply false. In spite of this P-values and confidence intervals, both truth based concepts, are calculated as a matter of course for this and other data sets, see Dezeure *et al.* (2015).

3. Approximation

John von Neumann (von Neumann (1947))

> "I think that it [mathematics] is a relatively good approximation to truth - which is much too complicated to allow anything but approximations."

The expression "assumed (revealed?) truth" is to be found in Tukey (1993d). This was Tukey's response to an early version of Davies (1995), the first paper the author wrote from the approximation point of view. Tukey writes:

> "Davies's emphasis on approximation is well chosen and surprisingly novel. While these will undoubtedly be a place for much careful work in learning how to describe the concept - - and its applications - - in detail, it is clear that Davies has taken the decisive step by asserting that there must be a formal admission that adequate approximation, of one set of observable (or simulated) values by another set, needs to be treated as practical identity. If, as is so convenient, we continue to use continuous models to describe - - or perhaps only to illuminate - - observed data, we should have to say that certain aspects of the data - - not typically, but unavoidably, including "Most (modelled) observations have irrational values!"- - are not to be used in relating conceptual (or simulated) samples to observed samples. Thought and debate as to just which aspects are to be denied legitimacy will be both necessary and valuable."

3.1. A formalization

The concept of approximation in Davies (1995, 2014, 2018a) is the following. Given a model P_Θ generate a sample the same size as the data with a specific value θ_0 of Θ and then compare the real sample with the simulated samples. The comparison compares the values of certain chosen features of data sets generated under the model with the values of the same features of real data. As Tukey pointed out above, the irrationality of data under the model will be denied legitimacy.

A formalization is given in Chapters 2.2 and 2.3 of Davies (2014). The following is for the specific case of the i.i.d. $N(\mu, \sigma^2)$ model for real data. The first decision is to decide on the statistics to be used. Suppose the standard deviation $sd(\boldsymbol{x})$ of the data $\boldsymbol{x}$ is chosen. Then under the model $N(\mu, \sigma^2)$ the statistic $(n-1)sd(\boldsymbol{X})^2/\sigma$ has a chi squared distribution with $n-1$ degrees of freedom. Given this a lower bound L_{sd} and an upper U_{sd} can be calculated such that $L_{sd} \leq sd(\boldsymbol{X})/\sigma \leq U_{sd}$ with probability α. The value of $sd(\boldsymbol{x})$ for the data is regarded as consistent with σ if $L_{sd} \leq sd(\boldsymbol{x})/\sigma \leq U_{sd}$. Turning this around gives the values of σ $sd(\boldsymbol{x})/U_{sd} \leq \sigma \leq sd(\boldsymbol{x})/L_{sd}$ which are consistent with the data $\boldsymbol{x}$. This is as in Section 1.4 of Davies (2014) but with the standard deviation instead of the mean.

The standard deviation has a weakness, it is very sensitive to outliers. This can be rectified using the median absolute deviation, the MAD. There is no exact expression for the distribution of the MAD but it is asymptotically normal (see Segers (2014)). Similarly the median is used and not the mean. Outliers can be included by considering the maximum absolute deviation from the median. Finally the shape of the data can be included using the Kuiper metric Kuiper (1962). In all four features of the data have been taken into account. The approximation region consists of all (μ, σ) which are consistent with the data for all four features. Spending α probability on each leads to $(3+\alpha)/4$ being spent on all four. Replacing α by $(3+\alpha)/4$ guarantees that the probability, that (μ, σ) lies in the approximation region for data generated under (μ, σ), is at least α.

3.2. Examples

3.2.1. The normal distribution

The data used in Figure 1 are the study times in semesters of 258 students. The histogram is the upper figures of Figure 1. The models are i.i.d. $N(\mu, \sigma^2)$. The lower figure shows the 0.95 approximation region based on the features in the last paragraph. The # denotes the standard deviation and mean of the data. which in this case do not belong to the approximation region. A further paper on approximation is Davies (2018b) in an issue of *Statistica Sinica* dedicated to the memory of Peter Hall.

For an excellent example of comparing data with simulations see Chapter 5.7 of Huber (2011) on "Modelling the length of the day"; an interplay of inspection, modelling, simulation, comparison, model fitting, parameter estimation and interpretation. See also Chapter 5.8 "The role of simulation" and Figure 1.7 of Davies (2014) for a comparison using box plots.

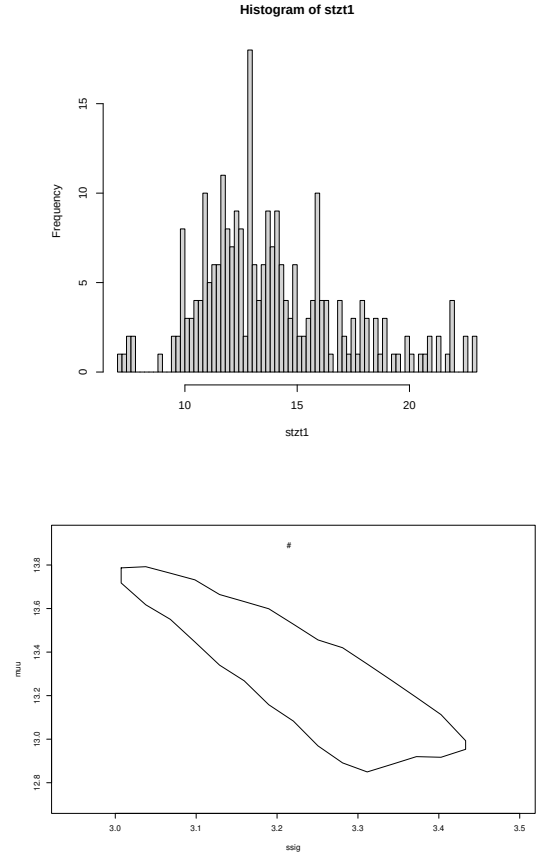

Figure 1 – The histogram and approximation region 258 student study times. The # in the bottom figure is the standard deviation and mean of the data.

3.2.2. Non-parametric regression

The second paper on approximation Davies and Kovac (2001) was concerned with controlling the number of peaks in non-parametric regression. The data were X-ray refraction data for thin films provided by Dieter Mergel, Department of Physics, University Duisburg-Essen. The idea is as follows. Given data $y(t_i) =, i = 1, \ldots n$ with the $t_i \in (0,1)$ ordered the model is

$$y(t_i) = f(t_i) + \sigma\varepsilon(t_i), \tag{1}$$

where ε is standard Gaussian white noise. For a given function f_n form the residuals $r_n(t_i) = y(t_i) - f_n(t_i)$. Supposing for the moment the $n = 2^m$ and calculate the multi

resolution coefficients

$$w_{j,k} = 2^{-j/2} \sum_{i=k2^j+1}^{(k+1)2^j}. \tag{2}$$

The residuals may be adequately approximated by white noise if

$$|w_{j,k}| \le \sigma_n \sqrt{2.5 \log n}, \tag{3}$$

where σ_n is an estimate of σ with default value

$$\sigma_n = \frac{1.48}{\sqrt{2}} \text{Median}\{|y(t_2] - y(t_1)|, \ldots, |y(t_n) - y(t_{n-1})|\}. \tag{4}$$

The goal is to minimize the number of local extremes of f_n subject to (3). This can be done and the function can be smoothed by minimizing the total variations of the first and second derivatives, the third leads to numerical errors; see Kovac (2007); Dümbgen and Kovac (2009) and the relevant R package *ftnonpar* which has been archived. The algorithms are very fast, less than 0.3 seconds for the whole data set of length 7001. The first row of Figure 2 shows the raw data and the function which minimizes the number of peaks for the first 2000 data points. The second row shows the functions which minimize the total variation of the second and third derivatives subject to the constraints of the peaks. The total variation of the third derivative was minimized using linear programming; it is very slow, requiring about 14 minutes.

3.3. Confidence regions and bands

A confidence region contains the true parameter values with a specified probability. For real data there are no true parameter values and the concept is meaningless. In contrast an approximation region specifies those parameter values which give an adequate approximation in a precisely defined sense to the data. This set may be empty.

Similarly in non-parametric regression a confidence bound makes no sense, there is no true function, there are only functions which approximate the data as in Figure 2.

4. GAUSSIAN COVARIATES

The concept of approximation in Section 3 cannot be used for linear regression because of (i) the large number of parameters and (ii) the lack of a definition of what an acceptable approximation is.

In Chapter 11.6 of Davies (2014) logistic regression is considered using the low birth weight data of Hjort and Claeskens (2003); Claeskens and Hjort (2003). In Davies (2014) an acceptable approximation is one with no irrelevant covariates. Relevance is operationalized in terms of P-values and they in turn are defined by replacing the covariates by random covariates as in the Gaussian covariate approach. In contrast to the Gaussian

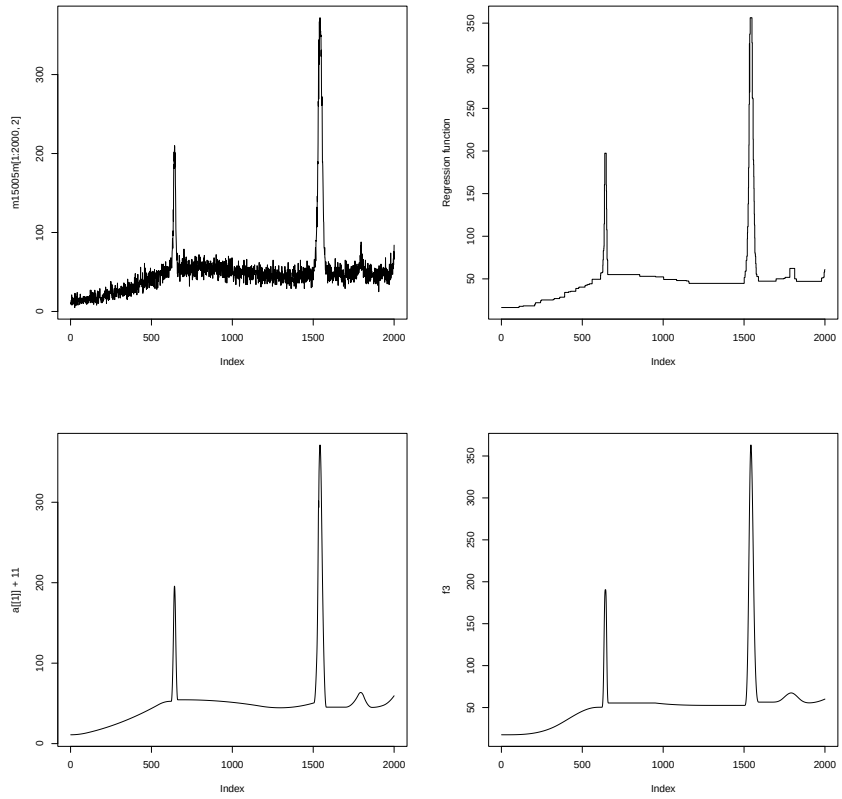

Figure 2 – X-ray refraction data for thin films provided by Dieter Merge, Physics, Universityät Duisburg-Essen.

covariate approach however, the random covariates were chosen to model the covariates they replace, for example, a 0-1 dummy covariate was modelled a binomial random variable. There were only 11 covariates and the method as such could not be extended to high dimensional data.

More thought on high dimensional regression lead to the concept of a Gaussian P-value. Given a subset $\boldsymbol{S}$ a covariate $\boldsymbol{x}_i \in \boldsymbol{S}$ is replaced by a Gaussian covariates $\boldsymbol{Z}$. On denoting by $rss_{\boldsymbol{S}}$ the sum of squared residuals when the dependent variable $\boldsymbol{y}$ is regressed on all covariates in $\boldsymbol{S}$, by $rss_{\boldsymbol{S}\setminus\boldsymbol{x}_i}$ when $\boldsymbol{y}$ is regressed on all covariates in $\boldsymbol{S}$ excluding $\boldsymbol{x}_i$ and by $RSS_{\boldsymbol{Z}\cup\boldsymbol{S}\setminus\boldsymbol{x}_i}$ when $\boldsymbol{y}$ is regressed on $\boldsymbol{Z}$ and all covariates in $\boldsymbol{S}$ but excluding $\boldsymbol{x}_i$. The Gaussian P-value of $\boldsymbol{x}_i$ is the probability that $\boldsymbol{Z}$ is better than $\boldsymbol{x}_i$, that is $P_{G,\boldsymbol{S}}(\boldsymbol{x}_i) = P(RSS_{\boldsymbol{Z}\cup\boldsymbol{S}\setminus\boldsymbol{x}_i} < rss)$. The first attempt at a theory was based on somewhat intuitive mathematics. The paper was sent to Lutz Dümbgen, who replaced the intuitive mathematics by theorems and proved

$$RSS_{\boldsymbol{Z}_i\cup\boldsymbol{S}\setminus\boldsymbol{x}_i}/rss_{\boldsymbol{S}\setminus\boldsymbol{x}_i} \sim \text{Beta}((n-k)/2, 1/2), \tag{5}$$

where k is the size of $\boldsymbol{S}$ and $\text{Beta}((n-k)/2, 1/2)$ denotes the Beta distribution with $(n-k)/2, 1/2)$ degrees of freedom. It is Lemma 1 of Dümbgen and Davies (2023) with $p0 = p-1$. The proof is about two pages long. All the work in Dümbgen and Davies (2023) is due to Lutz Dümbgen. A proof based on Cochran's theorem has been given by Joe Whittaker (Whittaker (2015)). This is the most important result in the theory of Gaussian P-values as it implies that the distribution is independent of the data, the subset $\boldsymbol{S}$ and the covariate $\boldsymbol{x}_i$. In other words, it is universally valid. The Gaussian P-value is given by

$$P_{G,\boldsymbol{S}}(\boldsymbol{x}_i) = \text{Beta}_{(n-k)/2,1/2}(rss_{\boldsymbol{S}}/rss_{\boldsymbol{S}\setminus\boldsymbol{x}_i}), \tag{6}$$

which inherits the universal validity of (5). From (6) it follows that

$$P_{G,\boldsymbol{S}}(\boldsymbol{x}_i) = P_{F,\boldsymbol{S}}(\boldsymbol{x}_i), \tag{7}$$

where $P_{F,\boldsymbol{S}}(\boldsymbol{x}_i)$ is the standard F distribution P-value. Both P-values are deduced from a standard Gaussian variable. In the case of Gaussian P-values it is $\boldsymbol{Z}$, in the case of F P-values it is ε in the standard model

$$\boldsymbol{y} = \sum_i \beta_i \boldsymbol{x}_i + \sigma\varepsilon. \tag{8}$$

The F P-value also requires that the covariates $\boldsymbol{x}_i$ are given, are independent of ε and that the data were in fact generated under the model (8). Thus F P-values are valid only for carefully designed simulations whereas Gaussian P-values are universally valid.

5. Approximation based inference

5.1. *Approximation regions*

An approximation base statistical inference will contain no truth based concepts such as F distribution P-values, confidence regions, likelihood, consistency, efficiency. In

Section 3 approximation regions require a P-value, the probability that a random sample generated under the model is accepted as being generated under the model as defined by the approximation, see Davies (2018b). The idea can be extended: an approximation region for the difference of the means of two samples is given by the differences of the means in the corresponding approximation regions. This requires a model to perform the simulations.

In Chapter 5 of Davies (2014) the location-scale is addressed without an explicit model. The approach is a functional one where location and scale values are defined by M-estimators: outliers are explicitly allowed. The discussion in Davies (2014) is not satisfactory but it does make the case for Fréchet differentiability, equation (5.18) of Davies (2014). Given data $\mathbb{P}_n$ one could specify a Kuiper metric ball of size δ and then defined the approximation region as the set of all (μ, σ) which are the location-scale values for some probability measure P with $d_{ku}(P, \mathbb{P}_n) < \delta$.. The choice of δ reflects the range of plausible samples.

For linear regression the situation is more complicated. A valid approximation is the least squares approximation when all Gaussian P-values are less than $p0$. Altering the least squares coefficients by a sufficiently small amount will also give a valid approximation. The author has some ideas how this can be done but the work is unpublished.

6. Indoctrination

In the Cambridge English Dictionary we read:

> indoctrination : the process of repeating an idea or belief to someone until they accept it without criticism or question

The first paper the author wrote on approximation is Davies (1995). It was rejected several times, in all there were about 12 referees' reviews of which only one was positive. Richard Gill, the editor of Statistica Neerlandica at the time , had two negative reviews and one positive. He published in any case.

Various versions of the present paper have also had multiple rejections, about eight in all including twice by the AoS and twice by the JRRS B . One Associate Editor of the AoS was at complete loss as were five statisticians who reviewed for the JRSS, they asked what the point of it was. It was also sent to several statisticians working in the area of high dimensional regression. Apart from Lutz Dümgen (see above) there was no response apart from emails thanking me for sending it.

In contrast to all this, two of the world's most eminent statisticians, John Tukey and Peter Hall, responded very positively. A first version of Davies (1995) was sent to Tukey who replied with the four papers Tukey (1993b), Tukey (1993d), Tukey (1993c) and Tukey (1993a). Part of Tukey (1993d) is cited in Section 3. I had several conversations with Peter Hall about statistical inference with the result that he arranged Davies (2008).

A student studying statistics will learn the "assumed (revealed?) truth" concepts of statistics, hypothesis testing, confidence intervals, significance, Bayesian statistics, prior

distributions, likelihood, AIC, BIC and so on, and this is all they will learn. Exercises and examinations will be formulated using these concepts and these concepts alone. Degrees will only be awarded to candidates who use this language, papers will only be accepted which are written in this language, and the successful students will have their own students and so on. At no point will there by any criticism or questioning All this is reflected in the language of statistics. In other words they have been indoctrinated, see above.

REFERENCES

Y. BENJAMINI, Y. HOCHBERG (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society. Series B, 57, no. 1, pp. 289–300.

G. CLAESKENS, N. L. HJORT (2003). *Focused information criterion*. Journal of the American Statistical Association, 98, no. 464, pp. 900–916.

L. DAVIES (1995). *Data features*. Statistica Neerlandica, 49, no. 2, pp. 185–245.

L. DAVIES (2014). *Data analysis and approximate models*, vol. 133 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

L. DAVIES (2018a). *Lasso, knockoff and gaussian covariates: A comparison*. arXiv:1805.01862 [math.ST].

L. DAVIES (2018b). *On P-values*. Statistica Sinica, 28, no. 5, pp. 2823–2840.

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

P. L. DAVIES, A. KOVAC (2001). *Local extremes, runs, strings and multiresolution*. The Annals of Statistics, 29, no. 1, pp. 1–65.

R. DEZEURE, P. BÜHLMANN, L. MEIER, N. MEINSHAUSEN (2015). *High-dimensional inference: confidence intervals, p-values and R-software hdi*. Statistical Science, 30, no. 4, pp. 533–558.

L. DÜMBGEN, L. DAVIES (2024). *Connecting model-based and model-free approaches to linear least squares regression*. Statistica, 84, no. 2, pp. 65–81.

L. DÜMBGEN, A. KOVAC (2009). *Extensions of smoothing via taut strings*. Electronic Journal of Statistics, 3, pp. 41–75.

L. DÜMBGEN, L. DAVIES (2023). *Connecting model-based and model-free approaches to linear least squares regression*. arXiv e-prints arXiv:1807.09633v4.

I. FALCONER (1999). *Henry cavendish: the man and the measurement*. Measurement Science and Technology, 10, pp. 470–477.

A. FRASER, M. BÈDARD, A. WONG, L. W. WEI, A. FRASE (2016). *Reproducibility and the quest for truth*. Statistical Science, 31, no. 4, pp. 578–590.

N. L. HJORT, G. CLAESKENS (2003). *Frequentist model average estimates*. Journal of the American Statistical Association, 98, no. 464, pp. 879–899.

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

A. KOVAC (2007). *Smooth functions and local extreme values*. Computational Statistics and Data Analysis, 51, no. 10, pp. 5155–5171.

N. H. KUIPER (1962). *On a metric in the space of random variables*. Statistica Neerlandica, 16, no. 3, pp. 231–235.

T. MERRICKS (2007). *Truth and Ontology*. Oxford University Press, New York.

J. SEGERS (2014). *On the asymptotic distribution of the mean absolute deviation about the mean*. arXiv e-prints arXiv:1406.4151.

J. W. TUKEY (1993a). *Discussion- Davies's data sets*. Princeton University, Princeton.

J. W. TUKEY (1993b). *Exploratory analysis of variance as providing examples of strategic choices*. In S. MORGENTHALER, E. RONCHETTI, W. A. STAHEL (eds.), *New Directions in Statistical Data Analysis and Robustness*, Birkhäuser, Basel.

J. W. TUKEY (1993c). *How Davies's data sets might reasonably be approached*. Princeton University, Princeton.

J. W. TUKEY (1993d). *Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper*. Princeton University, Princeton.

J. VON NEUMANN (1947). *The Mathematician*. In M. ADLER, R. HEYWOOD (eds.), *The Works of the Mind*, University of Chicago Press, Chicago.

J. WHITTAKER (2015). *Comment on arxiv:2202.01553: The distribution of a Gaussian covariate statistic*. arXiv e-prints arXiv:2503.11712.