© Grace Scientific Publishing

Journal of
Statistical Theory and Practice
Volume 3, No. 3, September 2009

The Life and Work of Michael A. Stephens: A Conversation with Richard A. Lockhart and John J. Spinelli

Richard A. Lockhart, *Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC V5A 1S6, Canada. Email: lockhart@sfu.ca*

John J. Spinelli, *Cancer Control Reseach, BC Cancer Agency, School of Population and Public Health, University of British Columbia Vancouver, BC, V5Z 1L3, Canada. Email: jspinelli@bccrc.ca*

Received: January 9, 2007 *Revised:* January 16, 2009

Abstract

In April 2007, Michael's colleagues held a two-day meeting to celebrate his 80th birthday. On that occasion Richard Lockhart had a conversation with Michael, which covered his early life. The conversation is on the web at `http://www.stat.sfu.ca/assessment/stephens.php`. It is summarized briefly here, and then John and Richard continue the conversation, focusing on Michael's professional life.

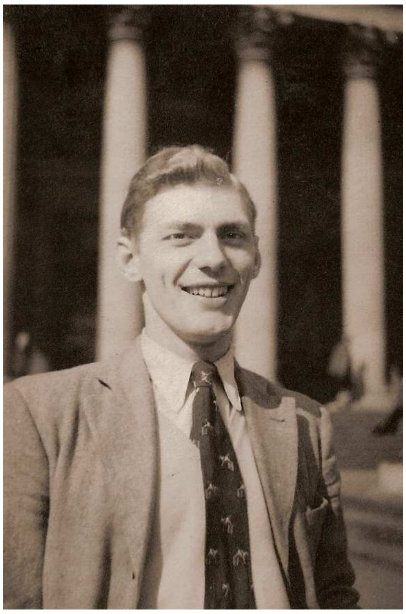
AMS Subject Classification: 01A70; 62-03; 62G10.

Key-words: Goodness-of-fit.

Michael Stephens was born on April 26, 1927 in Bristol, England. His parents were divorced and he lost contact with his mother very early on; later his father was killed in an air-raid during WW2, and Michael was brought up by his grandmother. In 1938, he won a scholarship to a very old boarding school, then dedicated to giving an education to 'poor boys and orphans' from Bristol.

In 1945, he obtained an Open scholarship to Bristol University, and took an Honours degree in Physics in 1948. He became the first holder of a Frank Knox fellowship to Harvard University, and completed his AM in 1949. He wanted to change from physics to mathematics, but with a physics degree there was no chance to enter one of the few and small British universities. He taught at Tufts College, then returned to teach at Woolwich and Battersea Polytechnics in London. Salaries were abysmal, so in 1956 he became an Instructor at Case Institute of Technology (now Case Western Reserve University) before entering, three

* 1559-8608/09-3/$5 + $1pp – see inside front cover



Michael at Bristol

years later, the University of Toronto to study statistics under Geoffrey Watson. He received his PhD in 1962. He was a faculty member at Toronto, McGill University and McMaster University in Canada and the University of Nottingham in England before moving to Simon Fraser University in 1976. Michael is a Fellow of the Institute of Mathematical Statistics and the American Statistical Association. He is a past President of the Statistical Society of Canada, and was awarded the Gold Medal of the Society in 1989. While he was in Toronto Michael met and married Evelyne Lucas, an immigrant from France. In 1967 their daughter Madeleine, who is a veterinarian, was born in Montreal, and she and her businessman husband, Malcolm, have four boys under 12. Madeleine has recently decided to be a stay-at-home mum. Michael's best-known research contributions are in the areas of directional data analysis and, of course, goodness-of-fit. In 1986, with Ralph D'Agostino, he edited the definitive book, *Goodness-of-Fit Techniques*. Michael has been a Professor Emeritus at Simon Fraser University since 1992, but has maintained an extremely active research program.

JJS-RAL: Your thesis work was on directional data.

MAS: Yes, for my thesis, Geof had introduced me to directions, a new field where there was a lot to do, and it got me off to a good start. I liked writing the early papers on directional data, and I have recently come back to the subject. As I've grown older, I've become more interested in providing practical results for statisticians, for example, writing on the effects of grouping data, which one often sees with directions, and fitting mixtures of distributions.

JJS-RAL: How did you get interested in Goodness-of-Fit?

Geoffrey Watson (left) and Michael at his
Ph.D. graduation in Toronto in 1962

MAS: While I was his student, Geof Watson introduced U^2, an adaptation of the Cramér-von Mises statistic for testing fit on the circle. I did some Monte Carlo (MC) studies, and became interested in finding the small-sample distribution. This involves Euclidean geometry, a subject I always loved.

JJS-RAL: That was in 1962?

MAS: Oh yes, a bumper year for me. I got my Ph.D., had my first two papers published, met Egon Pearson, and got married in December.

JJS-RAL: Evelyne comes from France. How did the two of you meet?

MAS: Yes, Evelyne was born in Le Havre, the big seaport in Normandy. As you can imagine, she had a very rough childhood during the war. Her schooling suffered as schools opened and closed, and her mother, sister and Evelyne moved from place to place. Later she trained as a couturiere and worked for a top fashion house in Paris. But the girls who make beautiful clothes are not the ones that make the money. So she decided to emigrate, and chose Toronto because she had met a Canadian student, Helen, in Paris. I knew Helen from a German course I was taking and I met Evelyne on her first weekend in Toronto. She started to make beautiful clothes for a small clientele, and I should add that she still does, but now only for herself and our daughter. And four young grandsons know where to go when they have a hole in their pants.

JJS-RAL: You mentioned meeting Egon Pearson in 1962.

MAS: Egon Pearson (ESP) was the editor of Biometrika and was getting interested in the Kolmogorov-Smirnov (KS) and Cramér-von Mises (CvM) statistics, and when I went to England on vacation after defending my thesis I told him I was interested in finding the distribution of U^2. He said, in his gentlemanly way, "it would be nice if you could find the moments, and then you could use my father's curves". I had no idea what his father's curves were, but found a good description in Kendall and Stuart and a book by Elderton. Karl Pearson, Egon's father, had invented a family of densities, which can be identified by the first four moments.The percentage points of the density can also be related to the moments, and a small set of such points had been given in Biometrika Tables, volume 1. At the time, ESP was getting very extended tables of percentage points of father's curves compiled in America (Johnson, Nixon, Amos and Pearson, Biometrika, 1963), repeated in Biometrika Tables, Volume 2 (Pearson and Hartley, 1972). So after London, I went to see Evelyne, who was visiting her family, and I stayed in her grandfather's small house — more like a cabin really — in a village where a very extended family had hidden from the bombings during the war. I sat on the terrace and set about finding the moments of U^2. Then I used Pearson Curves to approximate the distribution. ESP had given me his handwritten copy of the tables, and when I returned to Canada and put these on cards, I wrote a Fortran program to fit the Pearson curves. There are 17 tables, and each table took 45 minutes to type. I used the curves to approximate some very unpleasant densities in my directional work, and I still use the routine.

JJS-RAL: You became more interested in CvM statistics?

MAS: I continued my friendship with ESP, and he let me know he wanted to put tables of KS and CvM statistics in Biometrika Tables, Volume 2, which he was preparing in the '60s. These would be for what I call Case 0, when all parameters in the tested distribution are known, and the test reduces to a test for uniformity. I said that these tables (different for different sample sizes) would take a lot of room. His response was, "perhaps you could shorten them". I was able to do that (Stephens, 1970) and for years thereafter shortening tables by 'modifying' the original statistic became almost as much an obsession as finding moments.

JJS-RAL: You wrote a number of papers giving points for tests when parameters were unknown.

MAS: I realised that tables were also needed for cases where parameters must first be estimated from the sample, so I decided I'd better learn something about these statistics. In about 1970, at Stanford, I read Anderson and Darling's beautiful 1952 paper on CvM statistics (Anderson and Darling, 1952). This explains the asymptotic theory of CvM statistics, but does not carry through with any specific tests for the unknown-parameter case. I think I was rather lucky that they didn't; I started to work on these, and in the Annals in 1976 (after 5 years under review) I published percentage points for the asymptotic distributions of CvM statistics for testing normality and exponentiality (Stephens, 1976). This was the first of several papers giving points for other distributional tests.

Another impetus to learn more about goodness-of-fit came when ESP told me he had doubts about the power studies in the paper by Shapiro and Wilk in Biometrika (Shapiro and Wilk, 1965). They introduced their statistic for testing normality, which is based on a very appealing idea, to regress the order statistics of the sample on expected values of standard normal order statistics. They gave tables which appeared to show that the statistic was overwhelmingly more powerful than EDF or other older statistics. It turned out that their powers of EDF statistics had been calculated using Case 0 points, which are much larger than the points when parameters are estimated. When corrected, CvM powers were much closer to the SW powers.

JJS-RAL: This was given in your 1974 JASA paper (Stephens, 1974)?

MAS: Yes, but I always thought that the paper was incomplete. I should have given a bigger power study. But I did try to give lines of constant power on a plot of β_1 against β_2. I called these 'isodynes', after consulting at enormous length a scholar of classical Greek who apparently knew at least 50 words for power. I had had two years of classical Greek at my boarding school, but at last I learned the meaning of 'The Greeks have a word for it'. That paper was close to being rejected because, as the referee put it, "Stephens just chats about this and that", suggesting that there was nothing of substance in the paper. Perhaps he didn't like my Greek scholarship. No-one else has, either; 'isodynes' hardly rivals 'bootstrap' in the world of statistics. Fortunately, Norman Johnson, the editor, in sending the report of the referee, said that 'nevertheless I have decided to publish your paper'. It's been by far my most cited paper.

JJS-RAL: What other articles have you especially enjoyed writing?

MAS: The Shapiro-Wilk statistic is based on a straight line fit; they used the ratio of the variance estimate obtained by generalised least squares with the usual sample variance. Unfortunately this works well only for the normal distribution. Sarkadi (1975) showed that the technique gives inconsistent tests if applied to the exponential distribution, and John, you and I followed that up in Spinelli and Stephens (1987). So I wondered why it worked so well for the normal test. I had noticed that it was very close to the correlation coefficient for that case. This led to a paper in Biometrika (Stephens, 1975) where I found the asymptotic eigenvalues and eigenvectors of the covariance matrix of normal order statistics, and then could show the asymptotic equivalence of the SW statistic to the correlation coefficient. This was only heuristic. It was later shown rigourously, but not by me; the essential steps are by Julian Leslie in a paper given at a GoF conference in Hungary. This work led to more interest in the correlation coefficient on a Q-Q plot as a test statistic. I really enjoyed the paper we wrote, Richard, in one of the volumes of Balakrishnan and Rao (Lockhart and Stephens, 1998) where we gave the large-sample results for testing a number of distributions, with both full and censored samples. A great unread article, I fear.

JJS-RAL: You told a story at your 80th about a paper that nearly got published with the wrong title. You must have enjoyed that one.

MAS: That was a nice paper I wrote with Iain Currie 'Tests of Fit Based on Sample Moments', which came to me at the proof stage with 'movements' instead of 'moments'. As I said at my 80th, I was by then very tempted to change 'sample' to 'bowel', and send it to the New England Journal of Medicine. I'm sure I'd be on the national news, and fame at last!

JJS-RAL: What is the work you are most proud of?

MAS: That has to be the book, D'Agostino and Stephens (1986). I think 'research' books, (I am not thinking of textbooks, I am just envious of the authors of those when they sell well) are vital tools for research. I am thinking of the Kendall and Stuart volumes, or the works of Johnson and Kotz — they are the places you go to first, to see what has been done in the field, to get references, etc. In the early 80s, some colleagues came to me at an ASA meeting and suggested that we write a book about goodness-of-fit. This developed into a suggestion that Ralph D'Agostino and I should be editors. I insisted that the book should be about practical techniques of testing fit, because if we opened the door to including the theory, there would be several volumes. I'm afraid I took a long time satisfying myself that my own contributions were adequate and sufficient, until I had a meeting with Marcel Dekker himself (an 'uncle-nephew' meeting he called it), and I was pushed into submitting it with any imperfections that remained. Ralph D'Agostino was a superb editor and has remained a good friend. I still get a royalty cheque every year, more than enough to pay for a new pair of shoes. People around the world still ask me questions, and are very complimentary about the book. I wish I could see better and update it.

JJS-RAL: Tell us more about your use of Pearson curves to approximate distributions.

MAS: Oh, yes. I have had enormous mileage (kilometrage?) from Pearson Curves. At the time ESP introduced me to them, the curves had fallen out of favour because they were used to fit 'distributions' to sample data. The resulting distributional form would look absurd. But when the theoretical moments of a density are used to define an approximating Pearson curve, the approximation can be very good, especially in the long tail, usually the upper tail. Essentially, one is laying a curve on top of the true distribution, and making sure that the first four moments all agree. Your true distribution would have to have a very unpleasant density not to be well-fitted. Of course I have been criticised because there is no measure of accuracy, but I have had some fun by showing excellent agreement with much more tedious methods. In the years I was going to Stanford I used Pearson curves to find percentage points in a number of problems, which resulted in papers with Herb Solomon, Ingram Olkin and others. One important distribution where the cumulants are easy to find is the sum of weighted independent chi-squares, and it was matching points obtained by Herb Solomon in other ways that led to Solomon and Stephens (1978). I was asked for my program some years ago by a famous financial institution in New York, and was assured it would be used thousands of times a day. So when I see the stock market fluctuating wildly, I sometimes wonder if father's curves had something to do with it.

JJS-RAL: You continued your friendship with Egon Pearson?

MAS: Yes, I felt it a great honour to know him, and I still have the handwritten letters (no email in those days!) he wrote me. I feel that Egon Pearson has been under-appreciated as the practical British statistician that he was. I remained friendly with ESP after he went to live in a retirement home in beautiful countryside south of London where he intended to edit his father's papers. There were large files on a high shelf in his room, labeled 'KP papers'. They must have contained wonderful material concerning the early days of modern statistics, although I imagine appreciative letters from Fisher must have been lacking. I don't believe Egon ever got around to finishing that last labour of love, and it would be interesting to know if historians have made use of the papers.

JJS-RAL: You mentioned Stanford, and we know you went every summer. How did that come about?

MAS: I was visiting Geof Watson at Johns Hopkins in 1962 and 1963. He had invited a group of distinguished visitors, and I met C.R. Rao, E.J. Pitman, Jim Durbin, and Rupert Miller among others. Also, David Duncan was on the biostatistics faculty, and he and his wife were very kind to us. Later, Rupert invited me to give a seminar at Stanford, and there I met Herb Solomon. He did not attend the seminar, but he did go to the washroom, and Richard Olshen, in charge of seminars, introduced us when we met in the corridor. Herb invited me to come for the summer, and I continued those visits for 25 years. They were wonderful summer visits, and of course I met many great statisticians. Herb and his wife, Lotte, were always very hospitable, and working with Herb was fascinating. He used to hold brain-storming sessions to tackle problems brought by his granting agency and I learned a lot from my younger colleagues on his team.

JJS-RAL: You referred to poor eye sight when talking about the book; why was that?

MAS: For several years I have been suffering from macular degeneration, an eye problem which causes blind spots on the retina. These usually affect the centre of the vision; I have lost the sight of my right eye, and the left has the problem too. But I have been having a treatment recently which I hope will stabilize the situation. I can still work, thanks to big screens and large fonts, but naturally I am much slower and make mistakes. Most importantly I can no longer read normal print, and that is a real handicap, because I cannot read the Journals. I tell you this so that when I stick my neck out, I hope not to have the axe fall from learned colleagues who come to the conclusion that Stephens definitely should stop chatting about this and that.

JJS-RAL: But you do still enjoy statistics?

MAS: Yes, partly because I have no real hobbies. When I was young, the King of Sweden played tennis when he was 85 and I thought “That’s what I’ll do when I get old" but when I last tried to play the ball sailed past me even more often than when I could see better. I used to love to play squash too. I like putting flowers in pots, but that’s about it as far as gardening goes. I used to love reading but that’s not on nowadays. I can still enjoy the theater and TV, but I can’t see the faces too well. And of course there are concerts. My greatest thrill is still to work on a mathematical problem.

JJS-RAL: Any problems in particular?

MAS: Yes. For years I have been fascinated by the connections between methods of testing exponentiality using either ‘classical’ EDF tests or EDF tests based on the Total Time on Test. Both of these methods have a very understandable motivation, and I noticed in the book that they had similar powers. These connections help to explain why. There has been a lot of work done in GoF on Neyman’s smooth tests, including the book and overview by Rayner and Best (Rayner and Best, 1989, 1990) — two other excellent ‘research sources’ — and on expanding EDF statistics into components, very similar to Neyman’s components. I wrote a great unread paper in France when I spent a very pleasant year in Grenoble. The paper gives components of EDF statistics, following an idea I believe originated with Geof, and which was followed up in two interesting papers by Durbin and Knott (1972) and Durbin, Knott and Taylor (1975). More recently much work has been done in Europe and elsewhere, and I can quote only a few examples, to show the variety of the approaches: Henze and Klar (2002) for tests based on the empirical characteristic function (ECF), Fan (1996) and Fan and Huang (2001) using wavelets: Inglot, Kallenberg and Ledwina (1997) and Kallenberg and Ledwina (1997) for data-driven tests. As the different methods often give similar powers, I think it would be interesting to see why, and not just by Monte Carlo! New statistics are usually compared for power, although the applied worker doesn't always care!

JJS-RAL: How do you mean?

MAS: Well, many statisticians don't really want to make goodness-of-fit tests; they especially don't want to be told their observations are not normal or log-normal, for instance. So all that emphasis on getting the most power is wasted — back to good old chi-squared. For me, the emphasis on power was so that, if a test was significant for a small sample, you could think of a transformation — Box-Cox, or following the earlier work of Anscombe — so that the model would better fit the data. For small samples GoF tests can often be used to assess which of several models looks best. And that brings up the question; how nearly right do you want the model to be? As George Box said: "All models are wrong, but some are useful.". We know that a large enough sample will always lead to a rejection of a test of fit. (Of course, it will lead to rejection of any other fixed hypothesis too, but somehow this is always held against tests of it). But all the more reason to test small samples. A practical statistician knows that the model is only an approximation, and wants to feel that he or she may safely use the model without dangerous errors; if the sample is very large and the test rejects, the model might still be good enough for practical use.

JJS-RAL: You called this good-enoughness-of-fit sometime ago?

MAS: I called it good-enoughness-of-fit at the excellent Paris conference in 2000. I am still interested to know how we can use tests to decide if a model is good enough. For one or two-dimensional problems it is easy enough to see how a model change might influence your inference, but with the complicated models made possible by the computer, this is not so easy.

JJS-RAL: That will depend a lot on the inference procedures.

MAS: Yes and that is an important area of research: to look at how changes in the model will influence the inferences. Watson's ANOVA for directions provides a good example. I have seen papers in biology which apply Watson's F-test to see if the birds or butterflies know the way home, without bothering to test if the directions are von Mises, which is the assumption leading to the ANOVA technique. But if the test is robust against departures from von Mises, then the distribution is good enough to apply it.

JJS-RAL: The computer has revolutionised statistical practice during your career.

MAS: The computer was just coming into use when I got my Ph.D. at Toronto. I used an IBM 650 — you had to book time on it and watch that the air-conditioning didn't break down — and it took me all night to get nearly 20,000 MCs which would take seconds today. Then Fortran came on stream: I earned $100 for writing a program to calculate a complex-number function, not for mathematicians, but for chemists. Don't ask . . . .

Richard Lockhart, David Brillinger, Michael and Charmaine Dean

JJS-RAL: How do you think the computer has influenced GoF?

MAS: First of all, I believe statisticians should be helpful to other disciplines wherever we can. And for GoF, I think the applied worker wants tests with a natural motivation, perhaps coming from graphical display. For grouped data, comparing the observed with the expected is very natural, and similarly comparing the EDF with the tested distribution, or comparing components with their sample values. I notice that when complex models are proposed, the researcher can sometimes produce two variables which should give a straight line plot. This is often judged by eye, but I believe tests should be provided. The computer enables us to create very complex models and to perform huge Monte-Carlo studies; in short, to do things which statisticians knew they would like to do fifty or more years ago, but couldn't. But recently I've been finding there is another impact, something more subtle.

JJS-RAL: What is that?

MAS: The computer might point the way to a deeper understanding of the subject, the interconnections again. This has impressed me quite recently. As you know, Richard and I, with our friend Federico O'Reilly, have been investigating exact tests of fit based on sufficiency (Lockhart, O'Reilly and Stephens, 2007, 2009). We used the Gibbs sampler to get Monte-Carlo samples from the conditional distribution of the data, given the sufficient statistic, in order to find the distribution of a GoF test statistic. For the Gamma and von Mises distributions, we found a startling agreement of p-values with those given by the so-called parametric bootstrap, where one fits the tested distribution and draws samples from it. The correlation between p-values is truly remarkable and this of course suggests an underlying connection between these two very different techniques. I know that you, Richard, decided there was a theorem hidden away there (Oh, the advantages of a Berkeley education!), and I hope you have proved it. There you have the computer giving new research results which may never have occurred to us to look for if we did not have that power.

JJS-RAL: What major problems would you recommend young researchers to work on nowadays?

MAS: Helping to see if a complex model fits the data. The distribution theory behind these tests will probably be impossible to derive, but that is where the computer comes in with Monte Carlo studies. Of course, these studies must be correct. The problems arise with both frequentist and Bayesian analyses, since Bayesians often seem to be more concerned about whether their priors are acceptable but take little interest in testing if the likelihood fits the data. I get a bit tired of being told that a frequentist cannot 'get an answer', but a Bayesian can, after introducing a useful prior. Actually, Jin Zhang and I have recently introduced a prior to help get an estimate where MLE fails. I hope that gives me permission to use MCMC if I ever need it. I also feel that there are many situations where GoF statistics can be useful in a wider sense. For example, I worked with John Orwin (Orwin, McKinzey, Stephens and Dugmore, 2008) to examine the complete distributions of lichens in glaciology, whereas previously much less informative summary statistics had been used. I also worked with Fritz Scholz on testing whether fire stations responded to emergencies in a similar fashion, again using the comparison of the complete distributions rather than summary statistics. There are lots of opportunities to use GoF statistics in these more powerful ways.

JJS-RAL: Do you think interest in theory has declined?

MAS: Yes, for example, we had some difficulty getting both of those papers on exact versus approximate tests published. And surely the speed of rejections must provide a clue! But there's a new journal devoted to distribution theory, so that should help. I wonder if all the beautiful theory, developed by great statisticians in the last century, has been superseded by a computer which has no need of this theory. Is the age of elegance dead? I know that some colleagues think so. In the heady days of the '20s and '30s, when there were no computers, only brains, and Fisher and Neyman and Pearson were going at it hammer and tongs, it was natural to try to find optimum techniques, use sufficiency, try to see if you could get the minimum variance, etc. Actually, I sometimes wonder where we would be if they had got along better — perhaps we would not be frequentists or Bayesians, but all be fiducialists? But your eyes are glazing over . . . .

JJS-RAL: Yes, perhaps we'd better not go into that! There's something else . . . . In a tribute to you on your 80th birthday, Julian Leslie wrote that you were "the most entertaining statistician I ever met". We in Canada have often enjoyed your way with words, and there's been a few examples here. Also, when you give talks. Any comment?

MAS: H'm, save the most difficult question till the last! These are kind compliments, and I might only ruin it by commenting! Better just to say that I know my place in the statistical firmament — Canada has been good to me, and it's good to be appreciated, no matter what it's for!

JJS-RAL: Perhaps we should finish by a few remarks about the development of Statistics during your time in Canada?

MAS: There has been enormous progress. In the early '60s, there was very little activity in Statistics — Don Fraser was the outstanding star, but there were no really strong groups.

However, during the '60s things started to move — I suspect the Russians kindly putting up Sputnik loosened the purse-strings of the governments. The first statistics department was created at Waterloo, and good statisticians joined Universities across the country. Some were newcomers, and some were returning from the US, where previously, many of the best Canadian statisticians had made their careers. We also had the opportunity to belong to not one, but two statistical societies in Canada! Fortunately they finally merged, and we now have one flourishing society with a good journal. Statistics Canada, under first Martin Wilk and then Ivan Fellegi, has also become one of the outstanding statistical agencies in the world, renowned for its strict observance of confidentiality and for the breadth of its surveys. I served on the National Statistics Council for four years, and it was an exciting experience. Finally, it has been good to see the creation of Statistics departments separate from Mathematics; I have come to believe that this is essential for the subject to flourish in a University. Our granting agency seems to agree: for many years the statisticians have been judged by their own peers, and not by mathematicians. And at Simon Fraser University, although only 40+ years old, we now have a very strong group, including my interviewers, or should I say inquisitors? I am very proud to have been a part of that.

JJS-RAL: Thank you, Michael.

References

Anderson, T., Darling, D., 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23, 193–212.

D'Agostino, R., Stephens, M., 1986. *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Durbin, J., Knott, M., 1972. Components of Cramér - von Mises statistics. *Journal of the Royal Statistical Society, Series B*, 34, 290–307.

Durbin, J., Knott, M., Taylor, C., 1975. Components of Cramér - von Mises statistics: II. *Journal of the Royal Statistical Society, Series B*, 37, 216–237.

Fan, J., 1996. Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, 91, 674–688.

Fan, J., Huang, L., 2001. Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association*, 96, 640–652.

Henze, N., Klar, B., 2002. Goodness-of-fit tests for the inverse Gaussian distribution based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics*, 54, 425–444.

Inglot, T., Kallenberg, W., Ledwina, T., 1997. Data driven smooth tests for composite hypotheses. *Annals of Statistics*, 25, 1222–1250.

Johnson, N., Nixon, E., Amos, M., Pearson, E., 1963. Table of percentage points of Pearson curves, for given $\sqrt{\beta_1}$ and β_2, expressed in standard measure. *Biometrika*, 50, 459–498.

Kallenberg, W., Ledwina, T., 1997. Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, 92, 1094–1104.

Lockhart, R., O'Reilly, F., Stephens, M., 2007. Use of the Gibbs sampler to obtain conditional tests with applications. *Biometrika*, 94, 992–998.

Lockhart, R., O'Reilly, F., Stephens, M., 2009. Exact conditional tests and approximate bootstrap tests for the von Mises distribution. *Journal of Statistical Theory and Practice*, 3(3), 543–554. Accompanying paper.

Lockhart, R., Stephens, M., 1998. Probability plot: use of the correlation coefficient, 453–473. in *Order Statistics: Applications. Handbook of Statistics*, Vol. 17, Balakrishana, N., Rao, C.R. (editors). Elsevier Science, Amsterdam.

Orwin, J., McKinzey, K., Stephens, M., Dugmore, A., 2008. Identifying moraine surfaces with similar histories using lichen size distributions and the U^2 statistic, Southeast Iceland. *Geografiska Annaler: Series A*, 90, 151–164.

Pearson, E.S., Hartley, H., 1972. *Biometrika Tables for Statisticians.*, Volume 2. Cambridge University Press, New York.

Rayner, J., Best, D., 1989. *Smooth Tests of Goodness-of-Fit.* Oxford University Press, New York.

Rayner, J., Best, D., 1990. Smooth tests of goodness of fit: an overview. *International Statistical Review*, 58, 9–17.

Sarkadi, K., 1975. The consistency of the Shapiro-Francia test. *Biometrika*, 62, 445–450.

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

Solomon, H., Stephens, M., 1978. Approximations to density functions using Pearson curves. *Journal of the American Statistical Association*, 73, 153–160.

Spinelli, J., Stephens, M., 1987. Tests for exponentiality when origin and scale parameters are unknown. *Technometrics*, 29, 471–476.

Stephens, M., 1970. Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society, Series B*, 32, 115–122.

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

Stephens, M., 1975. Asymptotic properties for covariance matrices of order statistics. *Biometrika*, 62, 23–28.

Stephens, M., 1976. Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, 4, 357–369.