

## PETER HALL ON EXTREMES: RESEARCH, TEACHING AND SUPERVISION

A. H. Welsh

*Australian National University*

*Abstract:* We examine Peter Hall's early research, undergraduate teaching, and PhD supervision, using the theme of extreme order statistics to highlight interesting aspects of these activities. Focusing on this period allows us to see Peter when, like all academics in the early part of their careers, he was becoming an academic and still establishing himself. That he succeeded so greatly and rapidly began to make the many remarkable contributions that adorn his distinguished career, makes this early, formative stage particularly interesting to explore.

*Key words and phrases:* Endpoint of a distribution, order statistics, rates of convergence, regular variation, threshold selection.

### 1. Introduction

Peter Hall's many and wide-ranging contributions to statistics have been celebrated in a number of events (conference sessions, workshops and conferences) and in a special October 2016 issue of the *Annals of Statistics*. In addition, a Memoir of Peter's life and work has been published by Robinson and Welsh (2017). Nonetheless, some aspects of Peter's life and work, particularly in the very important early part of his career, have not been discussed in much detail. The purpose of this paper is to examine Peter's early research, undergraduate teaching, and PhD supervision, using the theme of extreme order statistics (hereafter just called extremes) to highlight their interesting aspects.

Peter's research on extremes uniquely straddles three important periods in his research career: his earliest research in probability; the period in which he redirected his research from probability to statistical theory; and his later work on statistical theory. This enables us to illustrate how what he did and the way he did it changed through these different periods. The research on extremes is interesting because it contains interesting ideas of value in themselves, as well as links to other areas of research interest to Peter. One key idea that forms a theme for Peter's research on extremes is that of using simple models to construct estimators and then deriving the properties of these estimators under different

approximate models. Particularly, later in his career, Peter incorporated methods and problems from other areas of his research (such as the iterated bootstrap for interval calibration or measurement error) into his research on extremes so this work provides some insights and a nice gateway into these other areas of his research. Since Peter did not publish as many papers on extremes as he did in some other fields of interest such as nonparametric function estimation or the bootstrap, his papers on extremes form a relatively manageable body of work (particularly by Peter's standards) for interested researchers to work with. These features make Peter's research on extremes uniquely useful for discussing the evolution of Peter's research, the development of his ideas, and the changes in the style of writing papers.

Peter's exceptional research productivity can make it easy to overlook other important activities in his early academic career. Peter's first academic appointments were as a lecturer (junior faculty member) at the University of Melbourne from August 1976 to mid-1978 and then in the Department of Statistics in the School of General Studies (formally renamed The Faculties in 1979) at the Australian National University (ANU) from mid-1978. The ANU had two statistics departments; the Department in the School of General Studies was responsible for undergraduate teaching and research while the Department in the Institute of Advanced Studies was a research department. Peter's appointment changed to a joint appointment between the two ANU departments in 1986, but Peter carried a full teaching load at ANU until then. That is, during the period that he redirected his research from probability to statistics and increased his productivity, Peter carried a full teaching load. In addition, he began to supervise PhD students. I audited one of Peter's advanced undergraduate courses (Order Statistics and Related Topics) in 1982 and worked under his supervision on estimating parameters of regular variation during my PhD (1982 – 1984). My direct experience of both of these important activities was therefore related to extremes. Discussing Peter's early teaching and supervision gives additional insight into Peter and an important stage of his career.

One of the conditions of Peter's appointment to ANU in 1978, was that he direct some of his research effort towards statistics. Peter took this seriously and started to work on statistical problems, beginning with nonparametric density estimation (Hall (1980a)). His statistical research on extremes started soon after, with Hall (1982a) treating estimation of the exponent of regular variation and Hall (1982b) treating estimation of the endpoint of a distribution. Peter had completed an undergraduate statistics degree at the University of Sydney, read

Table 1. Peter’s papers in refereed journals 1977–1982: “total” is the total number of papers, “statistics” is the number of papers in statistics (as opposed to probability), and “joint” is the number of papers with at least one other co-author.

year	total	statistics	joint
1977	2	0	0
1978	6	0	0
1979	8	0	1
1980	4	1	0
1981	17	8	2
1982	17	11	5

widely and was extremely smart. He sensibly worked on problems in statistical theory that made use of his strength, knowledge and ability in probability. Nonetheless, it is remarkable that he redirected his research so successfully and so quickly. Peter’s first promotion to Senior Lecturer came in 1983, by which time he had published over 54 papers in refereed journals and 2 books. This was followed by promotions to Reader in 1986, and then Special Professor in 1988.

We discuss Peter’s early probability research on extremes in Section 2 and his teaching in Section 3. We examine his statistical work on estimating the exponent of regular variation and the endpoint of a distribution in Sections 4 and 5, respectively; we discuss in particular the two papers Hall (1982a,b) and then the research that these papers led to. We describe Peter’s early PhD supervision in Section 6 and conclude with a brief discussion in Section 7.

## 2. Early Research on Extremes

A summary of Peter’s papers published in refereed journals between 1977 and 1982 is given in Table 1. The first column shows the number of papers, the second shows the number of them that are clearly in statistics as opposed to probability and the third shows the number that had additional authors. It is clear that Peter’s research in this period was mainly in probability, his papers were mostly single authored and he went very quickly from highly productive to exceptionally productive while at the same redirecting his research to statistics.

Peter’s early research on extremes was focused on establishing limit theorems and rates of convergence for extremes and functions of near extremes from samples of independent random variables. This fitted very well within Peter’s concurrent research on martingales and rates of convergence in the Central Limit Theorem. Indeed, this is a good reason for looking at the content and style of these papers.

Peter's first two papers on extremes established limit theorems and representations. In Hall (1978a), Peter considered independent and identically distributed (iid) random variables  $X_1, \dots, X_n$  in the domain of attraction of a stable law with exponent  $\alpha < 2$ . Let  $Y_{n1}, Y_{n2}, \dots, Y_{nn}$  denote the sample arranged in decreasing magnitude so  $|Y_{n1}| \geq |Y_{n2}| \geq \dots \geq |Y_{nn}|$ . Peter studied the asymptotic behaviour of the lower and upper sums  ${}^{(k)}S_n = Y_{n1} + \dots + Y_{nk}$  and  $S_n^{(k)} = Y_{n,k+1} + \dots + Y_{nn}$  as first  $n \rightarrow \infty$  and then  $k \rightarrow \infty$ . Suitably normalised, he showed that  $S_n^{(k)}$  converges in distribution to a random variable  $T^{(k)}$  and found the characteristic function of the limiting random variable. He used this result to show that  ${}^{(k)}S_n$  converges in distribution to  ${}^{(k)}T$  as  $n \rightarrow \infty$  and then that, as  $k \rightarrow \infty$ , for a suitable nonstochastic sequence  $\{c_k\}$ ,  ${}^{(k)}T - c_k$  has a limiting stable law.

In his next paper (Hall (1978b)), Peter considered i.i.d. random variables  $X_1, \dots, X_n$  with order statistics  $X_{n1} \leq X_{n2} \leq \dots \leq X_{nn}$ . Suitably normalised, the variables  $X_{nn}, X_{n,n-1}, \dots$  have a nontrivial limiting distribution  $\xi_1, \xi_2, \dots$ . Peter established a Rényi-like representation (Rényi (1953)) for  $\{\xi_n, n \geq 1\}$  and used it to obtain limit theorems for  $\xi_n$  as  $n \rightarrow \infty$ : Let  $Z_1, Z_2, \dots$  be independent standard exponential random variables, let  $\gamma$  be Euler's constant, and let

$$\zeta_n = \sum_{j=n}^{\infty} \frac{(Z_j - 1)}{j} + \gamma - \sum_{j=1}^{n-1} \frac{1}{j}.$$

Then the distribution of  $\{\xi_n, n \geq 1\}$  is the same as that of  $\{\tilde{\xi}_n, n \geq 1\}$ , where

$$\tilde{\xi}_n = \begin{cases} \exp\left(\frac{\zeta_n}{\alpha}\right) & \text{if } X_1 \text{ is in the domain of attraction} \\ & \text{of the Type 1 distribution} \\ -\exp\left(-\frac{\zeta_n}{\alpha}\right) & \text{if } X_1 \text{ is in the domain of attraction} \\ & \text{of the Type 2 distribution} \\ \zeta_n & \text{if } X_1 \text{ is in the domain} \\ & \text{of attraction of the Type 3 distribution} \end{cases}$$

and  $\alpha$  is the parameter in the limiting distribution.

The motivation given in Hall (1978b) was that "the process  $\{\xi_n\}$  deserves study in its own right". Later in the paper, Peter noted that the representation can be used simply to show that various random variables are independent. In fact, he used the results from Hall (1978a,b) in statistical work on extremes in Hall and Wang (2005). Thus, whatever the motivation given or omitted from the paper, the results were important. It also shows how Peter's early research in probability helped prepare him for and influenced him in his later research in statistics.

These first two papers were followed by several papers on rates of convergence for extremes.

Hall (1979a) considered the  $r$ th largest order statistic  $X_{nr}$  from a set of iid random variables. Let  $\{c_n\}$  be a sequence of constants such that for the smallest order statistic  $X_{n1}/c_n \rightarrow 1$  in probability, or almost surely, then  $X_{nr}/c_n \rightarrow 1$  in probability, or almost surely, for each  $r \geq 1$ . Peter obtained necessary and sufficient conditions for  $r(n) \rightarrow \infty$  sufficiently slowly such that  $X_{nr(n)}/c_n \rightarrow 1$  in probability, or almost surely, and then investigated the rate of convergence by establishing a central limit theorem and a law of the iterated logarithm.

For the largest order statistic  $X_{nn}$  from a sample of  $n$  iid standard normal random variables, Hall (1979b) showed that, with the optimal choice of normalising constants, the supremum metric and the Lévy distance between the normalized  $X_{nn}$  and its limiting extreme value distribution decrease like  $1/\log(n)$ .

Following on from Hall (1979b), Hall (1980b) considered the problem of estimating the distribution function  $\Pr(X_{nn} \leq x)$  of the largest order statistic  $X_{nn}$  from a sample of  $n$  iid standard normal random variables. The research was motivated by the result in Hall (1979b) showing that the rate of convergence of  $X_{nn}$  to the extreme value distribution is very slow. In this problem Peter first showed that taking powers of  $X_{nn}$  improves the rate of convergence to  $1/\{\log(n)\}^2$ , but no more. This means that using the limiting distribution to approximate the actual distribution is not accurate enough, so Peter derived a new approximation with relative error tending to zero as  $x$  and  $n$  tend to infinity.

Hall (1984) considered the sample mean of  $n$  iid observations from a symmetric (about zero) distribution after trimming the  $k$  largest (in absolute value) observations. If  $k/n \rightarrow 0$  as  $n \rightarrow \infty$ , the trimmed mean has the same asymptotic distribution as the mean. Peter showed that the rate of convergence can be made arbitrarily close to  $O(n^{-1})$  by an appropriate choice of  $k$ . He showed further that scaling by  $(n - k)^{1/2}/\sigma$ , where  $\sigma^2$  is the variance of the underlying distribution, as one ordinarily would, does not improve the rate of convergence; the scale factor needs to be a function of the last removed extreme value but is not usually available in practice. Thus Peter reached the negative conclusion that there is no practical way of improving the rate of convergence by trimming. Peter adopted a non-standard description of robustness (in terms of the trimmed estimator having a faster rate of convergence than the untrimmed estimator) so he expressed his conclusion in terms of robustness not being achievable by trimming.

In a paper motivated by research by Daryl Daley into a storage model, Daley and Hall (1984) derived limit laws for maxima of weighted  $\{w_n(\gamma)X_n\}$

and translated  $\{X_n - v_n(\delta)\}$  sequences of iid random variables. They assumed  $1 \geq w_n(\gamma) \geq w_{n+1}(\gamma) \rightarrow 0$  as  $n \rightarrow \infty$ ,  $w_n(\gamma) \rightarrow 1$  as  $\gamma \rightarrow 1$ , and  $0 \leq v_n(\delta) \leq v_{n+1}(\delta) \rightarrow \infty$  as  $n \rightarrow \infty$ ,  $v_n(\delta) \rightarrow 0$  as  $\delta \rightarrow 0$ . Daley and Hall showed that for  $w_n(\gamma) = \gamma^n$  and  $v_n(\delta) = n\delta$ , the limit law is one of Gnedenko's standard limit laws for extremes but that, in other cases, different limit laws can be obtained.

Peter's early papers (including those described above) reflect the style of the time. They are mostly short and very technical; reading them again now, it is not always clear from the papers themselves what the specific motivation for each paper might have been. This style was influenced by the "production process" of that time. The papers were handwritten and then placed in a queue to be typed. The typing was done by typists employed as secretaries (initially two and later three in the Faculties Department of Statistics) in between the other tasks they were required to do. They used IBM Selectric typewriters (known as "golf ball typewriters") which had interchangeable metal balls with different symbols and fonts on them to type the technical parts of the manuscripts. Revision at any stage was difficult. At the handwritten stage, for substantial changes, pages could be rewritten and then replaced; for small changes, pages could be cut up with scissors and then physically pasted with glue or tape and additional material added in between. The typists would correct mistakes they had made (using erasers, correction fluid and physical cutting and pasting) but were never pleased to redo pages if the author simply changed their mind. Peter was very efficient about the process and would usually write a single version of a paper. He thought out the structure and content in his mind before starting and he did not later change this or do much revision. Although Peter dominated the typing queue, access to typists caused tension and Peter could be impatient about having to wait for his papers. The contribution of the women (at that time they were all women) who did the typing was invaluable. (Other important contributions to research were made by Research Assistants who were hired by the Department to assist with finding and photocopying papers from the library and Programmers who wrote code and ran simulations.)

Typed papers were photocopied and then submitted to journals through the postal system. It took time for the necessary letters to be written and for the papers to be sent out for review and then returned. Reviewers saw the task of reviewing differently from how it is seen today; extensive revision was avoided whenever possible and minor issues of expression and style were certainly ignored. Peter was very quick and efficient at revising and resubmitting his papers. He expected editors and reviewers to be quick and efficient too. I have no idea

how Peter obtained his home telephone number, but I recall Peter telephoning Klaus Krickeberg, then editor of the *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* living in Paris, in the very early hours of the morning Paris-time to wake him up and demand a response to a paper he had submitted.

### 3. Teaching

In the early part of his career, while increasing his research productivity, Peter was also fully involved in undergraduate and graduate teaching. Peter did not document his teaching in his *curriculum vitae* and it is now quite difficult to work out exactly which courses he taught and when. Peter's papers in the ANU Archives include lecture notes and other course material which give some information. However, the information is incomplete and the dates on the lecture notes do not necessarily indicate exactly when the lectures were delivered. For example, during Peter's first academic appointment as a lecturer at the University of Melbourne from August 1976 to mid 1978, he taught courses in Stochastic Processes and Nonparametric and Introductory Statistics. It is very likely that he taught other courses too because, in a letter to Chip Heathcote at ANU on 15 December 1977, he wrote that he had the highest teaching load in the department at Melbourne.

On moving to ANU in mid 1978, Peter took up the then standard teaching load of three courses per year. In 1984 Peter voluntarily taught extra courses to ensure that he could take his first planned sabbatical leave in 1985. After this leave, Peter's appointment changed to a joint appointment with the research Department of Statistics in the Institute of Advanced Studies at ANU and he reduced his teaching load. This meant that, as his career advanced, Peter was able to travel more. He continued to teach the course in probability, although later this sometimes meant getting one of his post-doctoral research fellows to deliver the lectures.

When Peter started teaching at ANU, the standard undergraduate courses were labelled A0x, B0x and C0x with the letters denoting first, second and third year courses. In 1982 A01 and A02 split into an Economic Statistics stream and Statistical Techniques I and II; these were effectively new service courses designed to allow a different focus for economics and science students and to accommodate introducing the use of the computer to beginning students. There was an honours stream within the second and third year courses (involving extra material and extra questions in the assessment). There were honours (4th year

undergraduate) and masters courses too; these specialist and advanced courses usually had a specific title and were comparable to graduate courses. Most of the main courses were delivered by a single lecturer, typically for several consecutive years; the specialised and advanced courses were more often taught only once and may have had portions taught by several different people.

The archived teaching materials from Peter's early years at ANU suggest that he taught A02, B01 and C04. Around 1982, he moved from teaching A02 to teaching Statistical Techniques 1 and from teaching B01 to teaching Probability Theory. Statistics A02 was the second semester of the first year introductory statistics course, B01 was the introductory mathematical statistics course (many students started their study of statistics with this course) and C04 was a first course on stochastic processes. Statistical Techniques 1, the first year introductory statistics course included some basic probability, distributions and elementary statistics. Although he was initially a probabilist and his research was clearly in probability, Peter's undergraduate degree in statistics at the University of Sydney included courses in linear models and the design of experiments so he had the background to teach these statistics courses. Probability Theory was offered through the Department of Mathematics to attract mathematics as well as statistics students but it was a required course for statistics honours students. Peter had done a reading course on the 1968 edition of Chung's book "A Course in Probability Theory" at the University of Sydney, liked and always used the 1974 edition (Chung (1974)) for the course.

Peter taught the first year unit (most likely Statistics A02) in second semester 1980. The course had a large enough enrolment that each question in the final exam was marked by a different person. One of the markers, George Mailath, kindly shared the details of what followed. George placed the bags of ungraded exam scripts on the floor on one side, the current packet he was grading on the desk and the bags of graded exam scripts on the floor on the other side. The overnight cleaners judged the bags on the floor to be garbage and removed them for disposal. When this was discovered the following morning, Peter and George went to the Canberra rubbish tip to search for the missing bags but found nothing. Peter and George made a second trip (perhaps with some janitorial staff) and located the area where the ANU trash had been dumped. Peter and George returned to the tip for a third time with every available body (including academic colleagues) to help search for the bags of scripts. Very fortunately, they managed to find all of the bags! It is unclear what would have been done if they had not been found, but it would have been very embarrassing. There



were reminders every year for many years after that to avoid placing exam scripts anywhere where they could be construed as trash (near waste paper baskets or on the floor) - and these always made me think of Peter.

Peter also taught the more advanced courses “Stochastic Models and Processes” (1981), “Order Statistics and Related Topics” (1982), eight lectures on Poisson Processes in a Masters course on linear and nonlinear models, Poisson processes and geostatistics (1983), parts of “Rank tests, Multivariate Analysis and Time Series Analysis” (most likely the rank tests), “Goodness-of-fit tests and Techniques of Inference” (most likely the goodness-of-fit-tests), and “Statistics Special Topics” (1984).

In general, Peter’s lectures, like his seminars, were well-prepared, highly organised and presented on the blackboard. Some lecturers used overhead projectors, but Peter preferred the blackboard. The level in the lectures was uncompromising (so advanced students could learn a great deal) but Peter was generous in assessment and grading, so less well-prepared students could still get through. This was particularly important in Probability Theory where some of the statistics students were mathematically underprepared but could still get through the course and learn useful things from it. Peter’s preferred lecture times were first thing in the morning or at lunch time. My memory is that office hours were chosen to suit Peter far more than the students, but I cannot verify this now. In any case, to run teaching and research well required a very high degree of organisation and self-discipline and I think Peter carefully partitioned the activities so that they did not intrude on each other.

### 3.1. Order statistics and related topics

As a new PhD student, I audited the Honours level (4th year undergraduate) course Peter taught in the first semester of 1982. There were two students enrolled in the course and some Faculty members also auditing. Peter’s course description read as follows:

The aim of the course is to provide an introduction to a broad spectrum of statistical topics, constructed around a single theme. The subject of order statistics was chosen because it permits a relatively simple mathematical discussion, and yet is applicable in a great many areas. Thus, the emphasis is on variety, and we shall consider problems from extreme value theory, nonparametric statistics, empirical process theory, density estimation, U-statistics, location and scale estimation, near neighbour methods, etc.

This was very obviously research-led teaching, well before the phrase was in common use. As Neville Weber commented at the time, an alternative course title would have been "Things I have been working on recently"! This is particularly interesting because it reflects Peter's statistical interests precisely when he was redirecting his research from probability towards statistical theory. There was no text (there could not be), although Peter liked David (1970) and the early lectures were influenced by it. There were two 50 minute lectures per week, on Tuesday and Thursday at 1.00 pm, meaning that the course comprised about 26 lectures.

Peter began by defining the order statistics  $X_{n1} \leq X_{n2} \leq \dots \leq X_{nn}$ , the quantiles  $\xi_p$  as solutions of  $F(\xi_p) = p$ , where  $F$  is the underlying distribution function, and then the sample quantiles as  $\hat{\xi}_p = X_{n, [np]+1}$ , where  $[\cdot]$  is the integer part function. The first result Peter derived used the binomial distribution to construct a nonparametric confidence interval for the  $p$ th quantile  $\xi_p$ . The second derived Rényi's (1953) representation.

*Rényi's Representation: if  $(X_{ni}, i = 1, \dots, n)$  are the order statistics of a random sample with distribution function  $F$ , then the vector  $(X_{n1}, \dots, X_{nn})$  has the same distribution as  $(F^{-1}(\exp(-\sum_{j=1}^{n-i+1} (n-j+1)^{-1} Y_j))), i = 1, \dots, n)$ , where  $Y_j$  are independent standard exponential random variables.*

Peter really liked Rényi's representation and used it throughout the course (as well as in his research, in for example Hall (1982a,b)); it is an elegant result with a simple but clever proof (involving rearranging the sum in the exponent in the joint density of the order statistics from an exponential distribution) which Peter included in the course. Peter did not mention Bahadur's (1966) representation but established the closely related expansion

$$\hat{\xi}_p - \xi_p = -\Delta(p) \frac{p}{f(\xi_p)} + O_p(n^{-1}), \quad (3.1)$$

where  $n^{1/2} \Delta(p) \xrightarrow{D} N(0, (1-p)/p)$ , in proving the central limit theorem for a finite set of sample quantiles.

The proof of the expansion (3.1) from Rényi's representation that Peter gave illustrates his approach and style. Suppose  $F$  has a density  $f$  and  $f$  is a nonzero function with a bounded derivative in a neighbourhood of  $\xi_p$  for  $0 < p < 1$ . Set  $G(x) = F^{-1}(\exp(-x))$  for  $x > 0$ . Then  $F(G(x)) = \exp(-x)$ ,

$$f(G(x))G'(x) = -\exp(-x)$$

and

$$f'(G(x))G'(x)^2 + f(G(x))G''(x) = \exp(-x).$$

Hence  $G'(-\log(p)) = -p/f(\xi_p)$  as  $G(-\log(p)) = \xi_p$ , and  $G''(x)$  exists and is bounded in a neighbourhood of  $-\log(p)$ . Using a Taylor expansion about  $-\log(p)$ , we may deduce

$$G\{-\log(p) + \delta\} = \xi_p - \delta \frac{p}{f(\xi_p)} + O(\delta^2) \quad \text{as } \delta \rightarrow 0. \tag{3.2}$$

Let  $Y_1, \dots, Y_n$  be independent standard exponential random variables and recall we may write

$$\hat{\xi}_p = G \left\{ \sum_{j=1}^{n-[np]} (n-j+1)^{-1} Y_j \right\}. \tag{3.3}$$

Now  $\sum_{j=1}^{n-[np]} (n-j+1)^{-1} Y_j = \sum_{j=1}^{n-[np]} (n-j+1)^{-1} + \sum_{j=1}^{n-[np]} (n-j+1)^{-1} (Y_j - 1)$ ,

$$\sum_{j=1}^{n-[np]} (n-j+1)^{-1} = \sum_{j=[np]+1}^n i^{-1} = \int_{np}^n x^{-1} dx + O(n^{-1}) = -\log(p) + O(n^{-1}).$$

Furthermore  $\Delta(p) = \sum_{j=1}^{n-[np]} (n-j+1)^{-1} (Y_j - 1)$  is a sum of independent random variables with mean zero and so has variance

$$\begin{aligned} \text{Var}\{\Delta(p)\} &= \sum_{j=1}^{n-[np]} (n-j+1)^{-2} = \sum_{j=[np]+1}^n i^{-2} \\ &= \int_{np}^n x^{-2} dx + O(n^{-2}) = \frac{1-p}{np} + O(n^{-2}). \end{aligned}$$

By application of the Central Limit Theorem, we see that

$$n^{1/2} \Delta(p) \xrightarrow{D} N \left( 0, \frac{1-p}{p} \right).$$

Combining these results, we see that

$$\sum_{j=1}^{n-[np]} (n-j+1)^{-1} Y_j = -\log(p) + \Delta(p) + O(n^{-1}),$$

where  $\Delta = n^{-1/2}Z$ , say, where  $Z$  is normally distributed. We shall combine this result with (3.2) and (3.3), taking  $\delta = \Delta(p) + O(n^{-1})$  in (3.2). We obtain

$$\begin{aligned} \hat{\xi}_p &= \xi_p - \{\Delta(p) + O(n^{-1})\} \frac{p}{f(\xi_p)} + O(\{\Delta(p) + O(n^{-1})\}^2) \\ &= \xi_p - \Delta(p) \frac{p}{f(\xi_p)} + O(n^{-1}). \end{aligned}$$

Peter concluded with the remark that the central limit theorem for the quantiles could be proved for  $f$  continuous and not necessarily bounded.

The proof is set out like a proof in a research paper - students have to work

to understand the argument and the steps rather than having everything laid out for them. It is clear from the argument leading to (3.2) that (3.2) is relevant but it is not yet necessarily clear how; it helps to read the proof backwards but this can only be done after the entire proof has been presented. The second half of the proof shows some of Peter's technical facility in working with the argument of  $G$  in (3.3). Finally, Peter brought everything together and completed the proof by noting that the conditions can be weakened.

After proving the central limit theorem, Peter derived the asymptotic distribution of the spacing between two quantiles and showed how to use the spacing to construct a consistent estimator of the asymptotic variance of a sample quantile. He then lectured on Grenander's mode estimator and kernel density estimation.

Peter treated extreme order statistics very classically by proving the result that there are only three possible limiting distributions and discussing weak convergence in various cases. He also proved a strong law for extremes.

The next topic was linear combinations of order statistics or L-estimators. Peter approached this through the empirical process and applied the general result to Gini's mean difference statistic. He then considered Gini's mean difference as a U-statistic and used this to motivate U-statistics. He derived the central limit theorem for order 2 U-statistics and applied this to Gini's mean difference and the Wilcoxon signed rank statistic. He also discussed 2-sample U-statistics, considering the Wilcoxon rank sum statistic, and degenerate U-statistics.

Peter completed the course by again considering density estimation. He showed that the near neighbour density estimator can be written as a kernel density estimator and then applied results derived earlier in the course to near neighbour estimators. (Peter used the phrase "near neighbour" rather than "nearest neighbour" in the course.) He also considered near neighbour density estimation in the discrete case, focussing on the binary sample space in  $p$  dimensions,  $B = \{0, 1\}^p$ . Finally, Peter presented orthogonal series density estimators in univariate and multivariate problems with both continuous and discrete distributions. He discussed the idea of adding terms in the series to reduce the asymptotic mean integrated/summed squared error. The topic of choosing tuning parameters was to be very important in Peter's research on nonparametric estimation.

### 3.2. Poisson processes

Peter gave eight lectures in another advanced course in second semester 1983. Geoff Lee kindly lent me a copy of the notes he took from the course. This course

is interesting because of its connection to Peter's subsequent research on coverage processes; the course was very statistical in orientation, but it does show that Peter was thinking about mosaic processes at the time.

In the first two lectures, Peter defined the Poisson process and the inhomogeneous Poisson process, and established a relationship between them. He then gave three lectures on inference for Poisson processes. He derived the likelihood and discussed maximum likelihood estimation of the parameters. He also considered a single exponentially decaying Poisson process with intensity  $\lambda(t)a \exp(-\alpha t)$  and introduced the problem of testing whether  $\alpha = 0$  against the alternative that  $\alpha > 0$ . Peter concluded his lectures in the course by giving three lectures on mosaic processes. He introduced these processes by considering spheres of random radius centred on the points of a Poisson process and then reduced this to a 1-dimensional pattern of segments and spaces on a line. He discussed maximum likelihood estimation of the Poisson process parameter under the assumption of constant underlying segment length. He referenced Roach (1968), indicating that Roach had a different derivation and some typographical errors in the solution.

#### 4. Estimating the Exponent of Regular Variation

Peter first papers using extremes in statistical problems appeared in 1982. The first of these (Hall (1982a)) concerned the problem of estimating the exponent of regular variation, a key parameter describing the tails of distributions with regularly varying tails. In most of Peter's research on regular variation, the context is that we observe iid random variables with a distribution function  $F$  satisfying

$$1 - F(x) = Cx^{-\alpha}(1 + \delta(x)), \text{ as } x \rightarrow \infty, \quad (4.1)$$

where  $\alpha > 0$  is the unknown exponent of regular variation,  $C > 0$  is an unknown scale parameter and  $\delta(x) \rightarrow 0$  as  $x \rightarrow \infty$ . In particular, Hall (1982a) considered the cases  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$  with  $\beta > 0$ ,  $D$  a real number, and  $\delta(x) = D_1x^{-\alpha} + \dots + D_mx^{-m\alpha} + o(x^{-m\alpha})$  which holds, for example, for the extreme value distribution with  $F(x) = \exp(-x^{-\alpha})$ . The distribution (4.1) with  $\delta(x) = 0$  represents a distribution with Pareto tails. An interesting point made in Feuerverger and Hall (1998) is that the model with  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$ , becomes  $1 - F(x) = Cx^{-\alpha} + CDx^{-\alpha-\beta} + o(x^{-\alpha-\beta})$ , which can be interpreted as a mixture of Pareto models. The departure from an exact Pareto model is thus represented by an approximate mixture of Pareto models.

Under the assumption that  $\delta(x) = 0$  for  $x > d$ , it follows from Rényi's repre-

sensation that, given  $X_{n,n-r} > d$ , the scaled log-spacings  $U_i = i\{\log(X_{n,n-i+1}) - \log(X_{n,n-i})\}$  are independent exponential random variables with mean  $1/\alpha$ . In this case, Hill (1975) derived conditional maximum likelihood estimators of  $\alpha$  and  $C$ , given  $X_{n,n-r} > d$ , which are

$$\frac{1}{\hat{\alpha}_r} = r^{-1} \sum_{i=1}^r U_i = r^{-1} \sum_{i=1}^r \log(X_{n,n-i+1}) - \log(X_{n,n-r})$$

and  $\hat{C}_r = r(X_{n,n-r})^{\hat{\alpha}_r}/n$ .

By making the transformation  $x \rightarrow x^{-1}$ , the problem can be reformulated with the tail at the origin and the estimators mapped to using the smallest (on the transformed scale) rather than the largest order statistics. This transformation was used in Hill (1975), Hall (1982a), and in Peter's later papers, particularly in the proofs, to simplify notation, but the whole problem was formulated on the transformed scale in Feuerverger and Hall (1998) and Guillou and Hall (1999).

Hall (1982a) explored the properties of Hill's estimator under two forms for  $\delta(x)$ . For  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$  with  $D \neq 0$ , Hall (1982a) showed that is optimal to choose the threshold  $r = r(n)$  to increase at the rate  $O(n^\kappa)$ , where  $\kappa = 2\beta/(2\beta + \alpha)$ : if  $r$  increases more slowly, the variance of the estimator is of a larger order than the bias of the estimator; if  $r$  increases more quickly, the bias of the estimator is of a larger order of magnitude than the variance of the estimator. When  $\delta(x) = D_1x^{-\alpha} + \dots + D_mx^{-m\alpha} + o(x^{-m\alpha})$ , Hall (1982a) showed that an estimator with a faster rate of convergence ( $O(n^{m/(2m+1)})$  compared to  $O(n^{\kappa/2})$ ) can be obtained by combining several of Hill's estimators  $\hat{\alpha}_r$  computed with different thresholds  $r$ .

Hall and Welsh (1984) used the approach developed by Farrell (1972) for kernel density estimators to establish minimax rates of convergence for Hill's estimators. Under  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$  with  $D \neq 0$ , they showed that the minimax optimal rate of convergence for estimating  $\alpha$  is  $O(n^{\kappa/2})$ , where  $\kappa = 2\beta/(2\beta + \alpha)$ . Hall and Welsh (1985a) assumed that the optimal threshold  $r \sim hn^\kappa$ , with  $\kappa$  known, and proved invariance principles for  $A_n^*(h) = n^{\kappa/2}(\hat{\alpha}_{[hn^\kappa]} - \alpha)$  and  $C_n^*(h) = (\log n)^{-1}n^{\kappa/2}(\hat{C}_{[hn^\kappa]} - C)$ . The motivation for treating  $\kappa$  as known is that under  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$  with  $\beta = \alpha$  (as in the second form for  $\delta(x)$  used in Hall (1982a)),  $\kappa = 2/3$ . They then constructed a direct estimator of the value of  $h$  that minimises the asymptotic mean squared error of Hill's estimator. The direct estimator of the optimal  $h$  required three tuning parameters so was not particularly practical. However, it was a first improvement on Hill's (1975) suggestion of choosing  $r$  by increasing  $r$  until a test of the hypothesis that the

scaled log-spacings have an exponential distribution is rejected; Hall and Welsh (1985a) noted that this procedure overestimates the threshold  $r$  used to construct Hill's estimator so it is not consistent.

In Hall (1990), a general paper on using the bootstrap to estimate the mean squared error of an estimator to give us a criterion to minimise to choose tuning parameters, one of the examples discussed in detail is choosing the threshold  $r$  in Hill's estimator. Hall (1990) suggested using the  $m$  out of  $n$  bootstrap to estimate the mean squared error of Hill's estimator of  $\alpha$  as a function of  $r$ , minimising this estimate to obtain  $\hat{r}_m$  and then rescaling  $\hat{r}_m$  from sample size  $m$  back to the original sample size  $n$ . This last step is difficult in general. To make it feasible, Hall (1990) assumed, as in Hall and Welsh (1985a), that the optimal  $r \sim hn^\kappa$ , with  $\kappa$  known, so that we only need to estimate  $h$  and, more importantly, rescaling is achieved straightforwardly by multiplying  $\hat{r}_m$  by  $(n/m)^\kappa$ .

Hall and Weissman (1997) also considered the problem of choosing  $r$  in Hill's estimator as particular case in a general problem. The problem was that of estimating the probability of exceeding a given value  $x$  or of estimating a quantile that is beyond the observed data; following the paper, we focus on the first problem. Given an estimator  $\hat{\theta}(r)$  of  $\theta$ , where  $r$  is a tuning parameter, and a model  $F_\theta(x)$  such that  $\bar{F}(x)/\bar{F}_\theta(x) \rightarrow 1$  as  $x \rightarrow \infty$ , where  $\bar{F}(x) = 1 - F(x)$ , the general problem is to choose  $r$  to optimise  $E\{\bar{F}_{\hat{\theta}(r)}(x) - \bar{F}(x)\}^2$ . The value  $x$  is taken to be of larger order than  $O(n^{1/(2\beta+\alpha)+\epsilon})$  so that it is beyond the range of the data (the large order statistics under the model with  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$  are of order  $O(n^{1/\alpha})$  in probability). The approach used in Hall and Weissman (1997) is to estimate the mean squared error at  $(y, m)$ , where the value  $y \ll x$  and the sample of size  $m < n$ , rather than at  $(x_0, n)$ , to minimise the estimated mean squared error over  $r$  to obtain  $\hat{r}(y, m)$  and then back transform  $\hat{r}(y, m)$  to be an estimator on  $(x_0, n)$ . Here  $y$  is required to be within the range of the data, so effectively the problem is transformed to a less extreme version in such a way that  $\log(y)/\log(m) = \log(x)/\log(n)$ . This is achieved by setting  $y = O(n^{1/(2\beta+\alpha)-\epsilon_1})$  and  $m = O(n^{1-\epsilon_2})$  for  $\epsilon_1, \epsilon_2 > 0$ . Hall and Weissman (1997) used the  $m$  out of  $n$  bootstrap to estimate the mean squared error and concentrated on the Pareto type model (4.1) with  $\delta(x) = 0$ . They suggested estimating  $\hat{r}(y, m)$  for two values of  $m$ , regressing these on  $\log(m)$  to estimate the intercept  $\log(C)$  and the slope  $\gamma$  and then setting  $\hat{r}(x, n) = \hat{C}n^{\hat{\gamma}}$ . This means that the bootstrap requires two tuning parameters (compared to three in Hall and Welsh (1985a)).

Guillou and Hall (1999) proposed a different method for choosing the threshold  $r$  for Hill (1975) estimator that does not require specifying a particular

form for  $\delta(x)$  in (4.1). The method is based on the scaled log spacings  $U_i = i\{\log(X_{n,n-i+1}) - \log(X_{n,n-i})\}$  which we have noted are independent exponentially distributed random variables with mean  $1/\alpha$  if the model (4.1) holds with  $\delta(x) = 0$ , and approximately independent exponential random variables when  $\delta(x) \neq 0$ . Guillou and Hall (1999) showed that the order of the approximation to the asymptotic distribution of the estimator is determined by the bias of the estimator, so their idea was to choose  $r$  as large as possible subject to a finite bound on the bias of the estimator. They achieve this by finding the smallest integer  $\hat{r}$  such that  $|T_t| > c_{crit}$  for all  $t \geq \hat{r}$ , where  $c_{crit}$  is a fixed positive value (suggested from empirical experience to be between 1.25 and 1.5) and  $T_t$  is a symmetrically weighted sum of the  $U_i$  that is standardised under the assumption that the  $U_i$  are exactly exponentially distributed. The idea is to increase  $r$  until the bias has a significant effect on the approximation to the sampling distribution of Hill's estimator. As Guillou and Hall (1999) noted, this is similar in spirit to Hill's (1975) suggestion of testing the  $U_i$ 's for exponentiality but it treats the scaled log-spacings cumulatively rather than separately.

Most of Peter's research in regular variation was concerned with Hill's estimators. Peter did mention (in some of the papers described above) that the results for Hill's estimators could be extended to the kernel estimators proposed by Csörgő, Deheuvels and Mason (1985) as alternatives to Hill's estimators, but he did not publish any detailed work on these estimators. However, Feuerverger and Hall (1998) proposed and studied their own alternative estimators of the parameters  $\alpha$  and  $C$  in the model (4.1) with  $\delta(x) = Dx^{-\beta} + o(x^{-\beta})$ . Feuerverger and Hall (1998) considered the scaled log-spacings  $U_i = i\{\log(X_{n,n-i+1}) - \log(X_{n,n-i})\}$ . They suggested treating  $r$  of these as exponential with mean  $\alpha^{-1} \exp(D_1(i/n)^{\beta_1})$ , where  $D_1 = -\beta_1 C^{-\beta_1} D$  and  $\beta_1 = \beta/\alpha$ , and suggested using either maximum likelihood or regression estimators to estimate the unknown parameters. The regression estimator is obtained by minimising

$$\sum_{i=1}^r \left\{ \log(U_i) - \mu - D_1 \left( \frac{1}{n} \right)^{\beta_1} \right\}^2,$$

where the intercept  $\mu = -\log(\alpha) - \gamma$  and  $\gamma$  is Euler's constant. The suggested estimator of  $\alpha$  is either

$$\hat{\alpha} = \exp(-\gamma - \hat{\mu}),$$

which follows from the regression formulation, or



$$\frac{1}{\hat{\alpha}} = r^{-1} \sum_{i=1}^r U_i \exp \left( -\hat{D}_1 \left( \frac{i}{n} \right)^{\hat{\beta}_1} \right),$$

which follows from expressing the maximum likelihood estimator of  $\alpha$  as a function of the other unknown parameters and then replacing them by their estimators. The problem is much simpler if  $\beta_1$  is known (often  $\beta_1 = 1$ ) or a known function of  $\alpha$  (often  $\beta_1 = 1/\alpha$ ). Feuerverger and Hall (1998) suggested pilot estimates of the parameters (including  $\beta_1$ ) that can be computed to distinguish these cases.

Feuerverger and Hall (1998) considered the different problem of estimating tail parameters from record values. Record values are data recorded only if they exceed all previous values; simple examples are given by records in jumping and throwing events in field sports. Feuerverger and Hall (1998) adopted the slightly different model

$$F(x) = 1 - x^{-\alpha}K(x), \quad \text{where } K(x) = C(\log x)^\beta + o\{(\log x)^\beta\},$$

with  $0 \leq \beta < 1$ , as  $x \rightarrow \infty$ . If the record values are  $Y_1, \dots, Y_N$ , then the role of the scaled log spacings in the earlier analysis is here played by the log-spacings  $\{V_i\}$ , where  $V_1 = \log(Y_1) - \alpha^{-1} \log(C)$  and  $V_i = \log(Y_{i+1}) - \log(Y_i)$ ,  $i = 1, \dots, N - 1$ , which, when  $K(x) = C$ , are independent exponential random variables with mean  $1/\alpha$ . The maximum likelihood estimator of  $\alpha$  now satisfies

$$\begin{aligned} \frac{1}{\hat{\alpha}} &= N^{-1} \left[ \log(Y_1) - \alpha^{-1} \log(C) + \sum_{i=1}^{N-1} \{\log(Y_{i+1}) - \log(Y_i)\} \right] \\ &= N^{-1} \log(Y_N) - (N\alpha)^{-1} \log(C). \end{aligned}$$

The scale parameter  $C$  cannot be estimated consistently from the record values alone but the first term  $N^{-1} \log(Y_N) = 1/\tilde{\alpha}_0$ , say, is of larger order than the second term so  $\tilde{\alpha}$  is asymptotically equivalent to  $\tilde{\alpha}_0$ . In fact,  $\tilde{\alpha}_0$  is the best performing estimator of  $\alpha$  in Berred (1992); it is interesting that it is similar to Hill's estimator but does not involve a threshold parameter. Berred (1992) estimated the exponent of regular variation using record values but did not use the inter-record times. Feuerverger and Hall (1998) also considered including inter-record times. They showed that including inter-record times improves the rate of convergence of the estimator (by two orders of magnitude), but commented that the constant in the convergence result is much larger than that for  $\tilde{\alpha}_0$ , and so in small samples may outweigh the gains in rate of convergence. An interesting technical aspect of the problem is that, even under the simple models with  $K(x) = C$ , the maximum likelihood estimators are not regular and the

information matrix is not well-defined. This is treated by using the relationship between maximum likelihood and maximum probability estimators, a technique Peter had used earlier in Hall (1982b). Also, as noted in the abstract to the paper, if  $F$  is continuous then the joint distribution of any number of record times does not depend on  $F$  so there is no information about  $F$  in the record times. However, the record times and record values jointly contain more information about  $F$  than the record values alone.

## 5. Estimating Boundaries

The second of Peter's papers in 1982 to use extremes in statistical problems tackled the problem of estimating the endpoint of a distribution (Hall (1982b)). There are strong parallels between the two (1982) papers in terms of how Peter set up the problems and the approach he took to solving them. In Hall (1982b), Peter assumed that we observe iid random variables with a distribution function  $F$  satisfying

$$1 - F_\theta(x) = C(\theta - x)^{k+1}\{1 + \delta(\theta - x)\}, \text{ as } x \rightarrow \theta, \quad (5.1)$$

where  $C \geq 0$ ,  $k \geq 0$  and  $\theta$  is a location parameter representing the endpoint of the support of  $F$ , so  $F_\theta(x) = 1$  for  $x > \theta$ . For  $\delta(x) = 0$ , results of Polfeldt (1970a,b) and Woodroffe (1974) show that the maximum likelihood estimator of  $\theta$  based on the whole sample converges at the same rate as  $\hat{\theta} = X_{nn}$  when  $0 < k < 1$ . This means that, when  $0 < k < 1$ , estimators of  $\theta$  based on a fixed, finite number of order statistics cannot be improved on. This is not true when  $k \geq 1$ , the case that mainly interested Peter.

In Hall (1982b), Peter assumed that  $\delta(\theta - x) = 0$  for  $\theta - \epsilon < x < \theta$  with  $\epsilon > 0$ . Then, conditional on  $X_{n,n-r+1} > \theta - \epsilon$ , when  $k$  is known, Peter showed that the maximum likelihood estimator of  $\theta$  satisfies

$$\frac{k}{n} = \sum_{j=1}^{r-1} \frac{X_{n,n-j+1} - X_{n,n-r+1}}{\theta - X_{n,n-j+1}}, \quad \theta > X_{nn}$$

and  $\hat{C} = r/n(\hat{\theta} - X_{n,n-r+1})^{k+1}$ . He then studied the asymptotic distribution of the estimators under the more general model (5.1) with  $\delta(x) = O(x^\ell)$ ,  $\ell > 0$ . Let  $m = \min\{1, \ell/(k+1)\}$ . If  $k > 1$ ,  $r \rightarrow \infty$  and  $r/n^{2m/(2m+1)} \rightarrow 0$  as  $n \rightarrow \infty$ , Hall (1982b) showed that  $n^{1/(k+1)}r^{(k-1)/2(k+1)}(\hat{\theta} - \theta)$  is asymptotically normal. He also showed that if  $k = 1$  and  $r = O(n^{2m/(2m+1)})$ , then  $\{n \log(r)\}^{1/2}(\hat{\theta} - \theta)$  is asymptotically normal. Peter pointed out that there are situations in which  $2m/(2m+1) = 2/3$  or  $2/(k+3)$  so the conditions on  $r$  do not require detailed

knowledge of  $\ell$ . He noted that if we took  $\delta(x) = Dx^\ell + o(x^\ell)$ , we could derive asymptotic mean squared errors, but he did not present these results because they depend on  $C$  and  $D$  so are “of little practical value”. Finally, Peter considered the case with  $k > 1$  but otherwise unknown. The conditional maximum likelihood estimators are more complicated than when  $k$  is known but Peter showed that the estimator of  $\theta$  computed with  $k$  unknown has similar properties to that computed with  $k$  known, except that the variance is reduced by  $k^{-2}$ . The similarity in the approach to Hall (1982a) is striking.

One of the interesting features of the problem of estimating the endpoint for Peter is that it is irregular. Obviously, the endpoint of the distribution is an unknown parameter. If  $k > 2$ , we can differentiate the integral of the likelihood twice under the integral sign. For  $1 < k < 2$ , we can only differentiate once under the integral sign so the Fisher information has to be defined in terms of first rather than second derivatives. For  $k = 1$ , we cannot differentiate under the integral so Peter linked the estimator to the maximum probability estimator and showed that the maximum likelihood estimators of  $\theta$  when  $C$  is known and  $C$  is unknown have the same asymptotic distributions. i.e. information about  $C$  does not help in estimating  $\theta$ .

Peter returned to the problem of estimating the endpoint when the shape parameter is small in Hall and Wang (2005). They formulated the problem in terms of the lower endpoint, assumed that the observations have density function

$$f(x) \sim (x - \theta)^{\alpha-1}g(x; \theta, \omega), \quad \theta < x,$$

where  $\theta$  is the location parameter representing the endpoint of the distribution,  $\alpha$  is the shape parameter,  $\omega$  contains parameters other than  $\theta$  and  $g$  converges to a strictly positive constant as  $x \downarrow \theta$ , and studied the asymptotic properties of penalized likelihood estimators. Hall and Wang (2005) introduced the penalty as coming from the prior distribution with density  $p(\theta) = (X_{n1} - \theta)/(X_{n2} - \theta)$ . They interpreted this prior as “informative and empirical” near the smallest order statistic  $X_{n1}$  but uninformative elsewhere. Hall and Wang (2005) obtained asymptotic distributions for the penalized likelihood estimators of  $\theta$  in the cases  $0 < \alpha < 1$ ,  $1 \leq \alpha < 2$  and  $\alpha \geq 2$ . They showed, using results from Hall (1978a,b), that the penalized likelihood estimator achieves the optimal rate of convergence,  $O_p(n^{-1/\alpha})$ , the asymptotic distributions are normal for  $\alpha \geq 2$  and much more complicated for  $\alpha < 2$ . These results mean that we cannot use normal approximations or the conventional nonparametric bootstrap for inference about  $\theta$ ; we can use an m out of n bootstrap or a parametric bootstrap, which Hall

and Wang preferred because it avoids having to choose  $m$ . They noted that the intervals become more anti-conservative as  $\alpha$  increases and recommended using iterated bootstrap calibration to adjust the coverage of the bootstrap intervals.

Frontiers or boundaries generalise the concept of the endpoint of a distribution to more than one dimension and, not surprisingly, Peter was interested in this generalization of endpoint estimation. Frontier estimation is a key part of productivity analysis of firms. In the context of frontier analysis using longitudinal or panel data, Härdle, Hall and Simar (1995) considered a simple fixed effect regression model with different intercepts and different variances for each firm in which the problem of interest is to estimate the maximum intercept. They used weighted least squares estimates of the regression parameters and the residual bootstrap with the percentile method to set a confidence interval. They showed that consistency requires distinct intercepts; if the second largest intercept is close to the largest, then the coverage of the confidence interval can be poor. Härdle, Hall and Simar (1995) used the iterated bootstrap to estimate the true coverage and then found the nominal coverage which gives the desired coverage. They applied the method to railway data from different countries in which the response is number of employees per kilometre of network and there are various explanatory variables. This of course provided an opportunity for Peter to bring together his two great interests, trains and statistics. Hall, Park and Stern (1998) considered mostly parametric polynomial frontier estimators but also nonparametric local polynomial estimators in the two-dimensional case under the assumption that the data form a Poisson process inside the frontier. A key part of the approach is obtaining approximations to the asymptotic distributions of the estimators and then simulating from them to produce Monte Carlo estimates of the bias which can be used to adjust for bias. The asymptotic approximations involve joint extreme value distributions. Hall, Park and Stern (1998) also addressed the problem of constructing simultaneous confidence intervals for the frontier.

Hall and Simar (2002) considered the problem of boundary estimation when the data are subject to measurement error. In particular, they considered the problem of estimating the boundary of a distribution from independent data observed with an additive error which has variance  $\sigma^2$ . This paper shows how to reduce the  $L_1$  error from  $O(\sigma)$  as  $\sigma \downarrow 0$  to  $O(\sigma^2)$  or smaller. The main complication in the problem is identifiability as the error distribution is usually assumed to be symmetric, whereas in this context it is most often asymmetric. Hall and Simar (2002) dealt with this issue by first solving the problem where the density function is flat approaching its endpoint and the error density is unimodal

(with known mode at zero) with support extending beyond the endpoint. The known unique mode together with the assumed flatness ensures identifiability. They then treated the flat density assumption as approximate as  $\sigma \downarrow 0$ . The endpoint parameter is the argument that maximises the derivative of the density of the observed data and this leads to a natural kernel estimator. Hall and Simar (2002) considered both univariate and bivariate cases, explicitly bringing endpoint and frontier estimation together. It is striking that Hall and Simar used the approach of initially assuming a simple model to hold in order to construct an estimator and then studying its properties in the more general case that the model holds only approximately in a limit; this approach runs through much of Peter's work on extremes. This paper is also a key paper in Peter's research on measurement error and deconvolution. Carroll and Hall (1988) established minimax rates of convergence for nonparametric estimators of the density of interest in the additive measurement error model. According to Delaigle (2016), the very slow (logarithmic) rate of convergence when the underlying density is normal, discouraged Peter from working on deconvolution. However, the small variance ( $\sigma \downarrow 0$ ) asymptotics applied in Hall and Simar (2002) lead to faster (algebraic) rates of convergence and this result re-inspired Peter's interest in deconvolution problems.

## 6. PhD Supervision

During the period in the early 1980s when Peter was redirecting his research towards statistical theory, dramatically increasing his productivity and carrying a full teaching load, Peter also began to supervise PhD students. His first involvement with PhD supervision was with Brenton Clarke who completed in 1980. Brenton was a PhD student of Chip Heathcote but Chip's health deteriorated (he ended up having triple bypass surgery in 1981) and Peter stepped in to help while Brenton was writing up. I was next and became the first PhD student Peter listed on his *curriculum vitae*.

I first met Peter when I came to ANU in late 1981 to meet potential PhD supervisors and discuss potential projects. The PhD in Australia was (and largely still is) a pure research degree with no coursework so students needed to decide who they wanted to work with and on what topic either before starting or very soon thereafter. I thought Chip Heathcote would be a good supervisor so I applied to the Faculty Department of Statistics and they brought me to Canberra to meet the faculty members. I had seen Peter present a research seminar at the

University of Sydney earlier in the year but this was our first meeting. Peter was tall, had old-fashioned sideburns (shaved off in 1988), and at times a slight stutter (that he later overcame). My memory is that Peter showed me seven or eight preprints and reprints of his work, at least half of which concerned the law of the iterated logarithm for various density estimators. The list probably included Hall (1981a), Peter's first paper in the *Annals of Statistics*, Hall (1981b) and the two statistical papers on extremes Hall (1982a,b) that we have already discussed. Hall (1981b) is the only paper on the law of the iterated logarithm for density estimators that I can find in Peter's *curriculum vitae* that could have been on the list (showing the unreliability of memory) but the impression remains strong and is in some sense correct: I did not realise at that time that Peter was still very much a probabilist just starting to redirect his research to statistical theory. I did however find Peter's papers very mathematical and intimidating.

The ANU was just starting a system in which each PhD student had a supervision panel consisting of a supervisor and 2 advisors: Chip was my supervisor and Peter and Des Nicholls my advisors. I began working with Chip on empirical characteristic functions, extending work that had been done by Ray Chambers (during an MSc) and published in Chambers and Heathcote (1981). Chip was still recovering from his surgery so would come in to work in the morning and then go home to rest before lunch; in retrospect, what he did was amazing. One of the key ideas that Chip and Ray had had was that one could choose the argument of the empirical characteristic function to minimise the asymptotic variance of the estimator and that this minimising value should be non-zero when the error distribution is non-normal. Investigating this and then modifying the criterion a little led to a test for normality and this was the first paper out of my thesis work (Hall and Welsh (1983)). It was written with Peter rather than Chip which I think reflects the fact that I had some technical questions which Peter helped with and perhaps that Chip was not around so much at the time. I did manage to write two papers with Chip. I was at a bit of a loose end while Chip was going through the second manuscript, mainly I think because neither he nor I had an immediate idea of where to take the work next. At this point in 1983, Peter suggested I work with him on regular variation. I must have gained in confidence, perhaps from sitting in on his course on order statistics, from working with him, and from getting to know him because I agreed. Effectively, Peter took over directing me at that point, although Chip remained formally my supervisor.

Chip had given me a copy of Chambers and Heathcote and a preprint of Csörgő (1983) and suggested extending the paper; Peter gave me a copy of Hall

(1982a) and Farrell (1972) and suggested that I try to apply Farrell's argument to establish minimax rates of convergence for Hill's estimator. This was interesting and challenging; the result was published in Hall and Welsh (1983). I was interested in adaptive estimation (a key part of the empirical characteristic function work when the value of the empirical characteristic function argument is estimated) so I suggested we do this for the Hill estimator to allow for an estimated tuning parameter. This led to Hall and Welsh (1985a). The final paper at this time that I wrote with Peter (Hall and Welsh (1985b)) arose out of my robustness interests; we proved a central limit theorem for the median absolute deviation from the median. The proof used my favourite Bahadur's representation rather than Peter's favourite Rényi's representation and this later turned out to be useful for extending this and other results to regression problems. This was an extremely productive, stimulating time and I learned a great deal from Peter.

Peter's supervision style was to treat me as a (not very) junior colleague. At the start of my PhD he came into my office and said I should subscribe to *Biometrika* and join the Institute for Mathematical Statistics, the American Statistical Association and the Royal Statistical Society as the student rate was low and I would obtain copies of their journals. (I was already a member of the Statistical Society of Australia.) I dutifully did so and felt proud to join the wider statistical community. Peter was available for questions, turned work around quickly and expected me to get on with the work. He would help with specific questions but he did not do the work for me (or his other students) and he was open to my suggestions for problems to work on. Peter would also drop by to discuss other aspects of his work. Mostly Peter would write out his problem, solve it himself and then depart, thanking me for having been very helpful! Peter was keen to write papers as soon as we had enough material and encouraged me to do so too. I think this was both excellent advice and excellent training. One day Peter came into my office and said I had done enough work to write up a thesis and should now do so. This meant writing it out in long hand which took some time. Peter continued to provide support through the rest of my career and I am very grateful to him for this.

During my time as a PhD student, Phil Kokic completed a research masters with Peter on geometric probability while working at the Australian Bureau of Statistics, and Julian Wightwick came out from England to do a PhD on rates of convergence in the central limit theorem. The next students were Matt Wand and Michael Martin. I think Peter's supervision style suited us all very well, though

I think some later students found Peter's high expectations too demanding.

I married at the start of my PhD and my wife and I became friends with Peter and his wife, Jeannie. They entertained us at their home and we entertained them in our student flat. Initially, before I bought a car, going to their home required Peter to come and pick us up and then drop us back afterwards. As Peter and Jeannie lived at the then southern extremity of Canberra, this was very kind and generous. I discovered that Peter had a very sweet tooth (a gift of coconut ice one day may not have even made it home) and that he liked the soft spreadable walnut cheese that had two bands of walnuts in it that he bought from The Contented Cow, a cheese shop in Belconnen Shopping Centre.

## 7. Discussion

We have used the theme of extremes in Peter Hall's early research, teaching and PhD supervision to illustrate the period in his career when he was establishing himself before becoming famous. The recognition and rewards for his outstanding productivity and wide-ranging contributions were slow to come; neither the University of Melbourne nor ANU seemed at the time to appreciate the phenomenon that they had hired. Undoubtedly, this experience left its mark on Peter. But he did succeed and he continued to succeed spectacularly well throughout his career.

The focus on Peter's research has been on *early* research on extremes so we have not even covered all of Peter's research on extremes. Peter also published three papers with Nader Tajvidi on modelling bivariate extremes and estimating trends in extremes, in the early 2000s. The marginal distributions of bivariate extremes must be one of the three classical types of extreme value distribution but the link between them, the dependence function, can be quite general. Hall and Tajvidi (2000a) proposed nonparametric estimators of the dependence function in the bivariate case. One of their estimators modified an earlier proposal of Pickands (1981); they showed that this estimator and its derivative are both root-n consistent. Their other estimator used constrained smoothing splines with the tuning parameter selected by cross validation. Hall and Tajvidi (2004) used these methods to construct compact and semi-infinite bivariate as well as conditional prediction regions for extremes. They used bootstrap calibration to ensure correct coverage of the prediction regions. They compared their prediction regions with the prediction regions obtained from six parametric models for the dependence function to illustrate the advantages of flexibility and simplicity.



Hall and Tajvidi (2000b) worked on estimating temporal trends when fitting parametric models to extreme values from a weakly dependent time series. They developed a nonparametric local-likelihood approach with temporal cross-validation for selecting window width. They applied the method to Swedish wind-storm data and Victorian temperature data using models with approximately Pareto, generalized-Pareto, extreme-value or normal marginal distributions and their own time-varying probability plots to assess goodness of fit. One of the interesting results from the paper relates to the use of normal distributions: in cases where both location and scale are estimated together, the normal distribution was shown to have special features that permit it to play what Hall and Tajvidi described as a universal role as a “nominal” model for the marginal distribution. By this, they meant that fitting the normal distribution avoids inconsistency if the model is incorrect and avoids having to compute an initial undersmoothed mean to compute the variance.

These papers do not fit closely with Peter’s work on regular variation or estimating the endpoint of a distribution but they do fit well with his other research on nonparametric smoothing and methods for tuning parameter selection. As mentioned in the introduction, Peter’s research on extremes links in well with his research in other areas. These and Peter’s other later papers on extremes are also very different from his early papers; they are clearly in statistics, multi-authored, longer, and much more attention is paid to motivation. The changes in the requirements for publication and the way Peter embraced them are clear. One thing that did not change however, was the drive, exceptional work-ethic and outstanding productivity that characterised his whole career.

## Acknowledgment

I am grateful to Jeannie Hall for many helpful conversations. In addition, I am grateful for comments and insights into Peter’s early career provided by a number of colleagues, including Brenton Clarke, William Dunsmuir, George Mailath, Michael Martin, Terry O’Neill, John Robinson, Mervyn Silvapulle, Terry Speed, Neil Trudinger, and Sue Wilson. Finally, I am grateful to an associate editor for a number of suggestions for improvement.

## References

- Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math, Statist.* **37**, 577–580.  
Berred, M. (1992). On record values and the exponent of a distribution with regularly varying

- upper tail. *J. Appl. Probab.* **29**, 575–586.
- Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.
- Chambers, R. L. and Heathcote, C. R. (1981). On the estimation of slope and the identification of outliers in linear regression. *Biometrika* **68**, 21–33.
- Chung, K. L. (1974). *A Course in Probability Theory*. Academic Press.
- Csörgő, S. (1983). The theory of functional least squares. *J. Aust. Math. Soc.* **34**, 336–355.
- Csörgő, S., Deheuvels, P. and Mason, D. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.* **13**, 1050–1077.
- Daley, D. J. and Hall, P. (1984). Limit laws for the maximum of weighted and shifted i.i.d. random variables. *Ann. Probab.* **12**, 571–587.
- David, H. A. (1970). *Order Statistics* (1st Edition) New York: Wiley.
- Delaigle, A. (2016). Peter Hall and deconvolution. *Ann. Statist.* **44**, 1854–1866.
- Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43**, 170–180.
- Feuerverger, A. and Hall, P. (1998). On statistical inference based on record values. *Extremes.* **1**, 169–190.
- Feuerverger, A. and Hall, P. (1998). Estimating a tail exponent by modelling departure from a Pareto distribution. *Ann. Statist.* **27**, 760–781.
- Guillou, A. and Hall, P. (1999). A diagnostic for selecting the threshold in extreme value analysis. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **63**, 293–305.
- Härdle, W., Hall, P. and Simar, L. (1995). Iterated bootstrap with applications to frontier models. *J. Productivity Anal.* **6**, 63–76.
- Hall, P. (1978a). On the extreme terms of a sample from the domain of attraction of a stable law. *J. Lond. Math. Soc.* **18**, 181–191.
- Hall, P. (1978b). Representations and limit theorems for extreme value distributions *J. Appl. Probab.* **15**, 639–644.
- Hall, P. (1979a). On the relative stability of large order statistics *Math. Proc. Cambridge Phil. Soc.* **86**, 467–475.
- Hall, P. (1979b). On the rate of convergence of normal extremes *J. Appl. Probab.* **16**, 433–439.
- Hall, P. (1980a). Estimating a density on the positive half line by the method of orthogonal series. *Ann. Inst. Statist. Math.* **32**, 351–362.
- Hall, P. (1980b). Estimating probabilities for normal extremes *Adv. in Appl. Probab.* **12**, 491–500.
- Hall, P. (1981a). On trigonometric series estimates of a density. *Ann. Statist.* **9**, 683–685.
- Hall, P. (1981b). Laws of the iterated logarithm for nonparametric density estimators. *Z. Wahr. verw. Geb.* **56**, 47–61.
- Hall, P. (1982a). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **44**, 37–42.
- Hall, P. (1982b). On estimating the endpoint of a distribution. *Ann. Statist.* **10**, 556–568.
- Hall, P. (1984). On the influence of extremes on the rate of convergence in the central limit theorem. *Ann. Probab.* **12**, 154–172.
- Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32**, 177–203.

- Hall, P., Park, B. U. and Stern, S. E. (1998). On polynomial estimators of frontiers and boundaries. *J. Multivariate Anal.* **66**, 71–98.
- Hall, P. and Simar, L. (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *J. Amer. Statist. Assoc.* **97**, 523–524.
- Hall, P. and Tajvidi, N. (2000a). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli* **6**, 835–844.
- Hall, P. and Tajvidi, N. (2000b). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statist. Sci.* **15**, 153–167.
- Hall, P. and Tajvidi, N. (2004). Prediction regions for bivariate extreme events. *Aust. N. Z. J. Stat.* **46**, 99–112.
- Hall, P. and Wang, J. Z. (2005). Bayesian likelihood methods for estimating the end point of a distribution. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **67**, 717–729.
- Hall, P. and Weissman, I. (1997). On the estimation of extreme tail probabilities. *Ann. Statist.* **25**, 1311–1326.
- Hall, P. and Welsh, A. H. (1983). A test for normality based on the empirical characteristic function. *Biometrika* **70**, 830–834. Correction: *ibid.* (1984) **71**, 655.
- Hall, P. and Welsh, A. H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Ann. Statist.* **12**, 1079–1084.
- Hall, P. and Welsh, A. H. (1985a). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13**, 331–341.
- Hall, P. and Welsh, A. H. (1985b). Limit theorems for the median deviation. *Ann. Inst. Statist. Math.* **37**, 27–36.
- Hill B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.
- Pickands, J. (1981). Multivariate extreme value distributions. In *Bulletin of the International Statistical Institute: Proceedings of the 43rd Session (Buenos Aires)*, pp. 859–878. Voorburg, Netherlands: ISI.
- Polfeldt, T. (1970a). The order of the minimum variance in a nonregular case. *Ann. Math. Statist.* **41**, 667–672.
- Polfeldt, T. (1970b). Asymptotic results in nonregular estimation. *Skand. Aktuarietidskr* Supp. **1–2**, 667–672.
- Rényi, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hung.* **4**, 191–231.
- Roach, S. A. (1968). *The Theory of Random Clumping*. Methuen Monograph.
- Robinson, J. and Welsh, A. H. (2017). Peter Gavin Hall 1951–2016. *Historical Records of Australian Science* **28**, 171–182. Also published in *Royal Society of London.*, <http://rsbm.royalsocietypublishing.org/cgi/content/abstract/rsbm.2017.0035>.
- Woodroffe, M. (1974). Maximum likelihood estimation of a translation parameter of a truncated distribution II. *Ann. Statist.* **3**, 474–488.

Mathematical Sciences Institute, The Australian National University, Building 145 Science Road, Canberra ACT 2601 Australia.

E-mail: Alan.Welsh@anu.edu.au

(Received February 2017; accepted March 2017)