Stochastic Hydrol. Hydraul. 3 (1989) 281–292

Stochastic Hydrology and Hydraulics

Single station and regional analysis of daily rainfall extremes

D. L. Fitzgerald

Climatological Division, Irish Meteorological Service, Glasnevin Hill, Dublin 9, Ireland

Abstract: A peaks over threshold (POT) method of analysing daily rainfall values is developed using a Poisson process of occurrences and a generalised Pareto distribution (GPD) for the exceedances. The parameters of the GPD are estimated by the method of probability weighted moments (PWM) and a method of combining the individual estimates to define a regional curve is proposed.

Key words: Generalised Pareto distribution, Peaks over threshold, Probability weighted moments, Regionalisation

1 Introduction

Because of the high spatial variability of rainfall, the return period of a given daily total may be very different at neighbouring stations. Since the period of record of most stations is less than 50 years, the estimates of 100 to 200 year return periods commonly required by hydrologists have a high standard error. Regionalisation or pooling of records is an attempt to meet these difficulties. In the British Flood Studies Report (1975) a regional growth curve was defined based on Jenkinson's method of quartile analysis of the generalised extreme value distribution (GEV). This annual maximum (AM) method was evaluated by Hosking et al. (1985) and an alternative AM method based on PWM estimates of GEV parameters proposed. For individual stations Van Montfort and Witter (1986) employed a POT method and, using maximum likelihood (ML) parameter estimates, they concluded that the GPD was quite applicable. For extreme rainfall series the shape parameter, k, of the GPD is usually negative; for this case Hosking and Wallis (1987), in their comparison of ordinary moment, PWM and ML estimates of the GPD parameters, concluded that, while the ordinary moments provide the estimates with the lowest RMSE, in the most frequently encountered cases ($0>k>-0.2$) the PWMs have only slightly higher RMSE but have much lower bias. This latter property is most important to a regionalisation scheme and, as the main aim of this paper is the development of such a scheme, the PWM method of estimation was adopted.

2 Generalised Pareto distribution-PWM parameter estimates

Pickands (1975) showed that the GPD arises as a limiting form of independent exceedances of a high threshold. When $k<0$ the GPD arises as a compound of exponentials whose mean has a gamma distribution (Johnson and Kotz 1970). This lends plausibility to using the same distribution for both summer and winter rainfall as was done in

this study. A GPD variate has the form

$$y = \frac{d}{k}(1-(1-F)^k) \tag{1}$$

where $y(>0)$ is the exceedance of a set threshold, d is a positive scale parameter and k is a shape parameter which in applications is usually between +0.5 and -0.5 (Smith 1984). When k is positive y is bounded above and when $k=0$ the GPD reduces to the exponential form. The inverse form is

$$F = 1-(1-\frac{k}{d}y)^{\frac{1}{k}} \quad k\neq 0 \tag{2}$$

Hosking and Wallis (1987) give the GPD parameters in terms of the PWMs E(y) and E(y(1-F)), where E is the expectation operator, and also give an asymptotic expression for the covariance matrix of the two parameters.

Here the PWMs $B_0 = E(y) = \frac{d}{1+k}$ and $B_1 = E(yF) = \frac{d(k+3)}{2(k+2)(k+1)}$ are used giving

$$k = \frac{4(\frac{B_1}{B_0})-3}{1-2\frac{B_1}{B_0}} \quad \text{and} \quad d = -2\frac{B_0(1-\frac{B_1}{B_0})}{1-2\frac{B_1}{B_0}} \tag{3}$$

3 Sample estimates of the PWMs

Considering an ordered sample $(y_1, y_2, \cdots y_n)$, then after the manner of Kendall (1975)

$$E(\sum_{i=1}^{n}(i-1)y_i) = \sum_{i=2}^{n}\frac{n!}{(i-2)!(n-i)!}\int_z zF(z)^{i-1}(1-F(z))^{n-i}dF(z)$$

$$= n(n-1)\int_z zF(z)dF(z)\sum_{i=2}^{n}\frac{(n-2)!}{(i-2)!(n-i)!}F(z)^{i-2}(1-F(z))^{n-i};$$

$$\frac{1}{n(n-1)}E(\sum_{i=1}^{n}(i-1)y_i) = \int_z zF(z)dF(z) = E(yF)$$

Thus the GPD parameters may be estimated from the (ordered) sample values since

$$b_0 = \frac{1}{n}\sum_{i=1}^{n}y_i \quad \text{and} \quad b_1 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(i-1)y_i \tag{4}$$

are unbiassed estimates of the first two PWMs

Hosking and Wallis (1987) employ $\alpha_1 = \sum_{i=1}^{n}\frac{(n-i)}{n(n-1)}y_i$ as an estimator of $E(y(1-F)$ but since $\alpha_1+b_1=b_0$ their formulae and (3) are equivalent.

4 Covariance and variances of the PWMs and of the GPD parameters

Since the mean exceedance b_0 does not use the ordering of the values it is readily shown that

$$var(b_0) = \frac{d^2}{n(1+2k)(1+k)^2} \approx \frac{b_0^2}{n(1+2k)} \quad k \neq 0$$

$$= \frac{d^2}{n} \approx \frac{b_0^2}{n} \quad k = 0$$

Since b_1 weights the ordered sample values, we have from (4)

$$n^2(n-1)^2 var(b_1) = \sum_{i=1}^{n}(i-1)^2 var(y_i) + \sum\sum_{i \neq j}(i-1)(j-1)cov(y_i, y_j)$$

Using the properties of the GPD we get (Appendix A)

$$var(y_i) = \frac{d^2}{k^2}\left[\prod_{r=1}^{i}\frac{n-r+1}{n-r+2k+1} - \left(\prod_{r=1}^{i}\frac{n-r+1}{n-r+k+1}\right)^2\right] \quad k \neq 0$$

$$= d^2 \sum_{r=n-i+1}^{n}\frac{1}{r^2} \quad k = 0$$

and $i < j$

$$cov(y_i, y_j) = var(y_i)\left[\prod_{m=1}^{j}\frac{n-m+1}{n-m+k+1} \div \prod_{p=1}^{i}\frac{n-p+1}{n-p+k+1}\right] \quad k \neq 0$$

$$= var(y_i) \quad k = 0$$

From these expressions for the variances and covariances of the sample order statistics we can obtain var (b_1) and in addition

$$cov(b_0, b_1) = \frac{1}{n^2(n-1)}\left[\sum_{i=1}^{n}(i-1)var(y_i) + \sum\sum_{i \neq j}(i+j-2)cov(y_i, y_j)\right]$$

Now the usual method of finding a first order approximation of the variances and covariances of functions of random variables (Kendall and Stuart 1977) can be applied to the GPD parameters of eqn (3).

$$Putting \quad g = \frac{b_1}{b_0} \quad we\ get \quad var(k) = \frac{4var(g)}{(1-2g)^4} \tag{5}$$

where $var(g) = \frac{1}{b_0^2}(var(b_1) + (g)^2 var(b_0) - 2g\ cov(b_0, b_1))$, $\quad k \neq 0$

$$\text{and} \quad var(d) = \frac{4(A^2 var(b_0) + C^2 var(b_1) - 2AC\ cov(b_0, b_1))}{(2g-1)^2} \tag{6}$$

where

$$A = \frac{1+2g(1-g)}{2g-1} \quad \text{and} \quad C = \frac{1}{2g-1}$$

Finally

$$cov(d,k) = \frac{4b_1(4b_0b_1-2b_1^2-b_0^2)var(b_0)+4b_0^3var(b_1)-4b_0(5b_0b_1-2b_1^2-b_0^2)cov(b_0,b_1)}{(b_0-2b_1)^4} \quad (7)$$

For

$$k=0 \quad var(d) = \frac{1}{n}var(y) = \frac{d^2}{n} \quad (8)$$

In principle the expressions given here have the advantage of being valid for samples of size n; in practice the asymptotic expressions given in Hosking and Wallis (1987) or even the ML expressions (Smith 1984) gave similar results for the sample sizes (>40) encountered in the ensuing work.

5 Quantile estimation

Assume that the process generating the occurrences is Poisson with rate parameter λ independent of the process of exceedances of the threshold y_0 (with *CDF* F_0); then the number of exceedances of $y+y_0$ in T time units has the average value

$$N_y = \lambda_0 T(1-F_0(y)) \quad (9)$$

By definition if $N_y=1$ then $y=y_T$ is the exceedance having return period T. Hence

$$F_0(y_T) = 1-\frac{1}{\lambda_0 T}$$

For the GPD

$$y_T = \frac{d}{k}(1-(\lambda_0 T)^{-k}) \quad k\neq 0$$

$$= -d\ln(\lambda_0 T) \quad k = 0 \quad (10)$$

On the assumption that the covariances (d,λ) and (k,λ) are zero, the usual first order approximation about the true mean (m) yields

$$var(y_T) = (\frac{\partial y_T}{\partial d})_m^2 var(d)+(\frac{\partial y_T}{\partial k})_m^2 var(k)+(\frac{\partial y_T}{\partial \lambda})_m^2 var(\lambda)$$

$$+2(\frac{\partial y_T}{\partial d})_m(\frac{\partial y_T}{\partial k})_m cov(d,k) \quad (11)$$

All these quantities can be obtained from eqns (3) to (10) and by using $var(\lambda)=\lambda/t$ where t is the number of years of record and λ is expressed in events per year.
For $k=0$ we get

$$var(y_T) = (ln(\lambda T))^2 var(d)+\frac{d^2}{\lambda}var(\lambda) \quad (12)$$

6 Single station analysis

If y, the exceedance of the threshold y_0, is a GPD variate then for $y_s>y_0$ then CDF is readily seen, by insertion of the GPD form in the standard expression for a truncated distribution, to be again of GPD form with the same value of the shape parameter k but with scale parameter d_s given by

$$d_s = d_0(1-\frac{k}{d_o}s)$$

Since from (3) the ratio B_1/B_0 determines k, the ratio is not necessarily affected by a change of threshold. Also from (3) the mean exceedance of the higher threshold is given by

$$B_s = B_0(1-\frac{k}{d_0}s) \tag{13}$$

Davison (1984) suggests a plot of mean exceedance against threshold as a means of choosing a suitable base value above the (non-linear) lower portion of the graph. While the constancy of the ratio b_1/b_0 and (13) provided some guidance in the choice of threshold they were less than satisfactory and so it was decided to accept the lowest threshold for which the following three criteria held:

(a) the process of exceedances followed the GPD

Since we are especially concerned with the higher values of the CDF, the 'global' test of fit was based on its ordered values weighted by the mean values and was $\frac{1}{n}\sum_{i=1}^{n}\frac{i}{n+1}F_i$ which has mean $\frac{2n+1}{6(n+1)}$ and variance $\frac{(2n+1)(2n+3)}{180n(n+1)^2}$

Arbitrarily it was decided to accept only those values within one standard error of the mean; in all cases the CDF values generated by the PWMs were very close to the mean.

The pattern of F over (0,1) was examined by means of a test of spacings (Pyke 1965); he lists a number of tests including

$$H = \sum_{i=1}^{n+1}\frac{1}{n+1}\ln D_i, \quad D_i = F_i - F_{i-1}$$

Because of the number of ties it was decided to use instead a closely related statistic

$$S = \sum_{i=1}^{n+1} D_i \ln D_i$$

In Appendix B the following results are obtained:

$E(S) = -\sum_{i=2}^{n+1}\frac{1}{i} \approx -(ln(n+1.5)-1+\gamma)$ where γ is Euler's constant

$var(S) = \sum_{i=2}^{n+1}\frac{1}{i^2} - \frac{n}{n+2}(\frac{\pi^2}{6}-1)$ which is small even for modest sample sizes

S has its minimum when each $D_i = \frac{1}{n+1}$ and the sample CDF values have a perfectly regular pattern on (0,1).

Values of S near to the maximum of zero indicate highly clustered CDF values

Arbitrarily, $S > E(S)+\sqrt{var(S)}$ was regarded as too clustered while $S < E(S)-2\sqrt{var(S)}$ was rejected as too regular (superuniform).

(b) the process of counts was Poisson

The interoccurrence times were employed and, as well as the equality of the mean and standard error for exponential variates, the test of Hollander and Proschan (1972) was used. Again values had to be within one standard error of the mean.

(c) the fit of eqn (9) was tested by calculating r, the correlation coefficient, between $N_{y0}(1-F_0(s))$ and N_{ys}, the number of exceedances of the higher threshold s. The mean, standard deviation and also the individual values of the residuals for a series of s-values were taken into account, with special attention paid to the ten highest s-values. As r was usually higher than 0.99, the necessarily rather subjective examination of the residuals was the main element of this test.

6.1 Application

For their years of record in the period 1941-1986, daily totals in excess of 20 mm and their dates of occurrence were extracted for a selection of stations in the dry coastal strip of north Leinster (Fig. 1). The set of stations was chosen because the terrain is relatively uniform and the systems causing the rainfall usually affect the whole area about the same time. A daily total was regarded as a peak only if no higher daily fall occurred within the preceding or following three days; this was an attempt to reduce dependence between peaks, as the GPD has so far been shown to be valid only in the case of independent exceedances. There were generally about three peaks per year and these we may reasonably expect to be independent.

Next the variation of b_0 the mean exceedance and b_1/b_0, which determines the shape parameter, were found for a series of thresholds 21, 22..., the upper limit being where the number of exceedances was between 5 and 15. Perusal of the variation of b_0 and b_1/b_0 with threshold enabled a reasonable starting value to be chosen. Conformity to the criteria (a) to (c) was examined for thresholds at or above above this value and the lowest base conforming to the criteria accepted. For this threshold the parameters of the GPD plus their variances and covariance, the Poisson rate parameter plus its variance and the quantiles for 50, 100 and 200 years and their standard deviations were estimated from eqns (3) to (12). The results are given in Table 1 and were used to decide the (arbitrary) criteria for the regionalisation scheme.

7 Regionalisation-Pooling of records

The values of the 50-year return period, RP_{50}, were examined and only those within one standard deviation (SD) of the median of the set were used. Of this set only those stations whose shape parameter k was within one SD of the median were retained. The regional value of the shape parameter k_R was taken to be the weighted mean of the set. As weights n_i, the number of exceedances, were used.

The lowest threshold of the set was then considered as a base value. If more than one station had this base as threshold, then the one with the highest rate parameter was selected. Stations with thresholds higher than the base were regarded as having truncated versions of the basic curve for the region. For a station with threshold s units above the base y_0, the equivalent scale parameter was assumed to be

$$d_0 = \frac{d_s}{(1-\frac{k_R}{d_{0R}}s)} \simeq d_s(\frac{\lambda_0}{\lambda_s})^{k_R}$$

The mean (weighted by n_i) over the set of stations of the d_0 values then determined d_{0R}, the regional scale parameter for the base y_0.

The regional rate parameter λ_{0R} was then obtained as the weighted mean of the set of values

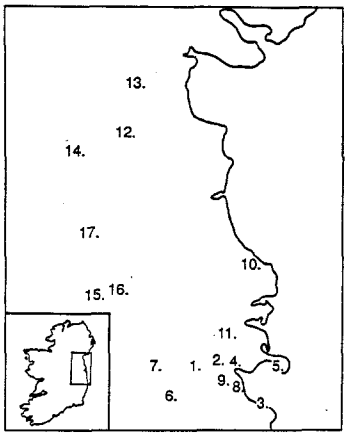

Figure 1. A selection of lowland stations from the dry coastal strip of North Leinster

$$\lambda_{0S} = \lambda_s(1-\frac{k_R}{d_{0R}}s)^{\frac{-1}{k_R}}$$

For the region we then have a base value (0) and the three parameters d_{0R},k_R and λ_{0R}. Since each station has contributed to the parameter values, the standard deviations of the parameters and $cov(k_R,d_{0R})$ can be estimated from the three sets of values. Quantiles for any return period and their standard deviations can then be generated.

7.1 Example

The 17 stations in Table 1 are all lowland stations in the dry eastern coastal strip of north Leinster (Fig. 1). The median value of RP_{50} is 63 mm and station number 12 was eliminated as its estimated RP_{50} was too low. The criterion for k eliminated station 8 whose highest fall was recorded during an intense thunderstorm in June 1963 which nearby station 9 largely escaped. The remaining 15 stations were then pooled giving the results shown in Table 2; these determine a reference curve for the region and enable us, when estimating the return period of a daily total at any lowland station of the region, to assume a base of 20 mm, a scale parameter of 6.390, a shape parameter of -0.102 and a rate parameter of 3.727. The SDs of the GPD parameters are less than those of the individual stations but the SD of the rate parameter is higher. The regionalised covariance is lower than that of the individual stations. The resulting SDs of the regional quantiles appear too low and, if the SDs of the parameters are accepted as realistic, the covariance would need to be lower in order to get more 'realistic' error estimates; it would be reasonable here to assume a covariance of zero.

8 Conclusions

The combination of a Poisson process of occurrences and a GPD distribution of exceedances is shown to lend itself readily to a regionalisation scheme. Even for individual stations the POT method of extreme value analysis has the advantage over the annual maximum (AM) method of enabling larger sample sizes to be extracted from the same period of record. Here the requirements of the theory were often met by a process

Table 1. Peak over Threshold Analysis of Daily Rainfall

Stn		Thr		PWM					Cov			RP_{50}	RP_{100}	RP_{200}
No.	Years	mm	N	b_0	d	s_d^2	k	s_k^2	(d,k)	λ	s_λ^2	mm	mm	mm
1	41-86	28.0	68	7.801	6.230	1.43	-0.201	0.022	0.116	1.478	0.032	71±12	82±18	94± 27
2	41-86	29.0	49	7.594	5.918	1.82	-0.221	0.032	0.156	1.065	0.023	67±12	77±18	90±28
3	41-86	25.0	101	7.836	7.350	1.31	-0.062	0.014	0.099	2.196	0.048	65±8	72±12	79±16
4	41-86	25.0	85	7.382	7.036	1.40	-0.047	0.016	0.110	1.848	0.040	61±7	67±10	73±14
5	41-86	21.0	143	6.331	5.264	0.49	-0.168	0.010	0.047	3.109	0.068	63±9	72±14	82±19
6	54-86	21.0	96	8.110	6.853	1.19	-0.155	0.015	0.089	2.909	0.088	73±14	83±19	95±27
7	50-86	20.0	114	7.409	7.062	0.92	-0.047	0.010	0.066	3.081	0.083	60±8	66±11	73±14
8	60-86	25.0	71	8.407	5.794	1.29	-0.311	0.027	0.118	2.630	0.097	81±27	112±43	136±65
9	48-86	21.5	120	8.029	7.177	1.11	-0.106	0.012	0.085	3.077	0.079	70±11	79±15	88±20
10	41-86	23.0	116	7.642	7.054	1.13	-0.077	0.013	0.090	2.522	0.055	64±9	72±12	79±16
11	41-86	21.0	156	7.263	6.441	0.65	-0.113	0.009	0.053	3.391	0.074	66±9	74±12	83±17
12	41-86	22.0	106	7.902	9.128	1.85	+0.155	0.014	0.136	2.304	0.050	53±5	56±6	58±7
13	44-86	25.5	67	7.507	7.098	1.93	-0.054	0.032	0.156	1.558	0.036	61±8	67±12	74±16
14	44-86	23.0	80	7.369	6.694	1.37	-0.092	0.017	0.111	1.860	0.043	61±9	68±12	76±17
15	64-86	20.0	75	6.943	6.358	1.28	-0.084	0.018	0.108	3.261	0.142	61±11	67±15	75±20
16	52-86	20.0	123	7.028	6.540	0.84	-0.069	0.011	0.070	3.541	0.100	61±9	67±12	74±16
17	44-86	21.0	127	6.650	6.089	0.79	-0.084	0.012	0.073	2.953	0.069	59±8	65±11	73±15

Key: Thr=threshold; N=number of exceedances; b_0 = mean exceedance; d=scale parameter of GPD; s_d^2 = estimated variance of d; k=shape parameter of GPD; λ = Poisson rate parameter; RP_{50} = 50-year return period; standard deviation follows ±

Table 2. Estimates of Regional Parameters for base value 20 mm

Stn.No	1	2	3	4	5	6	7	9	10	11	13	14	15	16	17
d_{20}	5.709	5.248	7.007	6.593	5.196	6.720	6.964	7.007	6.819	6.413	6.539	6.277	6.306	6.540	5.979
-k	0.201	0.221	0.062	0.047	0.168	0.155	0.047	0.106	0.077	0.113	0.054	0.092	0.084	0.069	0.084
λ_{20}	4.816	3.985	4.670	3.930	3.632	3.398	3.081	3.882	3.993	3.962	3.562	2.945	3.261	3.541	3.450

Weighted Means d_{0R}=6.390±0.325; k_R=–0.102±0.048; $cov(d_{0R},k_R)$=0.019; λ_{0R}=3.727±0.471
Return Period RP_{50}=64±2; RP_{100}=72±4; RP_{150}=77±5; RP_{200}=80±6

of occurrences of about 3 per year and this increased sample size should produce a marked reduction in the SDs of the quantile estimates compared with the AM method for a single station.

The pooling of the station estimates requires the broad assumptions of a basic threshold, Poisson rate parameter and GPD parameters which apply throughout the region. The results in Table 1 and in Table 2 show that, for a suitably homogeneous rainfall region, the assumptions are quite tenable. It is intended to extend the analysis to other larger areas.

Acknowledgements

I wish to thank my colleagues A. McManus, P. Lynch and A. McDonald for their help and also G. Ross of BMO for his careful reading of the article.

References

Davison, A.C. 1984: Modelling excesses over high thresholds, with an application. In: J. Tiago de Oliveira (ed.), Statistical extremes and applications, Reidel, 461-482

Hollander, M; Proschan, F. 1972: Testing whether new is better than used. Ann. of Math. Statist. 43 (4), 1136-1146

Hosking, J.R.M.; Wallis, J.R.; Wood, E.F. 1985: An appraisal of the regional flood frequency proceedure in the U.K. Flood studies report. Hydrol. Sci. J. 30 (1), 85-109

Hosking, J.R.M.; Wallis, J.R. 1987: Parameter and quantile estimation for the generalised Pareto distribution. Technometrics 29 (3), 339-349

Johnson, N.L; Kotz, S. 1970: Distributions in statistics: Continuous univariate distributions-1, Wiley, London, Ch. 19

Kendall, M.G. 1975: Rank correlation methods, London, Griffin, Ch. 10

Kendall, M.G.; Stuart, A. 1977: The advanced theory of statistics, Vol. 1, Griffin, London, Ch. 10

Natural Environment Research Council 1975: Flood studies report, NERC, London

Pickands, J. 1975: Statistical inference using extreme order statistics. Ann. of Statist. 3 (1), 119-131

Pyke, R. 1965: Spacings, J. R. Statist. Soc. B 27, 395-449

Smith, R.L. 1984: Threshold methods for sample extremes. In: J. Tiago de Oliveria (ed.), Statistical extremes and applications, Reidel, 621-638

Van Montfort, M.A.V.; Witter, J.V. 1986: The generalised Pareto distribution applied to rainfall depths. Hydrol. Sci. J. 31 (12), 151-162

Appendix A Variances and covariance of the order statistics of the GPD

For the i^{th} order statistic the p.d.f. is

$$f(y_i=y)=\frac{1}{B(i,n-i+1)}F^{i-1}(y)(1-F(y))^{n-i}dF(y) \text{ where } B(\,) \text{ is the } \beta \text{ function}$$

For the GPD with $k\neq 0$

$$E(y_i)=\frac{d}{k}\frac{1}{B(i,n-i+1)}\int_0^1(1-(1-F(y))^k)F(y)^{i-1}(1-F(y))^{n-i}dF(y)$$

$$=\frac{d}{k}\left[1-\frac{B(i,n-i+k+1)}{B(i,n-i+1)}\right]=\frac{d}{k}\left[1-\prod_{p=1}^{p=i}\frac{n-p+1}{n-p+k+1}\right]$$

Similarly

$$E(y_i^2)=\frac{d^2}{k^2}\left[1-2\prod_{r=1}^{r=i}\frac{n-r+1}{n-r+k+1}+\prod_{s=1}^{s=i}\frac{n-s+1}{n-s+2k+1}\right]$$

$$\text{Hence } var(y_i)=\frac{d^2}{k^2}\left[\prod_{r=1}^{r=i}\frac{n-r+1}{n-r+2k+1}-(\prod_{s=1}^{s=i}\frac{n-s+1}{n-s+k+1})^2\right]$$

Also

$$E(y_i,y_j)=\frac{d^2}{k^2}E(1-(1-F_i)^k-(1-F_j)^k+(1-F_i)^k(1-F_j)^k)$$

But

$$E((1-F_s)^k)=\frac{B(s,n-s+k+1)}{B(s,n-s+1)}=\prod_{r=1}^{r=s}\frac{n-r+1}{n-r+k+1}$$

Also for $r<s$ and with $C=\frac{n!}{(r-1)!(s-r-1)!(n-s)!}$

$$f(F_r,F_s)=CF_r^{r-1}(F_s-F_r)^{s-r-1}(1-F_s)^{n-s}dF_rdF_s$$

As suggested in David (1981) let $y_2=F_s$ and $y_1y_2=F_r$. Both y_1 and y_2 range 0 to 1 and the Jacobian of the transformation is y_2. Then we have

$$f(y_1,y_2)=C(y_1y_2)^{r-1}(y_2-y_1y_2)^{s-r-1}(1-y_2)^{n-s}y_2dy_1dy_2$$

$$E\left[(1-F_r)^k(1-F_s)^k\right]=\frac{d^2}{k^2}C\int_0^1\int_0^1(1-y_1y_2)^k(1-y_2)^k y_1^{r-1}y_2^{s-r-1}(1-y_2)^{n-s}dy_1dy_2$$

But $(1-y_1y_2)^k=\sum_{t=0}^{t=\infty}\frac{(-k)_t}{t!}(y_1y_2)^t$ where $(-k)_t=-k(-k+1)(-k+2)...(-k+t-1)$ and $(-k)_0=1$ gives

$$E\left[(1-F_r)^k(1-F_s)^k\right]=\frac{d^2}{k^2}C\sum_{t=0}^{t=\infty}B(s+t,n-s+k+1)B(r+t,s-r)\frac{(-k)_t}{t!}$$

$$\propto B(r,s-r)B(s,n-s+k+1)\left[1+\frac{(-k)r}{n+k+1}+\frac{(-k)(-k+1)r(r+1)}{2!(n+k+1)(n+k+2)}+\frac{(-k)(-k+1)(-k+2)r(r+1)(r+2)}{3!(n+k+1)(n+k+2)(n+k+3)}+...\right]$$

using $B(m+1,n-m)=\frac{m}{n}B(m,n-m)$

$$E\left[(1-F_r)^k(1-F_s)^k\right]=\frac{d^2}{k^2}CB(r,s-r)B(s,n-s+k+1)F(-k,r,n+k+1,1)$$

where $F(-k,r,n+k+1,1)=\frac{\Gamma(n+k+1)\Gamma(n+2k-r+1)}{\Gamma(n+2k+1)\Gamma(n+k-r+1)}$ is a hypergeometric function (Sneddon 1956)

Substituting for C we get

$$E\left[(1-F_r)^k(1-F_s)^k\right]=\frac{d^2}{k^2}\frac{\Gamma(n+1)\Gamma(n-s+k+1)\Gamma(n+2k-r+1)\Gamma(n+k+1)}{\Gamma(n-s+1)\Gamma(n+k+1)\Gamma(n+2k+1)\Gamma(n+k-r+1)}$$

$$=\frac{d^2}{k^2}\left[\prod_{p=1}^{p=s}\frac{n-p+1}{n-p+k+1}\left(\prod_{m=1}^{m=r}\frac{n-m+k+1}{n-m+2k+1}\right)\right]\qquad r<s$$

For $i<j$

$$E(y_i,y_j)=\frac{d^2}{k^2}\left[1-\prod_{p=1}^{p=i}\frac{n-p+1}{n-p+k+1}-\prod_{p=1}^{p=j}\frac{n-p+1}{n-p+k+1}+\prod_{p=1}^{p=i}\frac{n-p+k+1}{n-p+2k+1}\left(\prod_{n=1}^{m=j}\frac{n-m+1}{n-m+k+1}\right)\right]$$

Subtracting $E(y_i)E(y_j)$ we get

$$cov(y_i,y_j)=\frac{d^2}{k^2}\left[\prod_{m=1}^{m=j}\frac{n-m+1}{n-m+k+1}\left(\prod_{p=1}^{p=i}\frac{n-p+k+1}{n-p+2k+1}-\prod_{p=1}^{p=i}\frac{n-p+1}{n-p+k+1}\right)\right]\qquad i<j$$

$$=var(y_i)\left(\prod_{m=1}^{m=j}\frac{n-m+1}{n-m+k+1}\right)+\left(\prod_{p=1}^{p=i}\frac{n-p+1}{n-p+k+1}\right))$$

For k=0 there is the result $var(y_i)=d^2\sum_{m=n-i+1}^{m=n}\frac{1}{m^2}$ (Gumbel 1958) and from the above we have $cov(y_i,y_j)=var(y_i)$

References

David, H.A. 1981: Order statistics, Wiley, London, Ch. 3

Gumbel, E.J. 1958: Statistics of extremes, Columbia Univ. Press, New York, Ch. 4

Sneddon, I.N. 1956: Special functions of mathematical physics and chemistry, Wiley, London

Appendix B The Spacings Statistics: $S=\sum_{i=1}^{n+1}D_i\ln D_i$ and $H=\sum_{i=1}^{n+1}\frac{\ln D_i}{n+1}$

In the notation of Wilks (1962) spacings have a Dirichlet distribution $D(1,1,...1,n-k+1)$ for k spacings from a sample of n values, $k\le n$, $\sum_{i=1}^{n}D_i\le 1$. His eqn. (7.7.6) leads to

$$E(D_i^{r_i} D_j^{r_j} \cdots D_t^{r_t}) = \frac{\Gamma(1+r_i)\Gamma(r+r_j) \cdots \Gamma(1+r_t)\Gamma(1+n)}{\Gamma(1+r_i+r_j \cdots r_i+n)} \qquad \text{[B1]}$$

If we consider the n+1 spacings $D_1, D_2, D_3, \ldots D_{n+1}, D_{n+1}=1-\sum_{i=1}^{n} D_i$, the same distribution applies, but with degeneracy. Exploiting $\frac{\partial^r}{\partial\alpha^r} D_i^{\alpha} = D_i^{\alpha} (lnD_i)^r$ we have $E(\frac{\partial^r}{\partial\alpha^r} D_i^{\alpha}) = E(D_i^{\alpha} (lnD_i)^r), r=1,2\ldots$ Since $0<D_i<1$, the order of differentiation and integration can be changed and we get

$$\frac{\partial^r}{\partial\alpha^r} E(D_i^{\alpha}) = E(D_i^{\alpha} (lnD_i)^r)$$

Similarly $\frac{\partial^{r+s}}{\partial\alpha^r \partial\beta^s} D_i^{\alpha} D_j^{\beta} = D_i^{\alpha} D_j^{\beta} (lnD_i)^r (lnD_j)^s$ and so on.

Hence from [B1] we get

$$E(D_i \ln D_i) = (\frac{\partial}{\partial\alpha} \frac{\Gamma(1+\alpha)\Gamma(1+n)}{\Gamma(1+\alpha+n)})_{\alpha=1} \qquad \text{[B2]}$$

Using the notion of Scheid (1968) for the digamma function, $\Psi(\alpha) = \frac{1}{\Gamma(1+\alpha)} \frac{\partial}{\partial\alpha} \Gamma(1+\alpha)$

Then $\Psi(0) = -\gamma = -0.5772$ and $\Psi(n) = \sum_{i=1}^{n} \frac{1}{i} - \gamma$ when n is an integer.

Successive differentiation yields $\Psi^{(m)}(n) = (-1)^m m! [-\xi(m+1) + 1 + \frac{1}{2^{m+1}} + \frac{1}{3^{m+1}} + \cdots \frac{1}{n^{m+1}}]$ where $\Psi^{(1)}(n)$ is the trigamma function and $\xi(r) = \sum_{i=1}^{\infty} \frac{1}{i^r}$.

From [B2] $E(D_i \ln D_i) = \frac{1}{n+1} (\Psi(1) - \Psi(n+1))$

Hence $E(S) = -\sum_{i=2}^{n+1} \frac{1}{i}$ and is well approximated by $-(ln(n+1.5)+\gamma-1)$

Similarly, with $\alpha=0$ we get $E(H) = \Psi(0) - \Psi(n) = -\sum_{i=1}^{n} \frac{1}{i} = -(ln(n+0.5)+\gamma)$

Variances

$$E(S^2) = (n+1)E(D_i^2 (lnD_i)^2) + n(n+1)E(D_i D_j \ln D_i \ln D_j)$$

$$= (n+1)(\frac{\partial^2}{\partial\alpha^2} E(D_i^{\alpha}))_{\alpha=2} + n(n+1)(\frac{\partial^2}{\partial\alpha\partial\beta} E(D_i^{\alpha} D_j^{\beta}))_{\alpha=1,\beta=1}$$

$$= \frac{2}{n+2} \Psi^{(1)}(2) - \Psi^{(1)}(n+2) + \frac{2}{n+2} (\Psi(2) - \Psi(n+2))^2 + \frac{n}{n+2} (\Psi(1) - \Psi(n+2))^2$$

Since $E(S) = \Psi(1) - \Psi(n+1)$ we get

$$var(S) = \frac{2}{n+2} \Psi^{(1)}(1) - \Psi^{(1)}(n+1) = \sum_{i=2}^{n+1} \frac{1}{i^2} - \frac{n}{n+2} (\frac{\pi^2}{6} - 1)$$

As $n \to \infty$ $var(S) \to \frac{2}{n+2} (\frac{\pi^2}{6} - 1)$

The same method with $\alpha=0$ and $\beta=0$ yields

$$var(H) = \frac{1}{n+1} \Psi^{(1)}(0) - \Psi^{(1)}(n) = \sum_{i=1}^{n} \frac{1}{i^2} - \frac{n}{n+1} \frac{\pi^2}{6}$$

As $n \to \infty$ $var(H) \to \frac{\pi^2}{6(n+1)}$ while the result in Darling (1953) is $(\frac{\pi^2}{6} - 1)(\frac{1}{n+1})$

Third and fourth moments about the mean

Proceeding as above $\mu_3(H)=\frac{1}{(n+1)^2}\Psi^{(2)}(0)-\Psi^{(2)}(n)$ but the expression for S becomes more complicated:

$$\mu_3(S)=\frac{6}{(n+2)(n+3)}\Psi^{(2)}(1)-\Psi^{(2)}(n+1)+\frac{3n}{(n+2)^2(n+3)}\Psi^{(1)}(1)$$

The coefficient of skewness $\sqrt{b_1}$ has values for H of -0.49, at n=30 and of -0.27, n=100. For S the values are +0.54 and 0.32 respectively.

$$\mu_4(H)=\frac{1}{(n+1)^3}\Psi^{(3)}(0)-\Psi^{(3)}(n)+3\mu_2^2(H)$$

This gives a kurtosis excess for H of about 0.36 at n=30 and of 0.15 at n=100. For S values of a similar magnitude may be expected. This lends plausibility to using measures based on standard error as rough tests.

The advantage of S over H is that ties contribute only a small amount to S while for H the method of breaking ties is crucial. Since rainfalls are read to 0.1mm, it was decided to break ties (even if multiple) by assuming a difference of 0.04 mm between each pair.

References

Darling, D.A. 1953: On a class of problems relating to the random division of an interval. Ann. of Math. Statist. 24, 239-253

Scheid, F. 1968: Numerical analysis, McGraw Hill, New York, Ch. 18

Wilks, S.S. 1962: Mathematical statistics, Wiley, London, Chs. 7 and 8

Accepted July 30, 1989