Algorithm AS 217

Computation of the Dip Statistic to Test for Unimodality

By P. M. Hartigan

VA Medical Center, USA

[Received August 1984. Revised May 1985]

Keywords: Unimodality; Isotonic regression; Dip statistic

Language

Fortran 66

Description and Purpose

The dip statistic is the maximum difference between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference. The dip measures departure of the sample from unimodality and is proposed by Hartigan and Hartigan (1985) for use in a test of unimodality. Asymptotically the dip for samples from a unimodal distribution approaches zero and that for samples from any multimodal distribution approaches a positive constant.

The null distribution for the dip test is the uniform, as a "worst case" unimodal distribution. Asymptotically $\sqrt{n}$(dip) is positive for the uniform and zero for unimodal distributions whose densities decrease exponentially away from the mode. If the true unimodal distribution is not uniform, other ways of assessing the evidence for unimodality may be more powerful (Hartigan and Hartigan, 1985). Subroutine *DIPTST* calculates the dip statistic and the modal interval for the "best" fitting unimodal distribution for an ordered set of data.

Numerical Method

Let $x_1, x_2, \ldots, x_n$ be the ordered observations. The only possible endpoints for the estimated modal interval (x_L, x_U) are pairs of these observations. Consider the $n(n-1)/2$ possible modal intervals and compute for each (x_i, x_j) the greatest convex minorant (g.c.m.) of the empirical distribution function, F_n, in $(-\infty, x_i)$ and the least concave majorant (l.c.m.) of F_n in (x_j, ∞). Let d_{ij} be the maximum distance between F_n and these computed curves. Then twice the dip is the minimum value of d_{ij} over all modal intervals (x_i, x_j), such that the line segment from $[x_i, F^-(x_i) + 1/2\, d_{ij}]$ to $[x_j, F(x_j) - 1/2\, d_{ij}]$ lies in the set

$$\{x, y \mid x_i \leqslant x \leqslant x_j, F(x) - 1/2\, d_{ij} \leqslant y \leqslant F(x) + 1/2\, d_{ij}\}.$$

This condition ensures that the minorant, the modal segment, and the majorant can be pieced together to form a unimodal distribution.

The minorant and majorant computations are done once and for all in order n, and an order n algorithm exists to determine the dip and the modal interval.

(1) Begin with $x_L = x_1, x_U = x_n, D = 0$.
(2) Compute the g.c.m. G and the l.c.m. H for F_n in $[x_L, x_U]$; suppose the points of contact with F_n are respectively $g_1, g_2, \ldots, g_k$ and $h_1, h_2, \ldots, h_m$.

Present address: V. A. Cooperative Studies Coordinating Center, Medical Center, West Spring Street, West Haven, CT 06516, USA.

 0035–9254/85/34320 $2.00

(3) Suppose $d = \max \mid G(g_i) - H(g_i) \mid > \max \mid G(h_j) - H(h_j) \mid$ and that this maximum occurs at $h_j \leqslant g_i \leqslant h_{j+1}$.
Define $x_L^0 = g_i, x_U^0 = h_{j+1}$.

(4) Suppose $d = \max \mid G(h_j) - H(h_j) \mid \geqslant \max \mid G(g_i) - H(g_i) \mid$ and that this maximum occurs at $g_i \leqslant h_j \leqslant g_{i+1}$.
Define $x_L^0 = g_i, x_U^0 = h_j$

(5) If $d \leqslant D$, stop and set dip $= D/2$

(6) If $d > D$ set D max $\left(\sup_{x_L \leqslant x \leqslant x_L^0} \mid G(x) - F(x) \mid, \sup_{x_U \leqslant x \leqslant x_U^0} \mid H(x) - F(x) \mid \right)$

(7) Set $x_L = x_L^0, x_U = x_U^0$ and return to (2).

Table 1 has the percentage points of the dip statistic based on 9999 repetitions from the uniform. Since Hartigan and Hartigan (1985) have shown that $\sqrt{n}$*(dip) converges in distribution to the dip computed for a Brownian bridge it is suggested that interpolation be based on $\sqrt{n}$*(dip).

TABLE 1
Percentage points of the dip statistic in uniform samples

SAMPLE SIZE	PROBABILITY OF DIP LESS THAN TABLED VALUE[1]								
	.01	.05	.10	.50	.90	.95	.99	.995	.999
4	.1250	.1250	.1250	.1250	.1863	.2056	.2325	.2387	.2458
5	.1000	.1000	.1000	.1217	.1773	.1872	.1966	.1981[2]	.1996[2]
6	.0833	.0833	.0833	.1224	.1586	.1645	.1904	.2034	.2224
7	.0714	.0714	.0822	.1181	.1445	.1597	.1832	.1900	.2035
8	.0625	.0745	.0828	.1109	.1428	.1552	.1744	.1801	.1978
9	.0618	.0735	.0807	.1041	.1362	.1458	.1623	.1693	.1851
10	.0610	.0718	.0780	.0979	.1302	.1394	.1623[2]	.1699	.1828
15	.0544	.0606	.0641	.0836	.1097	.1179	.1365	.1424	.1538
20	.0474	.0529	.0569	.0735	.0970	.1047	.1209	.1262	.1382
30	.0395	.0442	.0473	.0617	.0815	.0884	.1012	.1061	.1177
50	.0312	.0352	.0378	.0489	.0645	.0702	.0804	.0842	.0926
100	.0228	.0256	.0274	.0355	.0471	.0510	.0586	.0619	.0987
200	.0165	.0185	.0197	.0255	.0341	.0370	.0429	.0449	.0496

[1] Based on 9999 simulations
[2] Repeated computations

Structure

SUBROUTINE DIPTST (X, N, DIP, XL, XU, IFAULT, GCM, LCM, MN, MJ)

Formal parameters

X	Real array (*N*)	input: the data, in ascending order
N	Integer	input: the number of observations

DIP	Real	output:	the dip statistic
XL	Real	output:	the lower end of the modal interval
XU	Real	output:	the upper end of the modal interval
IFAULT	Integer	output:	error indicator
MN	Integer array (N)	workspace:	contains the indices for the increasing (convex minorant) fit
MJ	Integer array (N)	workspace:	contains the indices for the decreasing (concave majorant) fit
GCM	Integer array (N)	workspace:	change points for the g.c.m.
LCM	Integer array (N)	workspace:	change points for the l.c.m.

Fault indications

$IFAULT = 0$ indicates successful execution of routine
$= 1$ indicates $N = 0$
$= 2$ indicates data are not sorted in ascending order.

Restrictions

The data input into the routine are assumed to be in ascending order and therefore to have no missing observations.

Timing

In Table 2 the run times for the subroutine are compared to a heap-sort sorting routine. Each time is an average of 30 runs. The dip computation appears to take approximately the same amount of time as the heap-sort routine.

TABLE 2
Run times on an IBM 4341, in seconds, for the dip computation and a heap-sort routine

N	1000	500	200	50	20
DIP	0.173	0.085	0.037	0.009	0.004
SORT	0.181	0.084	0.030	0.006	0.002

Acknowledgement

This work was supported in part by National Science Foundation Contract No. DCR 8401636.

References

Hartigan J. A. and Hartigan, P. M. (1985) The dip test of unimodality. *Ann. Statist.*, **13**, 70–84.

```
      SUBROUTINE DIPTST(X, N, DIP, XL, XU, IFAULT, GCM, LCM, MN, MJ)
C
C        ALGORITHM AS 217 APPL. STATIST. (1985) VOL.34, NO.3
C
C        DOES THE DIP CALCULATION FOR AN ORDERED VECTOR X USING
C        THE GREATEST CONVEX MINORANT AND THE LEAST CONCAVE MAJORANT,
C        SKIPPING THROUGH THE DATA USING THE CHANGE POINTS OF THESE
C        DISTRIBUTIONS. IT RETURNS THE DIP STATISTIC 'DIP' AND THE MODAL
C        INTERVAL (XL,XU)
C
      DIMENSION X(N)
      DIMENSION MN(N), MJ(N), LCM(N)
      INTEGER GCM(N), HIGH
C
C        CHECK THAT N IS POSITIVE
C
```

```
      IFAULT = 1
      IF (N .LE. 0) RETURN
      IFAULT = 0
C
C        CHECK IF N IS ONE
C
      IF (N .EQ. 1) GOTO 4
C
C        CHECK THAT X IS SORTED
C
      IFAULT = 2
      DO 3 K = 2, N
      IF (X(K) .LT. X(K - 1)) RETURN
    3 CONTINUE
      IFAULT = 0
C
C        CHECK FOR ALL VALUES OF X IDENTICAL
C        AND FOR 1.LT.N.LT.4
C
      IF (X(N) .GT. X(1) .AND. N .GE. 4) GOTO 5
    4 XL = X(1)
      XU = X(N)
      DIP = 0.0
      RETURN
C
C        LOW CONTAINS THE INDEX OF THE CURRENT ESTIMATE OF
C        THE LOWER END OF THE MODAL INTERVAL, HIGH CONTAINS
C        THE INDEX FOR THE CURRENT UPPER END
C
    5 FN = FLOAT(N)
      LOW = 1
      HIGH = N
      DIP = 1.0 / FN
      XL = X(LOW)
      XU = X(HIGH)
C
C        ESTABLISH THE INDICES OVER WHICH COMBINATION IS
C        NECESSARY FOR THE CONVEX MINORANT FIT
C
      MN(1) = 1
      DO 28 J = 2, N
      MN(J) = J - 1
   25 MNJ = MN(J)
      MNMNJ = MN(MNJ)
      A = FLOAT(MNJ - MNMNJ)
      B = FLOAT(J - MNJ)
      IF (MNJ .EQ. 1 .OR. (X(J) - X(MNJ)) * A .LT. (X(MNJ) - X(MNMNJ))
     *  * B) GOTO 28
      MN(J) = MNMNJ
      GOTO 25
   28 CONTINUE
C
C        ESTABLISH THE INDICES OVER WHICH COMBINATION IS
C        NECESSARY FOR THE CONCAVE MAJORANT FIT
C
      MJ(N) = N
      NA = N - 1
      DO 34 JK = 1, NA
      K = N - JK
      MJ(K) = K + 1
   32 MJK = MJ(K)
      MJMJK = MJ(MJK)
      A = FLOAT(MJK - MJMJK)
      B = FLOAT(K - MJK)
      IF (MJK .EQ. N .OR. (X(K) - X(MJK)) * A .LT. (X(MJK) - X(MJMJK))
     *  * B) GOTO 34
      MJ(K) = MJMJK
      GOTO 32
   34 CONTINUE
C
C        START THE CYCLING
C        COLLECT THE CHANGE POINTS FOR THE GCM FROM HIGH TO LOW
C
```

```
   40 IC = 1
      GCM(1) = HIGH
   42 IGCM1 = GCM(IC)
      IC = IC + 1
      GCM(IC) = MN(IGCM1)
      IF (GCM(IC) .GT. LOW) GOTO 42
      ICX = IC
C
C        COLLECT THE CHANGE POINTS FOR THE LCM FROM LOW TO HIGH
C
      IC = 1
      LCM(1) = LOW
   44 LCM1 = LCM(IC)
      IC = IC + 1
      LCM(IC) = MJ(LCM1)
      IF (LCM(IC) .LT. HIGH) GOTO 44
      ICV = IC
C
C        ICX,IX,IG ARE COUNTERS FOR THE CONVEX MINORANT
C        ICV,IV,IH ARE COUNTERS FOR THE CONCAVE MAJORANT
C
      IG = ICX
      IH = ICV
C
C        FIND THE LARGEST DISTANCE GREATER THAN 'DIP'
C        BETWEEN THE GCM AND THE LCM FROM LOW TO HIGH
C
      IX = ICX - 1
      IV = 2
      D = 0.0
      IF (ICX .NE. 2 .OR. ICV .NE. 2) GOTO 50
      D = 1.0 / FN
      GOTO 65
   50 IGCMX = GCM(IX)
      LCMIV = LCM(IV)
      IF (IGCMX .GT. LCMIV) GOTO 55
C
C        IF THE NEXT POINT OF EITHER THE GCM OR LCM IS
C        FROM THE LCM THEN CALCULATE DISTANCE HERE
C
      LCMIV1 = LCM(IV - 1)
      A = FLOAT(LCMIV - LCMIV1)
      B = FLOAT(IGCMX - LCMIV1 - 1)
      DX = (X(IGCMX) - X(LCMIV1) * A) / (FN * (X(LCMIV) - X(LCMIV1)))
     *  - B / FN
      IX = IX - 1
      IF (DX .LT. D) GOTO 60
      D = DX
      IG = IX + 1
      IH = IV
      GOTO 60
C
C        IF THE NEXT POINT OF EITHER THE GCM OR LCM IS
C        FROM THE GCM THEN CALCULATE DISTANCE HERE
C
   55 LCMIV = LCM(IV)
      IGCM = GCM(IX)
      IGCM1 = GCM(IX + 1)
      A = FLOAT(LCMIV - IGCM1 + 1)
      B = FLOAT(IGCM - IGCM1)
      DX = A / FN - ((X(LCMIV) - X(IGCM1)) * B) / (FN * (X(IGCM)
     *  - X(IGCM1)))
      IV = IV + 1
      IF (DX .LT. D) GOTO 60
      D = DX
      IG = IX + 1
      IH = IV - 1
   60 IF (IX .LT. 1) IX = 1
      IF (IV .GT. ICV) IV = ICV
      IF (GCM(IX) .NE. LCM(IV)) GOTO 50
   65 IF (D .LT. DIP) GOTO 100
```

```
C        CALCULATE THE DIPS FOR THE CURRENT LOW AND HIGH
C
C        THE DIP FOR THE CONVEX MINORANT
C
      DL = 0.0
      IF (IG .EQ. ICX) GOTO 80
      ICXA = ICX - 1
      DO 76 J = IG, ICXA
      TEMP = 1.0 / FN
      JB = GCM(J + 1)
      JE = GCM(J)
      IF (JE - JB .LE. 1) GOTO 74
      IF (X(JE) .EQ. X(JB)) GOTO 74
      A = FLOAT(JE - JB)
      CONST = A / (FN * (X(JE) - X(JB)))
      DO 72 JR = JB, JE
      B = FLOAT(JR - JB + 1)
      T = B / FN - (X(JR) - X(JB)) * CONST
      IF (T .GT. TEMP) TEMP = T
   72 CONTINUE
   74 IF (DL .LT. TEMP) DL = TEMP
   76 CONTINUE
C
C        THE DIP FOR THE CONCAVE MAJORANT
C
   80 DU = 0.0
      IF (IH .EQ. ICV) GOTO 90
      ICVA = ICV - 1
      DO 88 K = IH, ICVA
      TEMP = 1.0 / FN
      KB = LCM(K)
      KE = LCM(K + 1)
      IF (KE - KB .LE. 1) GOTO 86
      IF (X(KE) .EQ. X(KB)) GOTO 86
      A = FLOAT(KE - KB)
      CONST = A / (FN * (X(KE) - X(KB)))
      DO 84 KR = KB, KE
      B = FLOAT(KR - KB - 1)
      T = (X(KR) - X(KB)) * CONST - B / FN
      IF (T .GT. TEMP) TEMP = T
   84 CONTINUE
   86 IF (DU .LT. TEMP) DU = TEMP
   88 CONTINUE
C
C        DETERMINE THE CURRENT MAXIMUM
C
   90 DIPNEW = DL
      IF (DU .GT. DL) DIPNEW = DU
      IF (DIP .LT. DIPNEW) DIP = DIPNEW
      LOW = GCM(IG)
      HIGH = LCM(IH)
C
C        RECYCLE
C
      GOTO 40
C
  100 DIP = 0.5 * DIP
      XL = X(LOW)
      XU = X(HIGH)
C
      RETURN
      END
```