Dip Test Distributions, P-values, and other Explorations

Martin Mächler
ETH Zurich

Abstract

...

Keywords: MPFR, Arbitrary Precision, Multiple Precision Floating-Point, R.

1. Introduction

FIXME: Need notation

\[ D_n := \text{dip} \left( \text{runif(n)} \right); \]
but more generally,

\[ D_n(F) := D(X_1, X_2, \ldots, X_n), \quad \text{where } X_i \text{ i.i.d.}, X_i \sim F. \]  

Hartigan and Hartigan (1985) in their “seminal” paper on the dip statistic \( D_n \) already proved that \( \sqrt{n} D_n \) converges in distribution, i.e.,

\[ \lim_{n \to \infty} \sqrt{n} D_n \overset{D}{=} D_\infty. \]  

A considerable part of this paper is devoted to explore the distribution of \( D_\infty \).

2. History of the diptest R package

Hartigan (1985) published an implementation in Fortran of a concrete algorithm, where the code was also made available on Statlib\(^1\)

\(^1\)Statlib is now a website, of course, [http://lib.stat.cmu.edu/](http://lib.stat.cmu.edu/), but then was the preferred way for distributing algorithms for statistical computing, available years before the existence of the WWW, and entailing e-mail and (anonymous) FTP
On July 28, 1994, Dario Ringach, then at NY University, asked on Snews (the mailing list for S and S-plus users) about distributions and was helped by me and then about `dyn.load` problems, again helped by me. Subsequently he provided me with S-plus code which interfaced to (a `f2c`ed version of) Hartigan’s Fortran code, for computing the dip statistic, and ended the (then private) e-mail with

I am not going to have time to set this up for submission to StatLib. If you want to do it, please go ahead.

Regards, Dario

- several important bug fixes; last one Oct./Nov. 2003

However, the Fortran code file [http://lib.stat.cmu.edu/apstat/217](http://lib.stat.cmu.edu/apstat/217), was last changed Thu 04 Aug 2005 03:43:28 PM CEST.

We have some results of the dip.dist of before the bug fix; notably the “dip of the dip” probabilities have changed considerably!!

- see rcs log of .././src/dip.c

### 3. 21st Century Improvement of Hartigan’s Table

(((
Use listing package (or so to more or less “cut & paste” the nice code in .././stuff/new-simul.Rout-1e6))
))

### 4. The Dip in the Dip’s Distribution

We have found empirically that the dip distribution itself starts with a “dip”. Specifically, the minimal possible value of \( D_n \) is \( \frac{1}{2n} \) and the probability of reaching that value,

\[
P \left[ D_n = \frac{1}{2n} \right],
\]

is large for small \( n \).

E.g., consider an approximation of the dip distribution for \( n = 5 \),

```r
R> require("diptest") # after installing it ..
R> D5 <- replicate(10000, dip(runif(5)))
R> hist(D5, breaks=128, main = "Histogram of replicate(10'000, dip(runif(5)))")
```
which looks as if there was a bug in the software — but that look is misleading! Note how the phenomenon is still visible for $n = 8$,

```r
R> D8 <- replicate(10000, dip(runif(8)))
R> hist(D8, breaks=128, main = "Histogram of replicate(10000, dip(runif(8)))")
```

Note that there is another phenomenon, in addition to the point mass at $1/(2n)$, particularly visible, if we use many replicates,

```r
R> set.seed(11)
R> n <- 11
```
Dip Test Distributions, P-values, and other Explorations

R> B.s11 <- 500000
R> D11 <- replicate(B.s11, dip(runif(n)))

Dip D_{11}(U([0, 1])) B = 500'000 replicates

```
1/22
0.05
2/22
0.10
3/22
0.15
1
0.20
```

5. P-values for the Dip Test

Note that it is not obvious how to compute P-values for “the dip test”, as that means determining the distribution of the test statistic, i.e., D_n under the null hypothesis, but a natural null, H_0 : F \in \{ Fcadlag \mid f := \frac{dF}{dx} \text{Fisunimodal} \} is too large. Hartigans’(1985) argued for using the uniform U[0, 1] i.e., F'(x) = f(x) = 1_{[0,1]}(x) = \{0 \leq x \leq 1\} (Iverson bracket) instead, even though they showed that it is not quite the “least favorable” one. Following Hartigans’, we will define the P-value of an observed d_n as

\[ P_{d_n} := P[D_n \geq d_n] := P[\text{dip}(U_1, \ldots, U_n) \geq d_n] \], where U_i \sim U[0, 1], i.i.d. \quad (4) \]

5.1. Interpolating the Dip Table

Because of the asymptotic distribution, \( \lim_{n \to \infty} \sqrt{n} D_n \overset{D}{=} D_\infty \), it is makes sense to consider the “\( \sqrt{n}D_n \)”–scale, even for finite n values:

R> data(qDiptab)
R> dnqd <- dimnames(qDiptab)
R> (nn. <- as.integer(dnqd[["n"]]))
but we can use our table to compute a deterministic (but still approximate, as the table is from simulation too) P-value:

R> ## We are in this interval:
R> n0 <- nn.[i.n <- findInterval(n, nn.)]
R> n1 <- nn.[i.n +1] ; c(n0, n1)
[1] 200 500
R> f.n <- (n - n0)/(n1 - n0)# in [0, 1]
R> ## Now "find" y-interval:
Finally, in May 2011, after several years of people asking for it, I have implemented a `dip.test` function which makes use of a — somewhat more sophisticated — interpolation scheme like the one above, to compute a $P$-value. As `qDiptab` has been based on $10^6$ samples, the interpolation yields accurate $P$-values, unless in very extreme cases. Here is the small ($n = 63$) example from Hartigan:

```R
R> data(statfaculty)
R> dip.test(statfaculty)

Hartigans' dip test for unimodality / multimodality
data:  statfaculty
D = 0.0595, p-value = 0.08672
alternative hypothesis: non-unimodal, i.e., at least bimodal
```

where, from a $P$-value of 8.7%, we'd conclude that there is not enough evidence against unimodality.

### 5.2. Asymptotic Dip Distribution

We have conducted extensive simulations in order to explore the limit distribution of $D_\infty$, i.e., the limit of $\sqrt{n} D_n$, (2).

Our current R code is in `../../stuff/asymp-distrib.R` but the simulation results (7 Megabytes for each $n$) cannot be assumed to be part of the package, nor maybe even to be simply accessible via the internet.

### 6. Less Conservative Dip Testing

### 7. Session Info

```R
toLatex(sessionInfo())
```

- R Under development (unstable) (2014-11-22 r67037), x86_64-unknown-linux-gnu
- Base packages: base, datasets, grDevices, graphics, methods, stats, tools, utils
Other packages: diptest 0.75-6

References


Affiliation:

Martin Mächler
Seminar für Statistik, HG G 16
ETH Zurich
8092 Zurich, Switzerland
E-mail: maechler@stat.math.ethz.ch
URL: http://stat.ethz.ch/people/maechler