Dip Test Explorations

Martin Maechler
Seminar für Statistik
ETH Zurich, Switzerland
maechler@stat.math.ethz.ch

April 2009 (typeset on August 11, 2010)

Abstract

1 Introduction

FIXME: Need notation

D_n :=`dip( runif(n) )`;

but more generally,

$$D_n(F) := D(X_1, X_2, \ldots, X_n), \quad \texttt{where } X_i \text{ i.i.d. }, X_i \sim F. \tag{1}$$

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., $\lim_{n\to\infty} \sqrt{n}\, D_n \overset{\mathcal{D}}{=} D_\infty$.

A considerable part of this paper is devoted to explore the distribution of D_∞.

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]

- MM started in 1994, with S-plus code interfacing to Hartigan's Fortran

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

However, the Fortran code file `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 rcslog of ../../src/dip.c

3 21st Century Improvement of Hartigan^2's Table

((

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

))

[1] Statlib is now a website, of course, `http://lib.stat.cmu.edu/`, but then was *the* preferred way for distributing algorithm for statistical computing, available years before the existence of the WWW, and entailing e-mail and (anonymous) FTP

4 The Dip in the Dip's Distribution

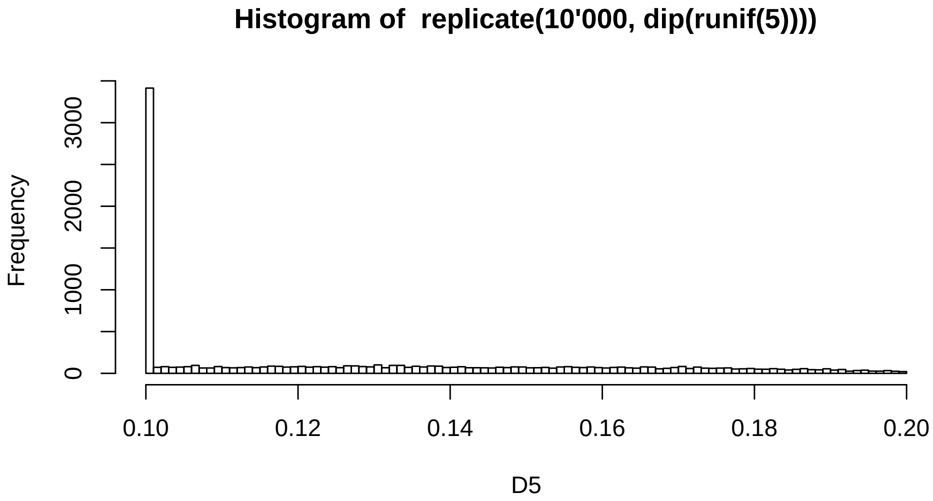
We have found empirically that the dip distribution itself starts with a dip. Specifically, the minimal possible value of D_n is $\frac{1}{2n}$ *and* the probability of reaching that value,

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

is large for small n.

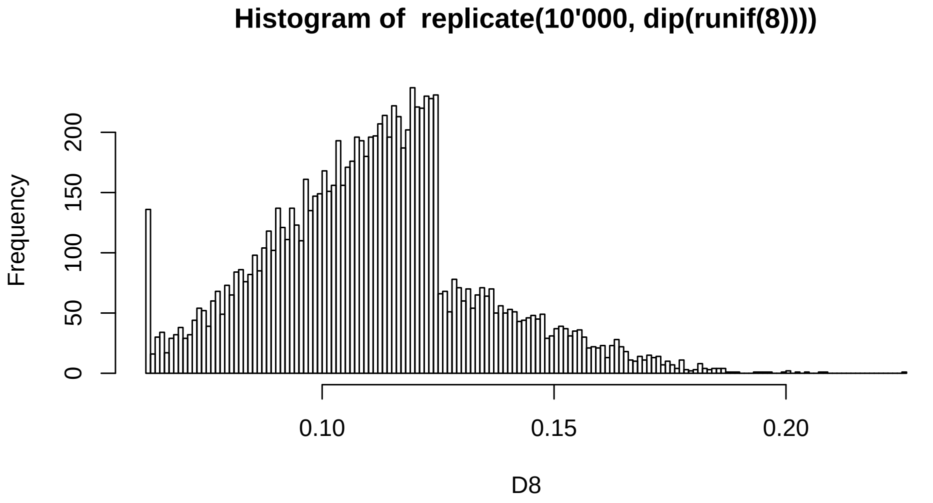
E.g., consider an approximation of the dip distribution for $n = 5$,

```
> D5 <- replicate(10000, dip(runif(5)))
> hist(D5, breaks=128, main = "Histogram of  replicate(10'000, dip(runif(5))))")
```



which looks like there was a bug in the software, and the phenomenon is still visible for $n = 8$,

```
> D8 <- replicate(10000, dip(runif(8)))
> hist(D8, breaks=128, main = "Histogram of  replicate(10'000, dip(runif(8))))")
```



5 P-values for the Dip Test

5.1 Interpolating the Dip Table

5.2 Asymptotic Dip Distribution

6 Less Conservative Dip Testing

7 Session Info

```
> toLatex(sessionInfo())
```

- R version 2.11.1 Patched (2010-08-09 r52694), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=de_CH.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=de_CH.UTF-8`, `LC_MONETARY=C`, `LC_MESSAGES=de_CH.UTF-8`, `LC_PAPER=de_CH.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=de_CH.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: diptest 0.25-3

References

J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Annals of Statistics*, 13:70–84, 1985.

P. M. Hartigan. Computation of the dip statistic to test for unimodality. *Applied Statistics*, 34:320–325, 1985.