18.465, Sept. 19, 2012

UNIMODALITY AND THE DIP STATISTIC

1. Unimodality

Definition. A probability density f defined on the real line $\mathbb{R}$ is called *unimodal* if the following hold:
(i) f is nondecreasing on the half-line $(-\infty, b)$ for some real b;
(ii) f is nonincreasing on the half-line $(a, +\infty)$ for some real a;
(iii) For the largest possible b in (i), which exists, and the smallest possible a in (ii), which also exists, we have $a \leq b$.

Here a largest possible b exists since if f is nondecreasing on $(-\infty, b_k)$ for all k and $b_k \uparrow b$ then f is also nondecreasing on $(-\infty, b)$. Also $b < +\infty$ since f is a probability density. Likewise a smallest possible a exists and is finite.

In the definition of unimodal density, if $a = b$ it is called *the* mode of f. It is possible that f is undefined or $+\infty$ there, or that $f(x) \uparrow +\infty$ as $x \uparrow a$ and/or as $x \downarrow a$.

If $a < b$ then f is constant on the interval (a, b) and we can take it to be constant on $[a, b]$, which will be called the *interval of modes* of f. Any $x \in [a, b]$ will be called *a* mode of f. When $a = b$ the interval of modes reduces to the singleton $\{a\}$.

Examples. Each normal distribution $N(\mu, \sigma^2)$ has a unimodal density with a unique mode at μ. Each uniform $U[a, b]$ distribution has a unimodal density with an interval of modes equal to $[a, b]$. Consider the gamma density $\gamma_a(x) = x^{a-1}e^{-x}/\Gamma(a)$ for $x > 0$ and 0 for $x \leq 0$, where the shape parameter $a > 0$. (The scale parameter $\lambda = 1$ for simplicity.) Differentiating the density for $x > 0$ with respect to x we get $[(a-1)x^{a-2} - x^{a-1}]e^{-x}$ which is 0 if and only if $x = a - 1$, but that is not possible for $x > 0$ if $a \leq 1$. For $a \leq 1$ the derivative is negative, so the density is decreasing for all $x > 0$ and has a unique mode at $x = 0$. We get the standard exponential distribution if $a = 1$. For $0 < a < 1$ the density has a sharp peak with $\lim_{x \downarrow 0} \gamma_a(x) = +\infty$. For $1 < a < \infty$ there is a unique mode at $x = a$.

Definition. A probability distribution P on $\mathbb{R}$ will be called *unimodal* if for some λ with $0 \leq \lambda \leq 1$, $P = \lambda\delta_x + (1-\lambda)Q$ where δ_x is a point mass at x, Q has a density f which is unimodal, and x is in the interval of modes of f.

If we observe real $X_1, \dots, X_n$ i.i.d. from an unknown P, there is a test, called the *dip test*, for whether P is unimodal. We will get to that in Section 4.

2. Convex and concave functions; minorants and majorants

Recall that a real-valued function G on an interval J, which may be a half-line or the whole line, is called *convex* on J if $G((1-t)x+ty) \leq (1-t)G(x)+tG(y)$ for any $x, y \in J$ and $0 \leq t \leq 1$. G is called *concave* if and only if $-G$ is convex.

For a given G defined on an interval J containing points $u < x$, the *chord* of the graph of G between u and x is defined as the line segment of all points

$$((1-t)u+tx, (1-t)G(u)+tG(x)) \ = \ (1-t)(u, G(u))+t(x, G(x))$$

for $0 \leq t \leq 1$. Thus for G to be convex on an interval containing u and x implies that the chord is above, or coincides with, the graph of G on $[u, x]$. Likewise, for a concave G, the chord is below, or coincides with, the graph. A linear function $G(x) \equiv a+bx$ for constants a and b is both convex and concave.

Let G be convex on an interval J containing a point x in its interior. Then for h small enough so that $x+h$ and $x-h$ are in J, and $0 < s < h$, we have the relation

$$\frac{G(x+h)-G(x)}{h} \ \geq \ \frac{G(x+s)-G(x)}{s} \ \geq \ \frac{G(x)-G(x-h)}{h}$$

because the point $(x+s, G(x+s))$ is on or below the chord joining $(x, G(x))$ to $(x+h, G(x+h))$, and the point $(x, G(x))$ is on or below the chord joining $(x-h, G(x-h))$ to $(x+s, G(x+s))$. Thus as $h \downarrow 0$, $(G(x+h)-G(x)/h$ decreases down to a finite limit, called $G'(x+)$, the right derivative of G at x. Similarly, the left derivative $G'(x-)$ of G at x also exists. By similar reasoning it follows that these one-sided derivatives are nondecreasing: if G is convex on an interval containing points $u < v$ in its interior, we have

$$G'(u+) \leq G'(v-) \leq G'(v+). \tag{1}$$

Likewise, a concave function on an interval J will have one-sided derivatives on the interior of J which will be nonincreasing. Existence of the one-sided derivatives implies that a convex or concave function G is continuous on the interior of J.

For example, let $G(0) = G(1) = 1$ and let $G(x) = 0$ for $0 < x < 1$. Then G is easily seen to be convex on the closed interval $J = [0, 1]$. It is continuous on the interior $(0, 1)$ but not at the endpoints.

2.1. **Greatest convex minorants and least concave majorants.**

Let $\mathcal{F}$ be a collection of convex functions on the interval J. Let

$$F_{\mathcal{F}}(x) = F_{\mathcal{F},J}(x) = \sup_{f\in\mathcal{F}} f(x)$$

for all $x \in J$. Also, $f \leq g$ on J will mean that $f(x) \leq g(x)$ for all x in J.

Theorem 1. (*a*) *For any non-empty set $\mathcal{F}$ of convex functions on J, if $F_{\mathcal{F}}$ has finite values, it is also convex on J.*
(*b*) *If for some real-valued function g on J, $f \leq g$ on J for all f in $\mathcal{F}$, then also $F_{\mathcal{F}} \leq g$ on J.*
(*c*) *Let g be a real-valued function of J such that there exists at least one convex function $f \leq g$ on J. Let $\mathcal{F}(g)$ be the set of all convex $f \leq g$ on J. Then $F_{\mathcal{F}(g)}$ is a convex function with $F_{\mathcal{F}(g)} \leq g$ on J.*

Proof. Let $x < y$ in J and $0 < \lambda < 1$. Then for any $f \in \mathcal{F}$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \leq \lambda F_{\mathcal{F}}(x) + (1-\lambda)F_{\mathcal{F}}(y).$$

Taking the supremum over $f \in \mathcal{F}$ on the left we get

$$F_{\mathcal{F}}(\lambda x + (1-\lambda)y) \leq \lambda F_{\mathcal{F}}(x) + (1-\lambda)F_{\mathcal{F}}(y),$$

so indeed $F_{\mathcal{F}}$ is convex on J, proving part (a). Part (b) is immediate. Part (c) then follows from parts (a) and (b). Q.E.D.

The function $F_{\mathcal{F}(g)}$ from part (c) of Theorem 1 is called the *greatest convex minorant* of g on J, abbreviated $\mathrm{GCM}(g) := \mathrm{GCM}(g, J)$. Likewise, for a function g such that there exists at least one concave $h \geq g$ on J, g has a *least concave majorant* $\mathrm{LCM}(g) := \mathrm{LCM}(g, J) \geq g$ on J.

The infimum of any non-empty collection of concave functions, if finite-valued, is concave. But the minimum of two convex functions is not convex in general, nor is the maximum of two concave functions concave (consider linear functions with different slopes).

3. Unimodality, convexity and concavity

The distribution function G of a unimodal distribution will itself be called unimodal. If Q has a unimodal density f, then the distribution function $F_Q(x) = Q((-\infty, x])$ is convex on $(-\infty, y]$ and concave on $[y, +\infty)$ for any y in the interval of modes of f. If P is unimodal with $P = \lambda\delta_x + (1-\lambda)Q$ for $\lambda > 0$, then y must be chosen equal to x.

Conversely, if G is a probability distribution function and for some m, G is convex on $(-\infty, m]$ and concave on $[m, +\infty)$, we'll see that G is unimodal. For any $u < m$, G, being convex in a neighborhood of u, has left and right derivatives $G'(u-)$ and $G'(u+)$ as shown around (1). Let $g(u) := G'(u+)$. This will equal $G'(u-)$ except at most for some sequence of values of u, not affecting integrals, and g will be a nondecreasing function on $(-\infty, m)$ by (1), with $G(u) = \int_{-\infty}^{u} g(v)dv$ for $u < m$. Consistently with convexity on $(-\infty, m]$, G may have a jump at m, say of height λ, with $\lambda = 0$ if G is continuous. Symmetrically, g will be nonincreasing for $u > m$ and we will then have $G(u) = 1 - \int_{u}^{+\infty} g(v)dv$. If $\lambda = 1$ then $g \equiv 0$. Otherwise if $\lambda < 1$ we see that the probability P with distribution function G is unimodal with $P = \lambda\delta_m + (1-\lambda)f$ where $(1-\lambda)f = g$. Because g is nondecreasing on $(-\infty, m)$ and nonincreasing on $(m, +\infty)$, m is a mode of f and P is indeed unimodal.

4. The dip functional, statistic, and test

The *dip functional* is defined by $D(F) = \inf_G \sup_x |(F-G)(x)|$, where the infimum is over all unimodal distribution functions G. Given an empirical distribution function F_n based on observations $X = (X_1, \ldots, X_n)$, the *dip test* for unimodality is based on the *dip statistic* D which is defined as $\mathrm{dip}(X) = D(F_n)$. The hypothesis will be rejected for large enough values of $\mathrm{dip}(X)$.

It will be shown in Proposition 3 that the largest possible value of $D(F_n)$ is 1/4. It occurs for any F_2 with $X_1 \neq X_2$. To see this, let G be unimodal. It must be continuous at X_j for at least one j, say at $X_{(1)}$, where F_2 has a jump of height 1/2 from 0 up to 1/2. To approach F_2 as closely as possible at this point and just left of it we must set $G(X_{(1)}) = 1/4$. We can actually let $G(X_{(2)}-) = 1/2$ and $G(X_{(2)}) = 1$ so that G matches up exactly with F_2 there, but still, we will have $\sup_x |(F_2 - G)(x)| = 1/4$.

To get highly non-unimodal sets of observations, we can take any number of points tightly clustered around 0, and another roughly equal number of points tightly clustered around 1, and few other observations, all in $[0, 1]$. Then to get a unimodal distribution function G as close to F_n as possible, since $F_n(0-) = 0$ and $F_n(x)$ rises to near 1/2 at some small $x > 0$, we'll need to take $G(y)$ about 1/4 for some y with $0 < y < x$ and so y also near 0, and/or $G(v)$ about 3/4 for some v close to 1, so the dip will be close to the maximum of 1/4.

The computed p-values and quantiles available for the dip test are based on Monte Carlo simulations with $X_1, \ldots, X_n$ i.i.d. $U[0, 1]$. This distribution was chosen because, although it's unimodal, the interval of modes is the entire interval on which the density is not zero. Commonly encountered unimodal distributions such as the normal and gamma examples mentioned above are "more unimodal" than $U[0,1]$ in that they have unique modes. Hartigan and Hartigan (1985) conjecture that $U[0,1]$ is the "asymptotically least favorable" unimodal distribution for the dip test, in other words asymptotically (for large n) the most difficult to distinguish from non-unimodal distributions.

5. GCM and LCM of probability distribution functions

On any interval J, if F is a probability distribution function, then since $0 \leq F \leq 1$, the constant 0 is a convex function $\leq F$, and so $\mathrm{GCM}(F) \geq 0$. On the other hand by definition of GCM, we have $\mathrm{GCM}(F) \leq F \leq 1$. Similarly for the LCM, so the GCM and LCM of a distribution function will each take values between 0 and 1.

A distance between bounded functions on J is defined by $\rho_J(f,g) = \sup_{x\in J} |(f-g)(x)|$. Suppose we're given a function F on J and want to approximate it as well as possible with respect to ρ_J by a convex function G. If $\rho_J(F,G) \leq h$ for some constant $h > 0$, then for one thing, $G \leq F + h$ on J. This implies that $G \leq GCM(F+h)$, or since $G - h$ is convex and $G - h \leq F$ that $G - h \leq GCM(F)$ and so $G \leq GCM(F) + h$. On the other hand we want that $G \geq F - h$ everywhere on J. The chances for this are maximized if we take $G = GCM(F+h) = GCM(F) + h$, which will be $\geq F - h$ if and only if if $F - GCM(F) \leq 2h$. Thus the closest approximation to F with respect to ρ_J by a convex G will be found by letting

$$h = \sup_{x\in J}(F - GCM(F,J))(x)/2,$$

which is finite (in fact between 0 and 1/2), and setting $G = GCM(F) + h$.

Symmetrically, to get the closest approximation to a given F on J by a concave function G with respect to ρ_J, set

$$h = \sup_{x\in J}(LCM(F,J) - F)(x)/2$$

and $G = LCM(F) - h$.

If J is a half-line $[x, +\infty)$, and G is the GCM of a probability distribution function F on J, then $G(t) \leq 1$ for all $t > x$. This implies that $G'(t+) \leq 0$, because if $G'(t+) > 0$ for some t, then $G'(u+) \geq G'(u-) \geq G'(t+)$ by (1) for all $u > t$, so $G(u) \to +\infty$ as $u \to +\infty$, which is impossible. So the GCM of F on J would be the constant $G(u) \equiv G(x)$ for $u \geq x$, which is not interesting. Thus the largest intervals on which it's reasonable to take the GCM of a distribution function F are half-lines $(-\infty, x)$ or $(-\infty, x]$.

Similarly, for a probability distribution function F, it's reasonable to consider its least concave minorant LCM(F) on a half-line $[m, +\infty)$ or a subinterval of such an interval. We saw in Section 3 that the distribution function G of a unimodal distribution is convex on some half-line $(-\infty, y]$ and concave on the half-line $[y, +\infty)$.

Recall that a distribution function F is right-continuous. It has limits from the left defined as $F(x-) = \lim_{y\uparrow x} F(y)$.

5.1. **The GCM of an empirical distribution function on a left half-line.** Let's evaluate the GCM G of the empirical distribution function F_n on a half-line $(-\infty, X_{(k)})$ for any k with $1 \le k \le n$. we clearly have $G \equiv 0$ on $(-\infty, X_{(1)})$. For $1 \le r < k$, if G has been found on $(-\infty, X_{(r)})$, for $X_{(r)} = X_{(k)}$, we are done. If $X_{(r)} < X_{(k)}$, consider the line segments joining $(X_{(r)}, F_n(X_{(r)}-))$ to $(X_{(j)}-, F_n(X_{(j)}-))$ with $X_{(r)} < X_{(j)} \le X_{(k)}$. We must have $G(X_{(i)}-) \le F_n(X_{(i)}-)$ for each i such that $X_{(r)} \le X_{(i)} \le X_{(k)}$. Therefore, the graph of G must be below or equal to the value on the line segment for each of these line segments. Since all the segments have left endpoint $X_{(r)}$, the most restrictive condition between $X_{(r)}$ and the next larger order statistic (which may be $X_{(r+1)}$, or may not in case of ties) comes from the line segment with smallest slope. If this slope occurs for more than one value of j, let j' be the largest value of j giving this smallest slope with $X_{(j')} \le X_{(k)}$. The graph of G on $[X_{(r)}, X_{(j')})$ must be the corresponding line segment. If $X_{(j')} < X_{(k)}$, iterate with the new $r = j'$. We find that $G = \mathrm{GCM}(F_n, (-\infty, X_{(k)}))$ is piecewise linear with changes of slope only at some points $X_{(j)}$ where the slope increases.

The evaluation of the LCM of F_n on $[X_{(k)}, +\infty)$ is symmetric, but with somewhat simpler expressions since F_n is right-continuous and we only need to consider values $F_n(X_{(j)})$ and not left limits.

To find the GCM and LCM of F_n on appropriate half-lines there is an algorithm, which can be computed (P. M. Hartigan, 1985; Maechler, 2004, 2009, 2010) but seems hard to describe in closed form.

6. Lower and upper bounds for the dip statistic

For any $X_1, \ldots, X_n$ with order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ the *spacings* s_j are defined as $X_{(j)} - X_{(j-1)}$ for $j = 2, \ldots, n$. First, let's consider lower bounds, depending on n. We have the following:

Proposition 1. *For any $n \ge 2$ and $x = (X_1, \ldots, X_n)$ such that not all X_j are equal, the dip statistic dip$(x) \ge 1/(2n)$. This value occurs* *when the spacings s_j are nonincreasing in j for $j = 2, \dots, k$ for some k and then nondecreasing in j for $j = k, \dots, n$. It is possible that $s_j = 0$ for some $j_0 \le j \le j_1$ where either $j_0 > 2$ or $j_1 < n$.*

Remarks. It seems natural that if the spacings decrease, so that points get closer and closer together, then increase, so that points get farther and farther apart, the sample is behaving in a very unimodal way.

Recall that for $0 < y < 1$, $F^{\leftarrow}(y) := \inf\{x : \ F(x) \ge y\}$. The minimum value $1/(2n)$ of $\mathrm{dip}(F_n)$ occurs with substantial probability for small n and X_j i.i.d. $U[0,1]$. As seen in the (non-adjusted) dip test quantile table, it happens with probability larger than 0.5 for $n = 4$, larger than 0.1 for $n = 5$ and 6, and larger than 0.01 for $n = 7$ and 8. In other words, in the $N = 10^6 + 1$ replications done by Maechler (2004), where the observations were N independent dip statistics, each based on n i.i.d. $U[0,1]$ observations for the given values of n, letting $F_{N,n}$ be the empirical distribution function of the N dip statistics for a given n, it was found that $F^{\leftarrow}_{N,4}(0.5) = 1/8$, $F^{\leftarrow}_{N,5}(0.1) = 1/10$, $F^{\leftarrow}_{N,6}(0.1) = 1/12$, $F^{\leftarrow}_{N,7}(0.01) = 1/14$, $F^{\leftarrow}_{N,8}(0.01) = 1/16$. Hartigan and Hartigan (1985), in their Table 1, found the same for their $N = 10^4 - 1$ replications. (From the full qDiptab table, in the diptest library, one might say a little more.) When a quantile equal to $1/(2n)$ appears in the q column of the dip test quantile table, then the p-value is not the usual $1 - q$ but rather 1.

For $n = 3$ the condition on spacings mentioned in the proposition will always hold vacuously, so the dip statistic will always equal 1/6, except in the extreme case when all the X_i are equal, when the dip is 0. For X_i i.i.d. $U[0,1]$, the dip will equal 1/6 with probability 1. So the dip test can only work for $n \ge 4$, which is why quantiles are only given for such n.

Proof. Since not all X_j are equal and any unimodal distribution function G has at most one discontinuity, there must be at least one order statistic $X_{(j)}$ at which G is continuous. We have for the empirical distribution function F_n based on $X_1, \dots, X_n$ that $F_n(X_{(j)}) \ge j/n$ and for the left limit there, $F_n(X_{(j)}-) \le (j-1)/n$. (In case there is just one X_i equal to $X_{(j)}$, both these inequalities become equalities.) So F_n has a jump of height at least $1/n$ at $X_{(j)}$. (If the X_i are all different, all these jumps are of height exactly $1/n$.) To approach F_n as closely as possible at $X_{(j)}$ and just below it, $G(X_{(j)})$ must equal $(F_n(X_{(j)}) + F_n(X_{(j)}-))/2$ and $\sup_x |(F_n - G)(x)| \ge 1/(2n)$ from values at and just below $X_{(j)}$.

Now suppose the spacings s_j are nonincreasing for $2 \le j \le k$ and nondecreasing for $k \le j \le n$. First suppose all the X_i are distinct, so that all spacings $s_j > 0$ for $j = 2, \ldots, n$, as will occur with probability 1 if X_i are i.i.d. with a continuous distribution such as $U[0, 1]$. In this case $F_n(x) = j/n$ for $X_{(j)} \leq x < X_{(j+1)}$ and $j = 1, \ldots, n-1$. As always, $F_n(x) = 0$ for $x < X_{(1)}$ and $F_n(x) = 1$ for $x \geq X_{(n)}$. Set $G(X_{(j)}) = (j - \frac{1}{2})/n$ for $j = 1, \ldots, n$. Let G be linear on each interval $[X_{(j-1)}, X_{(j)}]$ for $j = 2, \ldots, n$, where it will have slope $1/(ns_j)$. For $2 \leq j \leq k$ these slopes will be nondecreasing in j (vacuously if $k = 2$) since s_j are nonincreasing. Thus G will be convex on $[X_{(1)}, X_{(k)}]$. Extend the graph of G from $[X_{(1)}, X_{(2)}]$ to the left along the same line until it crosses the x axis at a point $(x_0, 0)$. Let $G(x) = 0$ for $x \leq x_0$. Then G will be convex on $(-\infty, X_{(k)}]$.

For $k \leq j \leq n$ the slopes will be nonincreasing since s_j are nondecreasing. Thus G will be concave on $[X_{(k)}, X_{(n)}]$. Likewise we can extend it to be concave on $[X_{(k)}, +\infty)$. We will have $\sup_x |(F_n - G)(x) = 1/(2n)$, on each interval $[X_{(j-1)}, X_{(j)}]$, on $(-\infty, X_{(1)})$, and on $(X_{(n)}, +\infty)$, and so on the whole line, finishing the proof when all $s_j > 0$.

If some $s_j = 0$ then by the assumptions, $s_j = 0$ if and only if $j_0 \leq j \leq j_1$ for some j_0, j_1 with $2 \leq j_0 \leq j_1 \leq n$ where either $2 < j_0$ or $j_1 < n$ (or both) since by assumption $s_j > 0$ for some j. Here $j_0 \leq k \leq j_1$.

Take the largest $i \leq k$ such that $s_i > 0$, if such an i exists. Then $i = j_0 - 1$ and $s_j > 0$ for $2 \leq j \leq i$. For $j = 1, \ldots, i-1$, $F_n(X_{(j)}) = j/n$ and $F_n(X_{(j)}-) = (j-1)/n$. Next,

$$X_{(j_0-1)} = X_{(j_0)} = \cdots = X_{(j_1)} = X_{(k)}.$$

Let their common value be m. We will define G to be convex on $(\infty, m]$ and concave on $[m, +\infty)$, with a jump at m. We will have $F_n(m) = j_1/m$ and $F_n(m-) = (j_0 - 2)/n$.

If $j_0 = 2$, let $G(x) = 0$ for $x < m$. Then G will be convex on $(-\infty, m]$ for any choice of $G(m) \geq 0$, and $G(x) = F_n(x) = 0$ for $x < m$.

If $j_0 \geq 3$, then i exists and equals $j_0 - 1$. As in the earlier part of the proof, let $G(X_{(j)}) = (j - \frac{1}{2})/n$ for $1 \leq j \leq i - 1$. Now set $G(m-) = (i - \frac{1}{2})/n$. Let G be linear on the closed intervals $[X_{(j-1)}, X_{(j)}]$ for $j = 2, \ldots, n-1$ (if any exist) and on the half-open interval $[X_{(i-1)}, m)$ Extend G to the left of $X_{(1)}$ as before. It will be convex on $(-\infty, m)$, and on $(-\infty, m]$ if $G(m)$, to be defined later, satisfies $G(m) \geq G(m-)$. Also, $\sup_{x<m} |(F_n - G)(x)| = 1/(2n)$ as in the case where all $s_j > 0$.

If $j_1 = n$, set $G(x) = 1$ for $x \geq m = X_{(n)}$. Then G is concave and equal to F_n on $[m, +\infty)$.

If $j_1 < n$, set $G(X_{(j)}) = (j - \frac{1}{2})/n$ for $j_1 < j \leq n$ and $G(m) = (j_1 - \frac{1}{2})/n$. Then G will be concave on $[m, \infty)$, with $\sup_{x \geq m} |(F_n - G)(x)| =$

$1/(2n)$, by symmetry to the case $j_0 \geq 3$ and similarly as in the case that all $s_j > 0$.

We need that $G(m) = (j_1 - \frac{1}{2})/n \geq G(m-) - (j_0 - \frac{3}{2})/n$, which is true. In fact at m, G will have a jump of height $(j_1 - j_0 + 1)/n \geq 1/n > 0$. The G as defined will have the desired properties including $\sup_x |(F_n - G)(x)| = 1/(2n)$, Q.E.D.

Recall that the *empirical measure* P_n based on given $X_1, \ldots, X_n$ is $\frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$, so that $F_n(x) = P_n(-\infty, x])$ for each x. An *atom* of P_n is an x such that $P_n(\{x\}) > 0$. Clearly, the atoms are the points $X_1, ..., X_n$, but there may be fewer than n distinct atoms in case of ties. Let m be the number of atoms. Let the distinct atoms be $y_1, ..., y_m$. Then the "sizes" $P_n(y_j)$ of the atoms have order statistics $p(1) \leq p(2) \leq \cdots \leq p(m)$. If $m = 1$ (all the observations are equal), let $p(m-1) = p(0) = 0$.

The dip is always at least $p(m-1)/2$. If there are no ties, then $m = n$, all $p(j) = 1/n$, and the following is a consequence of Proposition 1. If there are ties, then the proof is similar.

Proposition 2. *For any sample $x = (X_1, ..., X_n)$, we have dip$(x) \geq p(m-1)/2$.*

Proof. If $m = 1$, the inequality holds trivially (actually the dip equals 0). So let $m \geq 2$. We can assume that $P_n(y_j) = p(j)$ for $j = 1, ..., m$. A unimodal distribution function G must be continuous at at least one of y_{m-1} and y_m. So, as in the proof of Proposition 1,

$$\sup_x |(F_n - G)(x)| \;\geq\; \min(p(m-1), p(m))/2 = p(m-1)/2,$$

Q.E.D.

For any probability distribution P on the real numbers with distribution function F, *a median* of P or F is an x such that $F(x) \geq 1/2$ and also $P([x, +\infty)) \geq 1/2$ so that the left limit $F(x-) := \lim_{u \uparrow x} F(u) \leq 1/2$. If there is only one median it is called *the* median. If the median by the definitions so far is not unique, then there is an interval $[a, b]$ of medians and *the* median in that case will be defined as $(a + b)/2$. Thus if F_n is the empirical distribution function of a sample of size n, if $n = 2k + 1$ odd then the (sample) median is $X_{(k+1)}$, or if $n = 2k$ even then the (sample) median is $[X_{(k)} + X_{(k+1)}]/2$.

If the median of F is unique, it equals $F^{\leftarrow}(1/2)$. If there is an interval $[a, b]$ of medians with $a < b$ then $F^{\leftarrow}(1/2) = a$ (the smallest median). In the paper by Hartigan and Hartigan (1985), p. 78 (A) the following is stated and the idea of a proof is stated. A fuller and somewhat different proof will be given.

Proposition 3. *For any distribution function F, the dip functional $D(F) \le 1/4$. In particular for any finite sample (whose members may be distinct or not), the dip statistic is at most* $1/4$.

Proof. Let m be the median of F. A $\delta > 0$ will be chosen. Let $x_1 := F^{\leftarrow}(1/4)$. Define $x_0 := x_1 - 1/(4\delta)$. Let $x_4 := F^{\leftarrow}(3/4)$ and $x_5 := x_4 + 1/(4\delta)$. A distribution function G will be defined which will have a jump at m. Elsewhere G will have a density g which will equal δ on (x_0, m) and on (m, x_5) and be 0 elsewhere. Then by choice of x_0, G will increase from 0 at x_0 up to 1/4 at x_1- (at x_1 also, unless possibly if $x_1 = m$ which can happen). We will take $\delta > 0$ small enough so that $\delta(m - x_1) < 1/4$. It follows that $|(F - G)(x)| \le 1/4$ for $-\infty < x < x_1$ and for $x_1 \le x < m$. Symmetrically, taking δ also small enough so that $\delta(x_4 - m) < 1/4$, we will have $|(F - G)(x)| \le 1/4$ for $m \le x < x_4$ and for $x_4 \le x < \infty$.

From the definitions so far we will have $G(m) > G(m-)$, which implies that G must have a jump of height $G(m) - G(m-)$ at m, in other words the corresponding probability P has an atom of size $G(m) - G(m-)$ at m. We see that δ can be chosen small enough, satisfying just the two conditions put on it, and then the resulting distribution is unimodal and G is everywhere within 1/4 of F, Q.E.D.

Remark. The Hartigans seem to claim (p. 78, (A)) that we can take $G(m) = 3/4$ and $G(m-) = 1/4$ ("symmetric about the median... ...with an atom of size 1/2 at the median"), but I disagree in case F has an interval of medians $[a, b]$ with $a < b$, e.g. for a sample of even size, so the median of F is $(a+b)/2$. Then $F(a) = 1/2$ but $G(a) < 1/4$ so $|(F - G)(a)| > 1/4$.

Proposition 4. *For $n = 5$ and any sample $X_1, ..., X_5$ of 5 distinct real numbers, the dip $dip(X) \le 1/5$.*

Proof. The distribution function G of a probability distribution P will be defined to have a jump of height 1/5 at the sample median $X_{(3)}$ with $G(X_{(3)}) = 3/5$ and $G(X_{(3)}-) = 2/5$. Choose $\delta > 0$ small enough such that $\delta(X_{(5)} - X_{(1)}) < 0.2$. Set $X_{(0)} := X_{(1)} - 1/(5\delta)$ and $X_{(6)} := X_{(5)} + 1/(5\delta)$. A density g will be defined with $g(x) = \delta$ for $X_{(0)} < x < X_{(2)}$ or $X_{(4)} < x < X_{(6)}$, and $g(x) = 0$ for $x < X_{(0)}$ or $x > X_{(6)}$. Also, g will have a constant value γ_1 on $(X_{(2)}, X_{(3)})$, which must be

$$\gamma_1 = \frac{0.2 - \delta(X_{(2)} - X_{(1)})}{X_{(3)} - X_{(2)}}$$

in order that $G(X_{(3)}-) = 0.4$, and a constant value γ_2 on $(X_{(3)}, X_{(4)})$, which must be

$$\gamma_2 \ = \ \frac{0.2 - \delta(X_{(5)} - X_{(4)})}{X_{(4)} - X_{(3)}}$$

in order that $G(X_{(3)}+) = 0.6$ so that G is continuous from the right at $X_{(3)}$. By choice of δ, we have $\delta(X_{(3)} - X_{(1)}) < 0.2$ and $\delta(X_{(5)} - X_{(3)}) < 0.2$, from which it follows that $\min(\gamma_1, \gamma_2) > \delta$. Thus $g = 0.8f$ for a unimodal probability density f, so P is unimodal. We have $|(F_5 - G)(x)| \leq 1/5$ for all x similarly as in the previous proof. Q.E.D.

Remarks. For $n = 6$, we see that the dip statistic not only can be larger than $1/n = 1/6$, unlike for $n = 4$ or 5, but it can be larger than 1/5, specifically from the dip test quantile table (without adjustment) it was larger than 0.202 in about 5000 of Maechler's $10^6 + 1$ Monte Carlo simulations. This is an exception to the general pattern that the quantiles for a given q decrease as n increases. The Hartigans noticed this in the smaller simulations for their paper and repeated the simulation for $n = 5$ to confirm the results. The situation for $n = 5$ is explained by Proposition 4.

7. A further definition and theorems in the Hartigans' paper

The paper by Hartigan and Hartigan (1985), beside defining the dip statistic and test for the first time, proved several theorems about it. Let $\mathcal{W}$ be the class of all unimodal probability distribution functions, i.e. all distribution functions G such that for some m, G is convex on $(-\infty, m]$ and concave on $[m, +\infty$. Let $\rho(H, J) := \sup_x |(H - J)(x)|$ for any two bounded real functions H and J on $\mathbb{R}$. The dip functional defined above for a distribution function F was $D(F) = \inf_{G \in \mathcal{W}} \rho(F, G)$. The Hartigans defined an extended dip functional that I will call $DD(H)$ as follows. Let $\mathcal{V}$ be the class of functions f constant on $(-\infty, 0]$, with a possibly different constant value on $[1, +\infty)$, and such that for some $m \in [0, 1]$, f is convex on $[0, m]$ and concave on $[m, 1]$. Note that f need not be convex on $(-\infty, m]$ nor concave on $[m, +\infty)$. For a bounded function H from the real line into itself, define a functional $DD(H) = \inf_{g \in \mathcal{V}} \sup_x |(H - g)(x)|$. In this notation the Hartigans' theorems are as follows. Their Theorem 1 is:

Theorem 2. *If F is a distribution function with $F(0) = 0$ and $F(1) = 1$, then $DD(F) = D(F)$.*

Remark. Here F may be for example an empirical distribution function $\mathcal{U}_n$ for the $U[0,1]$ distribution, as used in setting the quantiles for the dip test.

Proof. For a given $G \in \mathcal{W}$, let it be convex on $(-\infty, m]$ and concave on $[m, +\infty)$. Let

$$H(x) = G(0)1_{x<0} + G(x)1_{0\le x\le 1} + G(1)1_{x>1}.$$

Then H has the constant value $G(0)$ for $x \le 0$ and the constant value $G(1)$ for $x \ge 1$. If $m < 0$, define $m_H = 0$. If $m > 1$, define $m_H = 1$. If $0 \le m \le 1$ define $m_H = m$. Then H is convex on $[0, m_H]$, because G is if $m_H > 0$, or trivially if $m_H = 0$. Likewise H is concave on $[m_H, 1]$, so $H \in \mathcal{V}$.

We have $\rho(F,H) = \sup_{0\le x\le 1} |(F-H)(x)|$, because for $x < 0$, $(F-H)(x) = (F-H)(0)$ and for $x > 1$, $(F-H)(x) = (F-H)(1)$. We also have

$$\sup_{0\le x\le 1} |(F-H)(x)| = \sup_{0\le x\le 1} |(F-G)(x)| \le \rho(F,G).$$

It follows that $DD(F) \le D(F)$.

We need to prove conversely that $D(F) \le DD(F)$. Recall that by Proposition 3, $D(F) \le 1/4$ for any F. Thus in the infimum defining $DD(F)$ we need only consider $G \in \mathcal{V}$ satisfying $\rho(F,G) \le 1/4$. In particular we will then have $c := G(0)$ and $d := G(1)$ satisfying $|c| \le 1/4$ and $|1-d| \le 1/4$, so $c \le 1/4 < 3/4 \le d$.

Now for such a $G \in \mathcal{V}$ let it be convex on $[0,m]$ and concave on $[m,1]$. Define

$$H(x) = c1_{G(x)<c} + G(x)1_{c\le G(x)\le d} + d1_{G(x)>d}.$$

Then evidently $c \le H(x) \le d$ for all x. We have the following lemma:

Lemma 1. *For the function H just defined,*
(*a*) *H is nondecreasing;*
(*b*) *$H \in \mathcal{V}$.*

Proof. Let $\xi := \sup\{x :\ G(x) \le c\}$ and $\eta := \inf\{x :\ G(x) \ge d\}$. Then because $c < d$, we have $\xi \le 1$ and $\eta \ge 0$. The following will be proved.

Claim (*i*): $H(x) = c$ for $-\infty < x < \xi$;

Claim (*ii*): $H(x) = d$ for $\eta < x < +\infty$;

Claim (*iii*): $\xi \le \eta$;

Claim (*iv*): For $\xi < x < \eta$ we have $c < H(x) = G(x) < d$.

To prove Claim (i), suppose it fails. Then there exist $0 < x < y \leq \xi \leq 1$ with $G(y) \leq c$, so $y < 1$, and $H(x) > c$. Then $G(x) > c$ because in this case $H(x) = \min(G(x), d)$. It follows that G can't be convex on $[0, y]$, so $m < y$. If $H(m) > c$ then as before $G(m) > c$, and G is concave on $[m, 1]$ but $G(y) < \min(G(m), G(1))$ violates this, a contradiction. So $H(m) \leq c$ and $H(m) = c$. Thus $G(m) \leq c$ and because G is convex on $[0, m]$ and $G(0) = c$, we have $G(x) \leq c$ and $H(x) = c$ for $0 \leq x \leq m$. Since $m < y < 1$, $G(1) = d$, and G is concave on $[m, 1]$, we have

$$G(y) \geq \frac{y-m}{1-m}c + \frac{1-y}{1-m}d > c,$$

contradicting $G(y) \leq c$. Thus Claim (i) is proved by contradiction.

Claim (ii) is proved symmetrically.

Claim (iii) then is immediate. Claim (iv) follows from the other claims and the definitions of ξ, η, and H. So all four claims hold.

Define $m_H = \xi$ if $m \leq \xi$, $m_H = \eta$ if $m > \eta$, and $m_H = m$ if $\xi < m < \eta$. To prove part (a) of the lemma, first suppose $\xi = \eta$. Then for $x < \xi < y$ we have by the Claims $H(x) = c \leq H(\xi) \leq H(y) = d$, so H is nondecreasing. So let $\xi < \eta$. We have $H(\xi-) = c \leq H(\xi) \leq H(\xi+)$ where the first inequality is clear and both inequalities hold with equality unless $m = \xi$, in which case G is concave on $[\xi, 1]$ and if it has a jump at its left endpoint it must be a jump upward, so the second inequality holds. Symmetrically, $H(\eta-) \leq H(\eta) \leq H(\eta+) = d$. For $\xi < x < \eta$, where $c < H(x) \equiv G(x) < d$ by Claim 4, suppose there is an x at which $H'(x+) = G'(x+) < 0$. If $x < m$, then by convexity of G on $[0, m]$ and so on (ξ, m), $G'(x+)$ is nondecreasing there, so it is just as negative on $[\xi, x)$. In this case G is continuous at ξ, and so $c = G(\xi) > G(x) > c$, a contradiction. There is a symmetrical contradiction if $x > m$. If $x = m$, we must have $G(m-) \leq G(m) \leq G(m+)$ because if G, convex on $[0, m]$, has a jump at m, it must be a jump upward, and likewise for G concave on $[m, 1]$. So, $H'(x+) = G'(x+) \geq 0$ for all $x \in (\xi, \eta)$ where it is defined (everywhere except possibly at m), and jumps if any are upward, so part (a) is proved.

To prove part (b), if $m_H = \xi$, then since H is constant on $(-\infty, m_H)$ with value c by Claim 1, and also $H(\xi) \geq c$, H is convex on $(-\infty, m_H]$ and in particular on $[0, m_H]$. On $(m_H, 1]$ we have $H > c$ and so $H = \min(G, d)$, the minimum of two concave functions there, which is concave. As $H(\xi+) \geq H(\xi)$, H is concave on $[m_H, 1]$. We have a symmetric proof of the desired properties of H if $m_H = \eta$. So suppose $\xi < m = m_H < \eta$. Then by Claim 4, $c < H(m-) \leq H(m) \leq H(m+) < d$.

As G can have a possible jump only at m, it is continuous, and so is H, at ξ and at η. Now G is convex on $[0, m]$ and so on $[\xi, m]$. As H is nondecreasing and $H(m) < d$ we have that $H = \max(c, G)$ on $(-\infty, m]$. Also, $G'(\xi+) = H'(\xi+) \geq 0$. Thus the two functions $H \equiv c$ on $(-\infty, \xi]$ and H on $[\xi, m]$, each convex on their domains, fit together to form a convex function on $(-\infty, m]$ where $m = m_H$ and so on $[0, m_H]$. Symmetrically, H is concave on $[m, +\infty)$ and so on $[m, 1]$, so $H \in \mathcal{V}$ and part (b) and the Lemma are proved. Q.E.D.

Next is

Claim (*v*). We can assume that $0 \leq c \leq 1/4$ and $3/4 \leq d \leq 1$.

To prove this claim, we already saw we can assume $|c| \leq 1/4$ and $3/4 \leq d \leq 5/4$. Let $H_1 = \max(H, 0)$. Then H_1 is clearly nondecreasing. It has the properties of H with the following changes if $c < 0$: ξ is now replaced by $\xi_0 = \sup\{x : \ G(x) \leq 0\}$, and $m_{H_1} = \xi_0$ if $m_H < \xi_0$, otherwise $m_{H_1} = m_H$, and H_1 is convex on $[0, m_{H_1}]$ and concave on $[m_{H_1}, 1]$. Symmetrically, we replace H_1 by $H_2 = \min(H_1, 1)$ with η replaced by $\eta_1 = \inf\{x : \ G(x) \geq 1\}$ and a corresponding definition of H_2. Then $H_2 \in \mathcal{V}$, $0 \leq H_2 \leq 1$, and clearly $\rho(F, H_2) \leq \rho(F, H)$, so Claim (v) is proved.

Another step is

Claim (*vi*). $\rho(F, H) \leq \rho(F, G)$.

To prove this, for any $x \leq 0$ we have $|(F - H)(x)| = |(F - H)(0)| = |(F - G)(0)|$ and likewise for $x \geq 1$, $|(F - H)(x)| = |(F - G)(1)|$. For $0 < x < 1$, now that $c \geq 0$ and $d \leq 1$ by Claim v, consider $|(F - H)(x)|$, which equals $|(F - G)(x)|$ except in two cases. One is $G(x) < c$, in which case it equals $|F(x) - c|$, which if $F(x) \geq c$ is clearly $\leq (F - G)(x)$, whereas if $F(x) < c$ it equals $c - F(x) \leq c = (G - F)(0)$. We have a symmetrical proof if $G(x) > d$, so the claim is proved.

For any $a \geq 1$, let $J = (-\infty, m)$ and on J let

$$G_a := GCM(1_{\{x \geq -a\}} H, J).$$

Define G_a as $LCM(1\{x \leq a\}H)$ on $[m, +\infty)$. Then $G_a(m) \geq H(m) \geq H(m-) \geq G_a(m-)$, so $G_a \in \mathcal{W}$ with the given m. We have necessarily $G_a(x) = 0$ for all $x < -a$ and since G_a is continuous in the interior of J, also $G_a(-a) = 0$. We must have $G_a(x) \leq H(x)$ for $0 \leq x \leq m$, and since G_a is convex, for $0 < \lambda < 1$ we must have $G_a(-a(1 - \lambda) + \lambda x) \leq \lambda H(x)$. In other words the graph of G_a on $[-a, x]$ must be below the straight line segment joining $(-a, 0)$ to $(x, H(x))$. These line segments have slopes depending continuously on x for $-a < x < m$. It will be shown that the infimum of their slopes is attained at some $x_1 \in [0, m]$, if we allow also left limits x_1-. If some slope is 0, as it is if and only if $c = 0$, then the minimum is 0 and attained at $x_1 = 0$. If $c > 0$, the slope goes to $+\infty$ as $x \downarrow -a$. We have $x_1 \geq 0$. If the slopes approach their minimum as $x \uparrow m$ but do not attain it at m, H must have a jump upward at m and we set $x_1 = m$ and use $m- = x_1-$. Once x_1 is chosen, the graph of G_a on $[-a, x_1]$ will be the above-mentioned straight line segment for $x = x_1$. For $c > 0$, it's possible that x_1 is not unique, as the minimum slope s may may be attained for all x in some interval $[u, v]$ with $0 \leq u < v < m$.

For $x_1 \leq x < m$ we simply have $G_a = H$. This is vacuous if $x_1 = m$. If $x_1 < m$, H on $[x_1, m)$ is convex, so it is its own GCM, and we can join the straight line segment to this function and preserve convexity, because $H'(x_1+)$ is at least equal to the slope of the line segment, otherwise the graph of H for x a little larger than x_1 would go below the extended line segment and we would contradict the choice of x_1.

Now let $a \to +\infty$. Then the slope s of the line segment which is the graph of G_a on $[-a, x_1]$ and so on $[-a, 0]$ will approach 0 because it is $\leq c/a$. Since $G'(x_1-) \leq s$, x_1 will decrease down to $\xi = \xi_0$, and $\sup_{x<m} |(G_a - H)(x)| \to 0$. By symmetry the same will occur for $x > m$. If F has a jump at m then we will want to have chosen G to be right-continuous at m. Then we will have

$$\rho(F, G_a) = \sup_{0 \leq x \leq 1} |(G_a - F)(x)| \to \sup_{0 \leq x \leq 1} |(H - F)(x)| = \rho(F, H),$$

which implies that $D(F) \leq DD(F)$ and finishes the proof of Theorem 2. Q.E.D.

The following theorem is easy. Unlike the Hartigans' statement, here $\beta \geq 0$ is not required. The following is otherwise their Theorem 2.

Theorem 3. *If H is a bounded function, constant on $(-\infty, 0]$ and constant on $[1, \infty)$, and $\mathcal{U}$ is the $U[0, 1]$ distribution function, then for any $\alpha \geq 0$ and any real β, $DD(\alpha H + \beta \mathcal{U}) = \alpha DD(H)$.*

Proof. We have by definition $DD(\alpha F + \beta \mathcal{U}) = \inf_{G \in \mathcal{V}} \rho(\alpha F + \beta \mathcal{U}, G)$, which equals $\inf_{\gamma \in \mathcal{V}} \rho(\alpha F + \beta \mathcal{U}, \alpha\gamma + \beta \mathcal{U})$ because for any real β and $\alpha > 0$, $\gamma \in \mathcal{V}$ if and only if $\alpha\gamma + \beta\mathcal{U} \in \mathcal{V}$. To see this, first, multiplication by $\alpha > 0$ preserves $\mathcal{V}$. Second, so does adding any constant times $\mathcal{U}$, which preserves constancy for $x \leq 0$ and for $x \geq 1$, and also does not change convexity on $[0, m]$ nor concavity on $[m, 1]$. If $\alpha = 0$ then all the quantities in the equations shown are 0 so the equations hold. Q.E.D.

Next is the Hartigans' Theorem 3:

Theorem 4. *For the empirical distribution functions $\mathcal{U}_n$ of $U[0,1]$ and a Brownian bridge B, $\sqrt{n}D(\mathcal{U}_n) \to DD(B)$ in distribution as $n \to \infty$.*

Proof. By Theorem 2, $D(\mathcal{U}_n) = DD(\mathcal{U}_n)$. Next, DD is homogeneous under multiplication by positive constants, so $\sqrt{n}DD(\mathcal{U}_n) = DD(\sqrt{n}\mathcal{U}_n)$, which by Theorem 3 equals $DD(\sqrt{n}(\mathcal{U}_n - \mathcal{U}))$. Next one applies a theorem on approximation of the empirical process $\sqrt{n}(\mathcal{U}_n - \mathcal{U})$ by Brownian bridge(s). Specifically, one can use the Komlós–Major–Tusnády theorem, in the Bretagnolle–Massart form, which appeared in 1989 (after 1985). If $\rho(H,J) < \delta$ for some H and J and $\delta > 0$, it's easily seen that $|DD(H) - DD(J)| < \delta$. The theorem follows, Q.E.D.

As the sample functions of the Brownian bridge are bounded (in fact continuous) with probability 1, $DD(B)$ will be a finite, well-defined random variable, although its distribution might be hard to find in closed form. Thus one can expect quantiles of $\sqrt{n}D(\mathcal{U}_n)$ to converge as $n \to \infty$.

8. Notes

The $GCM(F_n)$ and $LCM(F_n)$ on different intervals or half-lines are used in the proof of Theorem 2 (their Theorem 1) in Hartigan and Hartigan (1985) and also in the actual computation of dip statistics, for which P. Hartigan (1985) gave an algorithm. Maechler (2004) gives some documentation on obtaining relevant GCMs and LCMs for given data sets x via the dip(x,full.result=TRUE) option.

9. References

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

Hartigan, P. M. (1985). Computation of the dip statistic to test for unimodality. *Applied Statistics — J. Roy. Statist. Soc. Ser. C* **34**, 320-325.

Maechler, Martin (2012) (previous versions 2004, 2009, 2010). Package 'diptest.'
cran.r-project.org/web/packages/diptest/diptest.pdf