18.465 PS3 DUE FRIDAY FEB. 27, 2015

For this problem set you will need to be using the "diptest" library. After you get into R as usual you should type at an R prompt > library(diptest). The main function in the library, "dip", finds the dip of a data vector such as x = c($X_1, ..., X_n$). This is a test statistic for the hypothesis that the X_j are sampled from a unimodal distribution. One can find p-values for it using the unadjusted or adjusted quantile tables, and interpolation for $n \leq 5000$ not given in the table, or regression for $n > 5{,}000$ and $\alpha = 0.05$, 0.01, or 0.001.

On the course website are three related R codes "dipr," "diparg," and "diprad." The function "dipr" computes the same function as "dip," but in R directly rather than via Fortran code compiled into machine code. For large n, "dip" will run much faster than "dipr". If one wants to know where the dip occurs (which "dip" doesn't do), i.e. find an x such that $|(F_n - G)(x)| \doteq \sup_u |(F_n - G)(u)|$ for the unimodal distribution function G closest to F_n with respect to this supremum, then "diparg" and "diprad" provide both the dip (the supremum) and such an x. (For some data sets, the dip is attained at all non-tied observations except at most one, with a value $1/(2n)$, and then "diparg" gives an artificial answer "1e9" rather than an actual x.) You can download the codes into your R directory and then in R, type source("dipr"), source("diparg"), or source("diprad") to use them. The code "diprad" gives further outputs that will not be used in this pset.

1. As a normal density has a unique mode, it's more unimodal than $U[0, 1]$. So if one applies the dip test to a normal sample, at the 0.05 level, in other words rejecting it if the dip statistic is larger than the 0.95 quantile given in the table, one might expect to reject unimodality with probability even less than 0.05. The function dip(x) from the diptest library computes the dip statistic but does not provide a p-value.

An R function called "gausstry" is provided for you on the course website. After downloading it you can also type source("gausstry"). It's a function gausstry(n,N,qu) where n is the sample size, N is the number of replications of the experiment, and qu is the quantile. For whatever n you use, which will be one with tabulated quantiles, look up the quantile in the table and use it.

Date: 18.465, Feb. 20, 2015.

Specifically, try this for $n = 20$ and $N = 1000$. Read out with what relative frequency unimodality is rejected and report it. Do you find as expected that it's less than 0.05?

2. Consider mixtures of two normal distributions, both with variance 1 but with means μ and $-\mu$ for some $\mu > 0$. The distribution will then be $\frac{1}{2}[N(-\mu, 1) + N(\mu, 1)]$.
(a) Find for what values of μ this is unimodal. More specifically, find a μ_0 such that the distribution is unimodal for $0 \le \mu < \mu_0$ but not for $\mu > \mu_0$. *Hint*: by symmetry, the density has 0 derivative at 0. Find its second derivative at 0. If the second derivative is positive, there is a relative minimum at 0, so the distribution can't be unimodal. If the second derivative is negative, there is a relative maximum at 0. Proving that in that case 0 is the only mode is left as an extra credit part. Find μ_0 for which the second derivative equals 0.
(b) Let $\mu = 3\mu_0$, so that the distribution is rather far from unimodal, Another R program, called mixg, is also provided on the course website. It's a function mixg(mu,n,N,qu) where now mu= μ as in this problem, and n, N, and qu are as in Problem 1. Try for $n = 100$ and $N = 1000$ to see how often data sets from this non-unimodal distribution are rejected.
(c) Because the distribution is symmetric around 0, both $F_n(0)$ and $G(0)$ for its best-approximating unimodal distribution function will be about 1/2, so the dip will not occur there. Generate 10 data sets with $n = 100$ and the distribution as in part (b), and use diparg to find the dip and where it occurs. List your results. Do the dips tend to occur near one mode or the other, or between the modes, or where?

Suggestions: The μ in this part is $3\mu_0$ as in part (b). If you want to write an R function to do one of the $N = 10$ iterations, it's up to you. Some lines of the "mixg" code could be used. Since $n = 100$, $m = n/2$ can be found once and for all. Generate as in mixg

x = rnorm(m,mu,1) y = rnorm(m,-mu,1) u = sort(c(x,y))

so u is a vector of n = 2m components, in order. (The ordering is necessary for "diparg" to work. It may be that "dip" in the diptest library doesn't require this, I don't know. But it doesn't give the diparg anyhow.) (We do not want the vector (x+y)/2.) Do

diparg(u)

to find (a point "arg dip" where) where the dip is attained. Show the N=10 outputs (either the R outputs or the hand-collated results) for the arg dips.

If you note what the .95 unadjusted dip quantile is for n = 100, you can easily see whether each dip is significant (shows the non-unimodality of the distribution), but the problem does not ask this.
(d) (Extra credit) Show that in fact for $0 < \mu < \mu_0$ the distribution is unimodal with unique mode at 0.
(e) (Extra credit) For $\mu = \mu_0$, is the distribution unimodal or not?

3. Recall that a Beta(a, b) distribution has a density $x^{a-1}(1-x)^{b-1}/B(a, b)$ for $0 < x < 1$ and 0 elsewhere, where $a > 0$, $b > 0$, and $B(a, b)$ is the beta function $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ and is the right normalizing constant for this density.
(a) For what values of $a > 0$ and $b > 0$ is this density unimodal?
(b) In particular, for what $b > 0$ and $a = b$ is it unimodal?
(c) By adapting the gausstry code, you can add new arguments a and b to the function and put instead of rnorm(n), rbeta(n,a,b), which will generate n i.i.d. Beta(a, b) variables. For $a = b = 0.1$, verify that the dip test detects this non-unimodality with high frequency for n not too large.
(d) Generate v=rbeta(100,0.1,0.1), and use diparg to find the dip and where it occurs. Do this 10 times. Can you guess before doing it, by analogy with Problem 2(c), how the arg dip will behave?

4. Here's an application of the dip test to real data. Get into R and type library(MASS). This is a library of data sets, mentioned in Venables and Ripley. There is one called galaxies. Name it e.g. y = galaxies. It contains redshifts of 82 galaxies, in units of km/sec velocity of recession away from us. These are a sample from a paper by some astrophysicists. (The data are already in order as order statistics.) There are density estimates for this sample on p. 138 of Venables and Ripley. By the way, for any data set in their library, look under "Datasets" in their index, then under the particular data set to find on which pages they mention it. They say lower on p. 138 that the data "show evidence of at least four peaks."

From the original source paper (Postman, Huchra and Geller, 1986), I found there is one more galaxy with a lower redshift of 5607 in the same sky regions being sampled, so adjoin it by v = c(5607,y) to get a vector of 83 galaxy redshifts (which are already ordered). Also, there is a typo in "galaxies." The 78th entry (order statistic) in "galaxies," or the 79th after adjoining 5607, given as 26690, should be 26960. (This does not change which order statistic it is.) So type v[79] = 26960.

The correction from 26690 to 26960 is among the largest of the velocities, as there are just 83 total. If the dip doesn't come from the upper

region where the LCM is used, it could happen that it is unaffected by the change.

The programs in the diptest library don't require the data vector to be sorted. Apparently they sort it if necessary. However, my programs such as dipr, diparg... do require sorted data and otherwise give an error message.

(a) After also loading library(diptest), find dip(v). (*Hint*: if instead you use dip.test(v), it will find a p-value by interpolation, so that you don't need to go through the following steps.) Multiply it by the square root of 83 to get an adjusted dip statistic. To get a 0.95 quantile, interpolate between the n=50 and n=100 quantiles of the adjusted statistic, noting that 83 is about 2/3 of the way from 50 to 100. Is unimodality rejected at the 0.05 level by the dip test?

For parts (b) and (c) consider the following.

The dip function from library "diptest" has an option mentioned in Maechler's documentation, "Package 'diptest'", p. 2. Namely, you can instead of just dip(v), type dip(v,full.result=TRUE). That seems not to give the whole list mentioned in the documentation, but it does give "Modal interval [xL,xU] = ...". If arg dip is to the left of the modal interval, then the dip is $(F_n - GCM(F_n))(x)/2$, or if arg dip is to the right of the modal interval then it's the other expression in part (b) (with $x-$ instead of x). So from this you could answer part (b).

Now, when you list the data, you may see the arg dip among the data points (galaxy velocities). What is it about this data point that might make it the arg dip? Do you see on one side or the other of this point, a cluster of other values, or a gap without such values? What will these make F_n look like in the neighborhood? The GCM and LCM are each piecewise linear, so you need to find the appropriate line segment, which involves finding its endpoints, which are points on the graph of F_n, or limits of such points at jumps, i.e. points $(X_(j), F_n(X_(j)-))$.

(b) Note that the majority of the galaxies have redshifts between 18419 and 27000. Let G be convex for $x \leq m$ and concave for $x \geq m$. Use diparg on the data set. The dip is either $(LCM(F_n) - F_n)(x)/2$ if $x > m$ or $(F_n - GCM(F_n))(x)/2$ if $x < m$. Which is it?

(c) Find the line segment giving the graph of the LCM or GCM, whichever applies, on an interval containing arg dip. A method of finding the GCM of an empirical distribution function on a left half-line is described in Subsection 5.1 of the handout "Unimodality and the dip statistic," and for the LCM on a right half-line, there is a symmetric method.

5. (a) As the linear regressions of some adjusted dip quantiles on $x = 1/\sqrt{n}$, as given on the back of the table of adjusted quantiles, have concave residual patterns, do a quadratic regression (as in PS2):

for the 0.95 quantiles, if your last name begins with D through M;

for the 0.99 quantiles, if your last name begins with R or S;

for the 0.999 quantiles, otherwise.

See if the coefficient of x^2 is significantly different from 0.

(b) Even if it isn't, see how well the linear and quadratic regressions predict the adjusted quantiles for the largest two n with given quantiles not used in the regression, namely $n = 100$ and 200. If the quadratic regression does worse, it could indicate that the quadratic regression "overfitted" the four values for $n = 500$ on up. If it does better, then it may be preferable.