9.2 Variance of estimators for $w(\theta)$

The four estimators in question are as follows, with DD the number of pairs lying between θ and $\theta + \delta\theta$ for the real sky, RR the equivalent numbers in the random sky whose shape matches that of the real sky, and DR the cross-pair counts calculated from distance between each sky point and each random point.

$$w_1 = \frac{r(r-1)}{n(n-1)}\frac{DD}{RR} - 1 \quad (1)$$

$$w_2 = \frac{2r}{(n-1)}\frac{DD}{DR} - 1 \quad (2)$$

$$w_3 = \frac{r(r-1)}{n(n-1)}\frac{DD}{RR} - \frac{(r-1)}{n}\frac{DR}{RR} + 1 \quad (3)$$

$$w_4 = \frac{4\,n\,r}{(n-1)(r-1)}\frac{DD \times RR}{(DR)^2} - 1 \quad (4)$$

with w_1, w_2, w_3 and w_4 known as the Natural, Peebles, Landy-Szalay and Hamilton estimators. n is the number of points on the sky, r the number of randoms.

Note that correlations represent heavy computing: sorting out and binning $n(n-1)/2$ correlation distances for DD, $r(r-1)/2$ distances for RR, and nr distances for DR. Despite modern computing power you may need to think of optimized computational techniques to avoid minutes to hours of cpu.

A simple Fortran (non-optimized) loop to find and bin the $n(n-1)/2$ correlations in DD is

```
do 10 i=1,(n-1) !n points
   do 10 j=i+1,n
      arg1=alpha(i)-alpha(j)
      arg2=dec(i)-dec(j)
      dist=sqrt(arg1*arg1+arg2*arg2) !approximate short distances by Pythagoras
      index=ifix(dist/.05)+1 !bin every .05 degrees
      dd(index)=dd(index)+1.
10   continue
```

The results of the exercise for the given data file are shown in Figure 1.

The error in w_1 exceeds the Poisson error by a factor increasing with θ and the entire measurement is systematically offset. The effect is apparent to a lesser extent in w_2. The Poisson error is a good approximation to the error in w_3.

Try this with a random sky of your own making; after all if you've got to this point you will have needed to make 10 such random skies anyway. Chances are you won't get such an offset with w_1 and w_2; you'll get a result more like that of Figure 2. Why? Because the *average* of w_1 or w_2 over many realizations of 20000 data points has to be identically zero in all bins, and the most likely individual offsetis close to zero. Thus we confess; while the data set does indeed contain randomly and independently-distributed points, the data set was not chosen at random. Investigate how many 'random' skies we might have tried before finding one giving as much offset as that in Figure 1.

The estimator w_4 will differ little from w_3 certainly on this order of magnitude of points. Investigate, but read the original papers by Landy and Szalay (1993) and by Hamilton (1993) before taking this too far.

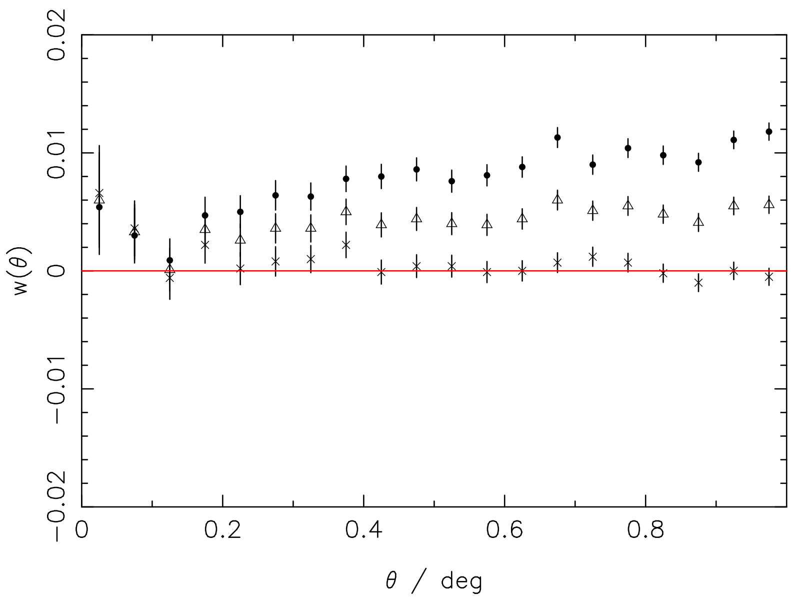

Figure 1: Variance of estimators for $w(\theta)$. The data consist of 20000 points generated in the region $0° < \alpha < 5°, 0° < \delta < 5°$, as given in the data file for the example. Here DD and DR have been averaged over 10 independent comparison sets of 20000 random points each. w_1 is indicated by dots, w_2 by triangles and w_3 by crosses. Poisson error bars $(1+w)/\sqrt{DD}$ are shown.

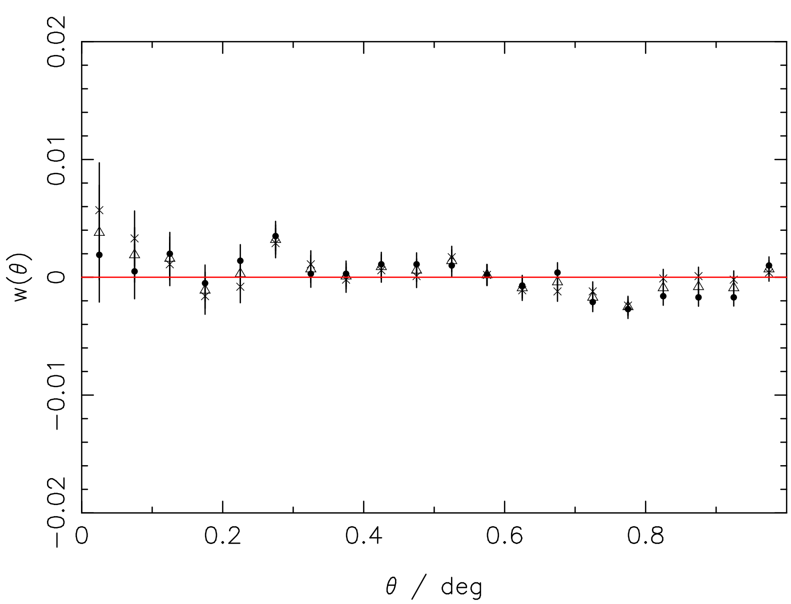

Figure 2: Variance of estimators for $w(\theta)$, using a different initial random sky but identical methodology; symbols as in Figure 1.