5.2 The Wilcoxon-Mann-Whitney test

Figure 1 in the solution for exercise 5.1 shows the histogram of the data tabled for this example.

Carrying out the Wilcoxon-Mann-Whitney test as described in the book requires that you rank the combined sample, preserving the m or n identity of each measurement, so that the combined m-ranks (or n-ranks) can be summed. A simple way to do this is as follows:

1. Sort each of the two data arrays, using a routine such as *shell.for* from *Numerical Recipes*.

2. Make a single-dimension rank array for each array, which after the sort is bound to be of the simple form $m(1) = 1, m(2) = 2, m(3) = 3, ...m(386) = 386; n(1) = 1, n(2) = 2, ...n(290) = 290$.

3. Interleave the smaller array into the larger one, adjusting the rank array for each accordingly. This can be done by hand; but the following few lines of Fortran work:

```
c data arrays are gals(386) and rand(290)

c rank arrays are irank_gals(386)and irank_rand(290)

      inc=0
      j=1
      do 70 i=1,290
76      continue
        if(gals(i).ge.rand(j).and.gals(i).lt.rand(j+1))then
            inc=inc+1
            irank_gals(i)=irank_gals(j)+inc
         do 74 k=j+1,386
            irank_rand(k)=irank_rand(k)+1
 74      continue
         goto 70
         endif
         j=j+1
         m=0
         goto 76
 70      continue
```

Following this procedure, we find the summed ranks for the data at galaxy positions to be $U_m = 131514$ for a combined sample of $N = m+n = 676$. This number is far beyond the reach of Table A2.9, and we therefore assess significance as described in the book using the Normal distribution, for which we get mean $\mu_m = m(N + 1)/2 = 130661$, square root of variance $\sigma = \sqrt{mn(N + 1)/12} = 2513$ and hence $z = (U_m - \mu_m)/\sigma_m = 0.339$. Reference to the erf function (integral Gaussian, Table A1) shows that this leaves an upper-tail area of 0.345. Our alternative hypothesis - that the distribution with m=386 members is greater than the distribution with 290 members - implies 'rejection' of the null hypothesis at the significance level of 0.345. Of course this is

no rejection at all. The Wilcoxon-Mann-Whitney test is a ranking test which should provide the most powerful discrimination when *location* is concerned. It gives us a null result, providing even less indication of a shift between the distributions than the K-S test of exercise 5.1.

Exercise 5.2 suggests trying the chi-square test, almost as an afterthought. Here comes trouble. The following table gives the chi-square results and in particular the contributions to chi-square from each bin of the two distributions. The expectation values, the $E(1,i)$ and $E(2,i)$ are computed from equation 5.17 of the book, and chi-square is computed from the expectation values according to equation 5.16.

bin i	centre mJy	m	E(1,i)	chisq	n	E(2,i)	chisq
1	-15.00	0	0.000	-	0	0.000	-
2	-14.00	0	0.000	-	0	0.000	-
3	-13.00	1	0.571	0.322	0	0.429	0.429
4	-12.00	2	1.142	0.645	0	0.858	0.858
5	-11.00	5	2.855	1.612	0	2.145	2.145
6	-10.00	7	5.710	0.291	3	4.290	0.388
7	-9.00	2	4.568	1.444	6	3.432	1.922
8	-8.00	11	7.423	1.724	2	5.577	2.294
9	-7.00	11	9.707	0.172	6	7.293	0.229
10	-6.00	7	10.278	1.046	11	7.722	1.392
11	-5.00	13	14.275	0.114	12	10.725	0.152
12	-4.00	11	16.559	1.866	18	12.441	2.484
13	-3.00	25	25.124	0.001	19	18.876	0.001
14	-2.00	39	35.402	0.366	23	26.598	0.487
15	-1.00	31	29.121	0.121	20	21.879	0.161
16	0.00	32	32.547	0.009	25	24.453	0.012
17	1.00	37	36.544	0.006	27	27.456	0.008
18	2.00	22	27.979	1.278	27	21.021	1.701
19	3.00	31	30.834	0.001	23	23.166	0.001
20	4.00	23	25.695	0.283	22	19.305	0.376
21	5.00	21	18.272	0.407	11	13.728	0.542
22	6.00	11	15.417	1.266	16	11.583	1.685
23	7.00	6	9.136	1.077	10	6.864	1.433
24	8.00	16	9.136	5.157	0	6.864	6.864
25	9.00	4	2.284	1.289	0	1.716	1.716
26	10.00	6	6.281	0.013	5	4.719	0.017
27	11.00	4	2.284	1.289	0	1.716	1.716
28	12.00	2	3.426	0.594	4	2.574	0.790
29	13.00	6	3.426	1.934	0	2.574	2.574
30	14.00	0	0.000	-	0	0.000	-
31	15.00	0	0.000	-	0	0.000	-
Totals (bin 6 to 26):				17.931			23.865

Following the strictures in the book, we chop off the 5 top and bottom bins as having such low expectation values as to destabilize chi-square. The result is a value of $\chi^2 = 41.80$ for $(21-1)(2-1) = 20$ degrees of freedom. Looking up Table A2.6 we find that under H_0, χ^2 is expected to exceed this only 0.5% of the time. Do we have a result at the 0.005 level of significance, despite the results of the foregoing (K-S and W-M-W) tests?

Look again at the table. By far the biggest contributions come from bin 24 for both samples. In the case of the larger (targeted) sample, there are 16 objects in this bin;

in the case of the random positions, there are zero. Suppose we remove this bin, but considering that this is a little too subjective, remove three more bins at the bottom and three at the top, so the chi-square is now computed on bins 9 to 23 inclusive. We find $\chi^2 = 18.67$ for 15 degrees of freedom, a level of significance at 0.13 and an inconclusive result as in the previous tests.

But you cry, the process has carefully removed just the bin in which the big difference shows up! Yes, but the values of chi-square should all have some measure of contribution to a large total; the distributions, if offset, are not offset on only one bin.

However, we did the extra bin removal *a posteriori* perhaps because we had some clear indications from more powerful tests that there was little result here. We have undoubtedly got ourselves into the realm of subjectivity, no matter what the result is. (Maybe we could have got away with retaining the data of the outer bins if we had made the bins twice as wide...and so forth.)

It is a problem of the chi-square test and of binning tests in general that some subjectivity to deal with bin sizes and distribution ends must be invoked. Alternatives should be sought.

There is a wider message. If you do one test and one alone, then examine very carefully where and how it is producing the result, if result there be. (For chi-square testing, it is imperative (a) never to do the test blindly and (b) always to consider the run of the individual contributions.) If you do more than one test - advisable if possible - then 'combining' the tests (as the trick question suggests) consists of examining their consistency, and how and where the inconsistencies arise. There is no formal mechanism for using a set of probabilities from different tests to produce a 'definitive' level of significance. Different tests are sensitive to different features - see Tables 5.5 and 5.6 - and have differing powers for hypotheses under consideration. The important aspect, as ever, is to choose the alternative hypothesis, H_1, *a priori*; and then to choose the test(s).

As an extension of exercises 5.1 and 5.2, use the bootstrap test, section 6.6, to assess the significance of the results for all three tests used.