Nonparametric tests

Mark Muldoon
School of Mathematics, University of Manchester

Overview

Last week we studied the t-tests, which assume the data are drawn from normal distributions. We also learned how to make qq-plots . . . what if these showed our data to be non-normally distributed?

- The sign test, a simple introduction to the idea of non-parametric tests
- The Mann-Whitney U test, an analogue of the two-sample t-test
- The Wilcoxon rank-sum test, an analogue of the paired-sample t-test
- Why would you want to use non-parametric tests?
- Why wouldn't you want to use them?

The sign test, motivation

Recall the first of H.H. Koh's macular pigment data sets

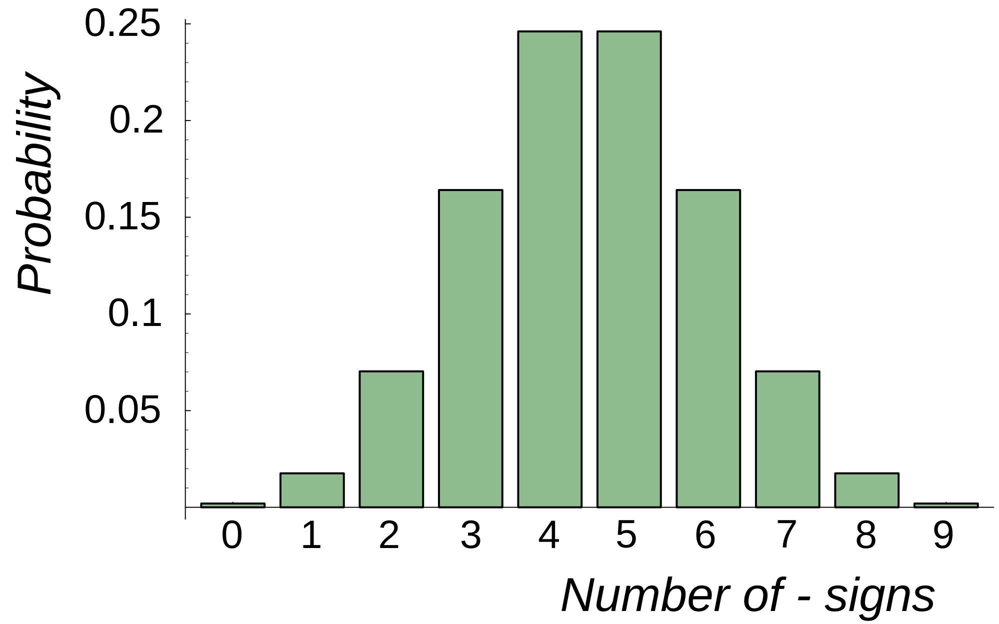
Patients	Controls	Difference	Sign of Diff.
0.189	0.377	-0.188	-
0.301	0.26	0.041	+
0.072	0.161	-0.089	-
0.242	0.119	0.123	+
0.271	0.295	-0.024	-
0.32	0.055	0.256	+
0.409	0.037	0.372	+
0.279	0.179	0.100	+
0.556	0.453	0.143	+

The idea of the sign test

Suppose the two groups were the same, in the sense that their MPOD levels were drawn independently from the same distribution.

- Positive and negative differences would be equally likely: each has probability 0.5.
- Say there are N pairs. Then the number of negative differences has a binomial distribution with $p = 0.5$. (Ignore differences of zero and reduce N accordingly).
- This is a sort of paired-sample test.

Distribution of number of − signs



Generally

$$P(\ k \text{ neg. diffs in } N \text{ pairs }) = p^k(1-p)^{N-k}\frac{N!}{k!\,(N-k)!}$$
$$= (0.5)^N\frac{N!}{k!\,(N-k)!}$$

MPOD data, conclusion

Returning to the MPOD data, one finds that if the null hypothesis were true . . .

- The probability of seeing 3 or fewer minus signs (a one-sided test) is about 0.254.
- The probability of seeing either 3 or fewer or 6 or more minus signs (the two-sided version) is about 0.508.
- It appears, from this test, that the patients and the controls have the same distribution of IOP.

The Mann-Whitney test

This is a non-parametric replacement for the two-sample t-test:

- Ingredients are two samples $\{x_1, x_2, \dots, x_{N_x}\}$ and $\{y_1, y_2, \dots, y_{N_y}\}$, not necessarily the same size.
- The null hypothesis is that the x's and y's are drawn from the same distribution.
- Alternative hypotheses include:
    - The two data sets are drawn from different distributions (two-tailed alternative)
    - The x's tend to be larger (one-tailed alternative)
    - The y's tend to be larger (another one-tailed alternative)

The idea behind Mann-Whitney test

The idea behind the test is clearest when the lists are very small, say 2 x's and 3 y's.

- Assemble all 5 numbers into a single list and arrange them in ascending order.
- If the null hypothesis is true—if these numbers really are all drawn from the same distribution—then all orderings of the x's and y's are equally likely.
- It would be somewhat odd if, say, the two x's were the first two entries in the ordered list.
- It is, of course, equally unlikely that they should fall in the last two entries.

Possible outcomes when ordering 5 letters

Pattern of x's & y's in ordered list	Ranks of the x's	Rank Sum
xxyyy	1, 2	3
xyxyy	1, 3	4
xyyxy	1, 4	5
yxxyy	2, 3	5
yxyxy	2, 4	6
xyyyx	1, 5	6
yyxxy	3, 4	7
yxyyx	2, 5	7
yyxyx	3, 5	8
yyyxx	4, 5	9

Remarks

- There are "five-choose-two",

$$\binom{5}{2} = \frac{5!}{2! \times 3!} = 10$$

ways to arrange the letters

- The sum of the ranks associated with the two x's ranges from 3 to 9.
- All the values of the rank-sum are reasonably likely:

Rank sum	3	4	5	6	7	8	9
Probability	0.1	0.1	0.2	0.2	0.2	0.1	0.1

Larger samples

For even moderate N_x and N_y there are large numbers of patterns (large compared to the range of the rank sum, that is), and some values of the rank sum are much, much more likely than others.

N_x	N_y	Num. patterns	Min. rank-sum	Max. rank-sum	Range of rank-sum
2	5	21	3	13	11
2	6	28	3	15	13
2	7	36	3	17	15
2	8	45	3	19	17
3	17	2280	6	57	52

Larger samples, details

The formulae are:

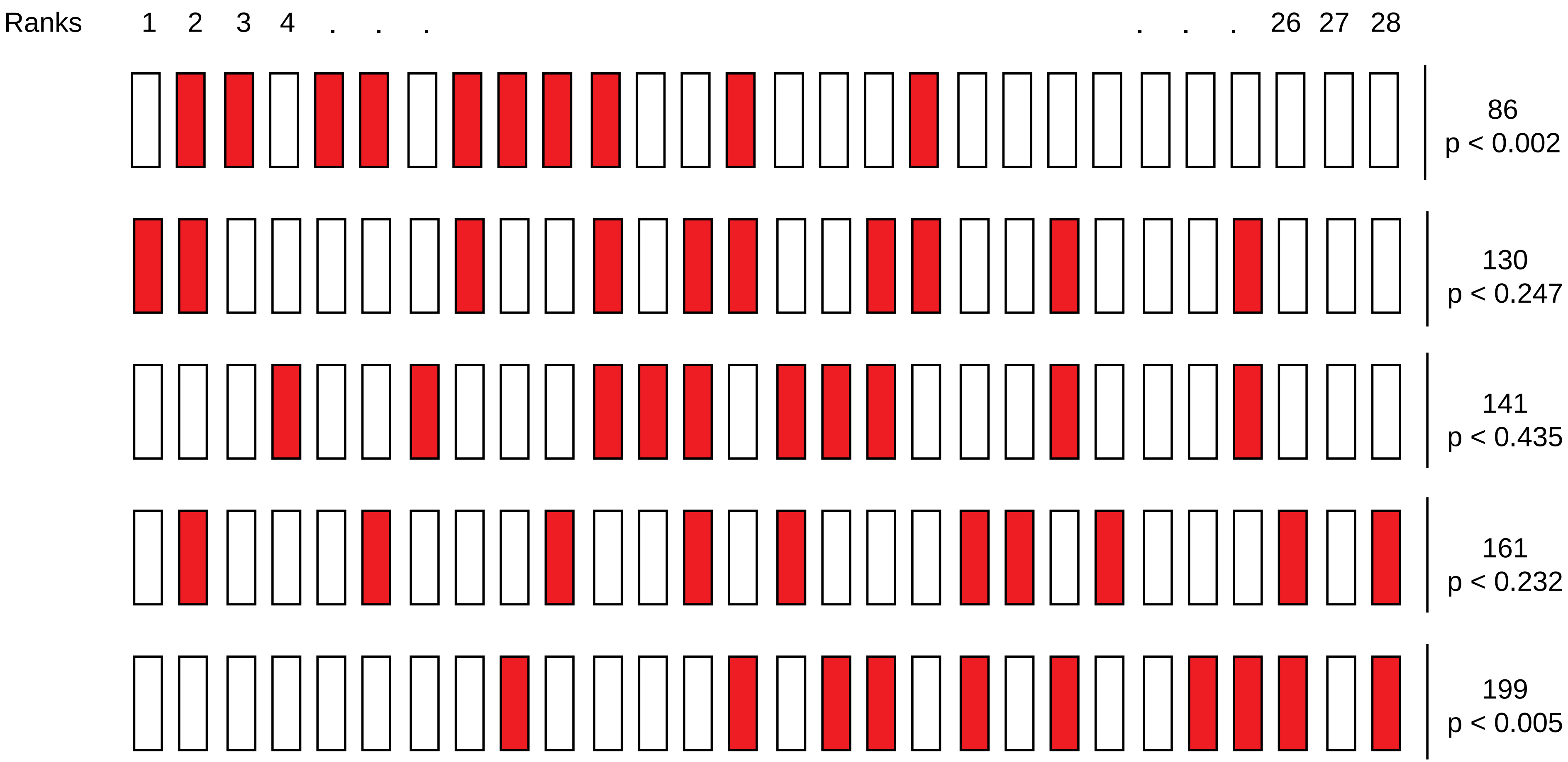
$$\text{Num. of patterns} = \frac{(N_x + N_y)!}{N_x! \times N_y!}$$

$$\text{Min. rank-sum} = \frac{N_x(N_x + 1)}{2}$$

$$\text{Max. rank-sum} = \frac{N_x(N_x + 2N_y + 1)}{2}$$

$$\text{Range of rank-sum} = N_x N_y + 1$$

Larger samples, in pictures



The figures at the end of each row give the rank sum for the red bars and the one-sided probability of a sum that size having arisen by chance if the null hypothesis were true.

The Mann-Whitney recipe

The ingredients are a confidence level C and two lists of values $\{x_1, x_2, \ldots, x_{N_x}\}$ and $\{y_1, y_2, \ldots, y_{N_y}\}$ where $N_x \leq N_y$. The null hypothesis is that these two lists are drawn from the same distribution.

- Assemble all the data into one large list.
- Sort the data into ascending order and assign ranks, averaging over ties.
- Add up the ranks of the x's and call this number R_x.
- Work out the probability of getting this sum when adding together N_x randomly chosen numbers from the list $\{1, 2, \ldots, (N_x + N_y)\}$.

The Mann-Whitney recipe, continued

The test statistic, U, is defined by $U = \min(u_x, u_y)$ where

$$
\begin{aligned}
u_x &= N_x N_y + \frac{N_x(N_x+1)}{2} - R_x \\
u_y &= N_x N_y - u_x \\
&= N_x N_y + \frac{N_y(N_y+1)}{2} - R_y
\end{aligned}
$$

where R_y is the rank sum for the y's.

The Mann-Whitney recipe, large samples

When both N_x and N_y are bigger than about 8 this statistic is approximately normally distributed with mean and variance given by:

$$\mu_U = \frac{N_x N_y}{2}$$

$$\sigma_U = \sqrt{\frac{N_x N_y (N_x + N_y + 1)}{12}}$$

One can thus check whether the U one measured is extreme by computing a z-score

$$z = \frac{U - \mu_U}{\sigma_U}$$

and checking it against the standard tables.

Large samples with many tied values

If the data contain many repeated values the standard deviation of the rank sum will be greatly diminished. In this case a more accurate expression for σ_U is this appalling formula

$$\sqrt{\left(\frac{N_x N_y}{(N_x + N_y)(N_x + N_y - 1)}\right)\left(\sum_j^{N_x+N_y} r_j^2\right) - \left(\frac{N_x N_y (N_x + N_y + 1)^2}{4(N_x + N_y - 1)}\right)}$$

Here the r_j are the ranks (after averaging) of all the data.

The Mann-Whitney recipe, small samples

For smaller values of N_y one consults the attached table and

- rejects the null hypothesis in favour of the two-sided alternative if U is *strictly less than* the value tabulated for $\alpha = (1 - C)/2$;
- rejects the null in favour of the one-sided alternative "the x's tend to be bigger" if u_x is *strictly less than* the tabulated value for $\alpha = (1 - C)$;
- rejects the null in favour of the one-sided alternative "the y's tend to be bigger" if u_y is *strictly less than* the tabulated value for $\alpha = (1 - C)$;

The Mann-Whitney test, an example

Consider the following random sub-sample from Catherine Collins's IOP data (the file `CDvsIOP.xls` that we have examined in computer labs).

Received treatment	22	29	40	34	26	
No treatment	20	27	16	18	48	22

Take the values from the Treatment group to be the x's as there are fewer of them. Then $N_x = 5$ and $N_y = 6$.

Sorting the data

Arranging the data in order one finds

Value:	16	18	20	22	22	26	27	29	34	40	48
Rank:	1	2	3	4	5	6	7	8	9	10	11

But it doesn't really make sense to give the two 22's different ranks: one should average the rank over tied values.

Value:	16	18	20	22	22	26	27	29	34	40	48
Rank:	1	2	3	4.5	4.5	6	7	8	9	10	11

Summing the ranks

The rank-sum corresponding to the x's is thus

$$R_x = 4.5 + 6 + 8 + 9 + 10 = 37.5$$

and the test statistics are

$$\begin{aligned} u_x &= N_x N_y + \frac{N_x(N_x+1)}{2} - R_x \\ &= 5 \times 6 + (5 \times 6)/2 - 37.5 \\ &= 7.5 \end{aligned}$$

and

$$u_y = N_x N_y - u_x = 22.5$$

Checking the table

Clearly the smaller of the two is u_x so $U = u_x$. Consulting the attached table in the rows for $n_1 = 5$ and $n_2 = 6$ we see that we cannot reject the null hypothesis: these 11 values could plausibly all have been drawn from the same distribution.

Paired samples: the Wilcoxon test

There is also a non-parametric analogue of the paired-sample t-test which looks at the distribution of the differences between the members of the pairs. The null hypothesis is that both members of each pair are drawn from the same distribution, though the distribution may vary from pair to pair.

If the null is true, the differences are equally likely to be positive or negative and should be distributed symmetrically. These ideas give rise to the Wilcoxon test . . .

The Wilcoxon recipe

The ingredients are a confidence level C and a list of paired values $\{(x_1, y_1), \ldots, (x_N, y_N)\}$; here there are N pairs. The null hypothesis is that the respective members of the N pairs are drawn from identical distributions.

- Compute the list of differences $\delta_j = (x_j - y_j)$.
- Sort the absolute values of the differences $\{|\delta_j|\}$ into ascending order.
- Add up the ranks assigned to the positive differences and call it w_+. Similarly, find the rank-sum for the negative differences and call it w_-.

The Wilcoxon recipe, large samples

For $N > 25$ there is an approximate test based on a z-score.
Find $T = \min(w_+, w_-)$ and compute

$$z = \frac{T - \mu_T}{\sigma_T}$$

where μ_T and σ_T are given by:

$$\begin{aligned} \mu_T &= \frac{N(N+1)}{4} \\ \sigma_T &= \sqrt{\frac{N(N+1)(2N+1)}{24}}. \end{aligned}$$

The Wilcoxon recipe, small samples

For small N one must do something more involved using the attached table

- Reject the null in favour of the two-sided alternative if $T = \min(w_+, w_-)$ is *less than or equal to* the tabulated value for $\alpha = (1 - C)/2$.
- Reject the null in favour of the one-sided alternative "the x's tend to be bigger" if w_- is *less than or equal to* tabulated value for $\alpha = (1 - C)$;
- Reject the null in favour of the one-sided alternative "the y's tend to be bigger" if w_+ is *less than or equal to* the tabulated value for $\alpha = (1 - C)$;

Example: Wilcoxon applied to the MPOD pairs

Here again are the differences, (patient - control), in the MPOD data:

δ	-0.188	0.041	-0.089	0.123	-0.024	0.256	0.372	0.100	0.143
$\lvert\delta\rvert$	0.188	0.041	0.089	0.123	0.024	0.256	0.372	0.100	0.143

where those among the $|\delta_j|$ that come from negative differences are highlighted in red.

Sorting these yields

$\lvert\delta\rvert$	0.024	0.041	0.089	0.100	0.123	0.143	0.188	0.256	0.372
Rank	1	2	3	4	5	6	7	8	9

MPOD pairs: rank sum

The two rank sums are

$$\begin{aligned} w_- &= 1 + 3 + 7 = 11 \\ w_+ &= 2 + 4 + 5 + 6 + 8 + 9 = 34 \end{aligned}$$

Thus $T = w_- = 11$ and one sees, from the attached table, that once again we cannot reject the null hypothesis with any substantial confidence.

Why use non-parametric tests?

There are two main reasons:

- They apply to more kinds of data than the t-tests. Suppose the observations can be ordered, but that differences between the observations do not have a consistent meaning. For example, anxiety is sometimes measured with sets of yes-no questions and the score is the number of yes answers. A set of 36 questions gives a scale 0-36, but the difference between scores of 1 and 2 is not the same as the difference between scores of 31 and 32.
- Even when t-testing is possible, the data may be so obviously non-normal (say, bimodal) as to rule it out.

Why not non-parametric tests?

In light of the preceding slide one might wonder, why not use non-parametric tests all the time?

- Non-parametric tests cannot give significant results for very small samples as all the possible rank-sums are fairly likely.
- They may require the use of fiddly tables and are not built into Excel (though they are available in R).
- They do not, without the addition of extra assumptions, give confidence intervals for the means or medians of the underlying distributions.

Downsides, continued

Other reasons not to use non-parametric tests include

- They are not *entirely* without assumptions: they assume that the data can be ordered and if there are lots of ties the power of the test will be much diminished.
- If the t-tests are applicable they will be more powerful: they will more often correctly detect small differences between two normally-distributed groups of measurements (whether paired or not).