Pairwise comparisons of treatment groups via Eulerian tours and Hamiltonian decompositions

Catherine B. Hurley[1] and R.W. Oldford[2]

[1] National Univ of Ireland Maynooth, [2] Univ of Waterloo Canada

Introduction

We present improved graphical displays for two classical data analysis problems- the comparison of treatments in a one-way layout, and the assessment of interaction between factors in a two-factor experiment. In both settings the key to the improved display is that all $\binom{n}{2}$ pairs of factor levels are compared, whereas conventional displays look at n-1 pairs only. The construction of these improved displays relies on concepts from graph theory, namely eulerian and hamiltonian traversals of complete graphs, We review these concepts and outline our constructions.

Boxplot comparison of treatment groups

Cameron and Pauling 1978 collected data on survival times of vitamin C treated cancer patients. How does survival time vary with cancer type?

Parallel boxplots of survival times for each cancer type (Figure 1a), are a typical way of comparing (here square root transformed) survival times. We describe some shortcomings of this display.

True or false?

Median survival for ovarian cancer is better than for colon cancer.

50% of breast cancer survivals exceed all stomach times

Can you tell from the boxplots below?

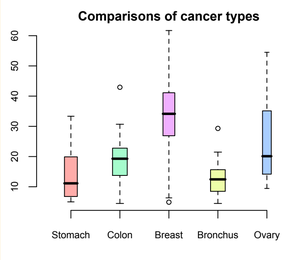

Fig 1a: Boxplots of groups

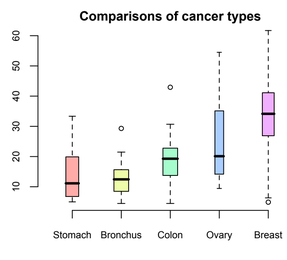

Fig 1b: Median sorted boxplots

Comparisons are easily made for groups that appear adjacently, such as Breast-Colon in Figure 1a. But a shortcomig is than visual comparisons between distant groups are tricky.

Note that taking the above two plots together, we have displayed all pairs of treatments adjacently, except Breast-Stomach, and Ovary-Stomach. In fact, we would need three permutations of the five groups to show all pairs of groups adjacently-- but then some pairings are duplicated.

All pairs boxplots

A more compact way to show all pairs of groups adjacently uses a longer sequence got by concatenating the sequences of Figure 1 and appending an extra `Bronchus': Stomach,...., Colon, Stomach,...., Breast, Bronchus. Of course, there are many sequences where every pair of groups appears adjacently; e.g. Figure 2 shows another.

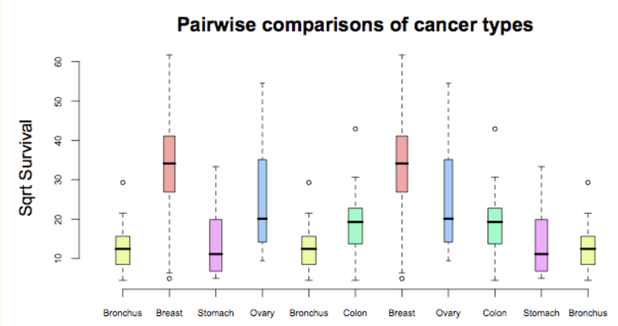

Figure 2: Boxplots of cancer data with all groups appearing adjacently

When the goal is comparing pairs of means, it is conventional to plot a separate display showing confidence intervals for each pair of means. We present a new, information rich visualization which combines pairwise confidence intervals and boxplots in a single dislay.

A new multiple comparison display

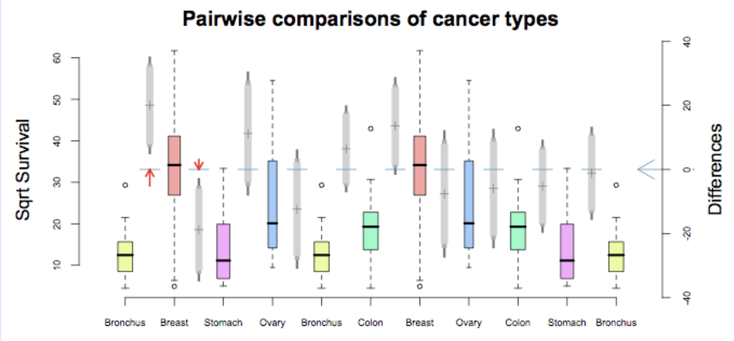

Figure 3 New multiple comparison display

The new display is constructed as follows:

Draw all-pairs boxplots groups, as in Figure 2. Insert confidence intervals (99%, Tukey HSD) for the difference in means of adjacent group. Add a second axis on the right hand side for the difference scale. Mark the zero line with disjoint blue line segments, so as not to interfere with the boxplots. A confidence interval not crossing the zero line shows that the pair of means is significantly different. Mark CIs for significantly different means with a red arrow, drawn on the opposite side of the zero line to the CI, whose length is determined by the distance of the CI to the zero line.

Note that each mean comparison is actually represented by a sequence of CIs of decreasing width, giving 90%, 95% and 99% confidence. For example, Colon-Breast are significantly different at 5% level but not at 1% level. We have designed the sequence of groups in such a way that the most significantly different pair occurs first, with p-values following an increasing trend theraftter.

With the boxplots and confidence intervals on a single display, the analyst can compare groups informally via boxplots and formally via confidence intervals. It should also be easy to assess whether significantly different pairs occur as a result of data outliers. This new display addresses major shortcomings in displays of groups such as (Figure 1a) and in conventional multiple comparison displays (not shown).

Interaction plots

Interaction plots are used to explore the presence of interactions between two factors. Figure 4 shows an interaction plot for the survival time of 48 rats, each given one of four treatments A, B, C, or D and one of three poisons P1, P2, or P3 (data from Box and Cox, 1964). Interaction is detected as a lack of parallelism in the profiles.

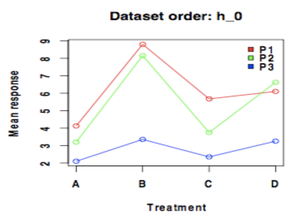

Figure 4: Interaction plot of Rat data

The profiles are similar in shape, indicating relatively strong main effects. The eye is drawn to possible interactions involving P1 and P2 with C and D where these profiles cross, but also to no interaction of P1 and P2 with A and B in the long nearly parallel line segments from A to B. in the long nearly parallel line segments from A to B.

Our perception of parallelism in the profiles depends on which line segments are compared, but in Figure 4 we can only compare segments connecting treatments in the order ABCD. To remedy this, we use treatment sequences where all levels appear adjacently. For 4 levels, two sequences (shown in Figure 5) suffice, and unlike the 5-group cancer example, no saving is had by concatenating the two sequences.

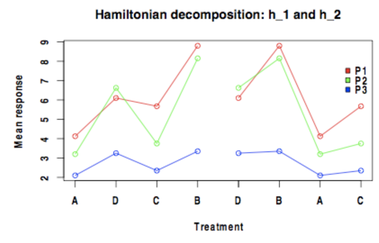

Figure 5: Interaction plots of Rat data, showing all levels adjacently

The first of these sequences had two crossings in the P1 and P2 profiles giving a stronger visual impression of interaction. By constrast, in the second sequence the P1 and P2 profiles zig-zag together, but the P3 profile is quite flat, suggesting that choice of treatment matters less for P3-dosed rats..

Graph traversals

Constucting visualizations for all pairwise comparisons of treatment groups requires an excursion to graph theory. Consider a graph where each node represents a treatment group. Figure 6 shows such a graph for the vitamin C data.

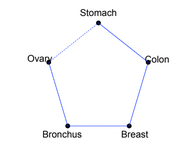

Figure 6: graph with 5 nodes

Here comparisons between every pair of groups is of interest so every pair of nodes is joined by an edge, resulting in a complete graph. For clarity these edges are not shown. The solid blue path visits all nodes exactly once and so is a **hamiltonian path**, corresponding to the sequence of boxplots shown in Figure 1a. Adding the dashed line segment forms a **hamiltonian cycle**. A path that visits every edge of the graph exactly once is an **eulerian path**. An **eulerian cycle** is just a closed eulerian path. The sequence of nodes visited in Figures 2 and 3 is such an eulerian cycle. In this example, the edge weights are p-values comparing node means. We constructed the eulerian using a classical algorithm modified to visit low-weight edges first.

Eulerian paths on complete graphs may also be formed from **hamiltonian decompositions**, namely the concatenation of edge-disjoint hamiltonian cycles. For example, the two sequences shown in Figure 1 form such a decomposition.

It is well-known that eulerian paths on complete graphs exist when the number of nodes n is odd. For even n, a path visiting all edges must necessarily visit n/2 -1 edges twice. Consider the interaction plot example. Here the underlying graph has n=4 nodes. When the two edge-disjoint hamiltonian paths of Figure 4 are concatenated, the resulting path visits all edges, but the BD edge joining the hamiltonians is visited twice.

Concluding remarks

In this presentation, we explored graphics for the one-way and two-way layouts as graph traversal problems. Conventional displays traverse all graph nodes and so visualize n-1 pairwise comparisons instead of all $\binom{n}{2}$ comparisons which may mislead or confuse the analyst. We recommend displays based instead on edge-traversals. In the case of the one-way layout, this motivates our design of a new, information-rich display. For two-way layouts, it leads to improved visual assessment of the presence of interaction.

In Hurley and Oldford(2008), we propose that the graph traversal metaphor is a useful notion for evaluating a plethora of statistical visualizations, leading to other improved or possibly new visualizations. Parallel coordinate displays are an obvious application. Here the underlying graph has one node per variable, and edge-weight is some measure of the importance of the pairwise relationship. Conventional PCPs visit all n nodes and thus display only n-1 pairwise relationships. By switching from node to edge-traversals, the resulting display shows all $\binom{n}{2}$ pairwise variable relationships.

The down side to visualizations based on edge traversals is that edge-traversing paths are much longer than node-taversing paths and visualizing them requires more physical space. For this reason we offer algorithms that construct edge-traversing paths where important edges, or possibly hamiltonians, occur first. The idea here is that when the path is rendered from let to right, our attention is naturally drawn to the first part.

Selected References

Hurley, C.B and Oldford R.W. 2008. "Pairwise display of high dimensional information via Eulerian tours and Hamiltonian decompositions", submitted.

Research supported in part by Research Frontiers grant from Science Foundation Ireland and a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

Presented at IBC July 2008, held in Dublin.