


References

- 
- [1] Eklof, M. D., and McDonnell, E. [eds.] (1993): Committee Draft 1: Programming Languages APL, Extended International Standards Organisation
 - [2] Kerf, J. de (1994): Extended APL versus ISO APL. APL CAM 16-(1), 74–95
 - [3] Pronk, C. (1993): Standaardisatie van programmeertalen. Informatie 36-(6), 398–405

How Much Difference Makes a Difference?

Dick Holt

3802 N. Richmond Street
Arlington, VA 22207 USA

Tel: 703-528-7624 home, 202-586-4449 office

Fax and BBS\APL: 703-528-7617, 1200-14440b, 24-hr

E-mail: dick.holt@acm.org

**Federal R&D Planning
with Non-Parametric Statistics in APL**

The Scenario

This is a story about APL in the everyday world of work.

Here's the picture: A group of high-level program directors at the Department of Energy has been asked to write a five-year strategic plan for transportation R&D. The plan is required by the Energy Policy Act of 1992. The audience for this plan is the Congress, the automobile industry, environmental groups and other lobbies, and the R&D program offices themselves.

Congress is concerned: California, plus a bloc of twelve eastern states, have mandated that 2% of all new car sales in 1999 be "zero-emission vehicles" because air quality in their urban areas doesn't meet EPA standards. In practice, this implies electric cars. But current (no pun intended) electric cars just don't hack it. Carefully targeted R&D is needed to improve range and performance, and to reduce costs.

Further, Congress is concerned that U. S. oil imports are growing at the same time that the domestic oil industry is in sharp decline. More importantly, people in many U. S. cities, as well as in many cities around the world, can't breathe because of automobile air pollution.

We need new thinking—from an energy, economic, and environmental point of view. The stakes are high.

The Panel

The Deputy Assistant Secretary for Transportation Technologies convened his R&D program managers and told them to develop a brand new strategic plan, from the ground up. To this committee, he added a couple of policy wonks like myself. For those unfamiliar with Washington jargon, let me explain that policy wonks are just people who try to figure out what to do next, whether or not it might work, and why.

The panel identified some thirty different approaches. After preliminary analysis, we boiled the field down to nine strategic alternatives. These strategies took on an identity of their own, and we gave them names: 21st Century Car, Ethanol, Current Diversified, Long Run Clean, Electric Vehicle, Regulation, Heavy Hitter, etc. Earlier strategies, such as Tree Hugger and UnkUnk (unknown unknowns), didn't make the cut.

Each strategy was then evaluated much more thoroughly in terms of cost, benefit, risk, technological status, employment impact, time-horizon, environmental effects, and other variables. We hired a professional facilitator to lead the planning meetings, and another contractor to do cost-benefit analyses.

In the end, we had pretty good analytic details on nine R&D strategies. Now came the hard part. The basic question: where do we play our R&D chips? About \$200 million/year for transportation R&D is on the table.

The bottom line? In the final analysis, each panelist was asked to rank the nine strategies. These rankings were judgement calls, based on all the data and analysis that we had worked on to date. Panelists were forced to mentally integrate all the information that had been developed during the entire planning process. Ties weren't allowed.

The Results

Here are the stark results. In the 9-by-10 matrix below, each of the nine rows is a strategy option, and each column is the ranking of these strategies by one of the ten panelists.

These aren't quite the original data. They are sorted in order of their increasing row sums ($M \leftarrow M[\uparrow + / M;]$) for reasons that will be explained later.

Rank data are inherently non-parametric. Descriptive parameters such as mean or standard deviation are useless with ordinal data because there's no quantitative measure of the amount by which rank X is preferred to rank Y .

		PANEL MEMBER									
		1	2	3	4	5	6	7	8	9	10
STRATEGY	1	1	1	1	3	1	1	1	1	4	5
	2	6	3	3	2	8	5	2	4	2	1
	3	5	8	6	1	2	3	6	9	1	4
	4	7	4	4	4	5	6	7	5	3	2
	5	2	2	5	7	7	8	5	6	7	3
	6	8	5	2	5	6	9	9	3	6	7
	7	3	6	7	9	4	7	8	2	8	8
	8	9	9	9	6	3	2	4	8	5	9
	9	4	7	8	8	9	4	3	7	9	6

The Role of APL

What can be made of this jumble of numbers? They sit there staring up at us like a fried egg on a plate. The rankings are all different, but how much difference *makes* a difference? How can we separate the signal from the noise in this judgmental process? Here's where APL comes to the rescue (well, sort of).

We can ask at least two interesting questions. First, we may ask: to what extent do panelists agree with one another? Second, we may ask: are the panel's rankings of Strategy X significantly different from their ranking of Strategy Y, when panel data are taken as a whole? Both questions lend themselves to quantitative tests of "significance" in the statistical sense of that word. We can estimate the odds that the results are, or are not, due to random chance at specified confidence levels.

The first estimate, the Spearman Rank Correlation Coefficient (Spearman's Rho) measures the extent to which panelists agree among themselves on the strategy rankings taken in pairs. This is a measure on columns. The second estimate, the Friedman Test, measures the extent to which panel rankings, when taken as a whole, can be considered to be distinct. This is a measure on rows.

In both cases, the comparison is made against random chance—what might have occurred had the panelists used 9?9 to rank their preferences.

How Well Did Panelists Agree Among Themselves?

The table below shows Spearman's Rho for all 45 combinations of ten panelists taken two at a time. That is, the correlation between panelist 3 and 8 (0.67) is shown in the third row and the eighth column (and, by symmetry, in the eighth row and the third column). In a perfect world, all entries would be 1.0. In a random world, individual entries could be anything between -1 and +1, but their overall average would be zero. Spearman's Rho helps to estimate where between these two extremes the data might be.

**“What can be made
of this jumble of numbers?
They sit there staring up at us
like a fried egg on a plate.”**

		PANEL MEMBER									
		1	2	3	4	5	6	7	8	9	10
1	—	0.55	0.20	-0.17	0.10	0.10	0.35	0.37	-0.27	0.18	
2	0.55	—	0.80	0.18	-0.10	-0.20	0.30	0.67	0.08	0.58	
3	0.20	0.80	—	0.53	0.10	-0.17	0.08	0.67	0.38	0.48	
4	-0.17	0.18	0.53	—	0.35	0.38	0.27	-0.10	0.95	0.57	
5	0.10	-0.10	0.10	0.35	—	0.53	-0.02	0.10	0.42	-0.30	
6	0.10	-0.20	-0.17	0.38	0.53	—	0.72	-0.23	0.37	-0.10	
7	0.35	0.30	0.08	0.27	-0.02	0.72	—	-0.02	0.17	0.27	
8	0.37	0.67	0.67	-0.10	0.10	-0.23	-0.02	—	-0.17	-0.02	
9	-0.27	0.08	0.38	0.95	0.42	0.37	0.17	-0.17	—	0.55	
10	0.18	0.58	0.48	0.57	-0.30	-0.10	0.27	-0.02	0.55	—	

I won't show the code for Spearman's Rho here. Statisticians know this stuff already, and non-statisticians would find it to be just messy arithmetic. For reference, I used *“Practical Nonparametric Statistics,”* Second Edition, by W. J. Conover, John Wiley and Sons, New York, 1980.

The APL needed to produce this matrix? A few lines of code using a nested loop, taking panelist 1 with 2-10, 2 with 3-10, etc. (...yes, I used loops—this isn't a story about code optimization). The table is symmetric about its major diagonal because I didn't take time to pretty it up. What you see is what went into my report.

Taken as a group, the average for 45 values of Rho was 0.23, with a standard deviation of 0.31. Taken as individuals however, some panelists show correlation at statistically significant levels.

For this number of strategies, any one value of Rho is significant at the 90% level if it's greater than 0.47 or less than -0.47 (a minus sign indicates a negative correlation). The value 0.47 comes from a table look-up in *“Quantiles of the Spearman Test Statistic.”* Don't worry if you're not familiar with this test. Most texts give an explanation and a recipe. Although the calculations are easy in APL, a knowledge of statistics is helpful in selecting what tests to use (and there are many).

Conclusion? Although there was no significant correlation when panel data are taken as a whole, about one quarter of all individual panel rankings were significantly correlated.

High agreement occurred between panelists 4 and 9 (0.95), between 2 and 3 (0.80), and between 6 and 7 (0.72). Panelist 8 also agrees well with both panelists 2 and 3 (0.67). Panelists 5 and 10 mildly disagreed (-0.30), although not at a significant level.

Lack of overall high correlation probably means that panelists pretty much thought for themselves. A good sign. High correlation between a few individuals could mean any number of things, about which I can't speculate. APL can't solve everything.

Are Any of the Strategies Significantly Different?

To estimate how sharply the panel ranked the nine strategies, I used the Friedman test (named for the noted economist Milton Friedman) on row sums. In the table below, Strategy 1 has the smallest row sum. Strategies 3 and 4 don't seem to be much different, and neither do strategies 6-9. How far apart must the row sums be to conclude that the panel rankings are significantly different?

	PANEL MEMBER										Row sum
	1	2	3	4	5	6	7	8	9	10	
STRATEGY	1	1	1	3	1	1	1	1	4	5	19
	2	6	3	3	2	8	5	2	4	2	36
	3	5	8	6	1	2	3	6	9	1	45
	4	7	4	4	4	5	6	7	5	3	47
	5	2	2	5	7	7	8	5	6	7	52
	6	8	5	2	5	6	9	9	3	6	60
	7	3	6	7	9	4	7	8	2	8	62
	8	9	9	9	6	3	2	4	8	5	64
	9	4	7	8	8	9	4	3	7	9	65

The Friedman test helps to estimate "how much difference *makes* a difference" in the ranking data. In a perfect world the row sums would be 10×19 . This is an uninteresting case because it tells us zero about the strength of those preferences—it's tantamount to one man's opinion. In a random world the nine row sums would, on average, all be 50, because $+ / 19$ is 45, and also because $(+ / + / M) = + / + / [1]M$ is 1 for any numeric matrix M of rank 2.

The Friedman test helps to estimate where between those two extremes the data might be. It turns out that two row sums in the table above are significantly different if they differ numerically by 18, 14, or 7 at the 95%, 90%, and 75% confidence levels, respectively. These values come from a table look-up of the F distribution.

The APL? Just a few lines of code to calculate the row sum differences that correspond to the 75%, 90% and 95% confidence levels. I won't show

it here—it's a space-consuming function of $+ / + / M$, $+ / M$, and ρM .

Strategy 1 is easily the panel's first choice, at a confidence level of almost 95%. The difference between the row sums for strategies 1 and 2 is 17, pretty close to the 95% critical value of 18. Strategy 2 is preferred to Strategy 3, but only at a confidence below 90% (the row sum difference is 9, below the 90% confidence level of 14, but above the 75% confidence level of 7). Iffy. Strategies 3 and 4, and 4 and 5 are indistinguishable.

Rank data usually become mush toward last choice. This is okay. We usually care only about top choices, not last ones. What would we have done if first and second choices were not so quantitatively distinguishable? We would have had to dig deeper to settle the matter.

These may seem like weak conclusions. Yet that's exactly what you get with non-parametric statistics. They're inherently weaker, because rank data contain inherently less information. Yet, among policy wonks, the motto is: "*Something* is better than *nothing*."

I'm not free to identify which strategy is which here. Besides, this article is about APL, not transportation R&D. Symmetrically, my report to the Deputy Assistant Secretary is about transportation R&D, not APL.

The Time Factor

In closing, I want to comment on the notion that APL operates at "the speed of thought." In real-world work situations, it just doesn't. It took a couple of days to get the ranking data (in Word-Perfect) from a contractor who didn't know how to transfer files electronically. It took two days part-time to do the analysis described here, including re-doing the Spearman stuff because I did it wrong the first time. It took another two days to write it up. All in all, it took more than two weeks, off-and-on, squeezed between other work and going to the annual APL conference.

Could other software do it better? I don't know. There was no competition. No one had any staff or software that might have done the job better or faster, and canned spreadsheets don't stand a snowball's chance in this kind of analysis. ■

Dick Holt is an Operations Research Analyst in the Office of Science Policy, U. S. Department of Energy. He received his Masters Degree in Physics from the Johns Hopkins University, and taught advanced calculus and differential equations at the California Institute of Technology.