Computer Output from Program BowlGraphMain.sas

This output illustrates key ideas behind the paper "The Optimal Threshold p-value for a Scientific Journal". The focus is on a two-sample t-test and on the false-positive and false-negative errors that the test makes.

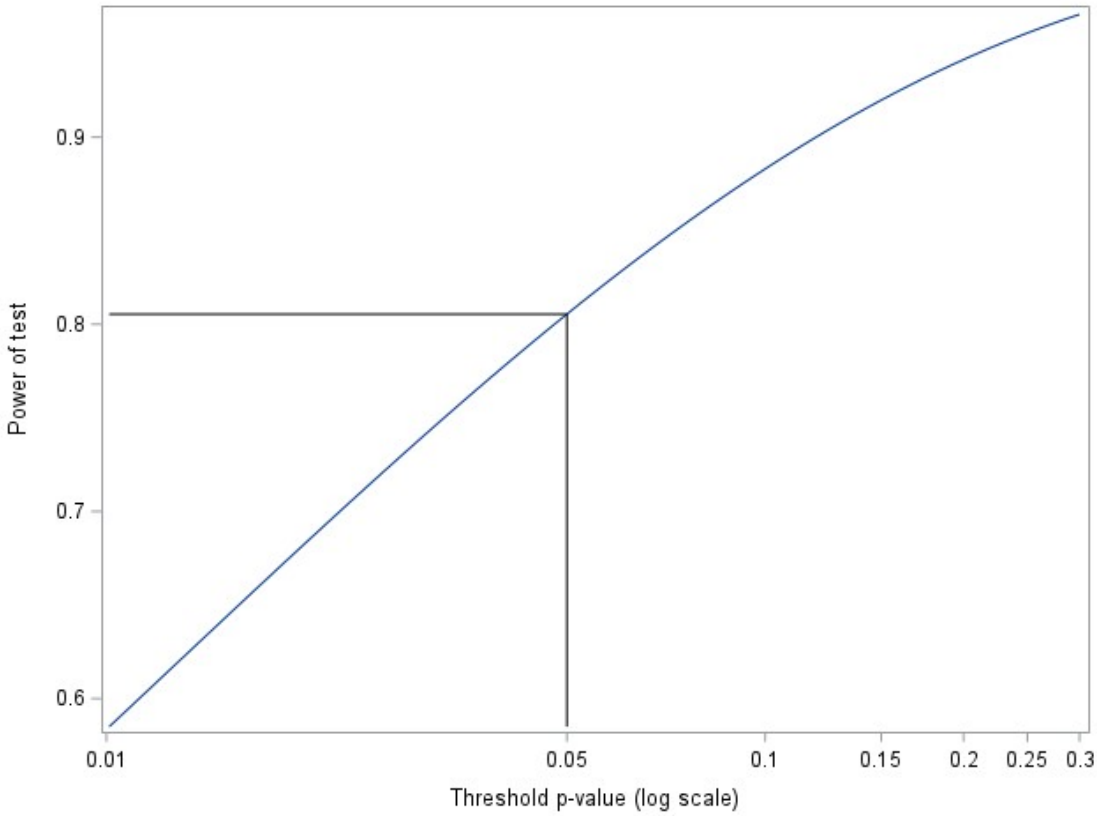
We examine the performance of the two-sample t-test in a specific case in which we assume that the population difference between the means is 20 units and the population standard deviation of the difference is 27 units. We assume that there are 30 entities in each of the two groups for a total of 60 entities in the research. These assumptions enable us to compute the "power" of the test. Here, computed two different ways, is the power of the test when the threshold p-value (which SAS refers to as "Alpha") is 0.05:

First is the output from the SAS Power procedure, which informs us that it is using the conditions specified above and then tells us the power of the test under the specified conditions.

Analysis	Index	Sides	Alpha	MeanDiff	StdDev	NPerGroup	NullDiff	Power	Error	Info
TwoSampleMeans	1	2	0.05	20	27	30	0	0.805412317043370		

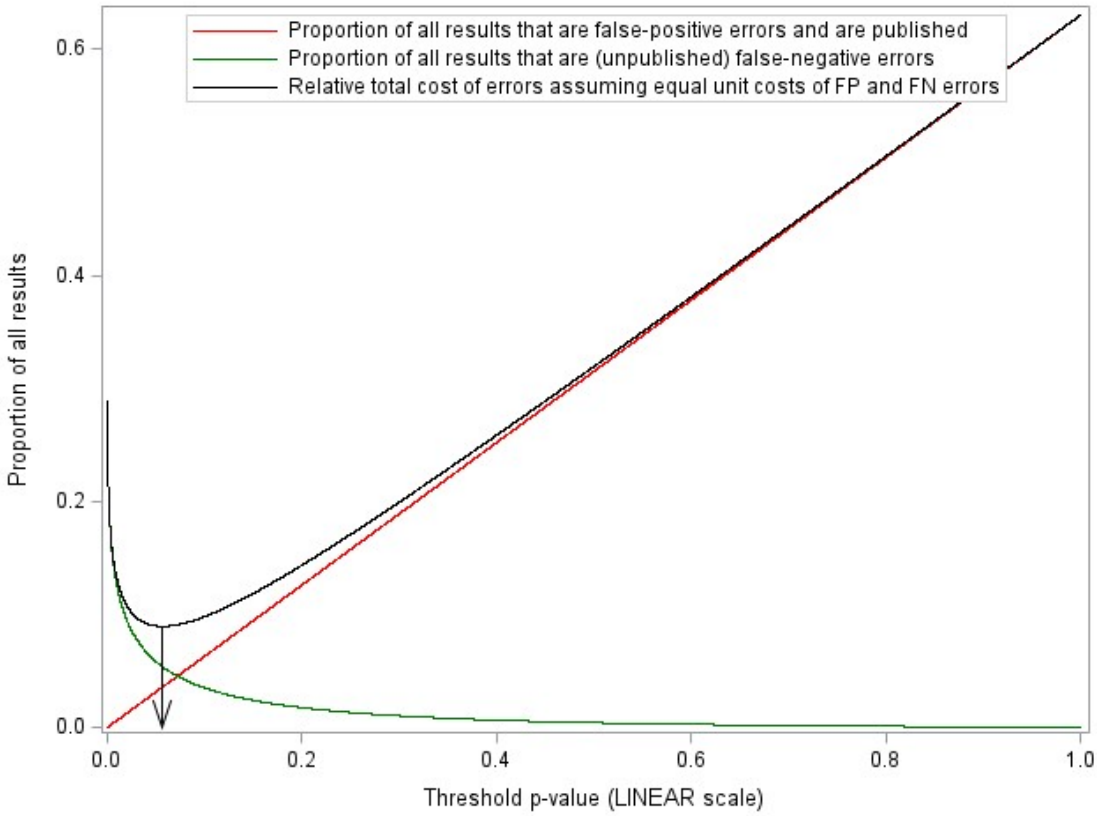
Next, to check for consistency, we can compare the power above with the power computed directly by the three lines of code in this program when the threshold p-value is 0.05, which (as extracted from the data file of this run of the program) is 0.805412317043370.

Note that the values 20, 27, and 2 x 30 = 60 were intentionally chosen to make the power roughly 0.8 when the threshold p-value is 0.05. This choice was made because it makes the situation under study resemble the ideal typical situation in which the power of the test is 0.8 for detecting the expected form of the studied effect. Of course, even higher power would be better, but that would be costly, so we must compromise to control research costs.



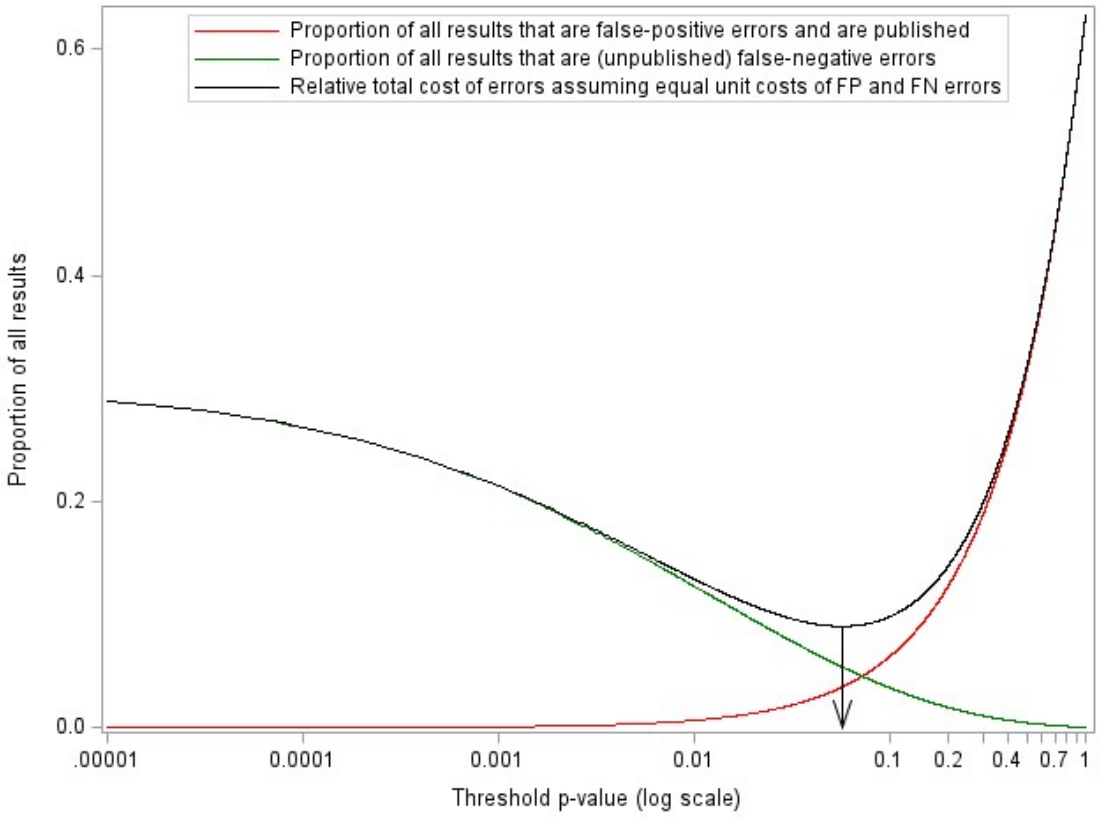
The above graph indicates the power of the two-sample t-test computed at each threshold p-value between 0.01 and 0.3, with the value at 0.05 highlighted, showing that the power of the test at this threshold p-value is roughly 0.805, as noted above.

The graph tells us the power of the test under the specified conditions (i.e., mean difference of 20, standard deviation of 27, number in each group of 30). However, ANY other standard statistical test for evidence of an effect will have a similar smoothly monotonically increasing line on the graph, though the location and shape of the line on the graph will vary from test to test. The relationship between the threshold p-value and the power is of interest because this relationship determines the locations of the green lines on the graphs below.



Here is figure M.6 from the paper, but with a LINEAR horizontal axis as opposed to a logarithmic horizontal axis, and using (almost) the entire possible range of potential threshold p-values between 0.0 and 1.0. Note how the false-postitive line (red line) is a straight-line relationship, as can also be seen in the simple algebra in the program that generates the data behind the line.

On the above and following graphs, the vertical arrow that runs down from the minimum point on the curving black line shows the theoretical optimal threshold p-value of roughly 0.06 for the journal for this specific research situation. This will become clearer when we zoom in on the arrow, which we do in a moment.

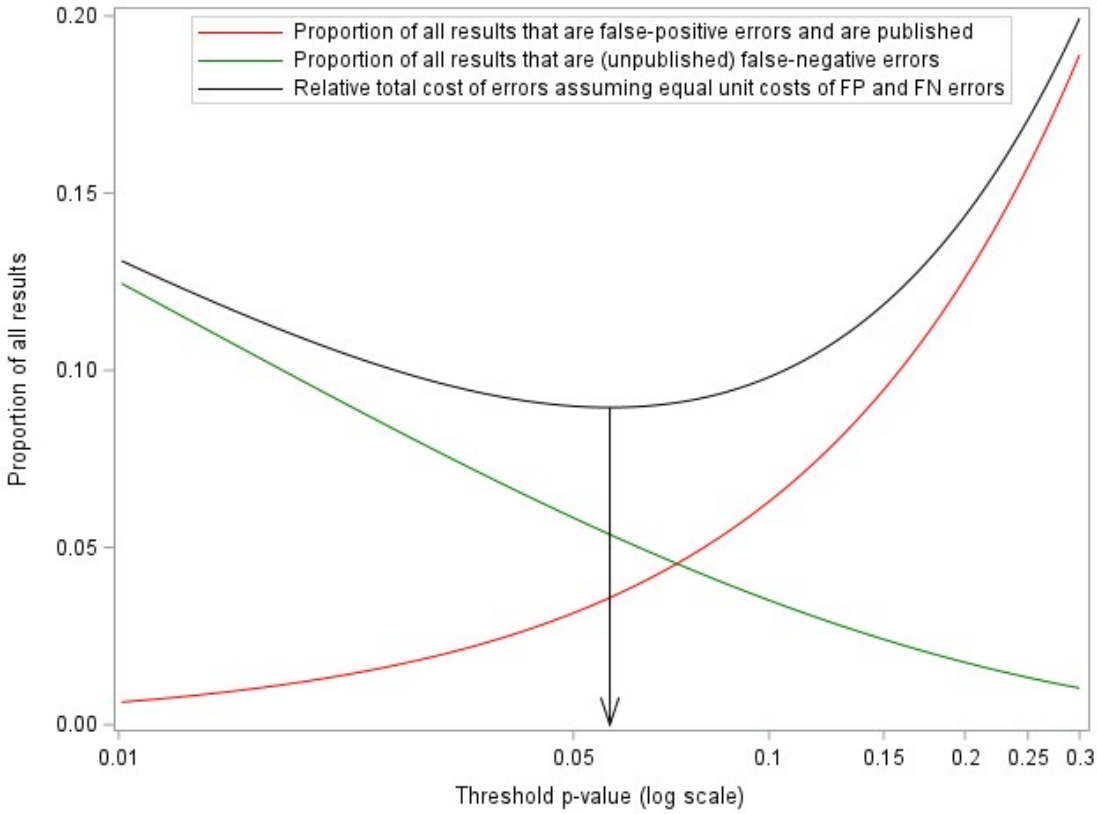


The above graph shows exactly the same data as the graph that precedes it, but using a logarithmic scale for the horizontal axis instead of a linear scale. This scale sensibly stretches things out at the lower end. Note how changing the axis scale has changed the false-positive (red) line from a straight line to a curved line. Note how changing the axis scale DOESN'T change the location on the horizontal axis of the minimum point of the bowl of roughly 0.06.

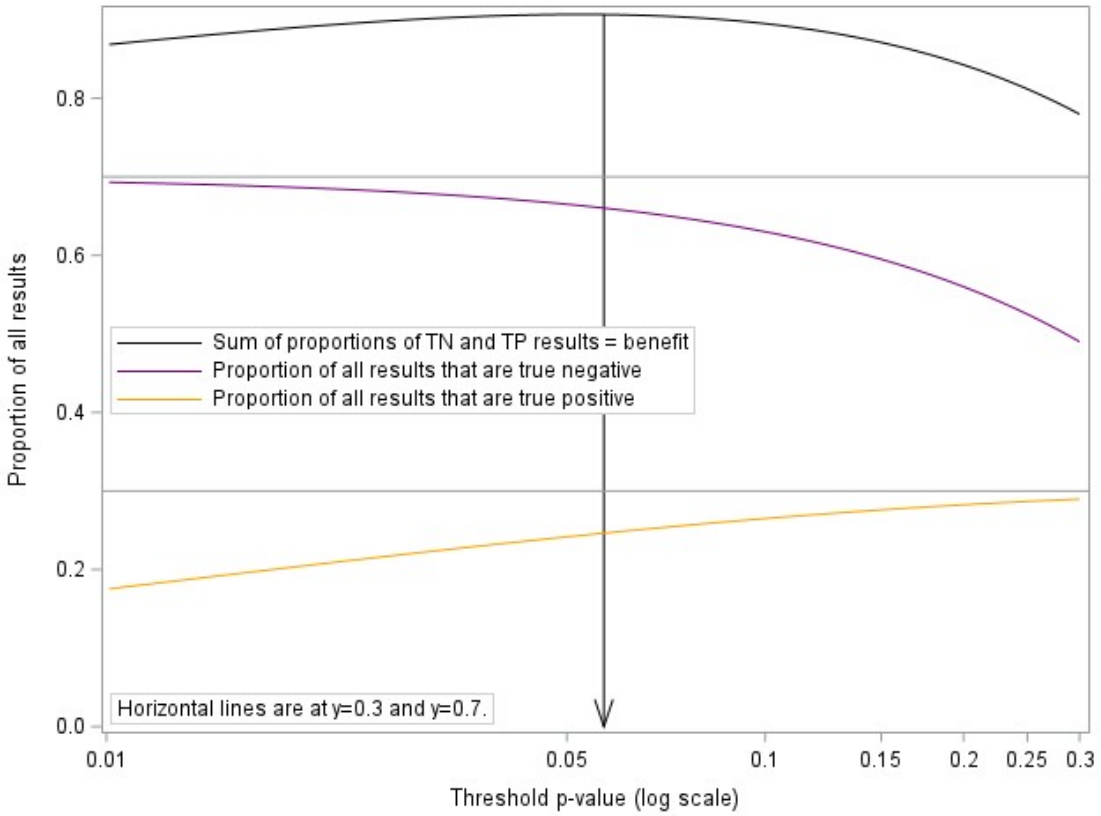
Note how the false-negative (green) line is a sigmoid function, but the false-positive (red) line is not.

Note how the black line theoretically peaks at the left end at 0.3, as defined by the value of PctTrue in the program. The black line peaks at the right end at 1 - 0.3 = 0.7 if PosPctPub is 100. It peaks at a little less than 0.7 if PosPctPub is less than 100. The current value of PosPctPub in this program is 90.

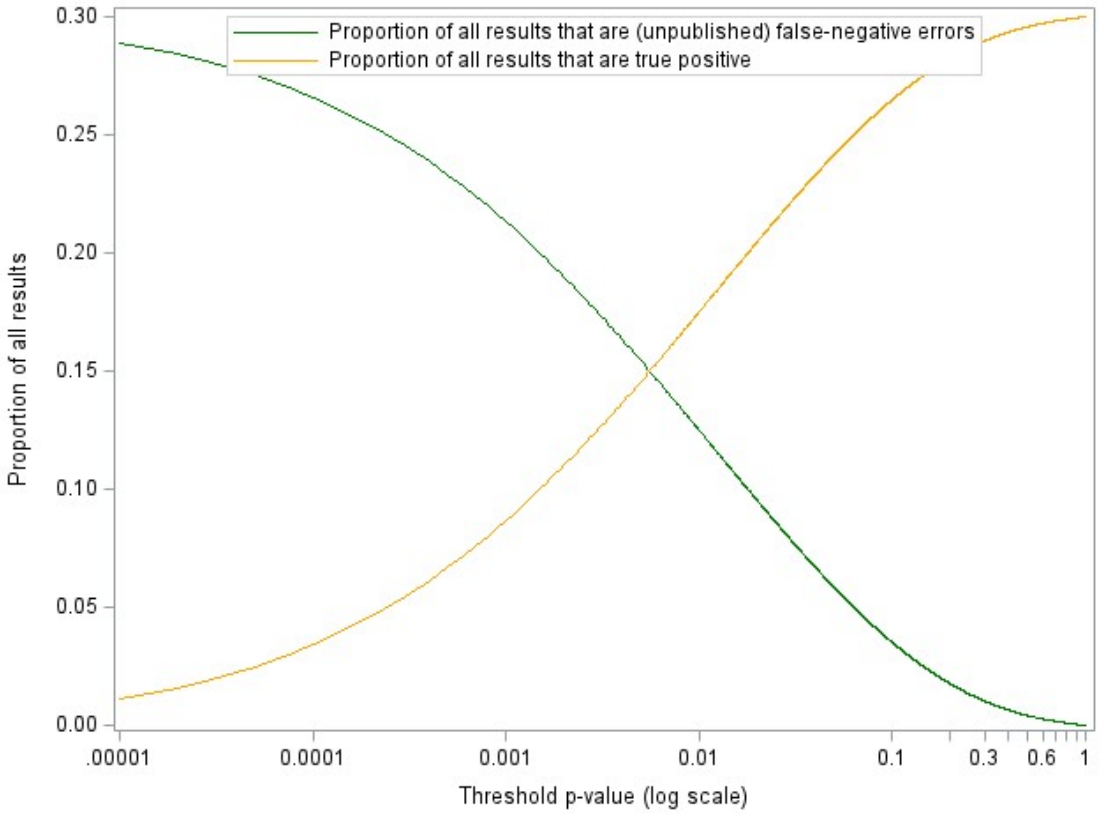
All the following graphs use a logarithmic scale on the horizontal axis, but we could draw the same conclusions if the scale were a linear scale, though data at the low end of the horizontal axis would be somewhat squished on the graphs in that case.



The above graph is figure M.6 in the paper. This graph is based on the same data as on the earlier graphs with the red and green lines. However, this graph is zoomed in on the small area of the horizontal axis of main interest. We can see more clearly here how the lowest point on the bowl is at roughly 0.06 on the horizontal axis. (The actual optimal threshold p-value extracted from the data file is 0.0569.)

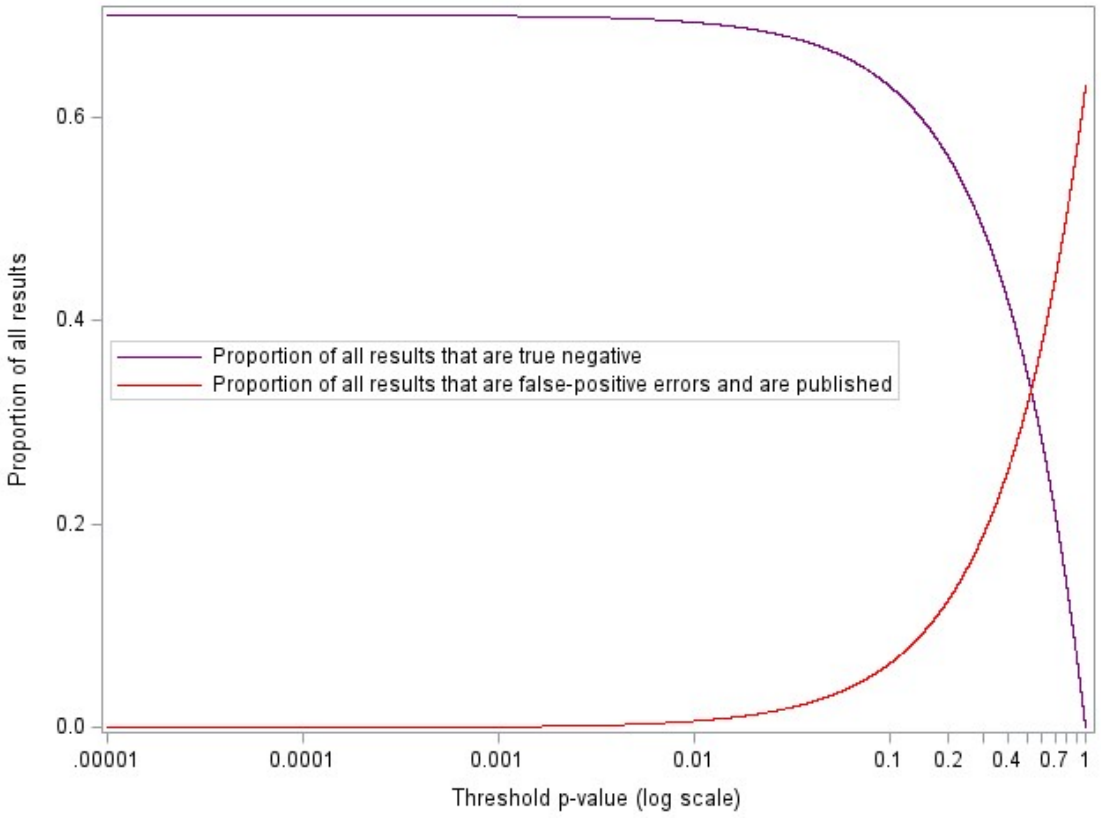


The above graph shows that we get the same optimum threshold p-value of roughly 0.06 on the horizontal axis by finding the point that maximizes the sum of the "benefits" from the true-positive and true-negative results. However, the "minimize the sum of the error costs" view seems more intuitive than the "maximize the sum of the benefits" view.



The mirror image of the green line is a meaningful line with a name -- it is the proportion of all the results that are true positive results. As also implied in an earlier graph above, the green line and its mirror image are sigmoid functions that peak (at opposite ends) at 0.3 on the vertical axis, as shown in the above graph. This is because since only 30% of the research studies are studying true hypotheses, therefore at most only 30% of the results can possibly be true positive results or false-negative errors.

Note that the yellow-orange line on the above graph and the yellow-orange line on the graph that precedes it are the same line. The yellow-orange line on the first of the two graphs is a snippet from the yellow-orange line immediately above. (The two yellow-orange lines appear to have different slopes because the scale on the vertical axis of the second figure is an expanded version of roughly a third of the scale on the vertical axis of the first figure.)

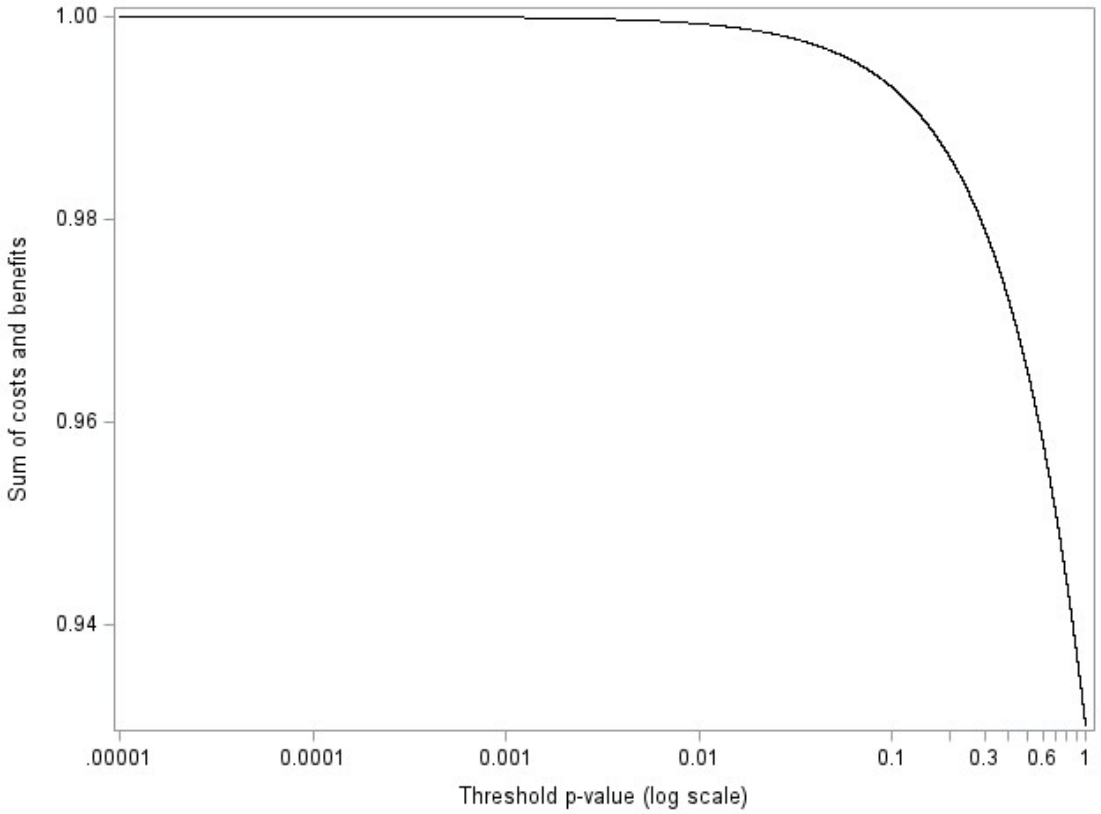


Above are the red line from the earlier graphs and its mirror image, which is the proportion of all the results that are true negative results. Using the value of PctTrue, which is 30, the red line peaks at 1 - 0.3 = 0.7 if the maximum threshold p-value on the horizontal axis is 1 and if PosPctPub is set at 100. The current value of PosPctPub in this program is 90.

The reason why the lines peak at 0.7 is that 100% - 30% = 70% of the research studies are studying relationships between variables that DON'T exist in the population, so at most 70% of the results can be true negative results or false-positive results.

The purple line on the graph three graphs above is a snippet from the purple line on the graph immediately above.

The line on the following graph is the sum of the "costs" and "benefits" computed by the program at each threshold p-value. This line will be horizontal at 1.0 on the vertical axis if PosPctPub equals 100. The current value of PosPctPub in this program is 90. This graph is a way of confirming that things are working as expected.


1.00
0.98
0.96
0.94
Sum of costs and benefits
.00001
0.0001
0.001
0.01
0.1
0.3
0.6
1
Threshold p-value (log scale)