Combining P-values

This posting is inspired by a question from a student. The question raises a common issue in what is called *meta-analysis*. Quoting from the Wikipedia entry,

> Meta-analysis combines the results of several studies that address a set of related research hypotheses.

There are many ideas and approaches to meta-analysis; in this posting I will address only one idea: How to combine P-values. And there are many ways to combine P-values; I will focus on just one method.

Here is the idea. Suppose that at $k = 10$ medical centers studies are done to compare a new treatment 'B' to a placebo for patients with some form of cancer. In each study the null hypothesis is that treatment B and the placebo are equally effective and the alternative is that treatment B is superior.

Note the following. The discussion below is equally valid regardless of the type of response: dichotomy, multi-category or numerical. Also, it does not matter whether the alternative is one- or two-sided.

Here is what does matter: The (null) sampling distribution of the test statistic is given by a pdf; it does not matter which pdf. The pdf could be the standard normal curve, a t-curve, a chi-squared curve or any of the many curves we have not studied.

To be specific, suppose that, after sorting, the $k = 10$ P-values are:

0.0645	0.0910	0.1148	0.1315	0.1319
0.1508	0.1537	0.1644	0.1766	0.1940

Let's examine these P-values. First, none of them is small enough to achieve statistical significance because each one is larger than 0.05. On the other hand, they are all pretty small, so we have the feeling that overall they support the alternative, that treatment B is superior to the placebo. But how do we measure this?

To date, we have focused on a single P-value as a measure of the strength of the evidence in support of the alternative. Now, with ten P-values, we need to think about the distribution of P-values. To this end, we have the following basic result.

On the assumptions that the null hypothesis is true and that the sampling distribution is given by a pdf, the distribution of P-values is uniform on the interval 0 to 1.

Recall that the uniform distribution is the pdf that is shaped like a rectangle. Thus, the pdf in question is the rectangle located on the interval $[0, 1]$. If you remember that a fundamental feature of a pdf is that its total area is one, then we see that the height of this pdf equals one. In words, this uniform distribution describes the result of selecting a number (a measurement) at random from between 0 and 1.

Looking back at our ten P-values, I am encouraged. Our values do **not** appear to be selected at random from the interval $[0, 1]$; they are too concentrated below 0.20.

The question becomes: How to we combine the information in these ten values to get one overall measure?

Here a math result comes to our aid. Suppose that X is a random variable with the uniform distribution on $[0, 1]$. Then the distribution of $Y = -2\ln(X)$ is chi-squared with two degrees of freedom.

Thus, if we have k P-values from different studies (and hence are independent):

$$X_1, X_2, \ldots X_k,$$

and we define

$$Y_i = -2\ln(X_i),$$

then the random variables

$$Y_1, Y_2, \ldots Y_k$$

are i.i.d. each with the chi-squared distribution with two degrees of freedom.

The final math result we need is that if we sum these i.i.d. chi-squared variables,

$$W = \sum Y_i$$

then the distribution of W is chi-squared with $2k$ degrees of freedom.

I will illustrate the utility of these ideas with my ten phony P-values above. For each of the ten P-values, the X_i's, calculate twice the negative of its natural log. Doing this we get:

5.482	4.794	4.329	4.057	4.051
3.784	3.746	3.611	3.468	3.280

If we sum these ten values we get $W = 40.602$.

To obtain our 'overall P-value' we just calculate the area under the chi-squared curve to the right of W. For our case, I go to the online chi-squared curve calculator located at

`http://stattrek.com/online-calculator/chi-square.aspx`

and enter 20 for degrees of freedom and put $W = 40.602$ in the second (CV) box. The overall P-value appears in the third box; it is 0.004.

Exercise: Find the overall P-value for the following six P-values: 0.0412, 0.0523, 0.0911, 0.1011, 0.1267 and 0.4500.

Answer: $W = 27.385$ and the overall P-value is 0.007.

Exercise: Repeat the above exercise after deleting the largest P-value from the set of six.

Answer: $W = 25.788$ and the overall P-value is 0.004.

Note: I want to thank Vladimir Babak at the Veterinary Research Institute in the Czech Republic for finding an error in my original version of this document.