Note on Look Elsewhere Effect in Exclusion Testing

Richard Lockhart, Simon Fraser University

August 2, 2010

In this note I consider the look elsewhere effect for exclusion testing. The conclusions are these:

- When testing multiple hypotheses a potential look elsewhere effect arises if it is possible for more than one of the null hypotheses to be correct at the same time.
- If it is *known* that at most k of these hypotheses can be true simultaneously then the relevant trials factor is at most k.
- In particular in a scan over some variable such as mass to exclude some particle which cannot exist at more than one such mass no multiplicity effect arises.
- This must be understood via the confidence set — hypothesis testing duality. The problem is not quite the same as testing the null hypothesis that a particle exists somewhere within a pre-specified range.
- There are other similar sounding problems where the frequency theory treatment is more complex. I consider the effect of somewhat different inference protocols. In particular I compare exclusion without testing for discovery (Protocol 1 below) with exclusion after testing for discovery (Protocol 3 below).

I consider a relatively simple context. The experimenter considers just two masses m_1 and m_2. At each mass she tests the null hypothesis that the Higgs particle is present at that mass (with the cross-section predicted by the standard model – the null hypothesis that it exists at any cross section at all at that mass is not testable by frequency methods). So there are two test statistics: T_1 and T_2 and corresponding critical values which produce level α tests of the two null hypotheses against the alternative that the particle does not exist at that mass.

Protocol 1: The analyst will declare as *excluded* each mass for which the null hypothesis is rejected at level α. In particular the analyst will declare the pair of values to be excluded if both null hypotheses are rejected by their corresponding test statistics. Let A_b be the (statistical) event defined by 'the analyst excludes both values' and for $i = 1, 2$ let A_i be the event that the analyst announces that mass m_i is excluded. If neither test rejects then the analyst says neither value is excluded; call this outcome A_n.

The frequency theory conclusion is that the type 1 error rate for this protocol is α – the probability that an exclusion statement is made and that this statement is wrong is no more than α regardless of which of three possible states of nature obtains. The three possiblities contemplated are: the Higgs particle exists at mass m_1; it exists at mass m_2 and it does not exist at either of these two masses.

I now want to be quite careful in the analysis. I consider in turn each of the three possible states of nature and evaluate the probability that I make an incorrect exclusion.

First if the particle does not exist at either mass I *cannot* make an incorrect exclusion; exclusion is correct at both masses. The probability that I incorrectly exclude the correct mass is 0.

If the particle exists at mass m_1 then I make an incorrect exclusion if either I say that I exclude mass m_1 or I say I exclude both masses m_1 and m_2. I make one of these two statements if and only if statistic T_1 causes me to reject the null hypothesis at mass m_1; the probability that this happens is α. The same argument applies for mass m_2. Notice that I am saying that the chance of A_b or of A_1 happening is α. In protocol 3 below I compare the probabilities of all 4 events for this protocol to those for the more realistic protocol 3 in one simplified example.

Summary: the probability that I make *any* incorrect exclusion is no more than α.

There is nothing special here about the choice of 2 hypotheses. The conclusion is the same even if I scan an infinite number of masses – say by demonstrating that all the test statistics for masses in some range would be significant.

The major point is that it is impossible for more than one of the null hypotheses to be true. In *any* such situation the set of rejected hypotheses might be called an exclusion set; its complement is a confidence set – except that here the possibility that none of the models here is correct is permitted. The coverage probability of the confidence set is $1 - \alpha$ – if the particle does exist at one of the masses scanned then the chance that the confides set includes that mass is $1 - \alpha$. The probability of an incorrect exclusion is α (if all the individual tests have level α). Notice that this is all just a rephrasing of the duality between confidence sets an hypothesis tests.

Protocol 2: A different problem arises if the set of masses to be examined is prespecified and this whole set is regarded as a null hypothesis. In this protocol the analyst computes a single test statistic to test the null hypothesis that the Higgs exists and its mass is one of m_1 and m_2. The analyst will either declare that the hypothesis has been rejected – mean the Higgs does not exist at either of these two prespecified masses — or accept the null hypothesis that it exists at one of those two masses. This classical hypothesis testing paradigm does not permit partial rejection of the kind where the analyst says it does not exist at mass m_1, say.

The two tests described for protocol 1 must somehow be combined into a single test of the composite null hypothesis and a suitable critical value derived for this combined test. The level is computed by calculating the probability that the combined statistic exceeds this critical level assuming in turn that the true mass is m_1 and that it is m_2. The larger of these two probabilities of rejection is the level α of the test.

This is a lot more work than Protocol 1. The extra work is required if you don't find halfway measures useful. In the Higgs particle context I do not see the value of this protocol.

Protocol 3: In fact the data are going to be analyzed in a different way entirely, I think, and this makes difficulties for frequentists. It is likely that you will look at a range of masses and test first the null hypothesis that the Higgs particle does not exist anywhere in this range. If this composite hypothesis is accepted – that is, you haven't already concluded it does exist – then you will go on to try to follow Protocol 1. The motivation for doing so is, of course, that acceptance of a null hypothesis is a very weak form of decision making.

The difficulty now is that the decision to use protocol 1 depends on the data – you only use it if you accept the first null hypothesis – that it doesn't exist anywhere in the range examined. The impact of this difference depends on the true state of nature:

1. The Higgs particle does not exist at either mass: With a 5σ standard for discovery you will almost always accept the null hypothesis of no Higgs and so the probabilities of outcomes A_b, A_1, A_2 and A_n above are virtually unchanged. In any case false exclusion rates are 0 because the particle does not exist.

2. The Higgs exists at mass m_1: The impact of discovery testing depends on how well separated the null and alternative hypotheses are.

   (a) If the tests are not very sensitive at mass m_1 because the cross-section at that mass is low then discovery is unlikely. As in the case where the particle does not exist the probabilities of the various outcomes are not much changed. Now, however, outcomes in which mass m_1 is excluded constitute errors. The overall error rate (the sum of the probabilities of A_b and A_1) is slightly less than guaranteed by the 2σ standard.

   (b) If the tests are very sensitive (have high power against the cross-section predicted at the masses examined) then having accepted the null hypothesis it is likely that every value in the alternative will be rejected – particularly in view of the 5σ rule for discovery and the 2σ rule for exclusion. In other words if the mass exists and the predicted cross section amounts to a 7σ difference or more then if the particle is not discovered it will be incorrectly excluded. Fortunately the power is high so that the incorrect exclusion rate remains down below 5%.

Summary: False exclusion rates remain controlled but if a discovery is missed the *conditional* error rate goes way up.

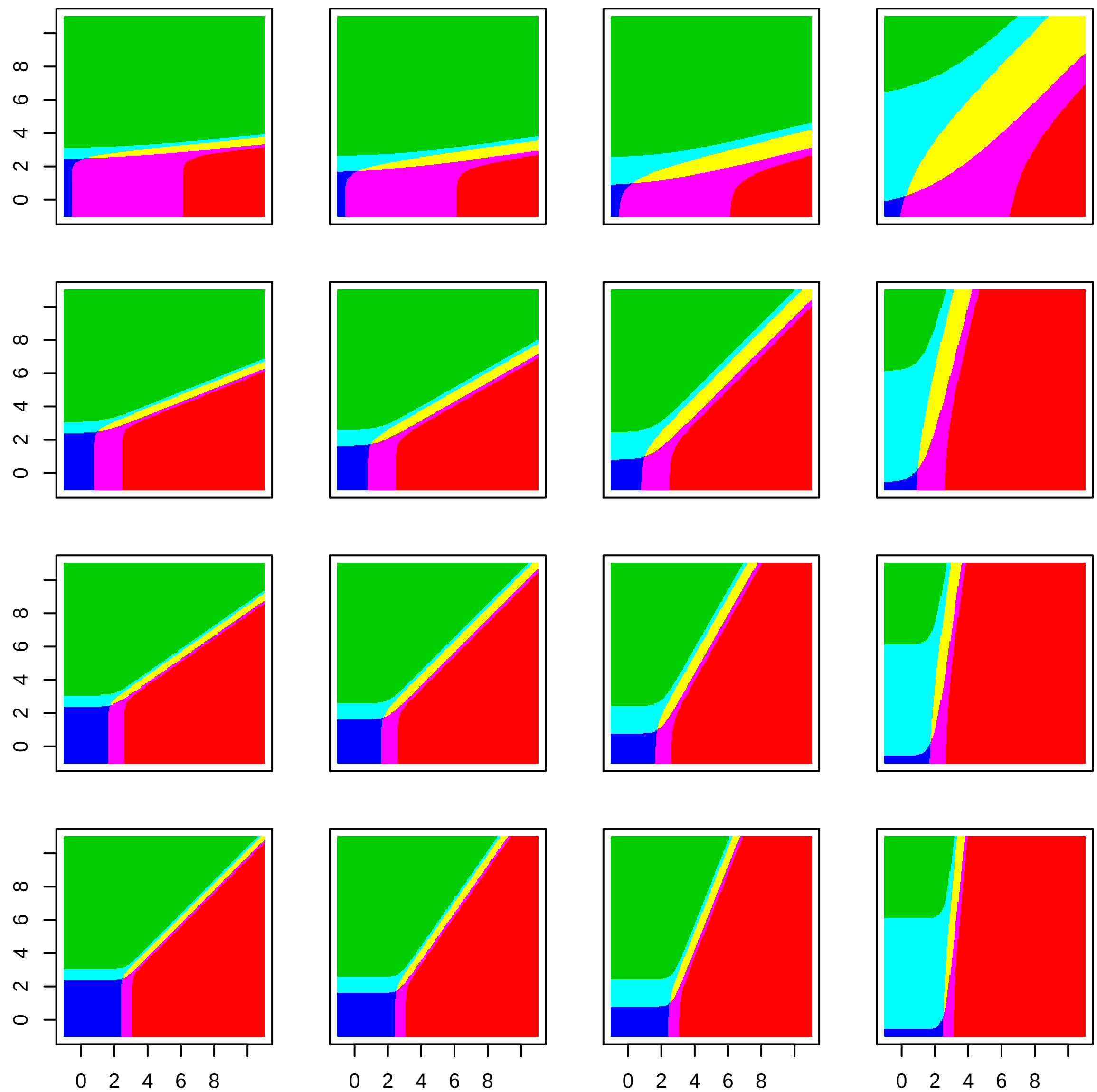
Protocol 4: One of the main topics of discussion in Banff was the notion that it is wrong to exclude a mass when there is no sensitivity at that mass. It is not possible to capture this concept within the hypothesis testing paradigm. Consider the following decision theory problem, however. You study two masses again and make one of the following decisions: discovery at mass m_1, discovery at mass m_2, no discovery and exclusion at mass m_1, no discovery and exclusion at mass m_2, no discovery and exclusion at both masses, no discovery and no exclusions. Call these events D_1, D_2, A_1, A_2, A_b and A_n. The states of nature which are possible are: no Higgs at either mass, Higgs at M_1 and Higgs at m_2. A loss function might be specified in a table like the following

	No Higgs	Higgs at m_1	Higgs at m_2
D_1	L_{fd}	0	L_{wd}
D_2	L_{fd}	L_{wd}	0
A_b	0	L_{fe}	L_{fe}
A_1	L_{me}	L_{fe}	L_{md}
A_2	L_{me}	L_{md}	L_{fe}
A_n	$2L_{me}$	$L_{md} + L_{me}$	$L_{md} + Lme$

In the table fd is 'false discovery', fe is 'false exclusion', wdd is 'wrong discovery', me is 'missed exclusion', and md is 'missed discovery'. The losses for 'missed exclusion' and 'missed discovery' are small; they amount to missed opportunities. The losses for 'false discovery' is very large if the 5σ rule is to be taken seriously. The loss for 'false exclusion' must be smaller than this though there is an argument to be made that exclusion is to be permitted more often because it is thought, a priori, that existence at any particular mass considered is unlikely.

I present here the Bayes procedure for a loss function of this form. I assume that there are two test statistics, T_1 and T_2, one for each mass, which are independent and have Gaussian distributions with mean 0 if the particle does not and with mean μ_i if the particle does exist at mass m_i. I have taken $L_{me} = 1$, $L_{fd} = L_{wd} = 10$, $L_{fe} = 5$ and $L_{md} = 2$ for illustration. I have evaluated for each value of μ_1 and μ_2 belonging to $\{0.5, 2, 3.5, 5\}$ the set of pairs (T_1, T_2) for which each of the 6 possible decisions will be taken. I used prior probabilities of 1/2, 1/4 and 1/4 for the three states of nature. In the figure the rows correspond to μ_2 with $\mu_2 = 0.5$ at the bottom. Columns correspond to μ_1 with $\mu_1 = 0.5$ on the left. The areas in green and red correspond to values of the statistics leading to discovery claims, those in blue lead to both masses being excluded, those in yellow to neither being excluded and the narrow cyan and magenta bands to areas where 1 mass is excluded but not the other.

It is possible to take this procedure and ask about its frequency properties. This would involve computing the probabilities of the 6 decisions as functions


0 2 4 6 8
0 2 4 6 8

of the parameters for the procedures shown above. I have not tried to do this yet.

It is worth noticing that large negative values of say T_1 lead to exclusion. It is clear that physicists are leery of exclusion in this situation – they feel that they should not be doing exclusion when the hypothesized value μ_1 is low and sensitivity is low. The problem is that for an observation of $T_1 = -3$ the likelihood at $\mu_1 = 1$ is *much* lower that the likelihood at 0. *If the model is right and $\mu_1 \geq 0$ is absolutely certain then mass m_1 should be excluded.* But the problem is that $T_1 = -3$ casts substantial doubts on the premises of that statement and so we should be reluctant. Nothing in the formal framework here or in hypothesis testing contemplates the effects of model failure.

Other ideas: The problem of testing for discovery and exclusion of Higgs in SUSY was discussed in Banff. This theory predicts, I think, that there are $k = 5$ Higgs particles so that if you consider a range of masses you could make up to $k = 5$ false exclusions. There is now a multiple comparisons problem but it is controlled by dividing the desired overall false exclusion rate by k. Suppose, for instance, that I want an overall 2σ or better false exclusion rate. If I use a 2.608σ standard for each mass I consider, and I consider any number of masses then:

$$P(\text{any false exclusions}) \leq \sum_{i=1}^{5} P(\text{false exclusion at mass } m_i) \leq 0.02275$$

where the m_i are the 5 true masses. (The number 0.02275 is the 2σ one-sided error rate.)

The key point is that the number of possible false exclusions bounds the trials factor.