Multiple t-tests vs. one-way ANOVA

I'm working on a classification problem and I have a very high F1 baseline of 85%. I have trained three classification models and I want to know which one is the best. How can I do so?

I tried two ways:

1. To compare each model against the baseline using paired t-test. So I have tests like:
   - baseline vs. model 1
   - baseline vs. model 2
   - baseline vs. model 3

   That tells me that only model 1 is significantly higher than the baseline and so I concluded that model 1 is the best. Is this a valid methodology given that usually classification models are compared against baselines?

2. To compare all models in one fell swoop with one-way ANOVA. So I entered the information of models 1-3 and the baseline which gave me a p-value of 0.02 indicating that there is a difference in means. Yet, with pairwise post-hoc tests, there is no significance between any of the pairs.

Which method is the correct one?

anova classification statistical-significance t-test multiple-comparisons

edited Jun 9 '13 at 14:07

asked Jun 9 '13 at 6:55
Sabba
30 1 5

2 Answers

If your goal is to see which methods are better than the baseline, then method 1 is correct. If your goal is to see which methods are better than each other, then method 2 is correct. Method 2 with Dunnett's test could also be used, this makes corrections for the multiple comparisons involved. However, the question of whether you need to make these corrections is subject to debate and differing opinion. The tag "Multiple comparisons" here on Cross Validated will find lots of posts on that subject.

answered Jun 9 '13 at 11:29
Peter Flom ♦
54.2k 9 58 136

One-way or two way ANOVA. Also, should I use the between groups MS and df or the within groups MS and df? – Sabba Jun 9 '13 at 14:30

One way, since you have only one independent variable. The other question should be taken care of by the software – Peter Flom ♦ Jun 9 '13 at 14:51

Do you recommend a specific software? – Sabba Jun 9 '13 at 15:00

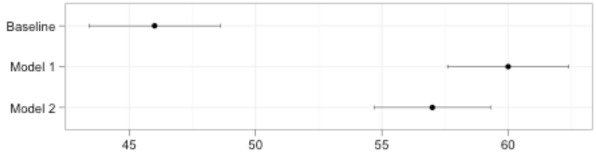
My favorite stat software is SAS, but that is very expensive. R is also good and can do this. Really any stat package should be able to. – Peter Flom ♦ Jun 9 '13 at 21:07

There are two problems with your first approach: Multiple testing and the interpretation of the difference between significant and non-significant.

First, when you perform several tests, you increase the overall error rate. If you set $\alpha = .05$, you will still reject the null hypothesis 5% of the time when it is actually true (i.e. when there is no difference whatsoever). However, the probability that you reject the null hypothesis at least once when performing several tests is higher because you are taking this 5% risk each time you run a test. The problem is not very acute with only three tests but it still does not make much sense to run a test if you don't care about controlling the error level.

A quick fix for this problem is to adjust the error level with the Bonferroni correction. Since it is based on a very general probability inequality, it doesn't matter what the tests are and you can always use it but you will loose power. For example, it's possible that, after applying the correction, none of your tests would be significant and you would be back to the results of your second approach. This would resolve the apparent inconsistency but would not be terribly useful. Generally speaking, the main problem with this technique is that yo are not using all the information you have and the correction is usually too conservative. That's one reason to choose an ANOVA approach when you have several groups.

The second problem is that the difference between significant and non-significant is not necessarily itself significant. If you rejected the null hypothesis that model 1 is equal to baseline but not the separate hypothesis that model 2 is different from baseline, you still haven't established that model 1 is different from model 2. Those are three different questions. Let's see how this could happen:



Now, let's imagine that the difference between baseline and model 1 is “barely” significant, say *p* = .04. The *p*-value from a test of the difference between baseline and model 2 would then be a little over the threshold, say *p* = .06, so not significant. Yet, at the same time, model 1 and model 2 appear very similar and the difference between the two is also obviously not significantly different from 0.

The thing is that the logic of statistical testing requires us to specify an error level but there is nothing exceptional about this threshold. We simply have a little less evidence than the score for model 2 is higher than baseline and maybe this evidence is not sufficient to rule out the null hypothesis at the specified error level. This is however not enough to conclude that it is different from model 1.

Ignoring the multiple testing issue, you could then conclude that you don't know if model 2 is in fact better or worse than baseline and you certainly don't know if it is better or worse than model 1. Intuitively, if your data indeed look like the data on the graph, I don't find this very satisfying because it means treating model 1 and model 2 differently based on evidence that is much thinner than the evidence that model 2 is in fact better than baseline. However, this type of thinking is quite different from the logic underlying statistical tests. Instead of blindly make binary decisions based on the tests, I would therefore rather look at some graphs and make a judgment call about the results.

Either way, the results you presented suggest that your models do in fact represent an improvement over the baseline but there is no way you can conclude that model 1 is better than model 2. This is also pretty much what you can conclude from the ANOVA. If you want to know more, you need to look carefully at the size of the difference and probably collect more data/test the models on a larger data set.

PS: All this ignores two potential additional problems that were not raised in the question, namely the nature of your response variable (is it a proportion?) and independence (are all the models tested on the same exemplars?) Depending on the answers to these questions, ANOVA/T-test might not be the best choice anyway (see also this previous question). Also, if you specifically want to compare each group to baseline, you can also achieve that in an ANOVA framework by using contrasts. You would have a theoretically sound approach with more power than post-hoc pairwise tests but still would not address the “which one is best” question.

answered Jun 9 '13 at 11:38
Gala
6,334 2 16 33

So it is more meaningful then to make pairwise significance tests between the models themselves model 1 vs. model 2 model 1 vs. model 3 model 2 vs. model 3 Would that be a better approach? As for independence, yes all models tested on the same exemplars. As for the nature of the variable, it's a group of exemplars with nominal labels. – Sabba Jun 9 '13 at 14:15

1 Well, it depends what you want to know. Comparisons with a baseline also seem meaningful but they won't tell you if one model is better than the other. The most important point is that you can't compare two models based on the significance of differences to the baseline. If you want to test if one model is better than the other, you have to look at the relevant difference directly but that's already what the post-hoc tests in your ANOVA are doing. – Gala Jun 9 '13 at 14:40

1 For the rest, it seems that McNemar's test (or, since you have several models, Cochran's Q) could be relevant. For more details, you can look at the question I mentioned in my answer or ask a follow-up question about this point. – Gala Jun 9 '13 at 14:48