Analysing rankings with the sign test, using p-values conditional on the rank order of the sample

Lars Bohlin [*]

April 18, 2016

Abstract

This paper deals with the problem of making inference from a survey question where the respondents are asked to rank a couple of objects from the best one to the worst. Pair wise sign tests could be used to investigate what object/objects that differs from the other but there is a problem to choose the combinations to be tested. The standard recommendation in statistics is to make these choices before the data is collected to avoid that the choice of method will depend upon the data in a specific sample. We show that the rejection frequencies under a true H_0 differ quite a lot from the level of significance if the choice of objects is dependent on the rank order of the sample and conventional p-values are used.

A method to calculate the p-values of the sign test conditional on the mean ranks in the specific sample is developed. The advantage with this methodology is, that the choice of objects to be compared may be done after the descriptive statistics are calculated and still yield consistent p-values. Since it is fairly difficult for the referees to control how authors choose what tests to make, results based upon conditional p-values would be much more trustworthy than results based on conventional p-values. The method of conditional p-values is evaluated in comparison with a strategy where all pairs of objects are tested and the p-values are adjusted with the Holm-Bonferoni method to avoid a too high family wise error rate.

Keywords: non-parametric methods, sign test, data mining, conditional p-values
JEL classification:

[*] CORRESPONDING AUTHOR. Mälardalen University College, P. O. Box 833, S-721 23 Västerås, Sweden. E-mail: lars.bohlin@mdh.se.

1 Introduction

One important assumption in statistics is that the researcher decides exactly what test that should be made and how the models should be formulated before they collect their data. This assumption is important since conventional methods of calculate p-values requires that the choice of method is not determined from the information in the specific sample. If the data in your sample suggest a reformulation of the model you should therefore always collect a second sample to fit the model suggested by the data in the first sample.

In the real world however, a lot of researchers finds it reasonable to base their choice of methodology on the data of their samples. It is very common to fit a model in the same sample that was used to formulate it. And you may argue that it sounds reasonable to let the information you have guide your analysis. As statisticians we have a difficult choice between trying to teach all these researchers that their way of thinking is wrong or trying to develop statistical methods that would be possible to use even if you formulate your model and/or decides what test you make, dependent on the data of a specific sample.

In this paper we use inference on rank orders as an example of how p-values could be calculated conditional on descriptive statistics from the actual sample in order to make it possible to make your choice of method dependent on your sample. Such methods would make it possible to draw more conclusions from the first sample and reduces the need for collecting new data.

We assume a survey question where the respondents are asked to rank k objects by assigning number 1 to the best object, number 2 to the second best down to k to the worst object. The answers to this kind of question give us the rank order between the k objects for this specific individual. From our sample we may calculate the rank order of the sample by comparing the mean answers of each object. But what conclusion could be drawn about the rank order between the k objects in the whole population?

The standard test for analysing this kind of question is Friedmans two way ANOVA analyses, also called the Friedmans test of randomized block designs (Friedman 1937). The null hypothesis of the Friedman test is that the k variables have the same distribution against the alternate hypothesis that at least one of them have a different distribution. In the context of this application one may say that the null hypothesis is that all the k objects are equally popular in the population and the alternate hypothesis is that at least one of them differ.

The drawback of Friedmans test is that the possible conclusion you may draw is not very strong. We would probably like to say something more about the rank order of the population than just whether all objects are equally popular or not. The next step would thus probably be to make some pairwise sign tests or signed-rank tests in order to find out which objects that differ in popularity. In this paper we limit the analysis to pairwise sign tests.

Table 1: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 20, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	2	3	4	5	2	3	4	5
1	0.0	0.1	0.3	1.9	0.6	2.4	6.9	20.5
2		0.0	0.0	0.3		0.1	1.0	6.9
3			0.0	0.1			0.1	2.3
4				0.0				0.6

Table 2: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 60, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	2	3	4	5	2	3	4	5
1	0.0	0.2	0.8	4.3	0.3	1.3	4.4	14.9
2		0.0	0.0	0.8		0.1	0.5	4.3
3			0.0	0.2			0.1	1.3
4				0.0				0.3

At this stage, however, you got a difficult problem. Either you choose to test all possible combination of objects. That strategy may, if there are a large number of objects, give you some significant results even if the null hypotheses is true just due to the fact that you make a lot of tests. You would need to deal with the problem of a high family wise error rate (FWER) and thus make some adjustment of the p-values that would reduce the power of the tests.

The other possibility is to choose a limited number of combinations to test. The problem at this stage is that the choice of objects to test should be made before you look at your data. Assume for example that you decide to test the object with the highest mean rank in your sample against the object with the lowest mean rank. In this case you let the rank order of your sample determine what test you make and the conventional p-values is not credible. In table 1 - 9 this problem is illustrated throw Monte Carlo analyses where 200 000 samples are drawn from a population where all the objects are equally popular.[1] The frequency of significant tests are reported conditional on the rank order between the objects in each specific sample[2].

[1] More specifically the answers of each person in each sample is generated by random numbers drawn from k identical normal distributions with a mean of 1 and standard deviation of 1. We than assign the answers 1 to k for the highest to the lowest random number.

[2] Of course some of that samples may not have definable sample rank orders since 2 or more objects may have the same means answers. In those cases the rank order between the ties where assigned randomly.

Table 3: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	2	3	4	5	2	3	4	5
1	0.0	0.3	1.2	5.7	0.6	2.3	6.6	20.2
2		0.0	0.1	1.2		0.1	1.0	6.7
3			0.0	0.3			0.1	2.2
4				0.0				0.6

Table 4: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 20, 6 objects to compare.

	frequency significant at 1%					frequency significant at 5%				
	2	3	4	5	6	2	3	4	5	6
1	0.0	0.0	0.1	0.5	2.5	0.5	1.8	4.4	10.1	24.9
2		0.0	0.0	0.0	0.5		0.1	0.6	2.2	10.1
3			0.0	0.0	0.1			0.1	0.5	4.4
4				0.0	0.0				0.1	1.8
5					0.0					0.5

Table 5: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 60, 6 objects to compare.

	frequency significant at 1%					frequency significant at 5%				
	2	3	4	5	6	2	3	4	5	6
1	0.0	0.1	0.4	1.3	5.5	0.2	0.9	2.6	6.5	18.5
2		0.0	0.0	0.1	1.3		0.0	0.3	1.2	6.5
3			0.0	0.0	0.4			0.0	0.3	2.6
4				0.0	0.1				0.0	0.9
5					0.0					0.3

Table 6: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 6 objects to compare.

	frequency significant at 1%					frequency significant at 5%				
	2	3	4	5	6	2	3	4	5	6
1	0.0	0.0	0.1	0.5	2.5	0.5	1.8	4.4	10.1	24.9
2		0.0	0.0	0.0	0.5		0.1	0.6	2.2	10.1
3			0.0	0.0	0.1			0.1	0.5	4.4
4				0.0	0.0				0.1	1.8
5					0.0					0.5

Table 7: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 20, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	2	3	4	5	6	7	2	3	4	5	6	7
1	0.0	0.0	0.1	0.2	0.7	3.2	0.5	1.5	3.2	6.7	13.5	29.0
2		0.0	0.0	0.0	0.1	0.7		0.1	0.4	1.3	3.9	13.3
3			0.0	0.0	0.0	0.2			0.0	0.3	1.3	6.7
4				0.0	0.0	0.1				0.0	0.4	3.2
5					0.0	0.0					0.1	1.5
6						0.0						0.5

Table 8: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 60, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	2	3	4	5	6	7	2	3	4	5	6	7
1	0.0	0.1	0.2	0.6	1.9	6.9	0.2	0.8	1.9	4.1	9.0	21.9
2		0.0	0.0	0.1	0.2	1.9		0.0	0.2	0.6	2.1	8.9
3			0.0	0.0	0.0	0.7			0.0	0.1	1.3	4.1
4				0.0	0.0	0.2				0.0	0.6	1.8
5					0.0	0.1					0.2	0.8
6						0.0						0.2

Table 9: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	2	3	4	5	6	7	2	3	4	5	6	7
1	0.0	0.1	0.4	1.0	2.7	9.2	0.4	1.3	3.0	6.3	12.9	28.6
2		0.0	0.0	0.1	0.4	2.7		0.1	0.4	1.2	3.6	12.9
3			0.0	0.0	0.1	1.0			0.1	0.2	1.2	6.4
4				0.0	0.0	0.4				0.1	0.4	3.0
5					0.0	0.1					0.1	1.3
6						0.0						0.4

From the tables we can see that comparing objects at the top of the rank order of the sample with objects in the bottom gives a frequency of significant tests above the significance level. If you already have collected your data and computed the descriptive statistics when you start to think about what analysis to make it is no longer possible to compare these objects using the conventional way of calculating p-values. This illustrates the importance of the assumption in statistics that everyone decides exactly what test that should be made before they look at their data.

So what could be made after the descriptive statistics are calculated? We may of course make a new survey, asking the same question to another sample of respondent's and test those objects that were far apart from each other in the first sample using the data of the second sample. But making another survey is costly so it would be good if we were able to get some more information out of the first sample.

We have already mention the option to compute all the possible sign tests and make a Holm Bonferoni (HB) adjustment of the p-values. In that case we do not do any choices after the descriptive statistics are calculated. Another option is to pick a fewer amount of combinations, to make the HB-adjustment smaller, but avoid those combinations that have a frequency of significant tests that is higher than the significance level. The researcher should run a Monte Carlo analysis on the number of objects and the sample size that he uses and compute a table like table 1 and then check what sign tests that are possible to make. The drawback of the first option is that the tests will have a very low power, the drawback of the second that some tests are not possible to make.

In some cases the low power of the Holm-Bonferoni adjusted p-values is not a very big problem. If you have a population with very high differences in preferences between the objects, or a very large sample size, you do not need a test with a high power. But sometimes you have the case that the objects you compare are pretty similar in popularity and you would like to investigate if there is at least a difference between the most popular and the least popular object. We cannot use the conventional p-values since they require that the choice of objects to compare is not based upon information from the sample. We need p-values that are calculated conditional on the rank order of the sample. The conditional p-values will tells us if the difference between the objects are big enough to reject the null hypothesis ***even if we take into account that we compare the best object with the worst***.

In the next chapter we define a couple of concepts that will be used in the paper and reviews the relevant literature. In chapter 3 we describe how the conditional p-values are determined for 5 objects and the sample sizes of 20, 60 and 200.[3] Chapter 4 evaluates the conditional p-values together with the HB adjusted p-values and chapter 5 concludes the findings of the paper.

2 Definitions and literature review.

In this paper we will define the individual rank order as the answer to the survey question from one specific respondent. (This should not be confused with the ranks used in non-parametric methods where all the answers in the sample are ranked.) The rank order of the sample will in this paper be defined as the rank order of the mean answer on each object in the actual sample. The rank order of the population is the difference in popularity between the objects in the total population. We will generate individual answers by drawing random numbers from normal distributions with the same standard deviation but with different means for each object. The rank order of the population is than defined as the rank order of the means in the normal distributions.

The family wise error rate is a crucial concept when several hypothesis are tested at the same time. The family wise error rate is defined as the probability to get at least one significant result if all the null hypotheses are true. The problem with family wise error rate have been discussed by for example Bonferroni (1936), Dunn (1961) and Holm (1979). Bonferoni suggested that the p-values should be multiplied with the number of hypotheses tests made. The Bonferoni adjustment has been criticized of being too conservative since it gives a very low power of the tests.

To increase the power a bit but still leave the family wise error rate below the significance level, in the case where all null hypothesis are true, Holm derived a variant of the method. Holm suggested that the lowest p-value should be multiplied by the number of tests while the second lowest should be multiplied by the number of tests minus 1, the third lowest with the number of tests minus 2 and so on. The Holm method could be described as a sequential

[3] Tables with p-values for other combinations of number of objects and sample sizes could be found on `http://www.natskolan.se/research/cond_p/signtest/sign.htm`

process where you take away the p-values one at a time, starting from the lowest p-value, and makes a Bonferoni adjustment of the rest of them. In the following we would refer to p-values calculated by the Holm version of the Bonferoni method as HB-adjusted p-values.

According to Chatfield (1995) it is quite common that researchers' runs a model, modify it based on the results and rerun the new model on the same sample until they find the model with the best fit.

"This iterative process can continue indefinitely, but still using the same data. Now diagnostic tests typically assume that the model is specified a priori and calculate a P-value as Probability(more extreme result than the one obtained l model is true). But, if the model is formulated, fitted and checked using the same data, then we should really calculate Probability(more extreme result than the one obtained l model has been selected as 'best' by the model formulation procedure). It is not clear in general how this can be calculated. What is clear is that the good fit of a best fitting model should not be surprising!" Chatfield (1995) page 427.

In complicated regression models the conditional p-values suggested by Chatfield would be quite difficult to calculate. In the much easier problem in this paper it is much easier to calculate p-values conditional on the way we choose what tests to make.

3 Defining p-values conditional on the rank order of the sample

In this section we derive p-values conditional on the rank order of the sample. We define the rank order of the sample as the rank order of the mean answer on each object. For each pair of objects we define the test statistic as the number of respondents that prefer that of the two object that had the first position in the rank order of the sample. Having asked the respondents to assign 1 to the best object the interpretation of this is, the number of respondents preferring that of the two objects that have the lowest mean answer in this sample.

One tail p-values

The relevant one-tail conditional p-value for this test statistic would be it's cumulative distribution. The cumulative distribution is reported in table 10 for a sample size of 20 and in table 11 for a sample size of 60. The distribution is simulated by Monte Carlo simulations based upon 2 million random samples where all objects are equally popular in the population, meaning that all different individual rank orders have the same probability of being drawn. As a comparison the last column of table 10 and table 11 reports the conventional p-value of the sign test.

It's a well known fact that the conventional one tail test would be inconsistent if the sign of the alternative hypothesis is determined from the outcome of the specific sample. If so the rejection frequency under a true null hypothesis would be twice as high as the level of significance. Using the conditional p-values below the one tail test will be consistent even if the sign of the alternative hypothesis is determined from the outcome of the specific sample.

Two tail p-values

In this paper we define the two tail p-values as the probability of getting this amount of respondents or more or an amount as far away from half of the sample on the other side of the distribution. The starting point for the two tailed p-values is a cumulative distribution cumulated from both tails. This cumulative distribution for the sample size of 20 are shown in table 12.

In table 12 the numbers in the above half of the table is the probability of getting a sample where this or a lower amount of respondents prefer the first object in the rank order of the sample. The number in the lower half of the table gives the probability of getting a sample where this or a higher amount of respondents prefer the first object in the rank order of the sample.

Table 10: P-values for x = number of respondents preferring the object on position v, n=20, 5 objects, one tail test.

	Position in the rank order of the sample										
v	1	1	1	1	2	2	2	3	3	4	
w	2	3	4	5	3	4	5	4	5	5	
x											
0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.979
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.942
8	0.999	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	0.999	0.868
9	0.983	0.995	0.999	1.000	0.979	0.994	0.999	0.979	0.995	0.983	0.748
10	0.885	0.955	0.985	0.996	0.859	0.944	0.985	0.859	0.955	0.885	0.588
11	0.628	0.804	0.911	0.972	0.562	0.763	0.911	0.562	0.804	0.628	0.412
12	0.316	0.530	0.719	0.880	0.236	0.459	0.720	0.237	0.530	0.316	0.252
13	0.112	0.257	0.445	0.685	0.062	0.191	0.445	0.063	0.257	0.112	0.132
14	0.030	0.091	0.205	0.427	0.011	0.054	0.206	0.011	0.091	0.030	0.058
15	0.006	0.023	0.069	0.205	0.001	0.010	0.069	0.001	0.023	0.006	0.021
16	0.001	0.004	0.017	0.073	0.000	0.001	0.017	0.000	0.004	0.001	0.006
17	0.000	0.001	0.003	0.019	0.000	0.000	0.003	0.000	0.001	0.000	0.001
18	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 11: P-values for x = number of respondents preferring the object on position v, n=60, 5 objects, one tail test.

	Position in the rank order of the sample										
v	1	1	1	1	2	2	2	3	3	4	
w	2	3	4	5	3	4	5	4	5	5	
x											
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.974
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.954
25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.922
26	0.999	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	0.999	0.877
27	0.995	0.999	1.000	1.000	0.994	0.998	1.000	0.994	0.999	0.995	0.817
28	0.979	0.993	0.998	1.000	0.974	0.991	0.998	0.973	0.993	0.979	0.741
29	0.934	0.976	0.992	0.998	0.918	0.969	0.992	0.918	0.976	0.934	0.651
30	0.839	0.931	0.974	0.993	0.804	0.914	0.974	0.804	0.931	0.839	0.551
31	0.690	0.843	0.931	0.979	0.631	0.809	0.931	0.632	0.844	0.689	0.449
32	0.507	0.709	0.850	0.946	0.431	0.655	0.851	0.431	0.709	0.506	0.349
33	0.330	0.544	0.728	0.884	0.253	0.474	0.729	0.253	0.544	0.329	0.259
34	0.192	0.377	0.576	0.787	0.127	0.304	0.576	0.127	0.376	0.191	0.183
35	0.100	0.235	0.417	0.659	0.055	0.172	0.417	0.055	0.235	0.100	0.123
36	0.048	0.133	0.274	0.514	0.021	0.086	0.274	0.021	0.132	0.047	0.078
37	0.021	0.068	0.164	0.370	0.007	0.038	0.163	0.007	0.067	0.021	0.046
38	0.008	0.031	0.088	0.245	0.002	0.015	0.088	0.002	0.031	0.008	0.026
39	0.003	0.013	0.043	0.149	0.001	0.005	0.043	0.001	0.013	0.003	0.014
40	0.001	0.005	0.019	0.083	0.000	0.002	0.019	0.000	0.005	0.001	0.007
41	0.000	0.002	0.008	0.042	0.000	0.000	0.008	0.000	0.002	0.000	0.003
42	0.000	0.001	0.003	0.020	0.000	0.000	0.003	0.000	0.001	0.000	0.001
43	0.000	0.000	0.001	0.008	0.000	0.000	0.001	0.000	0.000	0.000	0.001
44	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
45	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
46	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
47	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
48	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

To save space some rows with only ones and zeros are omitted.

Table 12: Cumulative distribution of x, cumulated from both tails. x = number of respondents preferring the object on position v, n=20, 5 objects, one tail test.

	Position in the rank order of the sample									
v	1	1	1	1	2	2	2	3	3	4
w	2	3	4	5	3	4	5	4	5	5
x										
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.001
8	0.017	0.005	0.001	0.000	0.021	0.006	0.001	0.021	0.005	0.017
9	0.115	0.045	0.015	0.004	0.141	0.057	0.015	0.141	0.045	0.116
10										
11	0.627	0.804	0.911	0.972	0.562	0.763	0.910	0.562	0.805	0.628
12	0.315	0.530	0.719	0.880	0.236	0.459	0.719	0.236	0.530	0.315
13	0.112	0.257	0.445	0.684	0.062	0.190	0.444	0.062	0.257	0.112
14	0.030	0.091	0.205	0.427	0.011	0.054	0.205	0.011	0.091	0.030
15	0.006	0.023	0.069	0.204	0.001	0.010	0.069	0.001	0.023	0.006
16	0.001	0.004	0.017	0.073	0.000	0.001	0.017	0.000	0.004	0.001
17	0.000	0.001	0.003	0.019	0.000	0.000	0.003	0.000	0.001	0.000
18	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000
19	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 13: P-values for x = number of respondents preferring object on position v, n=20, 5 objects two-tail test.

	Position in the rank order of the sample										
v	1	1	1	1	2	2	2	3	3	4	
w	2	3	4	5	3	4	5	4	5	5	
x											
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.001	0.003	0.019	0.000	0.000	0.003	0.000	0.001	0.000	0.003
4	0.001	0.004	0.017	0.073	0.000	0.001	0.017	0.000	0.004	0.001	0.012
5	0.006	0.023	0.069	0.205	0.001	0.010	0.069	0.001	0.023	0.006	0.041
6	0.030	0.091	0.205	0.427	0.011	0.054	0.206	0.011	0.091	0.030	0.115
7	0.113	0.257	0.445	0.685	0.063	0.191	0.445	0.064	0.257	0.113	0.263
8	0.332	0.535	0.720	0.880	0.257	0.465	0.721	0.257	0.535	0.332	0.503
9	0.743	0.849	0.926	0.975	0.703	0.820	0.926	0.703	0.849	0.744	0.824
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	0.743	0.849	0.926	0.975	0.703	0.820	0.926	0.703	0.849	0.744	0.824
12	0.332	0.535	0.720	0.880	0.257	0.465	0.721	0.257	0.535	0.332	0.503
13	0.113	0.257	0.445	0.685	0.063	0.191	0.445	0.064	0.257	0.113	0.263
14	0.030	0.091	0.205	0.427	0.011	0.054	0.206	0.011	0.091	0.030	0.115
15	0.006	0.023	0.069	0.205	0.001	0.010	0.069	0.001	0.023	0.006	0.041
16	0.001	0.004	0.017	0.073	0.000	0.001	0.017	0.000	0.004	0.001	0.012
17	0.000	0.001	0.003	0.019	0.000	0.000	0.003	0.000	0.001	0.000	0.003
18	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 13 reports two tailed conditional p-values for 5 objects and a sample size of 20. These are calculated from the probabilities in table 12 by adding up the probabilities at the same distance from 10. The interpretation of the number on the 6'th row is thus that it is the probability of getting a sample where 6 or fewer or 14 or more respondents prefer the object on position v before the object on position w. Table 14 reports two tailed conditional p-values for 5 objects and a sample size of 60.

The proposed method for the sign test with p-values conditional on the rank order of the sample is thus as follows:

Ask the respondents to rank the objects from the object they consider to have the highest quality to the object they consider to have the lowest. Construct the rank order of the sample by rank all the objects in the survey on the basis of the mean of the respondents' answers. For the combinations selected to be tested, count the number of respondents preferring the object on the highest position in the rank order of the sample. Find the p-value in the relevant column of the relevant table.

If there are ties in the rank order of the sample we suggest that the conventional p-value should be used if these two objects are to be tested against each other. A more difficult

Table 14: P-values for x = number of respondents preferring object on position v, n=60, 5 objects, two-tail test.

	Position in the rank order of the sample										
v	1	1	1	1	2	2	2	3	3	4	
w	2	3	4	5	3	4	5	4	5	5	
x											
13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	0.000	0.000	0.001	0.008	0.000	0.000	0.001	0.000	0.000	0.000	0.001
18	0.000	0.001	0.003	0.020	0.000	0.000	0.003	0.000	0.001	0.000	0.003
19	0.000	0.002	0.008	0.042	0.000	0.000	0.008	0.000	0.002	0.000	0.006
20	0.001	0.005	0.019	0.083	0.000	0.002	0.019	0.000	0.005	0.001	0.013
21	0.003	0.013	0.043	0.149	0.001	0.005	0.043	0.001	0.013	0.003	0.027
22	0.008	0.031	0.088	0.245	0.002	0.015	0.088	0.002	0.031	0.008	0.052
23	0.021	0.068	0.164	0.370	0.007	0.038	0.163	0.007	0.067	0.021	0.092
24	0.048	0.133	0.274	0.514	0.021	0.086	0.274	0.021	0.132	0.048	0.155
25	0.101	0.236	0.417	0.659	0.057	0.173	0.417	0.057	0.235	0.101	0.245
26	0.197	0.378	0.577	0.787	0.134	0.306	0.577	0.133	0.378	0.196	0.366
27	0.351	0.551	0.730	0.884	0.280	0.483	0.731	0.280	0.550	0.351	0.519
28	0.573	0.733	0.858	0.948	0.514	0.685	0.858	0.513	0.733	0.572	0.699
29	0.851	0.912	0.957	0.986	0.827	0.895	0.957	0.827	0.913	0.850	0.897
30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
31	0.851	0.912	0.957	0.986	0.827	0.895	0.957	0.827	0.913	0.850	0.897
32	0.573	0.733	0.858	0.948	0.514	0.685	0.858	0.513	0.733	0.572	0.699
33	0.351	0.551	0.730	0.884	0.280	0.483	0.731	0.280	0.550	0.351	0.519
34	0.197	0.378	0.577	0.787	0.134	0.306	0.577	0.133	0.378	0.196	0.366
35	0.101	0.236	0.417	0.659	0.057	0.173	0.417	0.057	0.235	0.101	0.245
36	0.048	0.133	0.274	0.514	0.021	0.086	0.274	0.021	0.132	0.048	0.155
37	0.021	0.068	0.164	0.370	0.007	0.038	0.163	0.007	0.067	0.021	0.092
38	0.008	0.031	0.088	0.245	0.002	0.015	0.088	0.002	0.031	0.008	0.052
39	0.003	0.013	0.043	0.149	0.001	0.005	0.043	0.001	0.013	0.003	0.027
40	0.001	0.005	0.019	0.083	0.000	0.002	0.019	0.000	0.005	0.001	0.013
41	0.000	0.002	0.008	0.042	0.000	0.000	0.008	0.000	0.002	0.000	0.006
42	0.000	0.001	0.003	0.020	0.000	0.000	0.003	0.000	0.001	0.000	0.003
43	0.000	0.000	0.001	0.008	0.000	0.000	0.001	0.000	0.000	0.000	0.001
44	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
45	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
46	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
47	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

To save space some rows with only ones and zeros are omitted.

question is what to do if one of the tie objects should be tested against another object. As an example consider the case when the first and second object in the rank order of the sample has the same sample mean answer. If one of those objects should be compared with the third object one may either choose the p-value for the positions 1 and 3 or the p-value from the positions 2 and 3. One possible solution may be to use the mean of the two different p-values.

4 Evaluation of different ways of calculating p-values

What is the relationship between the conditional and the conventional p-values? As an example consider a two tail test from a sample of 20 respondents. Suppose also that 14 respondents prefer the object in the first position in the rank order of the sample. The conventional two tailed p-value becomes 0.115. The conditional p-values are lower than the conventional for all cases except when the two objects are located on the first and fourth, first and fifth or the second and fifth places in the rank order of the sample. If the objects are closer to each other the conditional p-values would be higher than the conventional. With conditional p-values we have a lower probability of rejecting the Null hypothesis compared to the conventional sign test if the objects are far apart from each other but a higher probability of rejecting the Null hypothesis if they are close in the rank order of the sample.

This feature is illustrated in table 15 where rejection frequencies are reported for three strategies (Conventional p-values, HB adjusted p-values and conditional p-values) under the assumption that the null hypothesis is true, that is that all objects have the same popularity among the total population. The table is based on Monte Carlo simulation on 200 000 random samples.

From table 15 it can be seen that it is only the conditional p-values that have a rejection rate close to the significance level if the researcher makes his choice of test dependent on the rank order of the sample and only one test is made. The rejection rates are a bit lower than the significance level due to the fact that the test statistic has a discrete distribution.

The very low rejection frequencies for the HB adjusted p-values gives an illustration of the difficulty in testing this many hypothesis at the same time. The conditional p-values are a good alternative if only a few tests are to be made. If you make more than one test also the conditional p-values need to be HB adjusted and then of course also those p-values will be much lower.

one-tail versus two-tail test

If you want to use a one-tail test using the conventional method of calculating p-values, you must decide the direction of the inequality sign before you look at your data. A strategy

Table 15: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	2	3	4	5	6	7	2	3	4	5	6	7
1	0.0	0.1	0.4	1.0	2.7	9.2	0.4	1.3	3.1	6.4	13.1	28.7
	0.0	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.1	0.2	0.6	3.0
	0.7	0.8	0.7	0.6	0.6	0.8	3.2	3.4	4.8	4.2	4.2	4.4
2		0.0	0.0	0.1	0.4	2.7		0.1	0.4	1.2	3.7	13.0
		0.0	0.0	0.0	0.0	0.1		0.0	0.0	0.0	0.1	0.6
		0.7	0.7	0.6	0.8	0.6		4.3	3.8	3.6	3.7	4.2
3			0.0	0.0	0.1	1.0			0.0	0.3	1.2	6.4
			0.0	0.0	0.0	0.0			0.0	0.0	0.0	0.2
			1.0	1.0	0.7	1.0			3.5	3.2	3.5	4.2
4				0.0	0.0	0.4				0.0	0.4	3.1
				0.0	0.0	0.0				0.0	0.0	0.1
				0.9	0.7	0.7				3.4	3.8	4.8
5					0.0	0.1					0.1	1.3
					0.0	0.0					0.0	0.0
					0.7	0.8					4.3	3.5
6						0.0						0.4
						0.0						0.0
						0.7						3.3

The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

where the sign of the inequality sign is dependent on descriptive statistics in the sample would generally give a rejection rate twice as high as the significance level since one-tail p-values are twice as high as the two-tail p-values. This is a quite strong argument for not using one-tail tests in practice. There may be a too strong temptation to look in your data and decide the direction of the inequality sign dependent on what object that is the most popular in your sample. Once again it is very difficult for referees to control that the researcher in fact decided the direction of the inequality sign before the data was collected.

If we compare the conditional p-values in table 10 and table 11 with the corresponding two-tail p-values in table 13 and table 14, we see that all relevant p-values, that is all p-values close to the most common significance levels, are the same regardless of whether we use one tail or two tail test. It really doesn't matter what you choose. The conditional p-values will be robust also against this kind of peeping on the data. Mathematically the reason to this is that we almost always will reject the null hypothesis in the same side even in the two tail test and thus a very small probability on the other side of the distribution is added to the one-tail p-values.

Evaluation of the power under a false null hypothesis

To evaluate the power of the different ways of calculating p-values we generate random samples were we denote the different objects to be compared with letters. The rank order of the population is defined by using the alphabetic order with A for the most popular object. The individual rank orders are constructed by drawing random numbers from normal distributions with the standard deviations equal to 1 but where the different objects have different means. The random numbers for each individual are than ranked in order to achieve the rank order of the individual respondent. The differences in the means for the random numbers for the different objects will thus determine the strength in the preferences of the population.

Although table 15 and the tables in the rest of this paper looks very similar they differ in one important aspect. The row and columns in table 15 where defined by the rank order in each specific sample. In table 15 a specific object may end up in different rows in different samples since the rank order of the sample will differ between the different samples. In table 16 on the other hand, the rows and columns are defined from the specific object, meaning that a specific object will always be in the same row regardless of the rank order of each sample. And since the rows and columns are for specific objects they will reflect the rank order of the population.

When conditional p-values are used the comparison between object A and B will be determined from different distributions of the p-values in different samples since they will end up on different positions in the rank order of the sample in different samples. As an

Table 16: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 20, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	0,8	3,5	11	25,8	8,2	21,3	42,4	65,6
	0,2	0,9	3,7	10,9	0,8	3,5	11	25,8
	4,4	7,2	10,7	16,1	12,4	18,8	26,2	36,0
B		0,8	3,5	11,1		8,3	21,2	42,4
		0,2	0,9	3,7		0,8	3,5	11,1
		2,7	5,5	11,1		9,1	16,0	27,0
C			0,8	3,5			8,3	21,3
			0,2	0,9			0,8	3,5
			2,8	7,8			9,4	19,8
D				0,8				8,3
				0,2				0,8
				4,9				13,2

Means used in the generation of the random samples; 1,5-1,25-1-0,75-0,5. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 17: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 60, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	4,9	27,9	68,0	93,4	13,0	48,3	84,5	98,1
	1,4	12,5	44,9	78,9	4,7	26,9	65,1	90,8
	17,0	40,8	70,9	87,7	34,7	62,8	87,6	96,4
B		4,9	27,9	68,2		13,2	48,2	84,7
		1,4	12,4	45,1		4,7	27,0	65,2
		21,7	47,3	71,1		33,8	68,4	87,7
C			4,9	27,9			13,1	48,2
			1,3	12,5			4,7	26,8
			21,5	40,8			33,7	62,9
D				4,9				13,1
				1,4				4,7
				17,4				35,1

Means used in the generation of the random samples; 1,5-1,25-1-0,75-0,5. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 18: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 200, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	26,3	91,0	99,9	100,0	47,4	97,2	100,0	100,0
	13,0	78,7	99,7	100,0	30,4	91,4	99,9	100,0
	53,0	95,3	100,0	100,0	69,3	98,3	100,0	100,0
B		26,2	91,0	99,9		47,4	97,2	100,0
		13,1	78,7	99,7		30,2	91,7	99,9
		63,9	97,5	99,9		78,5	99,2	100,0
C			26,1	91,0			47,4	97,3
			13,1	78,6			30,1	91,5
			63,9	95,3			78,5	98,3
D				26,2				47,5
				13,0				30,3
				53,1				69,4

Means used in the generation of the random samples; 1,5-1,25-1-0,75-0,5. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

example, with a sample size of 60 and this strength in the preferences of the population, object A, the most popular object, ended up in the first position in 87,7 percent of the samples and in the second position in 10,7

Comparing the conventional p-value with the conditional p-value, the conventional has the strongest power when we compare objects far apart in the rank order of the population but the conditional has the strongest power comparing objects close to each other. But we should keep in mind that the conventional p-value is not viable for objects far apart in the rank order of the sample if the choice of objects is made conditional on the rank order of the sample. The conventional p-values in the test between object A and object E was viable only in a few samples where they ended up close to each other in the rank order of the sample.

At a first glance one may thought that the conditioning of the p-values totally would remove the effect that the power is higher when comparing objects far apart in the rank order of the population. But also the conditional p-values give a stronger power when comparing objects far apart in the rank order of the population. This is not strange if we keep in mind that we do not make the p-values conditional on the rank order of the population but on the rank order of the specific sample.

Comparing the HB adjusted conventional p-values with the conditional p-value the power is always much lower for the HB adjusted. But we should keep in mind that this comparison is only relevant if we choose between testing all possible combination with HB adjustment

Table 19: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 20, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	0,3	0,3	0,4	0,6	4,4	4,8	5,5	6,5
	0,0	0,1	0,1	0,1	0,3	0,3	0,4	0,6
	0,4	0,4	0,5	0,6	2,2	2,3	2,4	2,8
B		0,3	0,4	0,5		4,4	4,8	5,6
		0,0	0,1	0,1		0,3	0,4	0,5
		0,4	0,4	0,5		2,1	2,2	2,5
C			0,3	0,3			4,2	4,8
			0,0	0,0			0,3	0,3
			0,4	0,5			2,1	2,4
D				0,3				4,3
				0,0				0,3
				0,5				2,5

Means used in the generation of the random samples; 1,1-1,05-1-0,95-0,9. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

or just running one of them with the conditional p-value. If more than one test is made we will also need to HB adjust the conditional p-value.

With this strong preferences the power are very high for all different p-values in the sample size of 200. Table 19 - Table 21 report corresponding result from a population with weaker preferences. With weak preferences it is only the sample size of 200 that gives some reasonable power. In this case it is only the three cells in the upper left corner of the table where the conditional calculations of p-values give a stronger power. But in most of the samples the conventional p-values was not valid in these cells since they most often end up far apart in the rank order of the sample.

Table 20: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 60, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	0,7 0,1 1,2	1,2 0,1 1,4	1,9 0,2 1,7	3,1 0,4 2,2	3,1 0,3 5,4	4,2 0,5 5,8	6,2 1,0 6,8	9,1 1,6 8,5
B		0,7 0,1 1,1	1,1 0,1 1,3	1,9 0,2 1,7		3,2 0,3 4,8	4,2 0,5 5,4	6,2 1,0 7,0
C			0,7 0,1 1,1	1,1 0,1 1,4			3,1 0,3 4,8	4,2 0,6 6,0
D				0,7 0,1 1,2				3,1 0,3 5,7

Means used in the generation of the random samples; 1,1-1,05-1-0,95-0,9. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 21: Percentage frequency of significant tests, conditional on the object's positions in the rank order of the population, n = 200, 5 objects to compare.

	frequency significant at 1%				frequency significant at 5%			
	B	C	D	E	B	C	D	E
A	1,4 0,2 3,0	3,4 0,6 4,3	7,7 1,6 6,0	15,0 4,0 8,3	5,6 0,7 9,4	10,7 1,9 12,9	19,5 4,5 17,5	32,1 9,5 22,7
B		1,4 0,2 2,8	3,4 0,6 4,0	7,6 1,6 6,0		5,6 0,7 9,7	10,7 1,8 12,9	19,5 4,5 17,5
C			1,4 0,2 2,7	3,4 0,6 4,3			5,6 0,7 9,6	10,6 1,8 13,0
D				1,4 0,2 3,1				5,6 0,7 9,7

Means used in the generation of the random samples; 1,1-1,05-1-0,95-0,9. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 22: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 20, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
A	0.6	2.1	5.9	13.5	25.9	42.0	6.8	15.0	29.1	47.4	65.8	81.1
	0.1	0.5	1.7	4.7	11.0	21.2	0.4	1.3	3.5	7.9	15.1	25.1
	3.9	7.5	12.6	17.9	26.2	32.8	10.9	17.2	28.1	37.8	49.7	58.4
B		0.6	2.0	5.8	13.5	25.9		6.8	14.8	29.1	47.2	65.8
		0.1	0.5	1.7	4.7	10.8		0.3	1.1	3.3	7.8	15.1
		4.9	7.9	12.3	19.2	26.0		10.4	19.4	28.8	39.8	49.5
C			0.6	2.0	5.8	13.4			6.8	14.9	28.9	47.0
			0.1	0.5	1.7	4.7			0.3	1.1	3.3	7.8
			4.9	7.7	12.3	17.7			14.8	20.9	28.8	37.4
D				0.6	2.0	5.7				6.7	14.9	29.0
				0.1	0.5	1.7				0.3	1.1	3.5
				4.8	7.9	12.3				14.6	19.5	27.6
E					0.7	2.0					6.8	14.9
					0.1	0.5					0.4	1.2
					4.9	7.1					10.4	16.8
F						0.6						6.6
						0.1						0.4
						3.7						10.5

Means used in the generation of the random samples; 1.6, 1.4, 1.2, 1.0, 0.8, 0.6, 0.4. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Evaluating power in the case of 7 objects to compare

In table 22 it can be seen that the main patterns are the same when we have 7 objects as in the case of 5 objects to compare.

Table 23: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 60, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
	3.1	15.7	43.5	75.3	93.5	99.1	9.1	31.9	65.0	89.2	98.1	99.8
A	0.4	3.3	16.1	43.5	74.3	92.9	1.6	9.9	32.9	65.4	88.6	97.7
	12.7	33.2	59.0	80.6	92.8	96.9	28.3	55.7	79.6	93.1	98.1	99.4
		3.1	15.6	43.8	75.4	93.6		9.2	31.9	65.0	89.2	98.1
B		0.4	3.3	16.1	43.7	74.4		1.7	10.1	33.1	65.6	88.7
		16.0	38.7	65.1	84.6	92.8		34.8	62.1	83.7	93.5	98.1
			3.1	15.8	43.7	75.2			9.1	31.9	65.0	89.2
C			0.4	3.3	16.0	43.1			1.6	10.1	33.1	65.2
			17.2	40.1	65.2	80.6			36.5	63.3	83.7	93.1
				3.0	15.6	43.6				9.0	31.9	65.0
D				0.3	3.3	16.0				1.6	9.9	33.0
				17.0	38.7	58.7				36.4	62.3	79.7
					3.1	15.6					9.1	31.8
E					0.4	3.4					1.7	9.9
					15.9	33.0					34.7	55.9
						3.1						9.2
F						0.4						1.6
						12.6						28.2

Means used in the generation of the random samples; 1.6, 1.4, 1.2, 1.0, 0.8, 0.6, 0.4. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 24: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
A	15.1	70.9	98.3	100.0	100.0	100.0	32.3	86.9	99.7	100.0	100.0	100.
	4.1	43.2	91.7	99.8	100.0	100.0	11.9	64.5	97.2	100.0	100.0	100.
	37.3	84.8	99.2	100.0	100.0	100.0	54.3	93.0	99.8	100.0	100.0	100.
B		14.9	70.8	98.3	100.0	100.0		32.3	87.0	99.6	100.0	100.
		4.1	43.1	91.6	99.8	100.0		11.8	64.5	97.2	100.0	100.
		47.7	90.7	99.6	100.0	100.0		64.8	96.1	99.9	100.0	100.
C			15.0	71.0	98.4	100.0			32.3	87.0	99.6	100.
			4.1	43.5	91.6	99.8			11.8	64.8	97.2	100.
			53.7	92.6	99.6	100.0			65.2	96.0	99.9	100.
D				15.2	71.0	98.3				32.2	87.0	99.7
				4.1	43.2	91.6				11.9	64.7	97.2
				53.9	90.6	99.2				65.2	96.0	99.9
E					15.1	70.9					32.3	87.1
					4.2	43.1					12.0	64.5
					47.7	85.0					64.9	93.2
F						15.2						32.4
						4.1						12.0
						37.3						54.1

Means used in the generation of the random samples; 1.6, 1.4, 1.2, 1.0, 0.8, 0.6, 0.4. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 25: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 20, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
A	0.3 0.1 1.4	0.6 0.1 1.7	1.2 0.3 2.1	2.1 0.5 2.7	3.5 0.9 3.4	5.7 1.7 4.3	4.8 0.1 5.2	6.7 0.1 5.9	10.2 0.3 7.5	14.8 0.5 9.3	21.2 1.0 11.3	28.9 1.8 13.8
B		0.3 0.1 1.3	0.6 0.1 1.5	1.1 0.2 1.9	2.1 0.5 2.6	3.5 0.9 3.4		4.8 0.1 4.8	6.7 0.1 5.9	10.0 0.3 7.3	15.0 0.6 9.1	21.3 1.0 11.3
C			0.3 0.1 1.2	0.6 0.1 1.4	1.1 0.2 1.9	2.1 0.5 2.6			4.8 0.1 5.2	6.7 0.1 5.9	10.1 0.3 7.2	15.0 0.6 9.1
D				0.3 0.1 1.2	0.6 0.1 1.5	1.2 0.2 2.1				4.8 0.1 5.1	6.8 0.1 6.0	10.1 0.3 7.4
E					0.4 0.1 1.3	0.6 0.1 1.6					4.8 0.1 4.8	6.7 0.1 5.7
F						0.3 0.1 1.3						4.8 0.1 5.0

Means used in the generation of the random samples; 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 26: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 60, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
A	1.2	3.1	7.5	15.6	28.0	43.6	4.3	9.2	18.3	32.0	48.4	65.0
	0.1	0.4	1.2	3.3	7.9	15.8	0.3	1.1	2.9	6.8	14.0	24.5
	4.4	7.2	11.2	16.4	22.0	27.8	14.2	20.7	28.6	37.0	44.8	52.4
B		1.1	3.1	7.6	15.6	28.0		4.2	9.2	18.5	31.9	48.4
		0.1	0.4	1.2	3.4	8.0		0.3	1.0	2.9	6.9	14.0
		4.3	7.0	11.0	16.1	22.0		15.2	20.9	28.3	35.9	44.5
C			1.1	3.1	7.6	15.6			4.2	9.2	18.4	31.8
			0.1	0.4	1.2	3.4			0.3	1.0	2.9	6.8
			4.1	6.9	11.1	16.2			15.3	20.8	28.3	36.6
D				1.1	3.1	7.3				4.2	9.2	18.2
				0.1	0.3	1.2				0.3	1.0	2.8
				4.1	6.9	11.1				15.2	20.7	28.4
E					1.2	3.1					4.3	9.1
					0.1	0.4					0.3	1.1
					4.2	7.1					15.0	20.5
F						1.2						4.3
						0.1						0.3
						4.2						14.1

Means used in the generation of the random samples; 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

Table 27: Percentage frequency of significant tests conditional on the compared object's positions in the rank order of the sample, n = 200, 7 objects to compare.

	frequency significant at 1%						frequency significant at 5%					
	B	C	D	E	F	G	B	C	D	E	F	G
A	3.4 0.4 11.4	15.1 2.9 28.4	40.5 13.6 51.5	71.1 37.3 75.1	91.0 67.6 88.1	98.3 89.0 94.8	10.5 1.3 23.3	32.3 7.6 47.1	63.1 26.0 72.6	87.1 55.4 88.7	97.2 81.9 96.4	99.7 95.4 98.9
B		3.4 0.4 14.7	15.1 2.9 34.2	40.5 13.3 58.3	70.9 37.2 77.0	90.9 67.6 87.4		10.8 1.4 29.3	32.3 7.8 53.3	63.1 26.1 75.7	87.1 55.3 90.0	97.2 81.7 96.3
C			3.4 0.3 18.7	15.1 2.9 38.7	40.3 13.5 57.5	70.8 37.6 72.3			10.7 1.5 31.0	32.4 7.9 54.3	62.8 26.0 75.6	86.9 55.3 88.5
D				3.4 0.4 18.6	15.1 2.9 34.0	40.7 13.4 50.7				10.7 1.4 30.7	32.2 7.8 53.3	63.2 26.1 72.7
E					3.4 0.4 14.6	15.1 2.9 28.3					10.7 1.4 29.5	32.3 7.7 47.0
F						3.4 0.4 11.3						10.6 1.3 23.3

Means used in the generation of the random samples; 1.3, 1.2, 1.1, 1.0, 0.9, 0.8, 0.7. The first number in each cell refers to the conventional p-value, the second to the HB adjusted conventional p-value and the third to the conditional p-value.

5 Conclusions

Is the conditional p-values meaningful to use? Let us first consider the case where just one test is made. Remember that the conventional p-value is in most samples not viable in the upper right cells in table 4.3. Thus the conditional p-value would give a higher power than the conventional in most cases where both p-values are valid. The largest advantage with the conditional p-values is that you know that the test is always valid. If you use the conventional p-values and chooses what pairs to test after you have investigated descriptive statistics from your samples you need to run a Monte Carlo analysis to find out what combinations that are still valid to test with the conventional p-values.

If all tests are made, the conditional p-values would not be meaningful since they are derived from the choice of test to make. So in that case we go for the HB-adjusted conventional p-values. If more than one test are to be made it is quite difficult to say anything generally. The outcome would depend on how many and which combinations that are made, since that would determine the outcome of the HB adjustment. But if we use objects close to each other the higher power of the conditional p-values would of course also give a higher power after a HB adjustment. Evaluating different strategies for choosing which objects to test is an important question for future research but out of the scope of this paper.

Acknowledgements:
We are grateful for ...

References

Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilitÃ , Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze

Chatfield, C. (1995) "Model Uncertainty, Data Mining and Statistical Inference". Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 158, No. 3(1995), pp. 419-466

Dunn, O. J. (1961) "Multiple Comparisons Among Means". Journal of the American Statistical Association 56 (293) 52-64. doi:10.1080/01621459.1961.10482090.

Friedman, Milton (1937) "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". Journal of the American Statistical Association (American Statistical Association) 32 (200): 675-701

Holm, S. (1979) A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6 65-70.