Proceedings of the 2011 Winter Simulation Conference
S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.

A NOTE ON THE USE OF MULTIPLE COMPARISON SCENARIO TECHNIQUES IN EDUCATION AND PRACTICE

Kathryn A Hoad

Warwick Business School
University of Warwick
Coventry, UK

Thomas Monks

Peninsula College of Medicine and Dentistry
University of Exeter
Exeter, UK

ABSTRACT

Our main aim in this paper is to highlight current practice and education in multiple scenario comparison within DES experimentation and to illustrate the possible benefits of employing false discovery rate (FDR) control as opposed to strict family-wise error rate (FWER) control when comparing large numbers of DES experimentation scenarios in an exploratory manner. We present the results of a small survey into the current practice of scenario analysis by simulation practitioners and academics. Although the survey was small, the results indicated that the range of scenarios used in DES studies may prohibit the use of FWER control methods such as the Bonferroni Correction referred to in DES textbooks. Furthermore, 80% of our sample were not familiar with multiple comparison control. We provide a practical example of the FDR in action and argue that it is preferable to employ FDR instead of no multiple comparison control in exploratory style studies.

1 INTRODUCTION

Performing multiple comparisons, of simulation scenarios, can cause inflation of the probability of making a single type I error - more commonly known as the Familywise Error Rate (FWER) (Hochberg 1988, Benjamini and Hochberg, 1995). It is therefore advised in most standard discrete-event simulation (DES) texts to apply multiple comparison control (MCC) using a (classical) Bonferroni Correction to the alpha level used (see Law, 2007 p537; Robinson, 2004 p180, Banks et al., 2005 p449). A Bonferroni Correction ensures that the overall probability of making a Type I error remains at level alpha. This control of the FWER is especially important when choosing the best system design out of *n* possible designs; for example, selecting the process design change that maximises throughput in a factory. However, there are two issues with its application.

The first, well documented, issue relates to the feasible range of comparisons that can be made with a Bonferroni Correction. The consensus is that the upper bound is only 10 comparisons (Robinson, 2004; Swisher et al., 2003). Beyond this range confidence intervals convey little information and type II error rates drastically increase due to the strictness of hypothesis tests employed. The second issue relates to how appropriate it is to control the FWER. If a decision maker wishes to understand how (management) decisions or factors of a system affect the performance of a system then it is desirable to have a better trade-off between type I and II errors. This is particularly relevant when comparisons are used to explore a problem rather than to choose a 'best' scenario. Strong control of the FWER can produce more type II errors and hence lead to the failure to recognise the influence of some experimental factors on performance.

An arguably more suitable procedure for this type of exploratory experimentation - False Discovery Rate (FDR) control - was (re)introduced by Benjamini and Hochberg (1995). Previous work on this concept had been attempted but unpublished by Elkund, (Benjamini & Hochberg, 2000), and this work was reported by Seeger (1968). Here the problem of multiple comparisons is conceptualised differently to that seen in the DES literature. This approach controls the proportion of false rejections (discoveries) among all rejected hypotheses, as opposed to the previous methods that aimed to keep strict control of the FWER, i.e. making a single type I error (Benjamini and Hochberg, 1995; Verhoeven et al., 2005).

A useful property of FDR is that it also controls the FWER in the *weak sense* (Benjamini and Hochberg, 1995). Consider an experiment where M comparisons between scenarios are made. If all of the M null hypotheses are true (i.e. there is no evidence to suggest differences between scenarios) then FDR is equivalent to Bonferroni. These claims are backed up by several simulation studies (e.g. Benjamini and Hochberg, 1995, 2000). It has also been shown that the FDR approach can be a more powerful method than the Bonferroni method (Benjamini and Hochberg, 1995). Since that first paper many 'improved' or otherwise connected methods have been published (e.g. Verhoeven et al., 2005; Benjamini and Hochberg, 2000; Genovese and Wasserman, 2002; Storey, 2002; Storey & Tibshirani, 2003; Benjamini and Yekutieli, 2001).

This concept of FDR control has gained popularity in a number of disciplines e.g. evolution, ecology, biology, genetics (Garcia 2003,2004; Verhoeven et al., 2005; Benjamini et al., 2001), especially those that are more exploratory in nature, where large numbers of hypotheses are required to be tested, but where the strict control of the FWER can be relaxed (Black, 2004). It is this idea of exploration that is particularly compelling. In (DES) experimentation it can be argued that two main objectives exist: to explore the solution space in order to learn more about the 'important' factors and solution possibilities and/or to find the optimal solution or 'best of a subset' of possible solutions. It is important when your aim is to find the 'best' scenario, that the probability of making a type I error, an erroneous discovery, is kept to a minimum. The Bonferroni type methods strictly control the probably of making one or more type I errors, but at a cost of power, hence increasing the risk of overlooking 'real' differences between scenarios. It can also be argued that when the main aim is to explore the solution space by comparing large numbers of scenarios, it is important to try to reduce the number of type II errors, while still keeping some control of type I errors. The family of FDR controlling methods claim to achieve this aim.

In order to set this discussion in context, a survey of DES users was undertaken to explore the extent and details of scenario comparison usage in practice. The preliminary results of this survey are presented in the next section. We continue by describing the FDR method (as first proposed by Benjamini and Hochberg 1995) and one of its adaptations (Benjamini and Hochberg 2000) along with some of the more popular FWER controlling methods. We then present an example (using artificial data) of the use of several multiple comparison control methods in order to promote discussion regarding the performance and relevance of FDR and FWER techniques for DES scenario comparison.

2 A SURVEY OF THE USEAGE OF MCC IN PRACTICE

Before proceeding to a discussion of MCC procedures, it is important to understand the extent of their usage in practice. This section presents the results of a survey into the use of multiple comparison control in practice. The survey explored two main questions: How many scenarios are typically compared in DES studies? Are MCC procedures typically used in DES, and if not, why not? The survey was sent out to 60 (UK) simulation practitioners/academics and there was a 42% response rate. The majority of participants had experience of 6-20 DES studies (60% of respondents) while a number of respondents had participated in over 30 DES studies (24%).

2.1 Do simulation studies typically compare a large number of scenarios?

The first objective of the survey was to gauge if the typical range of scenarios analysed in DES studies was suitable for the application of classical MCC procedures, such as Bonferroni, or if the range effectively prohibited use of MCC. Figure 1 illustrates the typical number of scenarios respondents indicated they analyze. Given that the respondents indicated that they frequently used pair-wise comparisons in their DES studies (measured on a five point scale; 1 = Never, 5 = always; median = 4), it is notable that the majority of the responses (63%) were above six scenarios (and hence 10 comparisons) the point at which Bonferroni intervals lose the ability to convey useful information on differences. However, note that this means that 37% of respondents are typically operating in ranges where Bonferroni can be applied.

It is also important to note that 36% of respondents stated that they had conducted at least one DES study where the most scenarios they had analysed exceeded 100. We found very little evidence of alternative analysis methods such as meta-modelling (measured on a five point scale; 1 = Never, 5 = always; median = 1) or ranking and selection procedures were in regular use (median = 2); hence it is plausible to assume that direct scenario comparisons are used even with scenarios this high.

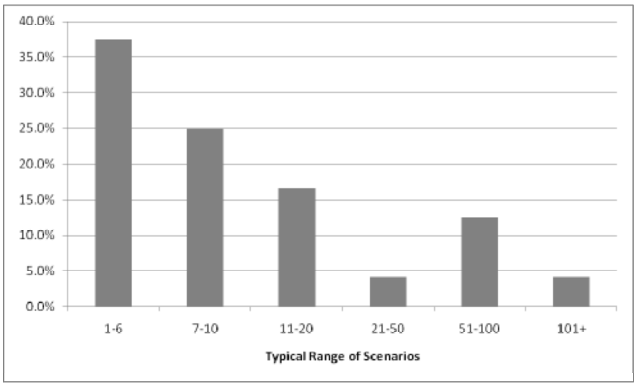

Figure 1: Typical ranges of scenarios for respondents

2.2 What if any multiple comparison control procedures are used?

Respondents were shown a list of commonly used multiple comparison control procedures (including the classical Bonferroni correction found in simulation texts) and asked if they were familiar with them. As suspected, only 20% of participants had heard of the MCC procedures listed. The respondents familiar with the procedures were then asked how frequently they employed them in practice. Of the five listed, the only MCC procedure found to be in use was the Bonferroni Correction and this appears to have been quite rare (median = 2.5).

2.3 Why are MCC procedures not used in practice?

The results of section 2.2. are not surprising, given the typical ranges of scenarios analysed by respondents (section 2.1). Although other authors discussing MCC have noted the breakdown of classical Bonferroni procedures beyond 10 comparisons, we wanted to explore other reasons blocking usage that incorporated both a statistical and practical worldview. Table **1** presents the results that respondents selected from a multi-choice list. As can be seen a large proportion of respondents were not familiar with MCC at all (40%); furthermore a large proportion of respondents were familiar with MCC, but were unfamiliar with the procedures.

Table 1: Reasons for not using MCC

Answer	Responses
I am not familiar with the term 'multiple comparison control'	40%
I am not aware of the need to apply multiple comparison control in DES studies	8%
I am familiar with the term 'multiple comparison control', but not with any of the procedures	20%
The multiple comparison procedures I am familiar with are too strict to be of practical use	16%
I do not have software that provides multiple comparison control functionality	8%
Multiple comparison control is unnecessary in DES studies	0%
Other e.g. “largely unnecessary, clients would not understand”	12%

2.4. Discussion of Survey Results

Clearly, given the sample size of the survey, it should be acknowledged that the results cannot be generalised as the characteristics of scenario analysis in all DES studies. However, as noted in section 2.1, the respondents were generally experienced DES modelers; hence it is believed that the results may hold some indications of typical practice worthy of further consideration.

The results showed that DES modelers appear largely unaware of the FWER and procedures that control it. This is perhaps surprising given in our sample the majority of respondents had a masters degree (56%) incorporating DES training (there was no evidence of an association between awareness of FWER & masters level training: $\chi^2(1) = .65, p > .1$). As an attempt to clarify the state of current UK DES teaching we contacted Simulation (UG/PG) module leaders at 7 different UK universities and asked whether formal scenario comparison methods and MCC procedures were covered within their modules. It appears from this informal survey that most courses do not go any further than formal comparison by t-test (as per Robinson 2004 or Law 2007) and that most modules do not currently cover MCC methods. Clearly there is some scope to improve simulation education on MCC.

In general the results show that DES projects typically involve ranges of scenarios outside of those feasible for classical MCC procedures such as the Bonferroni correction and that they are used rarely. This suggests that alternative procedures, such as FDR, that provide a better trade-off between Type I and Type II errors would be a useful addition to the DES literature. To address these points the next section of this paper provides a brief review of MCC methods to control the FWER and the FDR. This is followed by an example of the methods in action and how FDR provides a better trade-off between Type I and II errors than classical MCC procedures.

3 A REVIEW OF FDR AND FWER CONTROLLING METHODS

Firstly, let us clarify the difference between FDR and FWER. Consider carrying out m statistical hypothesis tests. Of these m individual tests, there are m_0 true null hypotheses and m_1 (= m-m_0) false null hypotheses. All the possible outcomes of these m tests are shown in table 2 (Benjamini and Hochberg, 1995). Obviously, in reality, only m and r are known. The FDR is defined by Benjamini and Hochberg (1995) as the expected proportion of type I errors (V) among all the significant results (r), i.e. E[V/r]. The FWER, however, is defined as the probability that the number of type I errors (V) is greater than or equal to one, i.e. $P(V \geq 1)$. That is, the FWER refers to the probability of making a single type I error in m comparisons.

Table 2: This table displays all possible outcomes of individual statistical hypothesis tests. V = the number of type I errors; T = the number of type II errors. Only m, r and m-r are observable. U, V, T, S and m_1 are unknown.

Truth	Decision		Total
	H_0 not rejected	H_0 rejected	
H_0 true	U	V	m_0
H_0 false	T	S	m_1
Total	m-r	r	m

Methods devised to control the FDR therefore keep the expected proportion of V/r at a chosen level. Methods devised to control the FWER therefore keep the $P(V \geq 1)$ at or below a desired level. Classical Bonferroni, is designed to keep the (FWER) probability of making even one type I error to α/m. As an example consider Table 3 that illustrates the individual confidence intervals needed to maintain an overall 5% level of significance when comparing different numbers of scenarios in a full pairwise fashion. The table illustrates how quickly the Bonferroni Correction effects the practical usefulness of results. For instance, beyond five scenarios the individuals confidence intervals are above 99.5%. Not only is this incredibly strict, but any useful information of the mean difference between scenarios is lost. Note that even if an overall 10% level of significance was set the individual confidence intervals become extremely strict and difficult to interpret (i.e. approximately 99.5% and above) beyond seven scenarios.

Table 3: Feasible and Infeasible use of a Bonferroni Correction in full pair-wise comparisons (α=0.05)

Scenarios	**Comparisons (m)**	**α (Bonferroni)**	**CIs**	
3	3	0.017	98.3	Interpretable, but wide CI's
4	6	0.008	99.2	
5	10	0.005	99.5	
6	15	0.003	99.7	Strict and CIs of little use
7	21	0.002	99.8	
8	28	0.002	99.8	
9	36	0.001	99.9	
10	45	0.001	99.9	

This is an issue present in all comparison problems, not just DES; hence the drive to produce less conservative versions of Bonferroni based on a sequential analysis of p-values (Holm, 1979; Hochberg, 1988; see Appendix for algorithms). However, the general consensus is that these methods are still on the conservative side (e.g. Verhoeven et al., 2005). This conservatism is particularly problematic in studies of a more *explorative* nature and has therefore led to the search for alternative methods that produce a better trade-off between type 1 and 2 errors. Thus the family of FDR methods was developed.

Like the step-down and up Bonferroni procedures, FDR is a sequential method that operates on the ranked p-values. The control algorithm for the original FDR procedure (Benjamini and Hochberg, 1995) is presented in Figure 2. There are several variations to this approach available and a full review of these methods is beyond the scope of this paper. However, a simple variation can be inferred by referring back to Table 2. Recall that the unknown value m_0 represents the total number of hypotheses that are 'true'. If the value of m_0 was in fact known then the FDR could be controlled at exactly level V/r. Hence if some of the m hypotheses are false ($m_0 < m$) then there is a benefit to statistical power if this number can be estimated (Benjamini and Hochberg, 2000). This estimation can be done empirically based on the resultant test p-values. A simple way to think of this result is that we are generating extra information based on test results and hence our tests gain more power. The procedure to perform the ‘adaptive’ FDR algorithm has both an algorithmic and graphical form and is very simple to implement. The algorithm is summarised in the Appendix; interested readers are referred to Benjamini and Hochberg (2000) for more details on the approach. Before discussing some theoretical results

Table 4 provides a simple overview of the differences in p-value rejection levels for the Bonferroni and FDR procedures discussed (adapted from Verhoeven, et al, 2005).

1. Carry out m hypothesis tests (i.e. scenario comparisons) and calculate m corresponding p-values.
2. Rank the p-values in ascending order: $p^{(1)} \leq p^{(2)} \leq p^{(3)} \leq \ldots \leq p^{(m)}$
3. Define $H^{(i)}$ as the null hypothesis associated with the p-value $p^{(i)}$
4. For i = m, m-1, m-2, … , 1, let n be the largest i for which $p^{(i)} \leq i\alpha/m$.
5. Reject all null hypotheses from $H^{(1)}$ up to and including $H^{(n)}$ (i.e. reject the null hypothesis corresponding to $p^{(n)}$ as well as all those having smaller p-values).

Figure 2: False Discovery Rate Algorithm (Benjamini and Hochberg, 1995)

Table 4: A comparison of p-value rejection levels for five multiple comparison methods (3 FWER controlling methods and 2 FDR controlling methods): (1) Holm, 1979; (2) Hochberg, 1988; (3) Benjamini and Hochberg, 1995; (4) Benjamini and Hochberg, 2000. (An extension of table from Fig.1 from Verhoeven et al., 2005).

Tests ranked by ascending p-value	Classical Bonferroni FWER control	(Step-down) Sequentially rejective Bonferroni FWER control (1)	(Step-up) Sharper Sequential Bonferroni FWER control (2)	FDR (3)	Adaptive FDR control, where $\hat{m}_0$ is the estimate of m_0. (4)
1	α/m	α/m	α/m	α/m	$\alpha/\hat{m}_0$
2	α/m	α/(m-1)	α/(m-1)	2α/m	$2\alpha/\hat{m}_0$
3	α/m	α/(m-2)	α/(m-2)	3α/m	$3\alpha/\hat{m}_0$
⋮	⋮	⋮	⋮	⋮	⋮
k	α/m	α/(m-k+1)	α/(m-k+1)	kα/m	$k\alpha/\hat{m}_0$
⋮	⋮	⋮	⋮	⋮	⋮
m	α/m	α	α	α	$m\alpha/\hat{m}_0$

4 SOME THEORETICAL RESULTS

When all the null hypotheses are true the FWER ≡ FDR as shown in table 5. When not all of the null hypotheses are true and $S > 0$, then the FWER ≥ FDR as shown in table 6 (Benjamini and Hochberg, 1995). Hence for the cases where $S > 0$, controlling the FDR at a set level can lead to fewer type II errors than controlling the FWER at the same level. Hence the power of each test can be greater under FDR control than equivalent FWER control. This advantage of FDR control increases as the number of false null hypotheses increases (Benjamini and Hochberg, 1995)

Table 5: The values that the FDR and FWER take when all the null hypotheses are true.

Number of type I errors	FDR values	FWER values
V = 0	FDR = E[V/r] = E[0] = 0	FWER = P(V ≥ 1) = 0
V > 0	V = r as all rejections of H_0 are false i.e. type I errors, hence FDR = E[1] = 1	FWER = P(V ≥ 1) = 1

Table 6: The values that the FDR and FWER take when not all the null hypotheses are true and S > 0.

Number of type I errors	FDR values	FWER values
V = 0	FDR = E[V/r] = E[0] = 0	FWER = P(V ≥ 1) = 0
V > 0	Since r = V + S, for V/r to equal 1, S must be 0. Hence for S > 0, 0 < FDR = E[V/r] < 1	FWER = P(V ≥ 1) = 1

This method has been shown to control the FDR when all m tests are independent (Benjamini and Hochberg, 1995) as well as when the tests are positively correlated (Benjamini and Yekutieli, 2001).
Benjamini and Yekutieli (2001) provides an alternative cut-off level for the above procedure that provides FDR control when the tests are either negatively correlated or have a more complicated dependence structure. The recommended modification is more conservative than the original procedure (Verhoeven et al., 2005) and involves substituting the following instruction for line 4 in the original procedure listed in Figure 1:

- For i = m, m-1, m-2, … , 1, let n be the largest i for which $p^{(i)} \leq i\,\alpha\, m \sum_{i=1}^{m} \frac{1}{i}$.

5 PRACTICAL EXAMPLE

In order to illustrate the application of the various methods discussed above we created a set of 55 artificial p-values, to represent the p-values that would be produced from testing differences between 10 scenarios in a full pair-wise comparison. Using Verhoeven et al.'s, (2005), artificial example as a guide, we drew the artificial p-values from two representative distributions: 15 p-values representing true differences between compared scenarios (true alternative hypotheses) were drawn from a t-distribution with a mean > 0; 40 p-values representing true null hypotheses (no difference between compared scenarios) were therefore drawn from a Uniform[0,1] distribution. Normally, the degrees of freedom for a t-test of the difference between two scenarios would come from the number of replications (or batched means) used to calculate the overall mean KPI of interest for each scenario. This was set at 29 (for all comparisons) for this artificial example. We applied five methods of interest (3 FWER methods and 2 FDR methods) to the set of p-values to sort into significant and non-significant results. The experiment was repeated 20 times using different random numbers to create 20 different sets of p-values. The decisions (mean & standard deviation) of each method are displayed in Figure 3. In these examples, the FDR control methods produced no type II errors in contrast to the FWER methods that failed to highlight approximately half of the significant alterative results. The FDR methods were predictably less strict regarding

Type I errors and consistently made more than the expected number, but were relatively close to the expected values in the majority of cases (see Figure 4).

Classical Bonferroni

FWER	Decision	
	Not Sig	Significant
True H_0	40 [0]	0 [0]
True H_a	7.8 [1.32]	7.2 [1.32]

Step-down sequential Bonferroni (Holm 1979)

FWER	Decision	
	Not Sig	Significant
True H_0	40 [0]	0 [0]
True H_a	7.2 [1.40]	7.8 [1.40]

Step-up sequential Bonferroni (Hochberg, 1988)

FWER	Decision	
	Not Sig	Significant
True H_0	40 [0]	0 [0]
True H_a	7.2 [1.40]	7.8 [1.40]

Benjamini & Hochberg FDR (Benjamini & Hochberg, 1995)

FDR	Decision	
	Not Sig	Significant
True H_0	38.15 [0.88]	1.85 [0.88]
True H_a	0 [0]	15 [0]

Adaptive Benjamini & Hochberg FDR (Benjamini & Hochberg, 2000)

FDR	Decision	
	Not Sig	Significant
True H_0	38.05 [0.83]	1.95 [0.83]
True H_a	0 [0]	15 [0]

Figure 3: The significance decisions (mean [sd]) made by five FWER & FDR methods acting upon (20 sets of) 55 artificial p-values, representing 15 true alternatives and 40 true nulls: *number of comparisons = 55, number of true alternatives = 15, number of true nulls = 40, effect size (i.e. true mean difference between scenario means) = 3, nominal alpha = 0.05, degrees of freedom for t-tests (i.e. number of replications or batched means – 1) = 29.*

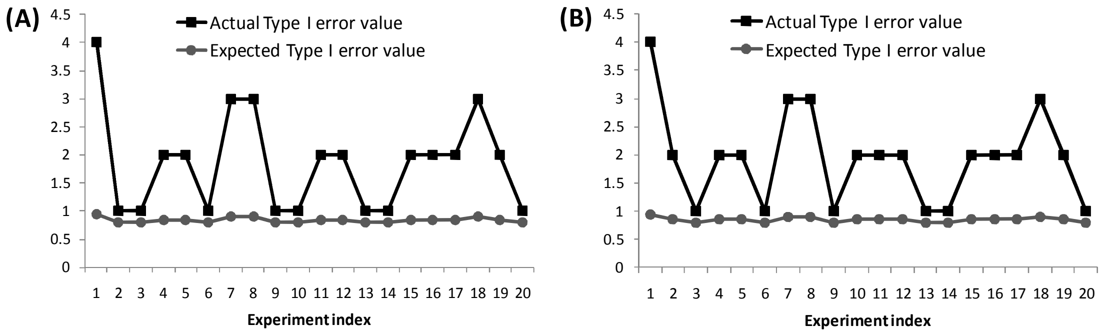

Figure 4: Actual Type I error values compared with expected Type I error values, for (A) *Benjamini & Hochberg FDR control method* (Benjamini & Hochberg, 1995) and (B) *Adaptive Benjamini & Hochberg FDR control method* (Benjamini & Hochberg, 2000).

6 DISCUSSION AND CONCLUSION

Our main aim in this paper was to highlight current practice in multiple scenario comparison and to illustrate the possible benefits of employing FDR control when comparing large numbers of DES experimentation scenarios in an exploratory manner. We believe that it is of benefit to simulation users to be aware of and be able to access the various techniques discussed in this paper among the many tools available when experimenting with DES models.

Given the small sample size of the survey detailed in section 2, the survey results should only be viewed as indications of scenario analysis and comparison in practice. The results highlighted the fact that DES users are comparing scenarios outside of the feasible region for a classical Bonferroni Correction. It also suggested that Bonferroni corrections and MCC in general are rarely used and that knowledge of MCC procedures is low.

In response to the first finding one option in multiple comparison problems in DES studies is to simply ignore it and perform all comparisons at an individual significance level. As one respondent to the survey pointed out 'any unexpected results should be investigated further by making additional runs'. This is an important difference between DES studies and other empirical sciences; as, for example, psychologists rarely have the luxury of 'cheap' collection of additional data. DES users also have access to the mechanism, i.e. the model, that generated the 'unexpected results' and can use it to help manage possible false discoveries. However, some of this effort could be reduced in exploratory studies, i.e. those that study the factors that influence performance, if an FDR type procedure was adopted. This provides a good balance between wrongly concluding differences are present due to an inflated FWER and missing important findings due to an overly conservative Bonferroni procedure.

This discussion also feeds into point three and the education of DES users on strategies to tackle the multiple comparison problem in practice. We believe that University courses and DES texts should provide some practical advice to students on the measures that they can take. Clearly the classical Bonferroni Corrections we found in DES texts are insufficient. One alternative is the FDR which provides an extra layer of protection against inference errors. To help students and practitioners we have provided a simple spreadsheet designed to be used in the comparison of scenarios in DES studies. It includes the Bonferroni and FDR procedures detailed in this paper and can be downloaded from here [http://www2.warwick.ac.uk/fac/soc/wbs/subjects/orms/people/katyhoad]:

REFERENCES

Banks, J., J.S.Carson II, B.L.Nelson, D.M.Nicol, *Discrete-Event System Simulation*, Prentice Hall Int., NJ, (4th Ed) 2005.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. - *J. Roy. Stat. Soc. B* 57: 289-300.

Benjamini, Y. and Hochberg, Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. - *J. Educ. Behav. Statist*. 25: 60-83

Benjamini, Y., D.Drai, G.Elmer, N.Kafkafi and I.Golani, (2001), Controlling the false discovery rate in behaviour genetics research, *Behavioural Brain Research*, 125, pp279-284.

Benjamini, Y. and D.Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Ann. Stat*. 29, pp1165-1188.

Black, M.A., 2004, A note on the adaptive control of false discovery rates, *J.R.Stat.Soc.B* Met. 66, pp207-304.

Garcia, L.V., 2003, Controlling the false discovery rate in ecological research, *Trends Ecol. Evol*. 18, pp553-554.

Garcia, L.V., 2004, Escaping the Bonferroni iron claw in ecological studies, *OIKOS*, 105, pp657-663.

Genovese, C and L.Wasserman, 2002, Operating characteristics and extensions of the false discovery rate procedure, *J.R.Stat.Soc.B*, 64, pp499-517.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. - *Biometrika* 75: 800-802.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. - Scand. *J. Stat*. 6: 65-70.

Law A.M, *Simulation Modeling & Analysis*, McGraw-Hill Int., NY, (4th Ed) 2007.

Robinson, S., Simulation: *The practice of model development and use*, John Wiley & Sons, Ltd, 2004

Seeger, P. (1968) A note on a method for the analysis of significances en mass. *Technometrics*, 10, pp586-593.

Storey, J.D., 2002, A direct approach to false discovery rates, *J.R.Stat.Soc.B*, 64, pp479-498.

Storey, J.D. and R.Tibshirani, 2003, Statistical significance for genomewide studies, *Proc.Natl.Acad.Sci.USA*, 100, pp9440-9445.

Swisher, J.R., S.H.Jacobson, S.H. and E.Yucesan, (2003) Discrete-Event Simulation optimization using ranking, selection and multiple comparison procedures: A survey. *ACM Transactions on Modeling and Computer Simulation*, Vol 13, No 2, pp134-154.

Verhoeven, K.J.F, K.L.Simonsen and L.M.McIntyre, Implementing false discovery rate control: increasing your power. *OIKOS* 108, pp643-647, 2005.

AUTHOR BIOGRAPHIES

KATHRYN A. HOAD is an assistant professor in the Operational Research and Management Sciences Group at Warwick Business School. She holds a BSc(Hons) in Mathematics and its Applications from the University of Portsmouth, an MSc in Statistics and a PhD in Operational Research from the University of Southampton. Her email address is <kathryn.hoad@wbs.ac.uk>

THOMAS MONKS is an associate research fellow in the Peninsula Medical School, University of Exeter and doctoral researcher at Warwick Business School supervised by Professor Stewart Robinson and Dr Kathy Kotiadis. He holds a BSc (Hons) in Computer Science and Applicable Mathematics from Staffordshire University and an MSc in Operational Research from Lancaster University. He has worked as both a Software Engineer in the private sector and an Operational Research Analyst within the public sector. His research interests include Discrete Event Simulation, Reuse and Management Learning. His e-mail address is <thomas.monks@pcmd.ac.uk>

APPENDIX

Adaptive FDR control algorithm (Benjamini & Hochberg, 2000)**:**

- Specify an acceptable False Discovery Rate, q, prior to carrying out comparisons.
- Carry out m hypothesis tests (i.e. scenario comparisons) and calculate m corresponding p-values.
- Rank the p-values in ascending order: $p^{(1)} \leq p^{(2)} \leq p^{(3)} \leq \ldots \leq p^{(m)}$
- Define $H^{(i)}$ as the null hypothesis associated with the p-value $p^{(i)}$
- For i = m, m-1, m-2, … , 1, if all $p^{(i)} \geq iq/m$ do not reject any hypotheses (stop).
- If this is not the case, starting with i = 1, calculate $S_i = (1 - p^{(i)})/(m+1-i)$, and continue calculating as long as $S_i \geq S_{i-1}$.
- Let n be the smallest i for which $S_i < S_{i-1}$.
- Set $\hat{m}_0$ = Minimum of $1/ S_n$ (rounded up to an integer value) and m.
- For i = m, m-1, m-2, … , 1, (starting at i = m), let k be the largest i for which $p^{(i)} \leq iq/\hat{m}_0$.
- Reject all null hypotheses from $H^{(k)}$ down to $H^{(1)}$.

Classical Bonferonni FWER control (e.g. Banks et al., 2005 p449; Robinson, 2007 p180)**:**

- Choose the desired overall significance level α.
- Carry out the m hypothesis tests (i.e. scenario comparisons), each at an individual significance level of α/m.
- Reject any null hypothesis for which the p-value < α/m.

Sequentially rejective (step-down) Bonferonni FWER control (Holm, 1979)**:**

- Carry out m hypothesis tests (i.e. scenario comparisons) and calculate m corresponding p-values.
- Rank the p-values in ascending order: $p^{(1)} \leq p^{(2)} \leq p^{(3)} \leq \ldots \leq p^{(m)}$
- Define $H^{(i)}$ as the null hypothesis associated with the p-value $p^{(i)}$
- For i = 1, …, m-2, m-1, m, (starting at i = 1) let n be the largest i for which $p^{(i)} \leq \alpha/(m-i+1)$.
- Reject all null hypotheses from $H^{(1)}$ up to and including $H^{(n)}$ (i.e. reject the null hypothesis corresponding to $p^{(n)}$ as well as all those having smaller p-values).

Sharper (Step-up) Sequential Bonferonni FWER control (Hochberg, 1988)**:**

- Carry out m hypothesis tests (i.e. scenario comparisons) and calculate m corresponding p-values.
- Rank the p-values in ascending order: $p^{(1)} \leq p^{(2)} \leq p^{(3)} \leq \ldots \leq p^{(m)}$
- Define $H^{(i)}$ as the null hypothesis associated with the p-value $p^{(i)}$
- For i = m, m-1, m-2, … , 1, (starting at i = m) let n be the largest i for which $p^{(i)} \leq \alpha/(m-i+1)$.
- Reject all null hypotheses from $H^{(n)}$ down to $H^{(1)}$ (inclusive).