A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure

Juliet Popper Shaffer, University of California,

Department of Statistics, 367 Evans Hall # 3860,

Berkeley, CA 94704-3860, USA

February 9, 1998

AMS classification: 62J15; 62F03; 62C12

Keywords: Hypthesis testing; Decision theory; Bayesian decision theory; Mean comparisons, False Discovery Rate

Abstract

Duncan's Bayesian decision-theoretic multiple comparison procedure requires a decision on the relative magnitudes of losses due to Type I and Type II errors. In this paper, the relative losses are chosen so that the procedure results in weak control of familywise error at the .05 level, i.e. the probability that all hypotheses are accepted is .95 when all hypotheses are true. Duncan's Bayesian formulation requires prior distributions and specification of associated hyperparameters for the variances

of the population means and of the errors. With noninformative priors, the required ratio of these values can be estimated from the sample. From a frequentist point of view, this obviates the necessity for any prior specification for these distributions. However, Duncan's assumption of a prior normal distribution for the population means is required and is retained. A simulation study then compares the modified method, with respect to Bayes risk and average power, to several frequentist-based multiple comparison procedures for testing hypotheses concerning all pairwise comparisons among a set of means. Results indicate considerable similarity in both risk and average power between Duncan's modified procedure and the Benjamini and Hochberg (1995) FDR-controlling procedure, with the same weak familywise error control. Both risk and power of these procedures are close to the risk and power of individual t-tests of the mean comparisons, and considerably superior on both measures to the properties of the best symmetric simultaneous testing procedure, based on the range of normally-distributed observations.

1 Introduction

1.1 General overview

The purpose of this paper is to compare Duncan's Bayesian multiple comparison procedure (Duncan, 1965; Waller and Duncan, 1969) with commonly-used non-Bayesian procedures. The specific context is the testing of hypotheses concerning the differences among all means in a one-way layout with samples from m populations $\mathcal{P}_j$, with means μ_j, $j = 1 \leq j \leq m$. It will be assumed that the population distributions are normal and that the sample means

have equal variances σ_M^2 around their true values μ. (In many applications, σ_M^2 equals σ^2/r, where r is the number of replications, and error variance σ^2 is assumed equal for each observation).

In the next part of the Introduction, Duncan's procedure will be described, and notation appropriate for this study will be established. The remainder of the Introduction will present some background on frequentist approaches to the multiple comparison problem and compare them with Duncan's Bayesian approach. In Section 2, a modification of Duncan's Bayesian approach is suggested that will make it appropriate to compare the procedure to selected non-Bayesian multiple comparison methods. Section 3 describes the design of a simulation study comparing the methods; the results regarding error control, Bayesian risk, and power are presented in Section 4. This is followed by a discussion of these results in Section 5, and some concluding remarks in Section 6.

1.2 Duncan's Bayesian Decision-Theoretic Procedure

The differences among all pairs of means are designated δ_i, $i = 1, \ldots, n$, where $n = m(m-1)/2$. The subscripts are chosen so that the corrresponding sample differences d_i are ordered from largest to smallest, i.e. $d_1 \geq d_2 \geq \cdots \geq d_n$.

In addition to the standard non-Bayesian assumptions, it is assumed that the means μ_j, $j = 1, \ldots, m$, are a random sample from a normal distribution. Without loss of generality the mean of that distribution can be assumed to be zero, since only differences are of interest; the variance will be designated τ^2.

The hypotheses of interest are formulated as follows: For each difference δ_i, $i = 1, \ldots, n$,

there is a pair of hypotheses:

$$H_{i1}\ \delta_i \leq 0,\ H_{i2}\ \delta_i \geq 0. \tag{1}$$

The possible decisions are: Reject H_{i1} (decide $\delta_i > 0$), reject H_{i2} (decide $\delta_i < 0$), or reject neither (make no decision). The loss functions are as follows:

Loss function if $\delta_i \leq 0$

- Do not reject H_{i1}: Loss $= 0$
- Reject H_{i1}: Loss $= k_1|\delta_i|$
- Do not reject H_{i2}: Loss $= k_2|\delta_i|$
- Reject H_{i2}: Loss $= 0$.

If $\delta_i \geq 0$, the loss structure is the same, with subscripts 1 and 2 interchanged above on H_i (but not on k). Losses are then summed over the decisions on the two hypotheses, giving:

Loss function if $\delta_i \leq 0$

- Reject neither hypothesis: Loss $= k_2|\delta_i|$
- Reject H_{i1}: Loss $= (k_1 + k_2)|\delta_i|$
- Reject H_{i2}: Loss $= 0$.

with appropriate modifications, interchanging subscripts 1 and 2 on H_i when $\delta_i > 0$.

The loss over the whole procedure, testing the n pairs of hypotheses, is the sum of the losses over the individual pairs.

Let $k = k_1/k_2$. Note that k can be thought of as the ratio of the loss due to a Type I error to the loss due to a Type II error in testing a single directional hypothesis. The expected loss, or risk, equals

$$k_2 NULL + (k_1 + k_2) DIR \tag{2}$$

where $NULL$ is the expected value of nonrejected differences and DIR is the expected value of differences rejected in the wrong direction. Duncan (1965) showed that the procedure minimizing the risk depends only on k and on the variance ratio τ^2/σ_M^2.

The procedure involves testing each pair of hypotheses H_{ij} using the corresponding Student t-test: Accept both hypotheses (or reject neither–see interpretation in Section 1.3) if

$$|t| \leq \sqrt{\Psi/(\Psi - 1)}\, t_\infty. \tag{3}$$

Here $\Psi = (\sigma_M^2 + \tau^2)/(\sigma_M^2) = E(MSB)/E(MSW)$, MSB and MSW are the between-treatment and the within-treatment mean squares, respectively, in a one-way layout analysis of variance, t_∞ is the value of z for which

$$\frac{\phi(z) + z\Phi(z)}{\phi(-z) - z\Phi(-z)} = k,$$

and ϕ and Φ are the standard normal density and cumulative distribution functions, respectively. If $\Psi = 1$, no hypotheses can be rejected.

Note that this procedure and its proposed modification are unrelated to Duncan's New Multiple Range Procedure, which is incorporated in some computer packages. Since Duncan has proposed other multiple comparison procedures, I shall refer to this Bayesian procedure as DUB.

1.3 Background for comparison of Non-Bayesian Procedures to Duncan's Bayesian Procedure

For many years, the most widely-used criterion in evaluating non-Bayesian multiple hypothesis testing has been control of the familywise error rate (FWE), i.e. the probability of one or more rejections of true hypotheses (Type I errors) in the family of hypotheses under consideration, at some small level α. Rejection of a hypothesis is a strong conclusion, based on statistical evidence against its truth, while acceptance is a weak conclusion, signifying only that the evidence is insufficient to decide whether the hypothesis is true or false. When the family consists of a small number of hypotheses, and the overall conclusions depend on the joint outcome of the tests, the FWE criterion, assuring high probability against error (only a Type I error is really an error from this point of view), seems reasonable. However, when the family size is large, the conclusions typically are less dependent on joint correctness, and in that case control of FWE may be unnecessarily stringent.

Recently, an alternative error-protection criterion has been proposed: the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995). The FDR is the expected proportion of true hypotheses *among those that are rejected*, i.e. the expected value of Q/R, where Q is the number of falsely rejected hypotheses (true hypotheses that are rejected), and R is the total number of rejections. (When $R = 0$. the ratio is defined to be zero.) In distinction to control of FWE, control of FDR implies that a certain number of errors are permissible with high probability if they represent a sufficiently small proportion of the strong conclusions reached. If the test statistics are independent and the FDR is set at α,

an FDR-controlling method limits the FWE to α when all hypotheses are true (termed *weak control* of FWE by Hochberg and Tamhane (1987)), and empirical evidence suggests that this weak error control at level α holds also for comparisons among means in a one-way layout using the Benjamini and Hochberg (1995) FDR-controlling method.

DUB is similarly based on an alternative criterion which permits some errors with high probability providing that the benefits of doing so are sufficiently great. It has a Bayesian decision-theoretic interpretation that is seemingly incommensurable with the error control approaches based on both FWE and FDR.

There are four differences in the overall approach of Duncan, as compared to most FWE- and FDR-controlling procedures, which must be addressed in order to achieve a reasonable comparison:

1. As in all Bayesian approaches, prior distributions on the parameters are assumed.

2. Since the prior distribution of the means is assumed to be normal, there are no mean differences of zero (with probability 1), so no point null hypotheses are true.

3. The magnitude of error is taken into account.

4. There is no concept corresponding to control of either FWE or FDR.

These four points are elaborated on below.

1. Most non-Bayesian procedures treat the true means as arbitrary fixed values. As noted, Duncan assumes a normal distribution of true means. Furthermore, Duncan

must make some assumption about the variance of this distribution (the mean is irrelevant, as noted above), as well as the error variance.

2. Most non-Bayesian procedures test hypotheses that the true means equal zero, an outcome with a priori probability zero under the Bayesian assumption. However, since directional conclusions are usually desired, an alternative to formulating the n null hypotheses $\delta_i = 0$, $i = 1, \ldots, n$, is to formulate the $2n$ null hypotheses $\delta_i \leq 0$ and $\delta_i \geq 0$, $i = 1, \ldots, n$ as does Duncan. Note that this formulation makes sense whether one believes that a difference of zero is possible or not. Recently, Williams, Jones, and Tukey (1994) have investigated procedures under the assumption that the null hypothesis is never true, using the directional formulation above, with the possibility $\delta = 0$ omitted, and substituting the level $\alpha/2$ for α in order to make the procedure equivalent in size and power to those under the point null hypothesis formulation.

3. In most non-Bayesian formulations of hypothesis testing, the magnitude of departure from the null hypothesis, while it determines power, has no formal representation in the testing procedure. Nonetheless, in most situations it is more serious to fail to decide a direction of difference when the difference is large, or to make a directional error if the difference is large in the opposite direction. Duncan's procedure takes both of these considerations explictly into account.

4. In Duncan's procedure, the ratio $k(= k_1/k_2)$ is chosen to reflect the relative seriousness of Type I and Type II errors. There is no obvious relationship between the choice of k and the choice of the FWE. (But see Section 4.1.)

2 Proposed Modifications of Duncan's Procedure

The following modifications and considerations, corresponding to the four points discussed in the previous section, provide a basis for comparing the properties of Duncan's procedure to those of frequentist-based procedures.

1. Duncan specifies a prior distribution for Ψ with a hyperparameter in such a way that the posterior distribution is a linear combination of a prior value and a estimate from the data. Note that a natural data-based estimate of Ψ is the usual F-ratio MSB/MSW. That estimate will be adopted in this comparison, making prior assumptions on the relative variances of the true means and the sampling errors unnecessary from a frequentist point of view. The criterion for rejecting a hypothesis H_i will therefore be taken to be

$$F > 1 \quad and \quad |t| \geq \sqrt{F/(F-1)}\, t_\infty \tag{4}$$

   where t_∞ is defined in Equation (3).

2. The n comparisons will be formulated as $2n$ directional hypotheses, as in (1).

3. The criterion for rejection in Duncan's case becomes more lenient as the variance of the true means increases. While non-Bayesian procedures have no formal representation of the magnitude of overall mean differences, the Benjamini-Hochberg FDR-controlling procedures, and the stepwise non-Bayesian FWE-controlling procedures, implicitly share the property of making rejection of any fixed observed difference more

likely as the overall set of differences increases in absolute magnitude.

4. The ratio k, instead of being fixed by relative error considerations as in Duncan's procedure, will be chosen so as to make the traditional Type I error equal to α in the complete null case. This makes it possible to compare the properties of the procedure with non-Bayesian procedures with weak control of Type I error at the same level. It necessitates different choices of k for each value of m. Note that in the directional formulation of Duncan, the loss when all pairwise hypotheses are true is zero. By adjusting the ratio k to make $FWE = .05$ under the overall null hypothesis, a loss is implicitly introduced for the traditional Type I error under the usual nondirectional formulation of null hypotheses in pairwise comparisons. (Under that traditional formulation, which is the basis of the FDR approach, directional errors are often referred to as Type III errors.)

3 Description of the Study

In this study, Duncan's Bayesian procedure with criterion specified in Equation (5) is compared with non-Bayesian procedures when the true means are a sample from a normal distribution. The sampling variance of a mean, σ_M^2, is set equal to 1 and assumed known; this should provide a good approximation for situations in which σ_M^2 is estimated with large degrees of freedom. Therefore, instead of the ratio F in Equation (4), the criterion with estimated error variance replaced by a known value is equivalent to

$$\sqrt{MSB/(MSB-1)}\, t_{\infty .}, \tag{5}$$

where MSB is computed as if treatment means were based on samples of size 1, and t_∞, because σ_M^2 is assumed known and equal to 1, is replaced by $d_i/\sqrt{2}$. The FWE is set at α = .05 in the complete null case, thereby determining the value of t_∞ and thus of k.

The procedures are compared on FWE, FDR, and average power (non-Bayesian concepts), and Bayes risk as defined in Equation (2). Average power is the expected value of the number of (false) hypotheses rejected divided by n, the number of false hypotheses. All power results reported will be in terms of average power.

3.1 Procedures

The non-Bayesian procedures compared with DUB are as follows.

1. The ordinary t test for each difference, ignoring multiplicity and assuming a known error variance $\sigma_M^2 = 1$:

$$Reject\ H_i\ if\ |d_i| > \sqrt{2}z_{.025},$$

where $z_{.025}$ is the upper .025 critical value on the standard normal curve. If H_i is rejected, make a directional decision corresponding to the sign of d_i. This procedure is designated SEP, since the hypotheses are treated separately without regard to their multiplicity, and of course it does not control either the FWE or the FDR.

2. The single-stage procedure based on the distribution of the range

$$Reject\ H_i\ if\ |d_i| > q_{m,.05},$$

where $q_{m,.05}$ is the upper .05 critical value of the range of m standard normal means. (This is the Tukey (1953) studentized range test adapted to known variance.) If H_i is

rejected, make a directional decision as in 1. This would be the optimal symmetric non-Bayesian simultaneous procedure with .05-level FWE control. This procedure is designated $RANGE$.

3. The FDR-controlling procedure in Benjamini and Hochberg (1995): Let p_i be the significance probability of $|d_i|$. Then the p_i are ordered from smallest to largest. Let j be the largest subscript i for which $p_i \leq i\alpha/n$. If there is no such subscript, accept all H_i. Otherwise reject all H_i with $i \leq j$. Make directional decisions as in the two methods described in Points 1 and 2. This procedure is designated $FDR1$, to distinguish it from other FDR-controlling procedures.

3.2 Conditions

All simulations were carried out using Splus on a Sun Ultra-1 workstation.

The number of means was varied from 2 to 100: See Table 1. (Although all procedures are equal when there are two means, the various power and risk measures were computed to assess the continuity between the cases of two and more than two means.) For each number m of means, the value of t_∞ was set to make the FWE in the null situation equal to .05, so that each of the procedures (except SEP) provided weak control of the FWE at level .05. The t_∞ values were approximated by simulating the null situation 100,000 times and using a value of t_∞ to two decimal points that gave the FWE closest to .05. The values of t_∞ and the corresponding values of k are given in Table 1.

For each number m of means, the variance τ^2 was set to make the average power for $RANGE$ approximately .50. Power was then approximated for each of the procedures

No of Means	t_∞	k(risk ratio)
2	1.69	91
3	1.91	178
4	2.01	244
6	2.09	315
8	2.12	347
12	2.12	347
16	2.09	315
20	2.07	296
30	2.02	252
40	1.97	215
50	1.93	190
100	1.78	120

Table 1: Number of means, t_∞, and k

based on 20,000 replications. Checks of other values of τ^2 indicated that the comparative powers of the procedures remained the same over a wide range of values of τ^2. A number of checks of the whole process at $\alpha = .10$ indicated that the comparative powers of the procedures behaved similarly at that FWE level.

4 Results

Results will be presented primarily in graphical form. Numerical values can be obtained from the author.

4.1 Values of t_∞ and k for which null $FWE = \alpha$

Let m equal the number of means. The values of t_∞ required to make the FWE of DUB equal to .05 in the complete null case vary relatively little within the range $m = 3$ to $m = 50$, increasing from 1.91 (for $m = 3$) to 2.12 (for $m = 8$ and $m = 12$) and then decreasing to 1.93 (for $m = 50$). The values are moderately lower for $m = 2$ and $m = 100$. If a single value $t_\infty = 2$ were to be used (corresponding to a constant value k of 237), the FWEs would vary from .04 to .065 within the range $m = 3$ to $m = 50$ (and would be approximately .025 and .022 for $m = 2$ and $m = 100$, respectively). Thus, a constant-risk-ratio procedure, as required under the Bayesian formulation, would not be too different from non-Bayesian procedures in weak control of the FWE within a wide range of values of m.

On the other hand, the risk ratio varies enormously as a function of t_∞, giving about a

2:1 ratio for the highest to the lowest value within the range $m = 3$ to $m = 50$. Duncan had suggested using a risk ratio of 100, which corresponds to $t_\infty = 1.72$; this gives approximately the usual .05 level for comparing two means (more exactly, $t_\infty = 1.69$ to the nearest two decimals gives the closest approximation to .05 for two means). However, for more than two means up to 50, a risk ratio of about 200 corresponds approximately to .05-level Type I error control when all means are equal.

4.2 Average power, FWE, and FDR

Figure 1 shows estimated average power results for the four procedures, at levels for which the average power of $RANGE$ is approximately .50. Although the power of the other three procedures is somewhat above that of $RANGE$ even for 3 means, their power increases rapidly relative to the power of $RANGE$ as the number of means increases. The power of SEP is higher than that of all others except at 50 and 100 means, where the power of DUB is greater than that of SEP. The powers of DUB and $FDR1$ are highly similar throughout, with the largest difference being about .03 when the number of means = 100. Both DUB and $FDR1$ are also close to SEP in power throughout the span of means; the largest difference between any pair of these is smaller than .04, while the differences between $RANGE$ and the three other methods are all greater than .24 when there are 100 means. (Note that, for 50 and 100 means, the critical value of t_∞ for DUB approaches 1.93 and 1.78, respectively, as $F \to \infty$, while the critical value of $|t|$ for SEP is 1.96. Thus, DUB is more powerful than SEP for 50 and 100 means for a sufficiently large variance of the population means. The same is true for DUB compared to $FDR1$, since the power of

the latter must be smaller than that of SEP.)

Although the FWE is .05 for all procedures except SEP when the complete null hypothesis holds, the power advantage of DUB and $FDR1$ over $RANGE$ is achieved at the cost of larger FWE when the complete null hypothesis is false. Figure 2 shows the estimated FWEs for the four procedures. Since there are no true mean differences of zero under the conditions of the simulation, FWE in this case equals the probability of one or more directional errors. The results are similar but not identical to those for power. The FWE of $RANGE$ is far below that for the other procedures, never rising above .002. SEP has the highest FWE except for 50 and 100 means, when the FWE of DUB is the highest. DUB and $FDR1$ are very similar from 3 to 20 means, after which they diverge somewhat with DUB having higher FWE levels; the largest difference is approximately .12 when there are 50 means.

Because there are no true mean differences of zero, the FDR is the expected value of the proportion of directional errors among the rejected hypotheses. Given average power .50 for $RANGE$, this proportion is extremely small, below .004 for all procedures over the whole span from 2 to 100 means.

4.3 Risk

Figure 3 shows estimated risk for the four procedures, with differences standardized to have unit variance and setting $k_1 = 1$, at levels for which the average power of $RANGE$ is approximately .50. The $RANGE$ procedure obviously has much higher risk than all the others, which again are very similar. The largest risk difference among the three other

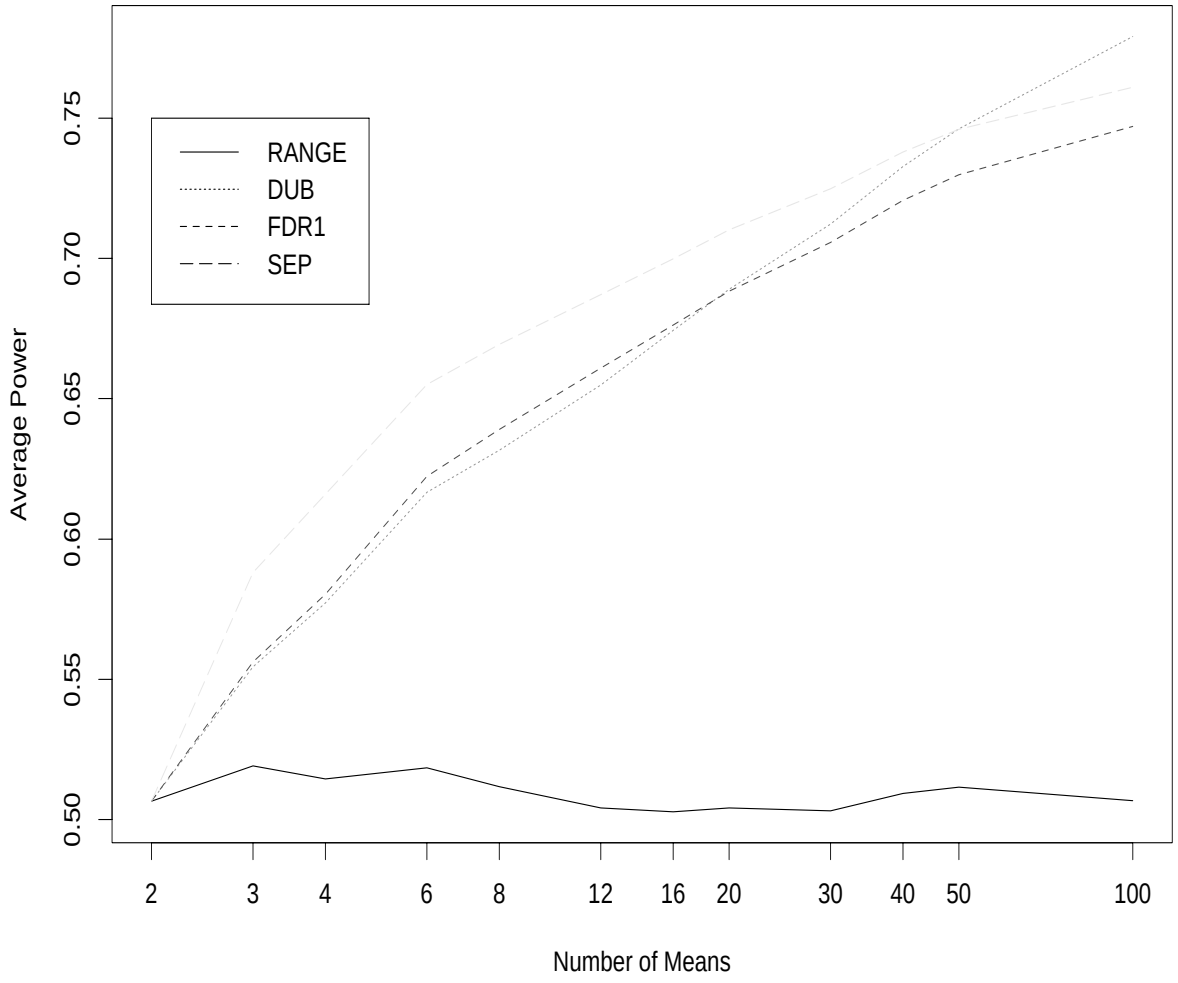

Figure 1: Average Power of RANGE, DUB, FDR1, SEP

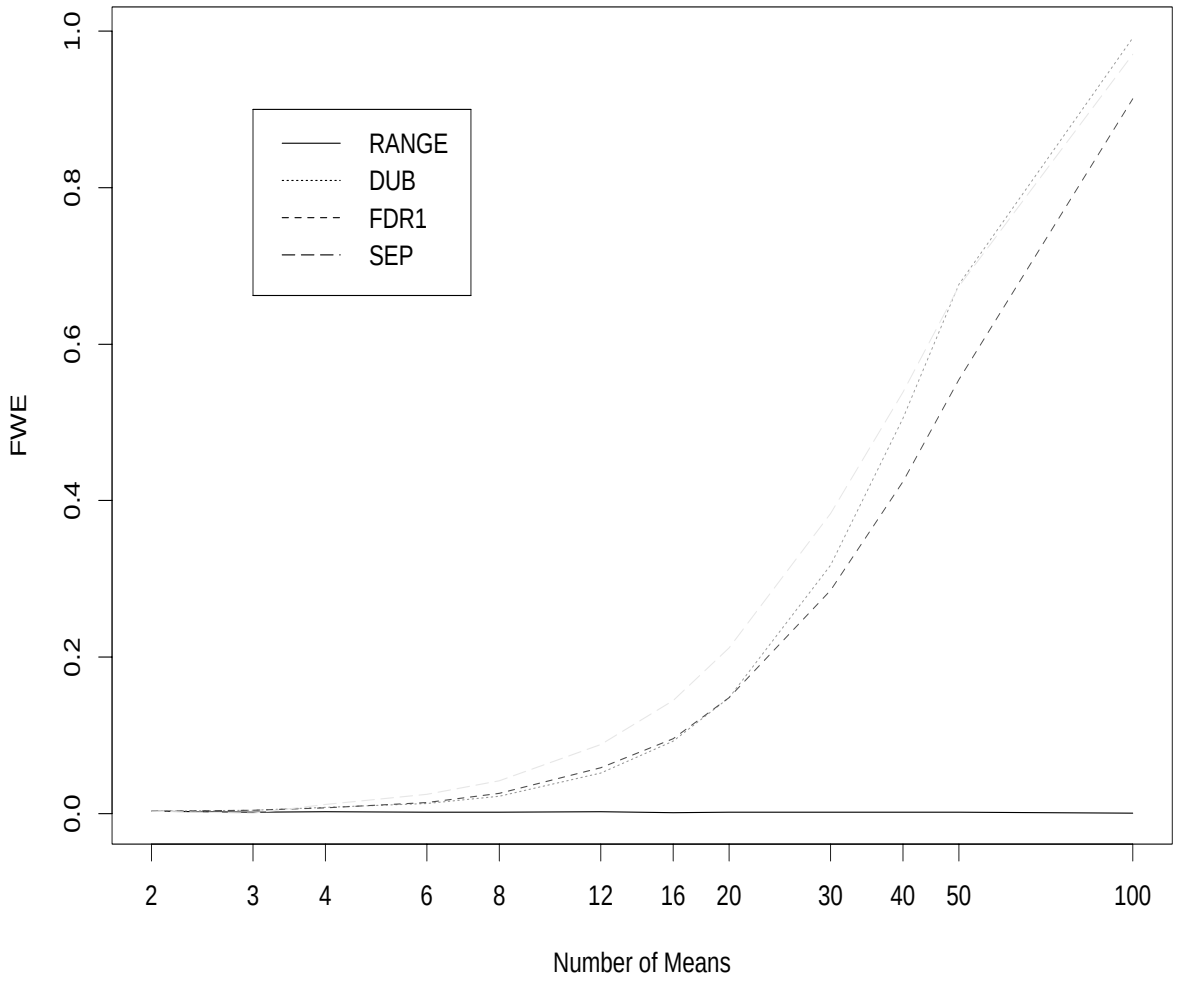

Figure 2: Famlywise Error (FWE) of RANGE, DUB, FDR1, SEP

methods is smaller than .03, while the differences between $RANGE$ and the three other methods are all greater than .70 at 100 means.

The risk is composed of two terms, involving the quantities $NULL$ and DIR (see Equation (2)). Estimates of these quantities are plotted separately for the four procedures in Figures 4 and 5, respectively. As would be expected, $RANGE$ has the largest value of $NULL$, since it is most likely to lead to a (false) failure to reject a hypothesis, while it has the smallest value of DIR, since the probability of directional errors is smaller using the more conservative $RANGE$ than using the other methods. On the other hand, SEP, the least conservative for small numbers of means, has the smallest value of $NULL$ and the largest value of DIR except for 100 means, where DUB has those properties.

4.4 Comparisons with minimum achievable risk

The minimum achievable risk could be attained if the value of Ψ, which might be considered the "true" F-ratio, were known. The procedure which would result in the minimum risk would involve substituting Ψ for F in Equation (4); this procedure is designated MIN. Figure 6 plots this minimum achievable risk and the risk under the three procedures DUB, $FDR1$, and SEP; the risk using $RANGE$ is so much higher than these risks (see Figure (3)) that including $RANGE$ in the figure would make it impossible to see differences among the other procedures. All three (DUB, $FDR1$, and SEP) have risks very close to the minimum risk; with at least three means, the largest difference between any of them and the minimum is less than .04. For 100 means, the difference in risk between $FDR1$ and minimum risk is about .02, between SEP and MIN about .009, and between DUB and

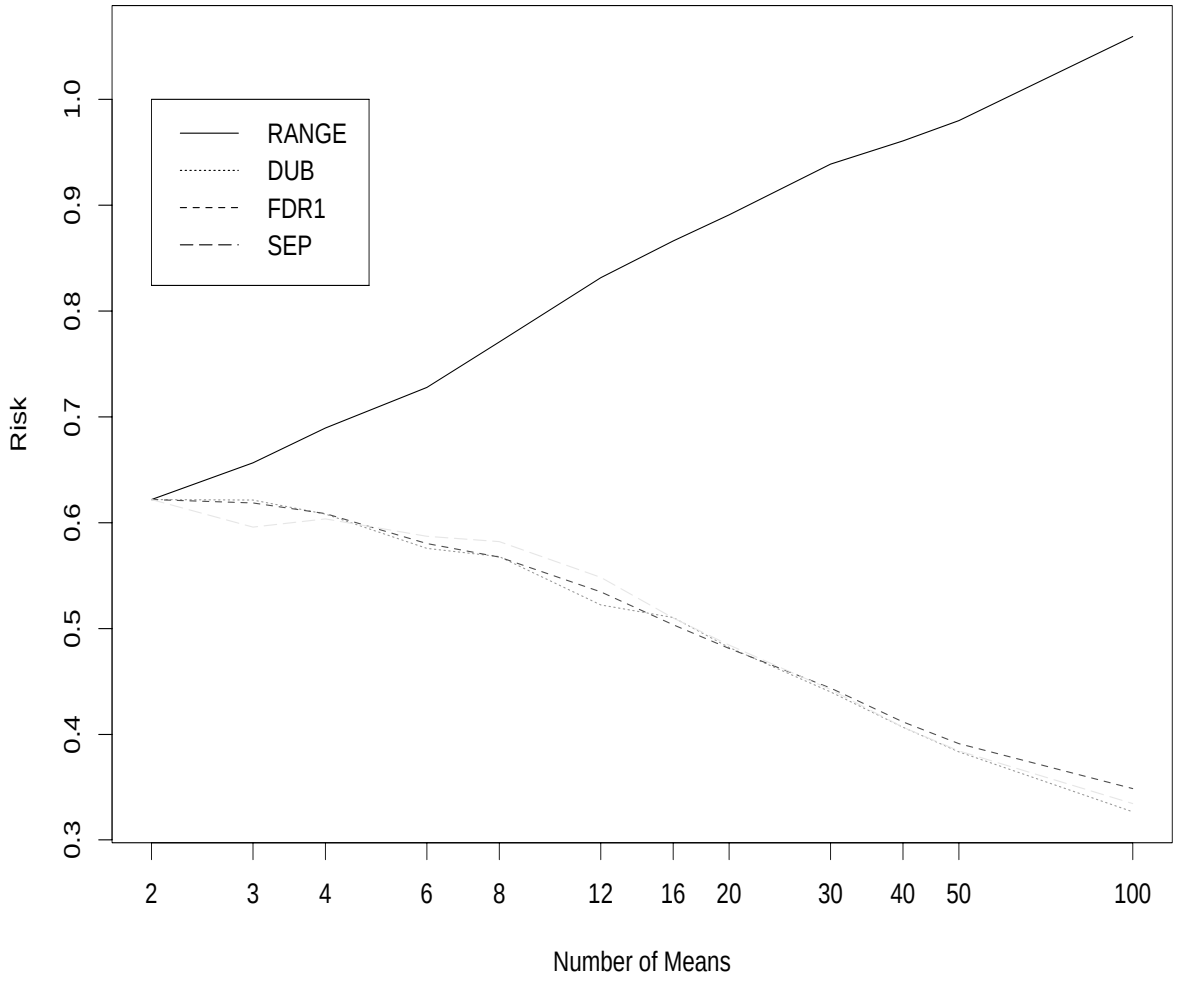

Figure 3: Risk of RANGE, DUB, FDR1, SEP

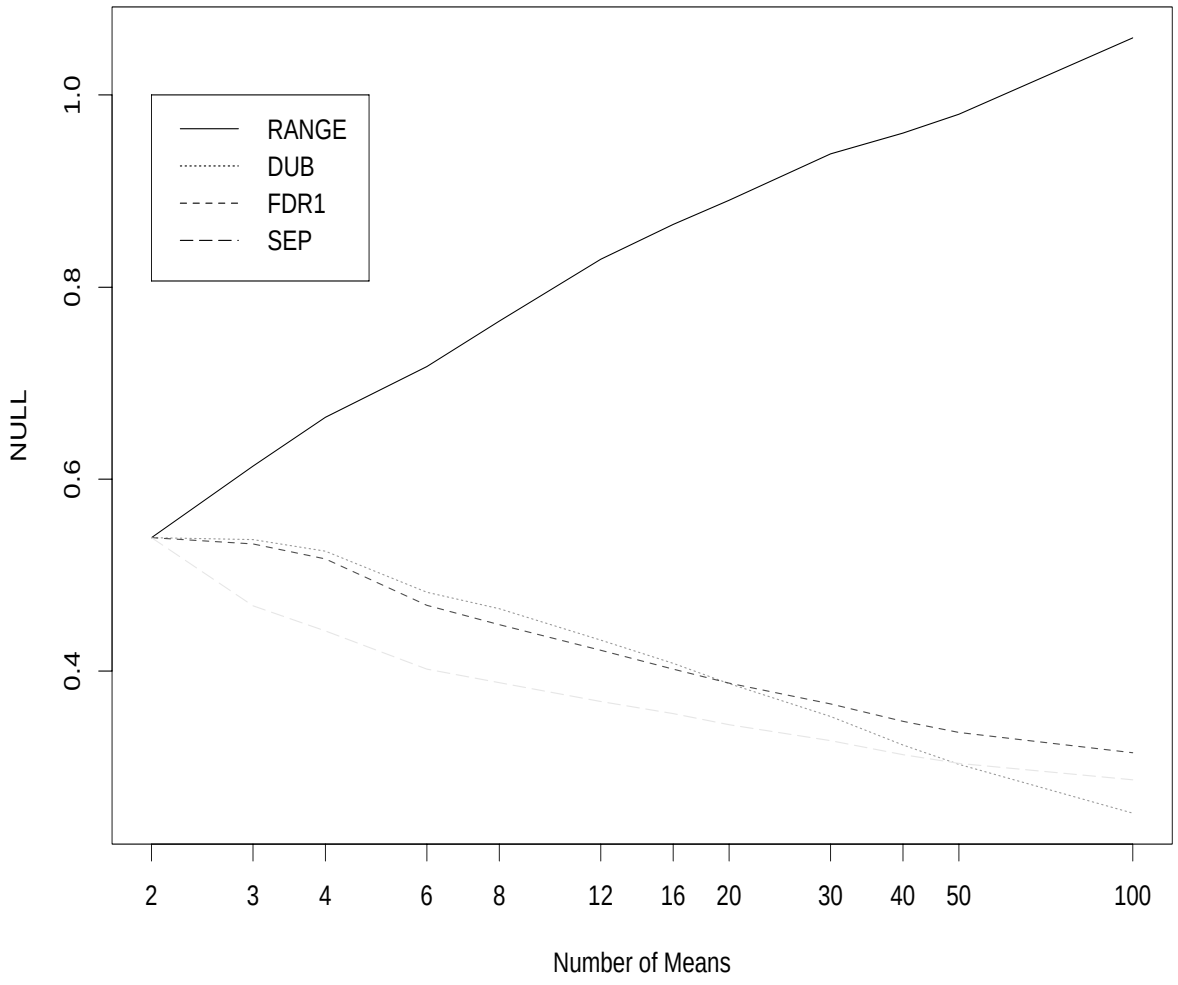

Figure 4: Null Risk ($NULL$) of RANGE, DUB, FDR1, SEP

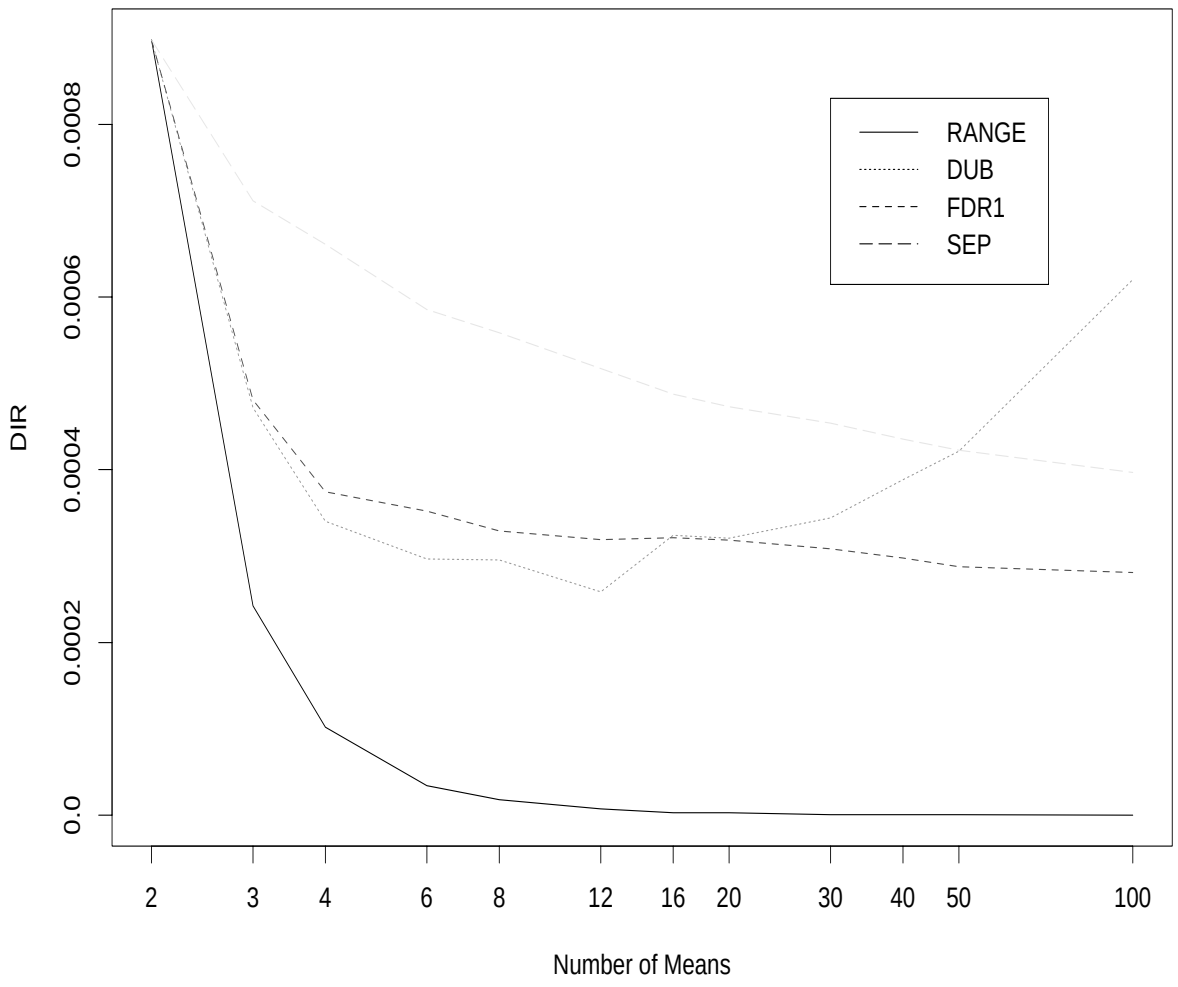

Figure 5: Directional Risk (DIR) of RANGE, DUB, FDR1, SEP

MIN about .0008. Duncan (1965) suggested that there would be little loss in using F to estimate Ψ for 15 or more means; in fact, for 16 means and above, the difference between the risk using DUB and the minimum risk is smaller than .001.

Figures 6 and 7 are plots of $NULL$ and DIR of DUB, $FDR1$, and SEP, along with $NULL$ and DIR, respectively, of MIN. For small numbers of means, SEP has smaller $NULL$ and larger DIR than MIN, while the $FDR1$ method shows the opposite pattern. The values of $NULL$ and DIR of DUB are closest overall to those of MIN.

5 Discussion

The fact that DUB does about as well as the ideal minimum-risk procedure when the number of means is as small as sixteen, although the estimate of Ψ must be rather crude with such a small number, provides support for using the empirical estimate of Ψ, thus avoiding the necessity for postulating a prior distribution for that quantity. Furthermore, the similarity of the risk values for DUB, $FDR1$, and SEP suggest that a number of procedures with generally different properties can achieve close to minimum risk.

A surprising result is the great similarity in both risk and power between DUB and $FDR1$, which also share weak control of the familywise error. Although SEP has somewhat better power than DUB and $FDR1$, the difference is relatively small when the average power of the range is about .50, and the disadvantage of SEP is that it does not have even weak control of FWE; in fact its familywise error rate increases sharply with the number of means, reaching a value greater than .99 under the complete null hypothesis with 50 and

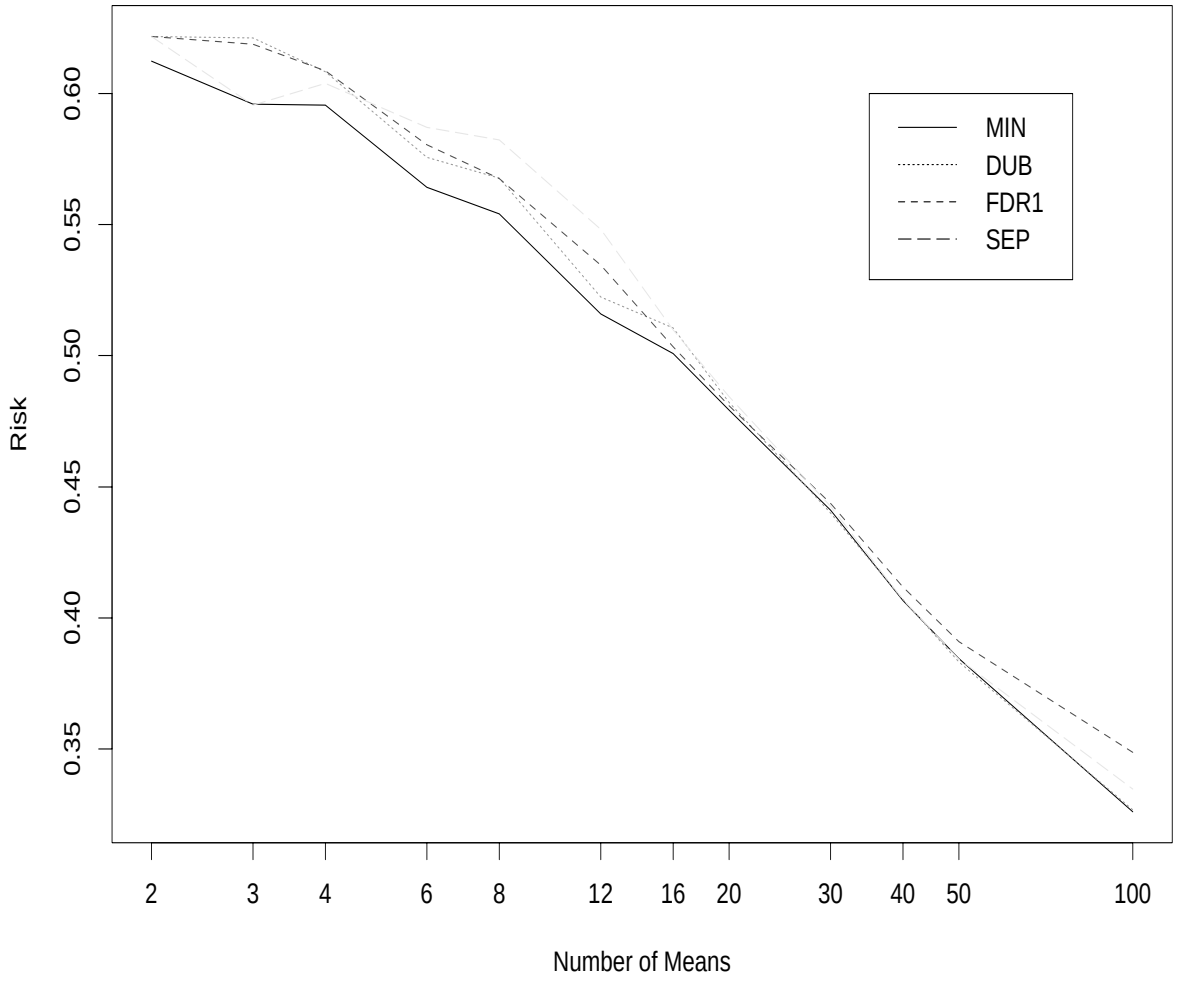

Figure 6: Risk of MIN, DUB, FDR1, SEP

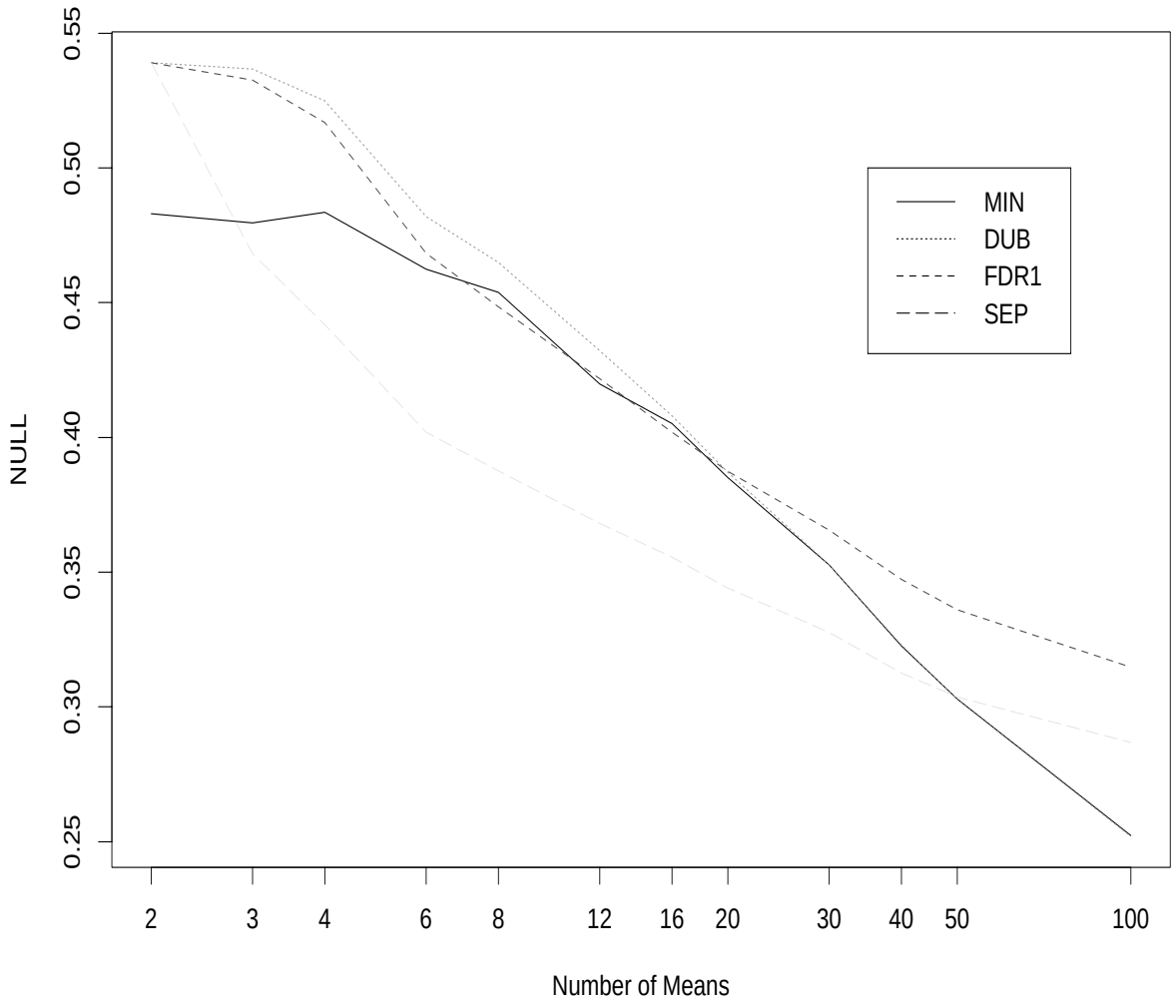

Figure 7: Null Risk ($NULL$) of MIN, DUB, FDR1, SEP

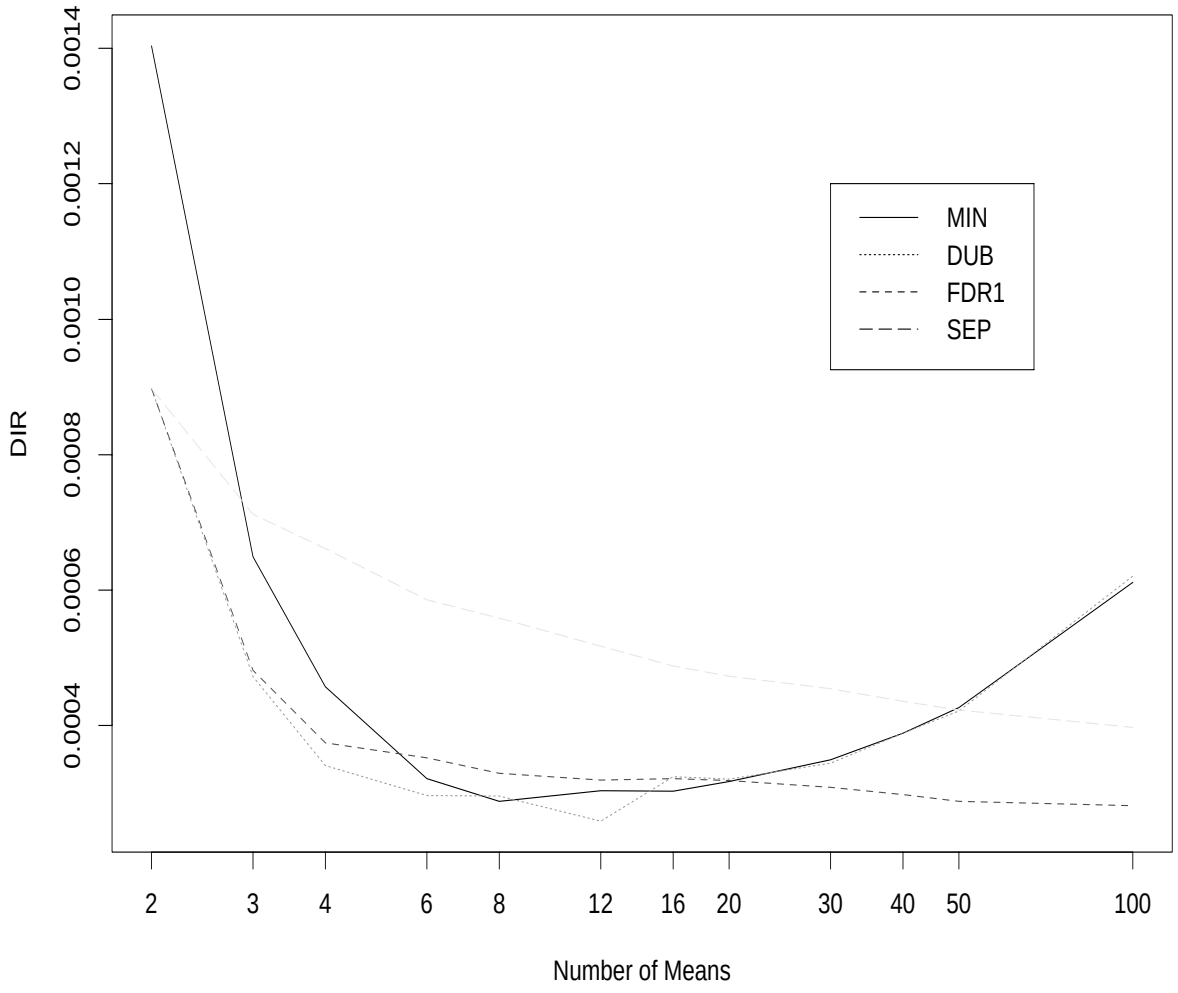

Figure 8: Directional Risk (DIR) of MIN, DUB, FDR1, SEP

100 means.

The $FDR1$, then, is an approximately minimum-risk method with good power properties, and has the advantage over DUB of not depending for weak familywise error control on an assumed normal distribution to generate the true mean values. It is also simple to apply when sample sizes and/or error variances are unequal. The robustness of DUB to moderate departures from normality have yet to be investigated, and it clearly would not be robust against drastic departures. Lewis (1984) refers to Duncan's assumption of a "single-cluster" model, although the normality assumption is more specific than that description implies. In principle, the approach can be generalized to allow for other distributions of the true mean values: See Berry and Hochberg (1998).

Furthermore, although DUB has the simple form of Equation (4) when means have equal variances, the form is considerably more complex when variances are unequal, and thus when sample sizes and/or error variances in the different groups are unequal. Whether simple approximate procedures would also weakly control FWE and be approximately minimum Bayes risk without the equal-variance restrictions remains to be determined.

On the other hand, an argument in favor of DUB is that Duncan has proposed a relatively simple confidence interval method corresponding to it, although the frequentist error properties of such intervals are not completely clear and must be investigated empirically. No simple method for forming confidence intervals corresponding to $FDR1$ has yet been proposed.

6 Conclusion

This study has shown that the Bayesian decision-theoretic method proposed by Duncan, modified to provide weak control of FWE and using an empirical estimate of the variance of the population means, has good properties both from the Bayesian point of view, as a minimum- risk method, and from the frequentist point of view, with good average power. These conclusions hold providing the assumption of a normal distribution of treatment means is satisfied, degrees of freedom for error are large, and the treatment sample means have equal variance around their true values. Although Duncan's loss function formulation is similar to that of Lehmann (1957a, 1957b), it differs in making the loss depend not only on the presence of a Type I or Type II error, but on the magnitude of departure from the null hypothesis, thus on the presumed seriousness of that error for practical purposes. The explicit dependence of the loss function on the magnitude of departures from null hypotheses is often a desirable feature, and is not shared by most frequentist-based multiple comparison procedures. Furthermore, even a completely Bayesian use of the method, keeping the risk ratio of Type II to Type I errors fixed at about 200 as the number of means varies, would approximately control the familywise error in the weak sense when the number of means is between 3 and 50, and thus should have good frequentist properties.

On the other hand, the $FDR1$ procedure, proposed by Benjamini and Hochberg using frequentist principles, has been shown to have approximately minimum Bayes risk, In fact, although very different in derivation and in formulation, both power and risk properties of $FDR1$ and DUB are remarkably similar, and both control FWE in the weak sense, under

the conditions of this study.

These results hold when true means are generated from a normal distribution, when they have equal variance, and when degrees of freedom for estimating variance are large. It remains to be seen whether similar properties hold when these conditions are not met. It is clear that DUB would break down under severe nonnormality. Whether it is robust to moderate nonnormality remains to be determined. The Duncan formulation can be generalized to produce methods for any specified mean and error distributions. It may be possible to extend the applicability of the Duncan approach by developing adaptive methods, based on approximating the distribution of means and errors from the sample. The robustness of the Duncan approach and/or possible modifications will be explored. If successful adaptations can be achieved, some of the advantages in flexibility of $FDR1$ over DUB would be reduced.

Acknowledgements

The author is grateful to John W. Tukey, Charles Lewis, and the referees for helpful suggestions. Much of this work was completed while the author was a Principal Research Scientist at Educational Testing Service, Princeton, NJ, USA.

References

Benjamini, Y., Hochberg, Y. (1995) "Controlling the false discovery rate" Journal of the Royal Statistical Society, 57, 289–300.

Berry, D., Hochberg, Y. (1998) "On Bayesian and quasi-Bayesian approaches to multiple comparison problems" Journal of Statistical Planning and Inference, in press.

Braun, H.I. (Ed.) (1994). *The Collected Works of John W. Tukey.* Vol. VIII: *Multiple Comparisons:1948-1983.* Chapman & Hall, New York.

Duncan, D.B. (1965) "A Bayesian approach to multiple comparisons" Technometrics, 7, 171–222.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple Comparison Procedures.* Wiley, New York.

Lehmann, E.L. (1957a) "A theory of some multiple decision problems, I" Annals of Mathematical Statistics, 28, 1–25..

Lehmann, E.L. (1957b) "A theory of some multiple decision problems, II" Annals of Mathematical Statistics, 28, 547–572.

Lewis, C. (1984) "Multiple comparisons: Fisher revisited" Unpublished manuscript, Univ. of Groningen, Netherlands.

Tukey, J.W. (1953) "The problem of multiple comparisons" Unpublished manuscript, reprinted in Braun, 1994, 1–300.

Waller, R.A., Duncan, D. B. (1969) "A Bayes rule for the symmetric multiple comparisons problem" Journal of the American Statistical Association, 64, 1484–1503.

Williams, V.S.L., Jones, L. V., Tukey, J. W. (1998) "Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement" Journal of Educational and Behavioral Statistics, in press.