Familywise Robustness Criteria Revisited for Newer Multiple Testing Procedures

A Dissertation

Submitted to

the Temple University Graduate Board

in Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

Charles W. Miller

April, 2009

ABSTRACT

Familywise Robustness Criteria Revisited for Newer Multiple Testing Procedures

Charles W. Miller

DOCTOR OF PHILOSOPHY

Temple University, April, 2009

Professor Burt Holland, Chair

As the availability of large datasets becomes more prevalent, so does the need to discover significant findings among a large collection of hypotheses. Multiple testing procedures (MTP) are used to control the familywise error rate (FWER) or the chance to commit at least one type I error when performing multiple hypotheses testing. When controlling the FWER, the power of a MTP to detect significant differences decreases as the number of hypotheses increases. It would be ideal to discover the same false null hypotheses despite the family of hypotheses chosen to be tested. Holland and Cheung (2002) developed measures called familywise robustness criteria (FWR) to study the effect of family size on the acceptance and rejection of a hypothesis. Their analysis focused on procedures that controlled FWER and false discovery rate (FDR). Newer MTPs have since been developed which control the generalized FWER (gFWER(k) or k-FWER) and false discovery proportion (FDP) or tail probabilities for the proportion of false positives (TPPFP). This dissertation reviews these newer procedures and then discusses the effect of family size using the FWRs of Holland and Cheung. In the case where the test statistics are independent and the null

hypotheses are all true, the Type R enlargement familywise robustness measure can be expressed as a ratio of the expected number of Type I errors. In simulations, positive dependence among the test statistics was introduced, the expected number of Type I errors and the Type R enlargement FWR increased for step-up procedures with higher levels of correlation, but not for step-down or single-step procedures.

ACKNOWLEDGEMENTS

I wish to acknowledge the help that I received while working on my Ph.D. When I started my educational journey at Temple, my initial goal was to complete the M.S. degree. During my first year of study, I had received advice from one of my professors, Burt Holland. He suggested that a Ph.D. would be more beneficial to a career in statistics. I want to thank him for telling me that I could do it. As my thesis advisor, he provided me with key support and ideas and showed enormous patience as I wrote my dissertation.

The members of my committee, Sanat Sarkar, Alan Izenman and Dror Rom, have given their time and expertise to better my work. I appreciate all of their helpful feedback and support.

I also wish to thank the professors whose classes I enjoyed and who were instrumental to my growth in the area of statistics: Burt Holland, Francis Hsuan, Boris Iglewicz, Alan Izenman, Milton Parnes, Damaraju Raghavarao, Sanat Sarkar, and William Wei.

Finally, I would like to thank Beth, Sydney and Jarod. Without the love and encouragement of my wife and children, I would have never been able to accomplish writing this dissertation.

To my family: Beth, Sydney and Jarod

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

With computer power and storage expanding at a fevering pace, increasingly large datasets are being made available. Often when using these datasets, one would like to make multiple tests. For instance, when analyzing gene microarrays to find significant genes as they relate to a particular medical condition, there is often some flexibility as to the choice of the family of hypothesis. Depending on the thresholds for deciding which genes to use, the family size could be 1000 or 2000. How does the size of the family affect the results given by a particular multiple testing procedure? A scientist would like a test to be as powerful as possible to detect significant genes. Of secondary importance, one would like a multiple testing procedure to be robust in the choice of a family. This property would make reproducibility more likely. The focus of this dissertation is the study of the robustness of newer multiple testing procedures (MTP) to the choice of the family of hypotheses.

1.1 Family

Lehmann and Romano (2005a) refer to the term "family" as a collection of hypotheses, $H_1, \ldots, H_m$ that is being considered for joint testing. This choice of a family can seem arbitrary at times. A scientist could consider a battery of tests performed in the morning and in the afternoon as belonging to one family since the same substrate was used or as belonging to two different families as determined by the time of day. Neither classification would be wrong, as long as there was a valid reason for the classification scheme. Westfall *et al.* (1999) provide guidelines for selecting a family and suggest that family determinations should be made *a priori*. Despite these guidelines, different experimenters could generate different results using the same MTP depending on how the family is defined.

Dudoit, Shaffer, and Boldrick (2003) state that gene prescreening is a common issue in expression and other large-scale biological experiments. This prescreening is often used to correct signal to background intensity and the proportion of missing values. Depending on the method used to prescreen the microarray data, one could have a dataset that is arbitrarily smaller or larger than someone else.

The number of adverse events studied during a clinical trial can vary greatly depending on the number of investigators and their interpretations of the definitions used to determine whether it is a valid observation of an adverse event. The investigators determine the size of the family as they classify events as adverse or not.

Ahmed (1991) discusses how family size affects the results of comparison studies at the National Center for Education Studies. In a study of the performance of eighth graders in mathematics, suppose that the state by state and one particular state versus rest comparisons were sought. With a total of 40 states participating in the study, a family size of $\binom{40}{2} = 780$ would be used in the state by state comparisons, and when comparing a particular state to the rest, then a family size of 39 would be appropriate. The choice of the family would be appropriate in both cases and yet the results could be drastically different. This type of situation can be found in many different areas. For example in finance, Ettredge *et al.* (2005) investigate the effect of adopting new disclosure rules on the market's ability to predict future earnings. They perform their analysis on all firm sizes, large firms and small firms, where large firms and small firms are a subset of the all firm sizes.

In all of these situations, one would hope that the same significant hypotheses would be discovered no matter which family classification scheme was used. While this goal is not attainable, we can still study the effects of family size on the different multiple testing procedures. To that end, a measure called familywise robustness is reviewed in the next chapter.

CHAPTER 2

History

2.1 Familywise Robustness Criteria

Holland and Cheung (2002) proposed familywise robustness criteria for assessing and comparing multiple comparison procedures. They focused on three Bonferroni style procedures that use p-values calculated from univariate distributions of test statistics. This dissertation uses these robustness criteria to evaluate newer multiple testing procedures that have been developed since 2002.

I begin with some notation. Let $\Omega_n = \{H_1, H_2, \cdots, H_n\}$ be a family of n hypotheses and let $\omega_m = \{H_{i_1}, H_{i_2}, \cdots, H_{i_m}\}$ be a subset of Ω_n, where $n \geq m$, *i.e.,* n is the size of the larger family and m is the size of the smaller family. We consider a particular hypothesis $H_i \in \omega_m \subset \Omega_n$ and study how the result of testing it changes if we either enlarge the family of hypotheses from ω_m to Ω_n, or contract the family from Ω_n to ω_m.

If the testing result (reject or accept) for H_i is the same regardless of whether $H_i \in \omega_m$ or Ω_n, we have made a family size consistent decision. If the testing result differs with the family selected, the decision is family size inconsistent.

Definition 2.1 *A testing procedure that tends to make family size consistent decisions is henceforth referred to as familywise robust.*

The abbreviation FWR is used for both familywise robust and familywise robustness.

The four types of FWR are notated:

Type R enlargement FWR

$$p_{i,rn|rm} = \Pr(H_i \text{ is rejected in } \Omega_n \mid H_i \text{ is rejected in } \omega_m),$$

Type A contraction FWR

$$p_{i,an|am} = \Pr(H_i \text{ is accepted in } \Omega_n \mid H_i \text{ is accepted in } \omega_m),$$

Type A enlargement FWR

$$p_{i,am|an} = \Pr(H_i \text{ is accepted in } \omega_m \mid H_i \text{ is accepted in } \Omega_n),$$

Type R contraction FWR

$$p_{i,rm|rn} = \Pr(H_i \text{ is rejected in } \omega_m \mid H_i \text{ is rejected in } \Omega_n).$$

Definition 2.2 *The testing procedure is said to be perfectly familywise robust when $FWR = 1$.*

The following theorems and definitions from Holland and Cheung (2002) are used in my dissertation and are stated next for convenience.

Theorem 2.1 *If an MTP is perfectly Type A enlargement FWR, then:*

1. *The MTP is perfectly Type R contraction FWR.*
2. $p_{i,rn|rm} = P(H_i$ *is rejected in* $\Omega_n)/P(H_i$ *is rejected in* $\omega_m)$.
3. $p_{i,am|an} = P(H_i$ *is accepted in* $\omega_m)/P(H_i$ *is accepted in* $\Omega_n)$.

Definition 2.3 *Let* $\mathcal{D}$ *be the class of step–down MTPs that satisfies the following two conditions:*

1. *There exists a monotonic increasing sequence of constants* $\{a_{j,n}\}$ *such that the hypotheses corresponding to* $P_{(1)}, P_{(2)}, ..., P_{(j)}$ *are rejected when* $P_{(h)} \leq a_{h,n}$ *for* $h = 1, ..., j$.
2. *For any hypothesis which is in both the smaller and larger families, having rank* i^* *among p-values in* Ω_n *and rank* i *among p-values in* ω_m,

$$a_{i^*,n} - a_{i,m} \leq 0.$$

Definition 2.4 *Let* $\mathcal{U}$ *be the class of step–up MTPs that satisfies the following two conditions:*

1. *There exists a monotonic increasing sequence of constants* $\{a_{j,n}\}$ *such that the hypotheses corresponding to* $P_{(j)}, P_{(j+1)}, \cdots, P_{(n)}$ *are retained when* $P_{(h)} > a_{h,n}$ *for* $h = j, ..., n$.
2. *The monotonic sequence* $\{a_{i,n}\}$ *for testing hypotheses in* Ω_n *and the monotonic sequence* $\{a_{j,m}\}$ *for testing hypotheses in* ω_m *satisfy* $a_{n-i+1,n} - a_{m-i+1,m} \leq 0$ *for all* $i = 1, \ldots, m$.

Theorem 2.2 *All MTPs that are in either* $\mathcal{U}$ *or* $\mathcal{D}$ *as defined above are perfectly Type A enlargement family-wise robust.*

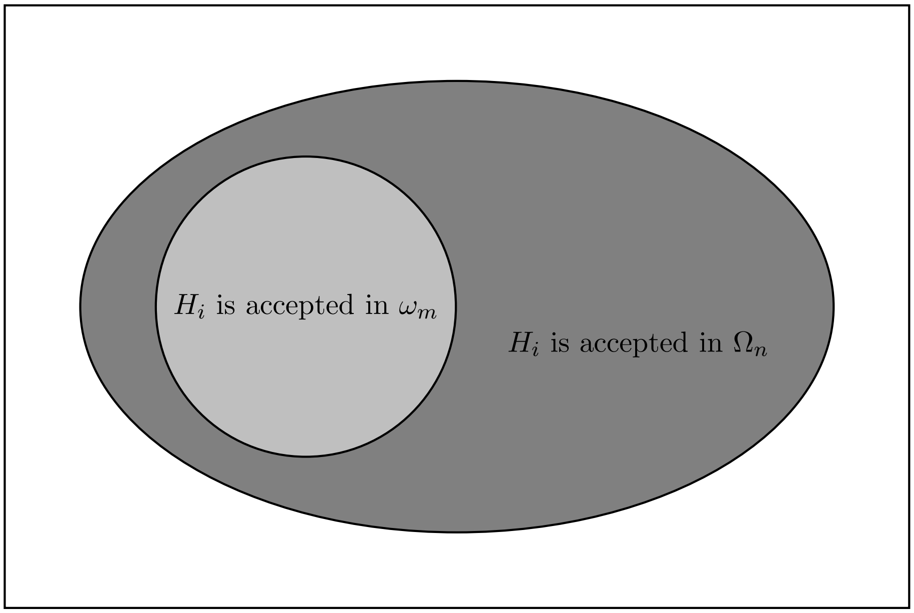

Figure 2.1: Multiple testing procedures that are in $\mathcal{U}$ or $\mathcal{D}$ have acceptance probabilities as depicted in this diagram.

The following measure can be used to assess a procedure's Type R enlargement FWR.

$$P_{rn|rm} = \frac{1}{m}\sum_{i=1}^{m} p_{i,rn|rm},$$

the probability that a hypothesis is rejected in Ω_n given that it is rejected in ω_m, averaged over all m hypotheses in the smaller family ω_m.

In simulations, $p_{i,rn|rm}$ can be estimated using

$$\widehat{p_{i,rn|rm}} = \frac{N_{i,m,n}}{M_{i,m}}$$

and $P_{rn|rm}$ with

$$\widehat{P_{rn|rm}} = \frac{1}{m}\sum_{i=1}^{m}\widehat{p_{i,rn|rm}} = \frac{1}{m}\sum_{i=1}^{m}\frac{N_{i,m,n}}{M_{i,m}}.$$

$N_{i,m,n}$ is the number of times that a hypothesis is rejected in both Ω_n and ω_m. $M_{i,m}$ is the number of times that a hypothesis is rejected in ω_m.

The other FWR measures can be calculated using simulations similarly.

Just as making a Type I error or rejecting a true hypothesis is considered more serious than making a Type II error or failing to reject a false hypothesis, the Type R familywise robustness criteria can be considered more important than Type A familywise robustness criteria. Since a familysize contraction or enlargement seems equally likely to occur, it would appear that the Type R enlargement and Type R contraction FWR are equally important. However, most of the procedures investigated are perfectly Type A enlargement and Type R contraction familywise robust as per Theorem 2.1. Therefore, Type R enlargement FWR is more informative than the Type R contraction FWR and thus can be considered the most important FWR measure of the four.

2.2 Analysis of Original MTPs

In their original paper, Holland and Cheung focused on three Bonferroni style procedures. They were Hochberg (*Hochberg*) (1988), Benjamini and Hochberg (*BH*) (1995), and the adaptive procedure of Benjamini and Hochberg (*adaptive*) (2000). In their analysis, they proved the following two theorems:

Theorem 2.3 *For a true null hypothesis, both Type A enlargement and Type A contraction FWR are bounded below by* $1 - \alpha$ *for Hochberg, and by* $1 - q*$ *for both BH and adaptive.*

Theorem 2.4 *For any procedure in either* $\mathcal{U}$ *or* $\mathcal{D}$ *as well as for BH and adaptive, if* m *is fixed and the test statistics are iid, Type R Enlargement FWR goes to zero as* $n \to \infty$.

Their overall conclusion was that the *BH* procedure was the best of the three analyzed. Although the *adaptive* procedure was better than the *BH* procedure, they chose the *BH* procedure overall since it had been shown to control false discovery rate (FDR) in more situations with dependent test statistics than the *adaptive* procedure.

2.3 Benjamini and Hochberg (BH) Procedure

Briefly, the *BH* procedure as given by Benjamini and Hochberg (1995) is a step up procedure which requires monotonicity of the critical values, $a_{m,m} \geq a_{m-1,m} \geq \ldots \geq a_{1,m}$ and it rejects $H_{(i)}$ for $i \leq i_0$ and accepts $H_{(i)}$ for $i > i_0$, where $i_0 = \max_{1 \leq i \leq m}\{i : p_{(i)} \leq a_{i,m}\}$, where $a_{i,m}$ is given by

$$a_{i,m} = \frac{i}{m}\alpha,$$

where $i = 1, 2, \ldots, m$.

The *BH* procedure controls FDR when the test statistics are independent or have positive dependence. Holland and Cheung proved that the *BH* procedure is not perfectly Type A enlargement robust and gave a counter example. They noted however that it was *almost* perfectly Type A enlargement robust with estimates of $p_{i,an|am} > 0.94$.

CHAPTER 3

Multiple Testing Procedures

Familywise error rate (FWER), the probability of rejecting at least one true hypothesis in the family, has been until recently the main criterion for controlling simultaneous Type I error. But as the size of the family of hypotheses to be tested increases, FWER becomes very restrictive and not very powerful at detecting false null hypotheses. When the family size is large, one might be willing to reject more than one true null hypothesis. Korn *et al.* (2004) and Lehmann and Romano (2005) suggest a new measure, k-FWER, which is an extension of FWER.

Definition 3.1 *Korn et al. (2004) and Lehmann and Romano (2005) define k-FWER as the following: k-FWER is the probability of rejecting at least k true null hypotheses.*

Suppose X is from a model $P \in \Omega$. A general hypothesis H is a subset ω of Ω. For testing $H_i : P \in \omega_i,\ i = 1, \ldots, m$, let $I(P)$ denote the set of true null hypotheses when P is the true probability distribution. Therefore, $i \in I(P)$ if and only if $P \in \omega_i$, and k-FWER which depends on P can be written as

$$k\text{-FWER} = Pr\{\text{reject at least } k \text{ hypotheses } H_i \text{ with } i \in I(P)\}$$

Control of the k-FWER requires that k-FWER $\leq \alpha$ for all P.

$$Pr\{\text{reject at least } k \text{ hypotheses } H_i \text{ with } i \in I(P)\} \leq \alpha \qquad \text{for all } P \tag{3.1}$$

In the following sections, I review procedures that control k-FWER and determine whether they belong to class $\mathcal{D}$ or $\mathcal{U}$ as defined in definition 2.3 and definition 2.4. The first three procedures, Bonferroni, Holm and Hochberg, known to control FWER, are generalized to control k-FWER. Please note that when the procedures, Bonferroni, Holm or Hochberg, are mentioned later in this dissertation, I mean them to be the generalized version.

3.1 Generalized Bonferroni Procedure

Lehmann and Romano (2005) generalized the Bonferroni (1936) method where, p_i is rejected if $p_i \leq \frac{k\alpha}{m}$, and m is the number of hypotheses tested in the family. There are no dependency restrictions on the test statistics associated with this procedure.

Corollary 3.1 *The Bonferroni k-FWER procedure belongs to the $\mathcal{D}$ class.*

Proof: Although, the Bonferroni procedure is not a step procedure, we can treat it as if it belongs to the $\mathcal{D}$ class. Since $n \geq m$, $a_{i^*,n} = \frac{k\alpha}{n} \leq \frac{k\alpha}{m} = a_{i,m}$, for all values of i^* and i. Therefore, $a_{i^*,n} - a_{i,m} \leq 0$. $\square$

3.2 Generalized Holm Procedure

The Holm (1979) procedure as generalized by Lehmann and Romano (2005) is a step down MTP which requires monotonicity of the critical values, $a_{1,m} \leq a_{2,m} \leq \ldots \leq a_{m,m}$ and it rejects $H_{(i)}$ for $i < i_0$ and accepts $H_{(i)}$ for $i \geq i_0$, where $i_0 = \min_{1 \leq i \leq m}\{i : p_{(i)} > a_{i,m}\}$, where $a_{i,m}$ is given by

$$a_{i,m} = \begin{cases} \frac{k}{m}\alpha & \text{if } i \leq k \\ \frac{k}{m+k-i}\alpha & \text{if } i > k \end{cases}$$

This procedure controls k-FWER under all test statistic dependency conditions.

Corollary 3.2 *The Holm k-FWER is a member of the $\mathcal{D}$ class.*

Proof: Since $a_{i,m}$ is defined in two regions, there are three cases to prove that the Holm k-FWER is a member of the class $\mathcal{D}$. I use the relationship $(n-m)-(i^*-i) \geq 0$, the difference in ranks between i^* and i can be at most $n-m$ in the proof.

Case 1: $i^* \leq k$ and $i \leq k$

$$\begin{aligned} n \geq& \quad m \\ \frac{k}{n}\alpha \leq& \quad \frac{k}{m}\alpha \\ \frac{k}{n}\alpha - \frac{k}{m}\alpha \leq& \quad 0 \\ a_{i^*,n} - a_{i,m} \leq& \quad 0 \end{aligned}$$

Case 2: $i^* > k$ and $i \leq k$

$$\begin{aligned}
n - m \geq &\ i^* - i \\
n + k - i^* \geq &\ m + k - i \\
n + k - i^* \geq &\ m \\
\tfrac{1}{n+k-i^*} \leq &\ \tfrac{1}{m} \\
\tfrac{k\alpha}{n+k-i^*} \leq &\ \tfrac{k\alpha}{m} \\
\tfrac{k\alpha}{n+k-i^*} - \tfrac{k\alpha}{m} \leq &\ 0 \\
a_{i^*,n} - a_{i,m} \leq &\ 0
\end{aligned}$$

Case 3: $i^* > k$ and $i > k$

$$\begin{aligned}
n - m \geq &\ i^* - i \\
n + k - i^* \geq &\ m + k - i \\
\tfrac{1}{n+k-i^*} \leq &\ \tfrac{1}{m+k-i} \\
\tfrac{k\alpha}{n+k-i^*} \leq &\ \tfrac{k\alpha}{m+k-i} \\
\tfrac{k\alpha}{n+k-i^*} - \tfrac{k\alpha}{m+k-i} \leq &\ 0 \\
a_{i^*,n} - a_{i,m} \leq &\ 0
\end{aligned}$$

The Holm k-FWER procedure is a member of the $\mathcal{D}$ class. □

3.3 Generalized Hochberg Procedure

The generalized Hochberg procedure by Lehmann and Romano (2005) is a step up procedure which requires monotonicity of the critical values, $a_{m,m} \geq a_{m-1,m} \geq \ldots \geq a_{1,m}$ and it rejects $H_{(i)}$ for $i \leq i_0$ and accepts $H_{(i)}$ for $i > i_0$, where $i_0 = \max_{1 \leq i \leq m}\{i : p_{(i)} \leq a_{i,m}\}$, where $a_{i,m}$ is given by

$$a_{i,m} = \begin{cases} \frac{k}{m}\alpha & \text{if } i \leq k \\ \\ \frac{k}{m+k-i}\alpha & \text{if } i > k \end{cases} \tag{3.2}$$

Just as the Hochberg (1988) procedure controls FWER for certain forms of positively dependent test statistics, the generalized Hochberg procedure controls k-FWER under the same dependency restrictions.

Corollary 3.3 *The Hochberg k-FWER is a member of the $\mathcal{U}$ class.*

Proof: From the definition of the $\mathcal{U}$ class, $a_{n-h+1,n} - a_{m-h+1,m} \leq 0$ for all $h = 1, 2, \ldots, m$. There are three cases to prove that the Hochberg k-FWER is a member of the class $\mathcal{U}$.

Case 1: $(n - h + 1) > k$ and $(m - h + 1) > k$

$$\begin{aligned} \frac{k\alpha}{k+h-1} - \frac{k\alpha}{k+h-1} &\leq 0 \\ \frac{k\alpha}{n+k-(n-h+1)} - \frac{k\alpha}{m+k-(m-h+1)} &\leq 0 \\ a_{n-h+1,n} - a_{m-h+1,m} &\leq 0 \end{aligned}$$

Case 2: $(n - h + 1) > k$ and $(m - h + 1) \leq k$

$$\begin{aligned} m - h + 1 &\leq k \\ k + h - 1 - m &\geq 0 \\ \frac{1}{k+h-1} - \frac{1}{m} &\leq 0 \\ \frac{k\alpha}{n+k-(n-h+1)} - \frac{k\alpha}{m} &\leq 0 \\ a_{n-h+1,n} - a_{m-h+1,m} &\leq 0 \end{aligned}$$

Case 3: $(n - h + 1) \leq k$ and $(m - h + 1) \leq k$

$$\begin{aligned} n &\geq m \\ \tfrac{1}{n} - \tfrac{1}{m} &\leq 0 \\ \tfrac{k\alpha}{n} - \tfrac{k\alpha}{m} &\leq 0 \\ a_{n-h+1,n} - a_{m-h+1,m} &\leq 0 \end{aligned}$$

The Hochberg k-FWER procedure is a member of the $\mathcal{U}$ class. □

3.3.1 Romano and Shaikh

Romano and Shaikh (2006) developed a generalized version of the Hochberg procedure that has no dependency restrictions associated with it. They propose using a stepup procedure with critical values $a'_{i,m} = a_{i,m}/D_1(k,m)$, where $a_{i,m}$ is defined in (3.2) and $D_1(k,m)$ is given as

$$D_1(k,m) = \max_{k \leq |I| \leq m} \left[|I| \frac{a_{m-|I|+k,m}}{k\alpha} + |I| \sum_{k<j\leq|I|} \frac{a_{m-|I|+j,m} - a_{m-|I|-1,m}}{j\alpha} \right] \tag{3.3}$$

Romano and Shaikh show that the critical values $a'_{i,m}$ are approximately half of the critical values $a_{i,m}$ given by (3.2). For example, for $m = 10$ and $k = 3$, the value of $D_1(3,10) = 1.90$. Values of $D_1(k,m)$ for other combinations of m and k are shown in Table 3.1. The value for $D_1(k,m)$ increases as m increases, until m becomes greater than optimizing value of $|I|$ for a given k. Then the value of $D_1(k,m)$ becomes constant. The optimizing value for $|I|$ when $k = 1$ is 17.

Corollary 3.4 *The RS procedure is a member of the $\mathcal{U}$ class.*

Proof: From Corollary 3.3, $a_{n-h+1,n} - a_{m-h+1,m} \leq 0$ for all $h = 1, 2, \ldots, m$.

m	k=1	k=3	k=5	k=15
10	2.11	1.90	1.60	
15	2.13	2.06	1.87	
20	2.13	2.12	2.00	1.28
30	2.13	2.16	2.12	1.63
40	2.13	2.17	2.16	1.83
50	2.13	2.17	2.18	1.95
60	2.13	2.17	2.18	2.03
80	2.13	2.17	2.18	2.12
100	2.13	2.17	2.18	2.16
200	2.13	2.17	2.18	2.19
500	2.13	2.17	2.18	2.19

Table 3.1: Values of $D_1(k,m)$, where $a_{i,m}$ is defined in (3.2).

Since $a'_{i,m} = a_{i,m}/D_1(k,m)$ and $D_1(k,n) \geq D_1(k,m)$, it follows that

$$\begin{aligned} D_1(k,n) &\geq D_1(k,m) \\ \tfrac{a_{n-h+1,n}}{D_1(k,n)} &\leq \tfrac{a_{n-h+1,n}}{D_1(k,m)} \leq \tfrac{a_{m-h+1,m}}{D_1(k,m)} \\ \tfrac{a_{n-h+1,n}}{D_1(k,n)} - \tfrac{a_{m-h+1,m}}{D_1(k,m)} &\leq 0 \\ a'_{n-h+1,n} - a'_{m-h+1,m} &\leq 0 \end{aligned}$$

The RS procedure is a member of the $\mathcal{U}$ class. □

3.3.2 Sarkar

Another generalized Hochberg procedure is given by Sarkar (2008). Assuming that the p-values have identical kth-order joint null distributions, the common cumulative distribution function of the maximum of any k of the p-values, G_k, can be used to derive critical values, $a_{i,m}$ and it rejects $H_{(i)}$ for $i \leq i_0$ and accepts $H_{(i)}$ for $i > i_0$, where $i_0 = \max_{1 \leq i \leq m}\{i : p_{(i)} \leq a_{max(i,k),m}\}$, where $a_{i,m}$ is given by

$$G_k(a_{i,m}) = \frac{k(k-1)\cdots 1}{(m-i+k)(m-i+k-1)\cdots(m-i+1)}\alpha,$$

with $i = k, \dots, m$.

The Sarkar procedure controls k-FWER strongly when the joint null distribution of the p-values is multivariate totally positive of order two (MTP_2). This procedure has been found to be a more powerful generalized Hochberg procedure than the one proposed by Lehmann and Romano (2005) when $2 \leq k \leq 1/\alpha$.

When the p-values are independent, $a_{i,m}$ is given by

$$a_{i,m} = \begin{cases} \left(\alpha \prod_{j=1}^{k} \frac{j}{m-k+j}\right)^{1/k} & \text{if } i \leq k \\ \left(\alpha \prod_{j=1}^{k} \frac{j}{m-i+j}\right)^{1/k} & \text{if } i > k \end{cases}$$

Corollary 3.5 *The Sarkar k-FWER procedure is a member of the $\mathcal{U}$ class.*

Proof: From the definition of the $\mathcal{U}$ class, $a_{n-h+1,n} - a_{m-h+1,m} \leq 0$ for all $h = 1, 2, \dots, m$. Since $a_{i,m}$ is defined in two regions, there are three cases to prove that the Sarkar k-FWER is a member of the class $\mathcal{U}$.

Case 1: $(n-h+1) > k$ and $(m-h+1) > k$

$$\begin{aligned} \left(\alpha \prod_{j=1}^{k} \frac{j}{h-1+j}\right) &\leq \left(\alpha \prod_{j=1}^{k} \frac{j}{h-1+j}\right) \\ \left(\alpha \prod_{j=1}^{k} \frac{j}{n-(n-h+1)+j}\right) &\leq \left(\alpha \prod_{j=1}^{k} \frac{j}{m-(m-h+1)+j}\right) \\ G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{n-(n-h+1)+j}\right) &\leq G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{m-(m-h+1)+j}\right) \\ a_{n-h+1,n} - a_{m-h+1,m} &\leq 0 \end{aligned}$$

Case 2: $(n-h+1) > k$ and $(m-h+1) \leq k$

$$\begin{aligned}
h &\geq m-k+1 \\
h-1+j &\geq m-k+1-1+j \\
\prod_{j=1}^{k}(h-1+j) &\geq \prod_{j=1}^{k}(m-k+j) \\
\prod_{j=1}^{k} \frac{j}{h-1+j} &\leq \prod_{j=1}^{k} \frac{j}{m-k+j} \\
G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{n-(n-h+1)+j}\right) &\leq G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{m-k+j}\right) \\
a_{n-h+1,n} - a_{m-h+1,m} &\leq 0
\end{aligned}$$

Case 3: $(n-h+1) \leq k$ and $(m-h+1) \leq k$

$$\begin{aligned}
n &> m \\
n-k+j &> m-k+j \\
\prod_{j=1}^{k} \frac{j}{n-k+j} &\leq \prod_{j=1}^{k} \frac{j}{m-k+j} \\
G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{n-k+j}\right) &\leq G_k^{-1}\left(\alpha \prod_{j=1}^{k} \frac{j}{m-k+j}\right) \\
a_{n-h+1,n} - a_{m-h+1,m} &\leq 0
\end{aligned}$$

The Sarkar k-FWER procedure is a member of the $\mathcal{U}$ class. $\square$

3.4 Stepdown False Discovery Proportion (FDP) Procedure

All of the MTPs introduced so far are based on the familywise criterion k-FWER. Korn *et al.* (2004) define the false discovery proportion (FDP) as the number of true hypotheses rejected divided by the total number of hypotheses rejected. It is defined to be zero if there are no rejections. For a given γ and α in $(0,1)$, they required $Pr\{FDP > \gamma\} \leq \alpha$.

The BH procedure introduced earlier is based on the familywise criterion of FDR, which is the expectation of FDP.

Lehmann and Romano (2005) proposed a stepdown method based on the false discovery proportion. Their procedure (StepDownFDP) is a step down MTP which requires monotonicity of critical constants, $a'_{1,m} \leq a'_{2,m} \leq \ldots \leq a'_{m,m}$. The critical constants are $a'_{i,m} = a_{i,m}/C_{\lfloor \gamma m \rfloor +1}$, where $a_{i,m}$ is given by

$$a_{i,m} = \frac{\lfloor \gamma i \rfloor + 1}{m + \lfloor \gamma i \rfloor + 1 - i}\alpha,$$

C_j by

$$C_j = \sum_{i=1}^{j} \frac{1}{i},$$

and $\lfloor x \rfloor$ is the greatest integer $\leq x$. This stepdown procedure is valid regardless of the dependency of the test statistics.

Remark 3.1 *The StepDownFDP procedure is not a member of the $\mathcal{D}$ class.*

I illustrate with a simple numerical example. Let $m = 50$, $n = 60$, $\alpha = 0.05$, and $\gamma = 0.05$. Now suppose that $i^* = i + n - m$ and $i = 1$ then

$$a'_{11,60} = \frac{\frac{4}{(60+4-11)}0.05}{2.083} = 0.001811 \geq 0.001573 = \frac{\frac{3}{(50+3-1)}0.05}{1.833} = a'_{1,50}.$$

Clearly, in this case, $a'_{i^*,n} - a'_{i,m} \nleq 0$ and hence the StepDownFDP procedure is not a member of the $\mathcal{D}$ class.

CHAPTER 4

Familywise Robustness of Newer MTPs

The FWR measures for the generalized versions of the Bonferroni, Holm and Hochberg procedures as well as the Sarkar, StepDownFDP and BH procedures are examined in this chapter, including theoretical results. The next chapter contains simulation results.

4.1 Type A Enlargement Familywise Robustness

The chapter begins with the Type A enlargement familywise robustness since the properties as set by Theorem 2.1 are used in the rest of this chapter. Although the newer MTPs are generalized multiple testing procedures which depend on the value of k, Theorem 2.2 can still be used to determine which of the newer MTP's are perfectly Type A enlargement familywise robust. All of the newer procedures that were found to be in class $\mathcal{U}$ or $\mathcal{D}$ are perfectly Type A enlargement familywise robust.

Corollary 4.1 *The generalized versions of Bonferroni, Holm, and Hochberg procedures as well as the RS and Sarkar procedure are all perfectly Type A enlargement familywise robust.*

The StepDownFDP and BH procedures are not a member of class $\mathcal{U}$ or $\mathcal{D}$ so they are not perfectly Type A enlargement familywise robust. Based on simulations shown in Chapter 5, the $\widehat{P_{an|am}}$ for both the StepDownFDP and BH procedures was always above 0.95 and generally above 0.999.

4.2 Type R Contraction Familywise Robustness

As set by Theorem 2.1, MTPs that are perfectly Type A enlargement familywise robust are also Type R contraction familywise robust.

Corollary 4.2 *The generalized versions of Bonferroni, Holm, and Hochberg procedures as well as the RS and Sarkar procedure are all perfectly Type R contraction familywise robust.*

StepDownFDP and BH are not perfectly Type R contraction FWR. In simulations, the estimates of $P_{rm|rn}$ were found to be above 0.94.

4.3 Type R Enlargement Familywise Robustness

4.3.1 Complete Null Hypothesis or Weak Control

In this section, we assume that all of the null hypotheses are true. Dudoit and van der Laan (2007) call this Complete Null Hypothesis. Lehmann and Romano (2005a) refer to this condition as weak control. We also assume that the tests are independent. The hypotheses, $\{H_1, H_2, \ldots, H_n\}$ have corresponding p-values, $\{p_1, p_2, \ldots, p_n\}$, where $p_i \sim$ iid U(0,1) for all $i = 1, 2, \ldots, n$.

Theorem 4.1 *Given a complete null hypothesis, a MTP that is perfectly Type A enlargement familywise robust has the probability of rejecting a hypothesis in the set Ω_n given it was rejected in the set ω_m or the Type R enlargement familywise robustness measure as follows:*

$$P_{i,rn|rm} = \lambda \frac{EV_n}{EV_m} \tag{4.1}$$

where $\lambda = m/n \in [0,1]$, $V_n = \sum_{i=1}^{n} I(H_i \text{ is rejected in } \Omega_n)$, and $V_m = \sum_{i=1}^{n} I(H_i \text{ is rejected in } \omega_m)$.

Proof: If a MTP is perfectly Type A enlargement familywise robust, then

$$P_{i,rn|rm} = Pr(H_i \text{ is rejected in } \Omega_n)/Pr(H_i \text{ is rejected in } \omega_m)$$

as per Theorem 2.1.

Assuming all of the null hypotheses are true and all of the p-values are independent, the probability of rejecting any hypothesis, H_i, is the same. If we define $X_i = I(H_i \text{ is rejected in } \Omega_n)$ and $V_n = \sum_{i=1}^{n} X_i$. Then $V_n \sim Bin(n,\ Pr(H_i \text{ is rejected in } \Omega_n))$ and $EV_n = n * Pr(H_i \text{ is rejected in } \Omega_n)$. Similarly, $V_m \sim Bin(m,\ Pr(H_i \text{ is rejected in } \omega_m))$ and $EV_m = m * Pr(H_i \text{ is rejected in } \omega_m)$. Therefore,

$$\begin{aligned} P_{i,rn|rm} &= \frac{Pr(H_i \text{ is rejected in } \Omega_n)}{Pr(H_i \text{ is rejected in } \omega_m)} \\ &= \frac{\frac{1}{n} EV_n}{\frac{1}{m} EV_m} \\ &= \lambda \frac{EV_n}{EV_m} \qquad \square. \end{aligned}$$

Finner and Roters (2002) give results for the expected Type I errors given independent p-values as follows:

For Single Step procedures,

$$EV_n = n\gamma_{n,n}. \tag{4.2}$$

For Step Down procedures,

$$EV_n = \sum_{i=1}^{n} F_n^{n-i}(\gamma_{n,n}, \ldots, \gamma_{n-i+1,n}). \tag{4.3}$$

For Step Up procedures,

$$EV_n = \sum_{i=0}^{n-1}(1 - F_n^i(1-\gamma_{1,n}, \ldots, 1-\gamma_{n-i,n})). \tag{4.4}$$

Where $\gamma_{i,n},\ i = 1, \ldots, n$, is a sequence of critical constants such that $1 \geq \gamma_{1,n} \geq \cdots \geq \gamma_{n,n} \geq 0$, and the joint cdf F_n^k is given by

$$F_n^k(1-\gamma_{1,n}, \ldots, 1-\gamma_{n-k,n}) = 1 - \sum_{i=0}^{n-k-1} \binom{n}{i} F_i(1-\gamma_{1,n}, \ldots, 1-\gamma_{i,n})\gamma_{i+1,n}^{n-i} \tag{4.5}$$

with $F_n^0 = F_n$ and $F_0^0 \equiv F_n^n \equiv 1$.

The expected number of Type I errors for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using equations (4.2), (4.3), and (4.4) are shown in Figure 4.1 using values of $k = \{1, 3, 5, 15\}$. In the plots in which $k \geq 3$, the Sarkar procedure has the highest value of EV_n and the StepDownFDP procedure has the lowest. In fact the expected number of Type I errors when $k = 15$ for the Sarkar procedure is at least five times higher than the other procedures.

Remark 4.1 *Finner and Roters (2002) suggest that a large expected number of Type I errors may be an indicator for good power performance. Since, the Sarkar procedure yields a larger expected number of Type I errors especially as k becomes larger, one might expect this procedure to be more powerful than other k-FWER procedures when the test statistics are independent.*

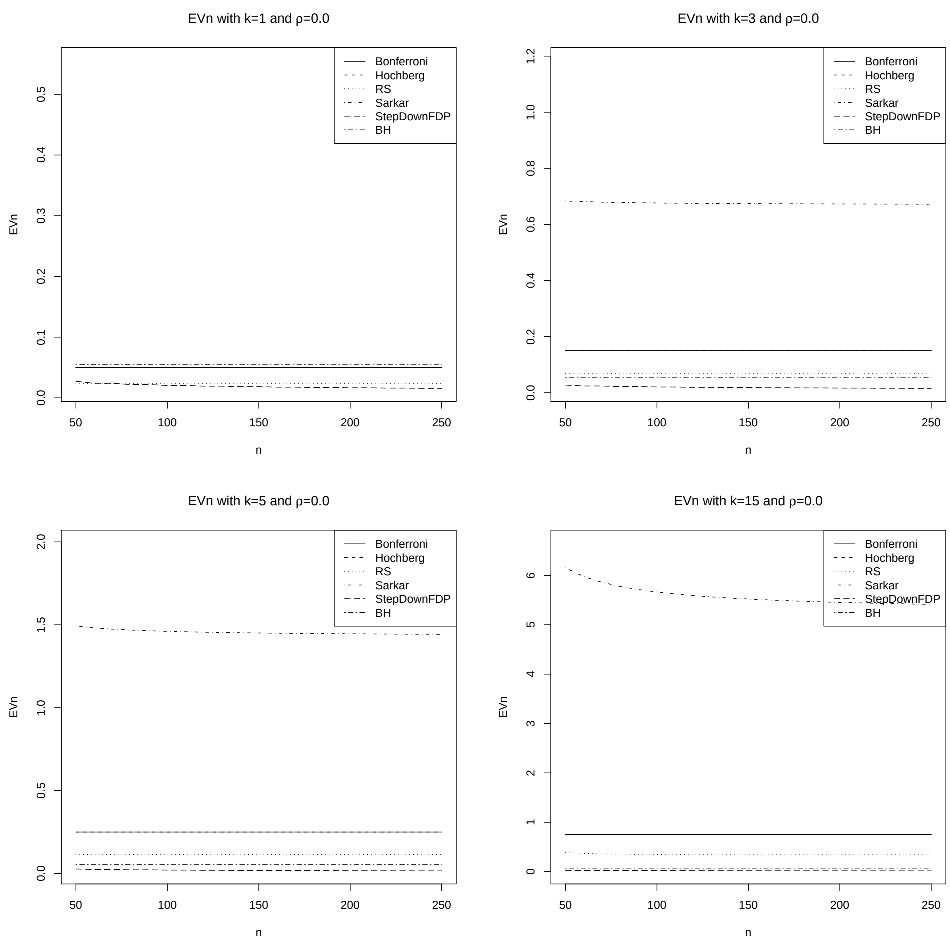

Figure 4.1: The expected number of Type I errors for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures.

Finner and Roters (2002) give another formula for the expected value of type I errors for the BH procedure as

$$EV_n = \alpha \sum_{i=0}^{n-1} \binom{n-1}{i} (i+1)!(\alpha/n)^i, \tag{4.6}$$

which is nondecreasing in n. By using monotone convergence, Finner and Roters prove that the limit of EV_n as $n \to \infty$ for the BH procedure is

$$\lim_{n\to\infty} EV_n = \frac{\alpha}{(1-\alpha)^2}. \tag{4.7}$$

They also give the limit of EV_n for the non-generalized forms of the Bonferroni, Holm and Hochberg procedures which have been generalized in the following lemma.

Lemma 4.1 *The limit of EV_n as $n \to \infty$ for the Bonferroni, Holm and Hochberg procedures is given by*

$$\lim_{n\to\infty} EV_n = k\alpha. \tag{4.8}$$

For the Romano and Shaikh procedure, the limit of EV_n is given by

$$\lim_{n\to\infty} EV_n = k\alpha/D_1(k,n), \tag{4.9}$$

and for the Sarkar procedure is given by

$$\lim_{n\to\infty} EV_n = (k!\alpha)^{\frac{1}{k}}. \tag{4.10}$$

Proof: From Finner and Roters (2002), we find that $\lim_{n\to\infty} n\gamma_{n-i,n} = c$ for all $i \in \mathbb{N}_0$ implies that $\lim_{n\to\infty} EV_n = c$. For the Bonferroni, Holm and Hochberg procedures, the $\lim_{n\to\infty} n\gamma_{n-i,n} = k\alpha$, which implies (4.8). For the RS procedure, $\lim_{n\to\infty} n\gamma_{n-i,n} = k\alpha/D_1(k,n)$, which implies (4.9), and for the Sarkar procedure, $\lim_{n\to\infty} n\gamma_{n-i,n} = (k!\alpha)^{\frac{1}{k}}$, which implies (4.10). $\square$

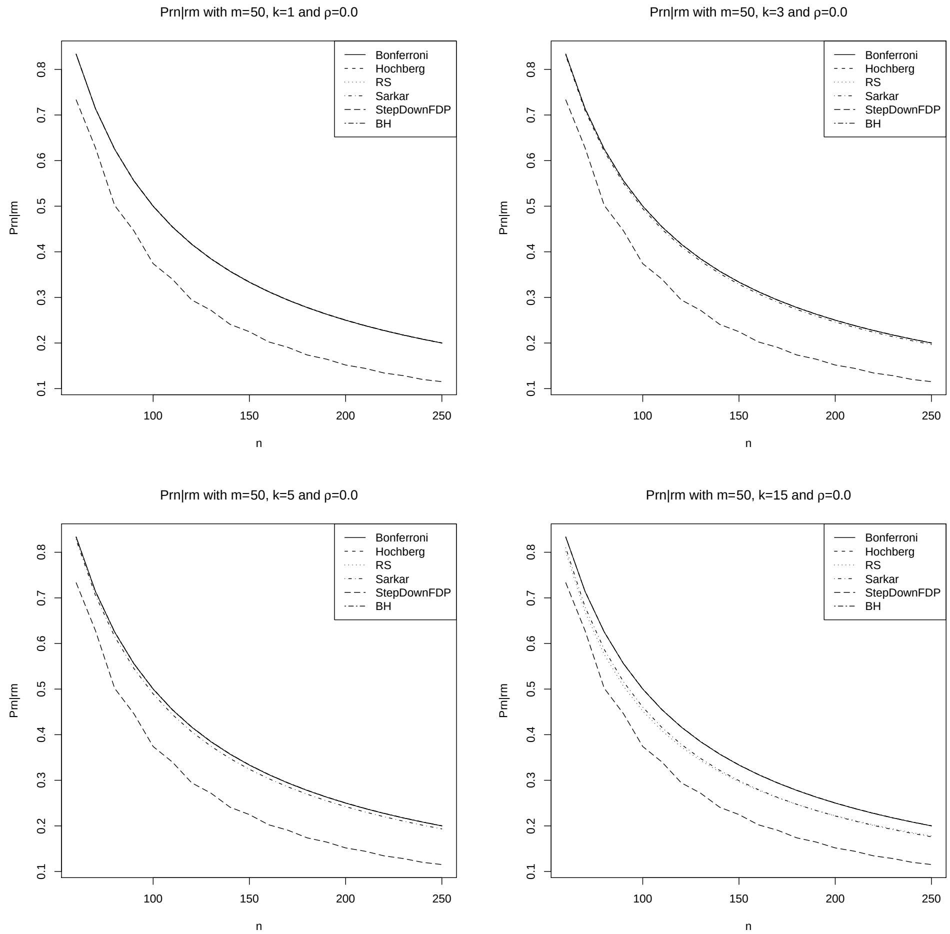

Figure 4.2: The Type R enlargement familywise robustness is plotted for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using Equation 4.1. It assumes that all of the null hypotheses are true.

Using equations (4.1), (4.2), (4.3) and (4.4), the Type R enlargement familywise robustness is plotted for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP and BH procedures and shown in Figure 4.2. From each of the plots in Figure 4.2, the StepDownFDP procedure appears to have the lowest Type R enlargement FWR amongst the procedures considered.

Theorem 4.2 *The Type R enlargement FWR for the StepDownFDP procedure is uniformly worse amongst all other procedures considered when* $n \geq m + max\{\lceil \frac{1}{\gamma} \rceil, 3\}$, $0 < \gamma < 1$, *and* $k < \lfloor \frac{m}{2} \rfloor$ *where* $\lceil x \rceil$ *is the smallest integer* $\geq x$ *and* $\lfloor x \rfloor$ *is the greatest integer* $\leq x$.

Proof: For all of the procedures considered except the BH procedure, a look at (4.3) and (4.4) shows that all of the terms of the sum appearing on the right side of the formula for EV_n are nonincreasing in n. As a result, the EV_n for the Holm, Hochberg, RS, Sarkar and StepDownFDP procedures are nonincreasing in n. The EV_n for the BH procedure was shown earlier to be nondecreasing in n.

I will use the RS procedure in this proof, since this procedure has the next lowest values of EV_n as indicated by Figure 4.1. But any of the remaining procedures can be used.

The smallest ratio EV_n/EV_m for the RS procedure occurs where EV_m has the largest value. This occurs when $m = 2$. The value of EV_n decreases as n increases until it reaches it limiting value of $k * \alpha/D_1(k, n)$. So, the numerator is the smallest when $n \to \infty$ and $k = 1$.

$$\begin{aligned}
\frac{EV^{SDFDP}_{2+max\{\lceil\frac{1}{\gamma}\rceil,3\}}}{EV^{SDFDP}_2} &< \frac{EV^{RS}_\infty}{EV^{RS}_2} = 0.6804 \\
\frac{EV^{SDFDP}_{m+max\{\lceil\frac{1}{\gamma}\rceil,3\}}}{EV^{SDFDP}_m} &< \frac{EV^{RS}_\infty}{EV^{RS}_m} \\
\frac{EV^{SDFDP}_n}{EV^{SDFDP}_m} &< \frac{EV^{RS}_n}{EV^{RS}_m} \\
\lambda\frac{EV^{SDFDP}_n}{EV^{SDFDP}_m} &< \lambda\frac{EV^{RS}_n}{EV^{RS}_m} \\
P^{SDFDP}_{i,rn|rm} &< P^{RS}_{i,rn|rm}
\end{aligned}$$

It follows that the Type R enlargement for the StepDownFDP procedure is always less than the Type R enlargement for any of the other procedures considered. □

Remark 4.2 *The StepdownFDP procedure is valid under all test statistics dependency conditions, so it is not too surprising that it doesn't perform as well as procedures that have dependency restrictions.*

From each of the plots in Figure 4.2, it is not clear that the BH procedure has the best type R enlargement FWR amongst the procedures considered. However, Tables 4.1 – 4.4 do show that the

n	Bonferroni	Hochberg	RS	Sarkar	StepDownFDP	BH
60	0.8333333	0.8331888	0.8332656	0.8331888	0.7332265	0.8336475
70	0.7142857	0.7140739	0.7141865	0.7140739	0.6284451	0.7147476
80	0.6250000	0.6247572	0.6248862	0.6247572	0.5016912	0.6255306
90	0.5555556	0.5553002	0.5554359	0.5553002	0.4459325	0.5561147
100	0.5000000	0.4997418	0.4998790	0.4997418	0.3740219	0.5005662
110	0.4545455	0.4542896	0.4544256	0.4542896	0.3400146	0.4551071
120	0.4166667	0.4164161	0.4165492	0.4164161	0.2944995	0.4172173
130	0.3846154	0.3843715	0.3845011	0.3843715	0.2718574	0.3851517
140	0.3571429	0.3569065	0.3570321	0.3569065	0.2408110	0.3576631
150	0.3333333	0.3331046	0.3332261	0.3331046	0.2247550	0.3338369
160	0.3125000	0.3122790	0.3123964	0.3122790	0.2024299	0.3129869
170	0.2941176	0.2939042	0.2940176	0.2939042	0.1905211	0.2945882
180	0.2777778	0.2775716	0.2776811	0.2775716	0.1737917	0.2782325
190	0.2631579	0.2629587	0.2630645	0.2629587	0.1646436	0.2635974
200	0.2500000	0.2498074	0.2499097	0.2498074	0.1517031	0.2504250
210	0.2380952	0.2379089	0.2380079	0.2379089	0.1444773	0.2385065
220	0.2272727	0.2270924	0.2271882	0.2270924	0.1342060	0.2276709
230	0.2173913	0.2172167	0.2173094	0.2172167	0.1283705	0.2177770
240	0.2083333	0.2081641	0.2082540	0.2081641	0.1200455	0.2087073
250	0.2000000	0.1998358	0.1999230	0.1998358	0.1152434	0.2003628

Table 4.1: The Type R enlargement familywise robustness for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using Equation 4.1 when $m = 50$ and $k = 1$

BH procedure does have the best type R enlargement FWR for the case when $m = 50$.

Theorem 4.3 *The Type R enlargement FWR for the BH procedure is uniformly better than other procedures considered.*

Proof: The BH procedure is the only one in which the ratio, (EV_n/EV_m), is greater or equal to 1 and increases as n increases. The Bonferroni procedure has a ratio, (EV_n/EV_m), which stays constant at a value of 1. All other procedures have a ratio, (EV_n/EV_m), less than or equal to 1 and the ratio decreases as n increases.

$$\begin{aligned} \frac{EV_n^{Others}}{EV_m^{Others}} &\leq 1 \leq \frac{EV_n^{BH}}{EV_m^{BH}} \\ \lambda\frac{EV_n^{Others}}{EV_m^{Others}} &\leq \lambda\frac{EV_n^{BH}}{EV_m^{BH}} \\ P_{i,rn|rm}^{Others} &\leq P_{i,rn|rm}^{BH} \end{aligned}$$

n	Bonferroni	Hochberg	RS	Sarkar	StepDownFDP	BH
60	0.8333333	0.8333320	0.8333332	0.8303577	0.7332265	0.8336475
70	0.7142857	0.7142837	0.7142855	0.7099290	0.6284451	0.7147476
80	0.6250000	0.6249977	0.6249998	0.6200100	0.5016912	0.6255306
90	0.5555556	0.5555531	0.5555553	0.5503096	0.4459325	0.5561147
100	0.5000000	0.4999975	0.4999997	0.4946972	0.3740219	0.5005662
110	0.4545455	0.4545430	0.4545452	0.4492937	0.3400146	0.4551071
120	0.4166667	0.4166642	0.4166664	0.4115240	0.2944995	0.4172173
130	0.3846154	0.3846130	0.3846151	0.3796122	0.2718574	0.3851517
140	0.3571429	0.3571406	0.3571426	0.3522936	0.2408110	0.3576631
150	0.3333333	0.3333311	0.3333331	0.3286431	0.2247550	0.3338369
160	0.3125000	0.3124979	0.3124998	0.3079683	0.2024299	0.3129869
170	0.2941176	0.2941156	0.2941174	0.2897408	0.1905211	0.2945882
180	0.2777778	0.2777758	0.2777776	0.2735505	0.1737917	0.2782325
190	0.2631579	0.2631560	0.2631577	0.2590738	0.1646436	0.2635974
200	0.2500000	0.2499981	0.2499998	0.2460523	0.1517031	0.2504250
210	0.2380952	0.2380934	0.2380951	0.2342772	0.1444773	0.2385065
220	0.2272727	0.2272710	0.2272725	0.2235777	0.1342060	0.2276709
230	0.2173913	0.2173896	0.2173911	0.2138128	0.1283705	0.2177770
240	0.2083333	0.2083317	0.2083332	0.2048651	0.1200455	0.2087073
250	0.2000000	0.1999984	0.1999998	0.1966363	0.1152434	0.2003628

Table 4.2: The Type R enlargement familywise robustness for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using Equation 4.1 when $m = 50$ and $k = 3$

n	Bonferroni	Hochberg	RS	Sarkar	StepDownFDP	BH
60	0.8333333	0.8333333	0.8320344	0.8273997	0.7332265	0.8336475
70	0.7142857	0.7142857	0.7131724	0.7056236	0.6284451	0.7147476
80	0.6250000	0.6250000	0.6240258	0.6151002	0.5016912	0.6255306
90	0.5555556	0.5555555	0.5546896	0.5451652	0.4459325	0.5561147
100	0.5000000	0.5000000	0.4992206	0.4895112	0.3740219	0.5005662
110	0.4545455	0.4545454	0.4538370	0.4441686	0.3400146	0.4551071
120	0.4166667	0.4166666	0.4160172	0.4065146	0.2944995	0.4172173
130	0.3846154	0.3846154	0.3840159	0.3747462	0.2718574	0.3851517
140	0.3571429	0.3571428	0.3565862	0.3475834	0.2408110	0.3576631
150	0.3333333	0.3333333	0.3328138	0.3240924	0.2247550	0.3338369
160	0.3125000	0.3125000	0.3120129	0.3035758	0.2024299	0.3129869
170	0.2941176	0.2941176	0.2936592	0.2855022	0.1905211	0.2945882
180	0.2777778	0.2777778	0.2773448	0.2694598	0.1737917	0.2782325
190	0.2631579	0.2631579	0.2627477	0.2551244	0.1646436	0.2635974
200	0.2500000	0.2500000	0.2496103	0.2422373	0.1517031	0.2504250
210	0.2380952	0.2380952	0.2377241	0.2305895	0.1444773	0.2385065
220	0.2272727	0.2272727	0.2269185	0.2200105	0.1342060	0.2276709
230	0.2173913	0.2173913	0.2170525	0.2103597	0.1283705	0.2177770
240	0.2083333	0.2083333	0.2080086	0.2015200	0.1200455	0.2087073
250	0.2000000	0.2000000	0.1996883	0.1933932	0.1152434	0.2003628

Table 4.3: The Type R enlargement familywise robustness for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using Equation 4.1 when $m = 50$ and $k = 5$

n	Bonferroni	Hochberg	RS	Sarkar	StepDownFDP	BH
60	0.8333333	0.8333333	0.8010673	0.8096798	0.7332265	0.8336475
70	0.7142857	0.7142857	0.6694156	0.6804564	0.6284451	0.7147476
80	0.6250000	0.6250000	0.5759196	0.5868819	0.5016912	0.6255306
90	0.5555556	0.5555556	0.5060501	0.5159657	0.4459325	0.5561147
100	0.5000000	0.5000000	0.4518176	0.4603573	0.3740219	0.5005662
110	0.4545455	0.4545455	0.4084734	0.4155788	0.3400146	0.4551071
120	0.4166667	0.4166667	0.3729911	0.3787450	0.2944995	0.4172173
130	0.3846154	0.3846154	0.3434016	0.3479127	0.2718574	0.3851517
140	0.3571429	0.3571429	0.3183340	0.3217249	0.2408110	0.3576631
150	0.3333333	0.3333333	0.2967999	0.2992052	0.2247550	0.3338369
160	0.3125000	0.3125000	0.2780976	0.2796330	0.2024299	0.3129869
170	0.2941176	0.2941176	0.2616912	0.2624651	0.1905211	0.2945882
180	0.2777778	0.2777778	0.2471528	0.2472838	0.1737917	0.2782325
190	0.2631579	0.2631579	0.2341448	0.2337632	0.1646436	0.2635974
200	0.2500000	0.2500000	0.2224375	0.2216449	0.1517031	0.2504250
210	0.2380952	0.2380952	0.2118453	0.2107213	0.1444773	0.2385065
220	0.2272727	0.2272727	0.2022159	0.2008241	0.1342060	0.2276709
230	0.2173913	0.2173913	0.1934239	0.1918150	0.1283705	0.2177770
240	0.2083333	0.2083333	0.1853646	0.1835797	0.1200455	0.2087073
250	0.2000000	0.2000000	0.1779500	0.1760225	0.1152434	0.2003628

Table 4.4: The Type R enlargement familywise robustness for the Bonferroni, Hochberg, RS, Sarkar, StepDownFDP, and BH procedures using Equation 4.1 when $m = 50$ and $k = 15$

The Type R enlargement FWR for the BH procedure is always greater than or equal to all of the other procedures considered. □

Next, the Type R enlargement familywise robustness measure will be examined under strong control.

4.3.2 Strong Control

In the previous section, analysis of the Type R enlargement FWR was performed under weak control. Using terms from Definition 3.1, weak control equates to $|I(P)| = m$, where cardinality of $I(P)$ is equal to the number of hypotheses tested in the family which is m. Therefore, strong control is one in which $|I(P)| \leq m$. Assuming that the test statistics are independent and identically distributed, the hypotheses, $\{H_1, H_2, \ldots, H_m\}$ have corresponding p-values, $\{p_1, p_2, \ldots, p_m\}$, and now these p-values must satisfy the following:

$$Pr\{p_i \leq u\} = u \text{ for any } u \in (0,1) \text{ and any } P \in \omega_i.$$

Using Theorem 2.4, the following is generated.

Corollary 4.3 *For the generalized versions of Bonferroni, Holm, and Hochberg procedures as well as the Sarkar and StepDownFDP procedure, if m is fixed and the test statistics are iid, Type R Enlargement FWR goes to zero as $n \to \infty$.*

As stated above, the Type R enlargement FWR for all of the procedures discussed in this dissertation go to zero as n approaches infinity, but the theorem doesn't address the rate of decline for each of the procedures. For that, the simulations presented in the next chapter will be used.

4.4 Type A Contraction Familywise Robustness

Given a complete null hypothesis, a MTP that is perfectly Type A enlargement familywise robust has the probability of rejecting a hypothesis in the set ω_m given it was rejected in the set Ω_n or the Type A contraction familywise robustness measure as follows:

$$P_{i,am|an} = \frac{1 - \frac{1}{m}EV_m}{1 - \frac{1}{n}EV_n} \tag{4.11}$$

Proof: Substituting for $Pr(H_i$ is rejected in $\omega_m)$ and $Pr(H_i$ is rejected in $\Omega_n)$ below,

$$\begin{aligned}
\text{Type A contraction FWR} &= Pr(H_i \text{ is accepted in } \omega_m | H_i \text{ is accepted in } \Omega_n) \\
&= Pr(H_i \text{ is accepted in } \omega_m)/Pr(H_i \text{ is accepted in } \Omega_n) \\
&= 1 - Pr(H_i \text{ is rejected in } \omega_m)/1 - Pr(H_i \text{ is rejected in } \Omega_n) \\
&= (1 - \tfrac{1}{m}EV_m)/(1 - \tfrac{1}{n}EV_n) \qquad \square
\end{aligned}$$

Using Equation 4.11, plots were generated in which m was 50, n varied from 60 to 250 and k was allowed values of 1, 3, 5, and 15. These plots are shown in Figure 4.3. One can see that the BH and StepDownFDP procedures perform the best and the Sarkar procedure the worst.

Theorem 4.4 *When all of the hypotheses are true, the Type A contraction FWR for MTPs that are in class $\mathcal{U}$ or $\mathcal{D}$ are bounded below by $1 - a_{m,m}$.*

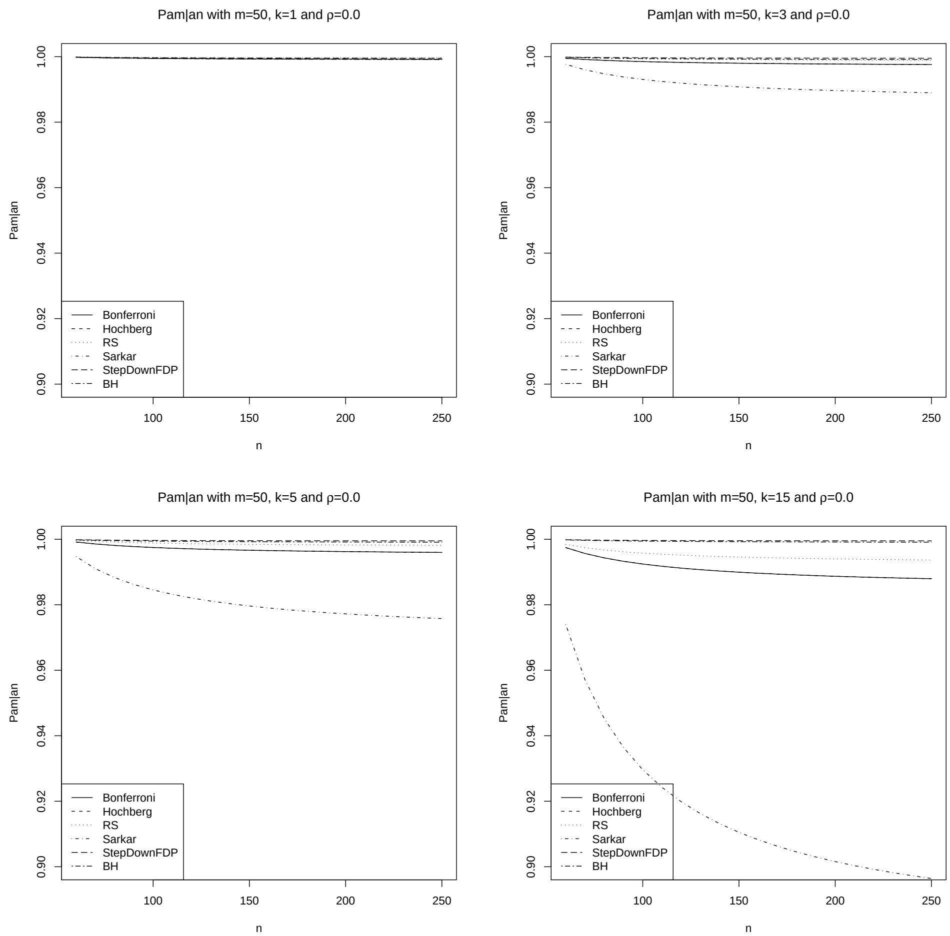

Figure 4.3: The Type A contraction familywise robustness was plotted for the Bonferroni, Hochberg, Sarkar, StepDownFDP, and BH procedures using Equation 4.11. It assumes that all of the null hypotheses are true.

Proof:

$$
\begin{aligned}
\text{Type A contraction FWR} &= Pr(H_i \text{ is accepted in } \omega_m | H_i \text{ is accepted in } \Omega_n) \\
&= Pr(H_i \text{ is accepted in } \omega_m)/Pr(H_i \text{ is accepted in } \Omega_n) \\
&\geq Pr(H_i \text{ is accepted in } \omega_m) \\
&\geq Pr(p_i > a_{i,m}) \\
&\geq Pr(p_i > a_{m,m}) \\
&\geq 1 - Pr(p_i < a_{m,m}) \\
&\geq 1 - a_{m,m} \qquad \square.
\end{aligned}
$$

Corollary 4.4 *The Type A contraction FWR for the generalized Bonferroni procedure is bounded below by* $1 - \frac{k\alpha}{m}$*. The Type A contraction FWR for the Holm and Hochberg procedures is bounded below by* $1-\alpha$ *and* $1-\alpha^{1/k}$ *is the lower bound for Type A contraction FWR for the Sarkar procedure.*

CHAPTER 5

Simulation

In the previous chapter, theoretical results were derived in the case when all of the null hypotheses were true and the test statistics were independent. In the first section of this chapter, we are interested in what happens to the familywise robustness measures when the test statistics are allowed to be positive dependent. Focus is limited to Type R enlargement and Type A contraction FWR since the other two measures, Type R contraction and Type A enlargement FWR are always close to 1.

In the second section of this chapter, we look at what effect strong control of familywise error rate has on the familywise robustness measures, Type R enlargement and Type A contraction FWR. Simulations were performed in which half of the null hypotheses were false and the test statistics were independent or positively dependent.

5.1 Complete Null Hypotheses or Weak Control of Familywise Error Rate

In this section, we still assumed that all of the null hypotheses are true, but we relaxed the assumption of independent test statistics. We looked at the positive dependence situation using two simulations, equicorrelated dependence and clumpy dependence. The first simulation was motivated by the computation of step-up critical constants by Dunnett and Tamhane (1992). It allows one to easily and quickly simulate dependent statistics. One-sided hypotheses testing was performed with null hypotheses $H_i: \ \mu_i = 0$ against the alternative hypotheses $H_i': \ \mu_i > 0$, for $i = 1, 2, \ldots, n$. $Z_i \sim \mathcal{N}(0,1)$ for $i = \{0, 1, 2, \ldots, n\}$, $U = \sqrt{\mathcal{X}_\nu^2/\nu}$, and $T_i = (\sqrt{1-\rho}Z_i - \sqrt{\rho}Z_0)/U$. Thus, the random variables, $T_1, T_2, \ldots, T_n$ have a Student's n-variate central t distribution with ν degrees of freedom and associated common correlation coefficient ρ.

The second simulation conducted was based on comments by Storey (2003). He hypothesized that the dependence structure in DNA microarrays is a "clumpy dependence". A "clumpy dependence" structure has dependence within a group, and no dependence between groups. To simulate clumpy dependence, one-sided hypotheses testing was performed with null hypotheses $H_i: \ \mu_i = 0$ against the alternative hypotheses $H_i': \ \mu_i > 0$, for $i = 1, 2, \ldots, n$. The test statistics were equally divided into ten groups and calculated using $T_i = (\sqrt{1-\rho}Z_i - \sqrt{\rho}Z_{0i})/U$, where $Z_i \sim \mathcal{N}(0,1)$ for $i = \{1, 2, \ldots, n\}$, $U = \sqrt{\mathcal{X}_\nu^2/\nu}$, and $Z_{0i} \sim \mathcal{N}(0,1)$ for $i = \{1, 2, \ldots, 10\}$. Thus, the random variables, $T_1, T_2, \ldots, T_n$ have a Student's n-variate central t distribution with ν degrees of freedom and associated common correlation coefficient ρ within a group and stochastically independent between groups.

The results from the 20,000 simulations were used to calculate the expected number of Type I errors, Type R enlargement FWR and Type A contraction FWR for the Bonferroni, Holm, Hochberg, RS, Sarkar, StepDownFDP and BH procedures. Simulations were performed with $k = \{1, 3, 5, 10, 15\}$, $m = \{50\}$, $n = m + 10, m + 30, \ldots, m + 190$, and $\rho = \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. For brevity, only the figures with $k = \{1, 5, 15\}$ were shown.

5.1.1 Expected Type I Errors

In Theorem 4.1, the Type R enlargement FWR was expressed as the ratio of two expected Type I errors (EV_n/EV_m). Theorem 4.1 and Equations 4.2, 4.3, and 4.4 which were used to calculate of the expected number of Type I errors required that the test statistics were independent. If the expected number of Type I errors for a procedure does not change when the test statistics are positively dependent, then the equations for the calculation of expected number of Type I errors might still be valid which implies that the Type R enlargement FWR can be calculated using Equation 4.1.

The expected number of Type I errors for each procedure was estimated using

$$\widehat{EV_n} = \frac{\text{total number of rejections per procedure}}{\text{number of simulations}}.$$

Each figure contains six plots in which the correlation coefficient varies from $\rho = 0.0$ to 0.9. The plot in the upper left corner of each figure corresponds to test statistics that are independent ($\rho = 0.0$). Each of these plots closely resembles the theoretical results produced in Figure 4.1.

As the correlation coefficient, ρ, is increased, the expected number of Type I errors appears to remain constant for the Bonferroni, Holm, and StepDownFDP procedures in both dependent structure cases. The expected number of Type I errors for all of the step up procedures increases with larger values of correlation in the equicorrelated dependence case. The BH procedure appeared to be the most sensitive to correlation. We start to see an increase in the expected number of Type I errors for the BH procedure when $\rho = 0.3$. The increase in the number of Type I errors is not seen in the other step up procedures until correlation increased above 0.5. In the clumpy dependence case, the increase in the expected number of Type I errors is only seen with the BH procedure. The effect with the BH procedure is still seen beginning with a $\rho = 0.3$, however the magnitude of the effect is much smaller than with the equicorrelated case.

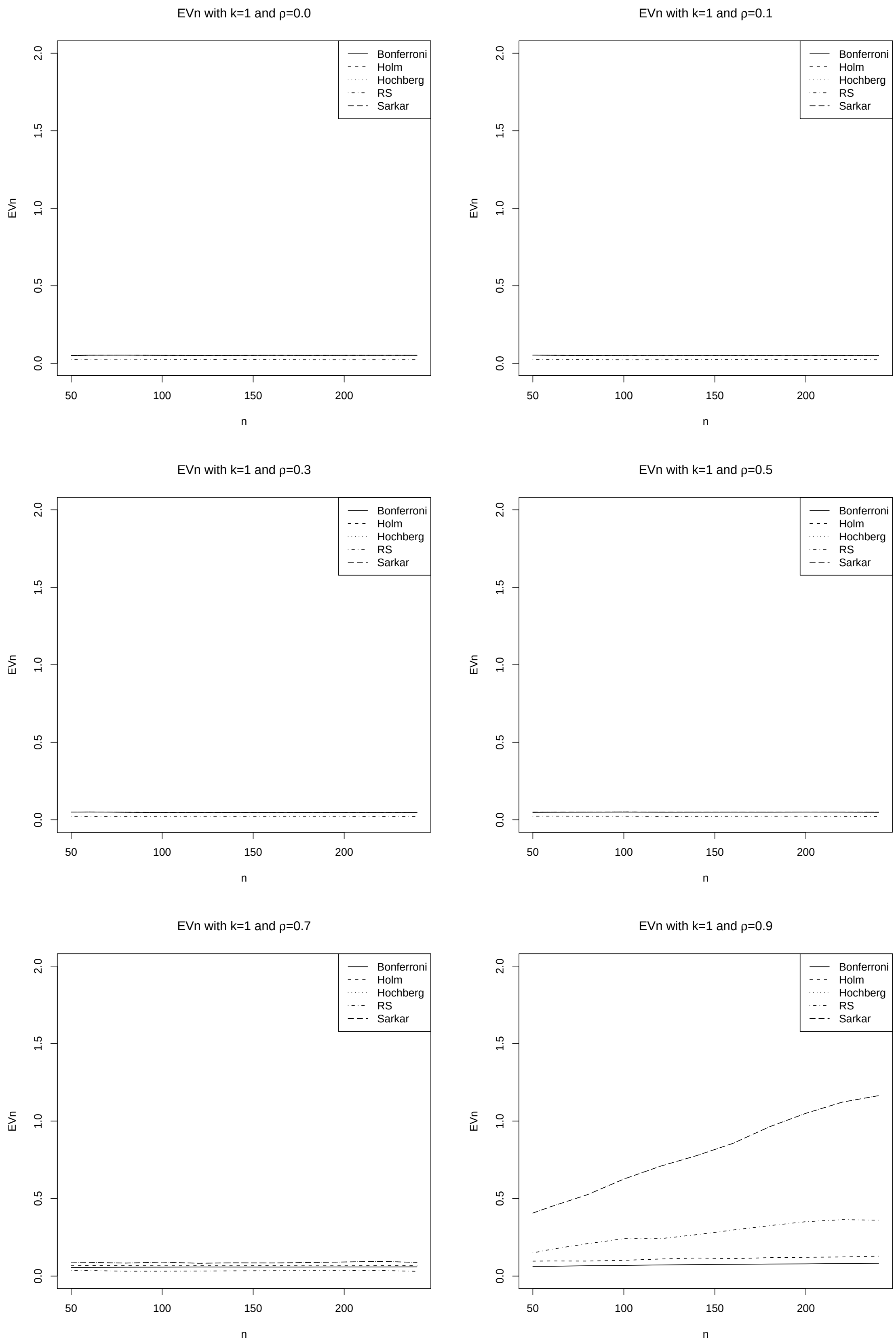

Figure 5.1: Expected number of Type I errors for k-FWER procedures with $k = 1$ and equicorrelated dependence

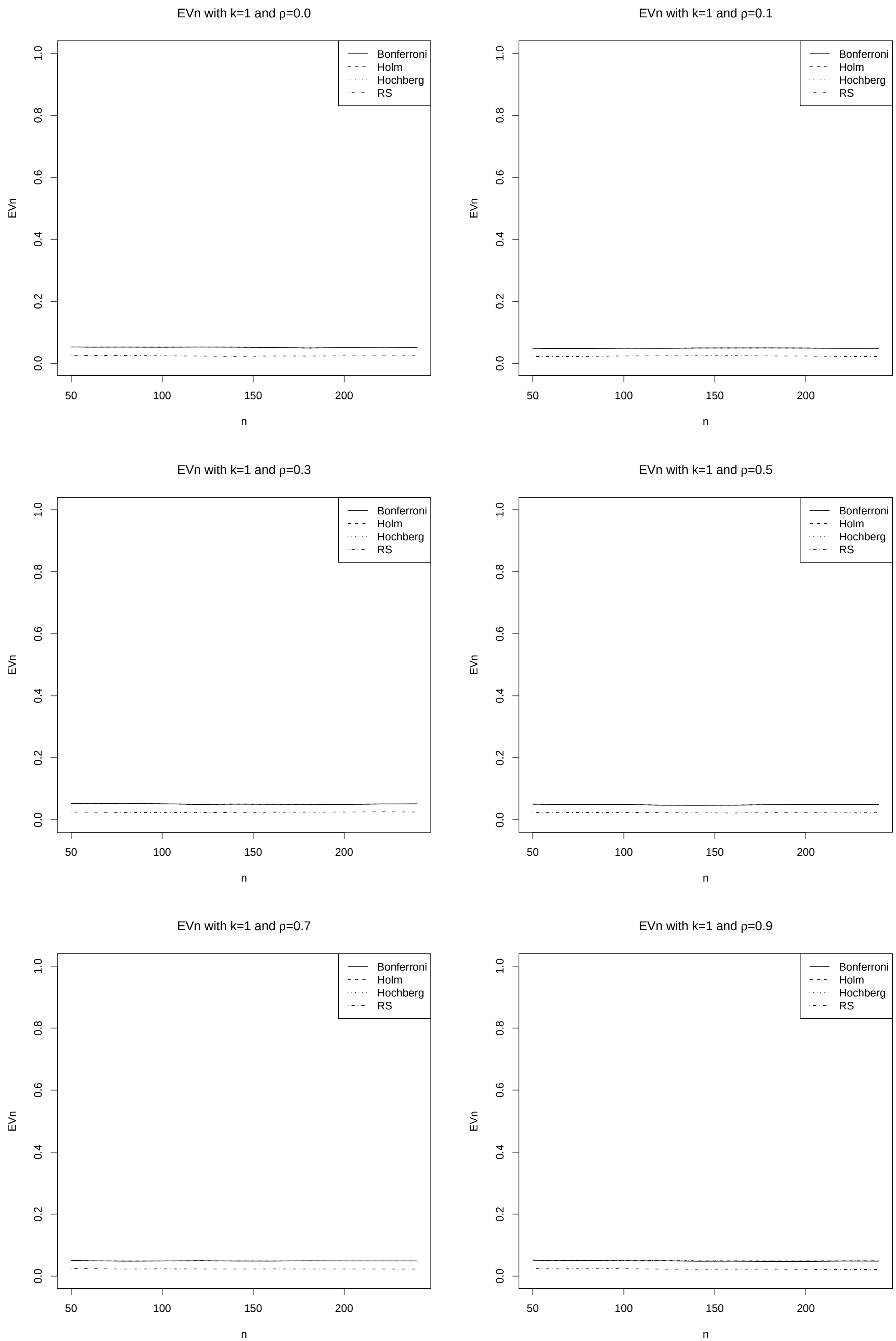

Figure 5.2: Expected number of Type I errors for k-FWER procedures with $k = 1$ and clumpy dependence

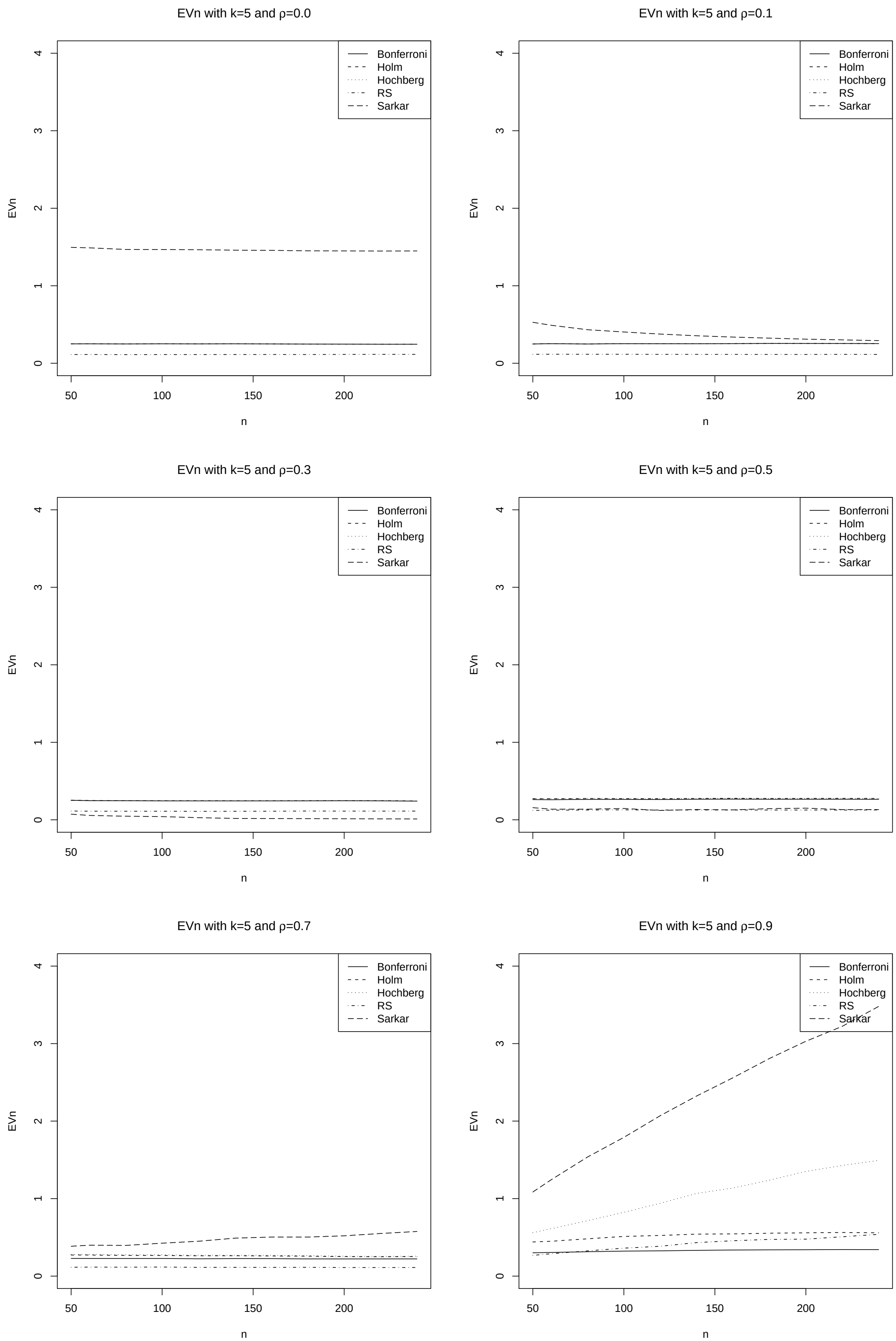

Figure 5.3: Expected number of Type I errors for k-FWER procedures with $k = 5$ and equicorrelated dependence

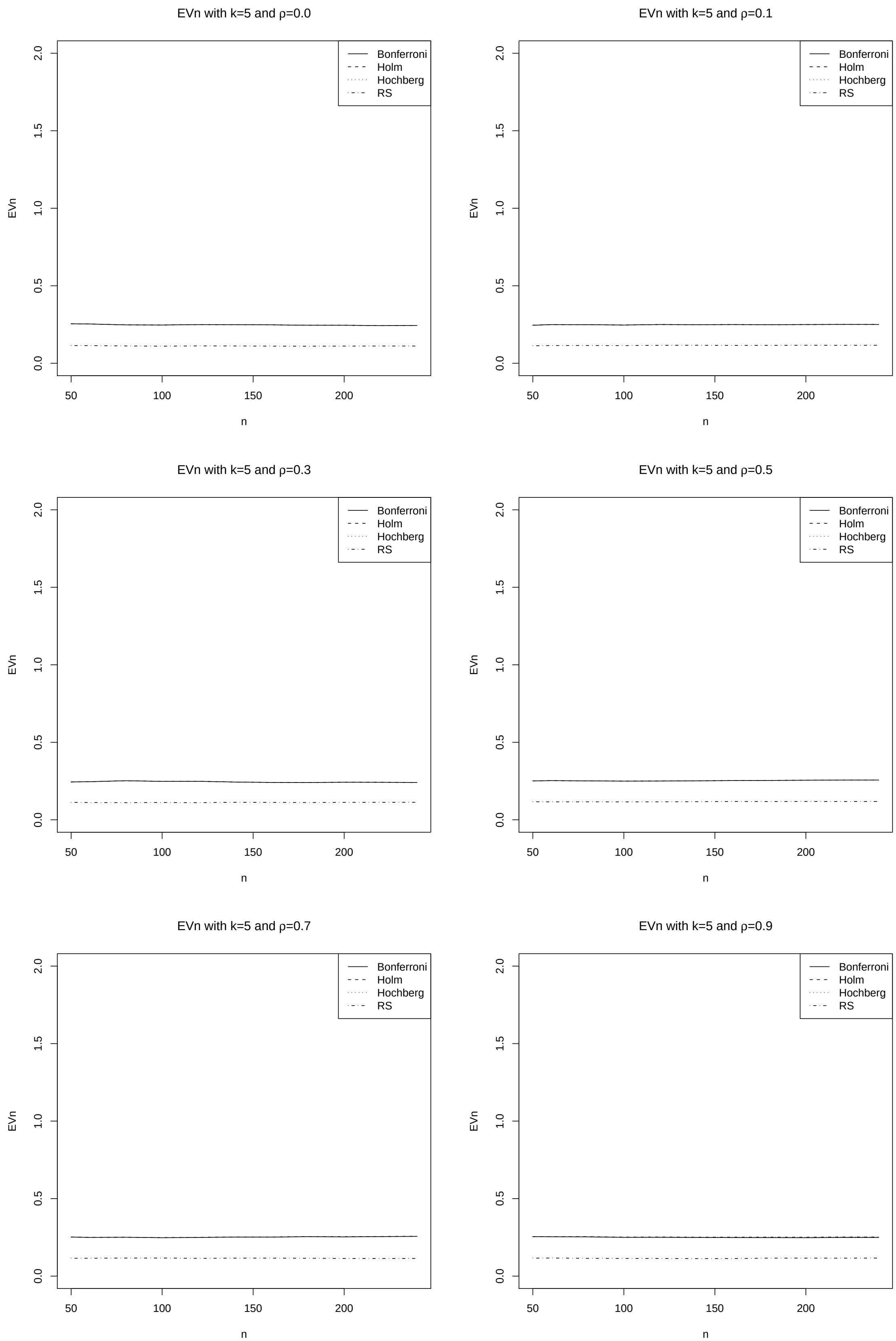

Figure 5.4: Expected number of Type I errors for k-FWER procedures with $k = 5$ and clumpy dependence

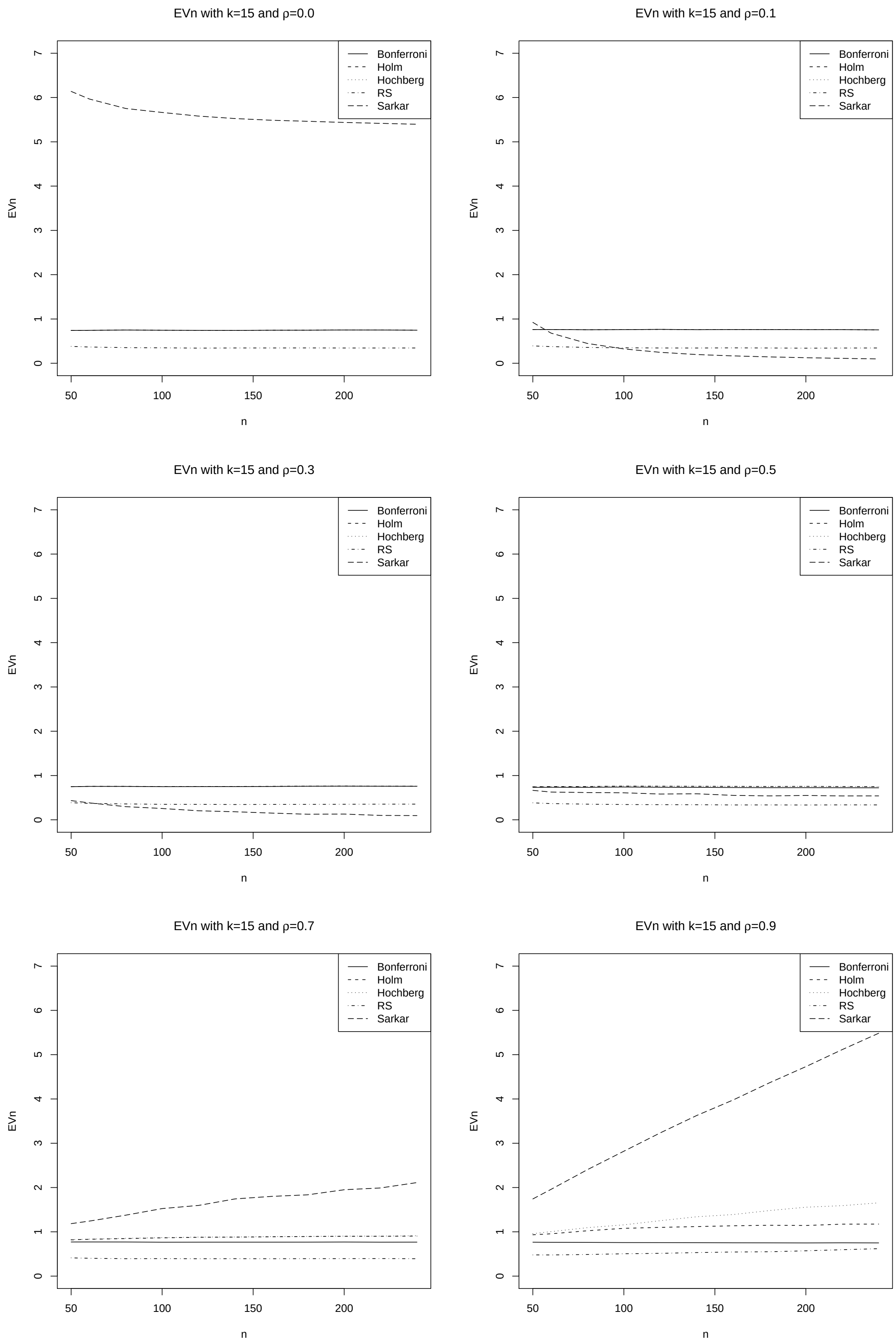

Figure 5.5: Expected number of Type I errors for k-FWER procedures with $k = 15$ and equicorrelated dependence

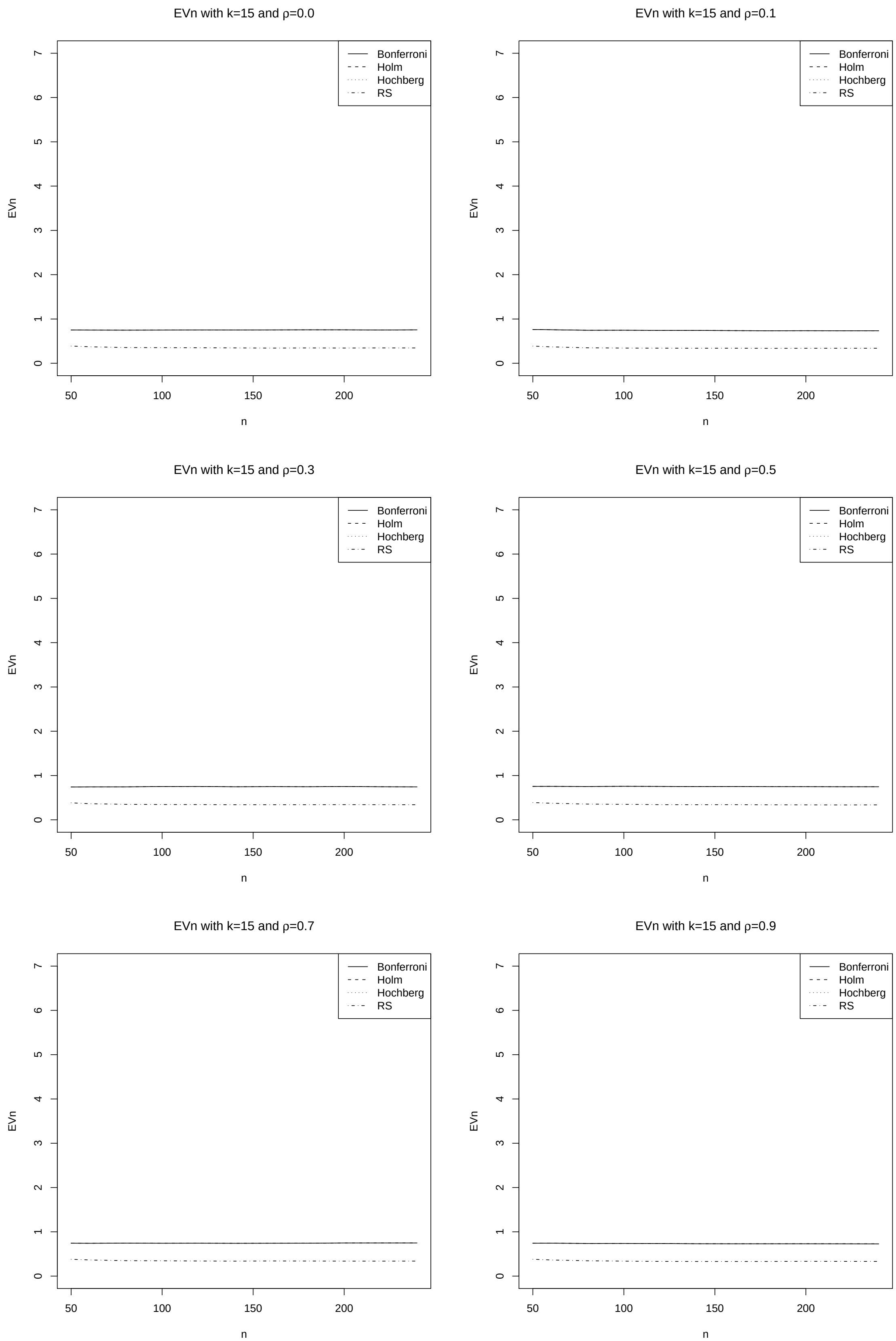

Figure 5.6: Expected number of Type I errors for k-FWER procedures with $k = 15$ and clumpy dependence

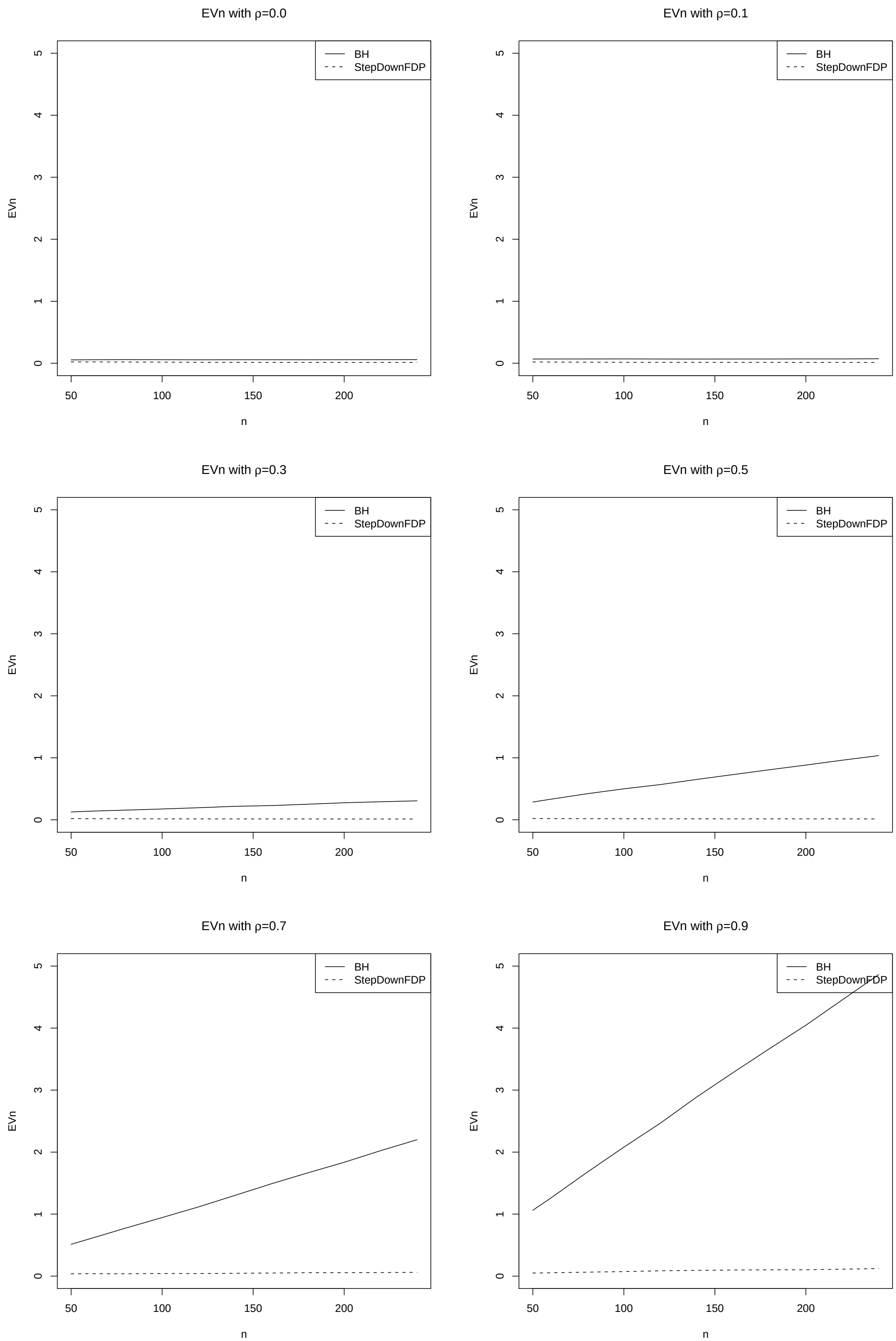

Figure 5.7: Expected number of Type I errors for BH and StepDownFDP procedures with equicorrelated dependence

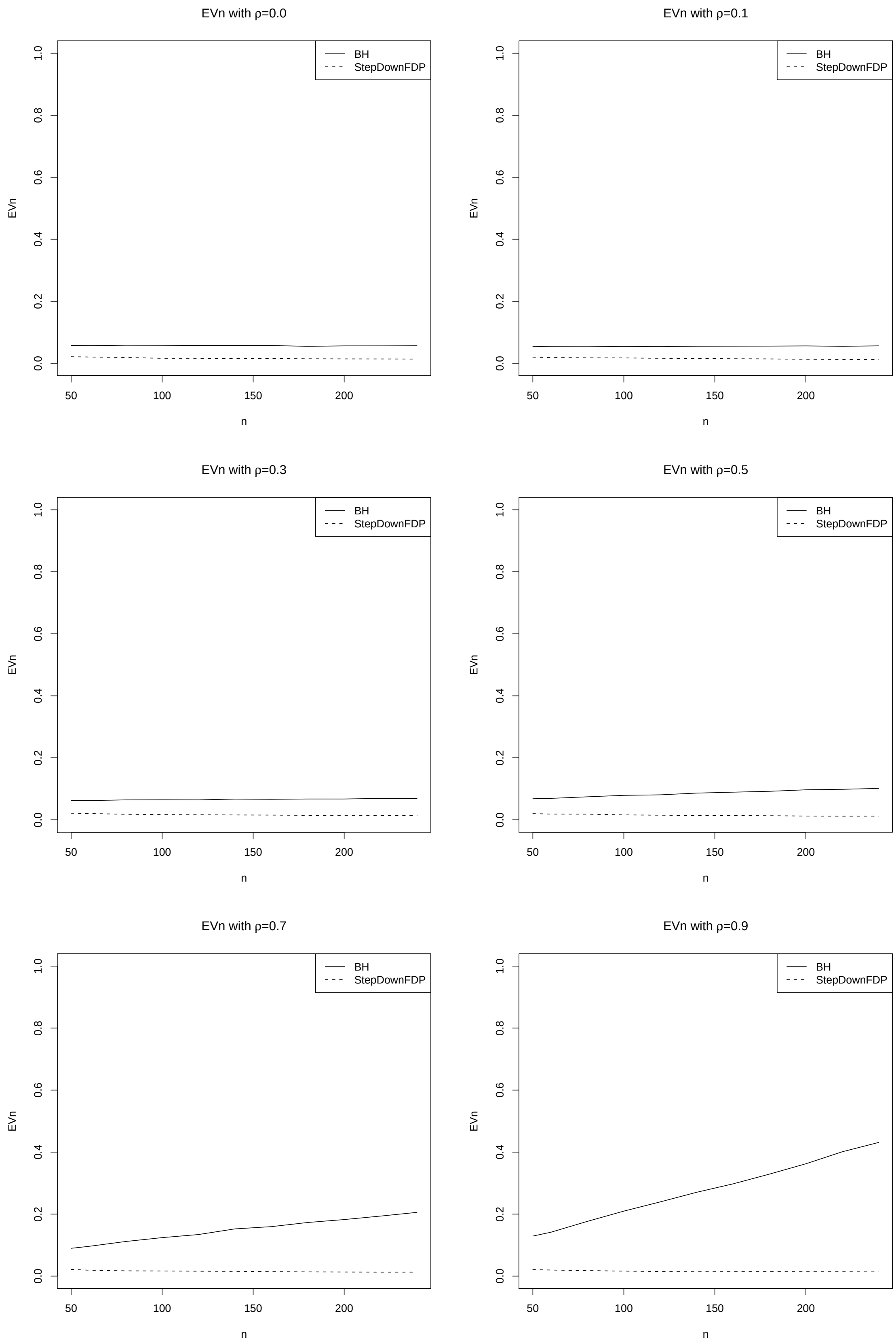

Figure 5.8: Expected number of Type I errors for BH and StepDownFDP procedures with clumpy dependence

5.1.2 Type R Enlargement FWR

The analysis of the Type R enlargement FWR follows similarly to the one for the expected number of Type I errors. The simulation plots which have a correlation coefficient of $\rho = 0$ resemble the corresponding theoretical plots in Figure 4.2. As the correlation coefficient, ρ, is increased, the Type R enlargement FWR appears to remain constant for the Bonferroni and Holm procedures in both dependent structure cases. In the simulation cases where there was an increase in the expected number of Type I errors, an increase in the Type R enlargement FWR was observed.

For the BH procedure, the Type R enlargement FWR when $\rho \geq 0.3$ is greater than the Type R enlargement FWR when $\rho = 0$. In the equicorrelated case, the Type R enlargement FWR for the step-up procedures, Hochberg, RS and Sarkar, increase over that seen in the independent case when $\rho > 0.5$. Again, this was exactly the condition in which the expected number of Type I errors increased.

With the clumpy dependence structure, the Type R enlargement FWR remains unaffected by correlation for all of the procedures except for the BH procedure. We expected to see this result given the simulation results for the expected number of Type I errors.

There was one departure from that seen with the expected number of Type I errors. The Type R enlargement FWR for the StepDownFDP procedure is higher than the independent case when $\rho \geq 0.7$. We did not see an increase in the expected number of Type I errors in this situation. I believe this is due to the very small value for the expected number of Type I errors for the StepDownFDP procedure when $\rho \geq 0.7$.

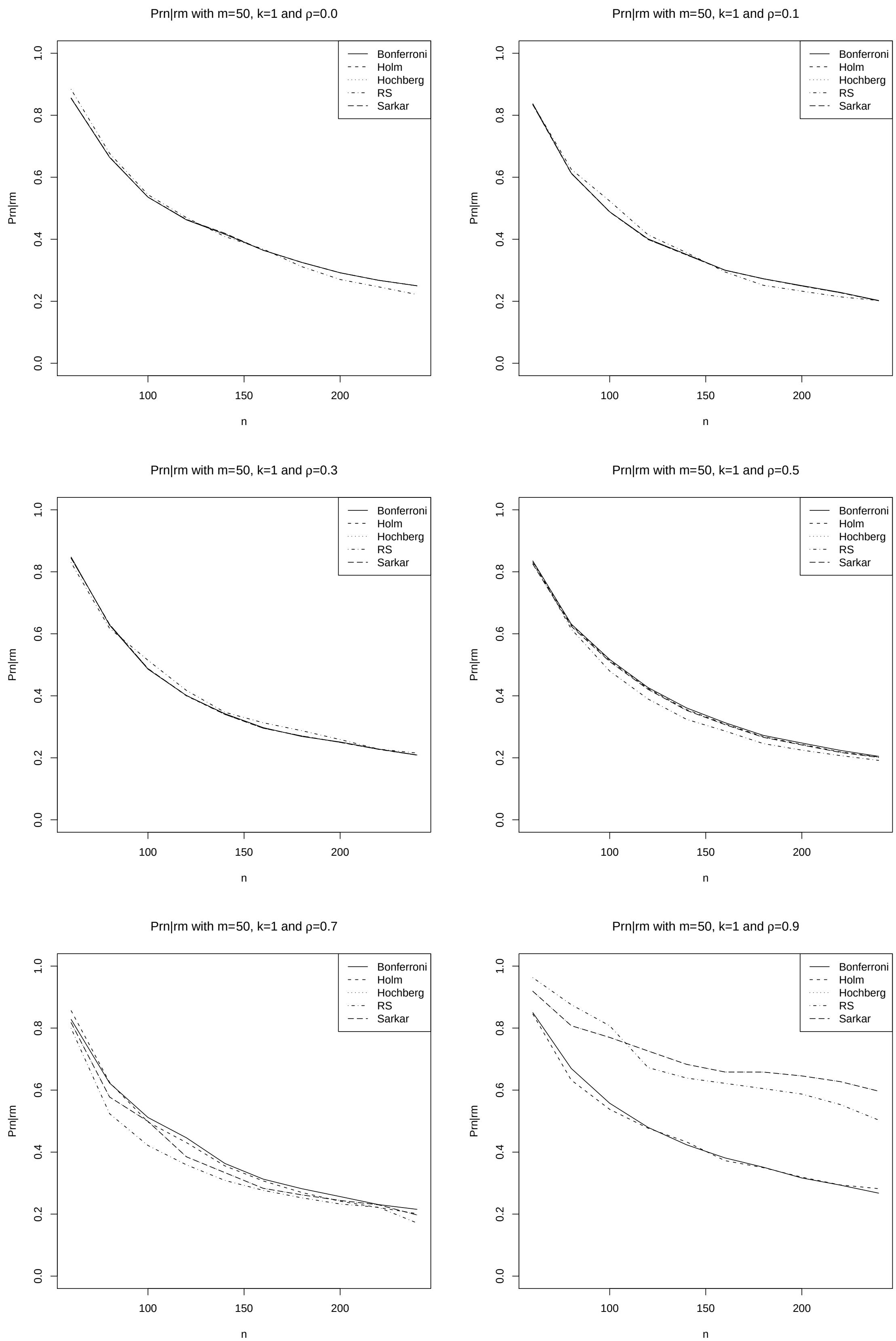

Figure 5.9: The Type R enlargement FWR for k-FWER procedures with $k = 1$ and equicorrelated dependence

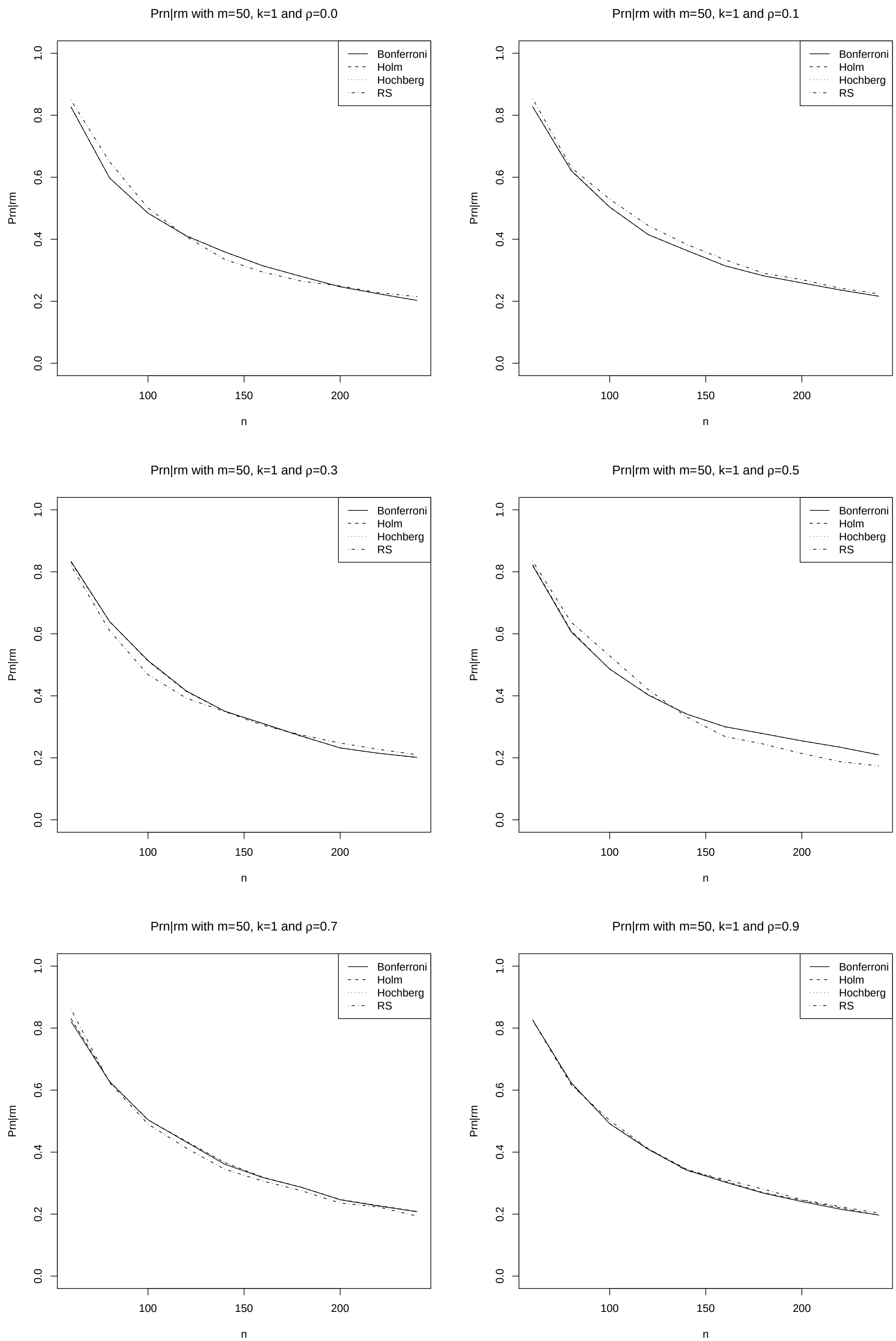

Figure 5.10: The Type R enlargement FWR for k-FWER procedures with $k = 1$ and clumpy dependence

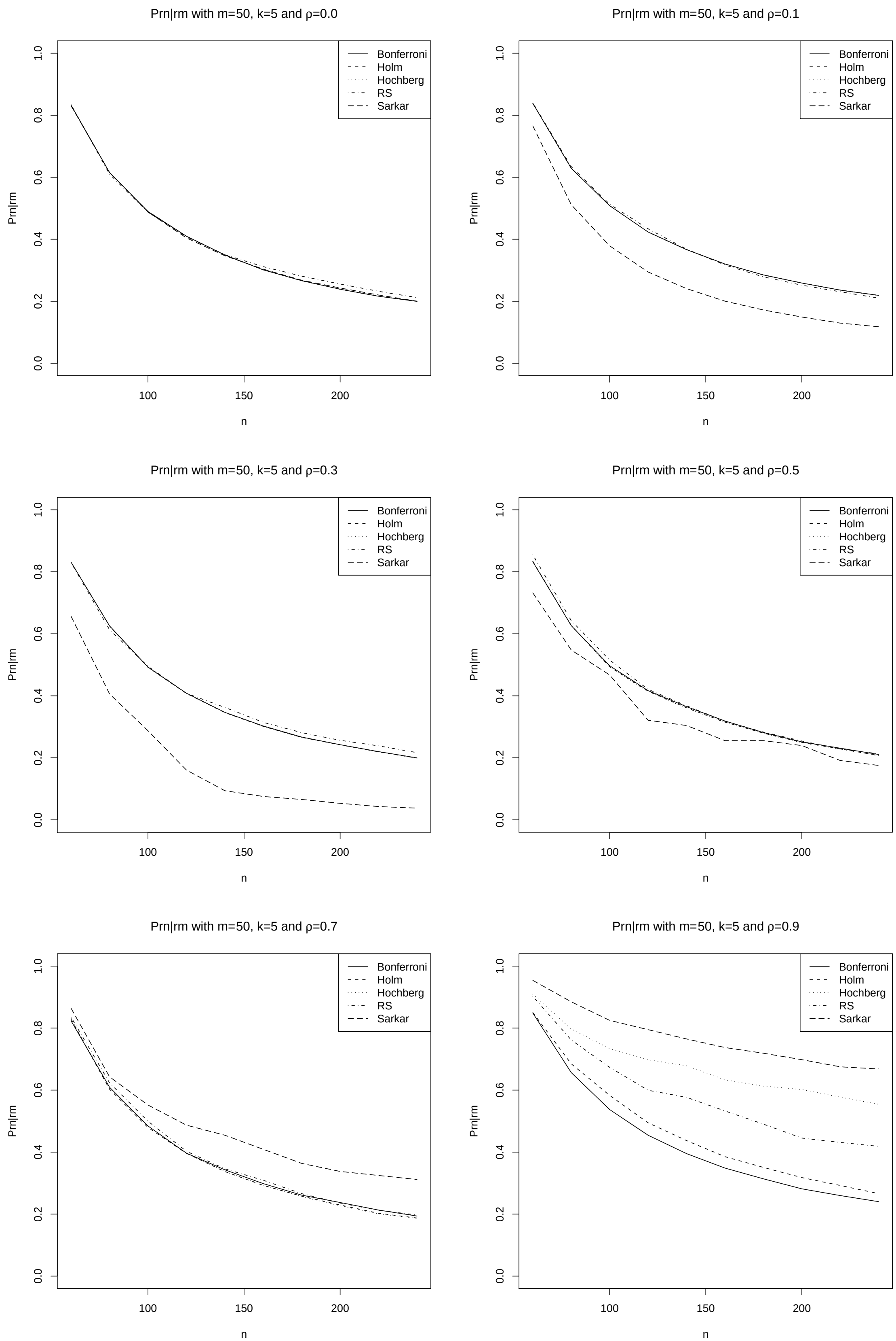

Figure 5.11: The Type R enlargement FWR for k-FWER procedures with $k = 5$ and equicorrelated dependence

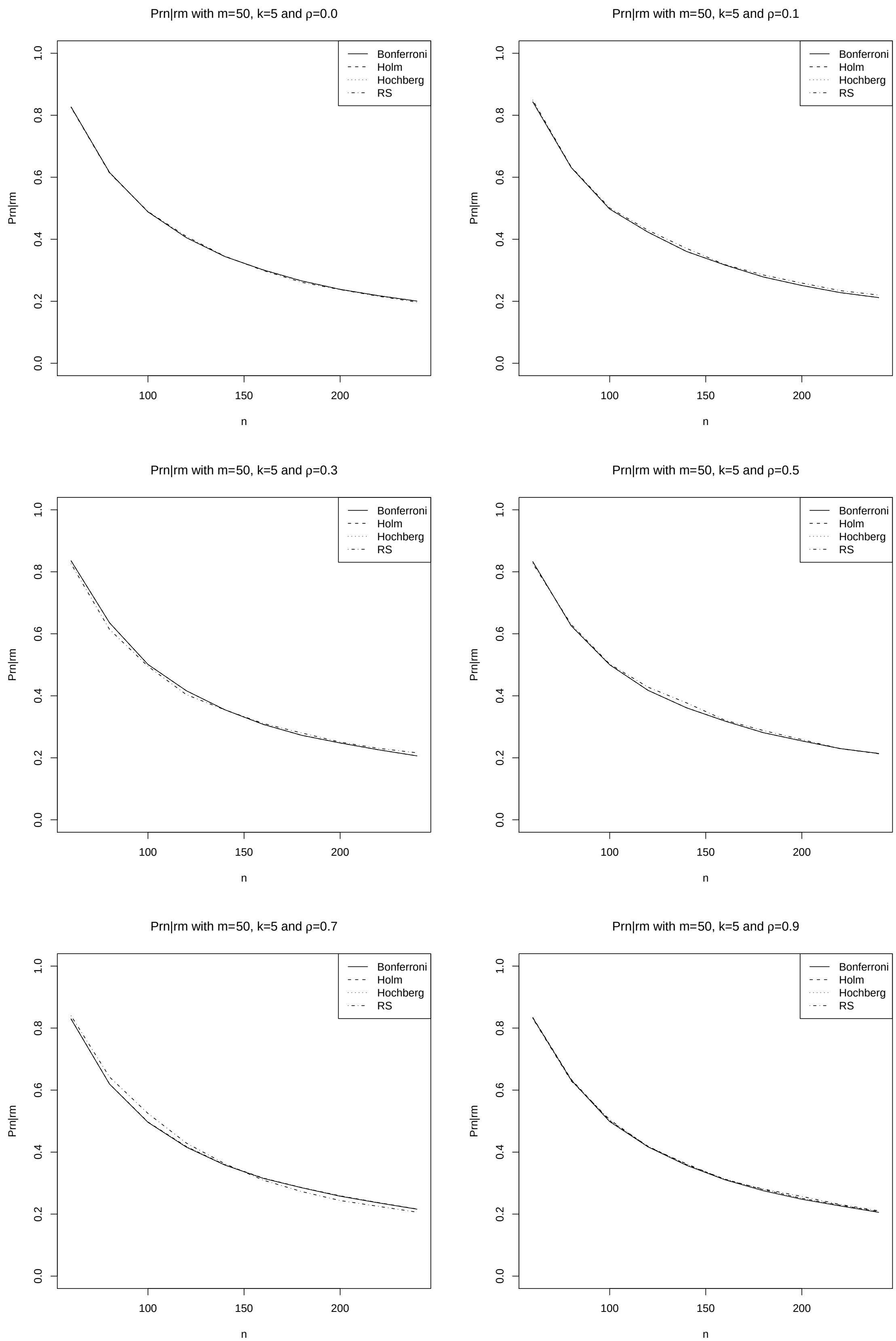

Figure 5.12: The Type R enlargement FWR for k-FWER procedures with $k = 5$ and clumpy dependence

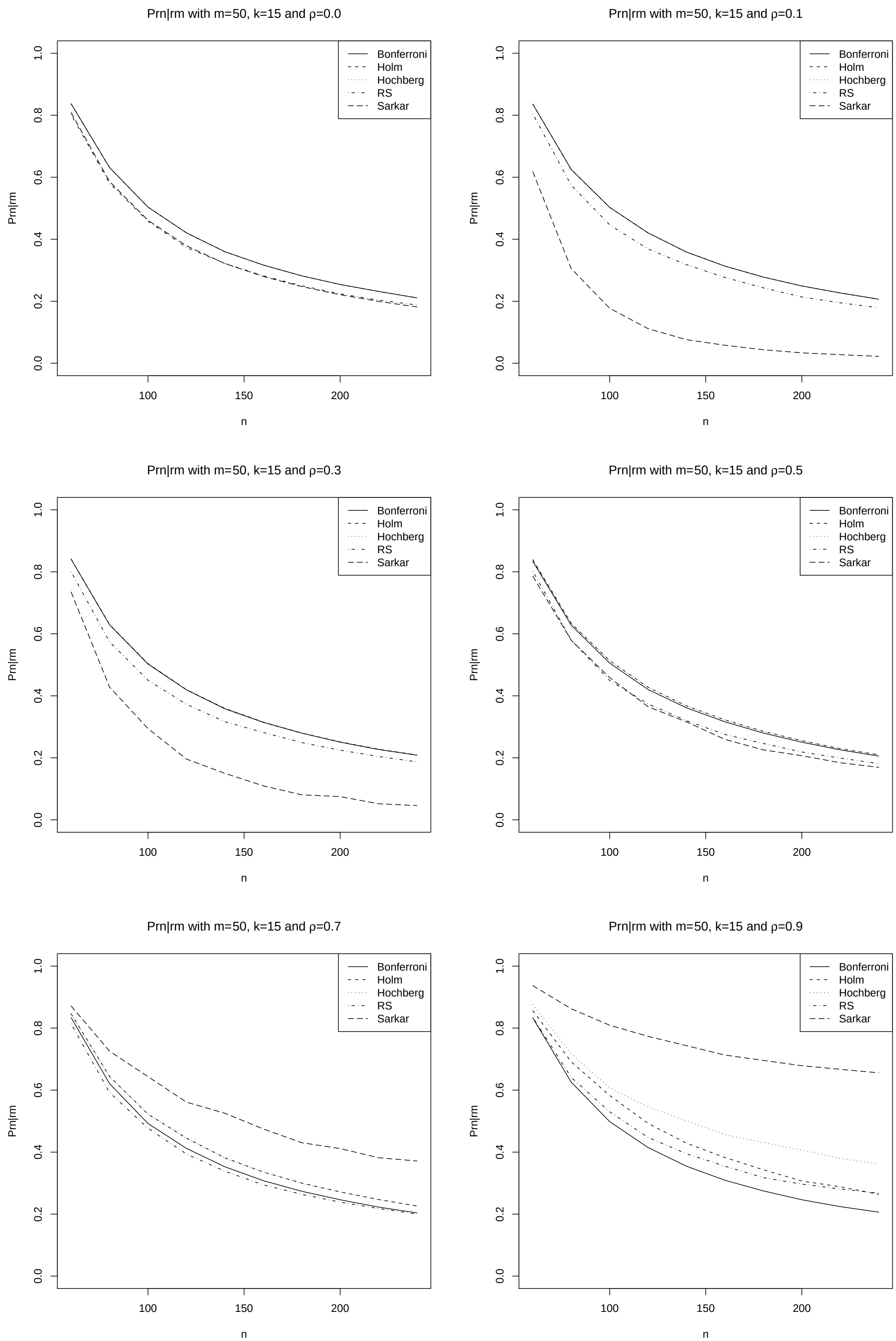

Figure 5.13: The Type R enlargement FWR for k-FWER procedures with $k = 15$ and equicorrelated dependence

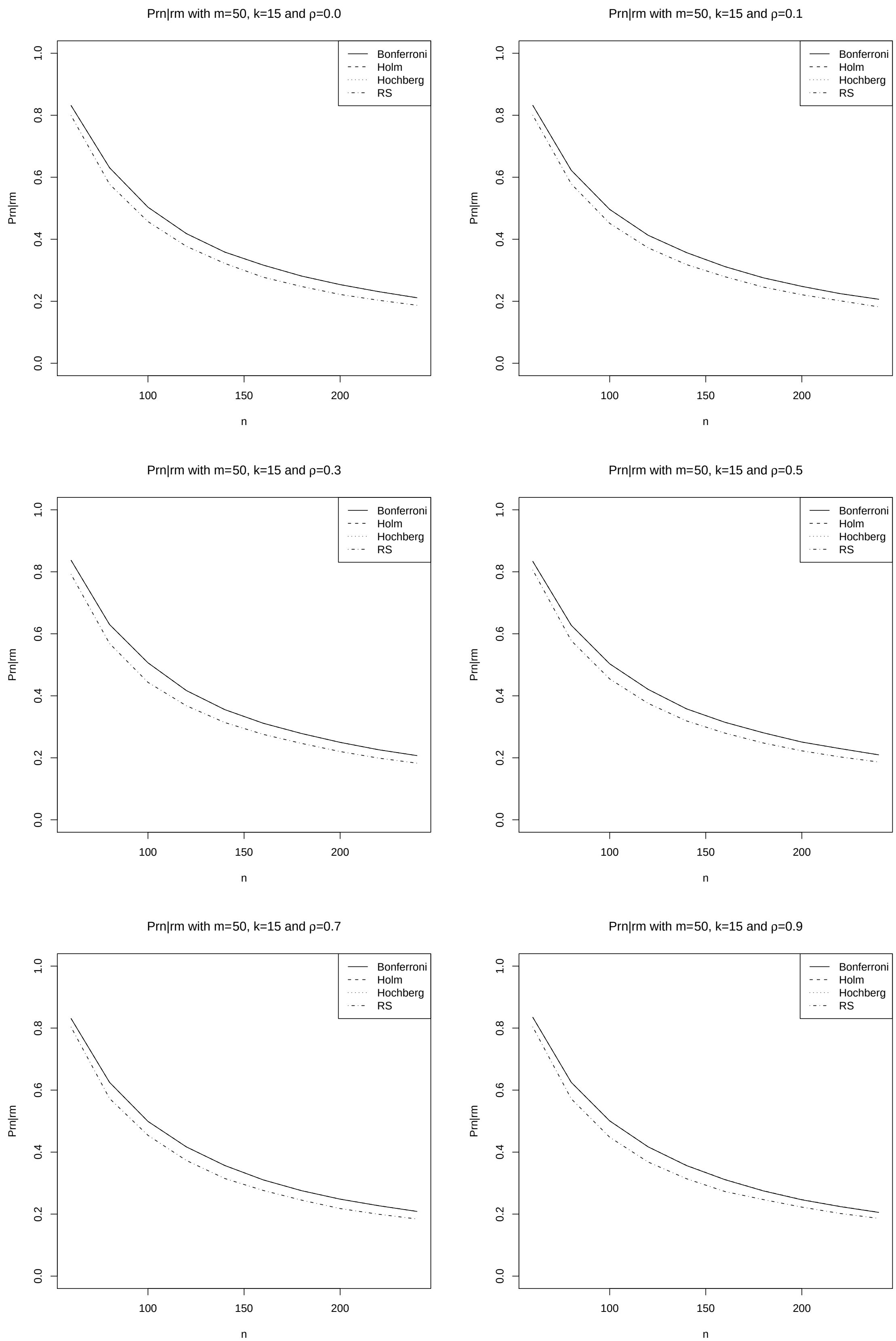

Figure 5.14: The Type R enlargement FWR for k-FWER procedures with $k = 15$ and clumpy dependence

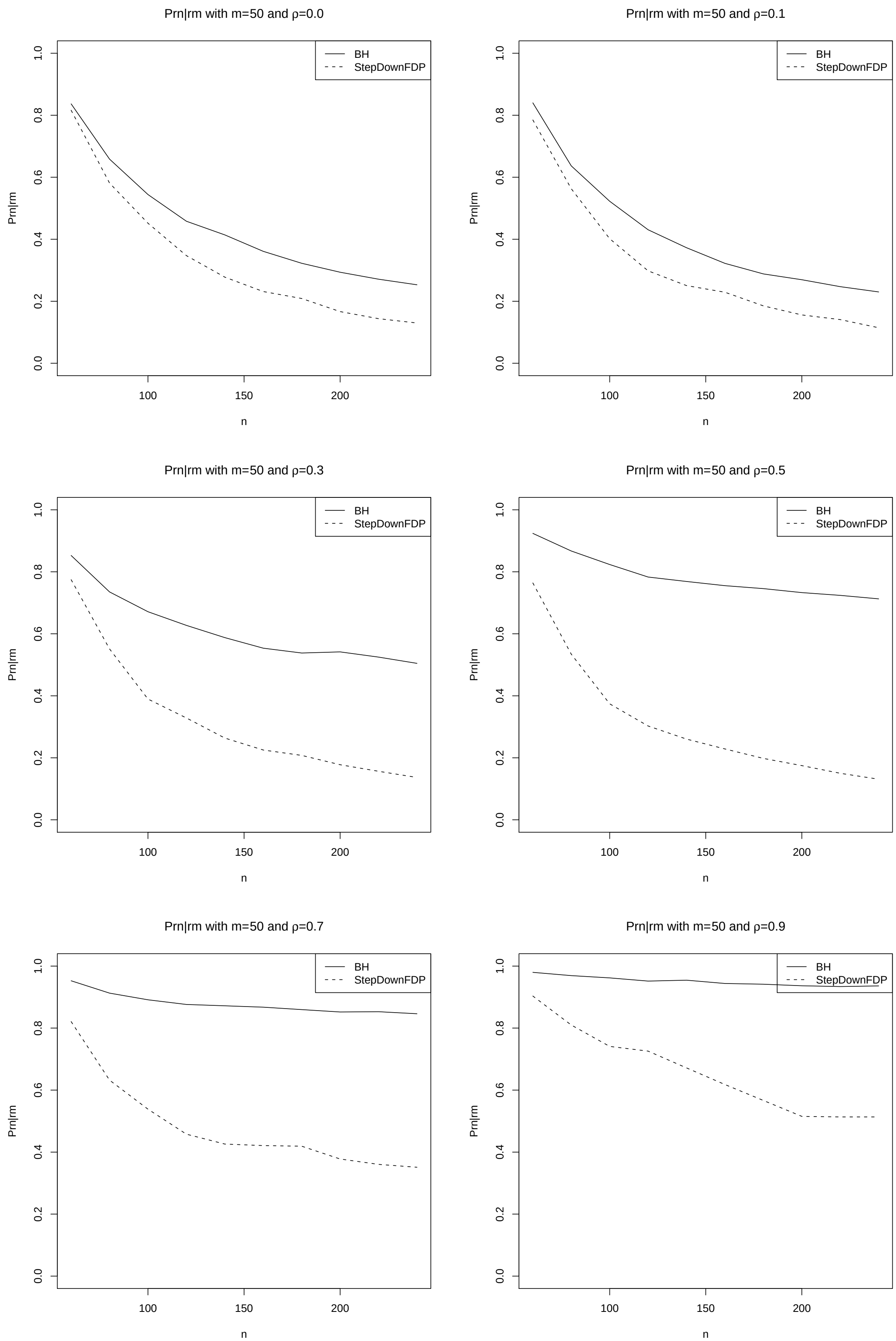

Figure 5.15: The Type R enlargement FWR for the BH and StepDownFDP procedures with equicorrelated dependence

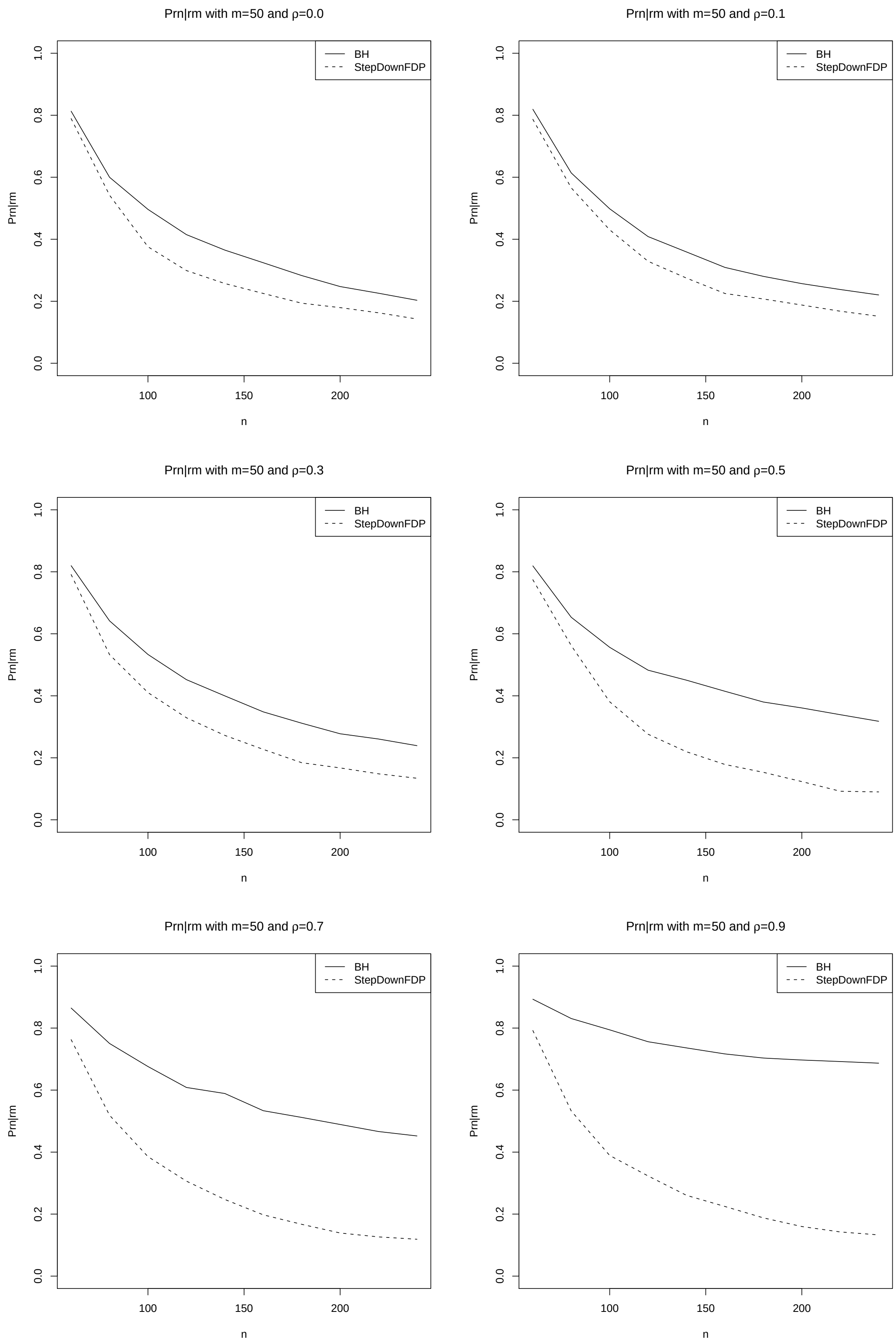

Figure 5.16: The Type R enlargement FWR for the BH and StepDownFDP procedures with clumpy dependence

5.1.3 Type A Contraction FWR

The effect of correlation on the Type A contraction FWR is similar to that found with the expected Type I error and Type R enlargement FWR. There doesn't appear to be any correlation effect on the Type A contraction FWR for the Bonferroni, and StepDownFDP procedures. The Type A contraction FWR for the Sarkar procedure which has different critical values for different correlations increases when the correlation coefficient is increased from 0 to 0.5 and then decreases as the correlation coefficient is increased above 0.5. This is exactly the opposite effect as was seen to occur to the Type R enlargement FWR for the Sarkar procedure.

The Type A contraction FWR for the BH procedure decreases with increased levels of correlation. Although the BH procedure was affected by the level of correlation, the BH and StepDownFDP procedures had the highest Type A contraction FWR among all of the procedures examined.

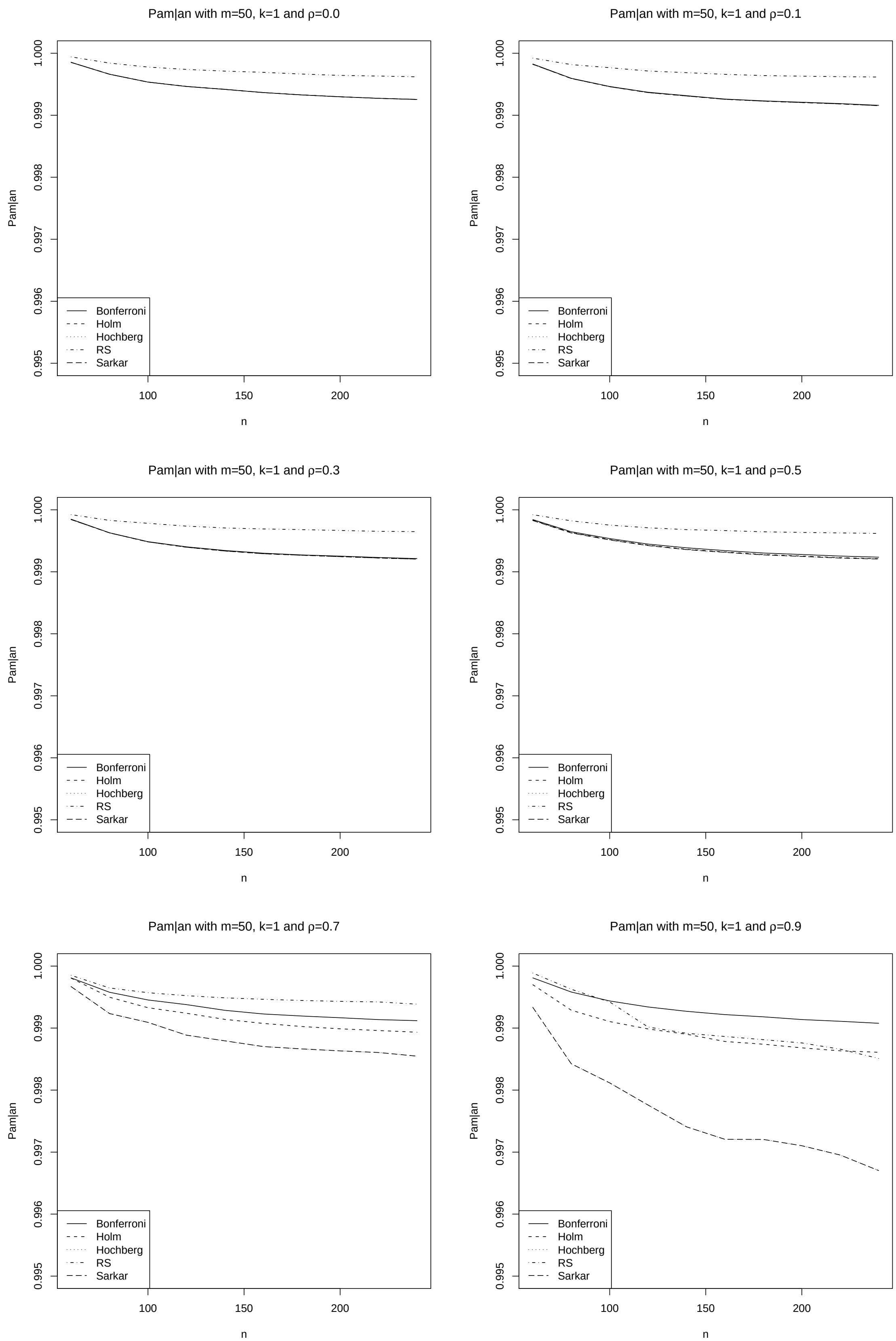

Figure 5.17: The Type A contraction FWR for k-FWER procedures with $k = 1$ and equicorrelated dependence

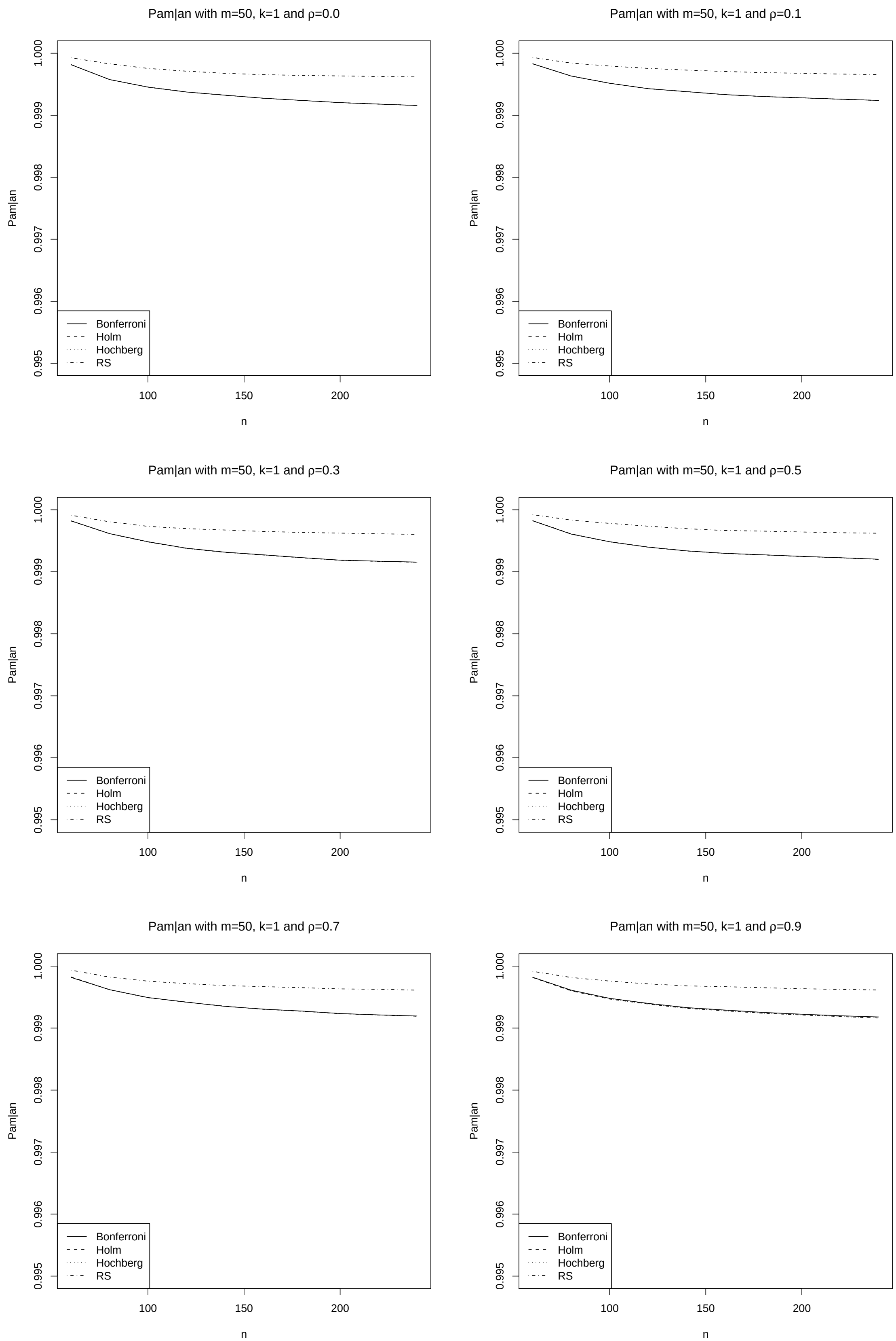

Figure 5.18: The Type A contraction FWR for k-FWER procedures with $k = 1$ and clumpy dependence

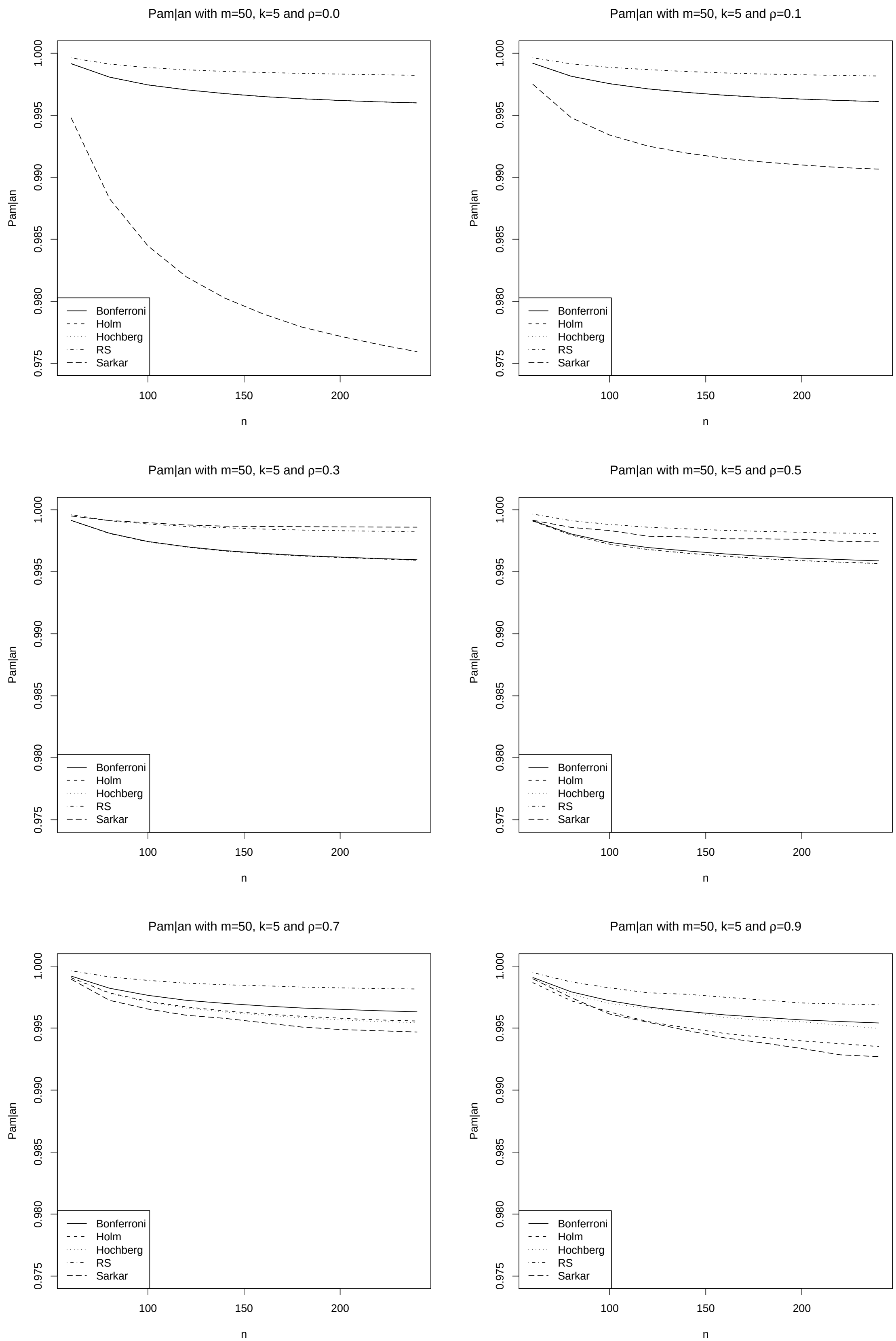

Figure 5.19: The Type A contraction FWR for k-FWER procedures with $k = 5$ and equicorrelated dependence

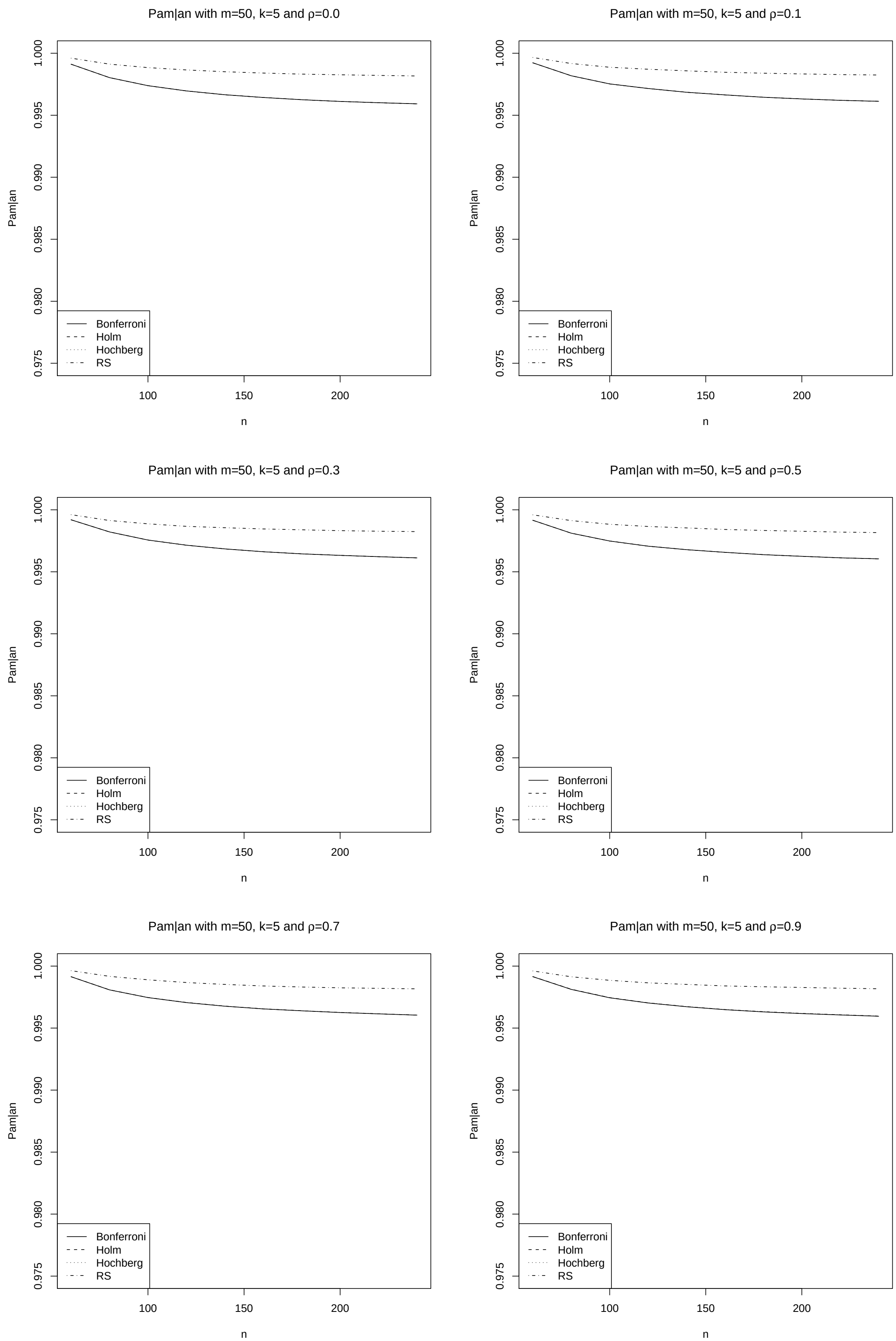

Figure 5.20: The Type A contraction FWR for k-FWER procedures with $k = 5$ and clumpy dependence

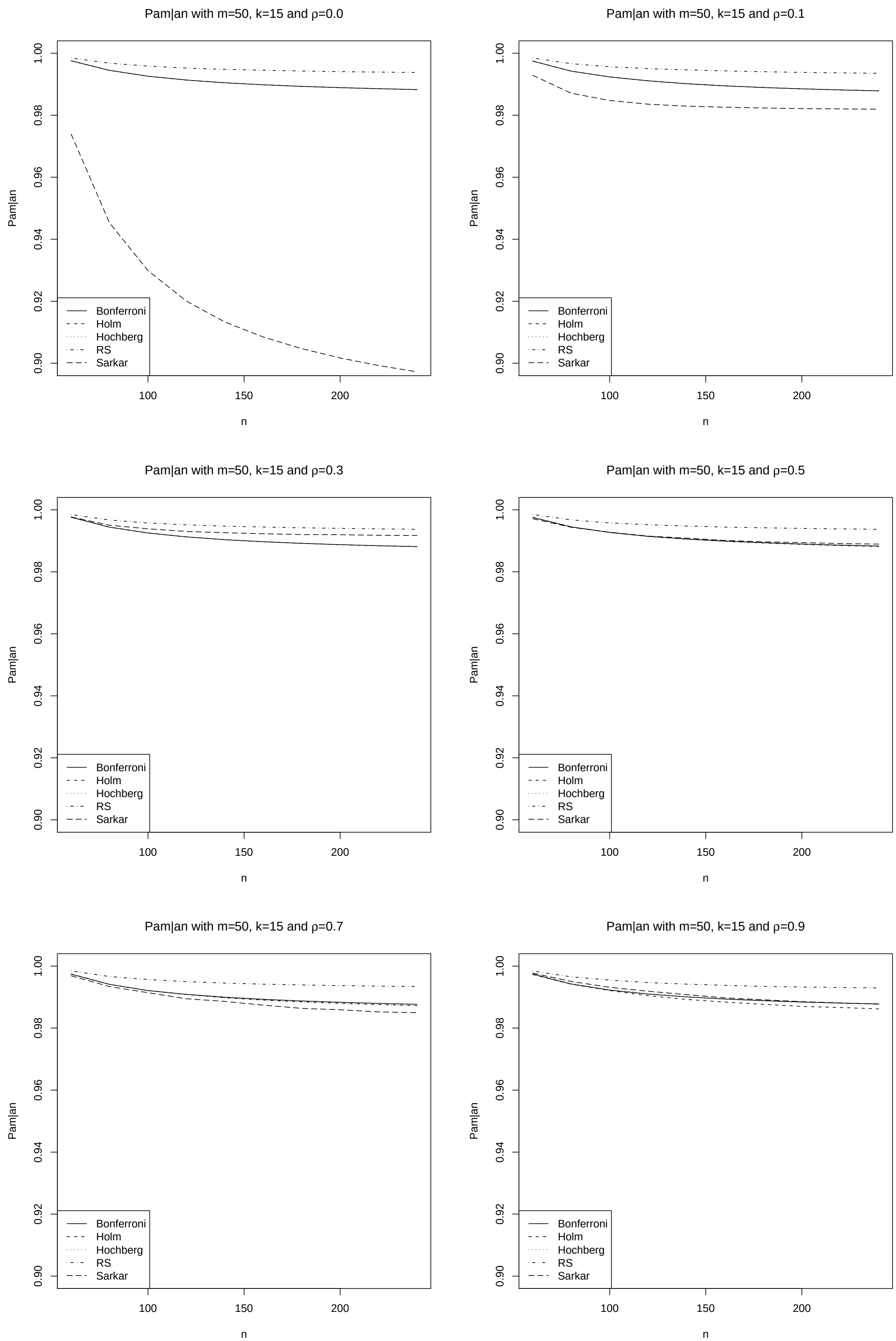

Figure 5.21: The Type A contraction FWR for k-FWER procedures with $k = 15$ and equicorrelated dependence

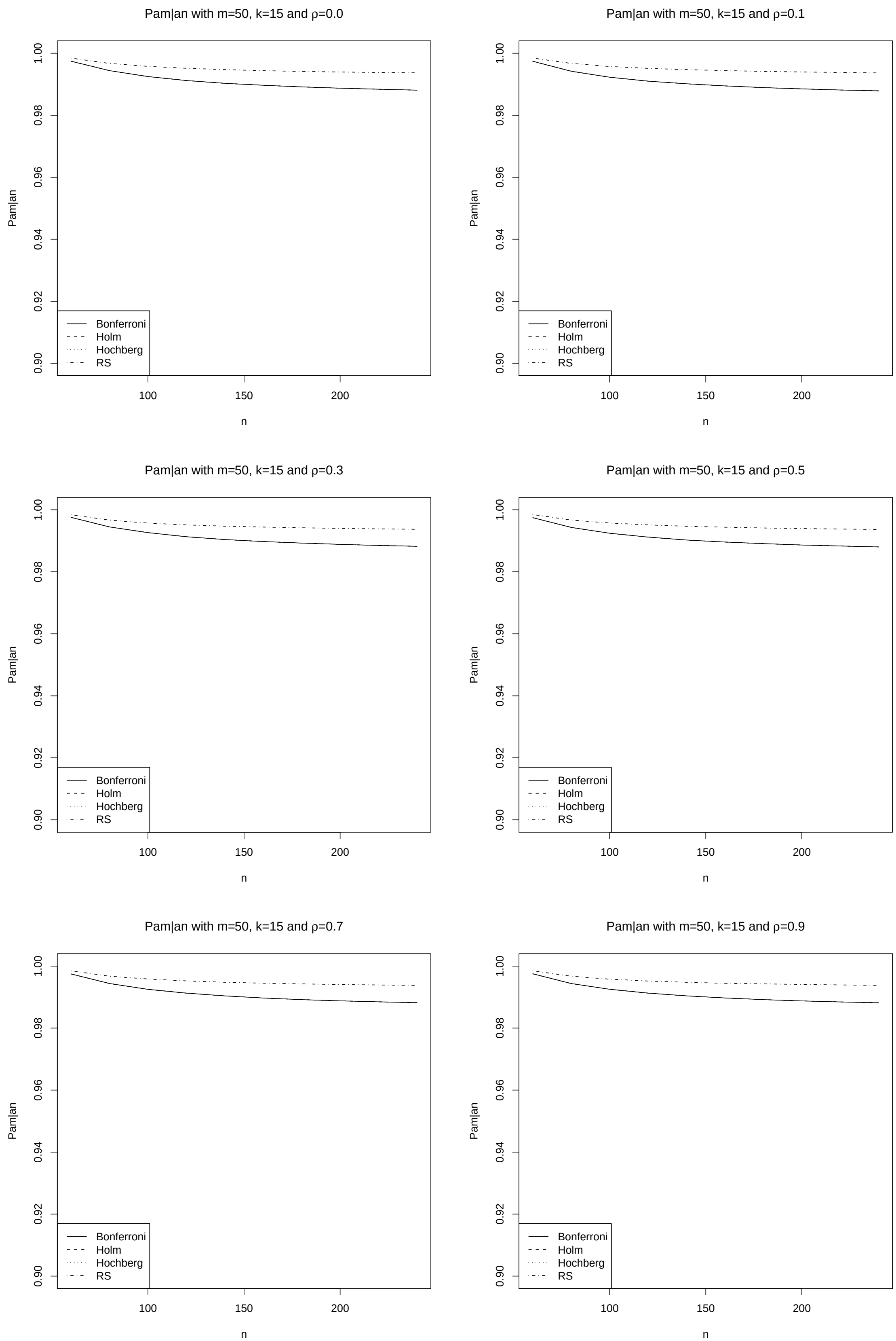

Figure 5.22: The Type A contraction FWR for k-FWER procedures with $k = 15$ and clumpy dependence

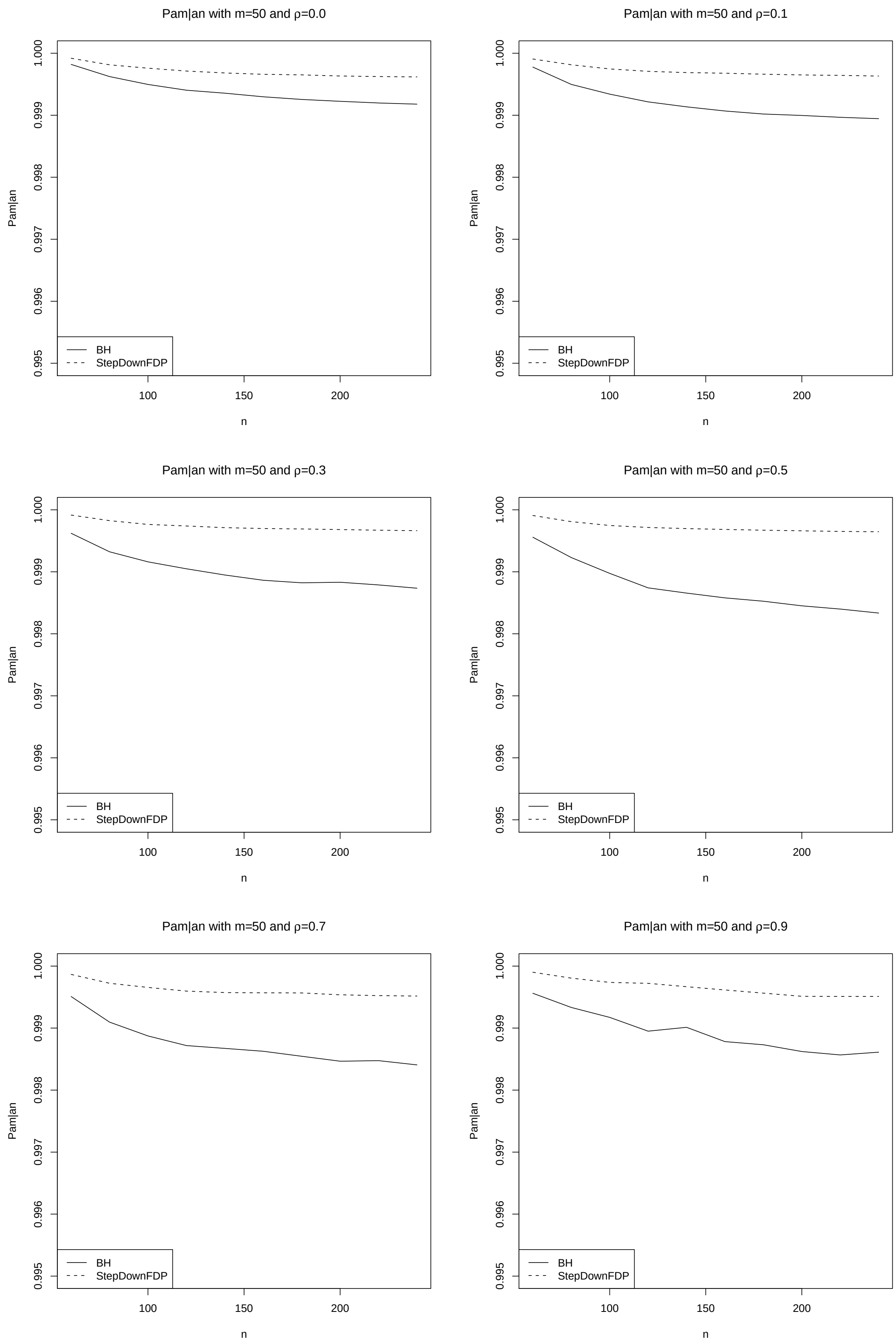

Figure 5.23: The Type A contraction FWR for the BH and StepDownFDP procedures with equicorrelated dependence

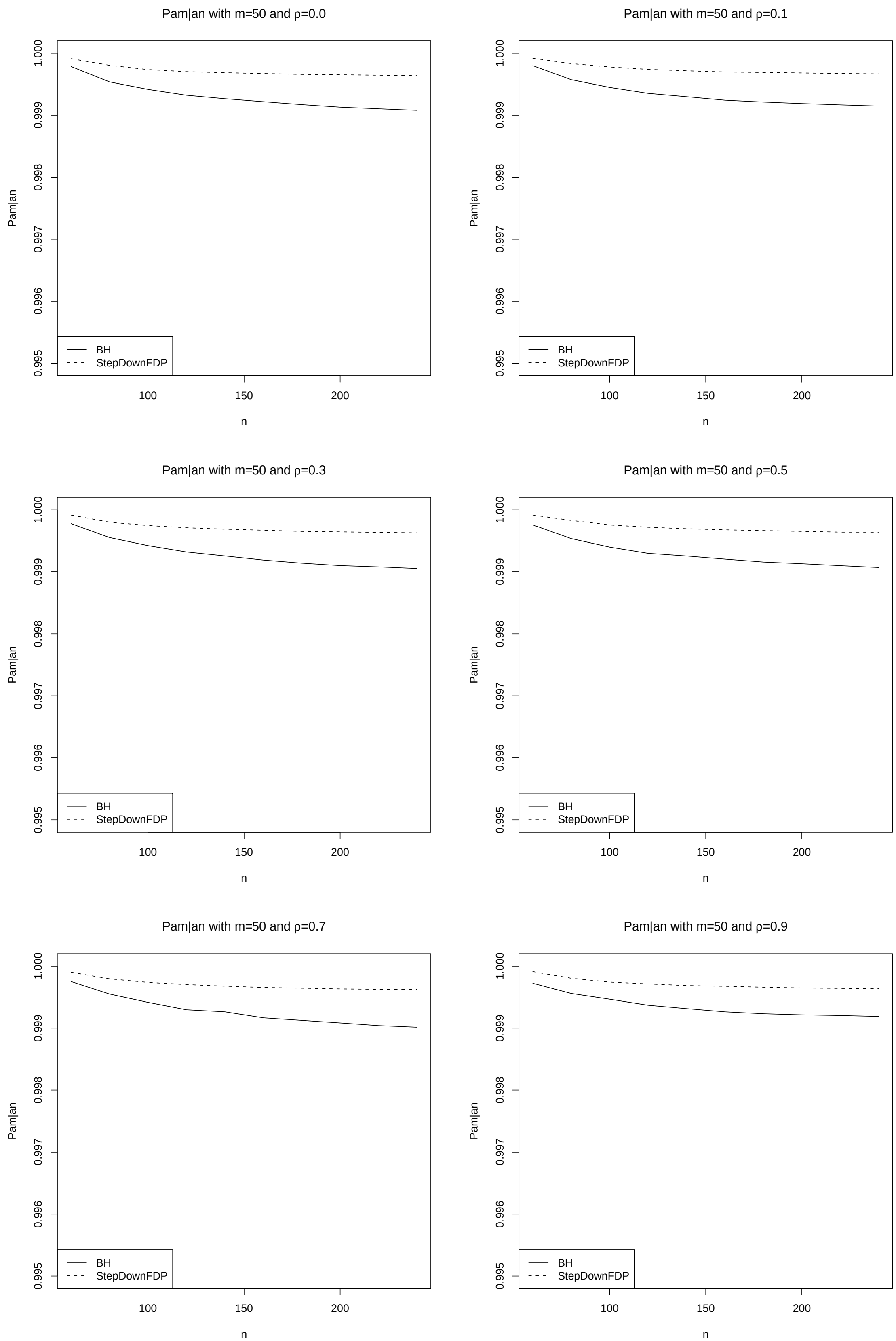

Figure 5.24: The Type A contraction FWR for the BH and StepDownFDP procedures with clumpy dependence

5.2 Strong Control of Familywise Error Rate

In this section, we want to see what happens to the familywise robustness measures as the familywise error rate (FWER) is controlled strongly. In each of the simulations which was conducted using 20,000 repetitions, half of the null hypotheses were assumed to be false. One-sided hypotheses testing with null hypotheses $H_i : \mu_i = 0$ against the alternative hypotheses $H_i' : \mu_i > 0$, for $i = 1, 2, \ldots, n$ was performed. $Z_i \sim \mathcal{N}(0, 1)$ for $i = \{0, 1, 2, \ldots, n\}$, $d_i = \{0\}$ for $i = \{1, 3, 5, \ldots, n-1\}$, $d_i = \{1, 2\}$ for $i = \{2, 4, 6, \ldots, n\}$, $U = \sqrt{\mathcal{X}_\nu^2/\nu}$, and $T_i = \{\sqrt{1-\rho}Z_i - \sqrt{\rho}Z_0\}/U + d_i$ for $i = \{1, 2, 3, \ldots, n\}$. Thus, the random variables, $T_1, T_2, \ldots, T_n$ have a Student's n-variate central t distribution with ν degrees of freedom and associated common correlation coefficient ρ.

5.2.1 Type R Enlargement FWR

In Figures 5.25-5.32, the Type R enlargement FWR was analyzed as FWER was controlled strongly. Half of the test statistics were simulated to come from a normal distribution with mean d. We were interested in the effect, if any, that changing the value of the correlation coefficient, ρ, and changing the size of the mean difference, d, had on the Type R enlargement FWR.

While holding k and d constant, the Type R enlargement FWR for the Bonferroni and Holm procedures appears to remain constant for the varied levels of correlation. The Type R enlargement FWR for the BH and StepDownFDP procedures increase as the correlation coefficient is increased. The increase is more pronounced for the BH procedure when $d = 1$ and for the StepDownFDP when $d = 2$. Finally, as the correlation increases, the Type R enlargement FWR for the Sarkar procedure decreases initially as correlation increases, but then it increases as correlation increases above 0.5.

While holding k and ρ constant, the Type R enlargement FWR for all of the procedures appeared to increase as d was increased from 1 to 2. This increase was dramatic for the BH and StepDown procedures and only slight for the others.

5.2.2 Type A Contraction FWR

The Type A contraction FWR while allowing the simulation parameters, k, d, and ρ to vary is shown in Figures 5.33-5.40. The effect of increasing k on the k-FWER measures was a reduction in the Type A contraction FWR holding ρ and d constant. The Type A contraction FWR for all of the procedures increases as ρ was increased. This was especially pronounced for the Sarkar procedure. Lastly, as d was increased, the Type A contraction FWR for all of the procedures decreased. The Type A contraction for the Sarkar procedure exhibited the most dramatic decrease amongst all of the procedures.

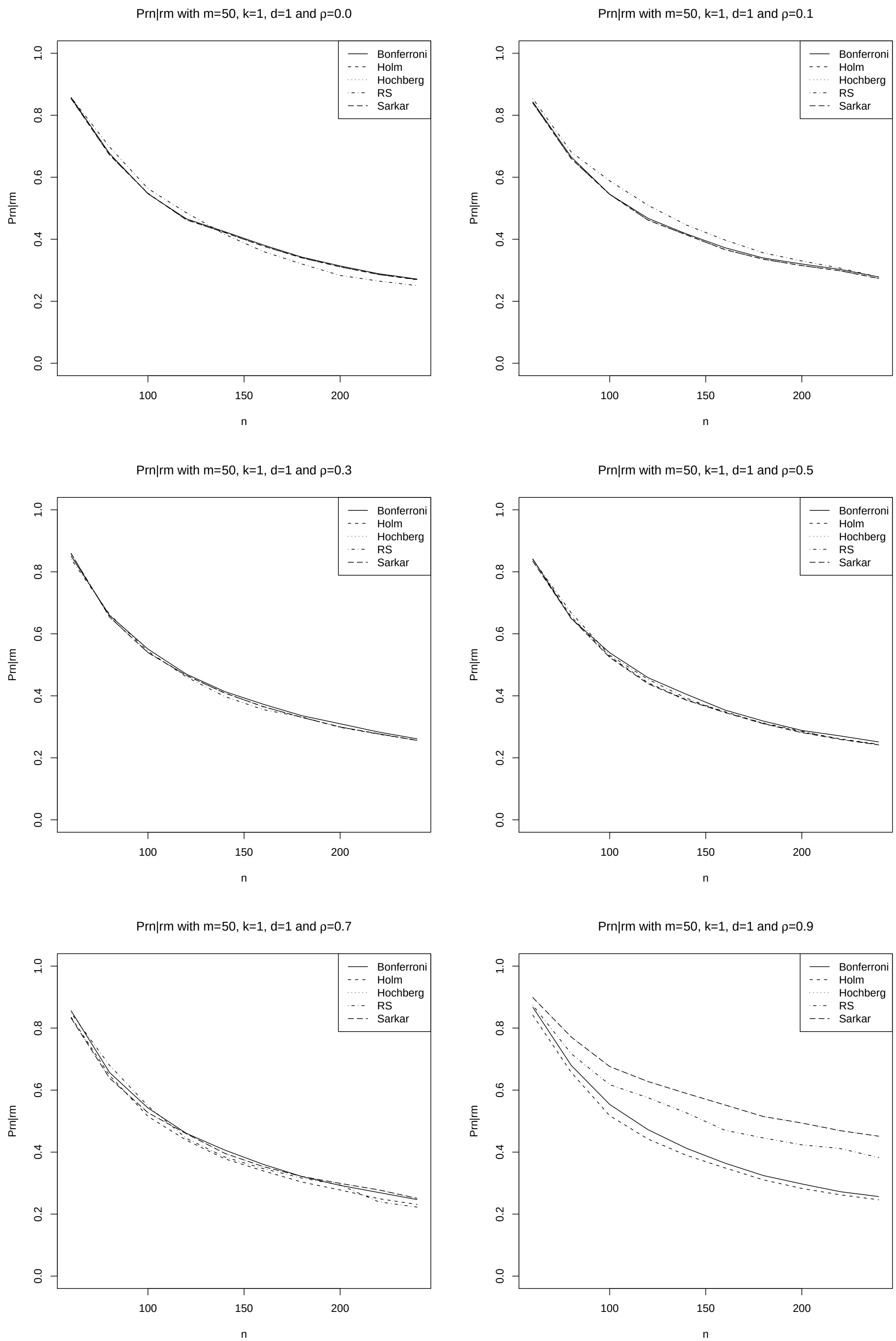

Figure 5.25: The Type R enlargement FWR while FWER is controlled strongly with $k = 1$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

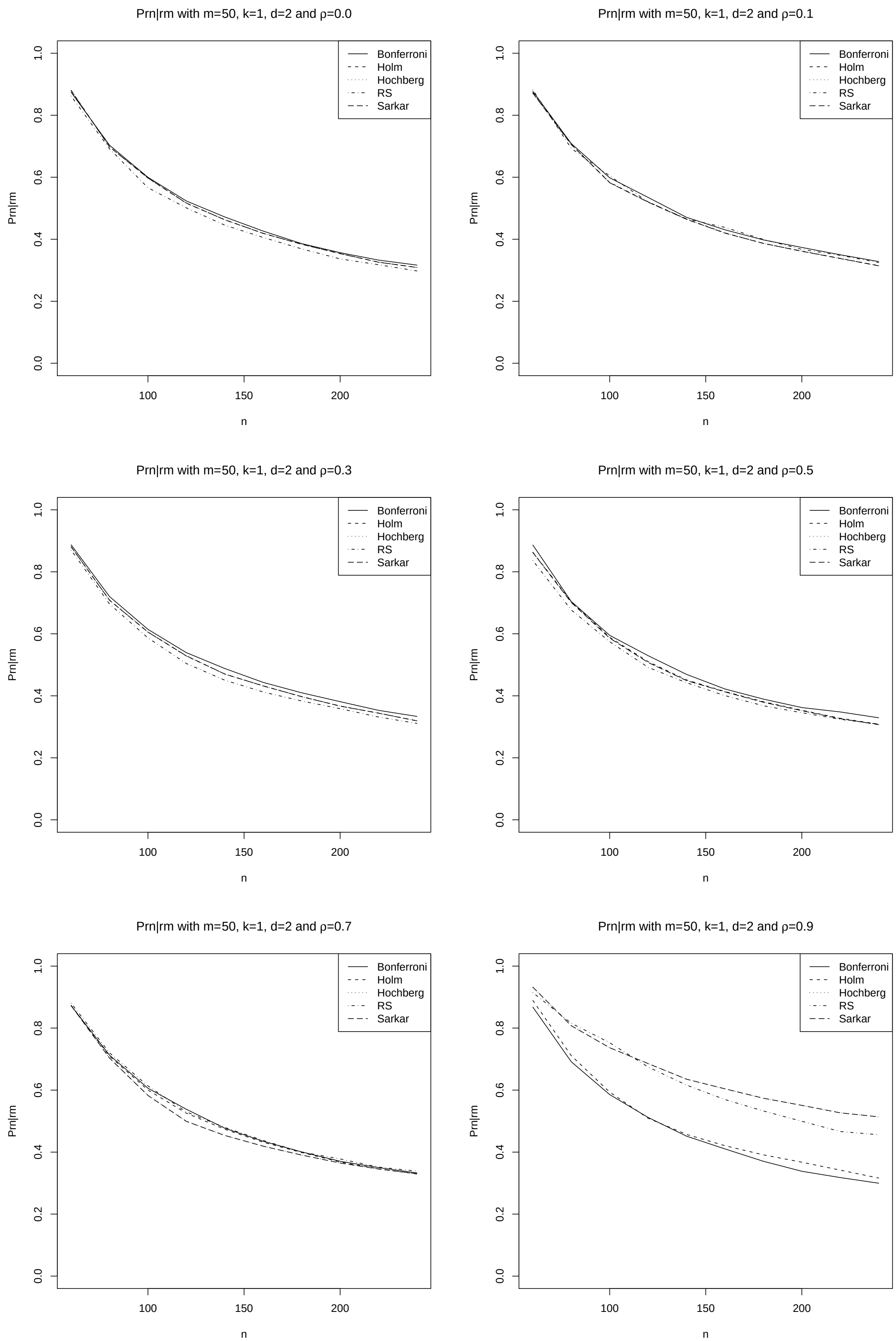

Figure 5.26: The Type R enlargement FWR while FWER is controlled strongly with $k = 1$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

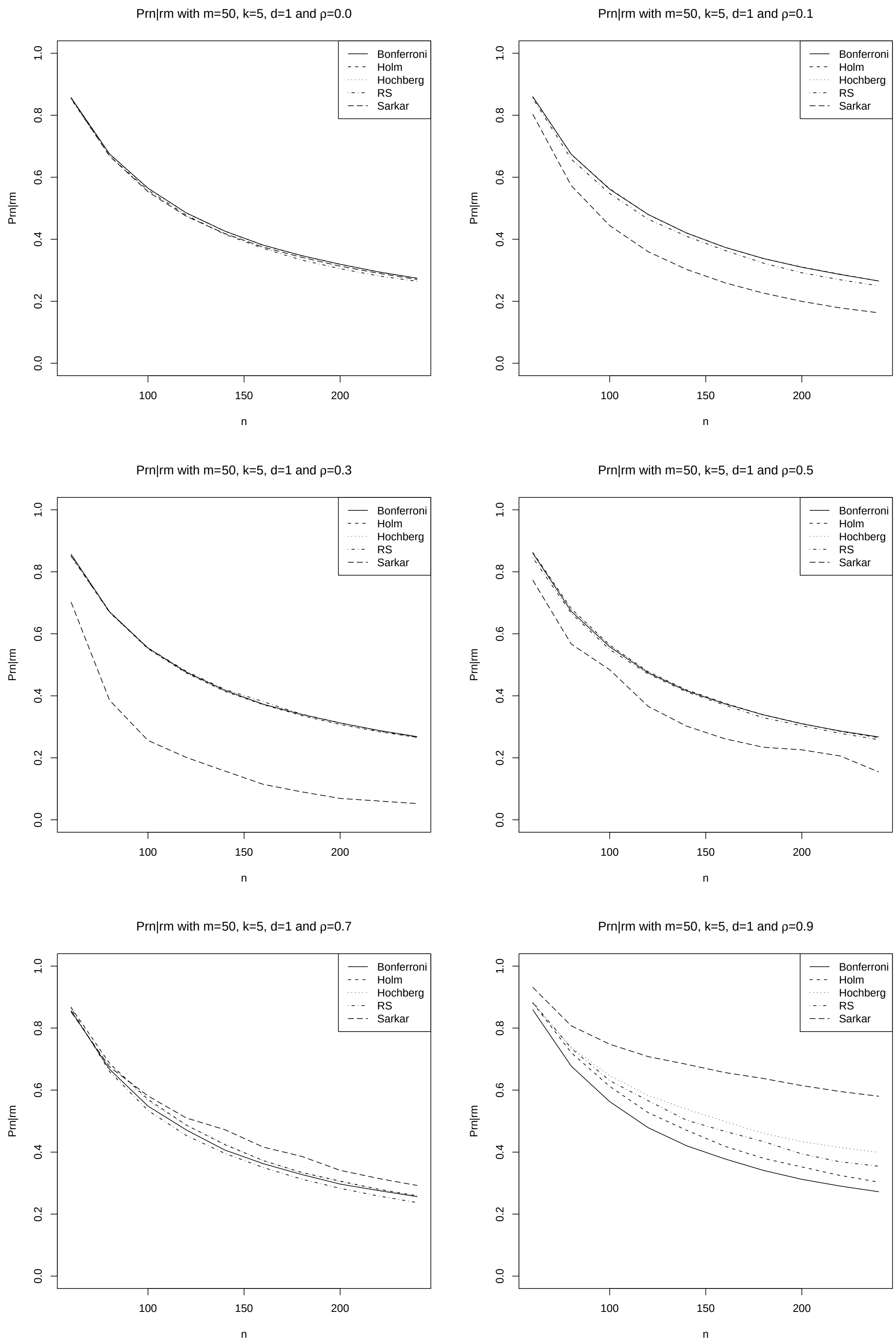

Figure 5.27: The Type R enlargement FWR while FWER is controlled strongly with $k = 5$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

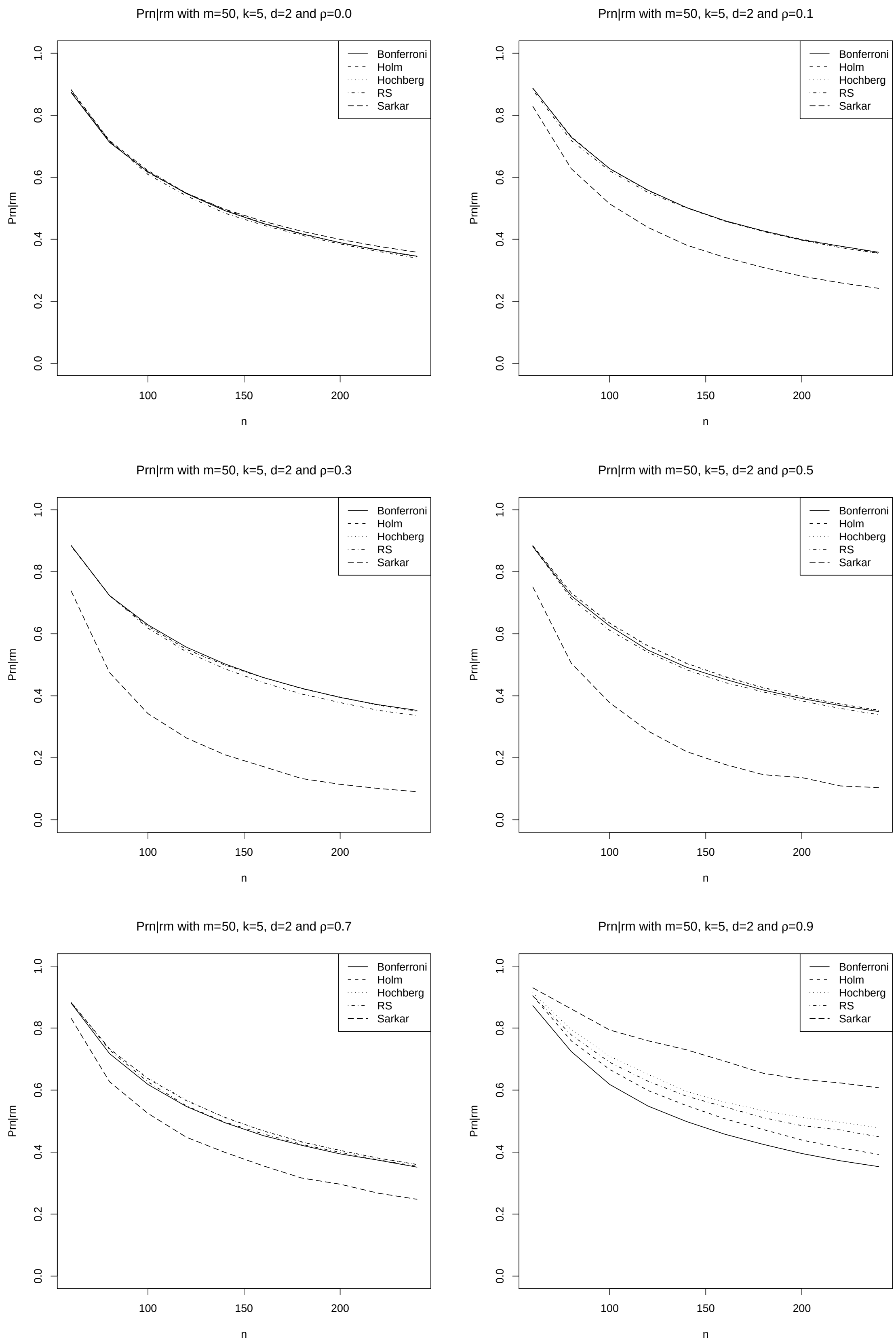

Figure 5.28: The Type R enlargement FWR while FWER is controlled strongly with $k = 5$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

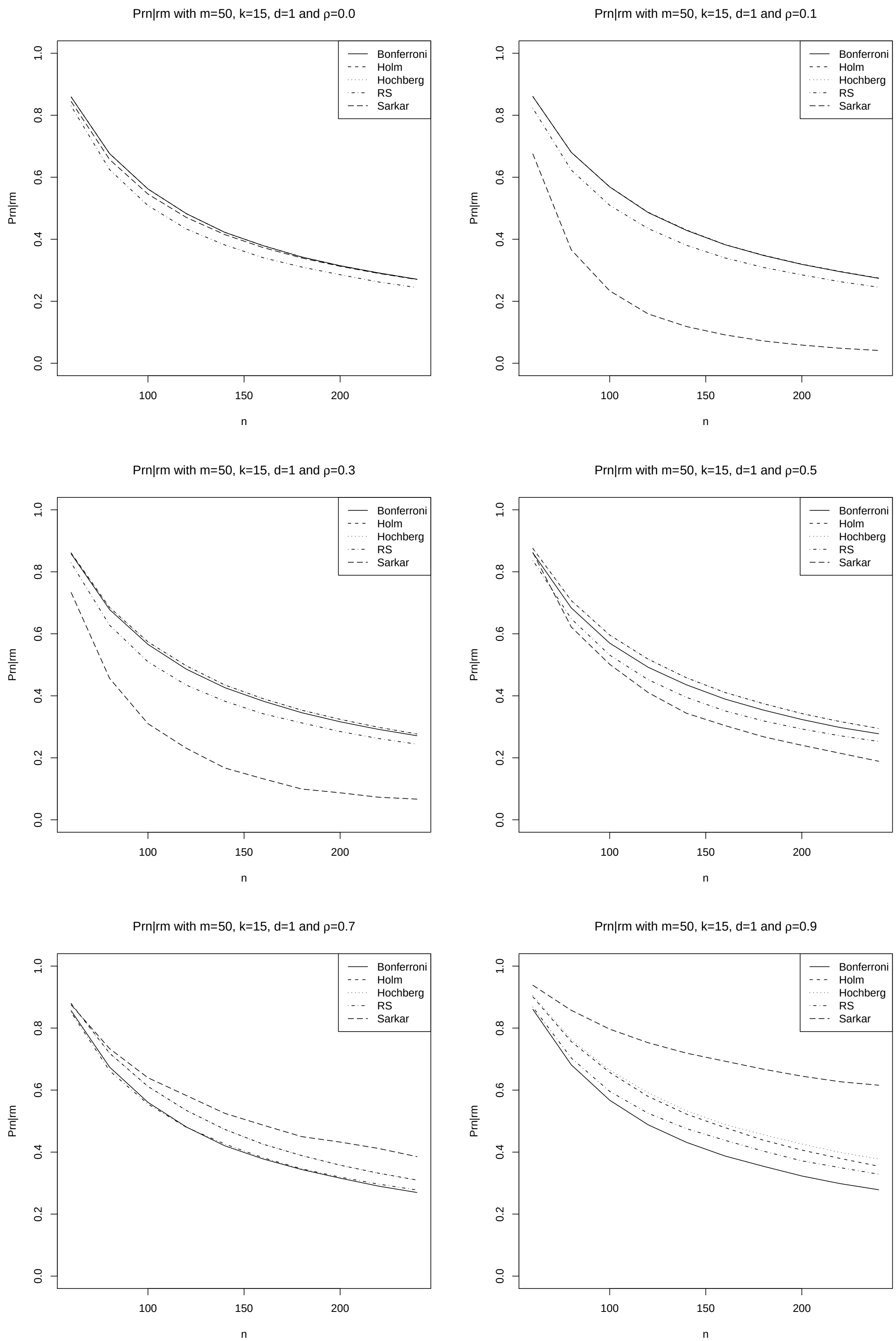

Figure 5.29: The Type R enlargement FWR while FWER is controlled strongly with $k = 15$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

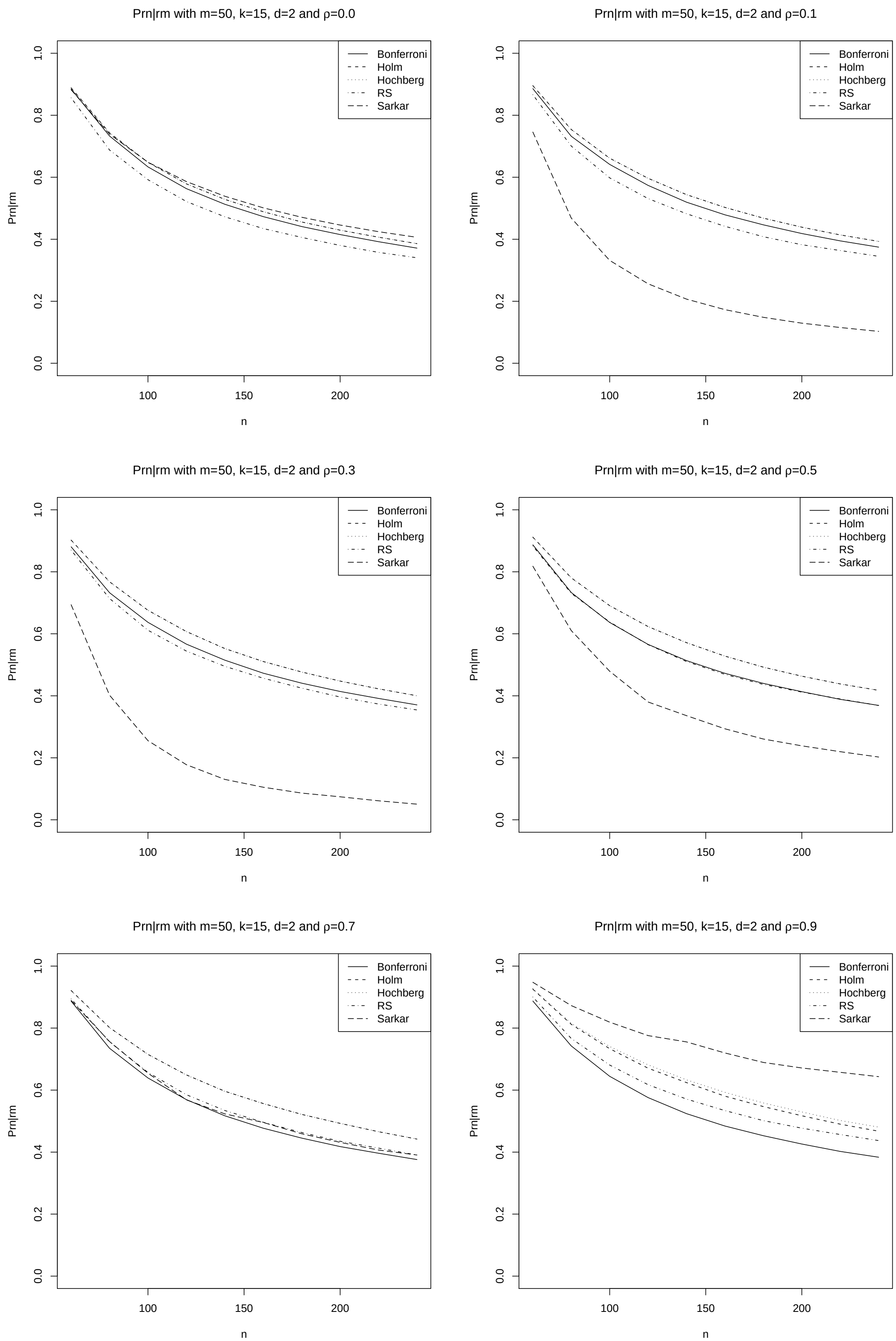

Figure 5.30: The Type R enlargement FWR while FWER is controlled strongly with $k = 15$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

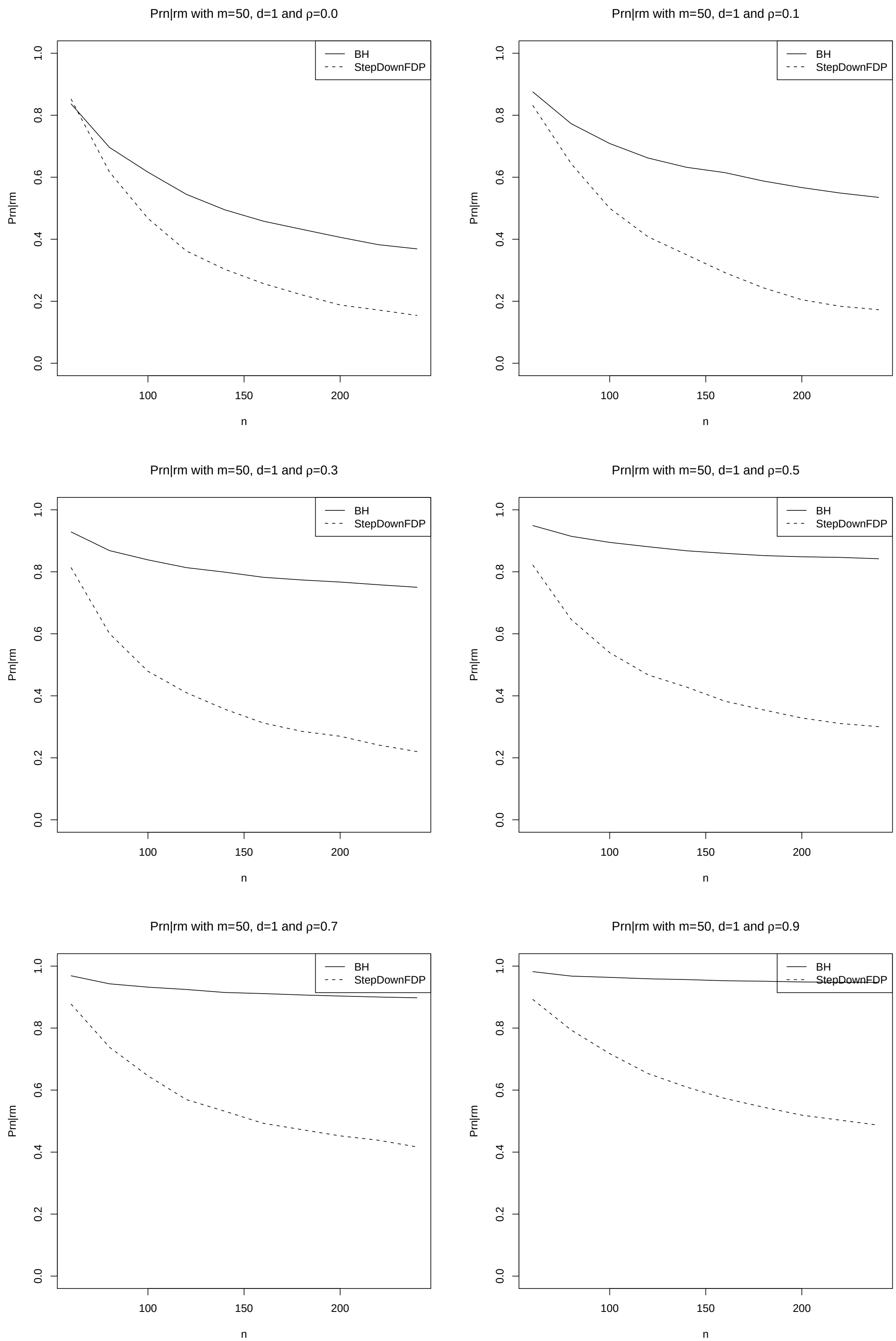

Figure 5.31: The Type R enlargement FWR for BH and StepDownFDP procedures with a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

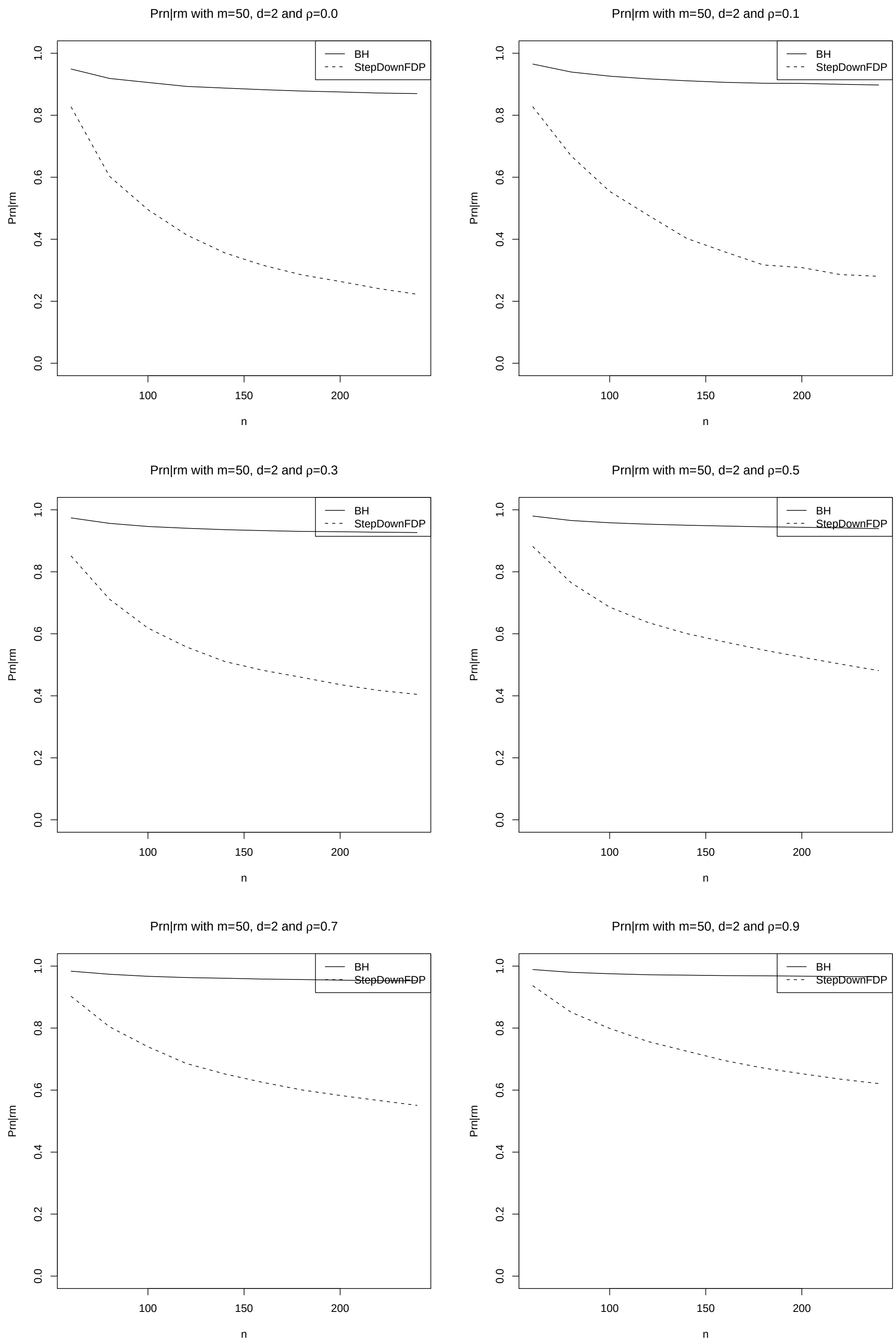

Figure 5.32: The Type R enlargement FWR for BH and StepDownFDP procedures with a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

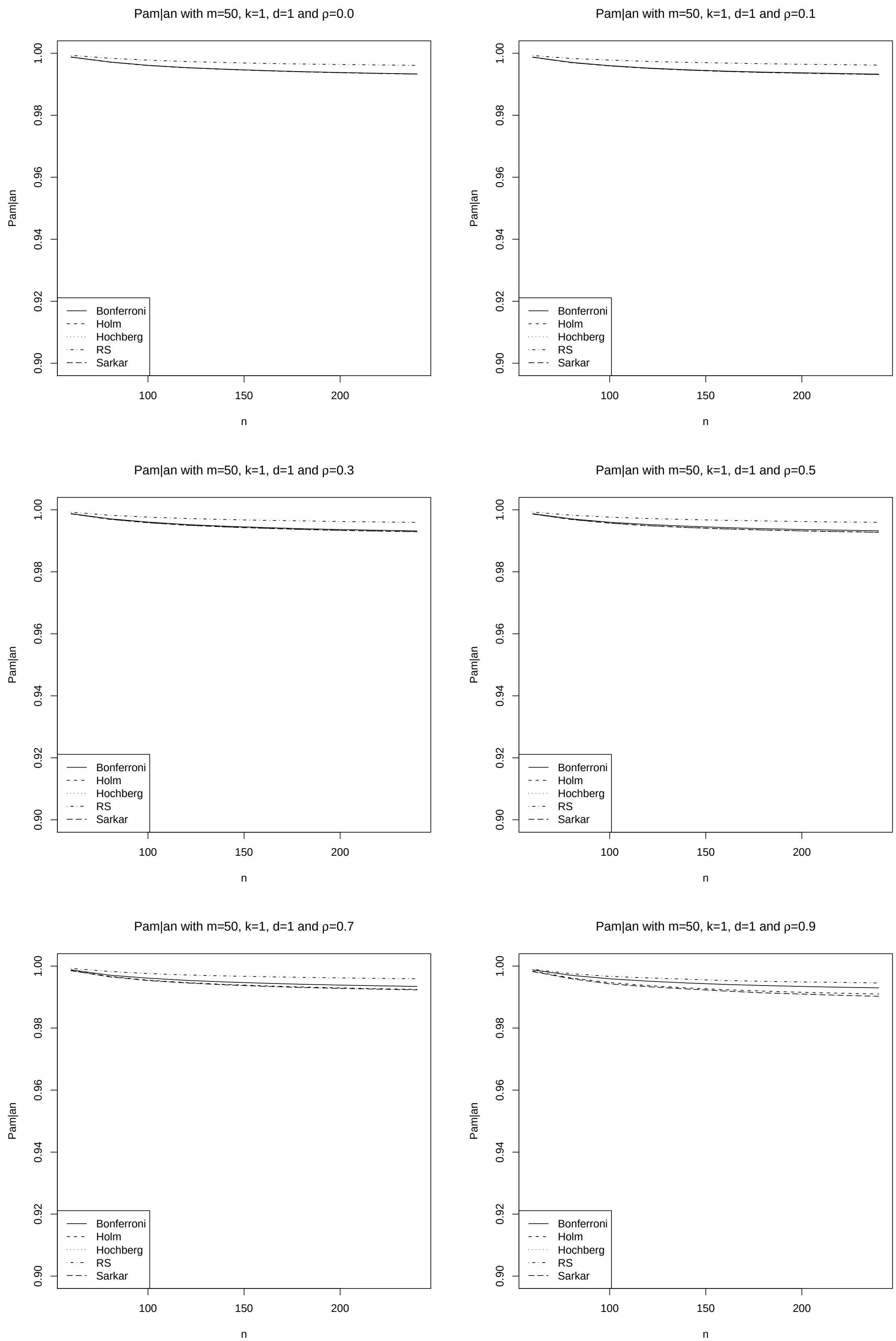

Figure 5.33: The Type A contraction FWR while FWER is controlled strongly with $k = 1$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

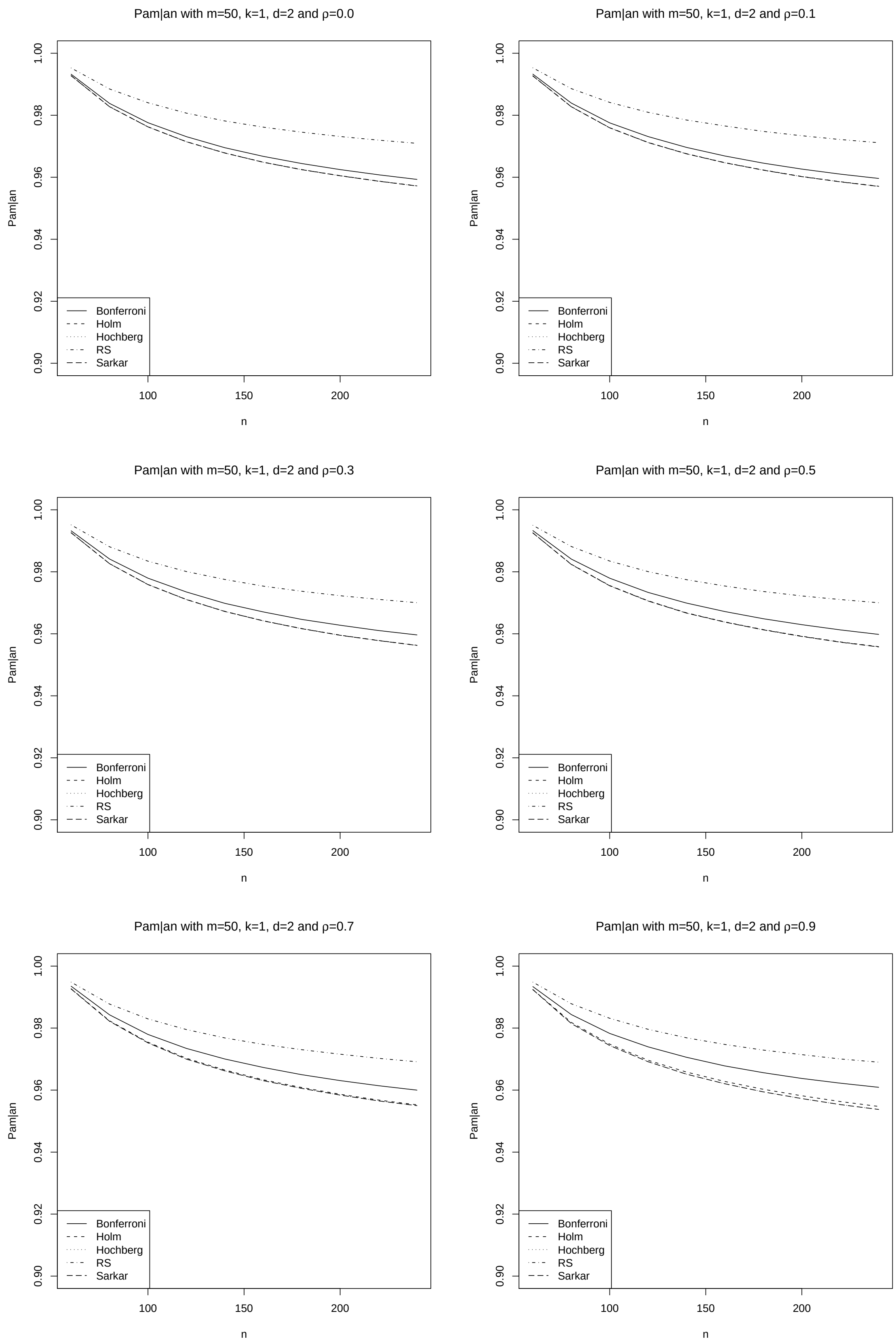

Figure 5.34: The Type A contraction FWR while FWER is controlled strongly with $k = 1$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

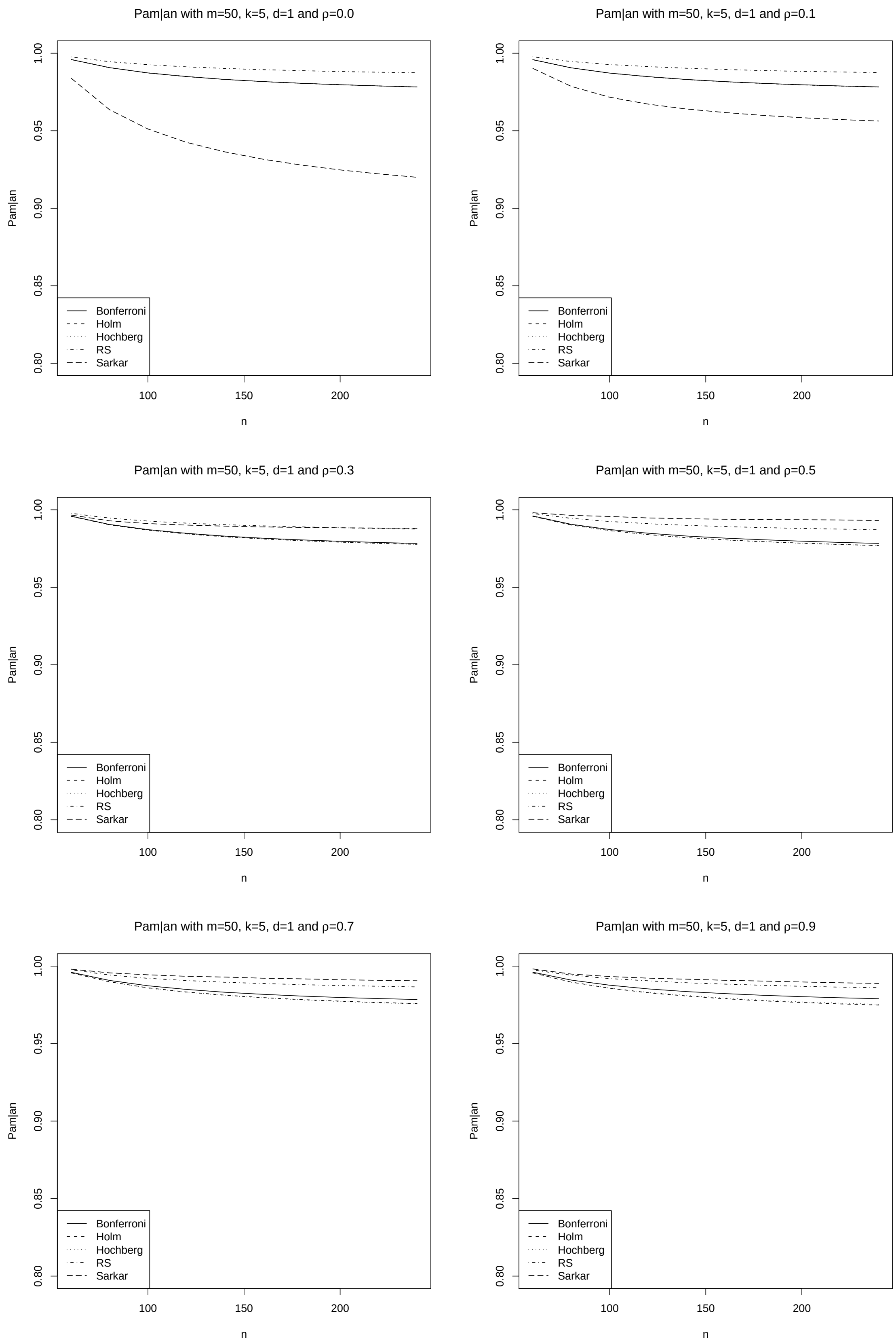

Figure 5.35: The Type A contraction FWR while FWER is controlled strongly with $k = 5$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

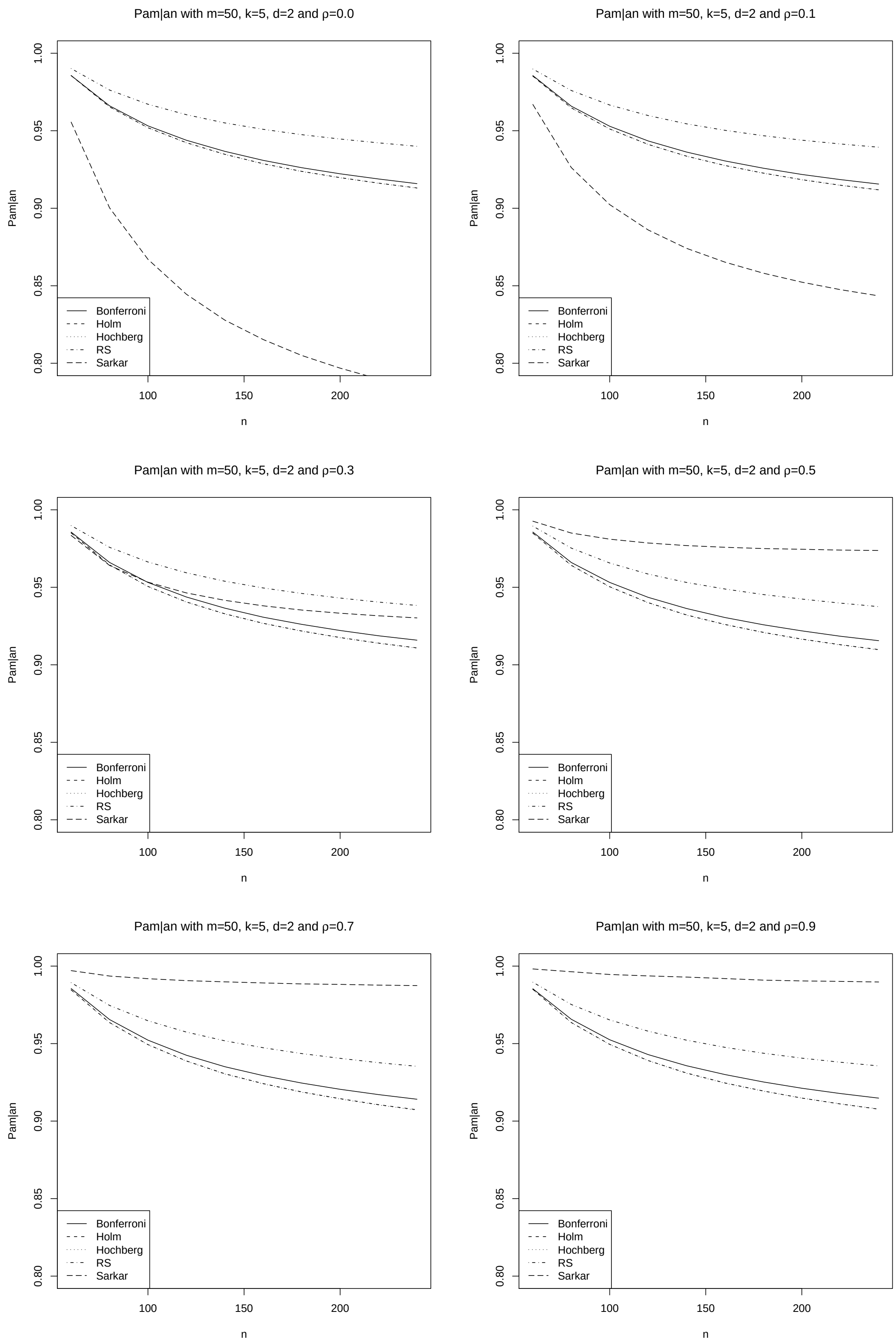

Figure 5.36: The Type A contraction FWR while FWER is controlled strongly with $k = 5$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

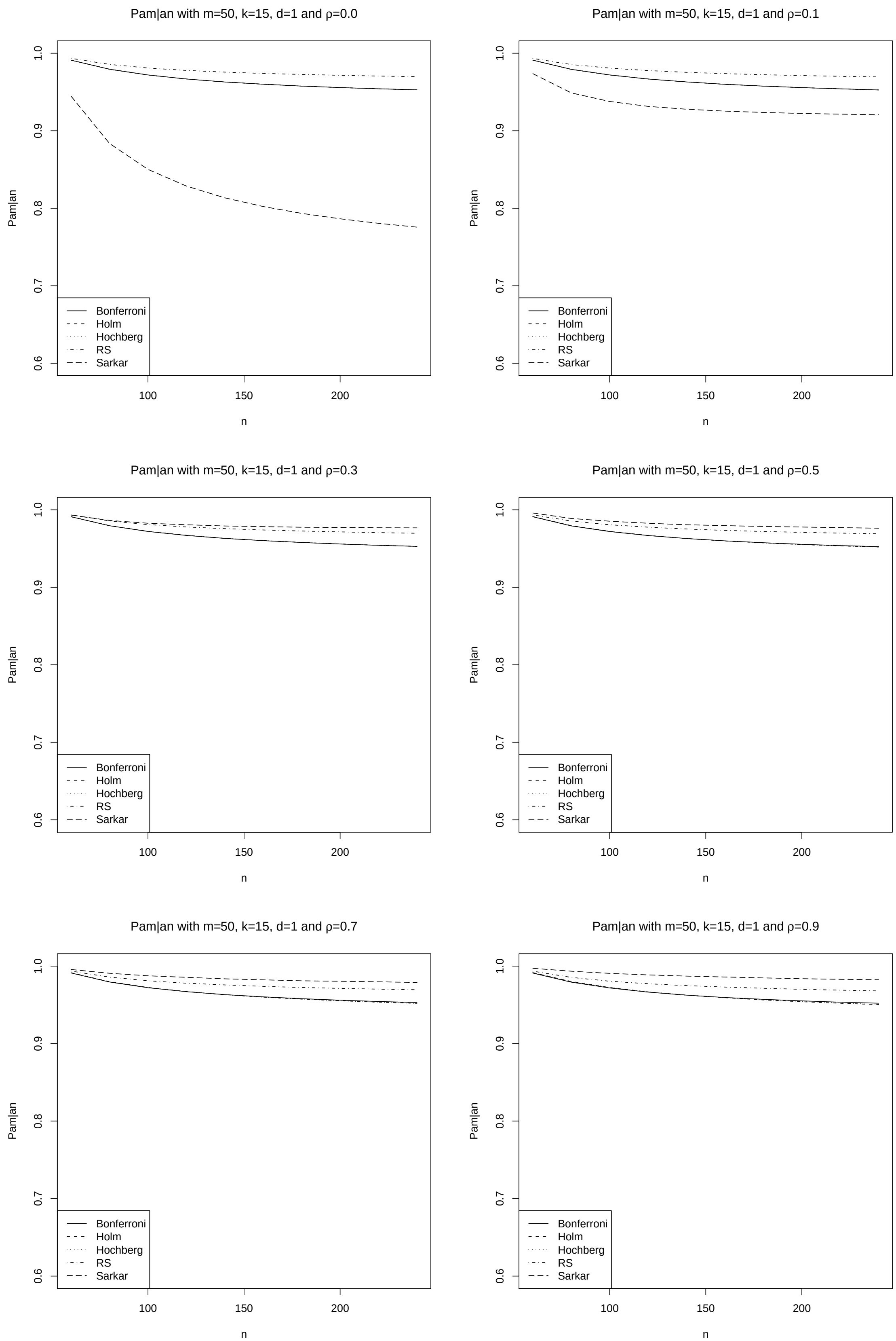

Figure 5.37: The Type A contraction FWR while FWER is controlled strongly with $k = 15$ and a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

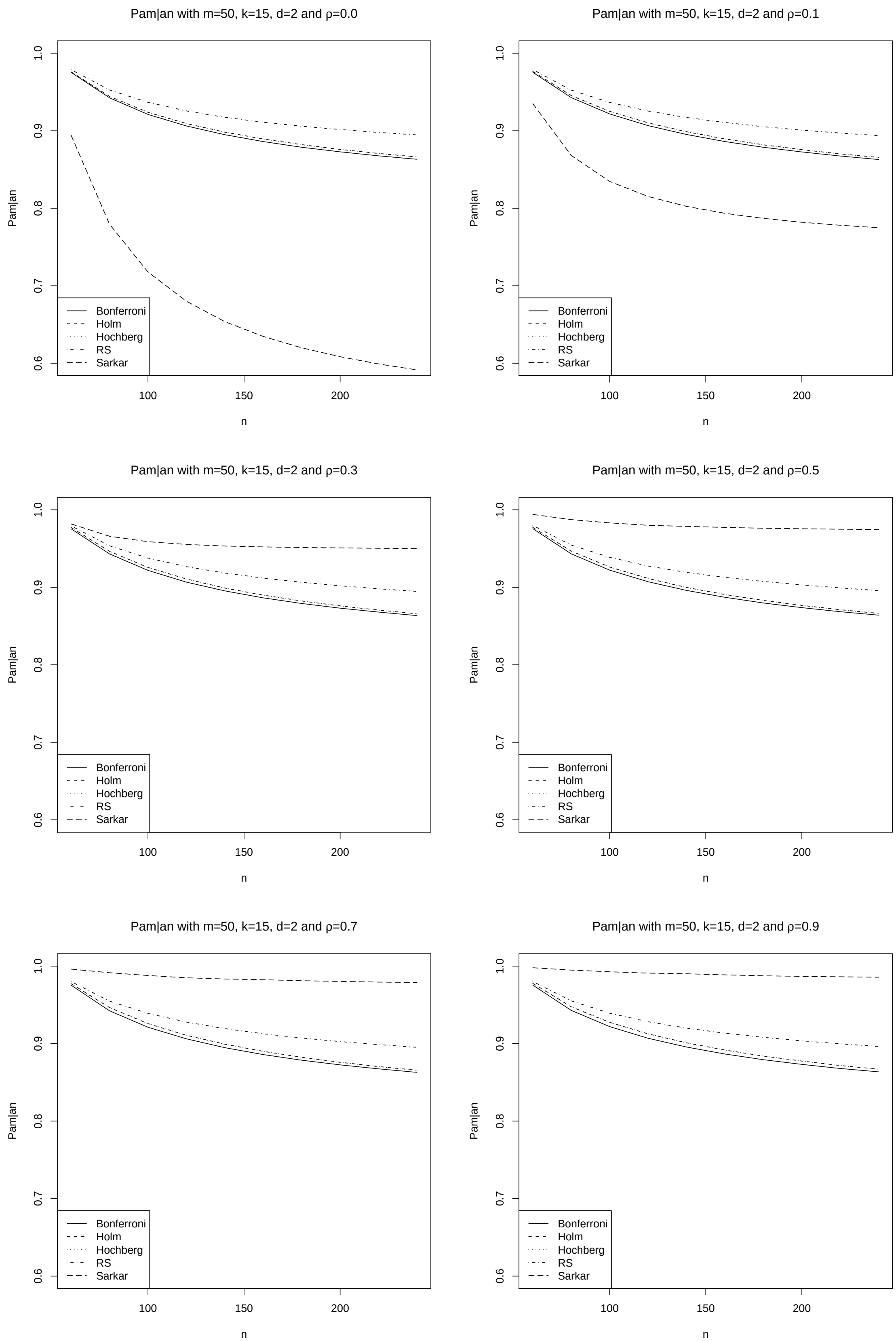

Figure 5.38: The Type A contraction FWR while FWER is controlled strongly with $k = 15$ and a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

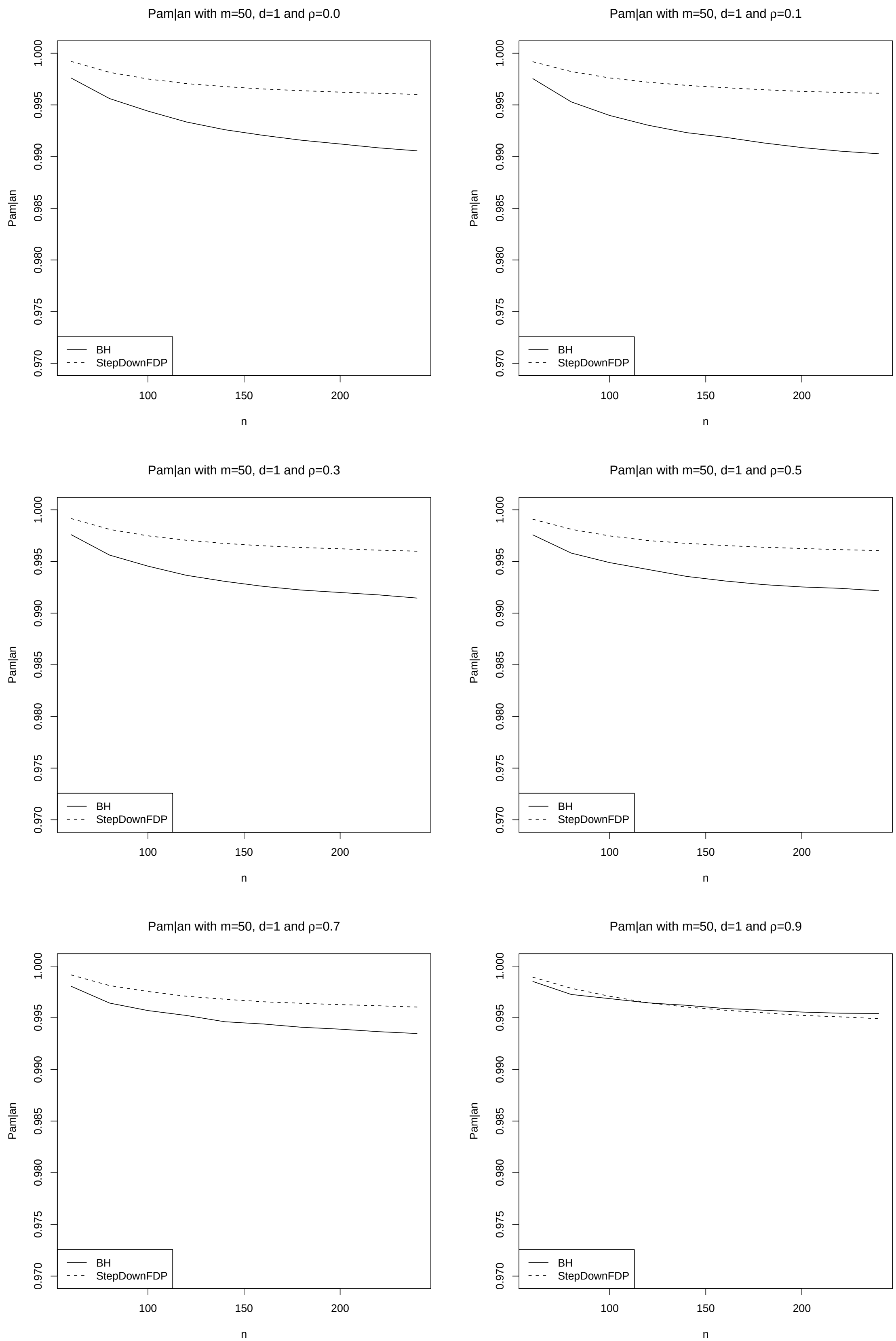

Figure 5.39: The Type A contraction FWR for BH and StepDownFDP procedures with a mean difference, d, of 1 for the test statistics corresponding to the false null hypotheses

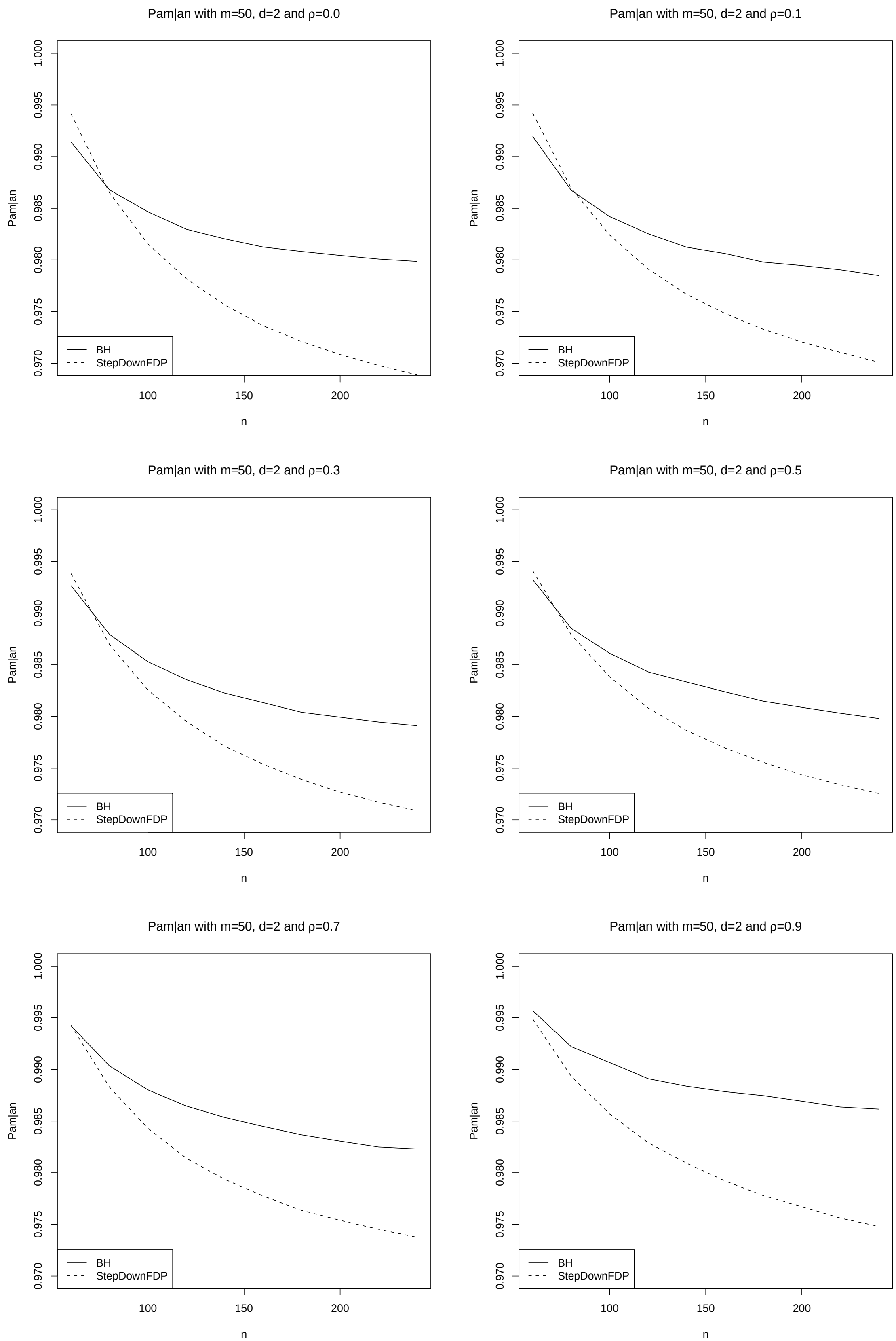

Figure 5.40: The Type A contraction FWR for BH and StepDownFDP procedures with a mean difference, d, of 2 for the test statistics corresponding to the false null hypotheses

CHAPTER 6

Conclusion

The focus of my dissertation was to investigate how the results of newer multiple testing procedures are affected by the size of a chosen family of hypotheses. Ideally, one would like the significant findings from multiple testing procedures to be the same regardless of the size of the chosen family. Although this property is unattainable, the robustness of such was studied with the familywise robustness measures as developed by Holland and Cheung (2002). Holland and Cheung looked at multiple testing procedures that controlled either familywise error rate (FWER) or false discovery rate (FDR). Since then, newer measures have been developed that control generalized familywise error rate (k-FWER) and false discovery proportion (FDP). This dissertation uses the measures of Holland and Cheung to measure the familywise robustness of these newer procedures.

When the test statistics were independent and all of the null hypotheses were assumed to be true, theoretical expressions for the Type R enlargement and Type A contraction FWR measures were derived using the expected Type I error for the multiple testing procedures. The expressions for the expected Type I error were obtained from Finner and Roters (2002). Finner and Roters suggested that a large expected number of Type I errors might be an indicator for good power performance. Because the Sarkar procedure yielded a larger expected number of Type I errors, especially as k became larger, one would expect this procedure to be more powerful than other k-FWER procedures when the test statistics were independent. The Bonferroni method had a Type R enlargement FWR which was equal to m/n. All of the procedures considered except the BH procedure were found to have an expected number of Type I errors which is nonincreasing in n. So their Type R enlargement FWR were all less than the Type R enlargement for the Bonferroni procedure. The BH procedure however has a expected number of Type I errors which is nondecreasing in n. This causes the Type R enlargement FWR for the BH procedure to be greater than that for the Bonferroni and all other procedures considered.

In the simulations, I wanted to see the effect of positive correlation on the expected number of Type I errors and the Type R enlargement FWR. In general, I saw no correlation effect for single-step and step-down procedures. In the equicorrelated case, the expected number of Type I errors and the Type R enlargement FWR for the step-up procedures were affected as correlation became larger. With clumpy dependence, correlation only existed within the blocks and there was no correlation between the blocks. The BH procedure was the only procedure in which correlation affected the expected number of Type I errors or Type R enlargement FWR in this case. The magnitude of this effect was much smaller than it was in the equicorrelated case.

I believe that the Type R enlargement FWR is the most important familywise robustness measure to use in the decision regarding which multiple testing procedure is the most robust. Based on the theoretical results and those provided in the simulations, I would recommend the Benjamini and Hochberg (1995) procedure. The BH procedure was consistently among the top procedures for Type R enlargement FWR. The BH procedure controls FDR and requires that the test statistics are multivariate totally positive of order two (MTP_2), so it might not be appropriate to use in all circumstances. When the BH procedure is not appropriate to use, I would recommend the Sarkar (2008) procedure if the test statistics are independent and $2 \leq k \leq 1/\alpha$. Under these conditions, the Sarkar procedure is more powerful than the other k-FWER procedures considered. Otherwise, I would suggest using the generalized Hochberg procedure by Lehmann and Romano (2005).

REFERENCES

[1] S.W. Ahmed. Issues Arising in the Application of Bonferroni Procedures in Federal Surveys. *American Statistical Association*, 1991.

[2] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] Y. Benjamini and Y. Hochberg. On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60, 2000.

[4] C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[5] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[6] S. Dudoit and MJ van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Springer Verlag, 2008.

[7] S. Dudoit, M.J. van der Laan, and K.S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1040, 2004.

[8] C.W. Dunnett and A.C. Tamhane. A step-up multiple test procedure. *Journal of the American Statistical Association*, 87(417):162–170, 1992.

[9] M.L. Ettredge, S.Y. Kwon, D.B. Smith, and P.A. Zarowin. The Impact of SFAS No. 131 Business Segment Data on the Market's Ability to Anticipate Future Earnings. *The Accounting Review*, 80(3):773–804, 2005.

[10] H. Finner and M. Roters. On the limit behaviour of the joint distribution function of order statistics. *Ann. Inst. Statist. Math.*, 46(2):343–349, 1994.

[11] H. Finner and M. Roters. Multiple hypotheses testing and expected number of type I errors. *Ann. Statist*, 30(1):220–238, 2002.

[12] W. Guo and J. Romano. A Generalized Sidak-Holm Procedure and Control of Generalized Error Rates under Independence. *Statistical Applications in Genetics and Molecular Biology*, 6(1):3, 2007.

[13] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[14] B. Holland and S.H. Cheung. Familywise robustness criteria for multiple-comparison procedures. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(1):63–77, 2002.

[15] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[16] E.L. Korn, J.F. Troendle, L.M. McShane, and R. Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398, 2004.

[17] E.L. Lehmann and J.P. Romano. Generalizations of the familywise error rate. *Ann. Statist*, 33(3):1138–1154, 2005.

[18] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[19] J.P. Romano and A.M. Shaikh. On stepdown control of the false discovery proportion. *Optimality: The Second Erich L. Lehmann Symposium*, 2006.

[20] J.P. Romano and A.M. Shaikh. Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist*, 34(4):1850–1873, 2006.

[21] S. Sarkar. Generalizing Simes' test and Hochberg's stepup procedure. *Ann. Statist*, 36(1):337–363, 2008.

[22] JP Shaffer. Multiple Hypothesis Testing. *Annual Reviews in Psychology*, 46(1):561–584, 1995.

[23] J.P. Shaffer. Recent developments towards optimality in multiple hypothesis testing. *Optimality: The Second Erich L. Lehmann Symposium*, 2006.

[24] J.D. Storey. Comment on 'Resampling-based multiple testing for DNA microarray data analysis' by Ge, Dudoit, and Speed. *Test*, 12(1):52–60, 2003.

[25] P.H. Westfall, R.D. Tobias, D. Rom, R.D. Wolfinger, and Y. Hochberg. *Multiple Comparisons and Multiple Tests*. SAS Publishing, 1999.