Unequal sample sizes according to the square-root allocation rule are useful when comparing several treatments with a control

Short running title: Sample sizes for comparisons with a control

Markus Neuhäuser[a*], Malwina M. Mackowiak[a], Graeme D. Ruxton[b]

[a] Department of Mathematics and Technology, RheinAhrCampus, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, D-53424 Remagen, Germany

[b] School of Biology, University of St Andrews, Dyer's Brae House, St Andrews KY16 9TH, UK

[*] Corresponding author (neuhaeuser@rheinahrcampus.de)

Abstract: A common situation in experimental science involves comparing a number of treatment groups each to a single reference (control group). For example, we might compare diameters of fungal colonies subject to a range of inhibitory agents to those from a control group to which no agent was applied. In this situation the most commonly applied test is Dunnett's test, which compares each treatment group separately to the reference whilst controlling the experiment-wise type I error rate. For analyses where all groups are treated equivalently statistical power is generally optimised by dividing subjects equally across groups. Researchers often still use balanced groups in the situation where a single reference group is compared to each of the others. In this case it is in fact optimal to spread subjects unequally: with the reference group getting a higher number of subjects (n_0) than each of the k treatment groups (n in each case). It has been previously suggested that a simple rule of thumb, the so-called square-root allocation rule $n_0 = \sqrt{k}\, n$ offers better power than a balanced design, without necessarily being optimal. Here we show that this simple-to-apply rule offers substantial power gains (over a balanced design) over a broad range of circumstances, and that the more-challenging-to-calculate optimal design often only offers minimal extra gain. Thus, we urge researchers to

consider using the square-root allocation rule whenever one control group is compared with a number of treatments in the same experiment.

Keywords: Dunnett’s test, power, sample size, square-root allocation rule, unbalanced samples

Introduction

When investigating a specified group, generally a control, and *k* other groups, called treatment groups hereafter, often the control group is compared with each of the other groups. In this many-to-one situation, Dunnett's test (Dunnett, 1955) is recommended for normally distributed data. Generalized Dunnett tests exist for other types of data, e.g. there are analogues for proportions and nonparametric approaches (Hothorn, 2016). However, here we focus on the original test proposed by Dunnett (1955). This test controls the experiment-wise type I error rate.

Dunnett’s test is based on *t* test statistics. However, the pooled variance estimate is based on data from all groups and the correlation between the *t* statistics is considered (see e.g. Bretz et al, 2011, pp. 71-75). Thus, the method is more powerful than a Bonferroni adjustment made after multiple two-sample tests (Bretz et al, 2011, p. 74).

When applying Dunnett’s test, appropriately chosen unbalanced sample sizes can give a more powerful test than a balanced design, even when there is no difference in variability between groups. Here, we seek to promote greater awareness of this. Groups of equal size are sometimes suggested very generally, see e.g. Curtis et al. (2018). Neuhäuser and Ruxton (2018) mentioned three situations where unequal sample sizes might be useful: (1) unequal variances between groups, (2) situations where one treatment involves more potential for suffering (or higher expense), and (3) the many-to-one situation considered here.

For Dunnett’s test, Dunnett (1955) recommended the square-root sampling allocation rule. Hence, this rule is not new, it is mentioned in textbooks (e.g. Hochberg and Tamhane, 1987;

Brock and Mounho, 2014; Rosenberger and Lachin, 2015; Green et al., 2018; Kieser, 2020). Nevertheless, the rule is rarely used. We searched for Dunnett's test in this journal *Ethology*. We found 14 papers since 2000, none of them applied or even mentioned the square-root allocation rule.

In the scenario where the control group is compared to each of the treatment groups, the control group plays a special role making it plausible to enlarge its sample size compared to each of the other groups (Hochberg and Tamhane, 1987). The design of the square-root allocation involves the same sample size n for each of the k treatment groups, but a larger sample size for the control group, namely $n_0 = \sqrt{k}n$. This design gives a test with higher power, or a test with the same power but a lower total sample size, than a design with equal group sizes $N/(k + 1)$ for every group (Liu 1997), where $N = n_0 + kn$ denotes the total sample size.

In order to illustrate the square-root allocation rule let's consider an example with $k = 4$ treatment groups, a control group and $N = 60$ observations in total. Here, $\sqrt{k} = 2$, thus the sample size of the control group is $n_0 = 2n$, twice the sample size of each treatment group. Hence, with $N = 60$ we have $n = 10$ observations per treatment group and $n_0 = 20$ observations in the control group, instead of an equal group size of $N/5 = 12$ observations for all $k + 1$ groups.

Methods

Here, we demonstrate the benefit of square-root allocation (SRA) in comparison to equal size allocation (ESA) based on a simulation study performed in R (version 4.0.4).

We do this on the basis of samples drawn from normal distributions. The standard deviation of these distributions is always one, but the mean value for the control group μ_0 is set to zero while a non-zero positive value μ_i is used for all of the k treatment groups. However, this mean value is the same for all treatment groups in our simulation. We explore the ability of Dunnett's test to detect this difference between the control and treatment groups as a function of total sample size

N. The values used for the treatment group means are given in Table 1, we make the value of μ_i lower as total sample size increases (since total sample size and this effect size have opposite effects on statistical power). We estimate statistical power as the fraction of 10,000 replicate simulations where the Dunnett's test detected the underlying difference in distributions, based on a nominal experiment-wise type I error rate of 0.05. The R code used for our simulations is available at www.hs-koblenz.de/profilepages/neuhaeuser/programme.

As well as SRA and ESA, we also considered the optimal allocation (OA), that is, the allocation that maximizes the power without changing the total sample size. The above-mentioned simulations were also used to search for the OA. To be precise, we determined the power for all possible allocations with a specific total sample size *N*. The allocation with maximum power is the OA. A user could obtain the OA by simulation or by programming a search procedure such as the one presented by Kwong et al. (2010), see Liu (1997) for mathematical details. Thus, for a user it is much easier to apply the square-root allocation than search for the optimal allocation.

There are several alternative ways in which statistical power could be measured in the case of multiple tests (Bretz et al., 2011). For our situation with $\mu_i > \mu_0$ we report on two of these:

- the probability of correctly rejecting all *k* false null hypotheses $\mu_i = \mu_0$, called the conjunctive power, and
- the probability of correctly rejecting at least one false null hypothesis $\mu_i = \mu_0$, called the disjunctive power.

That is, we report the fraction of replicate simulations where Dunnett's test suggests (correctly) that all treatment groups are different from the control, and the larger fraction where the test suggests that at least one of the treatment groups is different from the control.

In evaluating Dunnett's test we assume that we are dealing with a situation where negative values of μ_i are considered implausible and so one-sided testing is adopted.

Results and examples

The consistent pattern seen across the results shown in Table 2 and Figures 1 and 2 is that the square-root allocation rule consistently gives higher power values than equal sample sizes although the benefit is smaller for small numbers of treatment groups. Another consistent pattern is that the additional benefit due to the optimal allocation is very low in the scenarios we considered.

Having considered the theoretical advantages of SRA in an idealised simulation, we now want to explore its practical application. First, we re-examine a recent publication that used Dunnett's test (Adamo and McKee, 2017). In their figure 2, the number of eggs laid per day per cricket *Gryllus texensis* is shown for three groups with n = 40 per group; hence there are k = 2 treatment groups in their design. The maximum difference in means is around 6 eggs with a standard error of around 2. Thus, we can estimate a standard deviation of approximately $\sigma = 2\sqrt{40} \approx 12.6$. This standard deviation is approximately twice as large as the difference in means. If we take these values for the effect size and the standard deviation, we can estimate power in exactly the same way we did previously. We assume one-sided testing again, and a nominal experiment-wise type I error rate of 0.05.

For this scenario, with equal-size allocation (ESA) used by the original authors with 40 in each group the disjunctive power is 0.7682. With the square-root allocation (SRA), the sample sizes would be 35 for the two treatment groups and 50 for the control, and the disjunctive power is a little higher: 0.7916. The conjunctive power is 0.6102 for ESA and 0.6199 for SRA. From these calculations we see that adoption of SRA application would have offered these researchers more power, albeit only slightly more than with the ESA design that they used.

We can also use similar simulations to ask how much the sample size could be reduced by adopting SRA while still delivering the same power as adopting ESA. As mentioned above the disjunctive power is 0.7682 under ESA. If we require a power of at least 80% we would need 126 rather than 120 animals when applying ESA (i.e. 42 rather than 40 in each group), then the disjunctive power is 0.8040. With SRA 123 animals (51 in the control group and 36 in each treatment group) are sufficient for a power > 80%; with 51 in the control group and 36 in each of the other two groups, we predict disjunctive power of 0.8140. Thus, again we see in this example that with a low number of comparisons, since there are only two treatment groups – there is still a benefit to square-root allocation, but this benefit is small.

However, this benefit can be much larger when we have just a few more groups. We consider the hypothetical situation where the original study had two more treatment groups (four instead of two). Now with ESA we need a total sample size of 195 for a power > 80% (the estimated disjunctive power is 0.8211). With SRA a total sample size of 156 is sufficient for a power of 0.8020. If we require a power of 0.8211 (as for ESA), then 162 animals are sufficient (power = 0.8238). Thus, in this situation, SRA allows a big decrease in sample size without loss of power.

In order to illustrate how Dunnett's test can be applied in R we consider another example situation (using data presented by Dunnett, 1955, p. 1099). In this data-set, a blood count (millions of cells per cubic millimetre) was measured on three groups of animals, a control group (here group 0) and two groups treated with different drugs (groups 1 and 2). Dunnett's test is available with the function `glht` of the R package `multcomp`, and can be applied to this data as follows:

```
count <- c(7.4,8.5,7.2,8.24,9.84,8.32,9.76,8.8,7.68,9.36,12.8,9.68,
       12.16,9.2,10.55)
group <- as.factor(c(rep(0,6), rep(1,4), rep(2,5)))
anova.model <- aov(count ~ group)
library(multcomp)
summary(glht(anova.model, linfct = mcp(group = "Dunnett"), alternative =
       "greater"))
```

This yields the following output:

```
Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = count ~ group)

Linear Hypotheses:
             Estimate Std. Error t value  Pr(>t)
1 - 0 <= 0   0.6500     0.7584   0.857 0.32498
2 - 0 <= 0   2.6280     0.7115   3.694 0.00291 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

From this, we can infer that group 2 is significantly different from the control group (p = 0.0029), however, there is no significant difference between group 1 and the control (p = 0.3250). Simultaneous confidence intervals are computed when `summary` is replaced by `confint` in the R code above. Two-sided tests and confidence intervals can be obtained using `alternative = "two.sided"`. The use of the function `glht` is described in much more details by Bretz et al. (2011), including plots to display the results graphically. Bretz et al. (2011) also explain how Dunnett’s test can be combined with a closed testing procedure in order to increase power (see also Hayter and Tamhane, 1991).

Discussion

The square-root allocation, already proposed by Dunnett (1955), is rarely used. Instead equal group sizes dominate applications. However, the square-root allocation is convenient to implement, and increases power as shown above. It is also possible to obtain the same power with fewer experimental units, preferable due to ethical aspects for experiments with animals or humans. Sample size calculations might be performed using the web application developed by Grayling and Wason (2020), or by simulation (Colegrave and Ruxton, 2021).

Although the square-root allocation seldom yields exactly the optimal sampling allocation, it provides a reasonable approximation of the optimal allocation (Kwong et al., 2010). Kwong et al. (2010) presented a search procedure to obtain an optimal design, and they recommend its application if the cost of additional observations is relatively high. However, as shown here, the difference between the square-root allocation and the optimal allocation is often very small.

Our simulations only considered the case where underlying assumptions of Dunnett's test of normally distributed data and homogeneity of variance hold. However, the advantage of SRA over ESA even holds when variances increase in some treatment groups (Brock and Mounho, 2014). It is true that problems of unreliability due to heterogeneity of variance can be amplified by unequal sample sizes (Hothorn, 2016), but even the well-studied ANOVA F-test is known not to be always robust to variance heterogeneity when sample sizes are equal (Rogan and Keselman, 1977).

When a researcher has reason to question whether the homogeneity of variances is likely to hold in their system then switching to the robust procedure proposed by Herberich et al. (2010) seems useful. For this method no assumptions regarding distribution or variance homogeneity are necessary.

In this article we focus on the many-to-one situation where several treatment groups were compared to a control. There are other options, for example so-called Helmert contrasts which compare each group to the mean of preceeding groups. This approach might be useful when several dose groups are investigated (Hothorn, 2016). Helmert contrasts are orthogonal (i.e. uncorrelated) which makes the underlying computations numerically less complex. However, the research question should dictate which statistical tests to apply. And with the above-mentioned function `glht` of the R package `multcomp`, non-orthogonal contrasts can easily be handled. Nevertheless, we would like to note that the SRA also provides a reasonable approximation to the OA for the orthogonal contrasts investigated by Hayter and Tamhane (1991).

We recommend routine use of SRA in situations where several treatments are compared to a single reference group. This design requires trivial extra effort implement, and always offers some power benefit. Any benefit should be attractive, but the benefits can be substantial when the number of treatment groups is larger. The SRA may not be optimal, but our simulations suggest it may be recommended as “near optimal”.

All authors declare that there are no conflicts of interest. There was no funding for this manuscript. No animals were used for our study.

Ethics approval was not required for this research.

Acknowledgement: We thank the editor and two referees for useful comments.

References

Adamo SA, McKee R (2017). Differential effects of predator cues versus activation of fight-or-flight behaviour on reproduction in the cricket *Gryllus texensis*. Animal Behaviour 134: 1-8.

Bretz F, Hothorn T, Westfall P (2011). Multiple comparisons using R. CRC Press, Boca Raton, FL.

Brock WJ, Mounho B (2014). The role of the study director in nonclinical trials. Wiley, Hoboken, NJ.

Colegrave N, Ruxton GD (2021). Power analysis: An introduction for the life sciences. Oxford University Press, Oxford, UK.

Curtis MJ, Alexander S, Cirino G, Docherty JR, George CH, Giembycz MA et al. (2018). Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers. Br J Pharmacol 175: 987–993.

Dunnett CW (1955). A multiple comparison procedure for comparing several treatments with a control. J Amer Stat Assoc 50: 1096-1121.

Grayling MJ, Wason JMS (2020). A web application for the design of multi-arm clinical trials. BMC Cancer 20: 80.

Green JW, Springer TA, Holbeck H (2018). Statistical analysis of ecotoxicity studies. Wiley, Hoboken, NJ.

Hayter AJ, Tamhane AC (1991). Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparisons with a control. Journal of Statistical Planning and Inference 27: 271-290.

Herberich E, Sikorski J, Hothorn T (2010). A Robust Procedure for Comparing Multiple Means under Heteroscedasticity in Unbalanced Designs. PLoS ONE 5(3): e9788.

Hochberg Y, Tamhane AJ (1987). Multiple comparison procedures. Wiley, New York, NY.

Hothorn LA (2016). Statistics in toxicology using R. CRC Press, Boca Raton, FL.

Kieser M (2020). Methods and applications of sample size calculation and recalculation in clinical trials. Springer, Cham, Switzerland.

Kwong KS, Cheung SH, Wen MJ (2010). Sample size determination in step-up testing procedures for multiple comparisons with a control. Statist Med 29: 2743-2756.

Liu W (1997). On sample size determination of Dunnett's procedure for comparing several treatments with a control. J Stat Plann Infer 62: 255–261.

Neuhäuser M, Ruxton GD (2018). Some comments on the update to BJP guidance on experimental design and analysis. Br J Pharmacol 175, 3638-3639.

Rogan JC, Keselman HJ (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Journal 14: 493-498.

Rosenberger WF, Lachin JM (2015). Randomization in clinical trials. Wiley, Hoboken, NJ.

Table 1: Mean differences between groups used in the simulation study depending on N

Total sample size N	Difference in means $\mu_i - \mu_0$
< 25	1.2
26 to 50	0.95
51 to 75	0.85
76 to 100	0.75
101 to 150	0.7
151 to 200	0.675
201 to 250	0.65
251 to 300	0.625

Table 2: Disjunctive and conjunctive power $1 - \beta$ of equal-size allocation (ESA) and square-root allocation (SRA) for different values of k and N [see next page]

Figure captions

Fig. 1: Disjunctive and conjunctive power of equal-size allocation (ESA) and square-root allocation (SRA) for different values of k and N

A: k = 2 treatment groups

B: k = 3 treatment groups

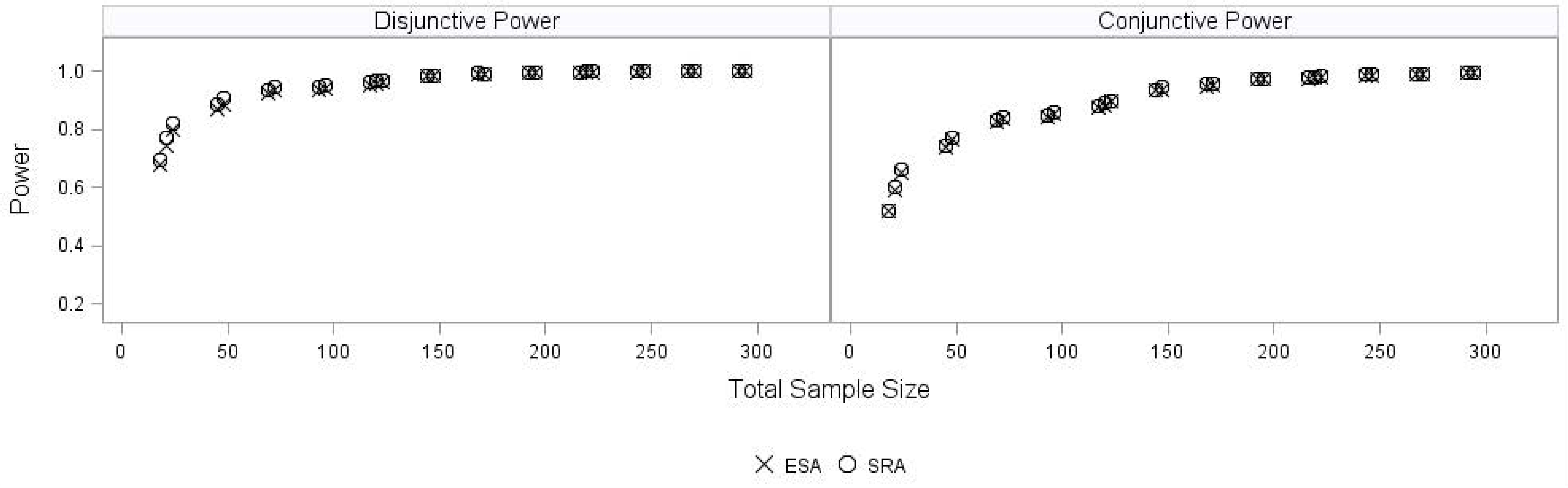
C: k = 4 treatment groups

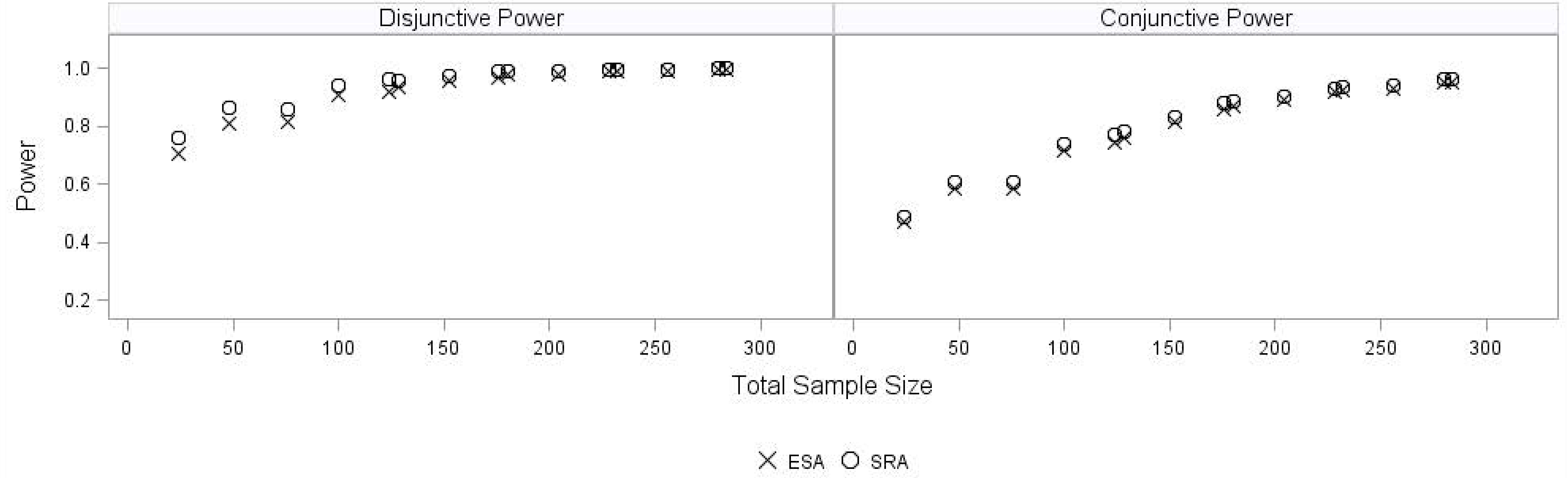
D: k = 5 treatment groups

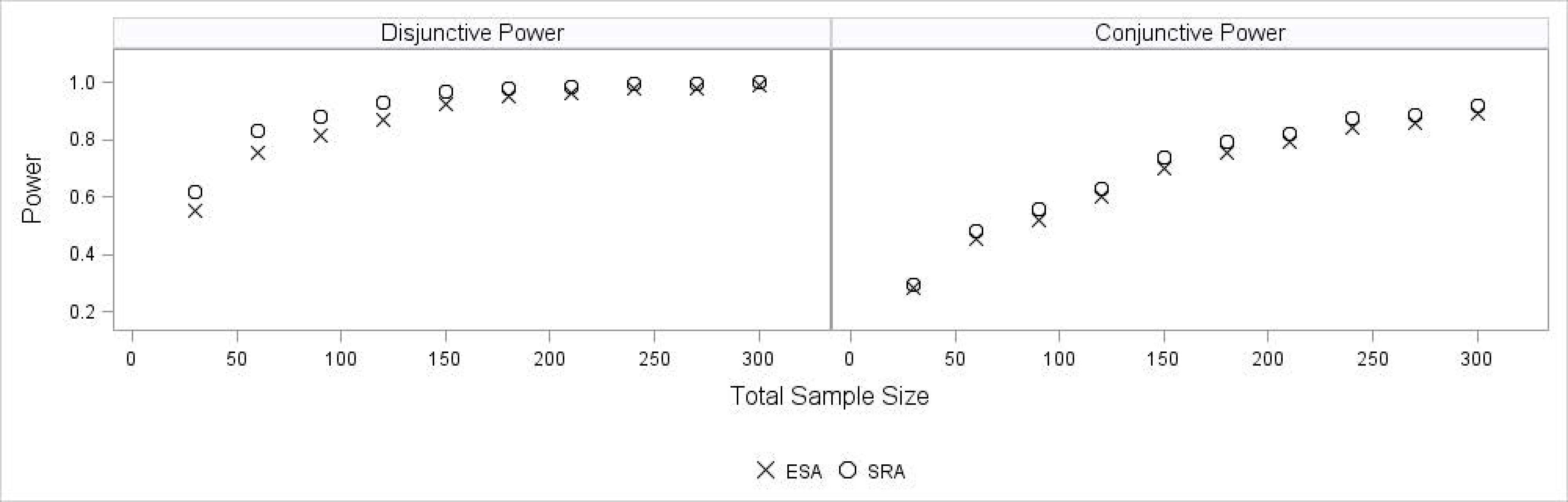
E: k = 6 treatment groups

Fig. 2: Comparison between disjunctive and conjunctive power for equal-size allocation (ESA), square-root allocation (SRA), and the optimal allocation (OA) for k = 5 and N = 102

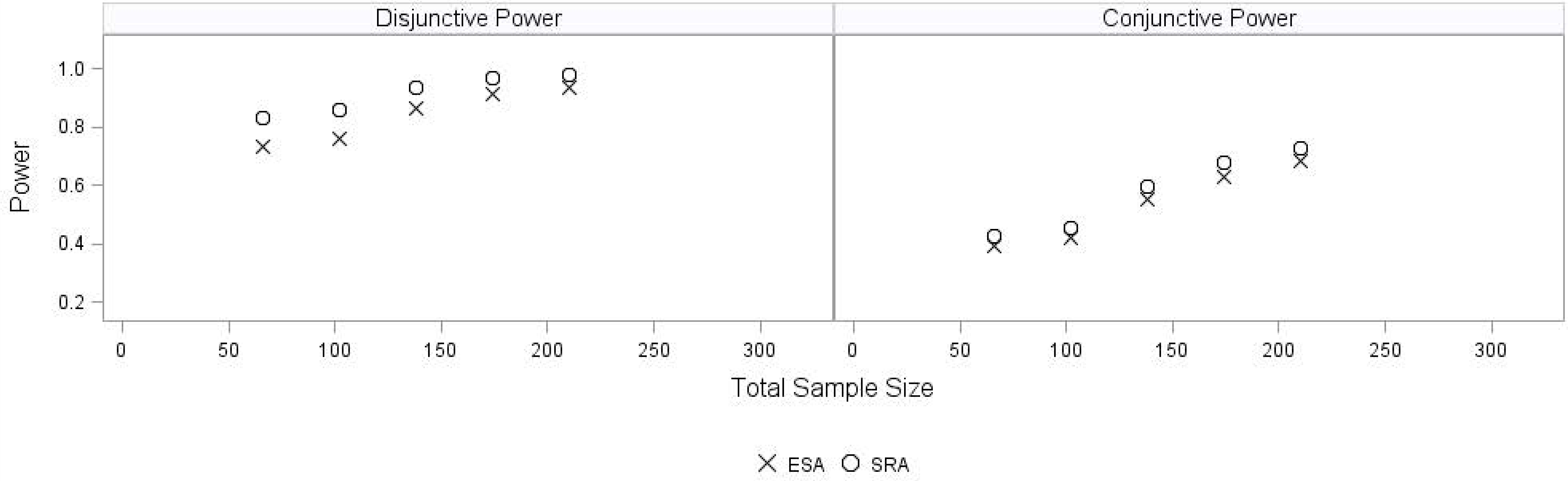
			Square root allocation				Equal size allocation			
k	N	μ_i	n_0	n	$1-\beta$		n_0	n	$1-\beta$	
					disjunctive	conjunctive			disjunctive	conjunctive
2	18	1.200	8	5	0.6930	0.5195	6	6	0.6760	0.5202
2	21	1.200	9	6	0.7696	0.6016	7	7	0.7449	0.5918
2	24	1.200	10	7	0.8197	0.6620	8	8	0.7998	0.6524
2	45	0.950	19	13	0.8837	0.7440	15	15	0.8675	0.7354
2	48	0.950	20	14	0.9048	0.7728	16	16	0.8879	0.7641
2	69	0.850	29	20	0.9350	0.8278	23	23	0.9218	0.8236
2	72	0.850	30	21	0.9438	0.8441	24	24	0.9341	0.8354
2	93	0.750	39	27	0.9479	0.8486	31	31	0.9371	0.8415
2	96	0.750	40	28	0.9523	0.8584	32	32	0.9426	0.8512
2	117	0.700	49	34	0.9639	0.8806	39	39	0.9506	0.8726
2	120	0.700	50	35	0.9659	0.8882	40	40	0.9585	0.8819
2	123	0.700	51	36	0.9696	0.8953	41	41	0.9646	0.8946
2	144	0.700	60	42	0.9850	0.9366	48	48	0.9824	0.9344
2	147	0.700	61	43	0.9861	0.9428	49	49	0.9824	0.9373
2	168	0.675	70	49	0.9926	0.9545	56	56	0.9876	0.9479
2	171	0.675	71	50	0.9919	0.9552	57	57	0.9876	0.9524
2	192	0.675	80	56	0.9963	0.9736	64	64	0.9941	0.9707
2	195	0.675	81	57	0.9961	0.9736	65	65	0.9939	0.9709
2	216	0.650	90	63	0.9971	0.9784	72	72	0.9940	0.9732
2	219	0.650	91	64	0.9976	0.9810	73	73	0.9970	0.9764
2	222	0.650	92	65	0.9977	0.9813	74	74	0.9966	0.9800
2	243	0.650	101	71	0.9981	0.9867	81	81	0.9974	0.9856
2	246	0.650	102	72	0.9983	0.9878	82	82	0.9978	0.9866
2	267	0.625	111	78	0.9990	0.9884	89	89	0.9982	0.9874
2	270	0.625	112	79	0.9993	0.9896	90	90	0.9986	0.9890
2	291	0.625	121	85	0.9990	0.9932	97	97	0.9994	0.9928
2	294	0.625	122	86	0.9997	0.9938	98	98	0.9991	0.9931
3	24	1.200	9	5	0.7606	0.4857	6	6	0.7057	0.4696
3	48	0.950	18	10	0.8615	0.6048	12	12	0.8097	0.5830
3	76	0.750	28	16	0.8602	0.6056	19	19	0.8130	0.5842
3	100	0.750	37	21	0.9410	0.7397	25	25	0.9055	0.7134
3	124	0.700	46	26	0.9595	0.7716	31	31	0.9204	0.7455
3	128	0.700	47	27	0.9576	0.7826	32	32	0.9319	0.7605
3	152	0.675	56	32	0.9751	0.8297	38	38	0.9552	0.8121
3	176	0.675	65	37	0.9867	0.8775	44	44	0.9693	0.8552
3	180	0.675	66	38	0.9875	0.8851	45	45	0.9757	0.8688
3	204	0.650	75	43	0.9912	0.9035	51	51	0.9804	0.8891
3	228	0.650	84	48	0.9962	0.9316	57	57	0.9881	0.9186
3	232	0.650	85	49	0.9966	0.9347	58	58	0.9890	0.9224
3	256	0.625	94	54	0.9959	0.9410	64	64	0.9901	0.9285
3	280	0.625	103	59	0.9982	0.9593	70	70	0.9944	0.9490
3	284	0.625	104	60	0.9981	0.9603	71	71	0.9946	0.9510
4	30	0.950	10	5	0.6179	0.2924	6	6	0.5486	0.2834
4	60	0.850	20	10	0.8328	0.4825	12	12	0.7565	0.4542
4	90	0.750	30	15	0.8819	0.5540	18	18	0.8124	0.5180
4	120	0.700	40	20	0.9289	0.6304	24	24	0.8706	0.5992
4	150	0.700	50	25	0.9677	0.7376	30	30	0.9258	0.7008
4	180	0.675	60	30	0.9803	0.7895	36	36	0.9491	0.7520
4	210	0.650	70	35	0.9851	0.8184	42	42	0.9640	0.7911
4	240	0.650	80	40	0.9942	0.8738	48	48	0.9772	0.8426
4	270	0.625	90	45	0.9952	0.8868	54	54	0.9809	0.8567
4	300	0.625	100	50	0.9982	0.9182	60	60	0.9889	0.8915
5	66	0.850	21	9	0.8310	0.4240	11	11	0.7296	0.3929
5	102	0.700	32	14	0.8556	0.4524	17	17	0.7567	0.4182
5	138	0.700	43	19	0.9342	0.5937	23	23	0.8637	0.5524
5	174	0.675	54	24	0.9660	0.6757	29	29	0.9138	0.6268
5	210	0.650	65	29	0.9794	0.7259	35	35	0.9354	0.6826
6	77	0.750	23	9	0.7674	0.3112	11	11	0.6416	0.2838
6	119	0.700	35	14	0.8833	0.4412	17	17	0.7715	0.3981
6	161	0.675	47	19	0.9389	0.5472	23	23	0.8510	0.4929
6	203	0.650	59	24	0.9673	0.6241	29	29	0.8942	0.5670

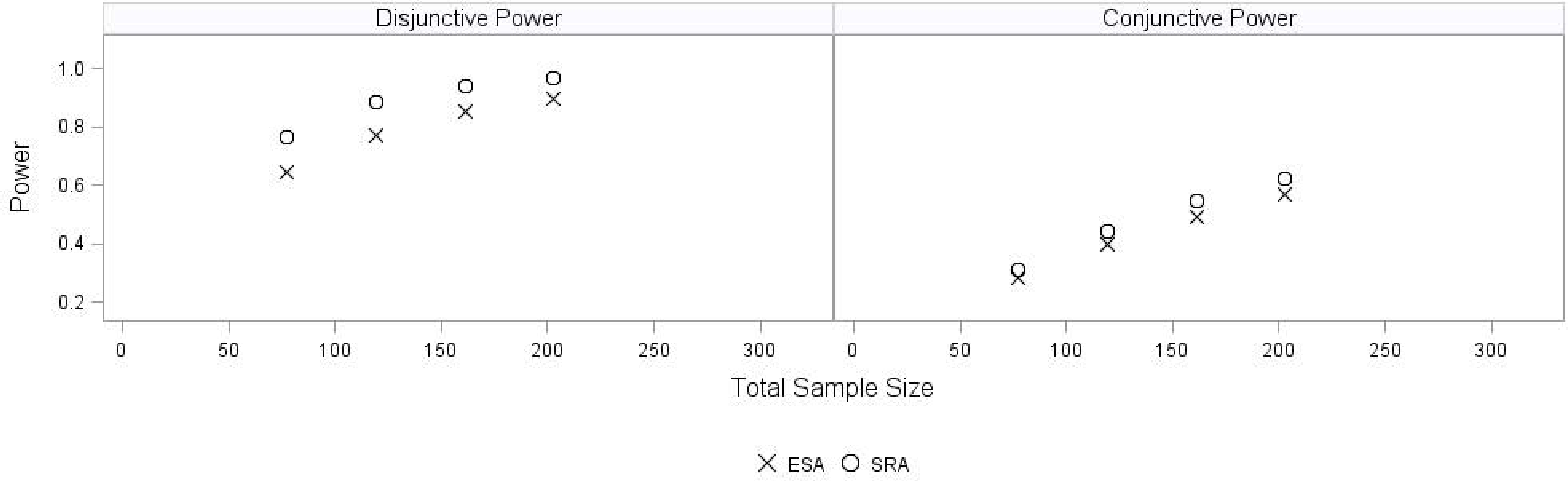

Disjunctive Power
Conjunctive Power
Power
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Total Sample Size
ESA
SRA

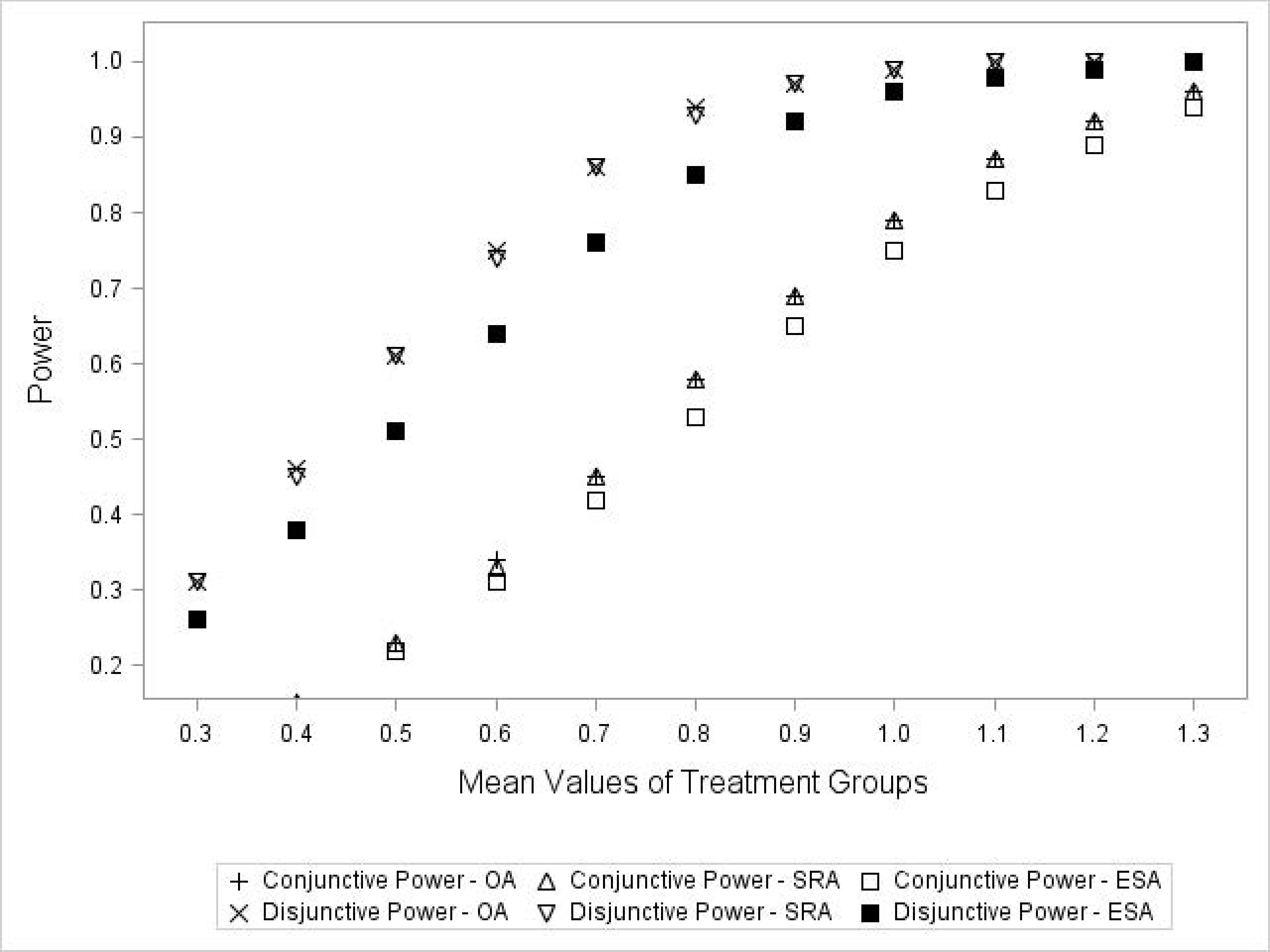

Disjunctive Power
Conjunctive Power
Power
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Total Sample Size
ESA
SRA



Disjunctive Power
Conjunctive Power
Power
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Total Sample Size
ESA
SRA


Disjunctive Power
Conjunctive Power
Power
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Total Sample Size
ESA
SRA


Disjunctive Power
Conjunctive Power
Power
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Total Sample Size
ESA
SRA


Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Mean Values of Treatment Groups
Conjunctive Power - OA
Conjunctive Power - SRA
Conjunctive Power - ESA
Disjunctive Power - OA
Disjunctive Power - SRA
Disjunctive Power - ESA