STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:859–866

Sample size tables for exact single-stage phase II designs

R. P. A'Hern*

Department of Computing and Information, Royal Marsden NHS Trust, Fulham Road,
London SW3 6JJ, U.K.

SUMMARY

Tables for single-phase II trials based on the exact binomial distribution are presented. These are preferable to those generated using Fleming's design, which are based on the normal approximation and can give rise to anomalous results. For example, if the upper success rate is accepted, the lower success rate, which the trial is designed to reject, may be included in the final confidence interval for the proportion being estimated. Copyright © 2001 John Wiley & Sons, Ltd.

INTRODUCTION

Single-stage phase II trials are frequently undertaken to determine whether a new procedure or treatment is likely to meet a basic level of efficacy before comparing it with the standard technique in a larger randomized phase III trial. Efficacy is commonly measured as a proportion. It might, for example, be the sensitivity of a new diagnostic procedure or the proportion of patients showing a tumour response (significant tumour shrinkage) when they are given a novel anti-cancer treatment. If the minimum required level of efficacy is p_1, the design of the trial focuses on demonstrating that this level is plausible given the trial results and the efficacy is greater than a second level (p_0) which would indicate that the procedure or treatment is clearly ineffective. This second level is the highest level of efficacy for which it is necessary to reject the new treatment. *The trial tests the null hypothesis* $\mathrm{H}_0 : P \leqslant p_0$ *against the alternative hypothesis* $\mathrm{H}_1 : P \geqslant p_1$. It consists of entering a predetermined number of subjects and deciding in favour of p_0 or p_1 based on the success rate observed by using an appropriate cut-off between p_0 and p_1. Such a trial might be designed to have a 90 per cent chance of demonstrating that the 95 per cent one-sided confidence interval for a tumour response rate excludes 5 per cent if the true rate is 20 per cent. In this case 20 per cent is the required level of efficacy and a rate of 5 per cent would give grounds for rejection.

Multi-stage designs, rather than single-stage designs, should be used in situations in which early termination is desirable if the treatment is ineffective; optimal two-stage designs have been presented by Simon [1] and three-stage designs have been proposed by Chen [2].

* Correspondence to: R.P. A'Hern, Department of Computing and Information, Royal Marsden NHS Trust, Fulham Road, London SW3 6JJ, U.K.

Received November 1999
Accepted May 2000

Designs for single-stage phase II trials commonly calculate the number of patients required using Fleming's single-stage procedure (see for example Machin *et al.* [3]). However, this method is based on a normal approximation to the binomial distribution and is therefore technically incorrect for small trial sizes, as becomes apparent if exact binomial distributions are applied to these designs. This method, for example, recommends a sample of size 34 to distinguish between the two rates 20 per cent (p_1) and 5 per cent (p_0) with a one-sided α of 5 per cent and 90 per cent power ($1-\beta$). The best cut-off to distinguish between these rates is $\geqslant 5$ successes to accept that the higher rate is more likely. However, this cut-off actually has an α_1 of 3 per cent, not 5 per cent, and a power of 84 per cent, not 90 per cent, that is, if the true rate is 5 per cent the probability that five or more successes will be observed is 3 per cent, and if the true rate is 20 per cent the corresponding probability is 84 per cent. The tables which are presented in this paper remedy the problems which may arise due to the use of the normal approximation by calculating the significance level and power based on the exact binomial distribution (except under circumstances detailed in the methods section below). In addition the appropriate cut-off to use for each design is presented so that the final decision rule is apparent at the outset. This has the benefit that once this cut-off has been achieved it is clear the trial must result in the rejection of the hypothesis that the true rate is the lower value, and planning of the follow-on phase III trial can commence or the trial may be terminated without waiting for the total accrual to be reached.

METHODS

The tables were calculated using Microsoft Excel® Office 97 SR-2, making use of the BINOMDIST function which provides binomial probabilities. BINOMDIST gives the probability and cumulative distribution functions for a specified binomial parameter (p), number of successes (r) and number of trials (N). The probability, for example, of getting exactly 7 out of 20 with $p=0.15$ is found by typing '=BINOMDIST(7,20,0.15,0)' into a cell in an Excel worksheet. Setting the final parameter to zero indicates that the probability is required; if this is set to 1 the cumulative distribution function value is returned. The values passed to this function can be references to other cells in the worksheet. These can be changed using a Microsoft Excel Visual Basic programme, enabling a large number of possible designs to be evaluated by testing whether α and β have specified sizes for different cut-off's and trial size. This function failed for some values of large N, in which case a normal approximation was employed. The first occasion this occurred was in calculating the probability of observing more than 304 successes in 1328 trials with a probability of success of 0.2. The effect of using a normal approximation was checked by calculating the probability of observing more than 303 successes in 1328 trials with the same probability of success using BINOMDIST; the exact value was 0.005148 and the normal approximation gave 0.005186, a difference of 0.000038. The normal approximation was therefore used when BINOMDIST failed.

Values of sample size (N) between $N=0.8\times F$ and $N=4\times F$, where F is the Fleming [4] sample size, were tested. The first design (and hence lowest N) which satisfied the design criteria was chosen. All possible cut-offs were tested if N was less than 30, but for larger values of N a restricted range was used. This range was based on the observation that the cut-off must fall between p_0N and p_1N, and will be approximately

$$C = N \times (p_0 + [(z_\alpha/(z_\alpha + z_{1-\beta})) \times (p_1 - p_0)])$$

where z_α and $z_{1-\beta}$ are the standardized normal deviates of the required significance level and power.

RESULTS

Table I shows the cut-off for accepting that p_1 is to be preferred to p_0, together with the total sample size for values of p_0 from 0.05 to 0.90 and for values of p_1 from $(p_0 + 0.05)$ to 0.95 Values for one-sided α equal to 0.01 and 0.05 and for power $(1-\beta)$ equal to 0.8 and 0.9 are shown.

EXAMPLES

A study was planned to determine whether it was possible to obtain adequate fluid from nipple aspirates to measure biochemical changes that may be related to cancer. For this technique to be considered for routine use a success rate in the region of 95 per cent or more would be desirable, but if it was 70 per cent or less the technique would be unacceptable. The study therefore adopted a single-stage design with a 95 per cent change ($\alpha_1 = 5$ per cent) of rejecting the method if the true percentage in whom adequate fluid can be obtained was 70 per cent and with a high chance (93 per cent) of concluding the method was worthwhile if the true percentage is 95 per cent or more. Nineteen aspirates would be taken and if 17 or more yielded adequate fluid the technique would be considered acceptable. If 17 aspirates were adequate the success rate would be 89.5 per cent, which has a lower one-sided 5 per cent confidence limit of 70.4 per cent.

A new anti-nausea drug was to be tested at phase II and it would only be considered worth undertaking a phase III trial if it reduced severe nausea from the expected level of 30 per cent to 15 per cent. In this case p_0 is above p_1 so the trial design can be found by considering the proportion without nausea (increasing from 70 per cent to 85 per cent). Using an α_1 of 5 per cent and power of 90 per cent gives a trial size of 65 with a cut-off of 52 patients not experiencing nausea to accept that a phase III trial should be undertaken.

A randomized phase II trial comparing two new treatments is being planned in which each arm will have a single-stage design. It has been decided that the probability that both new agents are taken on to phase III if they are both effective should be 90 per cent, but that this probability should only be 5 per cent if they are both ineffective. Table I can be used by noting that use of a 95 per cent power level for each of the arms will give a combined power of 90.2 per cent if both treatments are effective. Similarly, use of a 2.5 per cent significance level will ensure that the probability that both treatments go on to phase III if they are both ineffective is 4.94 per cent. If only one treatment is effective the probability it will be correctly identified is then 93 per cent.

Table I. Sample sizes and cut-offs for exact single-stage phase II trials.

p_0	p_1	$\alpha=0.05;\ 1-\beta=0.8$	$\alpha=0.05;\ 1-\beta=0.9$	$\alpha=0.01;\ 1-\beta=0.8$	$\alpha=0.01;\ 1-\beta=0.9$
0.05	0.10	14/169	18/233	23/267	28/346
	0.15	6/52	8/76	10/82	12/108
	0.20	4/27	5/38	7/44	8/57
	0.25	3/16	4/25	5/26	6/35
	0.30	3/14	3/16	5/21	5/25
	0.35	3/11	3/14	4/15	5/21
	0.40	2/7	3/12	4/13	4/15
	0.45	2/6	3/10	3/9	4/13
	0.50	2/5	2/7	3/8	3/9
	0.55	2/5	2/6	3/7	3/8
	0.60	2/4	2/5	3/6	3/7
	0.65	2/4	2/5	3/6	3/7
	0.70	2/4	2/4	3/5	3/6
	0.75	2/3	2/4	2/3	3/6
	0.80	1/1	2/4	2/3	3/5
	0.85	1/1	2/3	2/3	2/3
	0.90	1/1	1/1	2/2	2/3
	0.95	1/1	1/1	2/2	2/2
0.10	0.15	36/270	47/368	58/425	73/555
	0.20	13/78	17/109	21/122	26/160
	0.25	8/40	10/55	13/62	15/78
	0.30	6/25	7/33	9/37	11/49
	0.35	5/18	6/25	7/25	9/35
	0.40	4/13	5/18	6/19	7/24
	0.45	4/11	4/13	5/14	6/19
	0.50	3/8	4/12	5/12	5/14
	0.55	3/7	3/8	4/9	5/13
	0.60	3/6	3/7	4/8	4/9
	0.65	3/6	3/7	4/7	4/9
	0.70	3/5	3/6	3/5	4/8
	0.75	2/3	3/6	3/5	4/7
	0.80	2/3	3/5	3/4	3/5
	0.85	2/3	2/3	3/4	3/5
	0.90	2/2	2/3	2/2	3/4
	0.95	2/2	2/2	2/2	2/2
0.15	0.20	65/355	89/500	106/568	135/742
	0.25	22/101	28/136	35/157	44/206
	0.30	12/48	15/64	19/73	24/98
	0.35	8/28	10/38	14/47	16/58
	0.40	7/21	8/27	10/30	12/39
	0.45	5/14	7/21	8/21	10/29
	0.50	4/10	5/14	7/17	8/21
	0.55	4/9	5/13	6/13	7/17
	0.60	3/6	4/9	5/10	6/13
	0.65	3/6	4/9	5/9	5/10
	0.70	3/5	3/6	5/8	5/9
	0.75	3/5	3/6	4/6	5/9
	0.80	3/4	3/5	4/6	4/6

Table I. (Continued)

p_0	p_1	$\alpha = 0.05;\ 1-\beta = 0.8$	$\alpha = 0.05;\ 1-\beta = 0.9$	$\alpha = 0.01;\ 1-\beta = 0.8$	$\alpha = 0.01;\ 1-\beta = 0.9$
0.15	0.85	3/4	3/5	4/5	4/6
	0.90	2/2	3/4	4/5	4/5
	0.95	2/2	2/2	3/3	4/5
0.20	0.25	101/433	136/596	164/693	210/905
	0.30	31/116	41/160	51/186	64/242
	0.35	17/56	22/77	27/87	34/115
	0.40	12/35	15/47	18/52	22/67
	0.45	8/21	10/29	14/36	16/44
	0.50	7/17	8/21	11/26	12/30
	0.55	6/13	7/17	9/19	10/23
	0.60	5/10	6/13	8/16	9/19
	0.65	4/7	5/10	7/13	8/16
	0.70	4/7	5/9	6/10	7/13
	0.75	4/6	4/7	6/9	6/10
	0.80	3/4	4/6	5/7	6/9
	0.85	3/4	4/6	4/5	5/7
	0.90	2/2	3/4	4/5	4/5
	0.95	2/2	2/2	3/3	4/5
0.25	0.30	140/494	190/683	228/795	291/1031
	0.35	41/129	55/179	68/210	85/270
	0.40	22/62	28/83	35/97	44/127
	0.45	14/36	18/49	23/58	28/74
	0.50	11/26	13/33	17/39	20/48
	0.55	8/17	10/23	13/27	16/36
	0.60	7/14	8/17	10/19	13/27
	0.65	5/9	7/14	9/16	10/19
	0.70	5/8	5/9	8/13	9/16
	0.75	4/6	5/9	6/9	8/13
	0.80	4/6	4/6	6/9	6/9
	0.85	4/5	4/6	6/8	6/9
	0.90	4/5	4/5	5/6	6/8
	0.95	3/3	4/5	4/4	5/6
0.30	0.35	183/549	247/752	298/885	377/1133
	0.40	52/141	69/193	85/227	107/293
	0.45	27/67	36/93	43/104	53/133
	0.50	17/39	22/53	27/60	34/79
	0.55	12/25	16/36	20/41	24/51
	0.60	9/17	12/25	14/26	18/36
	0.65	8/14	9/17	12/21	14/26
	0.70	6/10	8/14	10/16	12/21
	0.75	6/9	6/10	8/12	10/16
	0.80	5/7	6/9	8/11	8/12
	0.85	4/5	5/7	7/9	8/11
	0.90	4/5	4/5	6/7	7/9
	0.95	3/3	4/5	4/4	6/7
0.35	0.40	224/584	305/806	363/938	462/1207
	0.45	62/148	84/206	102/240	130/313
	0.50	31/68	42/96	51/110	64/143

Table I. (Continued)

p_0	p_1	$\alpha = 0.05;\ 1-\beta = 0.8$	$\alpha = 0.05;\ 1-\beta = 0.9$	$\alpha = 0.01;\ 1-\beta = 0.8$	$\alpha = 0.01;\ 1-\beta = 0.9$
0.35	0.55	20/41	25/53	32/64	40/83
	0.60	14/26	18/36	23/42	26/50
	0.65	11/19	13/24	17/29	20/36
	0.70	8/13	11/19	14/22	15/25
	0.75	8/12	8/13	11/16	14/22
	0.80	7/10	8/12	10/14	11/16
	0.85	6/8	7/10	8/10	9/12
	0.90	5/6	6/8	6/7	8/10
	0.95	3/3	5/6	6/7	6/7
0.40	0.45	262/604	360/840	430/984	548/1266
	0.50	74/158	98/214	120/253	149/320
	0.55	36/71	46/94	58/113	71/142
	0.60	23/42	29/56	35/63	44/82
	0.65	16/28	19/34	25/42	31/54
	0.70	12/19	15/25	18/28	22/36
	0.75	10/15	12/19	15/22	17/26
	0.80	8/11	9/13	11/15	14/20
	0.85	6/8	8/11	10/13	11/15
	0.90	5/6	6/8	7/8	9/11
	0.95	4/4	5/6	7/8	7/8
0.45	0.50	299/618	412/861	490/1006	624/1292
	0.55	80/154	112/220	133/253	172/333
	0.60	39/70	53/98	65/115	81/147
	0.65	25/42	31/54	40/66	47/80
	0.70	16/25	22/36	26/40	33/53
	0.75	11/16	15/23	20/29	22/33
	0.80	10/14	11/16	15/20	17/24
	0.85	7/9	9/12	12/15	13/17
	0.90	6/7	7/9	10/12	12/15
	0.95	4/4	6/7	8/9	8/9
0.50	0.55	330/618	458/866	540/1005	693/1300
	0.60	90/158	119/213	144/250	183/323
	0.65	42/69	55/93	69/112	86/143
	0.70	24/37	33/53	42/64	51/80
	0.75	16/23	22/33	26/37	34/50
	0.80	13/18	16/23	20/27	23/32
	0.85	10/13	12/16	15/19	17/22
	0.90	7/8	9/11	12/14	14/17
	0.95	7/8	7/8	10/11	12/14
0.55	0.60	358/613	488/843	578/984	743/1274
	0.65	93/150	128/210	154/246	195/316
	0.70	46/70	59/92	70/105	90/138
	0.75	26/37	34/50	43/61	53/77
	0.80	18/24	23/32	28/37	33/45
	0.85	12/15	16/21	20/25	24/31
	0.90	10/12	12/15	16/19	18/22
	0.95	8/9	8/9	11/12	14/16

Table I. (Continued)

p_0	p_1	$\alpha = 0.05;\ 1-\beta = 0.8$	$\alpha = 0.05;\ 1-\beta = 0.9$	$\alpha = 0.01;\ 1-\beta = 0.8$	$\alpha = 0.01;\ 1-\beta = 0.9$
0.60	0.65	371/585	507/806	603/946	779/1230
	0.70	96/143	130/197	157/232	202/303
	0.75	44/62	59/85	74/103	90/128
	0.80	25/33	33/45	42/55	52/70
	0.85	17/21	21/27	26/32	33/42
	0.90	12/14	14/17	19/22	21/25
	0.95	9/10	9/10	13/14	16/18
0.65	0.70	373/545	519/764	612/890	793/1160
	0.75	96/133	128/180	157/216	201/280
	0.80	42/55	56/75	70/91	89/118
	0.85	25/31	33/42	40/49	50/63
	0.90	17/20	20/24	26/30	31/37
	0.95	11/12	14/16	15/16	19/21
0.70	0.75	368/501	504/691	599/812	772/1052
	0.80	92/119	125/164	151/194	190/247
	0.85	40/49	52/65	65/79	84/104
	0.90	24/28	31/37	38/44	45/53
	0.95	13/14	17/19	23/25	27/30
0.75	0.80	343/437	471/604	561/712	713/910
	0.85	85/103	113/139	136/164	173/211
	0.90	39/45	47/55	57/65	74/86
	0.95	21/23	26/29	34/37	39/43
0.80	0.85	300/359	411/495	493/588	633/759
	0.90	72/82	97/112	118/134	149/171
	0.95	28/30	40/44	51/55	57/62
0.85	0.90	249/281	334/379	401/451	513/580
	0.95	55/59	70/76	87/93	112/121
0.90	0.95	168/179	223/239	270/287	339/362

DISCUSSION

Exact one-stage designs are more accurate than those employing the normal approximation and avoid the anomaly that confidence intervals at the end of a trial which should apparently reject the lower proportion (p_0) may also include this value. The confidence interval which results when the number of successes equals the cut-off is useful to illustrate the purpose of the trial to investigators, since it excludes p_0 by a small amount and it is therefore important that this confidence interval is correct. Specification of the cut-off to be used also makes it clear at what stage a phase III trial becomes inevitable and allows planning of the phase III trial to commence. Termination of the phase II trial in favour of the randomized trial may also be carried out once this point has been reached, but early termination will affect the confidence interval for the estimate of the observed proportion because the denominator would be smaller. If the true rate is p_1 the expected trial size will be c/p_1, where c is the cut-off. The general observation that the cut-off must fall between p_0N and p_1N is a useful check of published trial designs. A design published by Ensign [5], for example, with $p_1 = 75$ per cent and $p_0 = 60$ per cent (5 per cent significance level, 90 per cent power), has a cut-off of 93 per cent (91/98) and can be seen to be incorrect.

Exact trial sizes are typically larger than those using the normal approximation. For example, this method gives a sample of size 555 to distinguish between the two rates 15 per cent (p_1) and 10 per cent (p_0) with a one-sided α of 1.0 per cent and 90 per cent power. This compares with a sample size of 535 using the Fleming design. The best cut-off for the latter design ($\geqslant 70$) has an α_1 of 1.2 per cent and a power of 90.5 per cent.

Machin *et al.* present an example derived from a phase II trial looking at the use of whole-body hypothermia in cancer therapy (Van der Zee *et al.* [6]). They asked how many lung cancer patients would be needed to see whether the treatment warranted further study given that the highest complete response rate at which it was required to reject the treatment was 15 per cent and that hypothermia would be worth developing further if the true response rate was 50 per cent (with an α_1 of 1 per cent 90 per cent power). Fleming's design gives $N = 18$ and with the best cut-off ($\geqslant 7$ for acceptance that the higher rate is more plausible) this has an α_1 of 1.1 per cent and power of 88 per cent. If exactly 7/18 successes are observed then the lower per cent one-sided confidence limit is 14.5 per cent and hence includes 15 per cent. This design therefore technically fails to meet the necessary criteria, but is clearly close. The exact design has 21 patients and uses a cut-off of $\geqslant 8$ for acceptance that the higher rate is more likely.

In the design of trials, values of power greater than or equal to 80 per cent are generally used, though with a power of 80 per cent there is still a one in five chance that the hypothesis H_1: $P \geqslant p_1$ will be incorrectly rejected if p_1 is the true value. Values of α_1 greater than 0.05 may be used if it is reasonable to undertake a phase III trial with less confidence about the activity of the treatment than would be gained by using a lower significance level.

REFERENCES

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
2. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 1997; **16**: 2701–2711.
3. Machin D, Campbell MJ, Fayers PM, Pinol AP. Phase II trials. In *Sample Size Tables for Clinical Studies*. Blackwell Scientific Publications: 1997; Section 10.1.
4. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**:143–151.
5. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three stage design for phase II clinical trials. *Statistics in Medicine* 1994; **13**:1727–1736.
6. Van der Zee, J, van Rhoon GC, Wike-Hooley JL, Faithful NS, Rheinhold HS. Whole body hyperthermia in cancer therapy: a report of phase I–II study. *European Journal of Cancer and Clinical Oncology* 1983; **19**:1189–1200.