MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 12, No. 4, pp. 441–456, 2002

A NOTE ON SAMPLE SIZE CALCULATION FOR MEAN COMPARISONS BASED ON NONCENTRAL *t*-STATISTICS

Shein-Chung Chow,[1] Jun Shao,[2,*] and Hansheng Wang[1]

[1]StatPlus, Inc., Heston Hall, Suite 206, 1790 Yardley-Langhorne Road, Yardley, PA 19067
[2]Department of Statistics, University of Wisconsin, Madison, WI 53706

ABSTRACT

One-sample and two-sample *t*-tests are commonly used in analyzing data from clinical trials in comparing mean responses from two drug products. During the planning stage of a clinical study, a crucial step is the sample size calculation, i.e., the determination of the number of subjects (patients) needed to achieve a desired power (e.g., 80%) for detecting a clinically meaningful difference in the mean drug responses. Based on noncentral *t*-distributions, we derive some sample size calculation formulas for testing equality, testing therapeutic noninferiority/superiority, and testing therapeutic equivalence, under the popular one-sample design, two-sample parallel design, and two-sample crossover design. Useful tables are constructed and some examples are given for illustration.

Key Words: Power; Therapeutic equivalence; Noninferiority; Superiority; Crossover design

*Corresponding author. E-mail: shao@stat.wisc.edu

DOI: 10.1081/BIP-120016229
1054-3406 (Print); 1520-5711 (Online)

INTRODUCTION

In clinical research, clinical trials are usually conducted for evaluation of the efficacy and safety of a test drug as compared to a placebo control or an active control agent (e.g., a standard therapy) in terms of mean responses of some primary study endpoints. The objectives of the intended clinical trials usually include (i) the evaluation of the effect, (ii) the demonstration of therapeutic equivalence/noninferiority, and (iii) the establishment of superiority. Typically, statistical inference on the drug effects is carried out through mean comparisons using various *t*-tests under a one-sample design, a two-sample parallel design, or a two-sample crossover design. During the planning stage of a clinical study, a crucial step is the sample size calculation, i.e., the determination of the number of subjects (patients) needed to achieve a desired power (e.g., 80%) for detecting a clinically meaningful difference in the mean drug responses.

For *t*-tests, sample size calculation formulas based on normal approximations (assuming that the sample sizes are large enough) can be found in statistical textbooks (e.g., Refs. [1,2]). These formulas may not be adequate when the sample sizes are small or moderate. The purpose of this article is to derive sample size calculation formulas based on the noncentral *t*-distributions. We consider three types of tests (testing equality, testing therapeutic noninferiority/superiority, and testing therapeutic equivalence) and three different designs (one-sample design, two-sample parallel design, and two-sample crossover design). Useful tables are constructed for a quick finding of the required sample sizes. Some examples are given for illustration.

ONE-SAMPLE DESIGN

Let x_i be the response from the ith sampled subject, $i = 1, \ldots, n$. In clinical research, x_i could be the difference between matched pairs such as the pre-treatment and post-treatment responses or changes from baseline to endpoint within a treatment group. It is assumed that x_is are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ^2. Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

be the sample mean and sample variance of x_is, respectively. Also, let $\epsilon = \mu - \mu_0$ be the difference between the true mean response of a test drug (μ) and a reference value (μ_0). Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) an indication of *improvement* (*worsening*) of the test drug as compared to the reference value.

Test for Equality

To test whether there is a difference between the mean response of the test drug and the reference value, the following hypotheses are usually considered:

$$H_0 : \epsilon = 0 \quad \text{vs.} \quad H_a : \epsilon \neq 0.$$

Since σ^2 is typically unknown, a commonly used test for this problem is the one-sample t-test, i.e., we reject the null hypothesis H_0 if

$$\left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{\alpha/2,n-1},$$

where $t_{a,n-1}$ is the upper ath quantile of the central t-distribution with $n-1$ degrees of freedom. Under the alternative hypothesis that $\epsilon \neq 0$, the power of the one-sample t-test is given by

$$1 - \mathcal{T}_{n-1}\left(t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}\epsilon}{\sigma} \right) + \mathcal{T}_{n-1}\left(-t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}\epsilon}{\sigma} \right),$$

where $\mathcal{T}_{n-1}(\cdot|\theta)$ is the cumulative distribution function of the noncentral t-distribution with $n-1$ degrees of freedom and the noncentrality parameter θ. When an initial value of ϵ/σ is given, the sample size needed to achieve power $1-\beta$ can be obtained by solving

$$\mathcal{T}_{n-1}\left(t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}\epsilon}{\sigma} \right) - \mathcal{T}_{n-1}\left(-t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}\epsilon}{\sigma} \right) = \beta.$$

By ignoring a small value $\leq \alpha/2$, the power is approximately

$$1 - \mathcal{T}_{n-1}\left(t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}|\epsilon|}{\sigma} \right).$$

Hence, the required sample size can also be obtained by solving

$$\mathcal{T}_{n-1}\left(t_{\alpha/2,n-1} \middle| \frac{\sqrt{n}|\epsilon|}{\sigma} \right) = \beta. \tag{1}$$

If the solution of Eq. (1) is not an integer, then the smallest integer that is larger than the solution of Eq. (1) should be taken as the required sample size. An initial value of $|\epsilon|/\sigma$ is needed to calculate the sample size according to Eq. (1). A lower bound of $|\epsilon|/\sigma$, usually obtained from a pilot study or historical data, can be used as the initial value. A lower bound of ϵ/σ can also be defined as the clinical meaningful difference between the response means relative to the standard deviation (SD) σ.

Table 1 lists the solutions of this equation for some values of α, β, and $\theta = |\epsilon|/\sigma$.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

Table 1. Smallest n with $\mathcal{T}_{n-1}(t_{\alpha,n-1}|\sqrt{n}\theta) \le \beta$

	$\alpha = 2.5\%$		$\alpha = 5\%$			$\alpha = 2.5\%$		$\alpha = 5\%$	
	$1-\beta =$		$1-\beta =$			$1-\beta =$		$1-\beta =$	
θ	80%	90%	80%	90%	θ	80%	90%	80%	90%
0.10	787	1053	620	858	0.54	29	39	23	31
0.11	651	871	513	710	0.56	28	36	22	29
0.12	547	732	431	597	0.58	26	34	20	27
0.13	467	624	368	509	0.60	24	32	19	26
0.14	403	539	317	439	0.62	23	30	18	24
0.15	351	469	277	382	0.64	22	28	17	23
0.16	309	413	243	336	0.66	21	27	16	22
0.17	274	366	216	298	0.68	19	25	15	20
0.18	245	327	193	266	0.70	19	24	15	19
0.19	220	293	173	239	0.72	18	23	14	18
0.20	199	265	156	216	0.74	17	22	13	18
0.21	180	241	142	196	0.76	16	21	13	17
0.22	165	220	130	179	0.78	15	20	12	16
0.23	151	201	119	164	0.80	15	19	12	15
0.24	139	185	109	151	0.82	14	18	11	15
0.25	128	171	101	139	0.84	14	17	11	14
0.26	119	158	93	129	0.86	13	17	10	14
0.27	110	147	87	119	0.88	13	16	10	13
0.28	103	136	81	111	0.90	12	16	10	13
0.29	96	127	75	104	0.92	12	15	9	12
0.30	90	119	71	97	0.94	11	14	9	12
0.32	79	105	62	86	0.96	11	14	9	11
0.34	70	93	55	76	0.98	11	14	8	11
0.36	63	84	50	68	1.00	10	13	8	11
0.38	57	75	45	61	1.04	10	12	8	10
0.40	52	68	41	55	1.08	9	12	7	9
0.42	47	62	37	50	1.12	9	11	7	9
0.44	43	57	34	46	1.16	8	10	7	8
0.46	40	52	31	42	1.20	8	10	6	8
0.48	37	48	29	39	1.30	7	9	6	7
0.50	34	44	27	36	1.40	7	8	5	7
0.52	32	41	25	34	1.50	6	7	5	6

Let Φ be the standard normal cumulative distribution function and z_α be the upper αth quantile of Φ. When n is sufficiently large, $t_{\alpha/2,n-1} \approx z_{\alpha/2}$, $t_{\beta,n-1} \approx z_\beta$, and

$$\mathcal{T}_{n-1}\left(t_{\alpha/2,n-1}\left|\frac{\sqrt{n}\epsilon}{\sigma}\right.\right) \approx \Phi\left(z_{\alpha/2} - \frac{\sqrt{n}\epsilon}{\sigma}\right). \tag{2}$$

This leads to

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\epsilon^2},$$

which is the formula given in many statistical textbooks.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

Test for Noninferiority/Superiority

The problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0 : \epsilon \leq \delta \quad \text{vs.} \quad H_a : \epsilon > \delta,$$

where δ is the superiority or noninferiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates superiority over the reference value. When $\delta < 0$, the rejection of the null hypothesis implies noninferiority against the reference value.

Using the one-sample t-test, the null hypothesis H_0 is rejected at the α level of significance if

$$\frac{\bar{x} - \mu_0 - \delta}{s/\sqrt{n}} > t_{\alpha,n-1}.$$

The power of this test is given by

$$1 - \mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\epsilon - \delta)}{\sigma}\right.\right).$$

The sample size needed to achieve power $1 - \beta$ can be obtained by solving

$$\mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\epsilon - \delta)}{\sigma}\right.\right) = \beta.$$

By letting $\theta = (\epsilon - \delta)/\sigma$, Table 1 can be used to find the required sample size. From approximation (2), the following approximate formula can be used to calculate the required sample size when n is sufficiently large:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\epsilon - \delta)^2}.$$

Test for Equivalence

The objective is to test the following hypotheses

$$H_0 : |\epsilon| \geq \delta \quad \text{vs.} \quad H_a : |\epsilon| < \delta.$$

The test drug is concluded to be equivalent to a gold standard on average if the null hypothesis is rejected at significance level α. In this problem, a commonly used test is the combination of two one-sided t-tests, i.e., the null hypothesis H_0 is rejected at significance level α if

$$\frac{\sqrt{n}(\bar{x} - \mu_0 - \delta)}{s} < -t_{\alpha,n-1} \quad \text{and} \quad \frac{\sqrt{n}(\bar{x} - \mu_0 + \delta)}{s} > t_{\alpha,n-1}.$$

The power of this test is

$$1 - \mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\delta-\epsilon)}{\sigma}\right.\right) - \mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\delta+\epsilon)}{\sigma}\right.\right).$$

Hence, the sample size needed to achieve power $1-\beta$ can be obtained by setting the power to $1-\beta$. Since the power is larger than

$$1 - 2\mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\delta-|\epsilon|)}{\sigma}\right.\right),$$

a conservative approximation to the sample size needed to achieve power $1-\beta$ can be obtained by solving

$$\mathcal{T}_{n-1}\left(t_{\alpha,n-1}\left|\frac{\sqrt{n}(\delta-|\epsilon|)}{\sigma}\right.\right) = \frac{\beta}{2},$$

which can be done by using Table 1 with $\theta = (\delta - |\epsilon|)/\sigma$. From approximation (2), the following approximate formula can be used to calculate the required sample size when n is sufficiently large:

$$n = \frac{(z_\alpha + z_{\beta/2})^2\sigma^2}{(\delta-|\epsilon|)^2}.$$

Examples

To illustrate the use of sample size formulas derived above, we first consider an example concerning a study of osteoporosis in post-menopausal women. Osteoporosis and osteopenia (or decreased bone mass), most commonly develop in post-menopausal women. The consequences of osteoporosis are vertebral crush fractures and hip fractures. The diagnosis of osteoporosis is made when vertebral bone density is more than 10% below what is expected for sex, age, height, weight, and race (see, e.g., Ref. [3]). Usually, bone density is reported in terms of SD from mean values. The World Health Organization (WHO) defines osteopenia as bone density value greater than one SD below peak bone mass levels in young women and osteoporosis as a value of greater than 2.5 SD below the same measurement scale. In medical practice, most clinicians suggest therapeutic intervention should be begun in patients with osteopenia to prevent progression to osteoporosis.

Suppose a pharmaceutical company is interested in investigating the effect of a test drug on the prevention of the progression to osteoporosis in women with osteopenia. We assume that the mean bone density before the treatment is 1.5 SD (i.e., $\mu_0 = 1.5\,\text{SD}$). Suppose that the mean bone density after treatment is expected to be 2.0 SD (i.e., $\mu_1 = 2.0\,\text{SD}$). We have $\epsilon = \mu_1 - \mu_0 = 2.0\,\text{SD} - 1.5\,\text{SD} = 0.5\,\text{SD}$. At $\alpha = 0.05$, the required sample size for having an 80% power (i.e., $\beta = 0.2$) for correctly detecting a difference of $\epsilon = 0.5\,\text{SD}$

change from pre-treatment to post-treatment can be obtained by normal approximation as

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\epsilon^2} = \frac{(1.96 + 0.84)^2}{(0.5)^2} = 31.36 \approx 32.$$

On the other hand, the sample size can also be obtained by solving Eq. (1). Note that

$$\theta = \frac{|\epsilon|}{\sigma} = 0.5.$$

By referring to the column under $\alpha = 2.5\%$ (two-sided test) at the row with $\theta = 0.5$ in Table 1, it can be found that the sample size needed is 34.

To illustrate the use of the sample size formula for testing equivalence, we consider another example concerning the effect of a test drug on body weight change in terms of body mass index (BMI) before and after the treatment. Suppose clinicians consider that a less than 5% change in BMI from baseline (pre-treatment) to endpoint (post-treatment) is not a safety concern for the indication of the disease under study. Thus, we consider $\delta = 5\%$ as the equivalence limit. The objective is then to demonstrate safety by testing equivalence in mean BMI between pre-treatment and post-treatment of the test drug. Assume the true BMI difference is 0 ($\epsilon = 0$) and the SD is 5% ($\sigma = 0.1$). With $\alpha = 0.05$, the sample size required for achieving an 80% power obtained by normal approximation is

$$n = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2}{\delta^2} = \frac{(1.64 + 1.28)^2 0.10^2}{0.05^2} = 34.11 \approx 35.$$

On the other hand, the sample size can be obtained by using Table 1. Note that

$$\theta = \frac{\delta}{\sigma} = \frac{0.05}{0.10} = 0.50.$$

By referring to the column under $\alpha = 5\%$ and $1 - \beta = 90\%$ at the row with $\theta = 0.50$ in Table 1, it can be found the sample size needed is 36.

TWO-SAMPLE PARALLEL DESIGN

Let x_{ij} be the response observed from the jth subject in the ith treatment group, $j = 1, \ldots, n_i$, $i = 1, 2$. It is assumed that x_{ij}, $j = 1, \ldots, n_i$, $i = 1, 2$, are independent normal random variables with mean μ_i and variance σ^2. Let

$$\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad \text{and} \quad s^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2$$

be the sample means for ith treatment group and the pooled sample variance, respectively. Also, let $\epsilon = \mu_2 - \mu_1$ be the true mean difference between a test

drug (μ_2) and a placebo control or an active control agent (μ_1). Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) as an indication of *improvement* (*worsening*) of the test drug as compared to the placebo control or active control agent. In practice, it may be desirable to have an unequal treatment allocation, i.e., $n_1/n_2 = \kappa$ for some κ. Note that $\kappa = 2$ indicates a 1–2 test-control allocation, whereas $\kappa = 1/2$ indicates a 2–1 test-control allocation.

Test for Equality

The objective is to test whether there is a difference between the mean responses of the test drug and the placebo control or active control. Hence, the following hypotheses are considered:

$$H_0 : \epsilon = 0 \quad \text{vs.} \quad H_a : \epsilon \neq 0.$$

A commonly used test for this problem is the two-sample t-test, i.e., the null hypothesis H_0 is rejected if

$$\left| \frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{\alpha/2, n_1+n_2-2}.$$

Under the alternative hypothesis that $\epsilon \neq 0$, the power of this test is

$$1 - \mathcal{T}_{n_1+n_2-2}\left(t_{\alpha/2,n_1+n_2-2} \left| \frac{\epsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right. \right) + \mathcal{T}_{n_1+n_2-2}\left(-t_{\alpha/2,n_1+n_2-2} \left| \frac{\epsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right. \right).$$

Thus, with $n_1 = \kappa n_2$, the sample size n_2 needed to achieve power $1-\beta$ can be obtained by setting the power equal to $1-\beta$.

After ignoring a small term of value $\leq \alpha/2$, the power is approximately

$$1 - \mathcal{T}_{n_1+n_2-2}\left(t_{\alpha/2,n_1+n_2-2} \left| \frac{|\epsilon|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right. \right).$$

Hence, the required sample size n_2 can also be obtained by solving

$$\mathcal{T}_{(1+\kappa)n_2-2}\left(t_{\alpha/2,(1+\kappa)n_2-2} \left| \frac{\sqrt{n_2}|\epsilon|}{\sigma\sqrt{1+1/\kappa}} \right. \right) = \beta.$$

Table 2 can be used to obtain the solutions for $\kappa = 1, 2$, and some values of $\theta = |\epsilon|/\sigma$, α, and β. When $\kappa = 1/2$, Table 2 can be used to find the required n_1 and $n_2 = 2n_1$.

Table 2. Smallest n with $\mathcal{T}_{(1+\kappa)n-2}(t_{\alpha,(1+\kappa)n-2}|\sqrt{n}\theta/\sqrt{1+1/\kappa}) \leq \beta$

	$\kappa = 1$				$\kappa = 2$			
	$\alpha = 2.5\%$		$\alpha = 5\%$		$\alpha = 2.5\%$		$\alpha = 5\%$	
	$1-\beta =$		$1-\beta =$		$1-\beta =$		$1-\beta =$	
θ	80%	90%	80%	90%	80%	90%	80%	90%
0.30	176	235	139	191	132	176	104	144
0.32	155	207	122	168	116	155	92	126
0.34	137	183	108	149	103	137	81	112
0.36	123	164	97	133	92	123	73	100
0.38	110	147	87	120	83	110	65	900
0.40	100	133	78	108	75	100	59	810
0.42	90	121	71	98	68	90	54	740
0.44	83	110	65	90	62	83	49	670
0.46	76	101	60	82	57	76	45	620
0.48	70	93	55	76	52	70	41	570
0.50	64	86	51	70	48	64	38	520
0.52	60	79	47	65	45	59	35	480
0.54	55	74	44	60	42	55	33	450
0.56	52	68	41	56	39	51	31	420
0.58	48	64	38	52	36	48	29	390
0.60	45	60	36	49	34	45	27	370
0.65	39	51	30	42	29	38	23	310
0.70	34	44	26	36	25	33	20	270
0.75	29	39	23	32	22	29	17	240
0.80	26	34	21	28	20	26	15	210
0.85	23	31	18	25	17	23	14	190
0.90	21	27	16	22	16	21	12	170
0.95	19	25	15	20	14	19	11	150
1.00	17	23	14	18	13	17	10	140
1.05	16	21	12	17	12	15	9	130
1.10	15	19	11	15	11	14	9	120
1.15	13	17	11	14	10	13	8	110
1.20	12	16	10	13	9	12	7	100
1.25	12	15	9	12	9	11	7	90
1.30	11	14	9	11	8	11	6	90
1.35	10	13	8	11	8	10	6	80
1.40	10	12	8	10	7	9	6	80
1.45	9	12	7	9	7	9	5	70
1.50	9	11	7	9	6	8	5	70

From approximation (2), the following approximate formula can be used when both n_1 and n_2 are large:

$$n_1 = \kappa n_2 \quad \text{and} \quad n_2 = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2 (1 + 1/\kappa)}{\epsilon^2}.$$

Test for Noninferority/Superiority

The problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0 : \epsilon \leq \delta \quad \text{vs.} \quad H_a : \epsilon > \delta,$$

where δ is the superiority or noninferiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates the superiority of the test drug over the control. When $\delta < 0$, the rejection of the null hypothesis indicates the noninferiority of the test drug against the control.

The usual two-sample t-test rejects the null hypothesis H_0 if

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha, n_1+n_2-2}.$$

Under the alternative hypothesis that $\epsilon > \delta$, the power of this test is given by

$$1 - \mathcal{T}_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2} \middle| \frac{\epsilon - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right).$$

The sample size needed to achieve power $1 - \beta$ can be obtained by solving the following equation:

$$\mathcal{T}_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2} \middle| \frac{\epsilon - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) = \beta.$$

By letting $\theta = (\epsilon - \delta)/\sigma$, Table 2 can be used to find the required sample size.

From approximation (2), the following approximate formula can be used to calculate the required sample size when n_1 and n_2 are sufficiently large:

$$n_1 = \kappa n_2 \quad \text{and} \quad n_2 = \frac{(z_\alpha + z_\beta)^2\sigma^2(1 + 1/\kappa)}{(\epsilon - \delta)^2}.$$

Test for Equivalence

The objective is to test the following hypotheses

$$H_0 : |\epsilon| \geq \delta \quad \text{vs.} \quad H_a : |\epsilon| < \delta.$$

The test drug is concluded to be equivalent to the control in average if the null hypothesis is rejected at significance level α.

A commonly used test for this problem is the combination of two one-sided t-tests, i.e., the null hypothesis H_0 is rejected at the α level of significance if

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -t_{\alpha,n_1+n_2-2} \quad \text{and} \quad \frac{\bar{x}_1 - \bar{x}_2 + \delta}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha,n_1+n_2-2}.$$

Under the alternative hypothesis that $|\epsilon| < \delta$, the power of this test is

$$1 - \mathcal{T}_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta-\epsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right) - \mathcal{T}_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta+\epsilon}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right).$$

Hence, with $n_1 = \kappa n_2$, the sample size n_2 needed to achieve power $1-\beta$ can be obtained by setting the power to $1-\beta$. Since the power is larger than

$$1 - 2\mathcal{T}_{n_1+n_2-2}\left(t_{\alpha,n_1+n_2-2}\left|\frac{\delta-|\epsilon|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right.\right),$$

a conservative approximation to the sample size n_2 can be obtained by solving

$$\mathcal{T}_{(1+\kappa)n_2-2}\left(t_{\alpha,(1+\kappa)n_2-2}\left|\frac{\sqrt{n_2}(\delta-|\epsilon|)}{\sigma\sqrt{1+1/\kappa}}\right.\right) = \frac{\beta}{2}.$$

Table 2 can be used to calculate n_1 and n_2.

From approximation (2), the following approximation formula can be used to calculate the required sample size when n_1 and n_2 are sufficiently large:

$$n_1 = \kappa n_2 \quad \text{and} \quad n_2 = \frac{(z_\alpha + z_{\beta/2})^2\sigma^2(1+1/\kappa)}{(\delta-|\epsilon|)^2}.$$

An Example

Consider an example concerning a clinical trial for evaluation of the effect of a test drug on cholesterol in patients with coronary heart disease (CHD). Cholesterol is the main lipid associated with arteriosclerotic vascular disease. The purpose of cholesterol testing is to identify patients at risk for arteriosclerotic heart disease. The liver metabolizes the cholesterol to its free form and transported in the bloodstream by lipoproteins. As indicated by Ref. [2], nearly 75% of the cholesterol is bound to low density lipidproteins (LDLs) and 25% is bound to high density lipoproteins (HDLs). Therefore, cholesterol is the main component of LDLs and only a minimal component of HDLs and very low density lipoproteins. LDL is the most directly associated with increased risk of CHD.

A pharmaceutical company is interested in conducting a clinical trial to compare two cholesterol-lowing agents for treatment of patients with CHD through a parallel design. The primary efficacy parameter is the LDL. Suppose that the pharmaceutical company is interested in establishing noninferiority of the test drug as compared to the active control agent. Similarly, we consider the difference of 5% is a difference of clinical importance. Thus, the noninferiority margin is chosen to be 5% (i.e., $\delta = -0.05$). Also, suppose the true difference in mean LDL between treatment groups is 1% (i.e., $\epsilon = \mu_2(\text{test}) - \mu_1(\text{control}) = -0.01$) and the standard deviation $\sigma = 0.10$. Consider $\alpha = 0.05$ and $\beta = 0.2$. Then, by normal approximation, the sample size by normal approximation can be determined by

$$n_1 = n_2 = \frac{2(z_\alpha + z_\beta)^2\sigma^2}{(\epsilon - \delta)^2} = \frac{2(1.64 + 0.84)^2 \times 0.1^2}{(-0.01 - (-0.05))^2} = 76.88 \approx 77.$$

On the other hand, the sample size can also be obtained by using Table 2. Note that

$$\theta = \frac{|\epsilon - \delta|}{\sigma} = \frac{0.04}{0.10} = 0.40.$$

By referring to the column under $\alpha = 5\%$ at the row with $\theta = 0.40$, $1 - \beta = 80\%$, and $k = 1$ in Table 2, it can be found that the sample size needed is 51.

TWO-SAMPLE CROSSOVER DESIGN

In this section, we consider a $2 \times 2m$ replicated crossover design comparing mean responses of a test drug and a reference drug. Let y_{ijkl} be the lth replicate or response ($l = 1, \ldots, m$) observed from the jth subject ($j = 1, \ldots, n$) in the ith sequence ($i = 1, 2$) under the kth treatment ($k = 1, 2$). The following model is considered:

$$y_{ijkl} = \mu_k + \gamma_{ik} + s_{ijk} + e_{ijkl}, \tag{3}$$

where μ_k is the kth treatment effect, γ_{ik} is the fixed effect of the ith sequence under treatment k, and s_{ijk} is the random effect of the jth subject in the ith sequence under treatment k. $(s_{ij1}, s_{ij2}), i = 1, 2, j = 1, \ldots, n$ are assumed to be i.i.d. as bivariate normal random variables with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{\text{BT}}^2 & \rho\sigma_{\text{BT}}\sigma_{\text{BR}} \\ \rho\sigma_{\text{BT}}\sigma_{\text{BR}} & \sigma_{\text{BR}}^2 \end{pmatrix}.$$

e_{ij1l} and e_{ij2l} are assumed to be independent normal random variables with mean 0 and variance σ_{WT}^2 or σ_{WR}^2 (depending on the treatment). Define

$$\sigma_{\text{D}}^2 = \sigma_{\text{BT}}^2 + \sigma_{\text{BR}}^2 - 2\rho\sigma_{\text{BT}}\sigma_{\text{BR}}.$$

σ_D^2 is the variability due to the effect of subject-by-treatment interaction, which reflects the heteroscedasticity of the subject random effect between the test drug and the reference drug.

Let $\epsilon = \mu_2 - \mu_1$ (test − reference),

$$\bar{y}_{ijk\cdot} = \frac{1}{m}(y_{ijk1} + \cdots + y_{ijkm}) \quad \text{and} \quad d_{ij} = \bar{y}_{ij1\cdot} - \bar{y}_{ij2\cdot}.$$

Under appropriate constraints on γ_{ik}s, an unbiased estimate for ϵ is given by

$$\hat{\epsilon} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} d_{ij}.$$

Under model (3), $\hat{\epsilon}$ follows a normal distribution with mean ϵ and variance $\sigma_m^2/(2n)$, where

$$\sigma_m^2 = \sigma_D^2 + \frac{1}{m}(\sigma_{WT}^2 + \sigma_{WR}^2).$$

An unbiased estimate for σ_m^2 can be obtained by

$$\hat{\sigma}_m^2 = \frac{1}{2(n-1)}\sum_{i=1}^{2}\sum_{j=1}^{n}(d_{ij} - \bar{d}_{i\cdot})^2,$$

where

$$\bar{d}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} d_{ij}.$$

Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) as an indication of *improvement* (*worsening*) of the test drug as compared to the reference drug. In practice, σ_m is usually unknown.

Test for Equality

The objective is to test the following hypotheses

$$H_0 : \epsilon = 0 \quad \text{vs.} \quad H_a : \epsilon \neq 0.$$

The null hypothesis H_0 is rejected at α level of significance if

$$\left|\frac{\hat{\epsilon}}{\hat{\sigma}_m/\sqrt{2n}}\right| > t_{\alpha/2,2n-2}.$$

Under the alternative hypothesis that $\epsilon \neq 0$, the power of this test is given by

$$1 - \mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}\epsilon}{\sigma_m}\right.\right) + \mathcal{T}_{2n-2}\left(-t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}\epsilon}{\sigma_m}\right.\right).$$

As a result, the sample size needed to achieve power $1 - \beta$ can be obtained by setting the power to $1 - \beta$ or, after ignoring a small term $\leq \alpha/2$, by solving

$$\mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}|\epsilon|}{\sigma_m}\right.\right) = \beta.$$

Table 2 with $\kappa = 1$ and $\theta = 2|\epsilon|/\sigma_m$ can be used to obtain n. From approximation (2),

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\sigma_m^2}{2\epsilon^2}$$

for sufficiently large n.

Test for Noninferiority/Superiority

Similar to test for noninferiority/superiority under a parallel design, the problem can be unified by testing the following hypotheses:

$$H_0 : \epsilon \leq \delta \quad \text{vs.} \quad H_a : \epsilon > \delta,$$

where δ is the noninferiority or superiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates the superiority of test drug against the control. When $\delta < 0$, the rejection of the null hypothesis indicates the noninferiority of the test drug over the control. The null hypothesis H_0 is rejected at the α level of significance if

$$\frac{\hat{\epsilon} - \delta}{\hat{\sigma}_m/\sqrt{2n}} > t_{\alpha,2n-2}.$$

Under the alternative hypothesis that $\epsilon > \delta$, the power of this test is given by

$$1 - \mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\epsilon - \delta}{\sigma_m/\sqrt{2n}}\right.\right).$$

As a result, the sample size needed to achieve power $1 - \beta$ can be obtained by solving

$$\mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\epsilon - \delta}{\sigma_m/\sqrt{2n}}\right.\right) = \beta,$$

which can be done by using Table 2 with $\kappa = 1$ and $\theta = 2(\epsilon - \delta)/\sigma_m$. When n is sufficiently large, approximation (2) leads to

$$n = \frac{(z_\alpha + z_\beta)^2\sigma_m^2}{2(\epsilon - \delta)^2}. \tag{4}$$

Test for Equivalence

The objective is to test the following hypotheses

$$H_0 : |\epsilon| \geq \delta \quad \text{vs.} \quad H_a : |\epsilon| < \delta.$$

The test drug is concluded equivalent to the control in average, i.e., the null hypothesis H_0 is rejected at significance level α when

$$\frac{\sqrt{2n}(\hat{\epsilon} - \delta)}{\hat{\sigma}_m} < -t_{\alpha,2n-2} \quad \text{and} \quad \frac{\sqrt{2n}(\hat{\epsilon} + \delta)}{\hat{\sigma}_m} > t_{\alpha,2n-2}.$$

Under the alternative hypothesis that $|\epsilon| < \delta$, the power of this test is

$$1 - \mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}(\delta - \epsilon)}{\sigma_m}\right.\right) - \mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}(\delta + \epsilon)}{\sigma_m}\right.\right).$$

As a result, the sample size needed to achieve power $1 - \beta$ can be obtained by setting the power to $1 - \beta$. Since the power is larger than

$$1 - 2\mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}(\delta - |\epsilon|)}{\sigma_m}\right.\right),$$

a conservative approximation to n can be obtained by solving

$$\mathcal{T}_{2n-2}\left(t_{\alpha/2,2n-2}\left|\frac{\sqrt{2n}(\delta - |\epsilon|)}{\sigma_m}\right.\right) = \frac{\beta}{2},$$

which can be done by using Table 2 with $\kappa = 1$ and $\theta = 2(\delta - |\epsilon|)/\sigma_m$. When n is large, approximation (2) leads to

$$n = \frac{(z_\alpha + z_{\beta/2})^2 \sigma_m^2}{2(\delta - |\epsilon|)^2}. \tag{5}$$

An Example

Consider a standard two-sequence, two-period crossover design ($m = 1$) for trials whose objective is to establish therapeutic equivalence between a test drug and a standard therapy. The sponsor is interested in having an 80% ($1 - \beta = 0.8$) power for establishing therapeutic equivalence with an equivalence margin $\delta = 25\%$. Based on the results from previous studies, it is estimated that the standard deviation is 20% ($\sigma_m = 0.20$). Suppose the true mean difference is -10% (i.e., $\epsilon = \mu_2(\text{test}) - \mu_1(\text{reference}) = -0.10$). According to normal approximation,

$$n = \frac{(z_\alpha + z_{\beta/2})^2 \sigma_m^2}{2(\delta - |\epsilon|)^2} = \frac{(1.64 + 1.28)^2 0.20^2}{2(0.25 - 0.10)^2} = 7.57 \approx 8.$$

On the other hand, the sample size calculation can also be performed by using

Table 2. Note that

$$\theta = \frac{2(\delta - |\epsilon|)}{\sigma_m} = \frac{2(0.25 - |-0.10|)}{0.20} = 1.50.$$

By referring to the column under $\alpha = 5\%$, $1 - \beta = 90\%$ at the row with $\theta = 1.50$ in Table 2, it can be found that the sample size needed is 9.

Remarks

Sample size calculation for assessment of bioequivalence under higher-order crossover designs including Balaam's design, two-sequence dual design, and four-period optimal design with or without log-transformation can be found in Ref. [4]. For assessment of bioequivalence, the U.S. Food and Drug Administration requires that a log-transformation of the pharmacokinetic (PK) responses be performed before analysis. Chow and Wang[5] examined the difference in sample size calculation under a crossover design with and without log-transformation.

In this section, we focus on $2 \times 2m$ replicated crossover designs. When $m = 1$, it reduces to the standard two-sequence, two-period crossover design. The standard 2×2 crossover design suffers the following disadvantages: (i) it does not allow independent estimates of the intra-subject variabilities because each subject only receives each treatment once, (ii) the effects of sequence, period, and carry-over are confounded and cannot be separated under the study design. The $2 \times 2m$ ($m \geq 2$) replicated crossover design, on the other hand, not only provides independent estimates of the intra-subject variabilities, but also allows separate tests of the sequence, period, and carry-over effects under appropriate statistical assumption.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*; John Wiley & Sons: New York, NY, 1998.
2. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Marcel Dekker: New York, 2000.
3. Pagana, K.D.; Pagana, T.J. *Diagnostic and Laboratory Tests*; Mosby, Inc.: St. Louis, MO, 1999.
4. Chen, K.W.; Chow, S.C.; Li, G. A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs. J. Pharmacokinet. Biopharm. **1997**, *25*, 753–765.
5. Chow, S.C.; Wang, H. On Sample Size Calculation in Bioequivalence Trials. J. Pharmacokinet. Pharmacodyn. **2001**, *28*, 155–169.