Article

SampSizeCal: The platform-independent computational tool for sample sizes in the paradigm of new statistics

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 16 July 2023; *Accepted 28 September 2023*; *Published online 23 October 2023*; *Published 1 June 2024*

Abstract
Dependent upon the statistical significance *p*-value and statistical power, the sample size estimation is widely used in various experimental sciences. Nevertheless, the *p*-value based paradigm, which has resulted in numerous fake conclusions that originate partly from insufficient sample sizes, has been widely criticized in recent years for serious problems. Therefore, I developed a platform-independent computational tool, SampSizeCal, for sample sizes in the paradigm of new statistics. In this tool, both default *p*-values and the maximum *p*-values were greatly enhanced, which will lead to the reasonable increase of sample sizes. The computational tool harbors more than 120 sample size methods for experimental designs. SampSizeCal includes both online and offline versions, and can be used for various computing devices (PCs, iPads, smartphones, etc.), operating systems (Windows, Mac, Android, Harmony, etc.) and web browsers (Chrome, Firefox, Sougo, 360, etc). It is currently the most comprehensive platform-independent computational tool for sample sizes, and can be used in experimental sciences such as medicine (clinical medicine, experimental zoology, public health, pharmacy, etc.), biology, ecology, agronomy, psychology and engineering technology.

Keywords sampling; sample size; computational tool; significance level; statistical power; new statistics.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Sample size refers to the number of subjects or survey subjects included in each sample in experimental research and survey research. A sample size that is too small will result in insufficient information, unstable indicators, large sampling errors, poor reliability of conclusions, and poor accuracy in inferring the population. A sample size that is too large will result in a waste of manpower, material resources, financial resources, and time. If it is larger, it will also increase the difficulty of quality control at work. In addition, the too large sample size may introduce more confounding factors, adversely affecting the research results. Therefore, it is of great significance to determine the sample size on the premise of ensuring that the research conclusions are scientific, authentic and reliable (Mace, 1964; Cochran, 1977; Pielou, 1977; Eberhardt, 1978; Southwood,

1978; Burnham et al., 1980; Seber, 1982; Zar, 1984; Fleiss, 1986; Downing et al., 1987; Cohen, 1988; Krebs, 1989; Desu and Raghavarao, 1990; Machin et al., 1997; Fleiss et al., 2003; Li and Fine, 2004; Ardilly, 2005; Good, 2005; Tille, 2006; Zhang, 2007; Chow et al., 2008; Ryan, 2013; Liang, 2014; Julious, 2020).

Dependent upon the statistical significance *p*-value and statistical power, the sample size estimation is widely used in various experimental sciences. The *p*-value is at the heart of the statistical significance testing and whether a research is statistically significant is mainly determined by using the *p*-value (Fisher, 1935; Yates, 1951; Sellke et al., 2001; Sun, 2016; Amrhein et al., 2019; Bergstrom and West, 2021; Zhang, 2022a-c). Nevertheless, the statistical significance paradigm has been substantially questioned in recent years because *p*-value is too sensitive, *p*-value is a dichotomous subjective index, and statistical significance is related to sample size, etc (Sellke et al., 2001; Trafimow and Marks, 2015; Baker, 2016; Wasserstein and Lazar, 2016; McShane and David, 2017; Amrhein et al., 2019; Tong, 2019; Wasserstein et al., 2019; Zhang, 2022a-c). Statistical significance paradigm was considered to be one of the sources of false conclusions and research reproducibility crisis (Ioannidis, 2005; Open Science Collaboration, 2015; Errington et al., 2021; Huang, 2021a-b, 2023; Kafdar, 2021; Nature Editorial, 2021; Vrieze, 2021; Huang, 2021a-b, 2023; Zhang, 2022a-c, 2023). To solve these problems, a new statistics, in which to use a stricter *p*-value rather than 0.05 is one of the choices, is suggested (Zhang, 2022a). Therefore, in addition to writing, publishing and using new statistical monographs and textbooks, the most urgent task is to revise and distribute various statistical software based on the new statistics for further use (Zhang, 2022a-c, 2023; Zhang and Qi, 2024).

The *p*-value based significance level has resulted in numerous fake conclusions that originate partly from insufficient sample sizes. In present article, based on Liang (2014) and other studies (Mace, 1964; Cochran, 1977; Pielou, 1977; Eberhardt, 1978; Southwood, 1978; Burnham et al., 1980; Seber, 1982; Zar, 1984; Fleiss, 1986; Downing et al., 1987; Cohen, 1988; Krebs, 1989; Desu and Raghavarao, 1990; Machin et al., 1997; Fleiss et al., 2003; Li and Fine, 2004; Ardilly, 2005; Good, 2005; Tille, 2006; Zhang, 2007; Chow et al., 2008; Ryan, 2013; Julious, 2020), I thus developed a platform-independent computational tool for sample sizes in the paradigm of new statistics. In this tool, both default *p*-values and the maximum *p*-values were greatly enhanced, which will lead to the reasonable increase of sample sizes. It is expected to be used in various experimental sciences including medicine, biology, ecology, agricultural sciences, etc.

2 Methods for Estimation of Sample Sizes

2.1 Comparing Means

2.1.1 One-Sample Design

2.1.1.1 Baseline Method: Significance Test

The baseline test is often used to compare the treatment group with the standard accepted value, or to compare the baseline data with the data after treatment. The sample size (n) is (Zar, 1984; Machin et al., 1997)

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

where d: the expected difference between the sample mean and the reference value; σ: standard deviation; z: z-test value. α: the probability of rejecting the truth, that is, the probability of rejecting the null hypothesis when there is no difference is true; β is the probability of taking the false, that is, the probability of accepting the null hypothesis when the difference is false. 1-β is the statistical power. z_{α}: standard normal deviation corresponding to α. z_{β}: standard normal deviation corresponding to β.

2.1.1.2 Baseline Method: Non-inferiority or Superiority Test

To make non-inferiority or superiority (NIS) test, the sample size (n) is

$$n = (z_\alpha + z_\beta)^2\sigma^2/(d - \delta)^2$$

where σ: standard deviation; δ: the margin of clinic significance. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, to make non-inferiority test on a drug for hypertension, for a community, known σ=20, d=9, and the margin is 11, then let δ=-11.

2.1.1.3 Baseline Method: Equivalence Test

For equivalence test, the sample size (n) is

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

2.1.1.4 One-Stage Test

For one-stage mean test, the sample size (n) for mean estimation is

$$n = (z_{\alpha/2}\,\sigma/d)^2$$

where σ: the standard deviation, $z_{\alpha/2}$: z-value, and d: the expected magin (absolute error of the mean, i.e., the half-width of confidence interval). The standard deviation, σ, can be estimated in some ways (Krebs, 1989; Zhang, 2007):

(1) Use the σ obtained in the past studies.

(2) σ is estimated from the previous explorative study.

(3) Using rules generalized from professional studies. In fish sampling, for example, suppose there are m observations and the observed values are approximately between a and b, we have σ= (b - a)*f_m, where f_m is obtained from Table 1 (Krebs, 1989).

Table 1 f_m table.

m	10	30	50	70	100	200	300	500
f_m	0.325	0.245	0.222	0.21	0.199	0.182	0.174	0.154

If the relative error is used, such as coefficient of variation (CV):

$$CV = \sigma/\bar{\mu}$$

The sample size for mean estimation is (Zhang, 2007):

$$n = (100*CV*z_{\alpha/2}/r)^2$$

where r: the expected relative error, i.e., the percent (0, 100) of half-width of confidence interval against the mean. Krebs (1989) argues that CV=0.7 for plankton, CV=0.4 for crabs, CV=0.4 for shellfish (0.4), and CV=0.8 for roadside sampling.

If the sample size (n) is large enough compared to the total population, the sample size that is actually used should be corrected to n^*:

$$n^* = n/(1 + n/N)$$

where N: the size of total population.

If a random variable does not strictly follow the normal distribution, it will approximately follow the normal distribution when the sample size is large enough, thus algorithms above hold also. Cochran (1977) proposes that the sample size is large enough if

$$n > 25(\sum(x_i - \bar{x})^3/(ns^3))^2$$

2.1.1.5 Two-stage Test

The sample size can be determined by using two-stage sampling, i.e., sampling with the pre-sampled sample size, n_1, and obtain the standard deviation σ_1, and then determine the sample size (Cochran, 1977):

$$n = (1 + 2/n_1)(z_{\alpha/2}\sigma_1/d)^2$$

2.1.1.6 Random Variable Follows the Poisson Distribution

If the random variable follows the Poisson distribution, the sample size (n) for mean estimation is

$$n = (100*z_{\alpha/2}/r)^2/\bar{\mu}$$

where $\bar{\mu}$: the estimated mean of the random variable, and r: permissible relative error, i.e., the percent (0, 100) of half-width of confidence interval against the mean.

Pielou (1969) proposes a method based on the absolute error d, as the following:

$$n = (z_{\alpha/2}/d)^2\ \bar{\mu}$$

2.1.1.7 Random Variable Follows the Negative Binomial Distribution

Spatial dispersion of organisms in particular invertebrates, usually follow the negative binomial distribution. The sample size (n) for mean estimation is

$$n = (100*z_{\alpha/2}/r)^2\ (1/\bar{\mu} + 1/k)$$

where $\bar{\mu}$: the estimated mean of the random variable, k: the parameter of the negative binomial distribution, and r: the expected relative error, i.e., the percent (0, 100) of half-width of confidence interval against the mean.

In addition, we may use the absolute error based estimation (Karandinos, 1976):

$$n = (z_{\alpha/2}/d)^2/((k\bar{\mu} + \bar{\mu}^2)/k)$$

2.1.1.8 Random Variable Follows the Binomial Distribution

For a random variable that follows the binomial distribution, the sample size (n) for mean estimation is

$$n = (z_{\alpha/2}/d)^2\ w\ (1 - w)$$

where $w=m/n$, m: the number of subjects with incidence; d: between-incidence difference.

2.1.1.9 Probabilistic Distribution Independent

$$n = (z_{\alpha/2}/d)^2[(\alpha' + 1)\,\bar{\mu} + (\beta' - 1)\,\bar{\mu}^2]$$

where α' and β' are the parameters in Iwao regression: $M^* = \alpha' + \beta'\bar{\mu}$.

2.1.2 Two-Sample Parallel Design

2.1.2.1 Significance Test

(1) Known the standard deviation of between-group difference

Suppose there are two groups, a treatment group and a control group (1:1 parallel control design; n observations are required for each group). To test the between-group difference significance, the sample size (n) for detecting between-group difference is (Machin et al., 1997; Chow et al., 2008; Julious, 2010)

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

where σ: the standard deviation; d: between-group difference; z: z-value.

(2) Known the standard deviations of two groups

Suppose the standard deviations of two groups are are σ_1 and σ_2 respectively, the sample size (n) (1:1 design; n observations are required for each group) is

$$n = z_{\alpha/2}^2(\sigma_1^2 + \sigma_2^2)/d^2$$

where d: between-group difference; $z_{\alpha/2}$: z value with confidence degree 1-α. σ_1 and σ_2: the standard deviations of two groups.

2.1.2.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) for detecting between-group difference is

$$n = 2(z_{\alpha/2} + z_{\beta})^2\sigma^2/(d - \delta)^2$$

where σ: standard deviation; d: between-group difference; δ: the margin of clinic significance, $\delta<0$ if non-inferiority test is made and $\delta>0$ if superiority test is made. For example, to make non-inferiority test on hypertension difference of two groups (1:1 parallel control design; n observations are required for each group), for a community, known σ=20, between-group difference d=8, and the margin is 10, then let δ=-10.

2.1.2.3 Equivalence Test

For equivalence design, the sample size (n) for detecting between-group difference is

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

2.1.2.4 Comparison of Paired Data

For the comparison of paired data, the sample size (n) (1:1 design; n observations are required for each group) is (Zar, 1984; Machin et al., 1997)

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

where σ: the standard deviation of between-group differences, d: between-group difference. Thus, n pairs of observations are required.

2.1.3 Two-Sample Crossover Design

2.1.3.1 Significance Test

For two-sample crossover design, the sample size (n) is

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

where σ: the standard deviation of the difference; d: between-group difference. For example, use the drugs A and B to treat hypertension. In the first trial, use A for a period and thereafter use B for the same period. In the second trial, use B first and thereafter A. If B reduces 5 mm Hg of blood pressure more than A, B is considered to be more effective. Set σ=10, α=0.0001, β=0.1, two-sided significance test, and calculate the sample size n (1:1 design; n observations are required for each group).

2.1.3.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) for detecting between-group difference is

$$n = (z_{\alpha} + z_{\beta})^2\sigma^2/(2(d - \delta)^2)$$

where δ: the margin of clinic significance, δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For the example above, to make two-sided non-inferiority test (1:1 design; n observations are required for each group). The margin is 1, i.e., δ=-1.

2.1.3.3 Equivalence Test

For equivalence design, the sample size (n) for detecting between-group difference is

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

Continue the example, for equivalence design, to make two-sided equivalence test (1:1 design; n observations are required for each group), δ=1.

2.1.4 Multiple-Sample One-Way ANOVA

2.1.4.1 Paralell Design

Suppose one factor with k≥3 groups (levels). We hope to statistically compare the difference between means of k populations represented by k groups. Known the variance of means

$$\Delta = 1/\sigma^2 \sum_{i=1}^{k}(\mu_i - \bar{\mu})^2$$

the sample size (n) for each group is: $n = \lambda/\Delta$, where λ is obtained from Table 2. For example, use five drugs to control a disease, the means of incidence reduction are 25, 30, 32, 20, 28; the standard deviation of incidence reduction of each drug is σ=5; two-sided test. Calculate the sample size of each group.

Table 2 λ table.

k	α=0.01	α=0.05	α=0.01	α=0.05
	β=0.1		β=0.2	
2	14.88	10.51	11.68	7.85

k	α=0.01	α=0.05	α=0.01	α=0.05
3	17.43	12.66	13.89	9.64
4	19.25	14.18	15.46	10.91
5	20.74	15.41	16.75	11.94
6	22.03	16.47	17.87	12.83
7	23.19	17.42	18.88	13.63
8	24.24	18.29	19.79	14.36
9	25.22	19.09	20.64	15.03
10	26.13	19.83	21.43	15.65
11	26.99	20.54	22.18	16.25
12	27.8	21.20	22.89	16.81
13	28.58	21.84	23.57	17.34
14	29.32	22.44	24.22	17.85
15	30.04	23.03	24.84	18.34
16	30.73	23.59	25.44	18.82
17	31.39	24.13	26.02	19.27
18	32.04	24.65	26.58	19.71
19	32.66	25.16	27.12	20.14
20	33.27	25.66	27.65	20.56

2.1.4.2 Pairwise Design

In this design, there are at least two groups and no control group is included. The sample size (n) is (Desu and Raghavarao, 1990; Fleiss, 1986)

$$n = \max\{n_{ij}\}$$

where

$$n_{ij} = 2\sigma^2 (z_{\alpha/(2T)} + z_{\beta})^2 / d_{ij}^2$$

where T: times of between-group comparisons, d_{ij}: between-group magin, $d_{ij}=\mu_i - \mu_j$. For example, the means of incidence reduction of two drugs (or dosages) are 18, 25, the value for the control group is 12, the σ for the incidence reduction of two drugs are 3.5 and 5. Conduct 1:1:1 parallel design, T=1; for n_{13}, σ=3.5, d_{13}=18-12=6; for n_{23}, σ=5, d_{12}=25-12=13. n= max$\{n_{13}, n_{23}\}$.

2.1.4.3 Multiple-Sample Williams Design

If the number of periods available for the crossover experiment is the same as the number of treatments, the crossover design that uses the generalized Latin square to balance the first-order lag effect with as few subjects as possible is the Williams design. Common Williams designs are three-group designs (a 6×3 crossover design) and four-group designs (a 4×4 crossover design). When the experimental groups (k) are odd, the design result is a $2k \times k$ crossover design, and when the experimental groups are even, the design result is a $k \times k$ crossover

design.

(1) Significance Test

The sample size (n) is

$$n = \max\{n_{ij}\}$$

where

$$n_{ij} = \sigma^2 (z_{\alpha/2} + z_{\beta})^2/(md_{ij}^2)$$

For example, use three drugs A, B and C to treat a disease. Their incidence reduction are 16, 19, 13 and σ=3 respectively. Use Williams three-crossover design, the treatments are ABC, ACB, BAC, BCA, CAB, CBA. m=6, d_{12}=16-19=3, d_{13}=16-13=3, d_{23}=19-13=6. Two-sided test. n=max{n_1, n_2, n_3}.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = \max\{n_{ij}\}$$

where

$$n_{ij} = \sigma^2(z_{\alpha} + z_{\beta})^2/(m(d_{ij} - \delta)^2)$$

The margin δ<0 if non-inferiority test is made and δ>0 if superiority test is made

(3) Equivalence Test

For equivalence design, the sample size (n) for detecting between-group difference (two-sided test) is

$$n = \max\{n_{ij}\}$$

where

$$n_{ij} = \sigma^2(z_{\alpha} + z_{\beta/2})^2/(m(\delta - |d_{ij}|)^2)$$

In the example above, m=6, d_{12}=16-19=3, d_{13}=16-13=3, d_{23}=19-13=6, σ=3. n=max{n_1, n_2, n_3}.

2.2 Comparing Variabilities

2.2.1 Estimation of Variance

If the the sample size is large enough, the sample size for variance estimation is (Mace, 1964):

$$n = 1.5 + z_{\alpha/2}^2\{[1/v + (1/v^2 - 1)^{1/2}]/v - 0.5\}$$

where $z_{\alpha/2}$: z-value at confidence level α, v: the permissible limit of variance ratio, represented by the ratio of confidence interval, as 0.35, 0.25, etc.

2.2.2 Repeated Parallel Controlled Design

2.2.2.1 Significance Test

Assume that there are two groups and n and m are the number of cases and observations in a group

respectively. The sample size (n) is calculated from the following formula

$$\sigma_T^2/\sigma_R^2 = F_{1-\beta,\, n(m-1),n(m-1)}/F_{\alpha/2,n(m-1),n(m-1)}$$

where σ_T^2 and σ_R^2: the variance for test group and reference group (control group). For example, a parallel control experiment with 2 groups, each repeated 3 times (3 replications). According to the pilot study, the within-subject standard deviation of the group T is 0.4, and the within-subject standard deviation of the group R is 0.6. 1:1 significance design. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.2.2.2 Non-inferiority or Superiority Test

The sample size (n) is calculated from the following formula

$$\sigma_T^2/(\sigma_R^2\ \delta^2) = F_{1-\beta,\, n(m-1),n(m-1)}/F_{\alpha,\, n(m-1),n(m-1)}$$

where δ: the margin. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, δ=-1.

2.2.2.3 Equivalence Test

The sample size (n) is calculated from the following formula

$$\delta^2\sigma_T^2/\sigma_R^2 = F_{\beta/2,\, n(m-1),n(m-1)}/F_{1-\alpha,\, n(m-1),n(m-1)}$$

where δ: the margin, for example, δ=-1.0.

2.2.3 Simple Random Effects Model

2.2.3.1 Significance Test

For significance test, the sample size (n) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2(\sigma_T^2 + \sigma_R^2)/(CV_T - CV_R)^2$$

where CV_T and CV_R: the coefficient of variation for group T (test group) and R (reference or control group) respectively; σ_T and σ_R: the standard deviation of group T and R respectively, and

$$\sigma_i^2 = \frac{CV_i^2}{2*m} + CV_i^4$$

For example, a two-group parallel control experiment with 2 replications. According to the pilot test, the coefficient of variation (CV) of the treatment group was 40%, and that of the control group was 60%. α=0.0001, β=0.10. 1:1 significance design. Calculate the sample size for each group, n.

2.2.3.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2(\sigma_T^2 + \sigma_R^2)/(CV_T - CV_R - \delta)^2$$

where δ: the margin of $CV_T - CV_R$, e.g., δ=0.2. δ<0 if non-inferiority test is made and δ>0 if superiority test is made.

2.2.3.3 Equivalence Test

For equivalence test, the sample size (n) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2(\sigma_T^2 + \sigma_R^2)/(\delta - |CV_T - CV_R|)^2$$

2.2.4 Comparison of Between-Subject Variation

2.2.4.1 Parallel Repeatable Design

(1) Significance Test

For significance test, the sample size (n) is

$$n = (z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/(\sigma_{BT}^2 - \sigma_{BR}^2)^2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + (\sigma_{BR}^2+\sigma_{WR}^2/m)^2 + \sigma_{WT}^4/(m^2(m-1)) + \sigma_{WR}^4/(m^2(m-1)))$$

For example, a parallel control experiment with two groups of 3 replications (m=3). According to the pilot study, the between-subject standard deviations of groups T and R were 0.2 (σ_{BT}) and 0.3 (σ_{BR}), respectively, and the within-subject standard deviations of groups T and R were 0.4 (σ_{WT}) and 0.5 (σ_{WR}), respectively. α=0.0001, β=0.10. 1:1 significance design. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2\sigma^2/(\sigma_{BT}^2 - \delta^2\sigma_{BR}^2)^2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + \delta^4(\sigma_{BR}^2+\sigma_{WR}^2/m)^2 + \sigma_{WT}^4/(m^2(m-1)) + \delta^4\sigma_{WR}^4/(m^2(m-1)))$$

where δ: the margin. 1:1 non-inferiority design. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above.

2.2.4.2 Crossover Repeatable Design

(1) Significance Test

For significance test, the sample size (n) is

$$n = ((z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/(\sigma_{BT}^2 - \sigma_{BR}^2)^2+2)/2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + (\sigma_{BR}^2+\sigma_{WR}^2/m)^2 - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2 + \sigma_{WT}^4/(m^2(m-1)) + \sigma_{WR}^4/(m^2(m-1)))$$

For example, a two-group cross-control experiment with 2 replications (*ABAB*, *BABA*). According to the pilot study, the between-subject standard deviations of groups T and R were 0.2 and 0.3, respectively, and the within-subject standard deviations of groups T and R were 0.4 and 0.5, respectively. α=0.0001, β=0.10, ρ=0.7. 1:1 significance design. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = ((z_{1-\alpha} + z_{1-\beta})^2\sigma^2/(\sigma_{BT}^2 - \delta^2\sigma_{BR}^2)^2+2)/2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + \delta^4(\sigma_{BR}^2+\sigma_{WR}^2/m)^2 - 2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2 + \sigma_{WT}^4/(m^2(m\text{-}1)) + \delta^4\sigma_{WR}^4/(m^2(m\text{-}1)))$$

1:1 non-inferiority design. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above.

2.2.5 Comparison of Overall Variation

Estimates of overall variation were obtained from standard 2x2 crossover/parallel designs or repeated 2x2 crossover/parallel designs.

2.2.5.1 Non-Repeated Parallel Controlled Trials

(1) Significance Test

For significance test, the sample size (n) is calculated from the following formula

$$\sigma_T^2/\sigma_R^2 = F_{1-\beta,\, n-1,n-1}/F_{\alpha/2,n-1,n-1}$$

where σ_T^2 and σ_R^2: the variance for test group and reference group (control group). For example, a non-repeated parallel controlled trial. According to the pilot study, the standard deviations of the groups T and R were 0.4 and 0.6. 1:1 significance design. α=0.0001, β=0.10. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

The sample size (n) is calculated from the following formula

$$\sigma_T^2/(\delta^2\sigma_R^2) = F_{1-\beta,\, n-1,n-1}/F_{\alpha/2,n-1,n-1}$$

where δ: the margin (e.g., 1.3, δ=-1.3). 1:1 non-inferiority design. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above.

(3) Equivalence Test

For equivalence test, the sample size (n) is calculated from the following formula

$$\delta^2\sigma_T^2/\sigma_R^2 = F_{1-\beta,\, n-1,n-1}/F_{\alpha/2,n-1,n-1}$$

where δ: the margin (e.g., 1.3, δ=-1.3). 1:1 equivalence design. Follow the example above.

2.2.5.2 Repeated Parallel Controlled Trials

(1) Significance Test

For significance test, the sample size (n) is

$$n = (z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \sigma_{TR}^2)^2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + (\sigma_{BR}^2+\sigma_{WR}^2/m)^2 + (m-1)\sigma_{WT}^4/m^2 + (m-1)\sigma_{WR}^4/m^2)$$

where $\sigma_{TT}^2=\sigma_{BT}^2+\sigma_{WT}^2$, $\sigma_{TR}^2=\sigma_{BR}^2+\sigma_{WR}^2$. For example, a two-group parallel control experiment with 3 replications (m=3). According to the pilot study, the between-subject standard deviations of groups T and R were 0.2 (σ_{BT}) and 0.3 (σ_{BR}), respectively, and the within-subject standard deviations of groups T and R were 0.4 (σ_{WT}) and 0.5 (σ_{WR}), respectively. α=0.001, β=0.10. 1:1 significance design. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \delta^2\sigma_{TR}^2)^2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + \delta^4(\sigma_{BR}^2+\sigma_{WR}^2/m)^2 + (m-1)\sigma_{WT}^4/m^2 + \delta^4(m-1)\sigma_{WR}^4/m^2)$$

where δ: the margin (e.g., 1.0, δ=-1.0). 1:1 non-inferiority design. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above.

2.2.5.3 Standard 2×2 Crossover Design

(1) Significance Test

For significance test, the sample size (n) is

$$n = ((z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \sigma_{TR}^2)^2+2)/2$$

where

$$\sigma^2 = 2(\sigma_{TT}^4 + \sigma_{TR}^4 - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2)$$

where $\sigma_{TT}^2=\sigma_{BT}^2+\sigma_{WT}^2$, $\sigma_{TR}^2=\sigma_{BR}^2+\sigma_{WR}^2$. For example, a 2×2 standard crossover control experiment. According to the pilot study, the between-subject standard deviations of groups T and R were 0.2 (σ_{BT}) and 0.3 (σ_{BR}), respectively, and the within-subject standard deviations of groups T and R were 0.4 (σ_{WT}) and 0.5 (σ_{WR}), respectively. α=0.0001, β=0.10, ρ=0.8. 1:1 significance design. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = ((z_{1-\alpha} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \delta^2\sigma_{TR}^2)^2+2)/2$$

where

$$\sigma^2 = 2(\sigma_{TT}^4 + \delta^4\sigma_{TR}^4 - 2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2)$$

δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above, 1:1 non-inferiority design, δ=-1 (the margin=1). Calculate the sample size for each group, n.

2.2.5.4 Repeated 2×2 Crossover Design

(1) Significance Test

For significance test, the sample size (n) is

$$n = ((z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \sigma_{TR}^2)^2+2)/2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + (\sigma_{BR}^2+\sigma_{WR}^2/m)^2 - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2 + (m-1)\sigma_{WT}^4/m^2 + (m-1)\sigma_{WR}^4/m^2)$$

where $\sigma_{TT}^2=\sigma_{BT}^2+\sigma_{WT}^2$, $\sigma_{TR}^2=\sigma_{BR}^2+\sigma_{WR}^2$. For example, a two-group cross-control experiment with 2 replications per subject (*ABAB*, *BABA*). According to the pilot study, According to the pilot study, the between-subject standard deviations of groups *T* and *R* were 0.2 (σ_{BT}) and 0.3 (σ_{BR}), respectively, and the within-subject standard deviations of groups *T* and *R* were 0.4 (σ_{WT}) and 0.5 (σ_{WR}), respectively. α=0.0001, β=0.10, ρ=0.8. 1:1 significance design. Calculate the sample size for each group, n.

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = ((z_{1-\alpha} + z_{1-\beta})^2\sigma^2/(\sigma_{TT}^2 - \delta^2\sigma_{TR}^2)^2+2)/2$$

where

$$\sigma^2 = 2((\sigma_{BT}^2+\sigma_{WT}^2/m)^2 + \delta^4(\sigma_{BR}^2+\sigma_{WR}^2/m)^2 - 2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2 + (m-1)\sigma_{WT}^4/m^2 + \delta^4(m-1)\sigma_{WR}^4/m^2)$$

δ<0 if non-inferiority test is made and δ>0 if superiority test is made. Follow the example above, 1:1 non-inferiority design, δ=-1 (the margin=1). Calculate the sample size for each group, n.

2.3 Large Sample Tests for Proportions

2.3.1 One-Sample Design

2.3.1.1 Significance Test

The sample size (n) for significance test of proportion estimation is

$$n = (z_{\alpha/2} + z_{\beta})^2 p(1 - p)/d^2$$

where p: the proportion of total population; d: the difference of proportion. Two-sided significance test. For example, an old method can reduce a disease incidence by 40% and the new method is expected to reduce it by 80%, thus d=0.8-0.4=0.3.

2.3.1.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) is

$$n = (z_{\alpha} + z_{\beta})^2 p(1 - p)/(d - \delta)^2$$

where δ: the margin of d. δ<0 if non-inferiority test is made and δ>0 if superiority test is made.

2.3.1.3 Equivalence Test

For equivalence design, the sample size (n) (two-sided test) is

$$n = (z_{\alpha} + z_{\beta/2})^2 p(1 - p)/(d - |\delta|)^2$$

where δ: the margin of d, e.g., δ=0.05.

2.3.2 Two-Sample Parallel Design

2.3.2.1 Significance Test

For the significance test of two-sample parallel design, the sample size (n_1) (two-sided test) for group 1 is (Chow et al., 2008)

$$n_1 = kn_2$$

and the sample size for group 2 is

$$n_2 = (z_{\alpha/2} + z_\beta)^2(p_1(1 - p_1)/k + p_2(1 - p_2))/d^2$$

where p_1 and p_2: the proportions for group 1 (treatment group) and group 2 (control group) respectively; d: the difference between p_1 and p_2, d=p_1-p_2.

2.3.2.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n_1) (two-sided test) for group 1 is

$$n_1 = kn_2$$

and the sample size for group 2 is

$$n_2 = (z_\alpha + z_\beta)^2(p_1(1 - p_1)/k + p_2(1 - p_2))/(d - \delta)^2$$

where d=p_1-p_2. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, non-inferiority design, δ=-5%.

2.3.2.3 Equivalence Test

For equivalence design, the sample size (n_1) (two-sided test) for group 1 is

$$n_1 = kn_2$$

and the sample size for group 2 is

$$n_2 = (z_\alpha + z_{\beta/2})^2(p_1(1 - p_1)/k + p_2(1 - p_2))/(\delta - |d|)^2$$

where d=p_1-p_2.

2.3.3 Two-Sample Crossover Design

2.3.3.1 Significance Test

For the significance test of two-sample crossover design, the sample size (n) (two-sided test; 1:1 design) is

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

where σ: the standard deviation of between-proportion difference. For example, use a test drug A and control drug B to treat a disease. Take the control drug for 1 month, wash out for 3 weeks, and then take the test drug

for 1 month, and vice versa for the other group. If the test drug is expected to have a 10% (d=0.1) higher effective rate than the control drug, the test drug is considered to have promotional value. The standard deviation of the pre-test difference σ=0.3. Choose α=0.0001, β=0.1, two-sided significance test.

2.3.3.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) (two-sided test; 1:1 design) is

$$n = (z_\alpha + z_\beta)^2\sigma^2/(2(d - \delta)^2)$$

where σ: the standard deviation of between-proportion difference (it can be determined in a pre-experiment); δ: the margin of between-proportion. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, if the reshold is 5%, then δ=-0.05 for non-inferiority test.

2.3.3.3 Equivalence Test

For equivalence design, the sample size (n) (two-sided test; 1:1 equivalence test) is

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

2.3.4 One-Way Analysis of Variance

In One-Way Analysis of Variance (One-Way ANOVA), if there is one factor only and the levels (groups) k≥3. it is a Multiple-sample Parallel Design.

2.3.4.1 Pairwise Design

In this design, there are at least two groups and no control group is included. The sample size (n) is (Desu and Raghavarao, 1990; Fleiss, 1986)

$$n = \max\{n_{ij}\}$$

where

$$n_{ij} = (z_{\alpha/(2T)} + z_\beta)^2(p_1(1 - p_1) + p_2(1 - p_2))/d_{ij}^2$$

where $d_{ij}=\mu_i - \mu_j$. For example, the incidence reduction of two treatment drugs are 40% and 60% respectively and the value for the control is 15%, $d_{13}=\mu_1 - \mu_3$=0.4-0.15=0.25, $d_{23}=\mu_2 - \mu_3$=0.6-0.15=0.45, 1:1:1 parallel design. The sample size of each group is $n = \max\{n_1, n_2\}$.

2.3.4.2 Overall Between-Proportion Comparison

If we want to compare the overall difference of multiple proportions, the sample size (n) for each group is (Cohen, 1988):

$$n = 1641.6\lambda/(\sin^{-1}p_{max}^{0.5} - \sin^{-1}p_{min}^{0.5})^2$$

where p_{max} and p_{min}: the maximum proportion and the minimum proportion respectively. α, β: as described above. k is the number of groups. λ: obtained from Table 2. As an example, for α=0.05, β=0.1, and k=3, λ=12.65. For example, we want to study the therapeutic effect of different intensities of pharmaceutical interventions on hypertension levels. It is estimated that the strong intervention group has a treatment rate of 85%, the weak intervention group has a treatment rate of 65%, and the control group has a treatment rate of 20%. A two-sided test is required, α=0.05, β=0.1, and the ratio of the sample size of the three groups is 1:1:1

(that is, the number of cases in the three groups is equal, n). Use the method above to calculate the sample size required.

2.3.4.3 Williams Design

(1) Significance Test

For the significance test of Williams Design, the sample size (n) (two-sided test) is

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

where σ: the standard deviation of between-proportion difference; d: between-proportion difference; k: the number of groups (Common Williams designs are three-group designs (a 6×3 crossover design) and four-group designs (a 4×4 crossover design). When the experimental groups (k) are odd, the design result is a $2k\times k$ crossover design, and when the experimental groups are even, the design result is a $k\times k$ crossover design). For example, the incidence reduction of two dosages (dosages 1 and 2) of a drug and the control are 75%, 65% and 20%. We are interesting in the difference between dosage 1 and the control (d=0.75-0.2=0.55), the standard deviation of proportion difference between dosage 1 and control is σ=0.5 (i.e., 50%); use Williams three-crossover design (m=6).

(2) Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (n) (two-sided test) is

$$n = (z_{\alpha} + z_{\beta})^2\sigma^2/(m(d - \delta)^2)$$

where σ: the standard deviation of between-proportion difference (it can be determined in a pre-experiment); δ: the margin of between-proportion difference. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, if the reshold is 10%, then δ=-0.1 for for non-inferiority test.

(3) Equivalence Test

For equivalence design, the sample size (n) (two-sided test) is

$$n = (z_{\alpha} + z_{\beta/2})^2\sigma^2/(m\ (\delta - |d|)^2)$$

2.3.5 Relative Risk - Parallel Design

2.3.5.1 Significance Test

Known that the proportions of treatment group and control group are p_t and p_c respectively, for the significance test, the sample size (two-sided test) for control group is

$$n_c = (z_{\alpha/2} + z_{\beta})^2(1/(kp_t(1 - p_t)) + 1/(p_c(1 - p_c)))/(\log OR)^2$$

where k=n_t/n_c, and for treatment group, the sample size is n_t=kn_c.

OR means odds ratio

$$OR= (p_t/(1 - p_t))/(p_c/(1 - p_c))$$

OR>1 means that treatment has a significant effect and OR<1 no significant effect. For example, the incidence reduction of a treatment drug in pre-experiment is 35%, and the value for control drug is 20%; use OR as the assess index for treatment drug' effect; two-sided test.

2.3.5.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (two-sided test) for control group is

$$n_c = (z_\alpha + z_\beta)^2(1/(k\, p_t(1 - p_t)) + 1/(p_c(1 - p_c)))/(\log OR - \delta)^2$$

where δ: the margin. For the margin of 0.1 (i.e., 10%), δ=-0.1. Follow the example above.

2.3.5.3 Equivalence Test

For equivalence design, the sample size (two-sided test) is

$$n_c = (z_\alpha + z_{\beta/2})^2(1/(k\, p_t(1 - p_t)) + 1/(p_c(1 - p_c)))/(\delta - \log OR)^2$$

2.3.6 Relative Risk - Crossover Design

2.3.6.1 Significance Test

Known that the proportions of treatment group and control group are p_t and p_c respectively, for the significance test, the sample size (n) (1:1 design and two-sided test) for each group is

$$n = (z_{\alpha/2} + z_\beta)^2\sigma^2/(\log OR)^2$$

where σ: the standard deviation of between-proportion difference. For example, the incidence reduction of a treatment drug and the standard method are 40% (p_t=0.4) and 25% (p_c=0.25) respectively. Use OR as the assess index for treatment drug' effect: OR=(p_t/(1-p_t))/(p_c/(1-p_c)). σ=0.3 (30%). 1:1 crossover control design and two-sided test.

2.3.6.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, the sample size (1:1 design and two-sided test) for each group is

$$n = (z_\alpha + z_\beta)^2\sigma^2/(\log OR - \delta)^2$$

where δ: the margin. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For the margin of 0.1 (i.e., 10%), δ=-0.1 for non-inferiority test. Follow the example above.

2.3.6.3 Equivalence Test

For equivalence design, the sample size (1:1 design and two-sided test) for each group is

$$n = (z_\alpha + z_{\beta/2})^2\sigma^2/(\delta - |\log OR)|)^2$$

where δ: the margin. For the margin of 0.1 (i.e., 10%), δ=0.1. Follow the example above.

2.3.7 Intervention-Control Comparison

For two independent populations with proportions p_1 and p_2 respectively, if p_1-p_2 follows normal distribution or the sample size is large enough, then the sample size (n) for each population is (Chow et al., 2008; Fleiss et al., 2003):

$$n = (p_1 + p_2)*(2 - p_1 - p_2)/2)\ (z_\alpha + z_\beta)^2/(p_1 - p_2)^2$$

where p_1 and p_2 can be estimated in advance.

2.3.8 Cohort Study

Known the incidence probability p_0 (e.g., 0.1) and p_1 (e.g., 0.2) in control group and treatment group respectively, the sample size for each group is (Machin et al., 1997; Chow et al., 2003; Fleiss et al., 2003):

$$n = (z_\alpha((p_0 + p_1)(2 - p_0 - p_1)/2)^{0.5} + z_{2\beta}(p_0(1 - p_0) + p_1(1 - p_1))^{0.5})^2/(p_1 - p_0)^2$$

2.3.9 Two-Stage Sampling For Proportion

Two-stage sampling can be used in proportion estimation. In the first sampling, n_1 samples are taken and p_1 is calculated; in the second sampling, n-n_1 samples are taken, and thus

$$n = p_1q_1/v + (3 - 8\,p_1q_1)/(p_1q_1) - (1 - 3\,p_1q_1)/(v\,n_1)$$

where n: the sample size, q_1=1- p_1, v=$d^2/z_{\alpha/2}{}^2$, d: the expected error (difference) of p.

2.4 Exact Tests for Proportions

2.4.1 Binomial Test

The sample size estimation of the binomial distribution test is suitable for accurate testing of smaller sample's count data. As an example table, Table 3 is for α=0.05, from which the sample size and margin (r) of the binomial distribution test can be achieved. For example, in the preliminary trial, the cure rate of a new anti-tumor drug was 60% (p_1) and the cure rate of standard treatment was 40% (p_0). Single-group design, binomial distribution test, α=0.05, β=0.1. Find the number of cases for each group in Table 3.

Table 3 Table for binomial distribution test (in part; α=0.05).

p_0	p_1	β=0.1		β=0.2	
		r	n	r	n
0.05	0.2	4	38	3	27
0.1	0.25	9	55	7	40
0.15	0.3	14	64	11	48
0.2	0.35	21	77	16	56
0.25	0.4	27	83	21	62
0.3	0.45	35	93	26	67
0.35	0.5	41	96	30	68
0.4	0.55	45	94	35	71
0.45	0.6	52	98	38	70
0.5	0.65	54	93	41	69
0.55	0.7	58	92	45	70
0.6	0.75	58	85	43	62
0.65	0.8	55	75	41	55
0.7	0.85	54	69	39	49
0.75	0.9	46	55	38	45
0.8	0.95	39	44	27	30
0.05	0.25	3	25	2	16
0.1	0.3	6	33	5	25
0.15	0.35	9	38	7	28
0.2	0.4	14	47	11	35

0.25	0.45	17	49	13	36
0.3	0.5	21	53	16	39
0.35	0.55	24	53	19	41
0.4	0.6	28	56	22	42
0.45	0.65	30	54	24	42
0.5	0.7	32	53	23	37
0.55	0.75	33	50	25	37
0.6	0.8	32	45	26	36
0.65	0.85	32	42	24	31
0.7	0.9	30	37	23	28
0.75	0.95	25	29	20	23
0.8	1	13	14	13	14

2.4.2 Fisher's Exact Test

In a two-group parallel control design, if the theoretical number in the four-cell table is less than 5, or the total number of observations is less than 40, Fisher's exact test is required. To accurately estimate the sample size, we need to query Table 4 to obtain the sample size for different proportions. For example, in the preliminary trial, the cure rate of a new anti-tumor drug for treating a certain cancer was 40% (p_1), and the cure rate of standard treatment was 10% (p_0). The two groups were parallel controlled in a 1:1 design, two-sided difference test, α=0.05, β=0.1. Find the number of cases needed for each group.

Table 4 Sample size table for Fisher's exact test.

p_0	p_1	α=0.05		α=0.10	
		β=0.1	β=0.2	β=0.1	β=0.2
0.05	0.3	42	34	33	25
0.1	0.35	52	39	41	31
0.15	0.4	60	46	48	34
0.2	0.45	65	49	52	39
0.25	0.5	71	54	56	40
0.3	0.55	72	55	57	41
0.35	0.6	77	56	57	41
0.4	0.65	77	56	57	41
0.45	0.7	72	55	57	41
0.5	0.75	71	54	56	40
0.55	0.8	65	49	52	39
0.6	0.85	60	46	48	34
0.65	0.9	52	39	41	31
0.7	0.95	42	34	33	25
0.05	0.35	33	25	26	20
0.1	0.4	39	30	32	23
0.15	0.45	45	34	35	26
0.2	0.5	47	36	39	28

0.25	0.55	51	37	40	29
0.3	0.6	53	41	40	29
0.35	0.65	53	41	40	33
0.4	0.7	53	41	40	29
0.45	0.75	51	37	40	29
0.5	0.8	47	36	39	28
0.55	0.85	45	34	35	26
0.6	0.9	39	30	32	23
0.05	0.4	25	20	21	16
0.1	0.45	31	24	24	19
0.15	0.5	34	26	28	20
0.2	0.55	36	27	29	23
0.25	0.6	36	30	29	24
0.3	0.65	40	31	33	24
0.35	0.7	40	31	33	24
0.4	0.75	36	30	29	24
0.45	0.8	36	27	29	23
0.5	0.85	34	26	28	20
0.55	0.9	31	24	24	19
0.6	0.95	25	20	21	16

2.4.3 Optimal Multiple-Stage Designs for Single Arm Trials

2.4.3.1 Optimal Two-Stage Designs

In this design, we allow the experiment to terminate after a certain number of failures. The sample size for this design can be obtained by consulting the Table 5. For example, a new anti-tumor drug is undergoing phase II clinical trials. The effectiveness of standard treatment is 30% (p_0), and if the effectiveness of the new drug reaches 50% (p_1), it is considered to have clinical value. Optimal Two-Stage Designs, α=0.05, β=0.1. The result of this example is: 8/24, 24/63, 7/24, 21/53. It menas that there are in total of 24 cases in the first phase, and 8 of them are effective, then the second phase of the trial can be carried out. In the second phase, 7 more cases need to be continued to reach 24 cases. If at least 7 cases are effective, further research can be conducted.

Table 5 Table for Optimal Two-Stage Designs (α=0.05).

p_0	p_1	β=0.1				β=0.2			
0.05	0.2	1/21	4/41	1/29	4/38	0/10	3/29	0/13	3/27
0.1	0.25	2/21	10/66	3/31	9/55	2/18	7/43	2/22	7/40
0.2	0.35	8/37	22/83	8/42	21/77	5/22	19/72	6/31	15/53
0.3	0.45	13/40	40/110	27/77	33/88	9/27	30/81	16/46	25/65
0.4	0.55	19/45	49/104	24/62	45/94	11/26	40/84	28/59	34/70
0.5	0.65	22/42	60/105	28/57	54/93	15/28	48/83	39/66	40/68
0.6	0.75	21/34	64/95	48/72	57/84	17/27	46/67	18/30	43/62
0.7	0.85	18/25	61/79	33/44	53/68	14/19	46/59	16/23	39/49
0.8	0.95	16/19	37/42	31/35	35/40	7/9	26/29	7/9	26/29

0.05	0.25	0/9	3/30	0/15	3/25	0/9	2/17	0/12	2/16
0.1	0.3	2/18	6/35	2/22	6/33	1/10	5/29	1/15	5/25
0.2	0.4	4/19	15/54	5/24	13/45	3/13	12/43	4/18	10/33
0.3	0.5	8/24	24/63	7/24	21/53	5/15	18/46	6/19	16/39
0.4	0.6	11/25	32/66	12/29	27/54	7/16	23/46	17/34	20/39
0.5	0.7	13/24	35/61	14/27	32/53	8/15	26/43	12/23	23/37
0.6	0.8	12/19	37/53	15/26	32/45	7/11	30/43	8/13	25/35
0.7	0.9	11/15	29/36	13/18	26/32	4/6	22/27	19/23	21/26

2.4.3.2 Flexible Two-Stage Designs

This design gives multiple choices for the number of cases in the two stages. The sample size and boundary value of the optimized and flexible two-stage design can be found in Table 6. For example, a new anti-tumor drug is undergoing phase II clinical trials. The effectiveness of standard treatment is 20% (p_0), and if the effectiveness of the new drug reaches 40% (p_1), it is considered to have clinical value. Optimal Flexible Two-Stage Designs, α=0.05, β=0.1. The result is: 4/18-20, 5/21-24, 6/25 ----- 13/48, 14/49-51, 15/52-55. It means that the first stage of the study requires 18-20 cases, and if at least 4 cases are effective, the second stage trial can be carried out. In the second stage, we will continue to do 28 to 30 cases, bringing the total number to 48. If 13 cases are effective, further research can be conducted.

Table 6 Table for Flexible Two-Stage Designs (α=0.05).

	p_0	p_1	r_i-n_i	R_j-N_j
β=0.1	0.05	0.2	1/17-24	4/41-46,5/47-48
	0.1	0.25	2/21-24,3/25-28	9/57-61,10/62-64
	0.4	0.55	16/38-39,17/40-41,18/42-44,19/45	49/104-105,50/106-107,51/108-109,52/110-111
	0.5	0.65	21/40,22/41-42,23/43-44,24/45-46,25/47	59/103-104,60/105-106,61/107,62/108-109,63/110
	0.6	0.75	20/32-33,21/34,22/36-36,23/37/24/38-39	61/90-91,62/92,63/93-94,64/95,65/96-97
	0.7	0.85	17/24,18/26,19/26,20/27-28,21/29,22/30,23/31	57/73-74,58/75,59/76-77,60/78,61/79,62/80
	0.8	0.95	10/12,11/13-14,12/15,13/16, 14/17,15/18,16/19	35/40,36/41,37/42,38/43,39/44, 40/45-46,41/47
	0.05	0.25	0/8-13,1/14-15	2/24,3/25-31
	0.1	0.3	1/12-14,2/15-19	6/36-39,7/40-43
	0.2	0.4	4/18-20,5/21-24,6/25	13/48,14/49-51,15/52-55
	0.3	0.5	6/19-20,7/21-23,8/24-26	21/55,22/56-58,23/59-60,24/61-62
	0.4	0.6	8/20,9/21-22,10/23-24,11/25-26,12/27	28/58,29/59-60,30/61-62,31/63,32/64-65
	0.5	0.7	10/19-20,11/21,12/22-23,13/24-25,14/26	33/55-56,34/57-58,35/59,36/60-61,37/62
	0.6	0.8	11/17-18,12/19,13/20-21,14/22,15/23,16/24	34/48-49,35/50-51,36/52,37/53-54,38/55
	0.7	0.9	7/10,8/11,9/12-13,10/14,11/15,12/16,13/17	27/34/28/35,29/36,30/37-38,31/39,32/40,33/41
	0.05	0.2	0/10-12,1/13-17	3/27-34
	0.1	0.25	1/13-15,2/16-20	6/40,7/41-45,8/46-47
	0.2	0.35	4/18-21,5/22-24,6/25	17/62-64,18/65-69,
	0.2	0.35	6/31,7/32-34,8/35-38	22/82-85,23/86-89
	0.3	0.45	7/23,8/24-25,9/26-29,10/30	27/73,28/74-76,29/77-78,30/79-80

	0.3	0.45	11/35-36,12/37-39,13/40-42	36/98-99,37/100-102,38/103-104,39/105
	0.4	0.55	11/25-26,12/27-29,13/30-31,14/32	37/78,38/79-80,39/81-82,40/83-85
β=0.2	0.5	0.65	12/23,13/24-25,14/26-27,15/28-29,16/30	45/77-78,46/79-80,47/81-82,48/83,49/84
	0.6	0.75	14/22-23,15/24,16/25	46/68,47/69,48/70-71
	0.7	0.85	9/13,10/14,11/15,12/16-17,13/18,14/19,15/20	44/56-57,45/58,46/59,47/60,48/61-62,49/63
	0.8	0.95	7/9,8/10,9/11,10/12, 11/13,12/14,13/15,14/16	25/28,26/29,27/30,28/31-32, 29/33,30/34,31/35
	0.05	0.25	0/5-10,1/11-12	2/17-22,3/23-24
	0.1	0.3	1/8-12,2/13-15	4/26,5/27-32,6/33
	0.2	0.4	2/10-12,3/13-15,4/16-17	10/33-35,11/36-40
	0.3	0.5	3/11,4/12-14,5/15-16/6/17-18	16/40-41,16/42-44,18/45-46,18/47
	0.4	0.6	5/12-13,6/14,7/15-16,8/17-19	22/44-45,23/46-47,24/48-49,25/50,26/51
	0.5	0.7	5/10,6/11-12,7/13-14,8/15,9/16-17	25/42,26/43-44,27-45,28/46-47,29/48,30/49
	0.6	0.8	5/8-9,6/10,7/11,8/12-13,9/14-15	25/35-36,26/37,27/38,28/39-40,29/41,30/42
	0.7	0.9	4/6,5/7,6/8,7/9,8/10-11,9/12,10/13	22/27,23/28-29,24/30,25/31,26/32-33,27/34

2.4.3.3 Optimal Three-Stage Designs

It is basically the same as the two-stage one. The sample content and margin can be found in Table 7. For example, a new anti-tumor drug is undergoing phase II clinical trials. The effectiveness of standard treatment is 20% (p_0), and if the effectiveness of the new drug reaches 40% (p_1), it is considered to have clinical value. Optimal Three-Stage Designs, α=0.05, β=0.1. The result is: 3/17 --> 7/30 --> 14/50. It means that the first stage of the study requires 17 cases, and if at least 1 case is effective, the second stage trial can be carried out. In the second stage, we will continue to do 13 cases, bringing the total number to 30. If at least 7 cases are effective, the third stage trial can be carried out. In the third stage, we will continue to do 20 cases, bringing the total number to 50 cases. If at least 14 cases are effective, continue the further reaserch.

Table 7 Table for Optimal Three-Stage Designs (α=0.05).

		β=0.1					β=0.2		
p_0	p_1	S_1: r_1/n_1	S_2: r_2/n_1n_2	S_3: $r_3/n_1n_2n_3$	p_0	p_1	S_1: r_1/n_1	S_2: r_2/n_1n_2	S_3: $r_3/n_1n_2n_3$
0.05	0.2	0/14	2/29	4/43	0.1	0.25	1/13	3/24	8/53
0.1	0.25	1/17	4/34	10/66	0.15	0.3	2/15	6/33	13/62
0.15	0.3	3/23	8/46	16/77	0.2	0.35	3/17	9/37	18/68
0.2	0.35	5/27	11/49	23/88	0.25	0.4	4/17	12/42	25/79
0.25	0.4	6/26	15/54	32/103	0.3	0.45	5/18	14/41	31/84
0.3	0.45	8/29	19/57	38/104	0.35	0.5	6/19	17/43	34/80
0.35	0.5	9/28	23/60	45/108	0.4	0.55	7/19	19/43	39/82
0.4	0.55	12/31	28/64	54/116	0.45	0.6	8/19	21/42	45/86
0.45	0.6	13/30	29/60	58/112	0.5	0.65	8/17	21/39	49/85
0.5	0.65	14/29	34/63	62/109	0.55	0.7	7/14	23/39	49/78
0.55	0.7	15/28	36/61	65/105	0.6	0.75	8/14	23/36	52/77
0.6	0.75	14/24	36/56	70/105	0.65	0.8	8/13	27/38	52/72
0.7	0.85	12/18	28/38	58/75	0.65	0.8	16/25	35/50	66/92
0.75	0.9	10/14	23/29	55/67	0.7	0.85	4/7	11/16	44/56
0.8	0.95	6/8	24/28	41/47	0.75	0.9	9/12	21/26	39/47

0.05	0.25	0/10	1/17	3/30	0.8	0.95	2/3	16/19	35/40
0.1	0.3	1/13	3/23	7/45	0.05	0.25	0/8	1/13	2/19
0.15	0.35	2/15	5/27	11/51	0.1	0.3	0/6	2/17	5/29
0.2	0.4	3/17	7/30	14/50	0.15	0.35	1/9	4/21	8/35
0.25	0.45	4/18	10/33	19/58	0.2	0.4	1/8	5/22	11/38
0.3	0.5	4/16	11/32	23/60	0.25	0.45	2/10	6/20	16/48
0.35	0.55	6/18	15/38	27/62	0.3	0.5	3/11	7/21	18/46
0.4	0.6	6/16	17/38	32/66	0.35	0.55	3/10	9/23	21/47
0.45	0.65	6/15	17/34	34/63	0.4	0.6	3/9	10/23	23/46
0.5	0.7	7/15	19/34	38/65	0.45	0.65	3/8	10/20	29/54
0.55	0.75	7/14	16/27	36/56	0.5	0.7	4/9	13/23	29/49
0.6	0.8	6/11	19/29	38/55	0.55	0.75	6/11	14/23	28/43
0.65	0.85	6/10	16/23	35/47	0.6	0.8	5/9	12/48	28/40
0.7	0.9	6/9	16/21	31/39	0.65	0.85	5/8	13/18	27/36
0.75	0.95	6/8	13/16	24/28	0.7	0.9	3/5	10/13	25/31
0.05	0.2	0/10	1/19	3/30	0.75	0.95	1/2	9/11	19/22

2.4.3.4 Minimum Three-Stage Designs

This method needs the minimal sample size. For example, a new anti-tumor drug is undergoing phase II clinical trials. The effectiveness of standard treatment is 20% (p_0), and if the effectiveness of the new drug reaches 40% (p_1), it is considered to have clinical value. Minimum Three-Stage Designs, α=0.05, β=0.1. The result is: 2/16 --> 6/28 --> 13/45. It means that the first stage of the study requires 16 cases, and if at least 2 cases are effective, the second stage trial can be carried out. In the second stage, we will continue to do 12 cases, bringing the total number to 28. If at least 6 cases are effective, the third stage trial can be carried out. In the third stage, we will continue to do 17 cases, bringing the total number to 45 cases. If at least 13 cases are effective, continue the further reaserch.

Table 8 Table for Minimum Three-Stage Designs (α=0.05).

β=0.1					β=0.2				
p_0	p_1	S_1: r_1/n_1	S_2: r_2/n_1n_2	S_3: $r_3/n_1n_2n_3$	p_0	p_1	S_1: r_1/n_1	S_2: r_2/n_1n_2	S_3: $r_3/n_1n_2n_3$
0.05	0.2	0/23	1/30	4/38	0.05	0.2	0/14	1/20	3/27
0.1	0.25	1/21	4/39	9/55	0.1	0.25	1/17	3/30	7/40
0.15	0.3	4/35	8/51	14/64	0.15	0.3	2/19	6/36	11/48
0.2	0.35	16/65	19/72	20/74	0.2	0.35	3/22	7/35	15/53
0.25	0.4	9/47	17/67	27/83	0.25	0.4	7/30	12/42	20/60
0.3	0.45	12/46	25/73	33/88	0.3	0.45	8/29	14/42	25/65
0.35	0.5	11/36	22/60	40/94	0.35	0.5	10/33	18/48	29/66
0.4	0.55	20/55	32/77	45/94	0.4	0.55	13/33	30/63	34/70
0.45	0.6	26/58	47/90	50/95	0.45	0.6	13/32	25/53	38/70
0.5	0.65	19/43	34/67	54/93	0.5	0.65	18/36	36/62	40/68
0.55	0.7	23/43	42/84	45/89	0.55	0.7	18/33	41/64	42/66
0.6	0.75	18/46	50/75	57/84	0.6	0.75	19/32	40/58	42/61

0.7	0.85	13/20	31/42	43/68	0.65	0.8	16/26	27/40	41/55
0.75	0.9	12/17	23/30	45/54	0.65	0.8	25/41	37/56	55/75
0.8	0.95	16/20	31/35	35/40	0.7	0.85	11/17	16/24	39/49
0.05	0.25	0/15	1/21	3/25	0.75	0.9	8/12	16/21	33/39
0.1	0.3	0/14	2/22	6/33	0.8	0.95	7/9	16/19	26/29
0.15	0.35	1/16	4/28	9/38	0.05	0.25	0/12	1/15	2/16
0.2	0.4	2/16	6/28	13/45	0.1	0.3	0/11	2/19	5/25
0.25	0.45	4/21	9/35	17/45	0.15	0.35	1/12	3/19	7/28
0.3	0.5	5/20	12/36	21/53	0.2	0.4	2/13	5/22	10/33
0.35	0.55	10/34	17/45	24/53	0.25	0.45	3/15	6/23	13/36
0.4	0.6	7/20	17/39	27/54	0.3	0.5	3/13	8/24	16/39
0.45	0.65	15/32	28/51	29/53	0.35	0.55	4/14	9/24	18/39
0.5	0.7	8/18	18/34	32/53	0.4	0.6	4/12	11/25	21/41
0.55	0.75	12/22	21/35	32/49	0.45	0.65	6/15	12/24	22/39
0.6	0.8	15/26	24/37	32/45	0.5	0.7	7/16	13/25	23/37
0.7	0.9	5/9	12/17	26/32	0.55	0.75	8/15	14/23	24/36
0.75	0.95	9/12	19/22	22/26	0.6	0.8	9/15	23/32	24/34
-	-	-	-	-	0.65	0.85	6/10	13/18	23/30
-	-	-	-	-	0.65	0.85	16/24	28/37	30/40
-	-	-	-	-	0.7	0.9	4/7	19/23	20/25
-	-	-	-	-	0.75	0.95	6/8	14/16	17/20

2.4.3.5 Flexible Designs for Multiple-Arm Trials

In this design, we need to specify a clinically meaningful boundary value $[-\delta, \delta]$ in advance. If the difference in proportions is greater than δ, then the group with the larger proportion will be selected. If the proportion difference is less than or equal to δ, other factors need to be considered in the selection. This design is not to compare the advantages and disadvantages between groups, but to maintain the existence of advantageous treatments as accurately as possible for further research.

(1) Flexible Designs for Two-Arm Trials

Suppose that λ is a pre-specified threshold, and δ=0.05. The sample sizes of the two groups of flexible designs can be found in Table 9 when ρ=0 or ρ=0.5. For example, the effectiveness of standard treatment is 25% (p_0), and the effectiveness of the new drug is 40% (p_1). Flexible Designs for Two-Arm Trials. Suppose that ρ=0, λ=0.9. The result is: 71. Each group needs 71 cases.

Table 9 Table for Flexible Designs for Two-Arm Trials.

p_0	p_1	λ=0.9, ρ=0	λ=0.8	λ=0.9, ρ=0.5
0.05	0.2	32	13	16
0.1	0.25	38	15	27
0.15	0.3	0.53	17	31
0.2	0.35	57	19	34
0.25	0.4	71	31	36
0.3	0.45	73	32	38
0.35	0.5	75	32	46
0.4	0.55	76	33	47

(2) Flexible Designs for Multiple-Arm Trials

Suppose that λ is a pre-specified threshold, and δ=0.05. The sample sizes of the multiple groups of flexible designs can be found in Table 10 when λ=0.8, 0.9, ρ=0, 0.5, r=3, 4, and d=0.2, 0.3, 0.4, 0.5.

Table 10 Table for Flexible Designs for Multiple-Arm Trials.

λ	d	ρ=0, r=3	ρ=0, r=4	ρ=0.5, r=3	ρ=0.5, r=4
0.8	0.2	18	31	13	16
0.8	0.3	38	54	26	32
0.8	0.4	54	73	31	39
0.8	0.5	58	78	34	50
0.9	0.2	39	53	30	34
0.9	0.3	77	95	51	59
0.9	0.4	98	119	68	78
0.9	0.5	115	147	73	93

d: Proportion's magin; *r*: number of arms.

2.5 Tests for Goodness-of-Fit and Contingency Tables

2.5.1 Test for Goodness-of-Fit

For Goodness-of-Fit test, the sample size (n) (One-group design and two-sided test) is

$$n = \delta_{\alpha,\beta}(\sum_{k=1}^{r}(p_k - p_{k,0})^2/p_{k,0})^{-1}$$

where p_k: the proportion of the category k, k=1,2,…,r; $p_{k,0}$: the proportion of the category k in the literature. $\delta_{\alpha,\beta}$: calculated from $F_{r\text{-}1}(\chi^2_{\alpha,r-1}|\delta)=\beta$, where $\delta=\lim_{n\to\infty}\sum_{k=1}^{r} n(p_k - p_{k,0})^2/p_{k,0}$. For example, to analyze the effect of a drug in the pilot study, preliminary trials have shown that the proportions of marked effect, effect and ineffect of the drug in treating the disease are about 20%, 55% and 25% respectively. According to literature reports, the proportions of marked effect, effect and ineffect of existing antihypertensive drugs are 15%, 50% and 20%, respectively.

2.5.2 Test for Independence - Single Stratum

For $r\times c$ contingency table data (two-way) without stratum, the following methods are commonly used for sample size estimation.

2.5.2.1 Pearson's Test

For Pearson's test, the sample size (n) (Two-group parallel design and two-sided test) is

$$n = \delta_{\alpha,\beta}(\sum_{i=1}^{r}\sum_{j=1}^{c}\left(p_{ij} - p_i p_j\right)^2/(p_i p_j))^{-1}$$

where r,c: the number of rows and columns in the table respectively; p_{ij}: the proportion of row category i and category j, i=1,2,…,r; j=1,2,…,c; $\delta_{\alpha,\beta}$: calculated from $F_{r\text{-}1,c\text{-}1}(\chi^2_{\alpha,(r-1)(c-1)}|\delta)=\beta$, and $\delta=\lim_{n\to\infty}\sum_{i=1}^{r}\sum_{j=1}^{c} n\left(p_{ij} - p_i p_j\right)^2/(p_i p_j)$. For example, to analyze the effect of a drug, preliminary trials have shown that the marked effective proportion, effective proportion and ineffective proportion of the drug in

treating the disease are about 15%, 58% and 25% respectively, and the marked effective proportion, effective proportion and ineffective proportion of the control are about 8%, 30% and 16% respectively.

2.5.2.2 Likelihood Ratio Test

The method is the same as Pearson's Test.

2.5.3 Test for Independence - Multiple Strata

The test for independence of multiple strata is used in multi-center (multi-stratum) clinical trials. The later can not only guarantee the repeatability and representativeness of experimental results, but also facilitate the selection of subjects within the expected time. Multi-center clinical trials produce multi-level contingency table data. When the response rate is binary data, the Cochran-Mantel-Haenszel Test is a commonly used method. Suppose that $n_{h,ij}$ is the number of response j in layer h (i.e., center h) after processing i (i.e., group i), $p_{h,\,ij}$ is the proportion of response j in layer h after processing i . The sample size (n) is

$$n = (z_{\alpha/2} + z_{\beta})^2/\delta^2$$

where

$$\delta = |(\sum_{h=1}^{H} \pi_h(p_{h,12} - p_{h,1.}p_{h,.2})/\,(\sum_{h=1}^{H} \pi_h p_{h,1.}p_{h,.2}p_{h,2.}p_{h,.1}))^{0.5}|$$

and π_h=n_h/n. For example, make a multi-center clinical trial for a drug and the control and observe the proportion of adverse events. Three strata are used (H=3). The data are as follows

Strata	Groups	Adverse events No	Adverse events Yes	Total
1	Treatment	0.25	0.25	0.50
	Control	0.15	0.35	0.50
2	Treatment	0.20	0.30	0.50
	Control	0.30	0.20	0.50
3	Treatment	0.35	0.15	0.50
	Control	0.15	0.35	0.50

Two-group 1:1 parallel design, and two-sided test. π_h=1/3.

2.5.4 Categorical Shift Test

In clinical trials, to study the changes in the data of the two categories before and after the trial, McNemar test and Stuart-Maxwell test are usually used.

2.5.4.1 McNemar Test

The McNemar test is suitable for comparisons before and after binary variables. For McNemar test, the sample size (n) is

$$n = (z_{\alpha/2}\,(\varphi+1) + z_{\beta}((\varphi+1)^2 - (\varphi-1)^2\pi_{Disordant})^{0.5})^2/((\varphi-1)^2\pi_{Disordant})$$

where $\varphi=p_{01}/p_{10}$, $\pi_{Disordant}= p_{01}+ p_{10}$, $p_{01}=P(x_1=0, x_2=1)$, $p_{10}=P(x_1=1, x_2=0)$. For example, use a drug to treat the disease. $p_{10}=0.6$ (60%), $p_{01}=0.2$ (20%). Two-sided significance test.

2.5.4.2 Stuart-Maxwell Test

The Stuart-Maxwell test is suitable for comparisons before and after multiple categorical variables. For Stuart-Maxwell test, the sample size (n) is

$$n = \delta_{\alpha,\beta}(\sum_{i<j}(p_{ij} - p_{ji})^2/(p_{ij} + p_{ji}))^{-1}$$

where $p_{ij}=n_{ij}/\sum\sum n_{ij}$, $\delta=\lim_{n\to\infty} n\sum_{i<j}(p_{ij} - p_{ji})^2/(p_{ij} + p_{ji})$. $F_{r(r-1)/2}(\chi_{\alpha,r(r-1)/2}|\delta)=\beta$. For example, to study the possibility of the effect-changing trend of using a drug to treat the disease, and the data are as follows

Before treatment	After treatment response 1	After treatment response 2	…	After treatment response r
Response 1	n_{11}	n_{12}	…	n_{1r}
Response 2	n_{21}	n_{22}	…	n_{22}
…	…	…	…	…
Response r	n_{r1}	n_{r2}	…	n_{rr}

2.5.5 Carry-Over Effect Test

Residual effects refer to some reasons caused by the previous stage of treatment (such as the withdrawal effect caused by drug resistance, psychological effects, and legacy effects caused by changes in the patient's physical condition due to medication) that interfere with the treatment effect of the next stage. To understand the residual effect, the sample size (n) is

$$n = (z_{\alpha/2} + z_{\beta})^2(\sigma_1^2 + \sigma_2^2)/\gamma^2$$

where σ_1: the standard deviation for from A to B; σ_2: the standard deviation for from B to A; γ: the difference of residual effects between the $A\rightarrow B$ and $B\rightarrow A$ orders. For example, in a trial for the drugs A and B, $\gamma=0.6$, $\sigma_1=3.6$, $\sigma_2=3.9$.

2.6 Time-to-Event

The result of some experiments is the time of an event. The time from the observation of the event is called the time-to-event. If the end point of the study is death, the time to the event is called the survival time. Survival time refers to the time elapsed from a certain starting point to the occurrence of an event. Survival probability indicates the probability that subjects who survived at the beginning of a unit period are still alive at the end of the period. Survival rate refers to the probability that the research object is still alive after a period of time, that is, the probability that the survival time is greater than or equal to. The survival function $S(t)$ (Survival

distribution function) is also called the cumulative survival rate, which is the probability that time is greater than a certain point in time. The death probability function $F(t)$ is referred to as death probability for short, which represents the death probability of an observed object from the beginning of observation to time t, and its relationship with the survival function is $F(t)$=1-$S(t)$. Hazard function refers to the probability of the surviving subjects at time t dying immediately after time t, which is a conditional probability.

2.6.1 Exponential Model

Survival time usually does not follow the normal distribution, and sometimes it approximately follows the exponential distribution, Weibull distribution, Gompertz distribution, etc. In most cases, it does not follow any regular distribution type.

2.6.1.1 Significance Test

Suppose that survival time follows the exponential distribution. We want to test the significance of difference between two groups of endpoints (survival rates). The sample size is

$$n_2 = (z_{\alpha/2} + z_{\beta})^2(\sigma^2(\lambda_1)/k + \sigma^2(\lambda_2))/(\lambda_1 - \lambda_2)^2$$

where n_1 and n_2: the sample size for group 1 and group 2 respectively, k=n_1/n_2; λ_1 and λ_2: the hazard ratio of group 1 and group 2 respectively; σ: standard deviation; T: The expected time for trials, that is, the time from the start to the end of trials; T_0: The expected time for all subjects to be enrolled (keep consistent with T), and

$$\sigma^2(\lambda_i) = \gamma^2 \left(1 + \lambda_i e^{-\lambda_i T}\left(1 - e^{-(\lambda_i-\gamma)T_0}\right)/\left((\lambda_i - \gamma)\left(1 - e^{-\lambda_i T_0}\right)\right)\right)^{-1}$$

For example, to study the effect of two treatment methods on the time to transformation of malignant tumor to cancer. The observation time lasted for 5 years (T=5, T_0=1). Assume that the hazard ratios of the two groups are λ_1=0.5 and λ_2=0.8, respectively. Estimate the sample size for each group.

2.6.1.2 Non-inferiority or Superiority Test

For non-inferiority or superiority test, we want to test if the difference between two groups of terminals (survival rates) is non-inferior or superior to the known margin. The sample size is

$$n_2 = (z_{\alpha} + z_{\beta})^2(\sigma^2(\lambda_1)/k + \sigma^2(\lambda_2))/(d - \delta)^2$$

where d: the difference between endpoints (survival rates) of two groups; δ: the margin; k=n_1/n_2; λ_1 and λ_2: the hazard ratio of group 1 and group 2 respectively; σ: standard deviation, and

$$\sigma^2(\lambda_i) = \gamma^2 \left(1 + \lambda_i e^{-\lambda_i T}\left(1 - e^{-(\lambda_i-\gamma)T_0}\right)/\left((\lambda_i - \gamma)\left(1 - e^{-\lambda_i T_0}\right)\right)\right)^{-1}$$

δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, the margin is 0.1 (i.e., 10%), δ=0.1 for superiority test. Follow the example above.

2.6.1.3 Equivalence Test

In equivalence design, we hope to test the equivalence of two groups of endpoints (survival rates). The sample size is

$$n_2 = (z_{\alpha} + z_{\beta/2})^2(\sigma^2(\lambda_1)/k + \sigma^2(\lambda_2))/(\delta - |d|)^2$$

where d: the difference between terminals (survival rates) of two groups; δ: the margin; k=n_1/n_2; λ_1 and λ_2: the

hazard ratio of group 1 and group 2 respectively; σ: standard deviation, and

$$\sigma^2(\lambda_i) = \gamma^2 \left(1 + \lambda_i e^{-\lambda_i T}\left(1 - e^{-(\lambda_i - \gamma)T_0}\right)/\left((\lambda_i - \gamma)\left(1 - e^{-\lambda_i T_0}\right)\right)\right)^{-1}$$

For example, the margin is 0.1 (i.e., 10%), δ=0.1. Follow the example above.

2.6.2 Cox Proportional Hazards Model

2.6.2.1 Significance Test

In significance test, adopt 1:1 two groups of parallel control design. Based on the Cox Proportional Hazards Model for survival analysis, test whether the difference between the two groups of endpoints is significant. The sample size is

$$n = (z_{1-\alpha/2} + z_{1-\beta})^2/(\log^2 b\, p_1 p_2 d)$$

where p_1 and p_2: the hazard rate of two groups respectively; b: the hazard ratio of two groups; d: the occurrence rate of specified event. For example, compare the therapeutic effect of a new method and a traditional method. In the pilot test, the hazard ratio of the traditional method and the new method is b=3, 70% of the patients will be observed local infection (d=0.7), when p_1=0.4, p_2=0.4, two groups 1:1 parallel control. Make significance test. Each group requires n cases.

2.6.2.2 Non-inferiority or Superiority Test

In non-inferiority or superiority test, adopt 1:1 two groups of parallel control design. Based on the Cox Proportional Hazards Model for survival analysis, test whether the difference between the two groups of endpoints is non-inferior or superior to the known margin. The sample size is

$$n = (z_{1-\alpha} + z_{1-\beta})^2/(\log b - \delta)^2 p_1 p_2 d)$$

where δ: the margin of log$b$. δ<0 if non-inferiority test is made and δ>0 if superiority test is made. For example, take superiority test, δ=0.4 for superiority test.

2.6.2.3 Equivalence Test

In equivalence design, adopt 1:1 two groups of parallel control design. Based on the Cox Proportional Hazards Model for survival analysis, test the equivalence of two groups of endpoints. The sample size is

$$n = (z_{1-\alpha} + z_{1-\beta/2})^2/(\delta - |\log b|)^2 p_1 p_2 d)$$

2.6.3 Logrank Test

The survival analysis based on the Logrank test (also known as the time series test), is based on the premise that the null hypothesis is established, and the difference between the actual number of deaths with two survival times and the theoretical number of deaths (expected number of deaths) calculated based on the number of initial observations and the theoretical death probability should not be large; if the difference is large, the null hypothesis is invalid, and the difference between the two survival curves can be considered to be statistically significant. For significance test, the sample size is

$$n = 2d/(p_1 + p_2)$$

where

$$d = (z_{1-\alpha/2} + z_{1-\beta})^2 (\sum_{i=1}^{N} w_i^2 \rho_i \eta_i)/(\sum_{i=1}^{N} w_i \rho_i \gamma_i)^2$$

Corresponding to w_i=1, n_i, and $n_i^{0.5}$, the test is Logrank test, Wilcoxon test and Tarone-Ware test. For example, in a trial of two years, it is assumed that the annual hazard rate of the experimental group is 0.8 (the annual event rate is $1-e^{-0.8}$), the annual hazard rate of the control group is 0.4 (the annual event rate is $1-e^{-0.4}$), and the annual loss rate was 2%, the annual non-compliance rate was 5%, and 8% of the patients in the control group chose other treatments (drop-in) similar to those in the experimental group. The total event rate was 86% in the treatment group and 60% in the control group. Calculate the sample size, n.

2.7 Group Sequential Methods

The traditional randomized controlled trial design requires the sample size to be determined before the trial begins. Sequential design usually does not fix the sample size in advance, but according to the order in which the subjects enter the experiment, one analysis is performed after one experiment is done, and the experiment is stopped immediately once the expected result is achieved. The group sequential design allows interim analysis of the accumulated data during the trial, such as evaluating the effectiveness and safety of the trial drug, and if there is enough evidence to prove that the trial drug is effective or ineffective, the trial can be terminated early. Compared with the traditional experimental design, because the interim analysis provides the possibility to end the trial early, the group sequential trial can often save the trial sample size, shorten the trial period, save money, and is more in line with the ethical requirements. In addition, the interim evaluation of data by group sequential design can also enable researchers to discover problems in the trial as early as possible, which is conducive to improving the quality of the trial.

The most used group sequential experiments are staged experiments. It is required to divide the whole experiment into k consecutive stages, and in each stage, $2n$ subjects join the experiment, and are randomly assigned to the experimental group and the control group, and each group has n subjects. When the ith ($i \le k$) stage test is over, the experimental results from stage 1 to stage i are accumulated for statistical analysis. If H_0 is rejected, the test can be ended, otherwise continue to the next stage of the test. If H_0 cannot be rejected after the end of the last kth stage, H_0 is acceptable. In group sequential experiments, repeated significance tests are required, and the significance level of each stage needs to be adjusted, and the adjusted significance level becomes the nominal significance level. In group sequential design, there are two conceptual ways of dividing time points, one is calendar time and the other is information time. Calendar time is based on the progress of the trial duration to determine when to conduct interim analysis; the meaning of information time refers to the percentage of the sample size observed at a certain observation point in the total sample size of the plan, measured by the amount of information that can be observed with which to decide when to conduct an interim analysis.

2.7.1 Pocock's Test

In Pocock's Test, the same margin and nominal significance level were used for each stage. For k-stage Pocock's Test, the total sample size is

$$n_{max} = R_p(k,\alpha,\beta)\ (z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)/(\mu_1 - \mu_2)^2$$

and the sample size for each stage is n_{max}/k. $R_p(k,\alpha,\beta)$ is obtained from the Table 11. Margin values for kth stage are listed in Table 12. For example, for k=5 and α=0.05, it is 2.413. So in a group sequential design with a significance level of 0.05 in 5 stages, a margin of 2.413 is used in each stage. Only when the nominal significance level in each stage is less than 0.0158218 can H_0 be rejected. As an example, for a 5-phase group sequential trial comparing the efficacy of a drug and a control, according to the pilot test, the overall standard

deviation is 3 ($\sigma^2=\sigma_1^2=\sigma_2^2=9$), $\mu_1-\mu_2=1$; Pocock design; calculate the sample size for each stage, n_{max}/k.

Table 11 $R_p(k,\alpha,\beta)$ table.

k	β=0.10			β=0.20		
	α=0.01	α=0.05	α=0.10	α=0.01	α=0.05	α=0.10
1	1	1	1	1	1	1
2	1.084	1.1	1.11	1.092	1.11	1.121
3	1.125	1.151	1.166	1.137	1.166	1.184
4	1.152	1.183	1.202	1.166	1.202	1.224
5	1.17	1.207	1.228	1.187	1.229	1.254
6	1.185	1.225	1.249	1.203	1.249	1.277
7	1.197	1.239	1.266	1.216	1.265	1.296
8	1.206	1.252	1.28	1.226	1.279	1.311
9	1.215	1.262	1.292	1.236	1.291	1.325
10	1.222	1.271	1.302	1.243	1.301	1.337
11	1.228	1.279	1.312	1.25	1.31	1.348
12	1.234	1.287	1.32	1.257	1.318	1.357
15	1.248	1.305	1.341	1.272	1.338	1.381
20	1.264	1.327	1.367	1.291	1.363	1.411

Table 12 Margin value table for Pocock's Test.

k	α=0.01	α=0.05	α=0.10
1	2.576	1.96	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	2.986	2.413	2.122
6	3.023	2.453	2.164
7	3.053	2.485	2.197
8	3.078	2.512	2.225
9	3.099	2.535	2.249
10	3.117	2.555	2.27
11	3.133	2.572	2.288
12	3.147	2.588	2.304
15	3.182	2.626	2.344
20	3.225	2.672	2.392

2.7.2 O'Brien and Fleming Test

This method adopts different margins for different stages, and the margin is set higher in the early stage and lower in the later stage. Table 13 lists the margin of the last stage of each stage. For example, the margin of the last stage of 13 stages is 2.04 (α=0.05), and the difference table can be used for the first 4 stages. The values are 4.562, 3.226, 2.634 and 2.281, respectively.

Table 13 Margin value table for O'Brien and Fleming Test.

k	α=0.01	α=0.05	α=0.10
1	2.576	1.96	1.645
2	2.58	1.977	1.678
3	2.595	2.004	1.71
4	2.609	2.024	1.733
5	2.621	2.04	1.751
6	2.631	2.053	1.765
7	2.64	2.063	1.776
8	2.648	2.072	1.786
9	2.654	2.08	1.794
10	1.66	2.087	1.801
11	2.665	2.092	1.807
12	2.67	2.098	1.813
15	2.681	2.11	1.826
20	2.695	2.126	1.842

For k-stage O'Brien and Fleming Test, the total sample size is

$$n_{max} = R_b\,(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)/(\mu_1 - \mu_2)^2$$

and the sample size for each stage is n_{max}/k. R_b is obtained from the Table 14.

Table 14 R_b table.

k	β=0.10			β=0.20		
	α=0.01	α=0.05	α=0.10	α=0.01	α=0.05	α=0.10
1	1	1	1	1	1	1
2	1.001	1.007	1.014	1.001	1.008	1.016
3	1.006	1.016	1.025	1.007	1.017	1.027
4	1.01	1.022	1.032	1.011	1.024	1.035
5	1.014	1.026	1.037	1.015	1.028	1.04
6	1.016	1.03	1.041	1.017	1.032	1.044
7	1.018	1.032	1.044	1.019	1.035	1.047
8	1.02	1.034	1.046	1.021	1.037	1.049
9	1.021	1.036	1.048	1.022	1.038	1.051
10	1.022	1.037	1.049	1.024	1.04	1.053
11	1.023	1.039	1.051	1.025	1.041	1.054
12	1.024	1.04	1.052	1.026	1.042	1.055
15	1.026	1.042	1.054	1.028	1.045	1.058
20	1.029	1.045	1.057	1.03	1.047	1.061

2.7.3 Wang and Tsiatis Test

It is an extension of Pocock Test and O'Brien and Fleming Test. It is the Pocock Test when ρ=0 and τ=0. It is the O'Brien and Fleming Test when ρ=0.5 and τ=0. Table 15 lists the margins for each stage.

Table 15 Margin value table for Wang and Tsiatis Test.

k	δ=0.10	δ=0.25	δ=0.40
1	1.96	1.96	1.96
2	1.994	2.038	2.111
3	2.026	2.083	2.186
4	2.05	2.113	2.233
5	2.068	2.136	2.267
6	2.083	2.154	2.292
7	2.094	2.168	2.313
8	2.104	2.18	2.329
9	2.113	2.19	2.343
10	2.12	2.199	2.355
11	2.126	2.206	2.366
12	2.132	2.213	2.375
15	2.146	2.229	2.397
20	2.162	2.248	2.423

For k-stage Wang and Tsiatis Test, the sample size is

$$n = R_{wt}\,(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)/(\mu_1 - \mu_2)^2$$

R_{wt} is obtained from the Table 16.

Table 16 R_{wt} table.

k	β=0.10			β=0.20		
	α=0.01	α=0.05	α=0.10	α=0.01	α=0.05	α=0.10
1	1	1	1	1	1	1
2	1.014	1.034	1.068	1.016	1.038	1.075
3	1.025	1.05	1.099	1.027	1.054	1.108
4	1.032	1.059	1.117	1.035	1.065	1.128
5	1.037	1.066	1.129	1.04	1.072	1.142
6	1.041	1.071	1.138	1.044	1.077	1.152
7	1.044	1.075	1.145	1.047	1.081	1.159
8	1.046	1.078	1.151	1.05	1.084	1.165
9	1.048	1.081	1.155	1.052	1.087	1.17
10	1.05	1.083	1.159	1.054	1.089	1.175
11	1.051	1.085	1.163	1.055	1.091	1.178
12	1.053	1.086	1.166	1.056	1.093	1.181
15	1.055	1.09	1.172	1.059	1.097	1.189
20	1.058	1.094	1.18	1.062	1.101	1.197

2.7.4 Inner Wedge Test

The above three group sequential tests can stop the test when H_0 is rejected and H_1 is accepted, i.e., the test can be stopped if the test drug is effective. The Inner Wedge Test is a method that stops the test when H_0 is accepted, that is, the test can be stopped when the test drug is ineffective.

For k-stage Inner Wedge Test, the sample size is

$$n = R_w (z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)/(\mu_1 - \mu_2)^2$$

R_w is obtained from the Table 17.

Table 17 R_w table (α=0.05).

β	δ	k	C_{w1}	C_{w2}	R_w
0.2	-0.5	1	1.96	0.842	1
0.2	-0.5	2	1.949	0.867	1.01
0.2	-0.5	3	1.933	0.901	1.023
0.2	-0.5	4	1.929	0.919	1.033
0.2	-0.5	5	1.927	0.932	1.041
0.2	-0.5	10	1.928	0.964	1.066
0.2	-0.5	15	1.931	0.979	1.078
0.2	-0.5	20	1.932	0.988	1.087
0.2	-0.25	1	1.96	0.842	1
0.2	-0.25	2	1.936	0.902	1.026
0.2	-0.25	3	1.932	0.925	1.04
0.2	-0.25	4	1.93	0.953	1.059
0.2	-0.25	5	1.934	0.958	1.066
0.2	-0.25	10	1.942	0.999	1.102
0.2	-0.25	15	1.948	1.017	1.12
0.2	-0.25	20	1.952	1.027	1.131
0.2	0	1	1.96	0.842	1
0.2	0	2	1.935	0.948	1.058
0.2	0	3	1.95	0.955	1.075
0.2	0	4	1.953	0.995	1.107
0.2	0	5	1.958	1.017	1.128
0.2	0	10	1.98	1.057	1.175
0.2	0	15	1.991	1.075	1.198
0.2	0	20	1.998	1.087	1.212
0.2	0.25	1	1.96	0.842	1
0.2	0.25	2	1.982	1	1.133
0.2	0.25	3	2.009	1.059	1.199
0.2	0.25	4	2.034	1.059	1.219
0.2	0.25	5	2.048	1.088	1.252
0.2	0.25	10	2.088	1.156	1.341
0.2	0.25	15	2.109	1.18	1.379
0.2	0.25	20	2.122	1.195	1.4

0.1	-0.5	1	1.96	1.282	1
0.1	-0.5	2	1.96	1.282	1
0.1	-0.5	3	1.952	1.305	1.01
0.1	-0.5	4	1.952	1.316	1.016
0.1	-0.5	5	1.952	1.326	1.023
0.1	-0.5	10	1.958	1.351	1.042
0.1	-0.5	15	1.963	1.363	1.053
0.1	-0.5	20	1.967	1.37	1.06
0.1	-0.25	1	1.96	1.282	1
0.1	-0.25	2	1.957	1.294	1.006
0.1	-0.25	3	1.954	1.325	1.023
0.1	-0.25	4	1.958	1.337	1.033
0.1	-0.25	5	1.96	1.351	1.043
0.1	-0.25	10	1.975	1.379	1.071
0.1	-0.25	15	1.982	1.394	1.085
0.1	-0.25	20	1.988	1.403	1.094
0.1	0	1	1.96	1.282	1
0.1	0	2	1.958	1.336	1.032
0.1	0	3	1.971	1.353	1.051
0.1	0	4	1.979	1.381	1.075
0.1	0	5	1.99	1.385	1.084
0.1	0	10	2.013	1.428	1.127
0.1	0	15	2.026	1.447	1.148
0.1	0	20	2.034	1.458	1.16
0.1	0.25	1	1.96	1.282	1
0.1	0.25	2	2.003	1.398	1.1
0.1	0.25	3	2.037	1.422	1.139
0.1	0.25	4	2.058	1.443	1.167
0.1	0.25	5	2.073	1.477	1.199
0.1	0.25	10	2.119	1.521	1.261
0.1	0.25	15	2.14	1.551	1.297
0.1	0.25	20	2.154	1.565	1.316

2.7.5 Between-Proportion Comparison

For this design, the fixed sample size is

$$n_{fixed} = (z_{1-\alpha/2} + z_{1-\beta})^2 (p_1(1 - p_1) + p_2(1 - p_2))/ (p_1 - p_2)^2$$

A 5-phase group sequential trial comparing the efficacy of a drug and a control. According to the overall effective rate of the test drug in the pre-test and 30% in the control group, Pocock Test, O'Brien and Fleming Test and Wang and Tsitis Test (δ=0.1) were carried out respectively. Calculate the number of cases for each stage (n_{fixed}* R_p (or R_b, R_w, R_{wt}, etc.)/k).

2.7.6 Survival Analysis

For the group sequential design with time events as the experimental results, the Cox proportional hazards model is used as an example. The sample size is

$$n = I_{max}/I_k$$

where

$$I_{max} = R_b\,((z_{1-\alpha/2}+ z_{1-\beta})^2/\theta^2$$

The sample size for each stage is $n=I_{max}/0.25$. For example, a 5-stage group sequential trial comparing the efficacy of a drug and a control, with time events as the test results, according to the pilot study, θ=0.2. Calculate the sample size for each stage.

2.7.7 Re-Estimation of Sample Size

In the interim analysis of some group sequential trials, it is necessary to re-estimate the sample size based on the accumulated data, and it should be noted that blind re-estimation may cause bias. Shih et al. proposed a random double-blind sample size re-estimation method based on the observed results after 50% of the samples were completed. The estimation formula is as follows

$$n = (z_{1-\alpha/2}+ z_{1-\beta})^2(\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2))/(\hat{p}_1 - \hat{p}_2)^2$$

where

$$\hat{p}_1 = (hp_1 - (1-h)p_2)/(2h\text{-}1)$$
$$\hat{p}_2 = (hp_2 - (1-h)p_1)/(2h\text{-}1)$$

For example, in the two-center clinical trial, center A assigns patients to the trial group with a probability of 60%, center B assigns patients to the trial group with a probability of 40%, and the entire trial assigns patients to the trial group with a probability of 45% (h). In the interim analysis, 50% sample size was completed, the effective rate of center A is 70% (p_1), and the effective rate of center B is 60% (p_2). Let α=0.0001, β=0.10, re-estimate the sample size required for the next stage, n.

2.8 Bioequivalence

Bioequivalence (BE) means that different preparations of the same drug are given the same dose under the same experimental conditions, and there is no significant difference in the degree and speed of absorption. It is mainly used to evaluate whether generic drugs (generic drugs) and patented drugs (brand-name drug) is equivalent. Bioequivalence is compared with the test product and the reference product, and the two should have similar dosage forms, and there is no significant difference in their absorption rate and absorption amount in the organism. The current experimental design and analysis of bioequivalence is based on the following assumptions: the absorption rate and absorption amount of two drugs are the same, that is, bioequivalence is considered, and their therapeutic effects should also be the same. Bioequivalence is for the population distribution in which these observations lie, and when it is a normal distribution or a lognormal distribution, it is sufficient to compare the mean and variation. That is, to see whether the bioavailability is equivalent requires a statistical inference of these availability values as a sample of population parameters in the two formulations. Chinchilli (1996) gave three definitions of bioequivalence, that is, population bioequivalence

(PBE): bioequivalence for the probability distribution function related to two drugs; average bioequivalence (ABE): bioequivalence for the mean or median of the probability distribution functions related to two drugs; subject bioequivalence (IBE): bioequivalence for most subjects in the population. Two drugs with subject bioequivalence have drug switchability, that is, after a patient takes a certain drug for a period of time, if he switches to another drug with subject equivalence, he can get the same drug. Drugs with population bioequivalence have prescribability, that is, doctors can choose arbitrarily when prescribing drugs to patients for the first time, which has the same effect on this population of patients.

2.8.1 Average Bioequivalence

In a standard 2 (sequential) × 2 (period) crossover experiment, namely two treatments *T* and *R*, subjects were randomly divided into two groups, the first group received *T* treatment in the first period and *R* in the second period, and the experimental order was *TR*. The second group received *R* treatment in the first period and *T* treatment in the second period, and the experimental order was *RT*. The sample size based on a 2 × 2 crossover design for average bioequivalence is

$$n = (z_{1-\alpha} + z_{1-\beta/2})^2\sigma^2/(2*(\delta - |d|)^2)$$

where δ: the margin of average bioequivalence ($\log(0.8)\leq\delta\leq\log(1.25)$); σ^2: within-subject variance; d: the difference between *T* and *R*. For example, predesign an average bioequivalence study comparing inhalation and subcutaneous administration of a drug in a 2×2 crossover design. According to the pilot study, the within-subject standard deviation is 0.3, the margin of average equivalence is δ=log(1.25), d=0.08, σ=0.3. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.8.2 Population Bioequivalence

The sample size based on a 2×2 crossover design for population bioequivalence is

$$n = \zeta(z_{1-\alpha} + z_{\beta})^2\sigma^2/\lambda^2$$

where

$$\zeta = 2\delta^2\sigma^2 + \sigma_{TT}^4 + (1+a)^2\sigma_{TR}^4 - 2(1+a)\rho^2\sigma_{TT}^2\sigma_{TR}^2$$

where ρ: between-subject correlation coefficient; a=1.74; δ: the margin , i.e., the average difference of AUC. For example, predesign a population bioequivalence study comparing inhalation and subcutaneous administration of a drug in a 2×2 crossover design. According to the pilot study, σ=0.3, σ_{TT}=0.4, σ_{TR}=0.4, ρ=0.8, δ=0, λ=-0.3, a=1.74. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.8.3 Individual Bioequivalence

The sample size, n, based on a 2×4 crossover design (*TRTR*, *RTRT*) for individual bioequivalence is calculated from the following formula

$$\hat{\gamma} + \hat{U}^{0.5} + \hat{U}_{1-\beta}^{0.5} \leq 0$$

where

$$U = ((|\hat{\delta}| + t_{\alpha,2*n-2}\frac{\hat{\sigma}_{a,b}}{2}(2/n)^{0.5})^2 - \hat{\delta}^2)^2 + \hat{\sigma}_{a,b}^4(\frac{2*n-2}{\chi_{1-\alpha,2*n-2}^2} - 1)^2 +$$

$$+ab\hat{\sigma}_{WT}^4\left(\frac{2*n-2}{\chi^2_{1-\alpha,2*n-2}} - 1\right)^2 + (1.5+\theta_{IBE})^2\hat{\sigma}_{WR}^4\left(\frac{2*n-2}{\chi^2_{\alpha,2*n-2}} - 1\right)^2$$

$$\gamma = \delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 - \theta_{IBE} max\{\sigma_0, \sigma_{WR}^2\}$$

$$\sigma_{a,b}^2 = \sigma_D^2 + a\sigma_{WT}^2 + b\sigma_{WR}^2$$

whereσ_{WT}^2 and σ_{WR}^2: within-subject variances of group *T* and *R* respectively; σ_{BT}^2 and σ_{BR}^2: between-subject variances of group *T* and *R* respectively; δ: the difference between *T* and *R*; σ_D^2: the interaction between subject and drug. For example, predesign an individual bioequivalence study comparing inhalation and subcutaneous administration of a drug in a 2×4 crossover design. According to the pilot study, σ_{BT}=0.2, σ_{BR}=0.1, σ_{WT}=0.3, σ_{WR}=0.4, ρ=0.8, δ=0, λ=-0.3, a=b=0.5, θ_{IBE}=4. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.8.4 In-vitro Trial

The sample size based on a 2×4 crossover design (*TRTR*, *RTRT*) for in-vitro trial (1:1 design) is calculated from the following formula

$$\hat{\xi} + \hat{U}^{0.5} + \hat{U}_{1-\beta}^{0.5} \le 0$$

where

$$U = \left((|\hat{\delta}| + z_\alpha\left(\frac{s_{BT}^2}{n}+\frac{s_{BR}^2}{n}\right)^{0.5}\right)^2 - \hat{\delta}^2\right)^2 + s_{BT}^4\left(\frac{n-1}{\chi^2_{1-\alpha,n-1}} - 1\right)^2 + (1-n_T^{-1})^2 s_{WT}^4\left(\frac{n(n_T-1)}{\chi^2_{1-\alpha,n(n_T-1)}} - 1\right)^2 +$$

$$+ (1+\theta_{BE})^2 s_{BR}^4\left(\frac{n-1}{\chi^2_{\alpha,n-1}} - 1\right)^2 + (1+c\,\theta_{BE})^2(1-n_R^{-1})^2 s_{WR}^4\left(\frac{n(n_R-1)}{\chi^2_{\alpha,n(n_R-1)}} - 1\right)$$

$$\gamma = \delta^2 + \sigma_T^2 - \sigma_R^2 - \theta_{BE} max\{\sigma_0^2, \sigma_{WR}^2\}$$

where σ_{WT}^2 and σ_{WR}^2: within-subject variances of group *T* and *R* respectively; σ_{BT}^2 and σ_{BR}^2: between-subject variances of group *T* and *R* respectively; δ: the difference between *T* and *R*; σ_D^2: the interaction between subject and drug. For example, predesign an in-vitro experiment without repeated parallel controls. According to the pilot study, σ_{BT}=0.4, σ_{BR}=0.5, σ_{WT}=0.5, σ_{WR}=0.5, δ=0, θ_{BE}=1.5. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.9 Dose Response Studies

The research on dose-response relationship mainly includes: the dose-response relationship between different dose groups, the shape of the dose-response relationship curve, and the optimal dose. Usually, a randomized parallel control design is used to study the dose-response relationship, and the effectiveness of the drug is proved by measuring the variance. The Williams method compares the minimum effective dose of the experimental group and the control group, demonstrates the dose-response relationship through a model, and illustrates the optimal dose through the maximum tolerable dose (MTD).

2.9.1 Continuous Response

The sample size is (1:1 design)

$$n = ((z_{1-\alpha} + z_{1-\beta})\sigma/d)^2 \sum_{i=0}^{k} c_i^2$$

where σ: standard deviation; d=$\sum_{i=0}^{k} c_i u_i$; $\sum_{i=0}^{k} c_i$=0; c_i: grouping, and c_0: control group; u_i: percent improvement from baseline for each group, and u_0: control group. Each group shares the same sample size;

For example, a four-group parallel controlled dose-response trial, including 1 control group and 3 test groups (k=3; doses are 10 mg, 20 mg, 30 mg respectively). According to the pilot study, σ=0.2, c_0=−6, c_1=1, c_2=2, c_3=3, u_0=0.05, u_1=0.1, u_2=0.2, u_3=0.25. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.9.2 Binary Response

The sample size is estimated from the following formula (1:1 design)

$$n = ((z_{1-\alpha}(\sum_{i=0}^{k} c_i^2\, \bar{p}(1-\bar{p}))^{0.5} + z_{1-\beta}(\sum_{i=0}^{k} c_i^2\, p_i(1-p_i))^{0.5})/d)^2$$

where d=$\sum_{i=0}^{k} c_i\, p_i$; $\sum_{i=0}^{k} c_i$=0; c_i: grouping, and c_0: control group; p_i: response rate of each group, and p_0: control group. For example, a four-group (k=3) parallel controlled dose-response trial, including 1 control group and 3 test groups (k=3; doses are 10 mg, 20 mg, 30 mg respectively). According to the pilot study, c_0=−6, c_1=1, c_2=2, c_3=3, p_0=0.05, p_1=0.1, p_2=0.2, p_3=0.25. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.9.3 Time-to-Event Endpoint

The sample size is estimated from the following formula (1:1 design)

$$n = ((z_{1-\alpha}\sigma_0(\sum_{i=0}^{k} c_i^2\,)^{0.5} + z_{1-\beta}(\sum_{i=0}^{k} \sigma_i c_i^2\,)^{0.5})/d)^2$$

where

$$\sigma^2(\lambda_i) = \lambda_i^2\,(1 + \mathrm{e}^{-\lambda_i T}(1 - \mathrm{e}^{\lambda_i T_0})/(T_0\lambda_i))^{-1}$$

where T_0: the inclusion time of the trial; T: total trial time; $\sum_{i=0}^{k} c_i$=0; c_i: grouping. For example, a phase II clinical trial of a drug is conducted. A control group, a low-dose group, a high-dose group and a combined treatment group are designed. Observe the patient's survival time. Assuming that the inclusion time of the trial is 9 months and the total trial time is 18 months. The median survival time of the four groups is estimated to be 12, 18, 20 and 22 months, and the corresponding risk ratios are 0.04/month, 0.03/month, 0.02/month and 0.03/month. α=0.001, β=0.10. Calculate the sample size for each group, n.

2.9.4 Minimum Effective Dose (MED)

The sample size for minimum effective dose based on the Williams test is

$$n = 2\sigma^2(t_\alpha(k)+ z_\beta)^2/\delta^2$$

For example, design a dose-response trial using the Williams test to detect the minimum effective dose. According to the pilot study, σ=0.4, k=3, δ=0.15. α=0.0001, β=0.10. Calculate the sample size for each group, n.

2.9.5 Cochran-Armitage Test for Trend

The sample size based on the Cochran-Armitage trend test (1:1 design) is

$$n = (n^*/4)\,(1 + (1+2\delta/A)^{0.5})^2$$

where

$$n^* = (z_{1-\alpha}((k+1)((k+1)^2\text{-}1)pd\,)^{0.5} + z_{1-\beta}(\sum_{i=0}^{k} p_i q_i b_i^2\,)^{0.5})^2$$

$b_i = i - (k+1)/2$

$d = \sum_{i=0}^{k} b_i p_i$

$A = \sum_{i=0}^{k} p_i (d_i - \hat{d})$

$p = \sum_{i=0}^{k} p_i/(k+1)$

$q = 1-p$

For example, design a four-group (1 control group; k=3 test groups) Cochran-Armitage trend detection dose-response trial. According to the pilot study, d_0=1, d_1=2, d_2=3, d_3=4, p_0=0.2, p_1=0.4, p_2=0.6, p_3=0.8, δ=1. α=0.0001, β=0.10. Calculate the sample size, n.

2.9.6 Dose Escalation Trials

The non-parametric escalation trial design is also called "M+N" design or "A+B" design. All subjects were randomly assigned to several groups, and each group contained several subjects. Subjects in the same group received the same dose level of the test drug. According to the random distribution scheme, the subjects entered the test process group by group, and the subjects in each group took a certain dose level of the test drug only once. Drug response results were recorded regardless of whether the expected drug response was detected. When the number of drug responders at a certain dose level satisfies the conditions for stopping the overall trial, the overall trial process ends and the explored target dose is obtained. When exploring dose limiting toxicity (DLT) and maximum tolerable dose (MTD), the "3+3" design of the climbing test is widely used, that is, each time there are 3 subjects who enter the test process, and a maximum of 6 subjects at a dose level will take the drug. The overall dose escalation rules for escalation trial design can be further divided into two different dose escalation strategies: TER strategy (traditional escalation rules) and STER strategy (strict traditional escalation rules). The biggest difference between the TER strategy and the STER strategy is that when a toxic response to the test drug is detected at the x_j dose level, phase I clinical trials that follow the TER dose escalation strategy do not allow subjects to continue to be enrolled at the x_{j-1} dose level, and it is required to stop the trial directly. At this time, the dose level x_j is considered to be the expected target dose level. However, clinical trials that follow the STER dose escalation rule require that subjects be continued to be included in the trial at the x_{j-1} dose level to observe the overall toxic response of the subjects at the x_{j-1} dose level, so that the target dose can be inferred.

The sample sizes of A+B TER strategy design are

$$n_j = \sum_{i=0}^{n-1} n_{ji} p_i^*$$

where

$$n_{ji} = \frac{Ap_0^j + (A+B)q_0^j}{p_0^j + q_0^j}, \quad j<i+1$$

$$n_{ji} = \frac{A(1-p_0^j - p_1^j) + (A+B)(p_1^j - q_0^j)}{1-p_0^j - q_0^j}, \quad j=i+1$$

$$n_{ji} = 0, \quad j>i+1$$

$$p_0^j = \sum_{k=0}^{C-1} \binom{A}{k} p_j^k (1-p_j)^{A-k}$$

$$q_0^j = \sum_{k=C}^{D} \sum_{m=0}^{E-k} \binom{A}{k} p_j^k (1-p_j)^{A-k} \binom{B}{m} p_j^m (1-p_j)^{B-m}$$

$$p_n^* = \prod_{j=1}^{n} (p_0^j + q_0^j)$$

The sample sizes of A+B STER strategy design are

$$n_j = n_{jn} p_n^* + \sum_{i=0}^{n-1} \sum_{k=i+1}^{n} n_{jik} p_{ik}$$

where

$$n_{jki} = \frac{A p_0^j + (A+B) q_0^j}{p_0^j + q_0^j}, \quad j<i$$

$$n_{jik} = A + B, \quad i \leq j < k$$

$$n_{jik} = \frac{A(1-p_0^j-p_1^j) + (A+B)(p_1^j - q_0^j)}{1-p_0^j - q_0^j}, \quad j=k$$

$$n_{jik} = 0, \quad j>i+1$$

$$p_i^* = \sum_{k=i+1}^{n} p_{ik}$$

$$p_i^k = (q_0^i + q_0^i)(1 - p_0^k - q_0^k) \prod_{j=1}^{i-1} (p_0^j + q_0^j) \prod_{j=i+1}^{k-1} q_2^j$$

$$p_n^* = \prod_{j=1}^{n} (p_0^j + q_0^j)$$

For example, the "3+3" design of the escalation trial. According to the pilot study, the dose-limiting toxicity of 6 doses (10, 13, 25, 38, 59, 68) of a certain drug were 0.02, 0.025, 0.035, 0.06, 0.2, 0.7. Calculate the sample size for each dose.

2.10 Microarray Studies

The sample size of microarray data is small and the number of variables is large. The traditional *t*-test and Wilcoxon test need to be adjusted when they are applied. There are FDR (fasle discover rate) control, FWER (family-wise error rate) control, etc., based on control indices; single-step method, step-wise method, resampling-based method, based on the control operation procedures, and frequency school method and Bayes school method, based on different schools. Multiple testing is an extension of the traditional concept of multiple comparisons. The null hypothesis H_0 is verified by repeated testing of multiple variables on the same question. This hypothesis is a series of hypotheses (a family of hypotheses), rather than a single hypothesis.

Assuming that m hypotheses are tested at the same time, among which m_0 are correct, and R represents the number of hypotheses with positive results, as indicated in the following table:

	Not reject H_0	Reject H_0	Total
H_0 is true	U	V	m_0
H_1 is true	T	S	$m-m_0$
Total	m-R	R	m

Among them, m is known before the hypothesis test, R is an observable random variable, and U, V, S, and T are unobservable random variables. Fasle Discover Rate (The proportion of occurring errors in the results of rejecting H_0) is defined as

$FDR = E(V/R)$, $R \neq 0$
$FDR = 0$, $R = 0$

Based on FDR design, the sample size of each group for one-sided fixed effect test is

$$n = ((z_{\alpha^*} + z_{\beta^*})^2/(a_1 a_2 \delta^2) + 1)/2$$

where

$$\alpha^* = r_1 f/(m_0(1-f))$$
$$\beta^* = 1 - r_1/m_1$$

The sample size for one-sided variable effect test can be calculated from the following formula

$$\sum_{j \in M_1} \phi\left(z_{\alpha^*} - \delta_j (n a_1 a_2)^{0.5}\right) - r_1 = 0$$
$$\alpha^* = r_1 f/(m_0(1-f))$$

where f: the false discovery rate; r_1: the number of actual rejections; a_k: the distribution ratio of two groups; m: the total number of tested genes; m_1: the number of prognostic genes, and δ: the size of the effect of prognostic genes.

Based on FDR design, the total sample size of each group for two-sided fixed effect test is

$$n = ((z_{\alpha^*/3} + z_{\beta^*})^2/(a_1 a_2 \delta^2) + 1)/2$$

≤

where

$$\alpha^* = r_1 f/(m_0(1-f))$$
$$\beta^* = 1 - r_1/m_1$$

The sample size for two-sided variable effect test can be calculated from the following formula

$$\sum_{j \in M_1} \phi\left(z_{\alpha^*/2} - |\delta_j| (n a_1 a_2)^{0.5}\right) - r_1 = 0$$
$$\alpha^* = r_1 f/(m_0(1-f))$$

where f: the false discovery rate; r_1: the number of actual rejections; a_k: the distribution ratio of two groups; m: the total number of tested genes; m_1: the number of prognostic genes, and δ_j: the size of the effect of prognostic genes.

One-sided fixed effect design. For example, design a microarray study of 2000 candidate genes (m=2000). It is estimated that there are 30 (m_1=30) genes that are differently expressed between the two groups, and the actual number of rejected genes is about 18 (r_1=18). FDR=0.01. δ=1, a_1=a_2=0.5. Calculate the sample size for each group, n.

One-sided variable effect design. For example, design a microarray study of 2000 candidate genes (m=2000). It is estimated that there are 30 (m_1=30) genes with different expression between the two groups, and the actual number of rejected genes is about 18 (r_1=18). FDR=0.01. δ_j=1, if $1 \leq i \leq 10$, and δ_j=0.5, if $11 \leq i \leq 30$.

a_1=a_2=0.5. Calculate the sample size for each group, n/2.

Two-sided fixed effect design. For example, design a microarray study of 2000 candidate genes (m=2000). It is estimated that there are 30 (m_1=30) genes with different expression between the two groups, and the actual number of rejected genes is about 18 (r_1=18). FDR=0.01. δ=1, a_1=a_2=0.5. Calculate the sample size for each group, n.

Two-sided variable effect design. For example, design a microarray study of 2000 candidate genes (m=2000). It is estimated that there are 30 (m_1=30) genes with different expression between the two groups, and the actual number of rejected genes is about 18 (r_1=18). FDR=0.01. δ_j=1, if 1≤i≤10, and δ_j=0.5, if 11≤i≤30. a_1=a_2=0.5. Calculate the sample size for each group, n/2.

2.11 Nonparametrics

Nonparametric tests are hypothesis tests that do not rely on statistical parameters. They are suitable for hypothesis testing of unknown distribution types, skewed data, hierarchical data, etc.

2.11.1 Test for Independence

The sample size for test of independence is

$$n = 4\ (z_{1-\alpha}/3 + z_{1-\beta}(2p_2 - 1 - (2p_1 - 1)^2)^{0.5})^2/(2p_1 - 1)^2$$

where p_1=$P((x_1-x_2)(y_1-y_2)>0)$, p_2=$P((x_1-x_2)(y_1-y_2)(x_1-x_3)(y_1-y_3)>0>0)$. For example, it has been observed that in a pilot study that as the x increases, y also tends to increase. A clinical trial is designed to verify the above conjecture. According to the pilot study, p_1=0.4, p_2=0.6. α=0.0001, β=0.10. Calculate the sample size, n.

2.12 Sample Sizes Calculation in Other Areas

2.12.1 ANOVA with Repeated Measures

The ANOVA with repeated measures can be repeated measures under the same condition, or repeated measures under different conditions. The ANOVA with repeated measures can be used to examine whether there are significant differences between various treatments, to find differences among subjects, or to find interaction between various treatments and groups of subjects. In parallel controlled clinical trials, it is mainly used to evaluate effectiveness and safety. The sample size for ANOVA with repeated measures is

$$n = 2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2/\delta^2$$

where σ^2: the sum of variances of all groups. For example, a test drug and a traditional drug are tested in parallel on experimental animals, and each experimental animal records disease scores repeated three times. According to the pilot study, σ^2=1.5, δ=1.2. α=0.001, β=0.10. Calculate the total sample size, n.

2.12.2 QT/QTc

The QT interval refers to the time course of ventricular depolarization and repolarization, that is, the time course from the starting point of the QRS complex to the end point when the T wave returns to baseline. Delayed cardiac repolarization will create a special cardiac electrophysiological environment in which arrhythmias are prone to occur, the most common of which is torsade de pointes (TdP), but other types of ventricular tachyarrhythmias can also occur. Since the degree of QT prolongation can be regarded as a relative biomarker of arrhythmogenic risk, there is usually a qualitative relationship between QT prolongation and TdP, and it is more important for those drugs that may cause QT prolongation. Since the QT interval is inversely related to heart rate, it is routine to correct the measured QT interval to a less heart rate dependent QTc interval through various formulas. However, it is unclear whether there is a necessary link between the occurrence of arrhythmia and an increase in the QT interval or the absolute value of QTc. Most drugs that cause TdP can

significantly prolong the QT/QTc interval (i.e., QT/QTc). Because QT/QTc interval prolongation is an electrocardiographic finding associated with increased sensitivity for detecting arrhythmias, adequate safety evaluation of new drugs before marketing should include a detailed characterization of their effects on the QT/QTc interval.

2.12.2.1 Parallel Control Design

The sample size for parallel control design is

$$n = 2\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 (\rho + (1-\rho)/K)/\delta^2$$

where $\rho=\sigma_b^2/(\sigma_b^2+\sigma_w^2)$; σ_b^2: between-subject variance; σ_w^2: within-subject variance; K: number of replications per subject; $\delta=d/(\sigma_b^2+\sigma_w^2)$; d: clinical difference; For example, a non-antiarrhythmic drug conducts a comprehensive ECG parallel control study to determine its effect on the QT/QTc interval. According to the prel-trial, σ_b=2.8, σ_w=0.5, d=2, K=5. α=0.0001, β=0.1. Calculate the sample size for each group, n.

2.12.2.2 Parallel Control Design with Covariates

The sample size for parallel control design with covariates is

$$n = (2+(v_1-v_2)^2/(\tau_1^2+\tau_2^2))\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \left(\rho + \frac{1-\rho}{K}\right)/\delta^2$$

where $\rho=\sigma_b^2/(\sigma_b^2+\sigma_w^2)$; σ_b^2: between-subject variance; σ_w^2: within-subject variance; K: number of replications per subject; $\delta=d/(\sigma_b^2+\sigma_w^2)$; d: clinical difference; v_1, v_2: means of two groups; τ_1, τ_2: the standard deviations of two groups. For example, a comprehensive electrocardiogram parallel control study was conducted on a non-antiarrhythmic drug. The C_{max} of the drug is known to have an impact on the QT/QTc interval to clarify its impact on the QT/QTc interval. According to the prel-trial, σ_b=3.2, σ_w=0.5, v_1=1.1, v_2=1.0, τ_1=1.8, τ_2=0.9, d=2, K=5. α=0.0001, β=0.1. Calculate the sample size for each group, n.

2.12.2.3 Crossover Control Design

The sample size for crossover control design is

$$n = \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 (\rho + (1-\rho)/K)/(\delta^2 - \gamma\left(z_{1-\alpha} + z_{1-\beta}\right)^2)$$

where $\rho=\sigma_b^2/(\sigma_b^2+\sigma_w^2)$; σ_b^2: between-subject variance; σ_w^2: within-subject variance; K: number of replications per subject; $\delta=d/(\sigma_b^2+\sigma_w^2)$; d: clinical difference; $\gamma=\sigma_p^2/(\sigma_b^2+\sigma_w^2)$; σ_p^2: Additional variation in crossover design. For example, a comprehensive ECG cross-control study was conducted on a non-antiarrhythmic drug to determine its effect on the QT/QTc interval. According to the prel-trial, σ_b=2.8, σ_w=0.5, σ_p=0.1; d=2, K=5. α=0.0001, β=0.1. Calculate the sample size for each group, n.

2.12.2.4 Crossover Control Design with Covariates

The sample size for crossover control design with covariates is

$$n = (1+(v_1-v_2)^2/(\tau_1^2+\tau_2^2))\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 (\rho + (1-\rho)/K)/(\delta^2 - \gamma\left(z_{1-\alpha} + z_{1-\beta}\right)^2)$$

where $\rho=\sigma_b^2/(\sigma_b^2+\sigma_w^2)$; σ_b^2: between-subject variance; σ_w^2: within-subject variance; K: number of replications

per subject; $\delta=d/(\sigma_b^2+\sigma_w^2)$; d: clinical difference; $\gamma=\sigma_p^2/(\sigma_b^2+\sigma_w^2)$; σ_p^2: Additional variation in crossover design; v_1, v_2: means of two groups; τ_1, τ_2: the standard deviations of two groups. For example, a comprehensive ECG cross-control study was conducted on a non-antiarrhythmic drug to determine its effect on the QT/QTc interval. According to the prel-trial, σ_b=2.8, σ_w=0.5, v_1=1.8, v_2=1.1, τ_1=1.8, τ_2=0.9, σ_p=0.1; d=2, K=8. α=0.0001, β=0.1. Calculate the sample size for each group, n.

2.12.3 Quality of Life (QOL)

Since chronic non-infectious diseases are difficult to cure, it is difficult to use cure rate to evaluate treatment effects, and the role of survival rate is also limited. Therefore, quality of life is used as an evaluation item for new drugs. The sample size for QOL analysis is

$$n = max\ \{c(z_{1-\alpha/2}\ +\ z_{1-\beta})^2/d^2,\ c(z_{1/2+\eta/2}\ +\ z_{1-\alpha/2})^2/(d-\phi)^2\}$$

where d: the difference; c: the constant. For example, a drug is undergoing a clinical trial based on the QOL. According to the pilot study, c=0.3, d=0.2, ϕ=0.1. η=0.1, α=0.0001, β=0.10. Calculate the sample size, n.

2.12.4 Bridging Studies

This design mainly evaluates the impact of "ethnic factors" on drugs, and provides relevant pharmacokinetics/pharmacodynamics or clinical trial data such as efficacy, safety, usage and dosage, so that clinical trial data can be extrapolated to reduce repeated clinical trials , quickly provide patients with medicines. Ethnic factors are generally defined as factors related to race or to a group of people with common traits and habits, and are usually divided into intrinsic and extrinsic factors. Chow et al proposed to use the Sensitivity Index as an indicator to extrapolate the experimental results of the placebo parallel controlled trial design. The sample size designed with sensitivity index is calculated from the following

$$\hat{P}_\Delta\ =\ E_{\delta,u}(1\ -\ \tau_{n-2}(t_{n-2}|\frac{\Delta\delta}{u})\ -\ \tau_{n-2}(-t_{n-2}|\frac{\Delta\delta}{u}))$$

where Δ: sensitivity index between ethnic groups. For example, a pharmaceutical company intends to promote a certain drug to another country, and conducts a bridging study with the parallel control design, using the sensitivity index as an assessment index for ethnic factors. According to the pilot test, Δ=2.1, $\hat{p}_\Delta$=0.60. α=0.0001. Calculate the sample size, n.

2.12.5 Vaccine Clinical Trials

The most important goal of evaluating a vaccine is its ability to prevent disease, which usually requires a large-sample placebo-controlled design. Relative reduction in disease incidence, $(p_C-p_T)/p_C$, where p_T is

2.12.5.1 Reduction in Disease Incidence

The sample size for the test of reduction in disease incidence is

$$n\ =\ z_{1-\alpha/2}^2((1-p_T)/p_T+(1-p_C)/p_C)/d^2$$

where $d=z_{1-\alpha/2}((1-p_T)/(np_T)+(1-p_C)/(np_C))^{0.5}$. For example, it is planned to implement a vaccine clinical trial, compared with placebo, and the index uses the reduction in disease incidence. According to the pilot study, the incidence rate of the vaccine group is 2% (p_T), and the incidence rate of the control group is 5% (p_C), d=0.1. α=0.0001. Two-group 1:1 parallel control. Two-sided test. Calculate the sample size for each group, n.

2.12.5.2 Evaluation of Vaccine Efficacy with Extremely Low Disease Incidence

The sample size for the test of the evaluation of vaccine efficacy with extremely low disease incidence is

$$n = (z_{1-\alpha}(\theta_0(1-\theta_0))^{0.5} + z_{1-\beta}(\theta(1-\theta))^{0.5})^2/((p_T + p_C)(\theta - \theta_0)^2)$$

where θ=(1-π)/(1-π+n_C/n_T), θ_0=(1-π_0)/(1-π_0+n_C/n_T), and π=(p_C-p_T)/p_C. p_T=0.01, p_C=0.02, θ=0.3, θ_0=0.5, follow the example above.

2.12.5.3 Composite Efficacy Measure

Composite Efficacy Measure index includes not only the evaluation of the occurrence of the disease, but also the evaluation of the infection of the disease. The sample size is

$$n = (z_{1-\frac{\alpha}{2}}\left(2\bar{\mu}^2\bar{p}(1-\bar{p}) + 2\bar{p}(\sigma_T^2 + \sigma_C^2)\right)^{0.5} + \\ + z_{1-\beta}(p_T(\sigma_T^2 + \mu_T^2(1-p_T)) + p_C(\sigma_C^2 + \mu_C^2(1-p_C)))^{0.5})^2/(\mu_T p_T - \mu_C p_C)^2$$

where μ_T, μ_C : the mean of test and control groups respectively; σ_T, σ_C: the standard deviation of test and control groups respectively. According to the pilot study, μ_T =0.2, μ_C=0.4, p_T =0.1, p_C=0.2, σ_T^2=σ_C^2=0.1. α=0.0001, β=0.1. Two-group 1:1 parallel control. Two-sided test. Calculate the sample size for each group, n.

2.12.6 Propensity Scores in Nonrandomized Clinical Trials

In nonrandomized trials, assignment of subjects is dependent on subject baseline covariates. For example, whether a patient will receive a drug may be affected by many factors. When these factors also affect the prognosis at the same time, they are potential interference factors. If the basic characteristics of treated and untreated patients are different, outcomes cannot be directly compared between the two groups. In the propensity score analysis, the propensity score is a probability (0~1), representing a patient's chance of receiving drug treatment under its existing basic characteristics (or interference factors). The propensity score focuses on the relationship between the basic characteristics of the object and the presence or absence of drug treatment, in an attempt to recreate a situation similar to random allocation. In a randomized trial, each subject should have a propensity score of 0.5 for treatment. In nonrandomized observational studies, propensity scores will vary according to the underlying characteristics of the patients. The most common propensity score comes from a logistic regression model: treatment or not is regarded as the dependent variable, and each factor of the basic characteristics is regarded as the independent variable. The propensity score method has been widely used in these non-randomized controlled trials to reduce the selection bias caused by confounding factors, so as to ensure that the baseline data between groups are balanced and comparable.

2.12.6.1 Weighted Mantel-Haenszel (WMH)

The sample size for weighted Mantel-Haenszel is

$$n = (\sigma_0 z_{1-\alpha/2} + \sigma_1 z_{1-\beta})^2/\delta^2$$

where

$$\delta = (1 - \phi)\ \sum_{j=1}^{m} w_j a_j\ b_{j1} b_{j2} \frac{p_{j1} - q_{j1}}{q_{j1} + \emptyset p_{j1}}$$

$$\sigma_1^2 = \sum_{j=1}^{m} w_j^2 a_j\ b_{j1} b_{j2} (b_{j2} p_{j1} q_{j1} + b_{j1} p_{j2} q_{j2})$$

$$\sigma_0^2 = \sum_{j=1}^{m} w_j^2 a_j\ b_{j1} b_{j2} (b_{j1} p_{j1} + b_{j2} p_{j2})(b_{j1} q_{j1} + b_{j2} q_{j2})$$

$$\phi = p_{j2} q_{j1} / p_{j1} q_{j2}$$

and m: the number of layers; k: two groups and one is the control, $k = 1,2$; b_{jk}=n_{jk}/n_j; $b_{j1}+b_{j2}$=2; a_j: the proportion of samples in ith layer *vs*. total samples, a_j=n_j/n; p_{jk}: the probability for kth group at jth layer. For example, a clinical trial comparing a drug with a traditional drug was designed with a weighted Mantel-Haenszel analysis, and the main evaluation index was cardiovascular events. The baseline variable was tested for between-group balance, and it was found that the baseline variable was unbalanced between the test drug and the control drug. The designed number of layers in clinical trial is 5. According to the pilot study, the proportions of samples in each layer to the total number of samples is 0.2, 0.2, 0.2, 0.2, and 0.2, the assigned proportions of each layer for test group is 0.5, 0.5, 0.5, 0.4, and 0.4, and in each layer the probabilities of occurring response are 0.6, 0.4, 0.8, 0.9, and 0.7. α=0.0001, β=0.10, ϕ=2. Calculate the sample size, n.

2.12.6.2 Unstratified Analysis

The sample size for unstratified analysis is

$$n = (\hat{\sigma}_0 z_{1-\alpha/2} + \hat{\sigma}_1 z_{1-\beta})^2 / (p_1 - p_2)^2$$

where

$$\sigma_1^2 = \sum_{j=1}^{J} a_j b_{j1} b_{j2} (b_{j2} p_{j1} q_{j1} + b_{j1} p_{j2} q_{j2})$$

$$\sigma_0^2 = \sum_{j=1}^{J} a_j b_{j1} b_{j2} (b_{j1} p_{j1} + b_{j2} p_{j2})(b_{j1} q_{j1} + b_{j2} q_{j2})$$

and J: the number of layers; k: two groups and one is the control, $k = 1,2$; b_{jk}=n_{jk}/n_j; $b_{j1}+b_{j2}$=2; a_j: the ratio of samples in ith layer *vs*. total samples, a_j=n_j/n; p_{jk}: the probability for kth group at jth layer. Follow the example above.

2.12.7 Sensitivity and Specificity Estimation

This is a single sample trial aimed to assess the value of a technique in finding a phenomenon (e.g., a disease). First, obtain the sensitivity and specificity (p) from previous studies or experiments, then the sample size is (Fleiss et al., 2003):

$$n = (z_{1-\alpha/2} (p(1 - p))^{0.5}/d)^2$$

where d: permissible error (e.g., 0.1), p: sensitivity (p_{se}, e.g., 0.8) or specificity (p_{sp}, e.g., 0.9). Use the maximum of n from p_{se} and p_{sp}.

2.12.8 Distance Based Sampling

In the distance based sampling, the CV for density estimation is (Seber, 1982)

$$CV = 1/(n*r - 2)^{1/2}$$

thus

$$n = (1/CV^2+2)/r$$

where n: the number of random points for measuring distance; r: the number of distance measuring for each

point. r=1, if only the distance from the nearst points is measured.

2.12.9 Linear Transect Sampling

The coefficient of variation for density estimation along a linear transect is (Eberhardt, 1978)

$$CV(D) = ((1 + CV^2(1/r_i))/n)^{1/2}$$

where n: the sample size, r_i: radial distance to each visible animal, $CV(1/r_i)$: the coefficient of variation of reciprocal of the radial distance. Usually $CV(1/r_i)$=1, and $CV(D)=(2/n)^{1/2}$. The sample size is thus

$$n = 2/CV(D)^2$$

In two-stage sampling, the length of a linear transect is (Burnham et al., 1980)

$$L = (b/CV^2(D))\ (L_1/n_1)$$

where b: a value in [1.5, 4] (usually b=3), L_1: the length of linear transect in the first sampling, and n_1: the number of subjects (e.g., animals) found in the first sampling. L_1/n_1 can be estinated in advance according to previous studies. For example, there are 20 objects/km, then L_1/n_1=0.05.

Assume that the CV for density estimation is 0.1, i.e., the half-width of confidence interval at α=0.05 is ±20% of the true density. 10 objects are found along a linear transect of 30 km. The length of linear transect to be investigated should be: $L = 3/0.1^2*(30/10) = 900$ km. If only 300 km is investigated, $CV(D)$=0.173 (i.e., $300 = 3/CV^2(D)*(30/10)$), i.e., ±17% of error, and the width of confidence interval at α=0.05 is ±34% of the true density (Krebs, 1989; Zhang, 2007).

2.12.10 Mark-Recapture Sampling

According to Seber (1982), for Petersen mark-recapture sampling, the number of marked animals recaptured in the second sampling (R) is

$$R = 1/CV^2$$

where CV is coefficient of variation for the estimation of population size (e.g., CV=0.03). Therefore the total number of marked animals in the first sampling should be larger than R.

For Schnabel mark-recapture sampling, the number of marked animals recaptured in the sequential samplings is

$$\sum R_i = 1/CV^2$$

where R_i: the number of marked animals recaptured in the i-th sampling. Given CV (e.g., 0.03, etc.), the time for stoping recapture can be calculated.

2.12.11 Stratified Random Sampling

Assume that the total samples will be proportionally assigned to sub-populations according to their sizes:

$$w_i = n_i/n = N_i/N$$

where n_i: the number of samples assigned to the i-th sub-population, n: the sample size of total population, N_i:

the size of the i-th sub-population, N:.the size of total population.

The sample size for total population is calculated by

$$n = m/(1 + m/N)$$

where

$$m = A*\sum w_i s_i^2/d^2$$

A=4 for α=0.05, A=7.08 for α=0.01, d: the permitted error, i.e., the half-width of of confidence interval for mean estimation at confidence level α, and s_i: the standard deviation for the i-th sub-population, which has been determined in the first sampling.

3 Simple Random Sampling

Once the same size is determined, one can implement the simple random sampling (or other advanced sampling techniques) to take the sample needed.

Probability Sampling Principle is a fundamental principle in sampling theory. It can be outlined as follows:

(1) Define a set of candidate samples S_i, i=1,2,…, each candidate sample contains some sampling units (subjects);

(2) Assign a selection probability to each candidate sample;

(3) With the help of the random number table, select the available sample from the candidate sample set S_i, i=1,2,…, through selection probability. By selecting a sample according to the above probability sampling principles, a suitable sampling theory can always be found to explain and analyze the data collected.

The Simple Random Sampling is a type of probability sampling and is defined as follows:

(1) Suppose the statistical population contains N sampling units;

(2) Select n sampling units (n is the sample size) from the statistical population (total population), and each sampling unit has an equal chance of being selected. The Simple Random Sampling is the basis of all random sampling techniques.

In sampling study, we assume that the sample size and the selection probability of all candidate samples are the same respectively. Randomly select one candidate sample from candidate samples and use it as the final sample to be taken.

4 Computational Tool: SampSizeCal

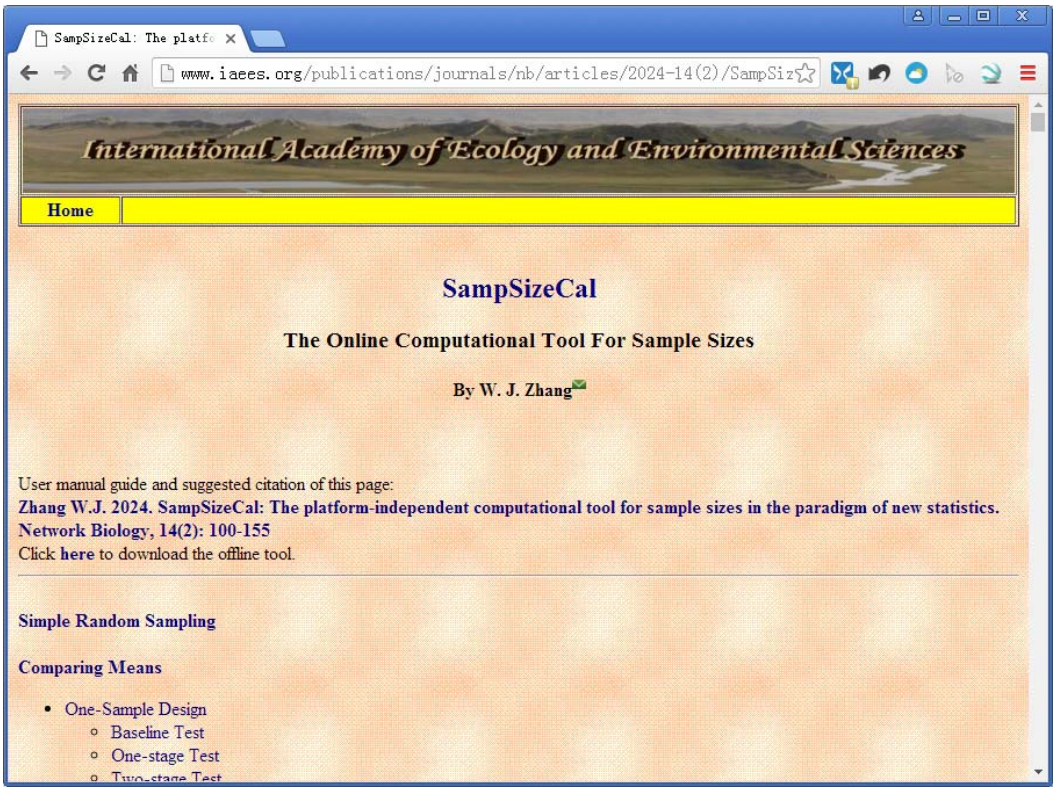
I delveloped a computational tool, SampSizeCal, to harbor more than 120 methods of sample size estimation described above. The SampSizeCal includes both online (http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/SampSizeCal.htm) and offline versions, and can be used for various computing devices (PCs, iPads, smartphones, etc.), operating systems (Windows, Mac, Android, Harmony, etc.) and web browsers (Chrome, Firefox, Sougo, 360, etc.). In this tool, both default p-value (in most cases, 0.0001, which is 200 times of the commonly used p-value, 0.05) and the maximum p-value (in most cases, 0.005, which is 10 times of the commonly used p-value, 0.05) were greatly enhanced Meanwhile, the default statistical power, 80%, was enhanced to 90% in SampSizeCal. These settings will lead to the reasonable increase of sample sizes. It is currently the most comprehensive platform-independent computational tools for sample sizes, and can be used in experimental sciences such as medicine (clinical medicine, experimental zoology, public health, pharmacy, etc.), biology, ecology, agronomy, psychology and

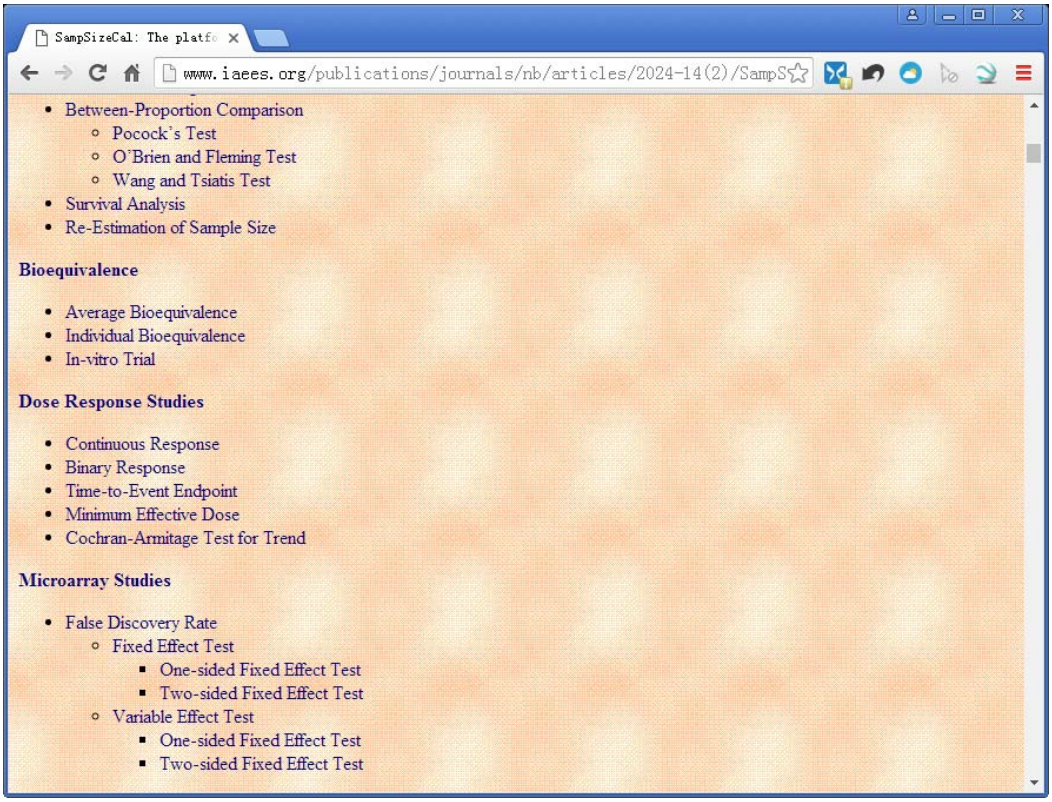
engineering technology.

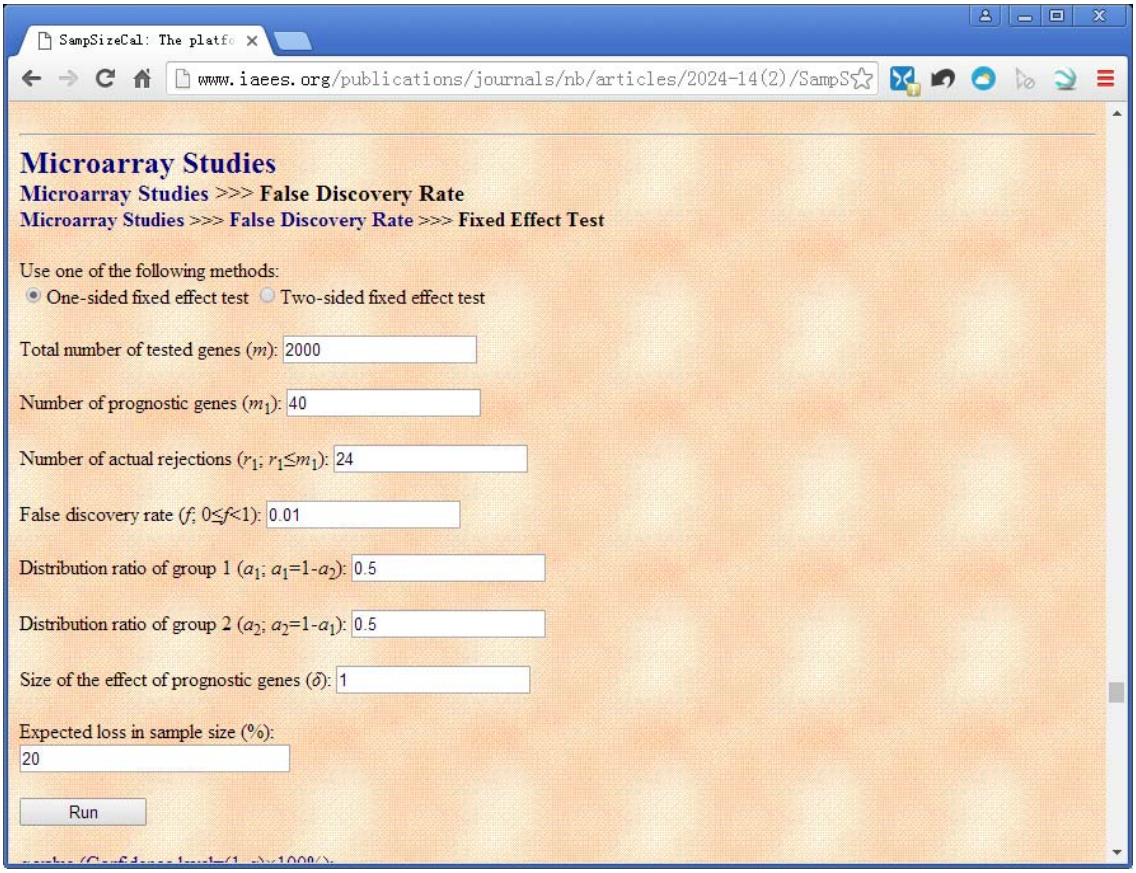
Both user manual guide and offline tool can be found at:

http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/e-suppl/SampSizeCal.rar

Double-click the offline tool, it will be opened in the default web browser.







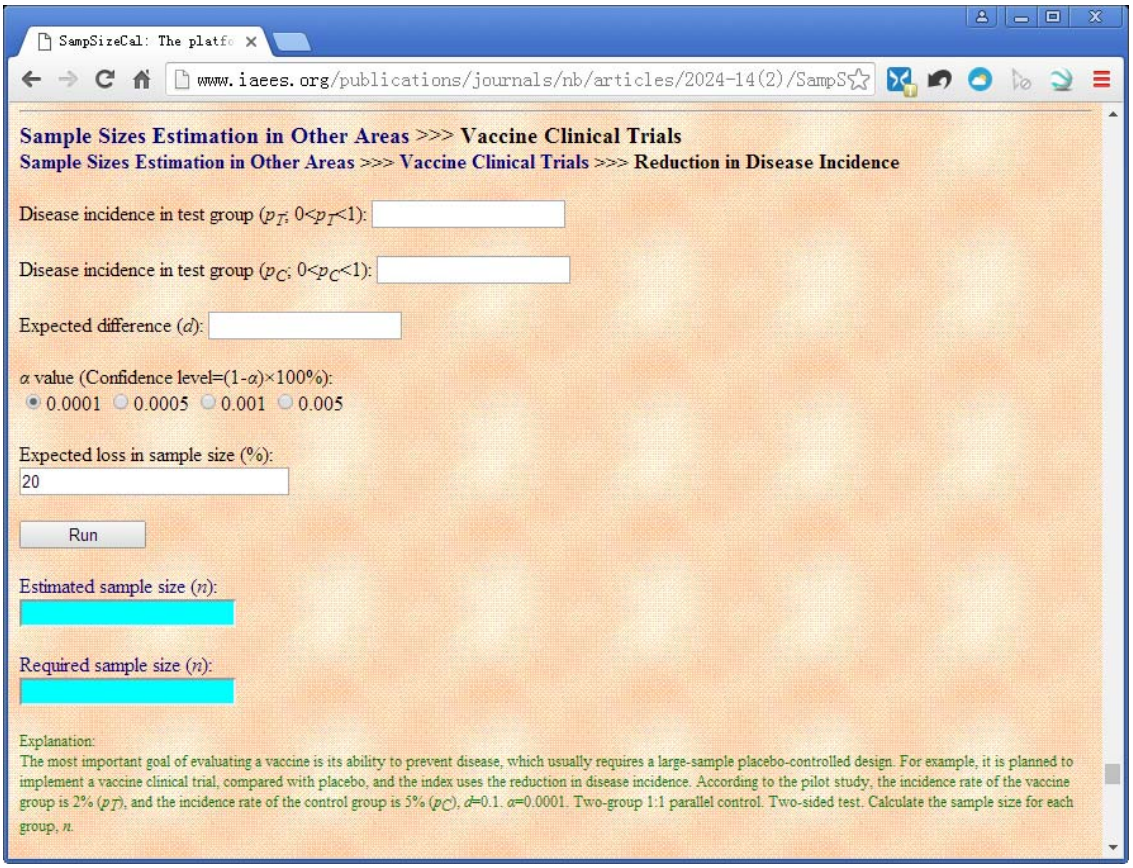


Fig. 1 Some page profiles in SampSizeCal.

5 Discussion and Explanation

5.1 Parameters Obtaining

Almost all of the methods for sample sizes require some parameters as standard deviations, differences, proportions, margins, etc. All these parameters can be obtained from small-scale pilot studys, past records, literature reports and reasonable estimations, etc.

5.2 Balanced Design

Some methods and software in the past allowed users to independently choose the distribution ratio of the total sample size in each group, resulting in arbitrariness, an unbalanced design, and reduced efficiency and credibility. In the present methods, in order to ensure efficiency and credibility, all adopt the balanced design, i.e., the total sample size is distributed in the same proportion among each group.

5.3 Significance Test, Non-inferiority or Superiority Test, and Equivalence Test

As indicated above, in the statistical hypothesis testing, α is the probability of rejecting the truth, i.e., the probability of rejecting the null hypothesis when there is no difference is true, and β is the probability of taking the false, i.e., the probability of accepting the null hypothesis when the difference is false. Sample size increases with the decrease of α and β. In addition, sample size increases with the decrease of specified difference and increase of the standard deviation achieved in pilot study. A small statistical power (1-β) will lead to the differences existing in the population cannot be detected, resulting in false negative results. The purposes of clinical studies are different, and the sample size estimation methods used are also different (Liang, 2014). In clinical trials, it is necessary to distinguish between significance tests and interval hypothesis tests. The significance test is used to infer whether two samples come from the same population. Its test hypothesis is the null hypothesis that the two groups are equal to each other, that is, the samples come from the same population. In clinical trials, for the evaluation of the therapeutic effects of two groups, the significance test results cannot evaluate the actual size of the difference, let alone whether the difference has practical clinical significance. They can only indicate whether the therapeutic effects of the two groups come from different populations. In clinical practice, it is often necessary to confirm whether a new drug is no worse than, equivalent to or even better than a standard effective drug, so non-inferiority/equivalence/superiority test is usually used. Their test hypothesis is an interval, so it can also be called "interval hypothesis" or "interval test". Interval hypothesis testing includes equivalence test, non-inferiority test and superiority test. A non-inferiority trial refers to a trial whose main research purpose is to show that the response to the experimental drug is not worse (non-inferior) than the control drug in a clinical sense. If the treatment difference (efficacy of the test drug - efficacy of the control drug) > 0, then the test drug is more effective. If the therapeutic effect of the experimental drug is less than 0, then the control drug is more effective; if the therapeutic effect of the experimental drug is allowed to be lower than that of the control drug within a certain range, the two drugs are still considered to be equally effective. That is, δ means the allowed maximum difference value that the therapeutic effect is not judged worse in a clinical sense. Then if the treatment difference >-δ, the test drug is non-inferior to the control drug. The δ is called the judgment margin (margin) of the non-inferiority test. Non-inferiority trials are usually used to compare a new treatment option with an effective drug or standard treatment regimen that is already on the market. Equivalence testing refers to a trial whose main purpose is to show that the magnitude of the difference between the responses of two or more treatments is not clinically important, usually by showing that the true difference is within the upper and lower bounds of clinically acceptable equivalence to confirm the original hypothesis. Only when both sets of hypotheses are established at the same time can it be considered equivalent. This is more common in bioequivalence of the same active ingredient and clinical equivalence verification when plasma cannot be measured. Superiority trials refer to the main research purpose of showing that the response of the drug under study is better than that of the

comparison preparation (positive or placebo control). Superiority trials are usually used for newly developed experimental drugs that have certain advantages and generally need to be compared with placebos undergo superiority trials to compare their true efficacy and safety to determine the benefits and risks of their marketing. If there are currently effective drugs that have been proven by superiority trials, they are often compared with them, and the efficacy of the drug to be verified is determined to be at least no worse (not inferior) to existing effective drugs as the minimum standard for its marketing. In clinical trials, the selection of clinical cutoff values should be jointly agreed upon by researchers and statisticians, and is based on the dual considerations of statistical reasoning and clinical judgment; if there is no recognized cutoff value, we may refer to Hou et al. (2009), and European Medicines Agency (2005), etc.

5.4 Randomized Clinical Trial (RCT)

Randomized clinical trial (RCT) is an important trial in clinical trials. According to the design scheme, it is often divided into parallel design, crossover design, factorial design and sequential design. Except for the sequential design, which does not require prior estimation of sample size, all other designs require estimation of sample size (Liang, 2014). (1) Parallel design: The research subjects are randomly assigned to two groups (or groups) and receive different treatments respectively. The two groups start the research at the same time and analyze and compare the research results at the same time. Double-blind randomized controlled trials with parallel designs are the gold standard for clinical trials. (2) Crossover design: A method in which four teams and two groups of subjects use two different treatment measures, and then the treatment measures are exchanged with each other, and finally the results are compared and analyzed. This design is more efficient than a parallel design and requires a smaller sample size. However, the intervention effects in the first stage may have an impact on the second stage, resulting in legacy effects or other interactive effects, making the design and analysis more complex. There are also shortcomings such as a long test period. (3) Factorial design: It is to combine the levels of two or more treatment factors and conduct experiments on various possible combinations to evaluate the individual effects of different treatments and the interaction of joint applications. Factorial design can analyze and deal with interactive factors, but the design and analysis are also more complex. (4) Sequential design: It does not specify samples before the test. They are assigned to the experimental group or the control group by randomization in order. After each test of one or a pair of subjects, analysis is carried out in a timely manner. Once it can be determined, the test can be stopped. The sequential design is in line with the reality that clinical patients seek medical treatment one after another. It is more suitable for paired comparisons of new drugs and old drugs or new drugs and placebos based on a single indicator, saving manpower and material resources. However, it is not suitable for experimental design of chronic diseases, multivariables, long-term follow-up, etc.

5.5 Consecutive Updates

The computational tool, SampSizeCal, is subject to consecutive updates in the future. Any constructive suggestions, corrections and supplements are encouraged. The possible updates can be found at:

http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/5-Zhang-Abstract.asp

References

Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. Nature, 567: 305-307. https://doi.org/10.1038/d41586-019-00857-9

Ardilly P. 2005. Sampling Methods: Exercises and Solutions. Springer, USA https://www.amazon.com/Sampling-Methods-Exercises-Solutions-2005-11-16/dp/B01A0B8NHA

Bergstrom CT, West JD. 2021. Manipulated P-values: Mathematical nosense in scientific papers.

https://www.laitimes.com/en/article/3km6i_41b77.html. Accessed 2022-4-23

Burnham KP, Anderson DR, Laake JL. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs, 72: 1-202. https://www.jstor.org/stable/3830641

Chow SC, Shao J, Wang H. 2008. Sample Size Calculations in Clinical Research (2nd ed). Chapman and Hall/CRC, Boca Raton, Florida, USA https://www.amazon.com/Calculations-Clinical-Research-Chapman-Biostatistics/dp/1584889829

Cochran WG. 1977. Sampling Techniques (3rd ed). Wiley, New York, USA. https://www.amazon.com/Sampling-Techniques-3rd-William-Cochran/dp/047116240X

Cohen J. 1988. Statistical Power Analysis for the Behavioral Science. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA https://www.amazon.com/Statistical-Power-Analysis-Behavioral-Sciences/dp/0805802835

Desu MM, RaghavaraoD.1990. Sample Size Methodology. Academic Press, New York, USA https://www.sciencedirect.com/book/9780122121654/sample-size-methodology

Downing JA, Perusse M, Frenette Y. 1987. Effect of interreplicate variance on zooplankton sampling design and data analysis. Limnology and Oceanography, 32: 673-680 https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1987.32.3.0673

Eberhardt LL.1978. Appraising variability in population studies. Journal of Wildlife Management, 42: 207-238 https://www.jstor.org/stable/3800260

Errington TM, Mathur M, Soderberg CK, et al. 2021. Investigating the replicability of preclinical cancer biology. eLife, 10: e71601. https://elifesciences.org/articles/71601

European Medicines Agency. 2005. Guideline on the choice of the non-inferiority margin. Doc. Ref. EMEA/CPMP/EWP/2158/99. London, UK https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf

Fisher RA. 1935. The Design of Experiments. Oliver and Boyd, Edinburg and London, UK. http://www.medicine.mcgill.ca/epidemiology/hanley/tmp/Mean-Quantile/DesignofExperimentsCh-III.pdf

Fleiss JL. 1986. The Design and Analysis of Clinical Experiments. John Wiley and Sons, New York, USA https://onlinelibrary.wiley.com/doi/book/10.1002/9781118032923

Fleiss JL. Levin B, Paik, MC. 2003. Statistical Methods for Rates and Proportions (3rd ed). John Wiley and Sons, New York, USA. https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428

Good PI. 2005. Resampling Methods: A Practical Guide to Data Analysis (3rd ed). Springer, USA https://www.amazon.com/Resampling-Methods-Practical-Guide-Analysis/dp/0817643869

Hou Y, Wu XY, Li K. 2009. Issues on the selection of non-inferiority margin in clinical trials. Chinese Medical Journal (Engl), 122(4): 466-470. https://pubmed.ncbi.nlm.nih.gov/19302756/

Huang HN. 2023. Statistics reform: practitioner's perspective. https://doi.org/10.13140/RG.2.2.31799.50084

Huang HN. 2021a. Statistics reform: challenges and opportunities. ScienceNet. https://blog.sciencenet.cn/blog-3427112-1318043.html. Accessed 2021-12-26

Huang HN. 2021b. What are the most misunderstood and misleading concepts or theories in statistics textbooks today? ScienceNet. https://blog.sciencenet.cn/blog-3427112-1269013.html. Accessed 2021-1-26

Ioannidis JPA. 2005. Why most published research findings are false. Plos Medicine, https://doi.org/10.1371/journal.pmed.0020124

Julious SA. 2020. Sample Sizes for Clinical Trials. Chapman and Hall/CRC, Boca Raton, Florida, USA

Kafdar K. 2021. Editorial: Statistical significance, *p*-values, and replicability. The Annals of Applied Statistics. https://doi.org/10.1214/21-AOAS1500

Krebs CJ. 1989. Ecological Methodology. HarperCollinsPublishers, New York, USA

Li HH. 2021. *p*-values are too sensitive, and the effect size is long and good to save. ScienceNet. http://blog.sciencenet.cn/blog-2619783-1286084.html. Accessed 2021-5-11

Li J, Fine J. 2004. On sample size for sensitivity and specificity in prospective diagonostic accuracy studies. Statistics in Medicine, 23: 2537-2550. https://pubmed.ncbi.nlm.nih.gov/15287083/

Liang XF. 2014. Sample size estimation. https://rstudio-pubs-static.s3.amazonaws.com/153235_a0277930a4924e46af765f4bbba3cdd6.html

Mace AE. 1964. Sample-Size Determination. Reinhold, New York, USA https://ui.adsabs.harvard.edu/abs/1965NucIM..34Q.179T/abstract

Machin D, Campbell M, Fayers P, Pinol A. 1997. Sample Size Tables for Clinical Studies (2nd ed). Blackwell, Malden, MA, USA https://www.deepdyve.com/lp/wiley/sample-size-tables-for-clinical-studies-2nd-edn-david-machin-michael-j-baWLRTYdpe

McShane BB, David Gal D. 2017. Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association, 112(519): 885-895. https://doi.org/10.1080/01621459.2017.1289846

Nature Editorial. 2021. Replicating scientific results is tough — but essential. Nature, 600: 359-360. https://doi.org/10.1038/d41586-021-03736-4

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science, 349(6251). https://doi.org/10.1126/science.aac4716

Pielou EC. 1977. Theoretical Ecology. WB Saunders, Philadelphia, USA https://www.amazon.com/Theoretical-Ecology-Applications-Robert-May/dp/0878935150

Ryan, Thomas P. 2013. Sample Size Determination and Power. John Wiley and Sons, New Jersey, USA https://www.wiley.com/en-us/Sample+Size+Determination+and+Power-p-9781118437605

Seber GAF. 1982. The Estimation of Animal Abundance and Related Parameters (2nd ed.). Griffin, London, UK https://www.amazon.com/Estimation-Animal-Abundance-G-Seber/dp/1930665555

Sellke T, Bayarri MJ, Berger JO. 2001. Calibration of p values for testing precise null hypotheses. The American Statistician, 55(1): 62-71. https://doi.org/10.1198/000313001300339950

Sun XJ. 2016. Improper use of *p*-values is a nonsense. ScienceNet. http://blog.sciencenet.cn/blog-41174-961169.html. Accessed 2016-4-5

Southwood TRE. 1978. Ecological Methods (2nd ed). John Wiley and Sons, New York, USA https://link.springer.com/book/10.1007/978-94-009-5809-8

Tille Y. 2006. Sampling Algorithms. Springer, USA https://www.amazon.com/Sampling-Algorithms-Springer-Statistics-Till%C3%A9/dp/0387308148

Tong C. 2019. Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. The American Statistician, 73(S1): 246-261. https://doi.org/10.1080/00031305.2018.1518264

Vrieze JD. 2021. Landmark research integrity survey finds questionable practices are surprisingly common. Science. https://doi.org/10.1126/science.abk3508

Wasserstein RL, Schirm AL, Lazar NA, 2019. Editorial: Moving to a world beyond "p<0.05". The American Statistician, 79: 1-19. https://doi.org/10.1080/00031305.2019.1583913

Yates F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46: 19-34. https://doi.org/10.2307/2280090

Zar JH. 1984. Biostatistical Analysis (2nd ed) Prentis Hall, New Jersey, USA https://www.amazon.com/Biostatistical-Analysis-4th-Jerrold-Zar/dp/013081542X

Zhang WJ. 2007. Methodology for Ecologigy Research. Sun Yat-sen University Press, Guangzhou, China https://book.douban.com/subject/2339027/

Zhang WJ. 2022a. Confidence intervals: Concepts, fallacies, criticisms, solutions and beyond. Network Biology, 12(3): 97-115. http://www.iaees.org/publications/journals/nb/articles/2022-12(3)/confidence-intervals-fallacies-criticisms-solutions.pdf

Zhang WJ. 2022b. Dilemma of t-tests: Retaining or discarding choice and solutions. Computational Ecology and Software, 12(4): 181-194. http://www.iaees.org/publications/journals/ces/articles/2022-12(4)/dilemma-of-t-tests.pdf

Zhang WJ. 2022c. *p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3): 80-122. http://www.iaees.org/publications/journals/ces/articles/2022-12(3)/p-value-based-statistical-significance-tests.pdf

Zhang WJ. 2023. A desktop calculator for effect sizes: Towards the new statistics. Computational Ecology and Software, 13(4): 136-181 http://www.iaees.org/publications/journals/ces/articles/2023-13(4)/4-Zhang-Abstract.asp

Zhang WJ, Qi YH. 2024. ANOVA-nSTAT: ANOVA methodology and computational tools in the paradigm of new statistics. Computational Ecology and Software, 14(1): 48-67 http://www.iaees.org/publications/journals/ces/articles/2024-14(1)/4-Zhang-Abstract.asp