—ヒト臨床試験（ヒト試験）における サンプルサイズ設計の統計学的留意事項—

鈴木　直子 (*SUZUKI Naoko*)[1*]　田中　瑞穂 (*TANAKA Mizuho*)[1]　野田　和彦 (*NODA Kazuhiko*)[1]
波多野　絵梨 (*HATANO Eri*)[1]　金子　拓矢 (*KANEKO Takuya*)[1]　中村　駿一 (*NAKAMURA Shunichi*)[1]
柿沼　俊光 (*KAKINUMA Toshihiro*)[1]　馬場　亜沙美 (*BABA Asami*)[1]　山本　和雄 (*YAMAMOTO Kazuo*)[1]

Key Words：ヒト臨床試験，ヒト試験，特定保健用食品，機能性表示食品，サンプルサイズ設計

Current Status and Issues of Clinical Trials for Efficacy and Safety Evaluation of Health Foods —Statistical considerations for sample size calculations in clinical trials—

Keywords: clinical trials, Foods for Specified Health Uses (FOSHU), Foods with Function Claims, sample size calculations

Authors:
Naoko Suzuki[1)]*, Mizuho Tanaka[1)], Kazuhiko Noda[1)], Eri Hatano[1)], Takuya Kaneko[1)], Shunichi Nakamura[1)], Toshihiro Kakinuma[1)], Asami Baba[1)], Kazuo Yamamoto[1)]
*Correspondence author: Naoko Suzuki

Affiliated institution
[1)] ORTHOMEDICO Inc.
2F Sumitomo Fudosan Korakuen Bldg., 1-4-1 Koishikawa, Bunkyo-ku, Tokyo, 112-0002, Japan.

はじめに

特定保健用食品や機能性表示食品は，一定の科学的根拠に基づいて特定の保健の目的が期待できる食品である。食品の機能性の科学的根拠を証明するためにヒト試験が実施されるが，その際は事前に研究デザインや解析手法などを厳密に計画する必要がある。試験計画の中でもサンプルサイズ設計は，アウトカムの評価や結果の解釈に影響を与えるため，試験計画における重要な要素の 1 つである。しかしながら，2017 年に消費者庁が実施した，機能性表示食品制度に届出済みの臨床試験の論文の調査では，34 編中 30 編が，サンプルサイズ設計の根拠が未記載となっていた [1)]。このような背景から，本稿では，サンプルサイズ設計の意義や統計学的留意事項，記載方法などをまとめた。

1. サンプルサイズとは？

サンプルサイズとは，研究対象の大元の集団である「母集団」からランダムに抽出した「サンプル（標本）」の数のことを言う。サンプルサイズは統計的仮説検定を行うにあたり重要な役割を担っている。次の例を用いて説明する。

母平均を μ，母分散を σ^2 とする正規分布 $N(\mu, \sigma^2)$ に従う母集団から n 個のデータ x_1，x_2，・・・，x_n を抽出し，μ について検定をしたとき，検定統計量 t_0 は次の式となる。

$$t_0 = \frac{\bar{x} - \mu_0}{\sqrt{V/n}} \quad (1)$$

[1] 株式会社オルトメディコ　* 責任著者
〒 112-0002 東京都文京区小石川 1-4-1 住友不動産後楽園ビル 2 階
Tel: 03-3818-0610 / Fax: 03-3812-0670

ここで，$\bar{x}$ と V は n 個のデータより求めた，標本平均値と標本分散であり，μ_0 は既知の値とする。この検定では，$|t_0|$ が自由度 $\phi = n-1$ の t 分布の両側 5% 点より大きければ，「有意差あり」と判定して，帰無仮説「H_0：$\mu = \mu_0$」を棄却し，対立仮説「H_1：$\mu \neq \mu_0$」を採用する。この時，注意する点が 2 つある。

まず 1 つ目は，式（1）に着目すると，検定統計量 t はサンプルサイズ n が大きくなると，$|t_0|$ の値も大きくなる。従って $\bar{x}$ と μ_0 との差が小さくても，サンプルサイズを大きくすればするほど，「有意差あり」と判定させることができる。2 つ目は，逆にサンプルサイズが小さい場合，$\bar{x}$ と μ_0 との差に実質的な差があるが，$|t_0|$ が大きくならず，「有意差あり」と判定されない場合がある点である。

1 つ目の状況においては，そこまで深刻な状況ではない。サンプルサイズが大きいならば，μ の点推定値として $\bar{x}$ をある程度信用できる。たとえ，検定で有意であったとしても，実質的に意味のある差なのかどうかを考察できる。一方，2 つ目の状況は深刻である。例えば，安全性の試験において，サンプルサイズが小さくて有意でないだけなのに，「これまで通り安全である（$\mu = \mu_0$）と判断したが，結果的には危険性が高まった」という事態を招いてしまうことが考えらえる。このような誤解を防ぎ，統計的検定結果を適切な判断に結びつけるためには，検出力やサンプルサイズ設計の知識が必要となる。

2. 2 種類の過誤とサンプルサイズ設計

本章では，まず，サンプルサイズ設計の重要な要素である，検定における 2 種類の過誤と検出力について説明する。続いて，サンプルサイズ設計の重要性について説明する。

2-1. 検定における 2 種類の過誤と検出力

仮説検定では，初めに帰無仮説（H_0）を設定し，それが棄却されたときの受け皿として対立仮説（H_1）を設定する。検定において，第 1 種の過誤とは，本当は H_0 が正しいのにも関わらず，H_0 を棄却してしまう過誤のことを指す。つまり，「差がない」のに「差がある」としてしまう過誤で，この過誤を犯す確率が有意水準（α）である。一般的に $\alpha = 0.05$ という小さな値が設定される。一方，本当は H_0 が正しくないにも関わらず，H_0 を棄却しない過誤のことを第 2 種の過誤といい，この過誤を犯す確率を β と表す。仮説検定では，あくまで α について扱っており，β については関心がもたれていない。検出力は β の逆であり，H_0 が正しくない時に，H_0 を棄却する確率である。つまり，第 2 種の過誤を犯す確率 β を 1 から引く（$1-\beta$）ことで表すことができる。β は 0.2 に設定することが推奨されていることから，検出力は $1-0.2 = 0.8$ となる [2]。これらを**表 1** にまとめた。

2 種類の過誤における基本的事項として，α と β はトレードオフ関係にあり，α を小さくしようとすると β が大きくなり，逆に α を大きくすると β が小さくなるといった特徴がある。また，サンプルサイズが大きいほど β は小さくなることも多くの検定で共通な基本的事項である。これらを含め，多くの検定で共通する基本的事項を以下にまとめた [3]。

1）帰無仮説は限定的であることから，第 1 種の過誤の確率（有意水準）α は 1 つに定まる。
2）対立仮説はパラメータが様々な値をとりうるので，第 2 種の過誤の確率 β はパラメータの値が異なると変化する。
3）α は小さく設定できたとしても，β は非常に大きな値になりうる。
4）α を大きくすると β は小さくなる。
5）サンプルサイズが大きくなると β は小さくなる。

表 1　検定における 2 種類の過誤と検出力

検定結果	本当に成り立っているのは	
	H_0 である	H_1 である
H_0 を棄却しない	正しい （その確率：$1-\alpha$）	第 2 種の過誤 （その確率：β）
H_0 を棄却する	第 1 種の過誤 （その確率：α = 有意水準）	正しい （その確率：$1-\beta$ = 検出力）

2-2. サンプルサイズ設計の重要性

サンプルサイズの大きさは，有意確率や第 2 種の過誤に影響を与えることを上述した。サンプルサイズを大きくすることは，推定値の精度を高めることだけでなく，第 2 種の過誤を犯す確率 β を小さくすることにより検出力を高めることができる。しかし，2 章で述べたようにサンプルサイズを大きくすることは，有意確率を簡単に下げることができ，ごく微量な差においても帰無仮説を棄却できるようになってしまうことに加え，第 1 種の過誤を犯す危険性が高くなる。

そこで，ある程度現実的な差を設定し，2 種類の過誤を犯さずにその差を検出するために，サンプルサイズの設計は重要となる。サンプルサイズに制限を加え，高い検出力を保持したまま，より少ないサンプルサイズで帰無仮説が棄却されるのなら，それは意味のある差があると判定できる。サンプルサイズ設計は，存在している意味のある差を見逃さないためや，意味のない差を検出しないためにある。

3. サンプルサイズ設計

試験計画におけるサンプルサイズは，試験デザイン，予算および実現可能性の重要な側面であり，通常，正式なサンプルサイズ計算を用いて決定される。サンプルサイズの設計は，一般的に 1 つの主要評価項目に基づいて行われ，その評価項目における差を高い確率で検出できるように，十分にサンプルサイズを大きくするべきという指針がある。

サンプルサイズを推定する際は，効果量を見積もり，α と $1-\beta$ の水準を設定する必要がある。ここで，効果量について説明する。

3-1. 効果量

効果量とは，群間での平均値の差の程度，変数間の関連の強さなど，研究関心の程度を表す値をデータの単位に左右されないよう標準化したものであり，2.2 節で述べた「ある程度現実的な差」のことを指す。また，差の大きさを表す効果量を d 族効果量，関連の強さを表す効果量を r 族効果量と言う。効果の推定はアウトカムの変数により異なる（**表 2**）。

アウトカムが連続変数の場合は，主に平均値の差の効果量（d）を表し，平均値の差を標準偏差で割ることにより効果量が計算される。例えば，独立した 2 群間の差の d は次の式で求められる。

$$d = \frac{\mu_1 - \mu_2}{S} \tag{2}$$

$$S^2 = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2} \tag{3}$$

ここで，群 1 および 2 のサンプルサイズは n_1 および n_2，平均値は μ_1 および μ_2，標準偏差は S_1 および S_2 とし，S は S_1 と S_2 をプールした標準偏差とする。ここで計算される d は，d 族効果量の中でも最もよく知られており，使用されている Cohen の d である [2]。d 族効果量を求める式は，長島俊輔（2018）の文献にまとめてあるので参考になる [4]。

効果量は相対的な数値であり，明確な基準はないが，Cohen は経験的な解釈として，小さな効果量とは $d = 0.2$，中程度の効果量とは $d = 0.5$，大きな効果量とは $d = 0.8$ と示している [2]。また，他の効果量と併せて，**表 3** に目安をまとめた。小さな効果量でも重要な意味を取りうる場合があるため，先行研究や関連する研究を踏まえて，最終的には研究者が判断する必要がある。

3-2. サンプルサイズの推定

以上より，サンプルサイズを推定する際に必要となる d, α および $1-\beta$ について説明した。これより，具体的なサンプルサイズ設計の推定方法の説明に移る。機能性表示食品制度では，プラセボ対照ランダム化比較試験が推奨されていることから，本稿では

表 2　サンプルサイズの設計におけるアウトカムの表示方法

項目	アウトカムの変数		
	二値変数	連続変数	イベント発生までの時間
各試験群の結果の表示	イベントの割合（%）	平均値と標準偏差	任意の時点でイベントが発生した割合（%）
効果の推定	オッズ比	平均値の差	ハザード比

両側 t 検定を用いた 2 群間比較試験の実施を想定した場合におけるサンプルサイズ設計の方法を紹介する。

対立仮説を $H_1: \mu_1 \neq \mu_2$ と設定し，$d > 0$ の時，高い検出力で H_0 を棄却したい。$n_1 = n_2$ として，必要なサンプルサイズ設計のための近似式は以下である。

$$n \approx 2\left(\frac{z_{\alpha/2} - z_{1-\beta}}{d}\right)^2 + \frac{z^2_{\alpha/2}}{4} \quad (4)$$

ここで，z_p は標準正規分布 $N(0, 1^2)$ における確率変数を示す。この確率変数の絶対値は，α および $1-\beta$ の設定によって変動する。$z_{\alpha/2}$ は α を小さくすると大きくなり，$z_{1-\beta}$ は，β を小さく，つまり $1-\beta$ を大きくすると大きくなる。仮に，$\alpha = 0.05$, $1-\beta = 0.8$ に設定し，$d = 0.8$ と見積もったとすると，サンプルサイズは

$$\begin{aligned} n &= 2\left(\frac{1.96 - (-0.842)}{0.8}\right)^2 + \frac{1.96^2}{4} \\ &= 24.528 + 0.960 \\ &\approx 26 \end{aligned} \quad (5)$$

となり，各群 26 名ずつ必要であることが分かった。サンプルサイズの推定式を見ればわかるように，d の大きさによっても変動する。式 (4) を用いて，α, $1-\beta$ および d を変化させた場合のサンプルサイズ設計の結果を**表 4** に示した。サンプルサイズを大きくすれば小さな差でも検出できるようになるが，統計学的な有意性と臨床的な有意性は異なることから，臨床的な有意性を考慮して検出したい差を設定する。その際，評価項目の妥当性が低ければ標準偏差が大きくなることが予測され，その結果，d を正しく評価できずに必要サンプルサイズが大きくなる可能性がある。従って，信頼性および妥当性の高い項目を主要評価項目として設定することが重要である。

4. プロトコルに記載する際の留意点

機能性表示食品制度において，最終製品を用いたヒト試験の結果を消費者庁に報告する際は，CONSORT 2010 声明に準拠した形式で査読付き論文として公表された論文を提出しなければならない[5, 6]。CONSORT は 25 項目のチェックリストから

表 3　Cohen による効果量の解釈の目安[2]

	効果量	効果小	効果中	効果大
平均の差の t 検定	d	0.20	0.50	0.80
相関	r	0.10	0.30	0.50
分散分析	η^2	0.01	0.06	0.14
カイ二乗検定	w	0.10	0.30	0.50
回帰分析	R^2	0.02	0.13	0.26
	f^2	0.02	0.15	0.35

表 4　有意水準 α，検出力 $1-\beta$ および効果量 d の設定に伴うサンプルサイズの変動

効果量 d	有意水準 α	検出力 $1-\beta$	n per group
0.80	0.05	0.80	26
	0.05	0.90	34
	0.01	0.80	39
	0.01	0.90	49
0.50	0.05	0.80	64
	0.05	0.90	86
	0.01	0.80	96
	0.01	0.90	121
0.20	0.05	0.80	394
	0.05	0.90	527
	0.01	0.80	586
	0.01	0.90	746

表 5　SPIRIT 2013 声明チェックリスト：14. サンプルサイズ（Sample size）[7]

Sample size	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations
サンプルサイズ	研究目的を達成するために必要な推計参加者数とその決定方法。サンプルサイズのすべての算定のもととなる，臨床的・統計学的仮定を含む。

表 6　サンプルサイズ設計の記載例

サンプルサイズは，●●を X 週間摂取することにより▲▲が改善する仮説に基づいて算出された。パイロット試験において，●●を X 週間後摂取した後の▲▲の実測値の群間差は $\Delta\mu_{1-2}$ であった（被験食品群；$N = n_1, \mu_1 \pm \sigma_1$, プラセボ群；$N = n_2, \mu_2 \pm \sigma_2$）（▲▲は高いほうが望ましい）。これらを踏まえて，群間の効果量の Cohen の d は 0.8 と見積った。統計学的有意水準（α）を両側 5%，統計学的検出力（$1-\beta$）を 80% とし，t 検定におけるサンプルサイズを計算すると，各群 26 名であった。10% の脱落を考慮して，各群 28 名とした。

構成されており，サンプルサイズについては，「どのように目標症例数が決められたか。」と記述されている[6]。これでは何を記述すればいいのか分かりにくいと思われるので，もう少し詳しく見てみる。

臨床試験では様々な指針が公表されており，SPIRIT 2013 声明は，すべての臨床試験のプロトコルに適合する 33 項目のチェックリストからなるガイドラインである[7]。また，SPIRIT は，CONSORT から適用可能な項目を正確に反映している。両チェックリストは，共通する項目の表現および構成が一貫しているため，SPIRIT に基づくプロトコルから CONSORT に基づく最終報告への移行は，容易であると考えられる[7]。ここで，SPIRIT のチェックリストのサンプルサイズについては**表 5** の通りに記載されている。以下に，記述する際の留意点をまとめた。

サンプルサイズ設計をプロトコルに記述する際には，一般的に，パイロット試験の各試験群における結果の測定値，アウトカムの検定方法，α，$1-\beta$，および算出された各試験群のサンプルサイズを含むべきである。また，各試験群に想定される結果の根拠や参考資料を提示することも推奨される。記載例を**表 6** に示した。しかし，実際にはパイロット試験そのものや希少疾患を対象にした試験など，パイロット試験を行うことができないが，サンプルサイズを設計したい場合がある。このような場合のプロトコルへの記述方法を 2 点まとめた。まず 1 点目は，研究に関する最新のエビデンスを用いて，目標の差を設定する方法である。理想的には，RCT のシステマティックレビューまたはメタアナリシスから得られたデータを用いると良い。RCT のデータがない場合には，観察研究からのデータを用いることができる。2 点目は，**表 3** に示した Cohen による効果量のカットオフ値を目安とし，サンプルサイズを設計する方法である。上記の様に，サンプルサイズが統計的に導き出されていない場合は，意図したサンプルサイズの理由とともに，その旨を明示する必要がある。

まとめ

本稿は，機能性表示食品制度下におけるサンプルサイズについて統計学的留意事項をまとめた。サンプルサイズは，有意水準，検出力および効果量の 3 つの要素の関連性から設計され，また，仮説検定の結果に影響を与える。そのため，サンプルサイズ設計の根拠をプロトコルに詳細に記述することは，その試験結果の解釈に繋がるため非常に重要である。

参考文献

1. 消費者庁．「機能性表示食品」制度における機能性に関する科学的根拠の検証　―届け出られた研究レビューの質に関する検証事業報告書
（2021 年 7 月 12 日アクセス可能：https://www.caa.go.jp/policies/policy/food_labeling/foods_with_function_claims/pdf/about_food_with_function_report_180416_0001.pdf）
2. Cohen, J.: Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates. 1988.
3. 永田靖：サンプルサイズの決め方．朝倉書店．2003.
4. 長島俊輔：看護学分野での統計改革を目指して：*t* 検定における *d* 族効果量の報告状況とその普及に向けた課題．日本看護研究学会雑誌．**41**(5): 1013-9, 2018.（DOI: https://doi.org/10.15065/jjsnr.20180422032）
5. 消費者庁．機能性表示食品の届出等に関するガイドライン（2021 年 3 月 22 日付け消食表第 120 号）
（2021 年 7 月 15 日アクセス可能：https://www.caa.go.jp/policies/policy/food_labeling/foods_with_function_claims/assets/foods_with_function_claims_210322_0002.pdf）
6. Schulz KF, Altman DG, Moher D.: CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *BMJ*. **152**(11): 726-32, 2010. (PMID: 20335313)
7. Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleza-Jeric K, Laupacis A, Moher D.: SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*. **346**: e7586, 2013. (PMID: 23303884)