解　説

看護研究におけるサンプルサイズの考え方

谷村　晋

Approaches to sample size determination in nursing research

Susumu Tanimura

Ⅰ．はじめに

研究調査データには必ず誤差が含まれる．誤差とは，真の姿とサンプル（標本）との乖離の程度のことをいう．誤差は「系統誤差」と「偶然誤差」の2つに大別され，この2つは全く異なる性質や特徴を持つ．一般的に，調査対象人数を増やせば増やすほど，信頼できる研究結果になる．その理由は，サンプルサイズを大きすくれば大きくするほど偶然誤差が小さくなり，精度（precision）や信頼性（reliability）が増大するからである．一方，どれだけサンプルサイズを大きくしても，系統誤差には影響しない．

サンプルサイズとその計算の根拠は，助成金申請書や研究倫理審査申請書の中で必ず示す必要がある．その主な理由は，(1) 安全性の確認されていない新しい介入による危険に多くの被験者を不必要にさらすべきでないという倫理的な観点，(2) 看護学的に意味のあるレベルを担保する精度以上に調査対象者を増やして，調査対象者のリソース（時間や労力）を無駄に消費するべきではないという倫理的な観点，(3) 研究にかかるコストを必要以上に大きくしないという経済的な観点による．

サンプルサイズを大きくすればするほど解析の精度が向上しP値が小さくなり，究極的には，看護学的に意味がないどんなに小さな差でも統計的に有意になる（Lantz, 2013）．例えば，2群の平均身長差が0.1mmでも，調査対象人数を増やし続ければいつかは有意になる．しかし，髪型を変えたら埋もれてしまうような微小な差は，看護学的に無意味であり，そのように大きなサンプルサイズにする合理性はなく，研究意義もない．そのため，近年急速に発展しているビッグデータなど大規模データを前提とした看護研究を例外として，一般的な看護研究ではちょうどいいサンプルサイズを見積もる必要が生じる．統計的検定で有意になる最低限のサンプルサイズはどれだけか，それを明らかにする手続きを例数設計またはサンプルサイズの決定という．サンプルサイズの計算方法について，さまざま方法が提唱されて乱立し，まだまだ開発途上の部分もあるものの，基礎的な統計手法については，定石といえる計算方法が出揃っている．

サンプルサイズの見積もりは，あくまでも大雑把な目安を得て概数を得る作業であり，絶対に正しい精緻な数値というものは存在しない．サンプルサイズの見積もりが正しかったのかどうかは，実際にデータを収集して分析した後に判明するものであり，不確定要素が多い事前の計算では限界がある．

しかしながら，サンプルサイズの計算を大きく誤り，大幅に少なく見積もってしまうと，サンプルサイズ不足となり解析の精度が落ち，目的とする看護学的に意味のある差や関連が存在していても，統計的有意性を確認できない事態に陥る．例えば，介入研究において介入群と対照群の差の推測を誤り，実際よりも大きな差が出ると見積もってしまうと，本当は介入効果が存在するはずなのに，それを統計的に検出できないという事態になる．そのため，事前の予測を行う際には，過去の文献や試験的なパイロット研究のデータなどに基づいて，慎重に行う必要がある．

看護研究の例数設計における方針には，3つの考え方がある．

1. 精度をもとに例数設計（信頼区間に基いて計算する）
2. 検出力をもとに例数設計（仮説検定に基いて計算する）
3. 実施可能性による例数設計（統計的推測とは無関係に計算する）

精度をもとに例数設計を行う方法は，信頼区間の幅

三重大学大学院医学系研究科

など精度の情報に基づいて必要サンプルサイズを求める方法である．探索的な研究の場合はこの方法を使わざるを得ない場合も多い．どちらかというと今後の研究に向けての情報収集などを目的としている場合に適しており，看護介入の有効性を検証する目的には向いていない．

検出力をもとに例数設計を行う方法は，統計的検定で有意になりやすい程度を考慮して必要サンプルサイズを求める方法であり，看護介入の有効性の検証を目的とする場合になどに適している．このアプローチにおいても，先行研究から予期される差，割合，相関係数，効果量など何らかの想定が必要になる．先行研究が全く見当たらず，手がかりがない場合はパイロット研究を実施し，自ら効果量など手がかりを求める必要がある．先行研究が見当たらず，パイロット研究も実施困難な場合は，Cohen の効果量（Cohen, 1988）から中程度の値を参考に計算する他はない．

実施可能性による例数設計を行う方法は，看護研究の実施可能性（実現可能性）に基づいて例数を決める方法であり，集められるだけ患者を集めるがそのときの打ち切りの目安を考える方法である．希少な疾患の患児など来院患者に限界がある場合，研究期間が限られている場合，研究資金や研究倫理の問題で多施設に展開できない場合など，そもそも大人数を集められないことがわかっている場合に用いる．

本稿は，サンプルサイズ推定法の学術的総括ではなく，看護学研究における実践的な手引き，あるいは計算方法を調査する手がかりの提供を目的としているため，看護学で用いられる研究デザインごとの分類整理を行う．また，サンプルサイズ計算の代表的な道具として，G*Power (Faul et al., 2009) や R (R Core Team, 2022) などがあるが，本稿では R を用いてサンプルサイズ計算の手順を解説する．

II. 実態調査

看護学研究に限らず実態調査はさまざまな分野で行われている．実態調査が全数調査である場合は標本を抽出するわけではないのでサンプルサイズの計算は不要である．しかし，標本調査の場合は，調査して得た標本割合や標本平均が母集団の母割合や母平均に十分に近いことを担保する必要があるため，サンプルサイズの見積もりが必要になる．信頼区間から導出したサンプルサイズの見積もりを行う場合は，信頼水準（confidence level），許容誤差（margin of error）を先に決めて，先行研究から得た割合や分散などを用いて計算を行う．ここで，信頼水準とは標本が調査対象の母集団をどれだけ正確に反映しているかを数値で表したものであり，許容誤差とはどの程度の誤差までなら許容できるのかを数値で表したものである．なお，仮説検定に基づくサンプルサイズの見積もりを行う場合，割合を求めるときには母比率の検定，平均値を求めるときには 1 標本 t 検定に基づいて計算を行う．ここでは，信頼区間から導出したサンプルサイズの見積もりについて解説する．

1. 割合を求める実態調査

特定の看護方法における実施割合や，法規制に対する賛成割合など，割合を求める実態調査の場合を想定する．割合を推定する場合の標本誤差 e は次のように定義される．

$$e = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

ここで $Z_{\alpha/2}$ は，有意水準が α であるときの Z 値であり，p は割合，n はサンプルサイズである．なお，信頼水準は $1-\alpha$ である．この標本誤差 e を許容誤差 d と読み替えて，n について解くと次にようになる．

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

慣習的に α や d は 0.05 とすることが多く，先行研究やパイロット研究から p を定めて n を求める．この n に回収率や有効回答率などによる水増しを行えば，最終的な必要最低限のサンプルサイズになる．R による計算例を付録 A.1 に示す．もし，先行研究が見当たらず，パイロット研究も困難であり，p について全く手がかりがない場合は $p = 0.5$ とする．事前の情報がないという理由で五分五分の割合を当てることは，無意味なサンプルサイズ計算に見えるが，$p = 0.5$ のときに n が最大になることから，最も安全な選択をしているといえる．

上述のサンプルサイズ計算は無限の母集団を前提としている．一方，例えば全国の特定認定看護師を母集団に想定する場合など，母集団のサイズが有限であると考えるほうが妥当な場合がある．有限母集団のサイズを N とすると，標本誤差 e は次のように定義される．

$$e = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

ここで$\sqrt{\frac{N-n}{N-1}}$は補正項とよばれるものであり，これにより標準誤差が割り引かれる．しかし，例えば，母集団のサイズ N が 10 万以上になると補正項はほぼ 1 になるため，実質的に無限母集団の場合と同じになる．さて，無限母集団の場合と同様に，標本誤差 e を許容誤差 d と読み替えてサンプルサイズ n について解くと，次のようになる．

$$n = \frac{N}{\left(\frac{d}{Z_{\alpha/2}}\right)^2 \times \frac{N-1}{p(1-p)} + 1}$$

この式を用いた R による計算例を付録 A.2 に示す.

2. 平均値を求める実態調査

平均値を推定するときの標本誤差 e は次のように定義される．前提条件の違いにより $Z_{\alpha/2}$ を用いる定義と $t_{\alpha/2}$ を用いる定義があるが，ここではは $Z_{\alpha/2}$ を用いることにする．

$$e = Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

ここで σ^2 は分散（既知の母分散）である．この標本誤差 e を許容誤差 d と見なして，必要最低限のサンプルサイズ n について解くと次のようになる．

$$n = \frac{Z_{\alpha/2}\sigma^2}{d^2}$$

この式は無限サイズの母集団の場合であり，有限母集団における標準誤差 e は補正項を用いて次のように定義される．

$$e = Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{N-n}{N-1}}$$

ここで標本誤差 e を許容誤差 d とおいて，サンプルサイズ n について解くと次のようになる．

$$n = \frac{N}{\left(\frac{d}{\sigma Z_{\alpha/2}}\right)^2 (N-1) + 1}$$

この式を用いた R による計算例を付録 A.3 に示す.

3. その他

看護学の実態調査では，割合や平均値の他にも，名義尺度（2 値変数や順序がない 3 値以上の変数）あるいは順序尺度を測定対象とする場合がある．名義尺度の場合は，割合のサンプルサイズ計算と同じ要領でサンプルサイズを求めることができる．順序尺度の場合は Whitehead（1993）が提唱する方法を用いる．この方法は，R の Hmisc パッケージに含まれる `posamsize()` に実装されている．

実態調査のサンプルサイズについて，ここで示した古典的な計算方法以外にも，これまでにさまざな計算方法が提唱されている．しかし，サンプルサイズの見積もりは，その計算の前提および仮定した数値に大きく左右されるため，算出すべき値から大きく乖離しないかぎり，計算方法やアルゴリズムの違いから生じる計算結果の僅少な差は問題にならないと考えられる．

III. 観察研究

2 つの変数の関連性や変数間の因果関係を調べる観察研究は最も一般的な看護研究の形態である．この研究デザインにおけるサンプルサイズの計算は，予定している統計的検定手法や，データ分析手法に基づいて計算する．複数の検定や分析を行う研究で，どの検定も等しく重要な場合は，念の為にそれぞれの手法における必要最低限のサンプルサイズを計算し，その中で最も大きいサンプルサイズを採用するようにする．例えば，t 検定で 30 人，fisher の正確確率検定で 25 人，重回帰分析で 45 人が必要な場合は，最も大きな 45 人を採用する．一方，それぞれの検定の重要性が等しくない場合は，重要な検定のみに基づいてサンプルサイズを見積もる．例えば，着目している 2 つの変数間の関連についてのみ興味があり，基本属性などその他の調査項目に関する統計的検定について検出力不足を許容できる場合は，その主要な検定に基づいてサンプルサイズを計算する．観察研究では回帰モデルが利用されることが多いが，交絡因子を含む調整変数（例えば，性・年齢など基本属性）としての役割を期待する変数に関しては，サンプルサイズの計算に組み入れる必要はない．

さらに，回収率や有効回答率による水増しをせずに，サンプルサイズの計算で得たサンプルサイズを最終的な必要最低限のサンプルサイズとしてしまうと，無回答や誤回答により，必要なサンプルサイズに届かないという事態に陥る危険があるため，注意が必要である．

1. 横断研究

短い期間で調査可能な横断研究デザインは看護学でも広く採用され，回帰モデルによる分析が行われる．モデル誤差に正規分布を仮定できる場合（つまり，目的変数が連続量変数である場合）は，古典的な線形回帰モデルが用いられ，正規分布以外の確率分布が仮定される場合は一般化線形回帰モデル（ロジスティック回帰モデル，ポアソン回帰モデル，コックス比例ハザードモデルなど）が用いられる．回帰モデルにおけるサンプルサイズの見積もりにはさまざまな提唱があるが，Cohen の効果量 f^2 に基づくサンプルサイズの計算方法（Cohen, 1988）が最も高い汎用性を有する．この方法は R の pwr パッケージの `pwr.f2.test()` に実装されている（付録 A.4）．重回帰モデルであれば，先行研究における R^2（決定係数）から効果量 f^2 を求める．モデル全体の効果量は次式で計算できる．

$$f^2 = \frac{R^2}{1-R^2}$$

特定の説明変数にのみ着目し，その他の調整変数などの説明変数を勘案する必要がなければ，特定の説明変数に着目した効果量を計算する．特定の説明変数を含めた場合の R^2 と含めない場合の R^2 をそれぞれ求め，その差分に基づいて効果量 f^2 を求める．

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$$

ここで R^2_{AB} は特定の説明変数を含めた場合の R^2，R^2_A は特定の説明変数を含めない場合の R^2 である．つまり，この式の分子は変数 B を追加した時の R^2 の差分を意味し，ΔR^2 と書く場合もある．効果量 f^2 を参照できる先行研究が見当たらない場合かつパイロット研究も実施できない場合は，上述の通り，Cohen の目安（Cohen,1988）から，中程度の効果量 f^2 を用いざるを得ない（付録 A.5）．なお，一般化線形モデルの場合は，重回帰モデルを線形予測子に置き換えて考えればよい．

効果量 f^2 を用いずに，先行研究の有病率比（prevalence ratio）やオッズ比（odds ratio, 以下 OR とよぶ）に基づいて横断研究におけるサンプルサイズを計算する方法も提唱されている（Woodward, 2013）．この方法は，R の epiR パッケージに含まれる `epi.ssxsectn()` に実装されている．

その他にも，ロジスティック回帰モデルのサンプルサイズ計算について，Peduzzi et al.（1996）は，目的変数 2 群の少ない方の割合を p，説明変数の項数を k としたとき，最小サンプルサイズを $n = 10k/p$ とすることを提唱している．例えば，説明変数が 3 つで，目的変数「あり」の割合が 20% のとき，$n = 10 \times 3/0.20 = 150$ となる．ただし，Long（1997）は，Peduzzi et al.（1996）の方法で 100 未満と計算された場合は 100 にすることを提案している．なお，これまでしばしば採用されてきた説明変数 1 つにつき 10 人で計算する方法は否定されている（van Smeden et al., 2016）ため，注意が必要である．

2. 共分散構造を解明する研究

看護学分野における観察研究では，1 つの目的変数と複数の説明変数から構成される単純な回帰モデルのみならず，複数の項目間の関係性やその程度を同時にモデル化する共分散構造解析（構造方程式モデリング，Structural Equation Modeling, 以下 SEM とよぶ）がよく用いられる．SEM は，複数の回帰モデル，相関分析，分散分析，因子分析を同時にモデリングできることから，応用性が大変高い．SEM のサンプルサイズ計算は，適合度指標と自由度に基づいて計算する方法が提唱されている（Moshagen & Erdfelder, 2016）．この方法は，Moshagen 自身の手により R の semPower パッケージ（Moshagen, 2021）に実装されている．このパッケージを用いた計算例を付録 A.6 に示す．

3. 症例対照研究

着目するアウトカムを有する群（症例群）と有さない群（対照群）を過去に遡って調査し，寄与要因を探る研究デザインとして症例対照研究があり，交絡因子の影響を取り除くために，共通する交絡因子を有する者同士でペア（あるいは 1: n の組）を組ませるマッチング法と，マッチング法を使わずにそのまま比較する方法がある．いずれの場合も，先行研究の OR を用いてサンプルサイズを見積もる（Dupont, 1988）．この方法は，R の epiR パッケージに含まれる `epi.sscc()` に実装されている．また，症例対照研究におけるサンプルサイズ計算の別の実装として samplesizelogisticcasecontrol パッケージ（Gail, 2021）がある．

4. コホート研究

着目する要因を有する群（曝露群）と有さない群（非曝露群）を前向きに調査し，アウトカムの発生頻度（罹患率あるいは死亡率など）から，着目する要因のリスクを評価する研究デザインとして，コホート研究がある．コホート研究におけるサンプルサイズの見積もりは，先行研究から曝露群のアウトカム発生頻度と非曝露群のアウトカム発生頻度を得て，それらをもとに計算する（Woodward, 2013）．この方法は，R の epiR パッケージに含まれる `epi.sscohortc()` に実装されている．

IV. 介入研究

看護介入の効果を測るためにしばしば介入研究が行われる．介入研究には，対照群がない前後比較試験，介入群と対照群を途中で入れ替えるクロスオーバー試験，介入群と対照群を無作為に割り付ける無作為化比較試験，無作為割付をしない非無作為化比較試験などがある．介入研究のサンプルサイズ計算は，介入効果の検出に十分な検出力を担保する目的で行われるため，その計算の根拠は，研究デザインではなく，使用予定の検定あるいは分析に基づく（クラスター無作為化比較試験などの例外を除く）．そのため，ここでは前後比較試験と無作為化比較試験を例に解説する．

1. 前後比較試験

前後比較試験は，介入前後でアウトカムが有意に変化したかを調べる研究デザインである．2 群の場合は対応のある t 検定（ただし，正規性を仮定できない場合は Wilcoxon の符号付順位検定）を用い，3 群以上の場合は繰り返し測定の分散分析（ただし，正規性を仮

定できない場合は Friedman 検定）を用いる．t 検定の場合は，先行研究やパイロット研究から，対応のある t 検定用の効果量 d を求め，それに基づいてサンプルサイズを見積もる．R では pwr パッケージに含まれる `pwr.t.test()` を用いて計算する（付録 A.7）．Wilcoxon の符号付順位検定の場合は，t 検定に基づいて計算したサンプルサイズを $\pi/3$ 倍すればよいが，モンテカルロ法を援用して計算する場合は，MKpower パッケージの `sim.ssize.wilcox.test()` を用いる．繰り返し測定の分散分析を利用する場合は，WebPower パッケージに含まれる `wp.rmanova()` を用いる．付録 A.12 に `wp.rmanova()` を用いた計算例を示す．

2. 無作為化比較試験

無作為化比較試験における統計的検定は，一般的に単純である．各群に割り付けたときに差が生じていないことを前提に，t 検定，fisher の正確確率検定，分散分析（一元配置分散分析，二元配置分散分析，繰り返し測定の分散分析）などを行う．多くの場合は，先行研究やパイロット研究から効果量を求め，その値に基づいてサンプルサイズを見積もる．

独立 2 標本の t 検定を用いる場合のサンプルサイズ計算を付録 A.9 に，fisher の正確確率検定を用いる場合のサンプルサイズ計算を付録 A.10 に，一元配置分散分析を用いる場合のサンプルサイズ計算を付録 A.11 に，繰り返し測定の分散分析を用いる場合のサンプルサイズ計算を付録 A.12 にそれぞれ示す．

t 検定において，等分散を仮定できず Welch の補正が必要な場合は，R の samplesize パッケージに含まれる `n.ttest()` を用いる．また，リッカート尺度の変数など同順位が多く出現するときの Wilcoxon 順位和検定に基づくサンプルサイズを計算するときには，R の samplesize パッケージに含まれる `w.wilcox.ord()` を用いる．

重回帰モデル（最小二乗法モデル）やロジスティック回帰モデルなど回帰モデルに基づくサンプルサイズ計算については，横断研究の節を参照されたい．

3. クラスター無作為化比較試験

病院勤務の看護師を対象にした環境型の介入研究など，個人単位の介入が難しく，病院などクラスター単位の介入を余儀なくされる無作為化比較試験は，クラスター無作為化比較試験とよばれる．この分析には，クラスター内環境の違いを考慮した線形混合モデルが必要になる．Raudenbush（1997）は，効果量 f，クラスター数，級内相関係数（intraclass correlation coefficients, ICC, 以下 ICC とよぶ）に基づくクラスター無作為化比較試験のサンプルサイズ計算を提唱しており，R では WebPower パッケージに含まれる `wp.crt2arm()` あるいは `wp.crt3arm()` に実装されている．

アウトカムの発生がゼロのクラスターが存在する場合，介入効果以外のメカニズムによって発生がゼロになる場合をゼロ過剰（zero inflation）という．そのような場合に，ポアソン回帰モデルや負の二項回帰モデルに基づく通常のサンプルサイズ計算では，正しい見積もりができない．そのため，Z. Zhou et al.（2022）はクラスター無作為化比較試験におけるゼロ過剰モデルに対応したサンプルサイズの計算方法を提唱し，Supplementary Materials にこの方法を実装した R コードを掲載している．

4. Stepped-wedge クラスター無作為化比較試験

Stepped Wedge デザインは，クラスター無作為化比較試験の新しい研究デザインとして注目されている．これは，介入開始時期をクラスター単位で無作為化し，観察期から介入期に移行する時期がクラスターによって異なる研究デザインであり，最終的にすべてのクラスターで介入が行われるため，非介入のために生じる研究倫理の問題を回避できる．この研究デザインにおけるサンプルサイズ計算は X. Zhou et al. (2018) や Liet al. (2018) などによって提唱されている．R では swdpwr パケージ（Chen, Zhou, Li, & Spiegelman,2022a; Chen, Zhou, Li, & Spiegelman.,2022b）がこれらの方法を実装している．

V. 尺度開発

看護学研究では，因子分析を用いて心理学的尺度の開発を行うことも多い．Mundfrom et al.（2005）はシミュレーションを用いて検討し，因子分析の必要最低限サンプルサイズについて，質問項目の 3 倍から 20 倍でかつ 100 から 1000 に収まる対象者数を提案している．尺度開発ではさまざまな角度から妥当性と信頼性を検証するが，その際に用いられる ICC に基づくサンプルサイズ推定について，Zou（2012）が提唱する手法が R の ICC.Sample.Size パッケージ（Rathbone et al., 2015）に実装されている．

VI. その他

サンプルサイズの見積もりを行う際に先行研究が全く存在しない場合は，ごく少数を対象としたパイロット調査を実施することになる．看護学研究の教科書を参照すると，パイロット調査のサンプルサイズは主研究のサンプルサイズの 10% とされている（Lackey & Wingate, 1997）こともあるが，そもそも，主研究のサ

ンプルサイズを見積もる手がかりとして，パイロット調査を行う場合は，いわゆる「缶詰の中の缶切りを取り出す」状態となり，本末転倒である．看護学研究におけるパイロット調査のサンプルサイズについて，Nieswiadomy and Bailey（2020）は 10 人，Hertzog（2008）は群ごとに 10 から 40 人と提案している．

看護学分野では，質的研究が行われることも多い．質的研究のサンプルサイズが小さすぎると，理論的飽和に達しなかったり，多すぎると捌き切れずに分析が甘くなり，分析の深さに支障が生じる可能性がある．そもそも事前に質的研究の必要最低限サンプルサイズを見積もることに妥当性がない（Marshall et al.,2013）という考え方がある一方で，質的研究のサンプルサイズについての検討が報告されている．質的研究の論文を調査した研究によると，サンプルサイズの分布（n = 560）は 20 人と 30 人の二峰性で中央値が 28 であった（Mason, 2010）ことから，Dworkin（2012）は 25 人から 30 人を推奨している．また，Boddy（2016）は，均一集団を対象とするのであれば 12 人で飽和に達する可能性が高く，30 人を超えると多すぎると指摘している．

VII. おわりに

看護学研究の例数設計について，ごく初歩的な検定に基づくサンプルサイズ計算について限定的に解説した文献はあるものの（Hayat, 2013; Ingram, 1998; Tam et al., 2020; Taylor & Muncer, 2000），網羅的に解説した文献は見当たらない．本稿では看護学研究で用いられる例数設計手法について，実践的観点から網羅的な整理を行った．

いずれの方法でも，サンプルサイズ計算は一定の確率で無作為に抽出をした標本を前提としている．例えば，サンプルサイズを見積もった結果，有限母集団 4,000 人の中から 30 人を抽出することになった場合，このサンプルサイズの計算結果は，選び出される確率はどれも 30/4000 = 0.0075 という確率[*1]で無作為に抽出されると前提を置いた場合の計算結果である．しかし，実際には，無作為抽出を実現することは困難な場合が多く，その場合は無作為抽出の前提が崩れている．このとき，データの偏りである選択バイアスの問題をよく検討するのみならず，検出力不足についても吟味する必要がある．

例数設計は研究計画段階で実施するものであるが，仮説検定に基づいて見積もった必要最低限のサンプルサイズが妥当であるか否かは，データ収集と分析を終えてみなければ判らない．実際，看護学論文について事後的な検出力を計算した研究では，大規模調査や効果量が大きい場合を除いて，多くの研究において検出力が 80% に満たないサンプルサイズであったことが明らかにされている（Polit & Sherman, 1990）．そのため，結果的にサンプルサイズが不足したか否かについて，事後的に算出をした検出力（$1-\beta$）を表中あるいは本文に書き添えることが推奨されている（Zhang et al., 2019）．なお，Hayat（2013）が指摘するように，事後的な検出力が不足していても，その看護研究あるいは論文が直ちに価値ないということにはならない．なぜならば，検出力が不足していても，それらの論文はメタアナリシスに組み入れることが可能であり，存在しないよりも少しでも情報がある方に意味があるからである．

統計学の発展は日進月歩であり，サンプルサイズの見積もりに関する数理的な基礎理論が不変であっても，既存のアルゴリズムが改良されたり，未対応だった状況に対応した新しい計算手法が提唱されるなど，進化が続いている．そのため，本稿で解説したサンプルサイズの計算方法はやがて別の方法に置き換わる可能性があり，その意味でサンプルサイズの計算について総括的整理の定期的な更新が必要であると考えられる．本稿では R を用いた具体的な計算例を付録に示しているが，とりわけ計算プログラムなど実践的なツールは変更が加えられやすく注意が必要である．

利益相反

本研究における利益相反は存在しない．

文献

Boddy, C. R. (2016). Sample size for qualitative research. *Qualitative Market Research: An International Journal, 19*(4), 426–432. https://doi.org/10.1108/QMR-06-2016-0053

Chen, J., Zhou, X., Li, F., & Spiegelman, D. (2022a). Swdpwr: A SAS macro and an R package for power calculations in stepped wedge cluster randomized trials. *Computer Methods and Pro-grams in Biomedicine, 213*, 106522. https://doi.org/10.1016/ j.cmpb.2021.106522

Chen, J., Zhou, X., Li, F., & Spiegelman., D. (2022b). *Swdpwr: Power calculation for stepped wedge cluster randomized trials* [R package version 1.7]. https://CRAN.R-project.org/package-swdpwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*

*1 ここでは説明を簡単にするため，復元抽出と非復元抽出の違いを無視する

(2nd ed.). Lawrence Erlbaum.

Dupont, W. D. (1988). Power calculations for matched case-control studies. *Biometrics, 44*(4), 1157–1168.

Dworkin, S. L. (2012). Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior, 41*(6), 1319–1320. https://doi.org/10.1007/s10508-012-0016-6

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gail, M. H. (2021). *Samplesizelogisticcasecontrol: Sample size and power calculations for case-control studies* [R package version 2.0.0]. https : / / CRAN . R -project . org / package – samplesizelogisticcasecontrol

Hayat, M. J. (2013). Understanding sample size determination in nursing research. *Western Journal of Nursing Research, 35*(7), 943–956. https://doi.org/10.1177/0193945913482052

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31*(2), 180–191. https://doi.org/10.1002/nur.20247

Ingram, R. (1998). Power analysis and sample size estimation. *NT Research, 3*(2), 132–139. https : / / doi . org / 10 . 1177 / 174498719800300210

Lackey, N., & Wingate, A. (1997). Advanced design in nursing research. SAGE Publications, Inc.

Lantz, B. (2013). The large sample size fallacy. *Scandinavian Journal of Caring Sciences, 27*(2), 487–492. https://doi.org/10.1111/j.1471-6712.2012.01052.x

Li, F., Turner, E. L., & Preisser, J. S. (2018). Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics, 74*(4), 1450–1458. https://doi.org/10.1111/biom. 12918

Long, J. (1997). *Regression models for categorical and limited depen-dent variables*. Sage Publications.

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does sam-ple size matter in qualitative research?: A review of qualitative interviews in is research. *Journal of Computer Information Systems, 54*(1), 11–22. https://doi.org/10.1080/08874417.2013.11645667

Mason, M. (2010). Sample size and saturation in phd studies using qualitative interviews. *Forum Qualitative Sozial-forschung / Forum: Qualitative Social Research, 11*(3). https://doi.org/10.17169/fqs-11.3.1428

Moshagen, M. (2021). Sempower: *Power analyses for sem* [R package version 1.2.0]. https://CRAN.R-project.org/package–semPower

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing struc-tural equation models. *Structural Equation Modeling: A Mul-tidisciplinary Journal, 23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *Inter-national Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Nieswiadomy, R. M., & Bailey, C. (2020). *Foundations of nursing re-search* (7th). Pearson.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per vari-able in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373–1379. https://doi.org/10.1016/ s0895-4356(96)00236-3

Polit, D. F., & Sherman, R. E. (1990). Statistical power in nursing re-search. *Nursing Research, 39*(6), 365–370.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rathbone, A., Shaw, S., & Kumbhare, D. (2015). *Icc.sample.size: Cal-culation of sample size and power for icc* [R package version 1.0]. https://CRAN.R-project.org/package–ICC.Sample.Size

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Tam, W., Lo, K., & Woo, B. (2020). Reporting sample size calcula-tions for randomized controlled trials published in nursing journals: A cross-sectional study. *International Journal of Nursing Studies, 102*, 103450. https://doi.org/10.1016/j.ijnurstu.2019.103450

Taylor, S., & Muncer, S. (2000). Redressing the power and effect of sig-nificance. A new approach to an old problem: Teaching statis-tics to nursing students. *Nurse Education Today, 20*(5), 358–364. https://doi.org/10.1054/nedt.2000.0429

van Smeden, M., de Groot, J. A., Moons, K. G., Collins, G. S., Alt-man, D. G., Eijkemans, M. J., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology, 16*(1), 163. https://doi.org/10.1186/s12874-016-0267-3

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine, 12*(24), 2257–2271. https://doi. org/10.1002/sim.4780122404

Woodward, M. (2013). Epidemiology: *Study design and data analysis* (3rd ed.). Chapman; Hall/CRC.

Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: Is it an informative and mean-ingful analysis? *General Psychiatry, 32*(4), e100069. https: // doi.org/10.1136/gpsych-2019-100069

Zhou, X., Liao, X., Kunz, L. M., Normand, S.-L. T., Wang, M., & Spiegelman, D. (2018). A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics, 21*(1), 102–121. https://doi.org/10.1093/biostatistics/kxy031

Zhou, Z., Li, D., & Zhang, S. (2022). Sample size calculation for cluster randomized trials with zero-inflated count out-comes. *Statistics in Medicine, 41*(12), 2191–2204. https : //doi.org/10.1002/sim.9350

Zou, G. Y. (2012). Sample size formulas for estimating intraclass cor-relation coefficients with precision and assurance. *Statistics in Medicine, 31*(29), 3972–3981. https://doi.org/10.1002/sim.5466

要　旨

看護研究を開始する前に必要最低限のサンプルサイズについて，その見積もりが求められる．サンプルサイズが必要サイズよりも小さければ検出力不足を引き起こし，データの中に存在する差や関連を検出できなくなる．一方，サンプルサイズを過剰に大きくすることは経済的あるいは倫理的な問題を引き起こしかねない．そのため看護学研究におけるサンプルサイズ計算の網羅的な整理が必要であるが，そのような文献はほとんど見当たらない．そこで，看護学研究でよく用いられる研究デザインや分析手法ごとにサンプルサイズの計算手法を整理することを目的として，実態調査，観察研究，介入研究，その他に分けて解説を行った．また，実践的な手がかりを提供することを目的として，R による計算の実践例と解説を加えた．

キーワード：看護研究，サンプルサイズ，効果量，検出力

付録 A　R による計算例

A.1　無限母集団を対象として割合を調べる実態調査のサンプルサイズ計算

先行研究で 30% の人が賛成とした調査を自分の地域でも実施する場合を想定する．許容誤差 d を 5%，信頼水準 λ を 90% としたとき，Z 値は次のようになる．

```
> Z <- abs(qnorm((1-0.9)/2))
> Z
[1] 1.644854
```

$Z_{0.9/2} \simeq 1.64$ であった．この値を用いて，必要最低限のサンプルサイズ n を次のように計算する．ここで，母集団のサイズが無限であることを仮定する．

```
> Z^2*0.3*(1-0.3)/0.05^2
[1] 227.2657
```

数字を切り上げる必要があるため，必要最低限のサンプルサイズは 228 人になる．次に，回収率や有効回答率による水増しを行って，最終的な必要最低限のサンプルサイズにする．

例えば，先行研究から質問票の返送・回収率が 40%，有効回答率が 95% として水増しを行うには次のようにする．

```
> 227.2657/0.4/0.95
[1] 598.0676
```

最終的な必要最低限のサンプルサイズは 599 人になる．

A.2　有限母集団を対象として割合を調べる実態調査のサンプルサイズ計算

信頼水準 95%，許容誤差 5%，先行研究に基づく割合を 0.3，母集団のサイズを 4000 である場合を想定する．まず Z 値を求め，続いてサンプルサイズを計算する．

```
> Z <- abs(qnorm((1-0.95)/2))
> Z
[1] 1.959964
```

手計算の場合は Z =1.96 としてよいが，ここではこの計算した Z の値をそのまま用いる．

```
> 4000/((0.05/Z)^2*((4000-1)/(0.3*(1-0.3)))+1)
[1] 298.6638
```

数字を切り上げる必要があるので，必要最低限のサンプルサイズは 299 人になる（回収率や有効回答率の水増しは省略）．

A.3　平均を調べる実態調査のサンプルサイズ計算

有限母集団について，標準偏差 σ = 4，サイズ N = 4000，許容誤差 d =1.5，信頼水準 95% の場合のサンプルサイズを計算する．Z 値を計算し，続いてサンプルサイズを計算する．

```
> Z <- abs(qnorm((1-0.95)/2))
> N <- 4000
> sigma <- 4
> d <- 1.5
> N/((d/(sigma*Z))^2*(N-1)+1)
[1] 27.13849
```

計算の結果，28 人が必要であることが判明した．

A.4　回帰モデルののサンプルサイズ計算

効果量 f^2，説明変数の自由度，有意水準，検出力から回帰モデルのサンプルサイズを見積もる．説明変数が全て連続変量の変数であれば，「説明変数の個数－1」が説明変数の自由度になる．カテゴリー変数を含む場合は，ダミー変数に変換してから数え上げることになる．例えば，性別（男，女），婚姻（既婚，未婚，その他），身長と 3

つの説明変数がある場合，性別の自由度は 2-1 で 1，婚姻の自由度は 3-1 で 2，身長の自由度は 1 となり，1+2+1-1=3 であるため最終的な説明変数の自由度は 3 になる．

効果量 f^2 が 0.6，自由度が 3，有意水準が 0.05，検出力が 0.8 であるとき，次のように計算する．

```
> pwr::pwr.f2.test(u = 3, f2 = 0.6, power = 0.8)

     Multiple regression power calculation

              u = 3
              v = 18.48919
             f2 = 0.6
      sig.level = 0.05
          power = 0.8
```

事前に library(pwr) を実行していれば，上記の「pwr::」は不要である．sig.level オプションの既定値は 0.05 であるため，省略可能である．検出力を指定する power オプションについて，既定値がないためこのオプションを省略することは出来ない．ここで，サンプルサイズ $n = v + u + 1$ であるため，$n = 3+18.48919+1 = 22.48919$ となり 23 人が必要であることが判明した．

A.5　Cohen の効果量の目安

先行研究が見当たらずパイロット研究も困難な場合は，Cohen の目安（Cohen, 1988）を用いる．pwr パッケージに含まれる cohen.ES() を用いて値を取り出すことができる．例えば，効果量 f^2 の中程度の値を取り出すには次のようにする．

```
> pwr::cohen.ES(test="f2", size = "medium")

     Conventional effect size from Cohen (1982)

           test = f2
           size = medium
    effect.size = 0.15
```

test オプションに効果量の種類を与える．効果量 f^2 の他に指定できる効果量は p, t, r, anov, shisq である．効果量の大きさは size オプションで指定し，small, medium, large の中から選ぶ．中程度の効果量 f^2 は 0.15 であった．

A.6　SEM のサンプルサイズ計算

適合度指標の RMSEA を 0.05 にするときの SEM のサンプルサイズを計算する．有意水準 $\alpha = 0.05$，検出力 $1 - \beta = 0.8$，自由度を 100 としたとき，サンプルサイズ計算は次のようになる．

```
> library(semPower)
> semPower(type = 'a-priori', effect = .05, effect.measure = 'RMSEA',
+           alpha = .05, power = .80, df = 100)
$type
[1] "a-priori"

[ 中略 ]

$requiredN
[1] 164

[ 後略 ]
```

ここで，type オプションに指定した a-priori とは，研究開始前の事前計算という意味である．調査データの回収後に検出力を計算する場合は，type オプションに post-hoc を指定する．計算の結果，164 人が必要最低限のサンプルサイズになった．

A.7 対応のある t 検定に基づくサンプルサイズ計算

先行研究などから得た効果量 d が 0.6，有意水準を 0.05，検出力を $1-\beta=0.8$ としたときの「対応のある t 検定」に基づくサンプルサイズ計算は次のようになる．

```
> pwr::pwr.t.test(d = .6, power = .8, type = "paired")

     Paired t test power calculation

              n = 23.79452
              d = 0.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

サンプルサイズを計算した結果，24 人が必要と判明した．

A.8 1 群の複数回測定した場合の分散分析に基づくサンプルサイズ計算

先行研究などから得た効果量 f を 0.43，有意水準 $\alpha=0.05$，検出力 $1-\beta=0.8$ とし，1 群を 3 回測定し，有意な変化を検討するときの分散分析に基づくサンプルサイズ計算は，次のようになる．

```
> WebPower::wp.rmanova(ng = 1, nm = 3, f = .43, power = .8, type = 1)

Repeated-measures ANOVA analysis

           n    f ng nm nscor alpha power
     53.6198 0.43  1  3     1  0.05   0.8
```

[後略]

ng オプションは群数，nm オプションは繰り返し測定回数をそれぞれ指定する．1 群を繰り返し測定する場合は type = 1 を与える．計算の結果，54 人が必要であることが判明した．

A.9 対応のない t 検定に基づくサンプルサイズ計算

先行研究などから得た効果量 d が 0.43，有意水準 $\alpha=0.05$，検出力 $1-\beta=0.8$ のときの「対応がない独立 2 標本の t 検定」に基づくサンプルサイズ計算は，次のようになる．

```
> pwr::pwr.t.test(d = .43, power = .8)

        Two-sample t test power calculation

                 n = 85.86943
                 d = 0.43
         sig.level = 0.05
             power = 0.8
       alternative = two.sided

NOTE: n is number in *each* group
```

type オプションを省略すると，規定値の独立 2 標本の t 検定になる．計算結果から 86 人ずつ合計 172 人が必要であることが判明した．次に，片方の群の人数が限られることが事前にわかっている場合を考える．例えば，第 1 群を 50 人と固定した場合は pwr.t2n.test() を用いて次のように計算する．

```
> pwr::pwr.t2n.test(n1 = 50, d = .43, power = .8)

     t test power calculation

             n1 = 50
             n2 = 292.011
              d = 0.43
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

上記により，第 2 群は 293 人が必要であることが判明した．

A.10 fisher の正確確率検定に基づくサンプルサイズ計算

fisher の正確確率検定は分割表の検定であるため，χ^2 検定のサンプルサイズ計算で代用できるが，ここでは四分割表を対象とした fisher の正確確率検定に基づくサンプルサイズ計算の手順を説明する．この計算には exact2x2 パッケージに含まれる ss2x2() を用いる．四分割表を 2 つの割合とみなして，それぞれ 0.2 と 0.8 としたときの必要サンプルサイズの計算は次のようになる．

```
> exact2x2::ss2x2(.2, .8, power = .8, approx = FALSE)
      Power for Fisher's Exact Test
       power = 0.8115276
          n0 = 12
          n1 = 12
          p0 = 0.2
          p1 = 0.8
   sig.level = 0.05
 alternative = two.sided
```

[後略]

library(exact2x2) を事前に実行していれば「exact2x2::」は不要である．sig.level オプションも省略可能である．計算の結果，24 人が必要であることが判明した．なお，2 つの割合を同人数で計算したくない場合は，n1.over.n0 オプションを用いる．

A.11 一元配置分散分析に基づくサンプルサイズ計算

先行研究などから得た効果量 f を 0.23，群数 k を 3，有意水準 α = 0.05，検出力 1 − β = 0.8 としたときの必要最低限サンプルサイズの求め方は，次のようになる．

```
> pwr::pwr.anova.test(f = .23, k = 3, power = .8)
     Balanced one-way analysis of variance power calculation
              k = 3
              n = 61.71968
              f = 0.23
      sig.level = 0.05
          power = 0.8
NOTE: n is number in each group
```

計算の結果，各群 62 人合計 186 人が必要であることが判明した．

A.12 繰り返し測定の分散分析に基づくサンプルサイズ計算

先行研究などから得た効果量 f を 0.43，有意水準 α = 0.05，検出力 1 − β = 0.8 とし，2 群を 5 回繰り返し測定し，交互作用を検討するときの分散分析に基づくサンプルサイズ計算は，次のようになる．

```
> WebPower::wp.rmanova(ng = 2, nm = 5, f = .43, power = .8, type = 2)
     Repeated-measures ANOVA analysis
       n    f   ng   nm  nscor  alpha  power
 65.76073 0.43    2    5      1   0.05    0.8
```

[後略]

付録と同じく，ng オプションは群数，nm オプションは繰り返しの測定回数である．type オプションについて，群間差を調べときは type = 0，郡内差を調べるときは type = 1，群間差と郡内差の交互作用を調べるときは type = 2 とする．nscor オプションは球面性の仮定が満たされていれば省略できるが，そうではない場合は，非球面性補正係数を与える．計算の結果，66 人が必要であることが判明した．