心理学のためのサンプルサイズ設計入門

村井潤一郎・橋本貴充

2017 年（講談社）

2　検定力分析に基づくサンプルサイズ設計（杉澤）

2.2　検定のロジックと検定力を考慮することの重要性

検定力について考えるための問題

[1] 赤玉と白玉が 100 個入っている袋からランダムに 1 つ取り出したところ、赤玉だった。袋の中身が

1. 赤玉 5 個、白玉 95 個
2. 赤玉 80 個、白玉 20 個

のいずれかとしたら、どちらだと考えるか。（→ 多分、2 番）

[2] また、袋の中身が

1. 赤玉 5 個、白玉 95 個
2. 赤玉 6 個、白玉 94 個

のいずれかとしたら、どちらだと考えるか。（→ 強いて言うなら 2 番か？）

検定力は低いが有意な結果が得られた、というのは後者の状態。
→ 帰無仮説のみに注目するのではなく、「帰無仮説のもとでは得られにくく、かつ、対立仮説のもとでは得られやすい結果」が得られたときに帰無仮説を棄却するという、多くの人が直感的に行う判断と同じロジックを取っている。

2.3　検定力分析の方法

2.3.4　母集団において期待される効果量

Cohen (1992) では、

- 測定尺度に依存しない
- 連続量である
- 0 以上の値を取る
- 帰無仮説のもとでは 0 となる

性質を持つものとして、8 種類の検定について効果量の指標を定義。

Table 2.1: 効果量の定義

検定の種類	効果量の指標	小	中	大
t 検定	$d = \frac{m_A - m_B}{\sigma}$	0.20	0.50	0.80
無相関検定	r	0.10	0.30	0.50
カイ 2 乗検定	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	0.10	0.30	0.50
分散分析	$f = \frac{\sigma_m}{\sigma}$	0.10	0.25	0.40
重回帰分析	$f^2 = \frac{R^2}{1-R^2}$	0.02	0.15	0.35

2.5 R による検定力分析の実行

2.5.4 1 要因分散分析

パッケージ `pwr` の関数 `pwr.anova.test()` を用いて、

```
pwr.anova.test(k=3, sig.level=0.01, f=0.25, power=0.8)
```

とすると、3 群の比較、有意水準 0.01、効果量 0.25、検定力 0.8 で検定を行うためのサンプルサイズ（群あたり）が得られる。

2.5.5 2 要因分散分析

パッケージ `pwr` の関数 `pwr.f2.test()` を用いて、

```
pwr.f2.test(u=1, f2=0.25^2, power=0.8)
```

とする。u は分子の自由度。
　必要なサンプルサイズは、アウトプット

```
     Multiple regression power calculation

              u = 1
              v = 125.5312
             f2 = 0.0625
      sig.level = 0.05
          power = 0.8
```

に分母の自由度 v として報告される。分母の自由度が

$$(\text{n of levels in A}) \times (\text{n of level in B}) \times (n-1)$$

によって求められることから、逆算して n を求める。
　重回帰分析の場合も、$u = 1, v = N - p - 1$ であることを利用すれば同じ。

3 信頼区間に基づくサンプルサイズ設計（石井）

3.2 1 群の平均値差の場合

1 群の母平均の CI は

$$[\overline{X} - t_1 s\sqrt{1/N}, \overline{X} + t_1 s\sqrt{1/N}] \tag{3.1}$$

なので、$t_1 s\sqrt{1/N}$ が CI の半幅[1]。

[1] s^2 は不偏分散であることに注意。

ここで、$t_1 s\sqrt{1/N}$ をデータ分布の SD s の何倍に抑えたいか考える。すると、

$$h_1 = \frac{t_1 s\sqrt{1/N}}{s} = t_1\sqrt{1/N} \tag{3.2}$$

を一定の値 h 以下にするような N を求めればよい。
つまり、

$$N \geq \frac{t_1^2}{h} \tag{3.3}$$

をみたす最小の N が求めるサンプルサイズとなる。

メモ

t_1 にも自由度として N が含まれているので、探索的に解く必要がある。

メモここまで

h はいくらであればよいのか？

$h = 0.5$ が一つの目安。

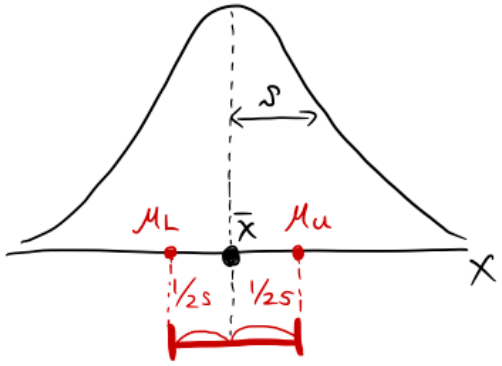
1. 相対評価では $1SD$ $(= \pm 0.5s)$ を区切りの目安としていた
2. h をあまりに小さくすると、実質的に同等のものを異なるものとして主張しかねない
   - 2 群の平均値差については、$h = 1/3s$ までの違いは同等とみなすことが多い　□ 要チェック

CI の半幅が s の $1/2$ に収まる（$h = 0.5$）とはどういう意味か？

母集団平均について、帰無仮説 $H_0 : \mu = \mu_0$ を考えたとき、

$$|\overline{X} - \mu_0| < 0.5s \Leftrightarrow \frac{|\overline{X} - \mu_0|}{s} < 0.5$$

であれば、H_0 は棄却されない。



母集団平均値として
データと整合的な範囲

$H_0 : \mu \in [\mu_L, \mu_u]$
は棄却されない。

3.3 対応のない 2 群の平均値差

平均値差の CI は、

$$\left[\overline{X}_1 - \overline{X}_2 - t_* s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 + t_* s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \tag{3.4}$$

である。ここで、$n_1 = n_2 = n$, $s_1 = s_2 = s_*$ と仮定すると[2]、

$$n \geq \frac{2t_*^2}{h^2} \tag{3.6}$$

によってサンプルサイズを決定できる。

メモ

CI の半幅は $t_* s_* \sqrt{2/n}$ なので、これを s_* の何倍に抑えたいか考えるとすると、

$$h_* = \frac{t_* s_* \sqrt{2/n}}{s_*} = t_* \sqrt{n/2}$$

となる。これを n について解けばよい。

メモここまで

はてな？

分散について、母集団値を $\sigma_1^2 = \sigma_2^2$ とおくのは仮定の話なので納得できるが、現実に得られるデータについて $s_1^2 = s_2^2$ とおくのは無理があるのでは。

本文では、等分散が仮定できないときは大きい方の *SD* にもとづいて h を決めるとある。しかし、そもそも等分散でなくても、s_* を基準に考えているのであれば式 (3.6) は成り立つはず。

はてな？ここまで

CI の半幅の s_* に対する比 $h = 0.5$ の解釈

母集団平均の差について帰無仮説 $H_0 : \mu_{\text{diff}} = \mu_{\text{diff}_0}$ を考えたとき、$\mu_{\text{diff}_0} \in [\mu_{\text{diff}_L}, \mu_{\text{diff}_U}]$ であれば棄却されない。つまり、

$$|\ \mu_{\text{diff}_0}\ | < 0.5 s_* \Leftrightarrow \frac{|\ \mu_{\text{diff}_0}\ |}{s_*} < 0.5$$

であれば、H_0 は棄却されない。

メモ

要は、平均値差の CI の下限と上限と標本平均との比が 0.5 に収まるようなサンプルサイズを決定するということ。

メモここまで

[2] s_1^2, s_2^2 は不偏分散。

3.4　対応のある 2 群の平均値差

平均値差の CI は、

$$\left[\overline{X}_1 - \overline{X}_2 - t_d s_d\sqrt{1/N},\ \overline{X}_1 - \overline{X}_2 + t_d s_d\sqrt{1/N}\right] \tag{3.7}$$

である。2 群の標本分散について $s_1^2 = s_2^2 = s^2$ と仮定すると、

$$N \geq \frac{2(1-r)t_d^2}{h^2} \tag{3.9}$$

を満たす最小の N が、CI の半幅の s に対する割合 h にもとづくサンプルサイズ。2 群の相関係数 r が小さいほど必要な N は大きくなる[3]。

メモ

CI の半幅は $t_d s_d\sqrt{1/N}$ なので、これを s の何倍に抑えたいか考えると、

$$h_* = \frac{t_d s_d\sqrt{1/N}}{s} = \frac{t_d\sqrt{s_1^2 + s_2^2 - 2rs_1s_2}\sqrt{1/N}}{s} = \frac{t_d s\sqrt{2(1-r)/N}}{s} = t_d\sqrt{\frac{2(1-r)}{N}}$$

となる。これを N について解けばよい。

メモここまで

CI の半幅の s に対する比 $h = 0.5$ の解釈

メモ

基本は対応のない 2 群と同じ。平均値差の CI の下限と上限と標本平均との比が 0.5 に収まるようなサンプルサイズを決定することになる。

メモここまで

はてな？

CI の半幅と s_d との比をもとに N を決定することもできる。その場合、$s_1^2 \neq s_2^2$ であってもよい。ただし、「群内の *SD* で標準化された平均値差」としての効果量ではなくなるので、h の意味はより不明確になる。

はてな？ここまで

3.5　相関係数

変換 $z = \operatorname{arctanh} r$ を考えたとき、r の CI は

$$\left[\tanh\left(z - z_0\sqrt{\frac{1}{N-3}}\right), \tanh\left(z + z_0\sqrt{\frac{1}{N-3}}\right)\right] \tag{3.10}$$

となる。この CI は r を中心に非対称であるので、CI の半幅は CI 幅の 1/2 と考え、

$$\frac{\tanh\left(z + z_0\sqrt{\frac{1}{N-3}}\right) - \tanh\left(z - z_0\sqrt{\frac{1}{N-3}}\right)}{2} < h_*$$

を考える。これを解くことで、必要な N が得られる。

[3] 負になることは考慮しない。

4　認知心理学研究におけるサンプルサイズ設計（井関）

4.2　既存の検定力分析の枠組みの限界

デザインの複雑さに関する認識が研究領域としての統計学とユーザとしての心理学者の間で違っている

- 心理学の論文では、2-3 要因は当たり前、4-5 要因も見かける
- 統計学の論文では、1 要因の被験者間バランスデザインを想定したものが多く、2 要因混合デザインですら複雑なものに相当する
  - 主効果と交互作用の違い、アンバランスデザイン、等分散性の仮定、球面性の仮定、など統計的な懸案事項が山ほどある

繰り返しに対する扱いの問題

同条件下で異なる刺激の提示を被験者内で繰り返した場合、測定値を平均して分析にかけることが多い。
　→「繰り返し」数は結局のところ「サンプルサイズ」としての単位には含まれず、被験者の人数のみが問題とされる。

4.3　一般的 ANOVA デザインにおける検定力分析

4.3.1　PANGEA

Westfall (2016) による *PANGEA* (Power ANalysis fo GEneral Anova desings) [4] [8]
　以下のものはすべて要因として扱う：

- 研究者にとって主要な関心のある変数（群、条件の別など）
- 従属変数のばらつきを説明しそうな他のグループ変数（参加者、材料の違いなど）

例えば、A 群と B 群のいずれかに割り付けられた被験者について、いずれの群も刺激 P と Q が提示されるが、材料 X と Y は群間でカウンターバランスを取って提示したとする。この場合、被験者は（いずれかの群にしか所属しないので）群要因にネストし、 材料も（群 A には X、群 B には Y というようになっているので）群要因にネストしており、群要因と刺激要因はクロスしていることになる。

被験者	群	刺激	材料
1	A	P, Q	X
2	A	P, Q	X
⋮	⋮	⋮	⋮
$N-1$	B	P, Q	Y
N	B	P, Q	Y

通常であれば、このデザインは一要因被験者内計画（刺激要因の水準 P, Q を被験者内で操作）として扱われることが一般的だが、実際はもっと複雑。

4.3.3　固定要因と変量要因

Westfall, Judd, & Kenny (2015) は、固定要因と変量要因を区別するための経験則として、追試するときに変えても良いと思えるかどうかで判断するという方法を提案。[9]
　さらに複雑なデザインについては、Judd, Westfall, & Kenny (2017) を参照。[2]

[4] https://jakewestfall.shinyapps.io/pangea/

4.3.4 繰り返し

PANGEA では、固定効果も変量効果も含むすべての要因を組み合わせた際の各セルの観測数を繰り返しという。

4.4 PANGEA

4.4.8 実験パラメータの設定

- 効果量
  - Cohen の $d = 0.45$ がデフォルトで入っている[5]
- サンプルサイズ
- 繰り返し数
- **分散分割係数** *variance partitioning coefficient*
  - 複数の変量要因が含まれる場合の、変量要因の分散の割合（変量要因の主効果、変量要因同士の交互作用、固定効果と変量要因の交互作用）
  - デフォルトでは、低次の要因ほど多くの分散が割り当てられる

効果量の値をわずかに変更しただけでも、検定力にかなりの影響が出る。

4.5 今後の課題とまとめ

- PANGEA は混合モデルを用いた分散分析を想定している
- 効果量を設定する際に関連する現象を全く見いだせないとしたら、それは研究しようとしている効果や現象を既存の研究の文脈に適切に位置づけられていないからではないか

5 臨床心理学研究におけるサンプルサイズ設計（国里）

5.1 実態

Journal of Consulting and Clinical Psychology では、APA の *Reporting Standards for Research in Psychology* に従って無作為化比較試験を行うことが求められている。

- 事前のサンプルサイズ設計の有無
- どのようにサンプルサイズを決定したのか報告

無作為化比較試験における報告ガイドライン CONSORT[6] 声明でも、サンプルサイズ設計の方法を報告するよう求めている。

5.2 補足事項

優越性試験・非劣性試験・同等性試験

無作為化試験は、

- 優越性試験 *superiority trial*: どちらの介入のほうが優れているか検証
- 非劣性試験 *non-inferiority trial*: ある介入が他の介入よりも明らかに劣ることがないことを示す[7]

[5] 社会心理学分野の大規模なメタ分析（Richard, Bond Jr., & Stokes-Zoota, 2003）[4] の結果に基づく。
[6] Consolidated Standards of Reporting Trials
[7] 副作用は少ないが、効果は劣らない新薬など

  - 非劣性マージンを設定
  - 有意差がない or CI が非劣性マージンに収まれば、非劣性と判断
- 同等性試験 *equivalence trial*: 複数の介入の間に効果の違いがないことを示す
  - 同等性マージンを設定
  - CI が同等性マージンに収まれば、同等と判断

に分類できる。

群間差の決定方法

Hislop et al. (2014)[8] による分類

- **重要な差** *important difference*: 治療者や患者が実際に差を感じるような意味付けのなされた差
  - アンカー: Global Rating Scale などの外的基準
  - 分布: 測定誤差を超えるさを重要な差とするなど
  - 医療経済学: 治療にかかるコストに見合うアウトカムが得られるか
  - 標準化効果量: Cohen のカットオフなど
- **現実的な差** *realistic difference*: 過去の研究における群間差を参考
  - パイロット研究: エビデンスがない場合に現実的な差を検討

いずれにも該当するものとして、意見聴取、エビデンスの展望、がある。

5.3 研究紹介

5.3.1 優越性試験のサンプルサイズ設計

クラスター無作為化試験において必要となる N を、完全無作為化試験において必要となる N を `pwr` パッケージの `pwr.t.test()` 関数で求めたのち、

$$1 + (m-1) \times \mathrm{ICC}$$

をかけて求める。[5]

ICC は級内相関係数で、不明な場合は 0.05 と設定されることが多い。

5.3.2 非劣性試験のサンプルサイズ設計

R の `TrialSize` パッケージの `TwoSampleMean.NIS()` 関数を用いる。

5.3.3 信頼区間に基づくサンプルサイズ設計

母比率の信頼区間

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$$

から、目標とする信頼区間の半幅を M とし、

$$M > z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$$

より

$$N > z^2_{1-\alpha/2}\frac{p(1-p)}{M^2}$$

としてサンプルサイズを求める。

[8] http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001645

7　発達心理学・教育心理学研究におけるサンプルサイズ設計（宇佐美）

7.1　2 つのモデル

実験群では $X = 0.5$、統制群では $X = -0.5$ となるようなエフェクトコーディングを考えて、以下のような 2 つのモデルを作る。

MRT

$$Y_{ij} = \beta_{j0} + \delta_m X_{ij} + e_{ij}, \quad \beta_{0j} = \beta_0 + u_{j0}$$

CRT

$$Y_{ij} = \beta_{j0} + \delta_c X_j + e_{ij}, \quad \beta_{0j} = \beta_0 + u_{j0}$$

7.2　効果量と平均値差の SE

7.2.1　級内相関 ρ と効果量 Δ

上記のモデルでいずれも

$$e_{ij} \sim N(0, \sigma_1^2), \quad u_{j0} \sim N(0, \sigma_0^2)$$

とする。このとき、級内相関は

$$\rho = \frac{\sigma_0^2}{\sigma_1^2 + \sigma_0^2} = \frac{\sigma_0^2}{\sigma^2}$$

であり、

- 0.05: 小
- 0.10: 中
- 0.15: 大

とする目安がある。[3]
　また、効果量を標準化された平均値差として

$$\Delta = \frac{\delta}{\sigma}$$

によって定義する。

7.2.2　平均値差 $\hat{\delta}$ の SE

帰無仮説 $H_0 : \delta = 0$ を検定するために、検定統計量

$$z = \frac{\hat{\delta}}{se(\hat{\delta})}$$

を用いる。
　このとき、$se(\hat{\delta})$ は

- MRT: $\sigma\sqrt{\frac{4(1-\rho)}{nJ}}$
  - ρ が大きいほど群間の等質性が高まる（$\mathrm{se}(\hat{\delta}_m)$ は小さくなる）
- CRT: $\sigma\sqrt{\frac{4[\rho(n-1)+1]}{nJ}}$

– ρ が大きくなると群間の等質性は満たされにくい（$\mathrm{se}(\hat{\delta}_c)$ は大きくなる）

である。

一般に

$$\mathrm{se}(\hat{\delta}_c) \geq \mathrm{se}(\hat{\delta}_m)$$

であり、また $\mathrm{se}(\hat{\delta}_c)$ の方が ρ の影響を受けやすい。

7.3 検定力分析

MRT

$$nJ > \frac{4(z_{1-\alpha/2} + z_\phi)^2(1-\rho)}{\Delta^2}$$

ただし、検定力を ϕ とする。

MRT の場合は、nJ が同じであれば検定力は変わらない。よって、$nJ = N$ として検定力分析できる。

CRT

$$n > \frac{4(z_{1-\alpha/2} + z_\phi)^2(1-\rho)}{\Delta^2 J - 4(z_{1-\alpha/2} + z_\phi)^2\rho}$$

$$J > \frac{4[\rho(n-1)+1](z_{1-\alpha/2} + z_\phi)^2}{\Delta^2 n}$$

- CRT では n と J の割合によって nJ が同じであっても検定力は変わってくる。
- 級内相関 ρ が大きく、かつ J が小さい場合、 n をいくら増やしても検定力 ϕ に到達しない。[6, 7]

7.4 正確度分析

効果量 Δ の CI 幅を L 未満に抑えることを考える。

MRT

$$nJ > \frac{16z_{1-\alpha/2}^2(1-\rho)}{L^2}$$

CRT

$$n > \frac{16z_{1-\alpha/2}^2(1-\rho)}{L^2 J - 16z_{1-\alpha/2}^2\rho}$$

$$J > \frac{16z_{1-\alpha/2}^2[\rho(n-1)+1]}{L^2 n}$$

- ρ によってサンプルサイズは大きく左右される
- J が小さく ρ が一定以上の場合、いくら n を増やしても CI 幅を L 以下に抑えることはできない

はてな？

この方法では、サンプルサイズが定まれば CI 幅は一位に定まるようになっている。検定力と異なり CI はそれ自体が確率変数であるから、どれくらいの確率でこの幅に収まるのか、評価できないと厳しい。

はてな？ここまで

7.5 まとめ

- ここではランダム切片モデルを扱っており、効果の集団間差は考慮していない。
- CRT で n を見積もってから J を求める場合、集団の大きさの**調和平均**を利用することが推奨されている。
- 分散成分については既知と考えている。未知とした場合は、OD Plus[9] や PowerUP! [10] などのソフトウェアを参照。
- 独立変数が連続的である場合のサンプルサイズ設計法の確立が必要。
- 級内相関 ρ についての経験的知見を積み重ねていくことが重要。

参考文献

References

[1] Hislop, J., Adewuyi, T. E., Vale, L. D., Harrild, K., Fraser, C., Gurung, T., ... & Norrie, J. D. (2014). Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review. *PLoS Med, 11*(5), e1001645.

[2] Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601-625.

[3] Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological methods, 5*(2), 199-213.

[4] Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363.

[5] 丹後俊郎 (2006). クラスター無作為化試験 丹後俊郎・上坂浩之（編）臨床試験ハンドブック — デザインと統計解析 —（pp. 456-463）朝倉書店

[6] 宇佐美慧 (2011). 階層的なデータ収集デザインにおける 2 群の平均値差の検定・推定のためのサンプルサイズ決定法と数表の作成 教育心理学研究, *59*(4), 385-401.

[7] Usami, S. (2014). Generalized sample size determination formulas for experimental research with hierarchical data. *Behavior research methods, 46*(2), 346-356.

[8] Westfall, J. (2016). PANGEA: Power ANalysis for GEneral Anova designs. Manuscript in preparation. Retrieved from http://jakewestfall.org/publications/pangea.pdf (April 25, 2017)

[9] Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*(3), 390-399.

[9] http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od
[10] http://web.missouri.edu/~dongn/PowerUp.htm