CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Institut für Biometrie

und klinische Epidemiologie

Sample Size Planning for the Wilcoxon-Mann-Whitney Test with Clustered Data

Erin Sprünken

erin-dirk.spruenken@charite.de

Symposium Translational Studies

21.03.2024

SAMPLE SIZE PLANNING

- Sample size is generally an issue
- The more – The better
- Feasibility and ethical aspects?

$\Rightarrow$ Planning, such that n is minimal but the test reaches certain power

SAMPLE SIZE PLANNING

- Sample size is generally an issue
- The more – The better
- Feasibility and ethical aspects?

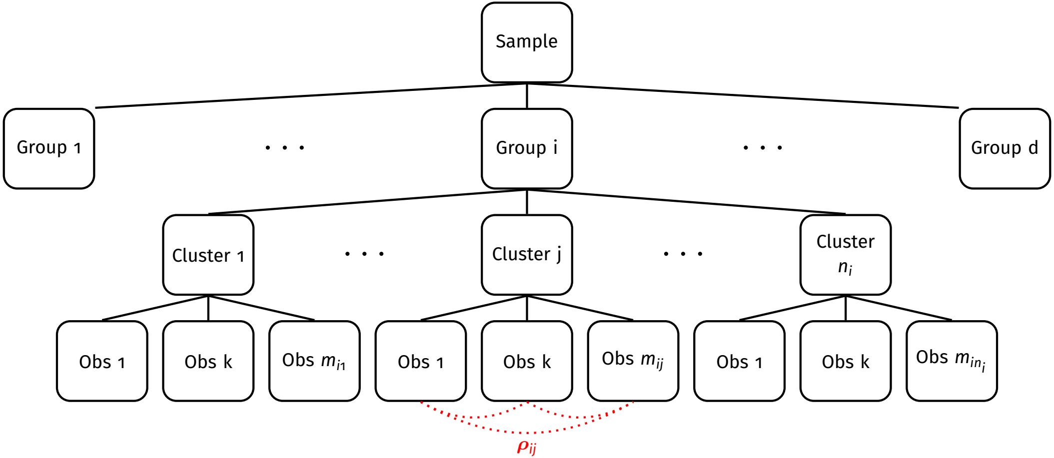
$\Rightarrow$ Planning, such that n is minimal but the test reaches certain power

Planning sample sizes is well established for common procedures (e.g. t-test, Wilcoxon-Mann-Whitney test), **but** what if we have clustered data ...

WHAT IS CLUSTERED DATA? – I

- Social Sciences and Economics
- Neurology and Neurosurgery
- Ophtalmology
- Dentistry
- Animal Trials

WHAT IS CLUSTERED DATA? – II



EXISTING METHODS

General methods exist for ...

- Clustered Binary Data
- Clustered Categorical Data
- Clustered Continuous Data

... with extensions to specific settings (e.g. with restrictive hypotheses)

EXISTING METHODS

General methods exist for ...

- Clustered Binary Data
- Clustered Categorical Data
- Clustered Continuous Data

... with extensions to specific settings (e.g. with restrictive hypotheses)

Our aim is

- Providing a method with little assumptions
- Providing a method robust to different distributions
- A method that can be easily used or implemented by practitioners!

MODEL

$$X_{ijk} \sim F_i, \quad i = 1, 2;\ j = 1, \ldots, n_i;\ k = 1, \ldots, m_{ij};\ n = n_1 + n_2$$

$$\mathbf{X}_{ij} = \left(X_{ij1}, \ldots, X_{ijm_{ij}}\right)^{\mathsf{T}}$$

$$\mathbb{C}or\left(\mathbf{X}_{ij}\right) = \rho_{ij}\mathbf{1}_{m_{ij}}\mathbf{1}_{m_{ij}}^{\mathsf{T}} + (1 - \rho_{ij})\mathbf{I}_{m_{ij}}$$

MODEL

$$X_{ijk} \sim F_i, \quad i = 1, 2;\ j = 1, \dots, n_i;\ k = 1, \dots, m_{ij};\ n = n_1 + n_2$$

$$\mathbf{X}_{ij} = \left(X_{ij1}, \dots, X_{ijm_{ij}}\right)^{\mathsf{T}}$$

$$\mathbb{C}or\left(\mathbf{X}_{ij}\right) = \rho_{ij}\mathbf{1}_{m_{ij}}\mathbf{1}_{m_{ij}}^{\mathsf{T}} + (1 - \rho_{ij})\mathbf{I}_{m_{ij}}$$

Assumptions

(A1) $m_{ij} = m < \infty \; \forall\, i, j$

(A2) m <u>is known a priori and not subject to planning</u>

(A3) $\rho_{ij} = \rho \; \forall\, i, j$

(A4) ρ <u>is known a priori and not subject to planning</u>

(A5) $X_{ijk} \perp X_{ij'k'} \; \forall\, j \neq j',\ k, k'$

(A6) $X_{ijk} \perp X_{i'j'k'} \; \forall\, i \neq i',\ j, j', k, k'$

THE (NONPARAMETRIC) RELATIVE EFFECT

- The relative effect is defined as $p = \int F_1 \mathrm{d} F_2$
- The relative effect p is the probability $\mathbb{P}\left(X_{1jk} < X_{2jk}\right) + \frac{1}{2}\mathbb{P}\left(X_{1jk} = X_{2jk}\right)$
- Describes, whether X_{2jk} tends to larger/smaller values than X_{1jk}.
- Thus:
  - $p = 0.5$ means, there is no stochastic tendency (neither random variable tends to larger values)
  - $p > 0.5$ means, X_{2jk} tends to larger values than X_{1jk}
  - $p < 0.5$ means, X_{2jk} tends to smaller values than X_{1jk}.

TEST STATISTIC

Hypothesis

$$\mathrm{H}_0 : F_1 = F_2 \quad \text{vs.} \quad \mathrm{H}_1 : F_1 \neq F_2$$

Statistic

$$\begin{aligned} \sqrt{n}\left(\widehat{p} - p\right) &\overset{\text{asy}}{\sim} N(0, \mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p}\right)) \\ \Rightarrow t = \sqrt{n}\frac{\widehat{p} - p}{\sqrt{\mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p}\right)}} &\overset{\text{asy}}{\sim} N(0, 1) \end{aligned}$$

THE REQUIRED SAMPLE SIZE – THE FORMULA

Solving the equations...

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\widehat{p} < -\widetilde{c} \mid H_0\right) + \mathbb{P}\left(\sqrt{n}\widehat{p} > \widetilde{c} \mid H_0\right) &= \alpha \\ \mathbb{P}\left(\sqrt{n}\widehat{p} < -\widetilde{c} \mid H_1\right) + \mathbb{P}\left(\sqrt{n}\widehat{p} > \widetilde{c} \mid H_1\right) &= 1 - \beta\end{aligned}$$

... yields...

$$n = \left(\frac{\Phi^{-1}\left(\beta\right)\sigma_{\hat{p}(1)} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sigma_{\hat{p}(0)}}{p_0 - p}\right)^2,$$

where we have the inverse standard normal CDF (Φ^{-1}), standard deviations ($\sigma_{\hat{p}(0)}, \sigma_{\hat{p}(1)}$) and the assumed effects under the null (p_0) and alternative hypothesis (p).

THE REQUIRED SAMPLE SIZE – VARIANCE COMPONENT

The general (asymptotic) variance is:

$$\mathbb{V}\text{ar}\left(\sqrt{n}\widehat{p}\right) = \frac{2 \cdot (1 + (m-1) \cdot \widetilde{\rho})}{m} \left(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2\right)$$

$$\widetilde{\sigma}_i^2 = \mathbb{V}\text{ar}\left(F_i(X_{i'jk})\right),$$

where $\widetilde{\rho}$ is a transformation of ρ.

THE REQUIRED SAMPLE SIZE – VARIANCE COMPONENT

The general (asymptotic) variance is:

$$\mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p}\right) = \frac{2 \cdot (1 + (m-1) \cdot \widetilde{\rho})}{m}\left(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2\right)$$

$$\widetilde{\sigma}_i^2 = \mathbb{V}\mathrm{ar}\left(F_i(X_{i'jk})\right),$$

where $\widetilde{\rho}$ is a transformation of ρ. But its specific form might differ between null and alternative hypothesis, as the variance under the alternative is unknown:

$$\sigma^2_{\hat{p}(0)} = \mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p} \mid H_0\right) = \frac{2 \cdot (1 + (m-1) \cdot \widetilde{\rho})}{6m}$$

$$\sigma^2_{\hat{p}(1)} = \mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p} \mid H_1\right) = \ ?$$

THE REQUIRED SAMPLE SIZE – VARIANCE COMPONENT

The general (asymptotic) variance is:

$$\mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p}\right) = \frac{2 \cdot (1 + (m-1) \cdot \widetilde{\rho})}{m} \left(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2\right)$$

$$\widetilde{\sigma}_i^2 = \mathbb{V}\mathrm{ar}\left(F_i(X_{i'jk})\right),$$

where $\widetilde{\rho}$ is a transformation of ρ. But its specific form might differ between null and alternative hypothesis, as the variance under the alternative is unknown:

$$\sigma^2_{\hat{p}(0)} = \mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p} \mid H_0\right) = \frac{2 \cdot (1 + (m-1) \cdot \widetilde{\rho})}{6m}$$

$$\sigma^2_{\hat{p}(1)} = \mathbb{V}\mathrm{ar}\left(\sqrt{n}\widehat{p} \mid H_1\right) = \ ?$$

Using the upper boundary derived by Birnbaum and Klose (1957), we approximate the alternative variance (cf. Master's Thesis of C. Abele):

$$\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2 \overset{H_1}{=} w(p)p(1-p) + (1 - w(p))\frac{1}{6}$$

THE REQUIRED SAMPLE SIZE – CORRELATION COMPONENT

The variance involves the term $\widetilde{\rho}$, which we defined as:

$$\widetilde{\rho}_i = \mathbb{C}\text{or}\left(F_i(X_{i'jk}), F_i(X_{i'jk'})\right),$$

and assuming/approximating that $\widetilde{\rho}_1 = \widetilde{\rho}_2 = \widetilde{\rho}$.

THE REQUIRED SAMPLE SIZE – CORRELATION COMPONENT

The variance involves the term $\widetilde{\rho}$, which we defined as:

$$\widetilde{\rho}_i = \mathbb{C}\mathrm{or}\left(F_i(X_{i'jk}), F_i(X_{i'jk'})\right),$$

and assuming/approximating that $\widetilde{\rho}_1 = \widetilde{\rho}_2 = \widetilde{\rho}$.

- Model NP-1: Compute $\widetilde{\rho} = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right)$, i.e. we are assuming that the null hypothesis is true and then compute the correlation using copula theory.

THE REQUIRED SAMPLE SIZE – CORRELATION COMPONENT

The variance involves the term $\widetilde{\rho}$, which we defined as:

$$\widetilde{\rho}_i = \mathbb{C}\mathrm{or}\left(F_i(X_{i'jk}), F_i(X_{i'jk'})\right),$$

and assuming/approximating that $\widetilde{\rho}_1 = \widetilde{\rho}_2 = \widetilde{\rho}$.

- Model NP-1: Compute $\widetilde{\rho} = \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right)$, i.e. we are assuming that the null hypothesis is true and then compute the correlation using copula theory.
- Model NP-2: Replace $\widetilde{\rho}$ with $\mathbb{C}\mathrm{or}\left(\Phi(Y_{ijk}), \Phi(Y_{ijk'})\right)$, where

$$\begin{pmatrix} Y_{ijk} \\ Y_{ijk'} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

  i.e. we approximate any F_i with the CDF of the standard normal distribution

THE REQUIRED SAMPLE SIZE – CORRELATION COMPONENT

The variance involves the term $\tilde{\rho}$, which we defined as:

$$\tilde{\rho}_i = \mathbb{C}\mathrm{or}\left(F_i(X_{i'jk}), F_i(X_{i'jk'})\right),$$

and assuming/approximating that $\tilde{\rho}_1 = \tilde{\rho}_2 = \tilde{\rho}$.

- Model NP-1: Compute $\tilde{\rho} = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right)$, i.e. we are assuming that the null hypothesis is true and then compute the correlation using copula theory.
- Model NP-2: Replace $\tilde{\rho}$ with $\mathbb{C}\mathrm{or}\left(\Phi(Y_{ijk}), \Phi(Y_{ijk'})\right)$, where

$$\begin{pmatrix} Y_{ijk} \\ Y_{ijk'} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

  i.e. we approximate any F_i with the CDF of the standard normal distribution
- Model NP-3: Replace $\tilde{\rho}$ with the value 1, i.e. we are assuming the observations within a cluster are perfectly correlated

SIMULATION PROCEDURE

The idea is to compute $\Delta n = \widehat{n} - n$, where $\widehat{n}$ is estimated and the true n is given as a parameter in the simulation.

Simulation Steps:

1. Provide distribution, n, m, p, ρ, α as parameters
2. Compute required distribution parameters such that $p = \int F_1 \mathrm{d}F_2$ is satisfied
3. Simulate the empirical power with an appropriate test using 10^4 simulations
4. Check whether empirical power is within the interval $(0, 1)$
5. Compute the estimate $\widehat{n}$ using the empirical power $1 - \widehat{\beta}$

SIMULATION SETTINGS

We included...

- cluster sizes $m \in \{1, \ldots, 6\}$
- correlations $\rho \in \{0, 0.1, \ldots, 1\}$
- true relative effects $p \in \{0.6, 0.65, \ldots, 0.85\}$
- Normal (N(0,1) vs. N(θ,1)), Beta (Beta(2,4) vs. Beta(2,θ)), Exponential (Exp(1) vs. Exp(θ)), Poisson Distribution (Pois(2) vs. Pois(θ))

In Addition to the proposed models NP-1 to NP-3, we provide results for:

- Model NP-0: Planning sample size for standard WMW-test (i.e. ignoring clusters)
- Model T-0: Planning sample size for the standard t-test (i.e. ignoring clusters)
- Model T-1: Planning sample size for an adjusted t-test

SIMULATION SETTINGS – II

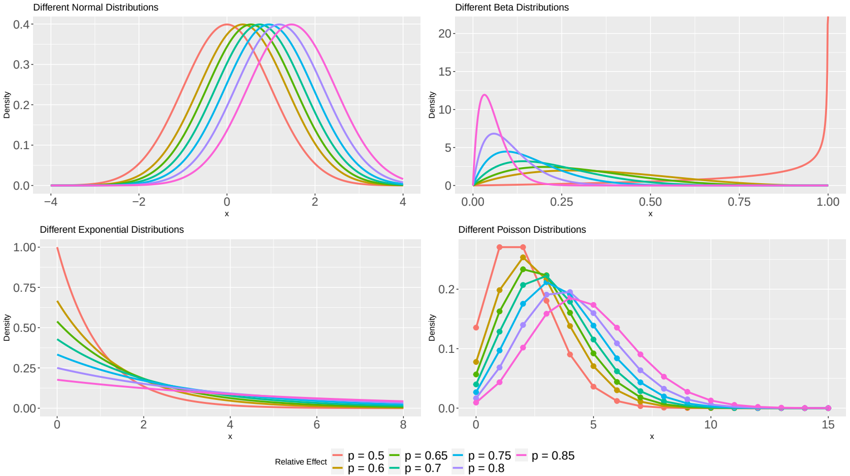


Figure 1: Distributions used in Simulation

RESULTS – I

Table 1: MAE of Models Stratified by Distribution for true $n = 20$

	NP-0	NP-1	NP-2	NP-3	T-0	T-1
Normal	17.56	2.87	2.86	17.56	11.37	1.30
Beta	17.14	2.35	2.36	17.14	41.01	17.20
Exponential	17.45	2.56	2.59	17.45	12.04	13.97
Poisson	19.53	4.04	4.01	19.53	8.19	5.92
Total	17.92	2.95	2.95	17.92	18.04	9.54

Table 2: RMSE of Models Stratified by Distribution for true $n = 20$

	NP-0	NP-1	NP-2	NP-3	T-0	T-1
Normal	25.54	3.41	3.39	25.54	19.65	1.70
Beta	25.20	2.75	2.76	25.20	57.32	21.80
Exponential	25.60	2.96	2.97	25.60	13.11	14.19
Poisson	27.48	4.40	4.41	27.48	13.51	6.06
Total	25.97	3.44	3.44	25.97	31.56	13.32

CONCLUSION

- Models NP-1 and NP-2 are robust methods that can be used by practitioners
- The only parameters to know before the study are:
  - Assumed true effect p
  - Cluster size m
  - Moment correlation of the data ρ
- Contrary to the t-test, our method is not only robust but also requires no knowledge about the variance
- The incorporation of the dependency structure is crucial to obtain good results

(SOME) REFERENCES

- Happ et al.: Optimal sample size planning for the Wilcoxon-Mann-Whitney test, *Statistics in Medicine* (2018)
- Birnbaum and Klose: Bounds for the Variance of the Mann-Whitney Statistic, *Annals of Mathematical Statistics* (1957)
- Donner: Sample Size requirements for stratified cluster randomization designs, *Statistics in Medicine* (1992)
- Jung et al: Sample size calculations for clustered binary data, *Statistics in Medicine* (2001)
- Rosner and Glynn: Power and Sample Size Estimation for the Clustered Wilcoxon Test, *Biometrics* (2011)

Thank You!

RESULTS – II

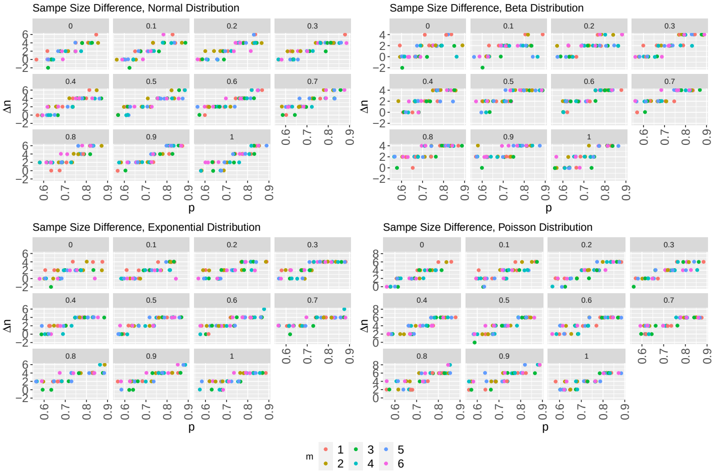


Figure 2: Detailed Results for Model NP-1 and true $n = 20$

RESULTS – III

Figure 3: Detailed Results for Model T-1 and true $n = 20$