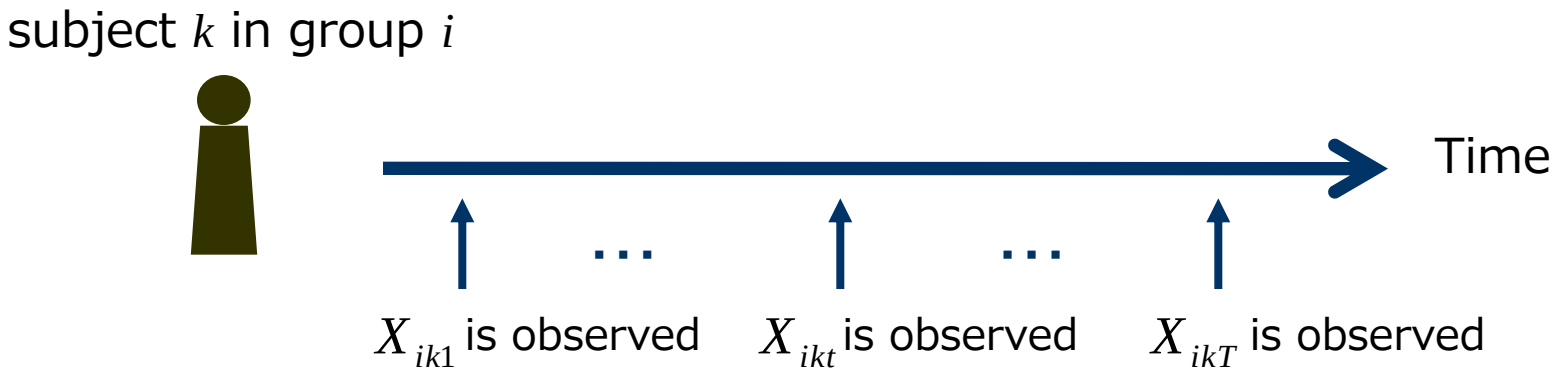
Longitudinal Data Analysis Based on Ranks and Its Performance

Takashi Nagakubo
Asubio Pharma Co. Ltd.
Masashi Goto
Biostatistical Research Association NPO.

Outline

- Introduction
  - Longitudinal Data
- Rank Empirical Distribution Method
  - Relative Effects
  - Hypothesis
  - Estimation of Relative Effects
  - Test Statistic
- Case Study
  - γ-GTP Study (Brunner *et al*., 2002)
- Simulation
- Conclusion

Longitudinal Data



X_{ikt} : response at time t from subject k in group i .

Response X_{ikt}

i : Group $i = 1, \ldots, I$
k : Subject $k = 1, \ldots, n_i$
t : Time $t = 1, \ldots, T$

Longitudinal Data

Responses from a single subject

Group	Subject	Observation: Time 1	Observation: Time ⋯	Observation: Time T	Marginal distribution: Time 1	Marginal distribution: Time ⋯	Marginal distribution: Time T
	$k=1$	X_{111}	⋯	X_{11T}	F_{11}	⋯	F_{1T}
$i=1$	⋮	⋮		⋮	⋮		⋮
	$k=n_1$	X_{1n_11}	⋯	X_{1n_1T}	F_{11}	⋯	F_{1T}
				⋮			
	$k=1$	X_{I11}	⋯	X_{I1T}	F_{I1}	⋯	F_{IT}
$i=I$	⋮	⋮		⋮	⋮		⋮
	$k=n_1$	X_{In_11}	⋯	X_{In_1T}	F_{I1}	⋯	F_{IT}

Longitudinal Data Analysis

- Longitudinal data
  - Individuals are measured repeatedly through time
- Repeated measures ANOVA (Winer *et al*., 1991)
  - Can quantitatively evaluate main effects and interactions of variation factors
  - Interpretation is easy
  - Assumes normality

The assumption of normality is not always satisfied.

"Rank Empirical Distribution Method" (Brunner *et al*., 2002)

This method is based on relative effects determined using distribution functions. The relative effect is estimated by ranks.

- Robust with respect to outliers
- May be utilized for arbitrary data types
- Results are invariant under arbitrary monotone transformations of the data
- Experimental designs with completely at random missing data may be analyzed
- Absence of variability in some trial groups is admitted
- Very accurate approximations for small sample sizes

Relative Effects 1

Nonparametric effects for two samples (Mann & Whitney, 1947)

For independent random variables $Y_1 \sim F_1$ and $Y_2 \sim F_2$.

$$p = \Pr(Y_1 \le Y_2) = \int F_1 dF_2$$

The nonparametric effect is the probability that Y_2 is greater than Y_1.

Extend this effect to longitudinal data (Thompson, 1991).

Relative Effects

Relative effect at time t in group i

$$p_{it} = \int H dF_{it}$$

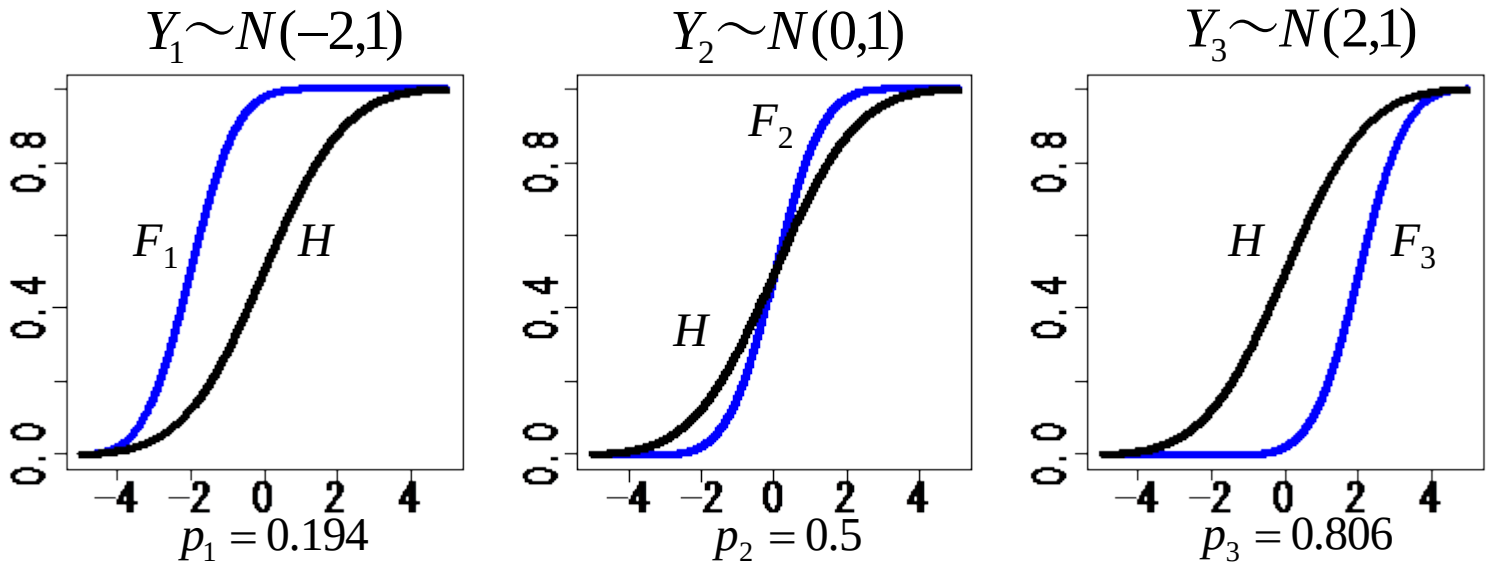
i : Group t : Time
F_{it} : Marginal distribution function at time t in group i
H : Weighted mean of marginal distribution functions

Relative Effects 2

Relative effects

$$p_{it} = \int H dF_{it}$$

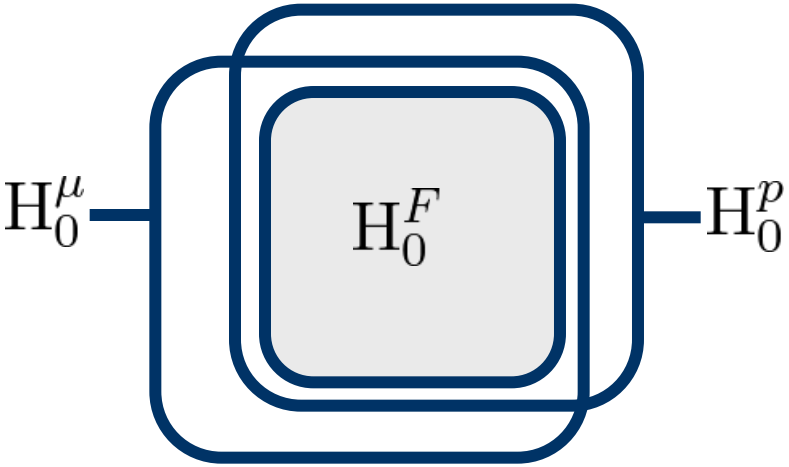
- This effect is the probability that a random variable distributed according to F_{it} is greater than a random variable distributed according to H.
- This effect indicates the tendency of the observations.



Hypothesis 1

Relation between hypothesis on the means, relative effects, and distribution functions

Means	$\mathrm{H}_0^{\mu} : \boldsymbol{C\mu} = \boldsymbol{0}$
Relative effects	$\mathrm{H}_0^{p} : \boldsymbol{Cp} = \boldsymbol{0}$
Distribution functions	$\mathrm{H}_0^{F} : \boldsymbol{CF} = \boldsymbol{0}$



$$\boldsymbol{C\mu} = \boldsymbol{C}\int x d\boldsymbol{F}(x) = \int x d(\boldsymbol{CF}(x))$$

$$\boldsymbol{Cp} = \boldsymbol{C}\int H(x) d\boldsymbol{F}(x) = \int H(x) d(\boldsymbol{CF}(x))$$

Hypothesis 2

Formulation of hypothesis by means of distribution function

Hypothesis of no interaction $\mathrm{H}_0^F(\mathrm{g}\times\mathrm{t}): F_{it} - \overline{F}_{i\cdot} + \overline{F}_{\cdot t} - \overline{F}_{\cdot\cdot}$

Hypothesis of no group effect $\mathrm{H}_0^F(\mathrm{g}): \overline{F}_{1\cdot} = \cdots = \overline{F}_{I\cdot}$

Hypothesis of no time effect $\mathrm{H}_0^F(\mathrm{t}): \overline{F}_{\cdot 1} = \cdots = \overline{F}_{\cdot T}$

Using matrix notation

Vector of the distribution function

$$\boldsymbol{F} = (F_{11}, \ldots, F_{1T}, \ldots, F_{I1}, \ldots, F_{IT})^{\mathrm{T}}$$

Centering matrix

$$\boldsymbol{P}_a = \boldsymbol{I}_a - \frac{1}{a}\boldsymbol{J}_a$$

$\boldsymbol{I}_a$: Identity matrix

$\boldsymbol{J}_a$: Matrix of 1

Hypothesis

Interaction

$$\mathrm{H}_0^F(\mathrm{g}\times\mathrm{t}): \boldsymbol{C}_{\mathrm{g}\times\mathrm{t}}\boldsymbol{F} = \boldsymbol{0}$$

$$\boldsymbol{C}_{\mathrm{g}\times\mathrm{t}} = \boldsymbol{P}_I \otimes \boldsymbol{P}_T$$

Group effect

$$\mathrm{H}_0^F(\mathrm{g}): \boldsymbol{C}_{\mathrm{g}}\boldsymbol{F} = \boldsymbol{0}$$

$$\boldsymbol{C}_{\mathrm{g}} = \boldsymbol{P}_I \otimes \frac{1}{T}\boldsymbol{1}_T^{\mathrm{T}}$$

Time effect

$$\mathrm{H}_0^F(\mathrm{t}): \boldsymbol{C}_{\mathrm{t}}\boldsymbol{F} = \boldsymbol{0}$$

$$\boldsymbol{C}_{\mathrm{t}} = \frac{1}{I}\boldsymbol{1}_I^{\mathrm{T}} \otimes \boldsymbol{P}_T$$

Estimation of Relative Effects 1

$$p_{it} = \int H dF_{it} \Rightarrow \hat{p}_{it} = \int \hat{H} d\hat{F}_{it} = \frac{1}{n_i} \sum_{k=1}^{n_i} \hat{H}(X_{ikt})$$

Relative effects are estimated by replacing the distribution functions with the corresponding empirical distribution functions.

Empirical distribution function and its weighted mean

$$\hat{F}_{it}(x) = \frac{1}{n_i} \sum_{k=1}^{n} c(x - X_{ikt})$$

$$\hat{H}(x) = \frac{1}{N} \sum_{i=1}^{I} \sum_{t=1}^{T} n_i \hat{F}_{it}(x) = \frac{1}{N} \sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{k=1}^{n} c(x - X_{kt})$$

Empirical distribution functions are defined using the counting function

$$c^{-}(x) = \begin{cases} 0, x \le 0 \\ 1, x > 0 \end{cases}, \; c^{+}(x) = \begin{cases} 0, x < 0 \\ 1, x \ge 0 \end{cases}, \; c(x) = \frac{1}{2}[c^{-}(x) + c^{+}(x)]$$

Estimation of Relative Effects 2

Ranks of observations

$$R_{ikt} = \frac{1}{2} + \sum_{j=1}^{I} \sum_{l=1}^{n_i} \sum_{u=1}^{T} c(X_{ikt} - X_{jlu})$$

Ranks are also defined using the counting function

Ranks and empirical distribution functions

$$R_{ikt} = \frac{1}{2} + N\hat{H}(X_{ikt})$$

There is a relationship between the ranks and the means of the empirical distribution functions

Estimators for relative effects

$$\hat{p}_{it} = \int \hat{H} d\hat{F}_{it} = \frac{1}{n_i} \sum_{k=1}^{n_i} \hat{H}(X_{ikt}) = \frac{1}{N}(\bar{R}_{i\cdot t} - \frac{1}{2})$$

Relative effects are estimated using ranks

Asymptotic Distributions of the Estimators

Vector of the relative effects $\hat{\boldsymbol{p}} = (\hat{p}_{11}, \ldots, \hat{p}_{1T}, \ldots, \hat{p}_{I1}, \ldots, \hat{p}_{IT})^{\mathrm{T}}$

Vector of the rank means $\overline{\boldsymbol{R}}_{\cdot} = (R_{1\cdot 1}, \ldots R_{1\cdot T}, \ldots, R_{I\cdot 1}, \ldots R_{I\cdot T})^{\mathrm{T}}$

$$\min n_i \to \infty \quad \mathrm{H}_0^F : \boldsymbol{CF} = \boldsymbol{0}$$

Under the assumption

$$\sqrt{n}\boldsymbol{C}\hat{\boldsymbol{p}} = \sqrt{n}\boldsymbol{C}\frac{1}{N}\overline{\boldsymbol{R}}_{\cdot} = \sqrt{n}\boldsymbol{C}\frac{1}{N}(\overline{R}_{1\cdot 1}, \ldots, \overline{R}_{I\cdot T})^{\mathrm{T}}$$

This statistic has asymptotically multivariate normal distribution with expectation vector 0 and covariance matrix $\boldsymbol{C}\boldsymbol{V}_n\boldsymbol{C}^{\mathrm{T}}$.

Estimation of covariance matrix

$$\hat{\boldsymbol{V}}_n = \oplus_{i=1}^{I} \hat{\boldsymbol{V}}_i, \qquad \hat{\boldsymbol{V}}_i = \frac{1}{N^2(n_i - 1)} \sum_{k=1}^{n_i} (\boldsymbol{R}_{ik} - \overline{\boldsymbol{R}}_{i\cdot})(\boldsymbol{R}_{ik} - \overline{\boldsymbol{R}}_{i\cdot})^{\mathrm{T}}$$

Test Statistics

Quadratic form of $\sqrt{n}C\hat{\boldsymbol{p}}$

$$Q_n = n\hat{\boldsymbol{p}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}[\boldsymbol{C}\hat{\boldsymbol{V}}_n\boldsymbol{C}^{\mathrm{T}}]^{-}\boldsymbol{C}\hat{\boldsymbol{p}}$$

Wald type statistic

The statistic has asymptotically a central χ^2_f distribution with $f = \mathrm{rank}(C)$.

If the covariance matrix is singular, the use of this statistic is not recommended (Brunner *et al.,* 2002).

ANOVA type statistic

$$F_n = \frac{n\hat{\boldsymbol{p}}^{\mathrm{T}}\boldsymbol{T}\hat{\boldsymbol{p}}}{\mathrm{tr}(\boldsymbol{T}\boldsymbol{V}_n)}$$

$$\hat{f} = \frac{[\mathrm{tr}(\boldsymbol{T}\hat{\boldsymbol{V}}_n)]^2}{\mathrm{tr}(\boldsymbol{T}\hat{\boldsymbol{V}}_n\boldsymbol{T}\hat{\boldsymbol{V}}_n)} \qquad \boldsymbol{T} = \boldsymbol{C}^{\mathrm{T}}[\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}]^{-}\boldsymbol{C}$$

distributed according to F with degrees of freedom (f, ∞)

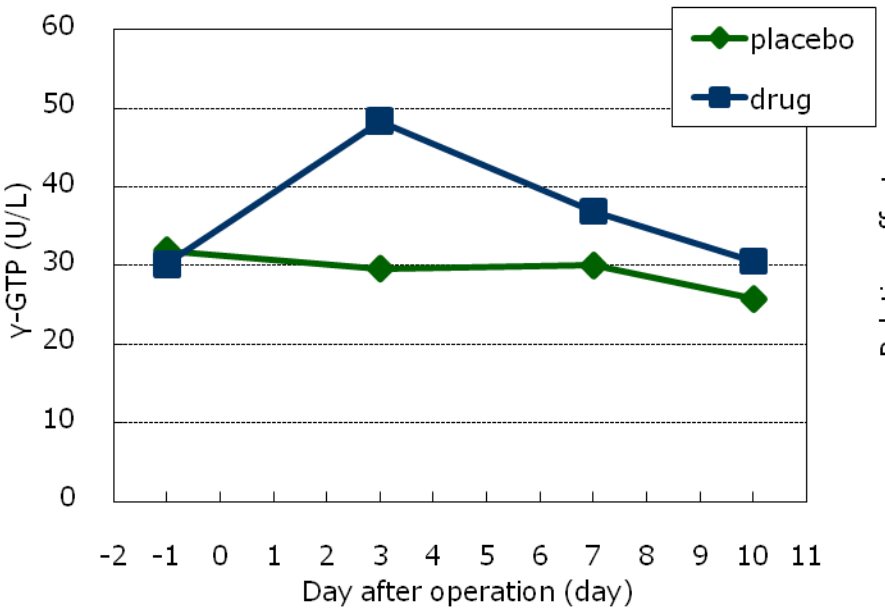
Case Study

γ-GTP Study (Brunner *et al.*, 2002)

- For 50 patients whose gall bladders had to be removed due to cholelithiasis, γ-GTP levels were investigated.
- 26 patients were treated with a specific drug; 24 patients received a placebo.
- The γ-GTP level of each patient was measured before the operation and on days 3, 7, and 10 after the operation.
- The efficacy of the drug is represented by the existence of an interaction between treatment group and time.

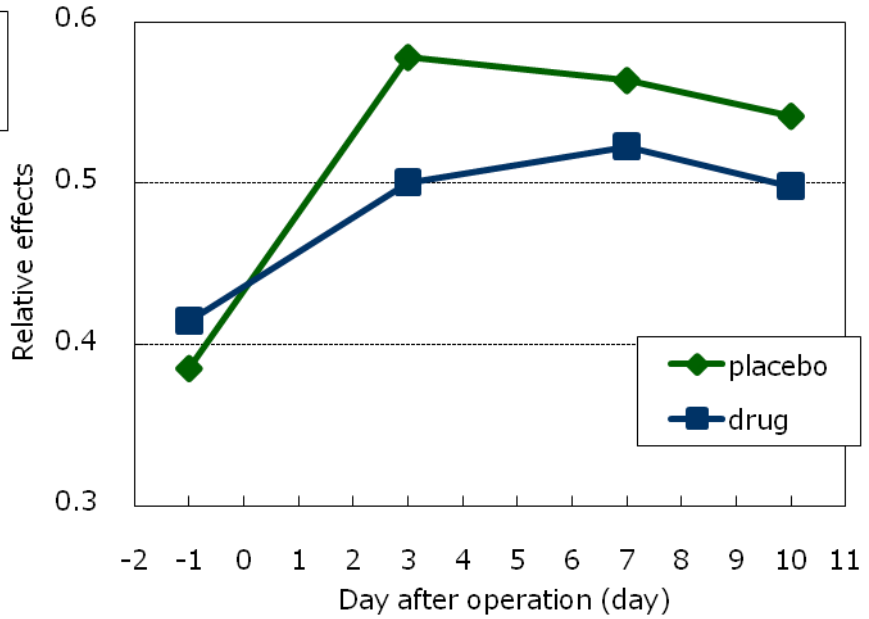
Case Study: Results

Repeated Measures ANOVA



	F	p-value
Group	0.670	0.4173
Time	1.477	0.2234
Interaction	1.280	0.2836

Rank Empirical Distribution Method



	F	p-value
Group	0.227	0.6340
Time	8.471	0.0004
Interaction	0.930	0.3845

Case Study: Conclusion

- Since the interaction is not significant, γ-GTP does not necessarily decrease earlier in the drug group.
- Group differences are not significant, either.
- The time effect is not significant using RM-ANOVA, but is significant using the RED method.
- Thus, different results were obtained depending on the RED method or RM-ANOVA used.

Simulation

Comparison between power of the rank empirical distribution method and repeated measures ANOVA

Model $X_{ikt} = \mu_i + \tau_t + (\mu\tau)_{it} + e_{ikt}$

Mean structure

$$\boldsymbol{\mu} = (0, 1)^{\mathrm{T}}$$

$$\boldsymbol{\tau} = (0, 1, 2, 3)^{\mathrm{T}}$$

$$\boldsymbol{\mu\tau} = \begin{pmatrix} 0 & -1 & -2 & -3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Covariance structure

$$\boldsymbol{V}_{ik} = (v_{rc}), (r, c = 1, \ldots, T)$$

$$v_{rc} = \sigma_e^2 \rho^{|r-c|}$$

Group, Time : $I = 2, T = 4$ Sample size : $n_i = 20, 30, 40$

Error variance : $\sigma_e^2 = 16, 25, 36$ Correlation coefficient : $\rho = 0.4, 0.6, 0.8$

Distribution : Normal, Power normal $\lambda = 0, 0.5, 1$

Number of simulations: 10,000

Simulation Results: Group Effects

Group effects $n_i = 30$

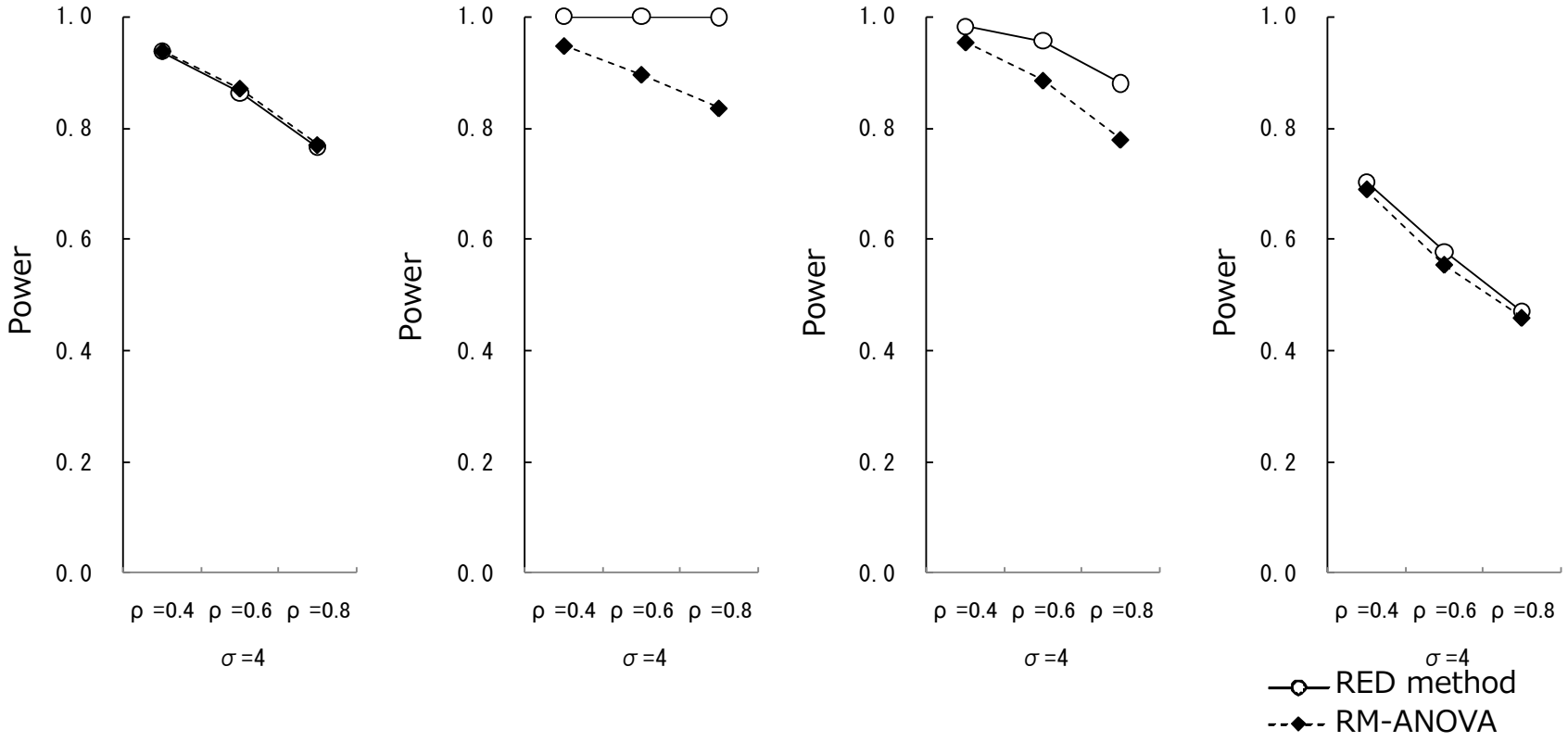
Power normal distribution

Normal distribution

$\lambda = 0$

$\lambda = 0.5$

$\lambda = 1$



Simulation Results: Group Effects

ANOVA table

Factor	df	F-value	p-value	Contribution rate(%)
Method	1	415.1	near 0	3.89
Latent distribution	3	1816	near 0	51.20
Sample size	2	726.4	near 0	13.64
Correlation coefficient	2	278.9	near 0	5.23
Error variance	2	894.3	near 0	16.80
Method * Latent distribution	3	158.7	near 0	4.45
Method * Sample size	2	5.32	0.01	0.08
Method * Correlation coefficient	2	3.00	0.05	0.04
Method * Error variance	2	16.92	near 0	0.30
Latent distribution * Sample size	6	18.48	near 0	0.99
Latent distribution * Correlation coefficient	6	5.27	near 0	0.24
Latent distribution * Error variance	6	22.34	near 0	1.20
Sample size * Correlation coefficient	4	0.10	0.98	near 0
Sample size * Error variance	4	0.74	0.57	near 0
Correlation coefficient * Error variance	4	0.27	0.90	near 0

The factor with the highest contribution rate was 51.20% in a latent distribution.

Also, the contribution rate of error variance and the sample size was high.

The contribution rate for the interaction between the method and latent distribution was 4.45%.

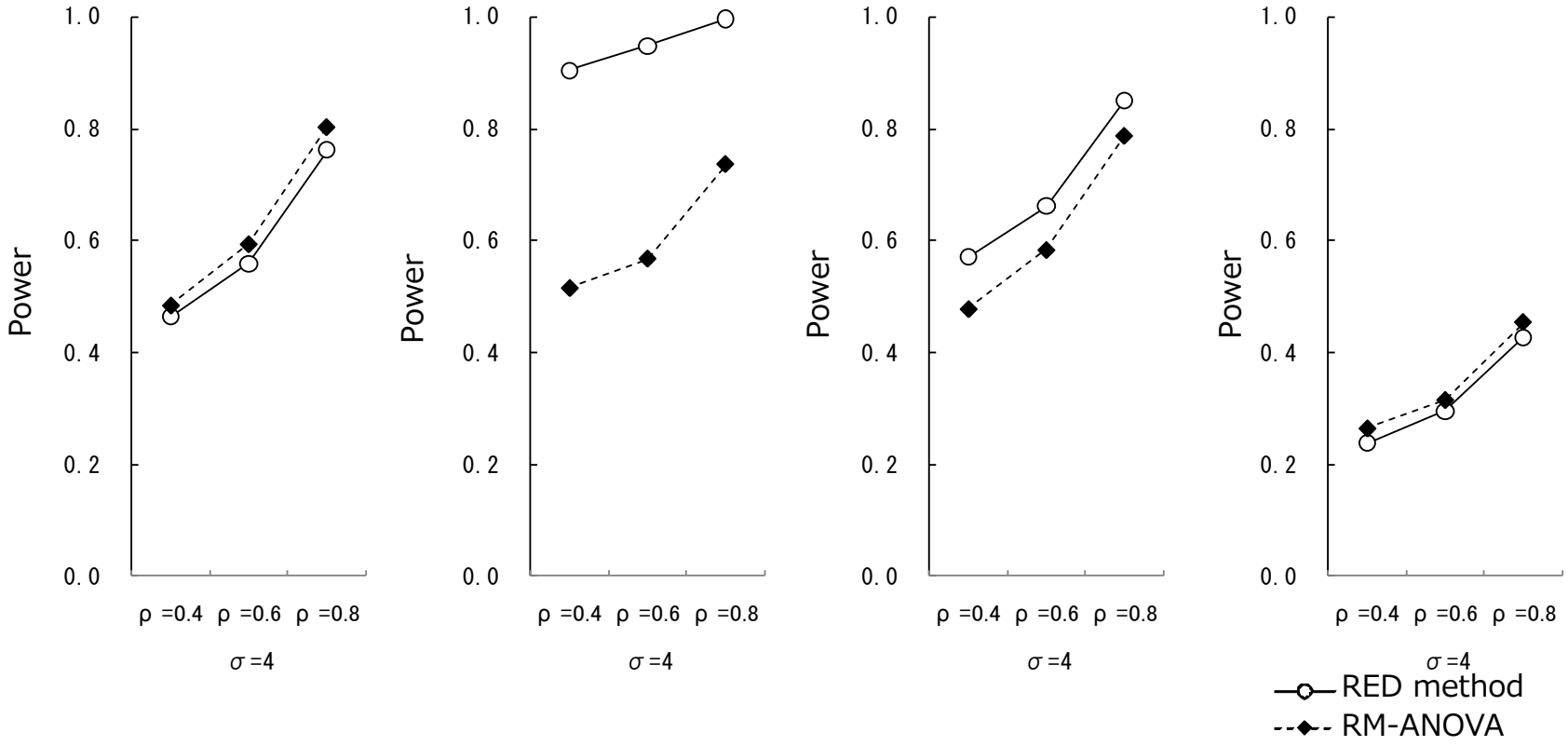
It was suggested that the power of the rank empirical distribution method was different from the power of repeated measures ANOVA.

Simulation Results: Time Effects

Time effects $n_i = 30$

Power normal distribution

Normal distribution $\lambda = 0$ $\lambda = 0.5$ $\lambda = 1$



Simulation Results: Time Effects

ANOVA table

Factor	df	F-value	p-value	Contribution rate(%)
Method	1	199.6	near 0	3.47
Latent distribution	3	595.2	near 0	31.08
Sample size	2	328.8	near 0	11.44
Correlation coefficient	2	733.2	near 0	25.53
Error variance	2	392.9	near 0	13.67
Method * Latent distribution	3	169.4	near 0	8.83
Method * Sample size	2	1.54	0.22	0.04
Method * Correlation coefficient	2	8.86	near 0	0.29
Method * Error variance	2	7.44	near 0	0.24
Latent distribution * Sample size	6	3.71	near 0	0.34
Latent distribution * Correlation coefficient	6	12.36	near 0	1.25
Latent distribution * Error variance	6	4.95	near 0	0.47
Sample size * Correlation coefficient	4	0.69	0.60	0.02
Sample size * Error variance	4	0.22	0.93	near 0
Correlation coefficient * Error variance	4	1.49	0.21	0.07

The factor with the highest contribution rate was 31.08% in a latent distribution.

The contribution rate of the correlation coefficient was higher than in the case of the group effect.

The contribution rate for the interaction between the method and latent distribution was 8.83%.

It was suggested that the power of the rank empirical distribution method was different from the power of repeated measures ANOVA.

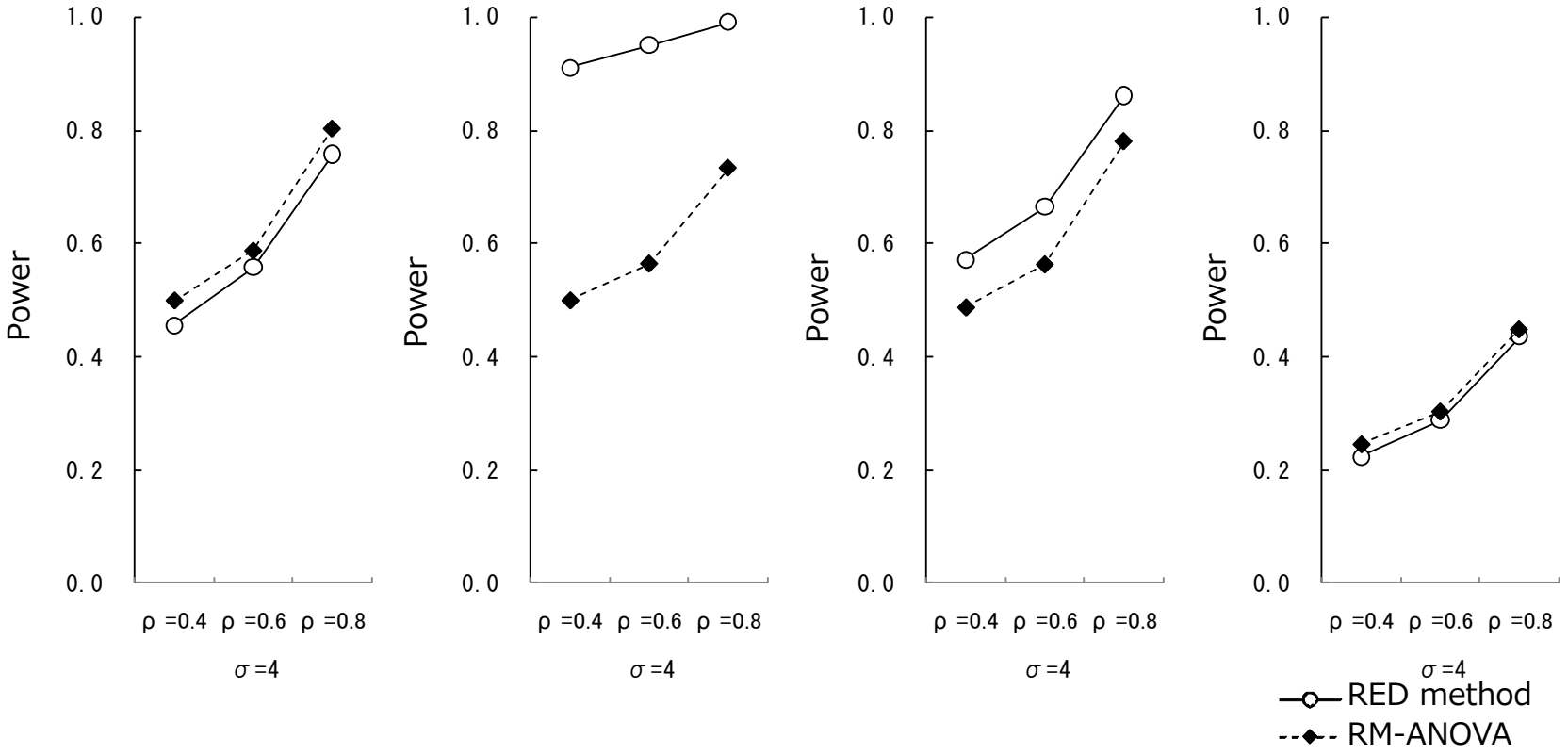
Simulation Results: Interaction

Interaction $n_i = 30$

Normal distribution

Power normal distribution

$\lambda = 0$ $\lambda = 0.5$ $\lambda = 1$



Simulation Results: Interaction

ANOVA table

Factor	df	F-value	p-value	Contribution rate(%)
Method	1	206.7	near 0	3.54
Latent distribution	3	605.4	near 0	31.12
Sample size	2	330.1	near 0	11.30
Correlation coefficient	2	745.5	near 0	25.55
Error variance	2	398.2	near 0	13.64
Method * Latent distribution	3	173.2	near 0	8.88
Method * Sample size	2	1.48	0.23	0.04
Method * Correlation coefficient	2	9.59	near 0	0.31
Method * Error variance	2	7.18	near 0	0.23
Latent distribution * Sample size	6	3.78	near 0	0.34
Latent distribution * Correlation coefficient	6	12.75	near 0	1.27
Latent distribution * Error variance	6	5.10	near 0	0.48
Sample size * Correlation coefficient	4	0.73	0.57	0.02
Sample size * Error variance	4	0.24	0.92	near 0
Correlation coefficient * Error variance	4	1.58	0.18	0.08

Similar to the case of the time effect.

Conclusion

- We explained the rank empirical distribution method based on relative effects.
- The RED method and RM-ANOVA were applied to a case study with differing results.
- We next conducted a simulation study.
- The simulation indicated that the power of both methods is almost the same for normally distributed data.
- The power of the RED method is higher than that of RM-ANOVA for skewed distributed data.

References

- Brunner, E., Domhof, S. and Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. John Wiley & Sons.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, **18**, 50-60.
- Nagakubo, T. and Goto, M. (2009). Longitudinal data analysis based on ranks (in Japanese). Bulletin of the Computational Statistics of Japan, **22**, 109-129.
- Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, **33**, 410-419.
- Winer, B., Brown, D. and Michels, K. (1991). *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill.

Thank you for your kind attention