Statistical Inference Under Subjective Ranking of Experimental Units

Omer Ozturk

The Ohio State University

Department of Statistics

June 23, 2011

Outline

Outline

1. Experimental unit information: Subjective, objective
2. Order Restricted Randomized Designs
3. Two-Sample Model
4. Linear models
5. Examples
6. Empirical Power
7. Concluding Remark

Outline

1. Experimental unit information: Subjective, objective
2. Order Restricted Randomized Designs
3. Two-Sample Model
4. Linear models
5. Examples
6. Empirical Power
7. Concluding Remark

Outline

Outline

1. Experimental unit information: Subjective, objective
2. Order Restricted Randomized Designs
3. Two-Sample Model
4. Linear models
5. Examples
6. Empirical Power
7. Concluding Remark

Outline

Outline

Subjective versus objective information

- In many scientific investigations, the investigators have a wealth of information on potential experimental units. This information can be categorized into two classes
- I: Well-defined, numerically quantifiable information: We use model based approach to analyze this type of information.
- II: Not-well defined, subjective, incomplete information: It is not clear how we use this kind of information in statistical analysis.
- One of the technique used in the analysis of subjective information is called Ranked Set Sampling, introduced by McIntyre (1954) and republished in (2005).

Standard Ranked Set Sampling

- Select k units at random from a specified population.
- Rank these k units with some expert judgment without measuring them.
- Retain the smallest judged unit and return the others.
- Select the second *k* units and retain the second smallest unit judged.
- Continue to the process until k ordered units are measured.
- Note: These *k* ordered observations $X_{(1)i}, ..., X_{(k)i}$ are called a cycle.
- Note: Process repeated $i = 1, \cdots, n$ cycle to get nk observations. These *nk* observations are called a standard ranked set sample.

Diagram

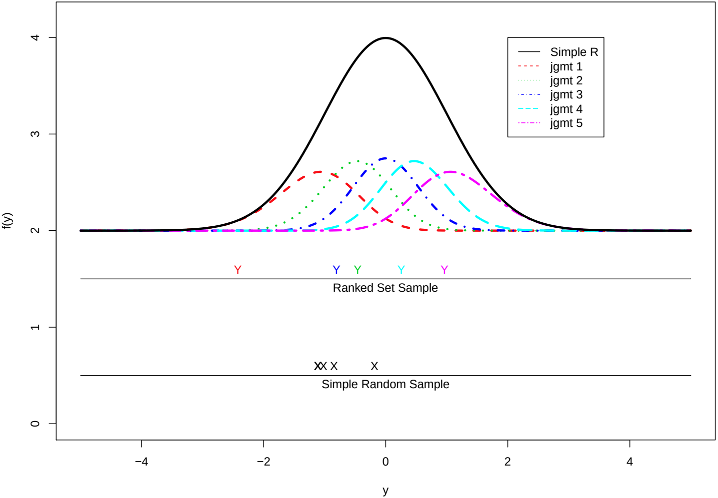
Let $k = 4$ and $n = 2$.

Cycle	Judgment Rank 1	2	3	4
1	$X_{[1]1}$	$X_{[2]1}$	$X_{[3]1}$	$X_{[4]1}$
	$X_{[1]1}$	$X_{[2]1}$	$X_{[3]1}$	$X_{[4]1}$
	$X_{[1]1}$	$X_{[2]1}$	$X_{[3]1}$	$X_{[4]1}$
	$X_{[1]1}$	$X_{[2]1}$	$X_{[3]1}$	$X_{[4]1}$
2	$X_{[1]2}$	$X_{[2]2}$	$X_{[3]2}$	$X_{[4]2}$
	$X_{[1]2}$	$X_{[2]2}$	$X_{[3]2}$	$X_{[4]2}$
	$X_{[1]2}$	$X_{[2]2}$	$X_{[3]2}$	$X_{[4]2}$
	$X_{[1]2}$	$X_{[2]2}$	$X_{[3]2}$	$X_{[4]2}$

$X_{[1]1}, \cdots, X_{[4]2}$ is called a ranked set sample.

- For each fully measured unit, we need $k - 1$ additional units for ranking.
- Measured units are all independent.
- Under a stable ranking condition, observations from the same judgment class are identically distributed.

Why ranked-set sampling?



Why Ranked-set Sampling?

- Consider a ranked-set sample of size k. For simplicity assume $n = 1$.
- Let $Y_1, \cdots, Y_k$ and $Y_{(1)1}, \cdots, Y_{(k)1}$ denote the simple random and ranked-set samples, respectively. Then

$$
\begin{aligned}
var_{SRS}(\bar{Y}) &= \frac{1}{k^2} var(\sum_{j=1}^{k} Y_{(j)}) = \frac{1}{k^2} \sum_{j=1}^{k} \sigma^2_{(j)} + \frac{2}{k^2} \sum \sum_{i<j} \sigma_{ij} \\
&= var_{RSS}(\bar{Y}) + \frac{2}{k^2} \sum \sum_{i<j} \sigma_{ij} \\
var_{SRS}(\bar{Y}) &\geq var_{RSS}(\bar{Y}).
\end{aligned}
$$

- Variance of a ranked-set sample mean is always less than the variance of a simple random sample mean.
- If the judgment ranking is at random (the worst case scenario), then $var_{SRS}(\bar{Y}) = var_{RSS}(\bar{Y})$.

Problems and Concerns in Ranked Set Sampling

- Waste of experimental units.
- Role of randomizations
- Analysis
- Appropriate statistical inference
- Availability of appropriate software

A Review of Completely Randomized Design

Assume that we have a two-sample problem

- Control (C) and treatment (T) regimes.
- N available experimental units, $N = 12$
- Objective is to make inference for $\Delta = \mu_C - \mu_T$.
- Randomization: Randomly assign T and C to these N experimental units.

C	T	T	C
T	C	C	T
C	C	T	T

Order Restricted Randomization

- Assume that again we have N available units and let $h = 2$ be the set size.
- Divide N units randomly to $N/2$ sets and rank the units within each set.

pairs	1		2		3	
sets	1	2	1	2	1	2

- Group these $N/2$ sets in pairs

pairs	1		2		3	
sets	1	2	1	2	1	2
	R_1	R_1	R_2	R_1	R_1	R_1
	R_2	R_2	R_1	R_2	R_2	R_2

- Within each pair (replication) randomly assign C and T to ranked units in the first set.
- Do the opposite assignment in the second set.

1		2		3	
1	2	1	2	1	2
R_1 C	R_1 T	R_2 T	R_1 T	R_1 C	R_1 T
R_2 T	R_2 C	R_1 C	R_2 C	R_2 T	R_2 C

Illustration of ORR design

1		2	
1	2	1	2
R_1 $Y_{1[1]11}$	R_1 $Y_{2[1]12}$	R_2 $Y_{2[2]12}$	R_1 $Y_{2[1]22}$
R_2 $Y_{2[2]11}$	R_2 $Y_{1[2]21}$	R_1 $Y_{1[1]12}$	R_2 $Y_{1[2]22}$

where $Y_{i[j]rt}$ is the response measurement from the i-th treatment, j-th judgment ranked unit r-th set and the t-th replication. Some naive estimators for $\Delta = \mu_C - \mu_T$

- ORR design, $\hat{\Delta}_{ORR} = \bar{Y}_{1[.]..} - \bar{Y}_{2[.]..}$.
- RSS design, $\hat{\Delta}_{RSS} = \bar{Y}_{1[.]..} - \bar{Y}_{2[.]..}$
- SRS design, $\hat{\Delta}_{SRS} = \bar{Y}_{1.} - \bar{Y}_{2.}$

Some observations

- All three estimators are unbiased

$$E\hat{\Delta}_{RSS} = E\hat{\Delta}_{ORR} = \hat{\Delta}_{SRS} = \mu_C - \mu_T.$$

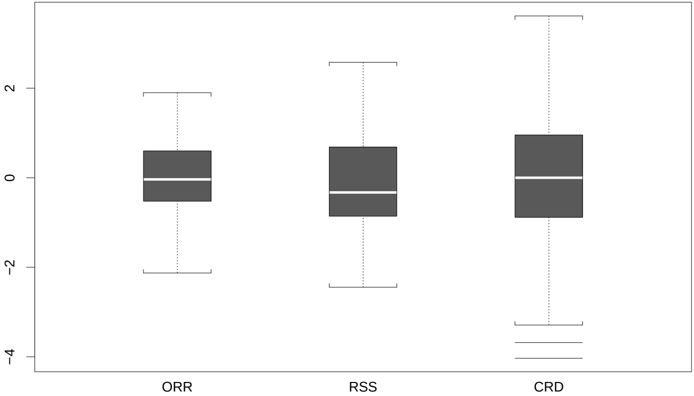
- Variances of these estimators are:

$$\begin{aligned} V(\hat{\Delta}_{SRS}) &= \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} = 2\sigma^2/n \\ &= (\sigma_{[11]} + \sigma_{[22]} + 2\sigma_{[12]})/n \\ V(\hat{\Delta}_{RSS}) &= (\sigma_{[11]} + \sigma_{[22]})/n, \\ V(\hat{\Delta}_{ORR}) &= (\sigma_{[11]} + \sigma_{[22]} - 2\sigma_{[12]})/n. \end{aligned}$$

- Variance reductions:

$$V(\hat{\Delta}_{SRS}) - V(\hat{\Delta}_{RSS}) = 2\sigma_{[1,2]}/n$$
$$V(\hat{\Delta}_{RSS}) - V(\hat{\Delta}_{ORR}) = 2\sigma_{[1,2]}/n$$

Box plots for the ORR and CRD estimates of Δ



Large Set Sizes $h > 2$

- For simplicity, we again consider two treatment combinations, C and T. Let set size $h \geq 2$.
- Select two sets each of size h and rank the units within each set. For example, for $h = 4$

Set I	R_1	R_2	R_3	R_4
Set II	R_1	R_2	R_3	R_4

- Divide ranks within each set into two disjoint sets.

$$\boldsymbol{a} = \{a_1, \cdots, a_k\}, \quad \boldsymbol{b} = \{b_1, \cdots, b_{h-k}\},$$

where k is the largest integer less than or equal to $h/2$.

- Perform randomization to assign control and treatment regimes to the units in ***a*** or ***b*** in set 1.
- Do an opposite assignment in the second set.

Control			Treatment		
a_1	$\cdots$	a_k	b_1	$\cdots$	b_{h-k}
$Y_{1[a_1]11}$	$\cdots$	$Y_{1[a_k]11}$	$Y_{2[b_1]11}$	$\cdots$	$Y_{2[b_{h-k}]11}$
b_1	$\cdots$	b_{h-k}	a_1	$\cdots$	a_k
$Y_{1[b_1]21}$	$\cdots$	$Y_{1[b_{h-k}]21}$	$Y_{2[a_1]21}$	$\cdots$	$Y_{2[a_k]21}$

Example: h=4, $\boldsymbol{a} = \{R_1, R_3\}$, $\boldsymbol{b} = \{R_2, R_4\}$.

Set I	R_1, C	R_2, T	R_3, C	R_4, T
Set II	R_1, T	R_2, C	R_3, T	R_4, C

Main feature of this design can be summarized as follows

- While responses between sets are all independent, responses within each set are positively correlated.
- All h ranks are used within each treatment regime so that the design is balanced.
- This design is unique for $h = 2$, but the number of possible designs increases with h. For a general h, $2^{h-1} - 1$ designs are possible.
- This design is motivated for the inference of the contrast parameter $\Delta = \mu_C - \mu_T$.
- If $h > 2$, then the design $\boldsymbol{a} = \{1, 3, 5, \cdots, h-2, h\}$ minimizes the variance of $\hat{\Delta}(ORR)$.

Variance of $\hat{\Delta}_{ORR}$ for normal distribution

h	$\boldsymbol{a}$	ORR	$SRS - RSS$	$RSS - ORR$
2	$\{1\}$	0.727	0.637	0.637
3	$\{1\}$	0.825	0.954	0.219
	$\{2\}$	0.530	0.954	0.515^*
4	$\{1,2\}$	0.687	1.178	0.135
	$\{1,3\}$	0.336	1.178	0.485^*
	$\{1,4\}$	0.386	1.178	0.437
5	$\{1,2\}$	0.732	1.278	-0.010
	$\{1,3\}$	0.449	1.278	0.273
	$\{1,4\}$	0.353	1.278	0.369
	$\{1,5\}$	0.470	1.278	0.252
	$\{2,3\}$	0.424	1.278	0.298
	$\{2,4\}$	0.277	1.278	0.445^*
	$\{2,5\}$	0.353	1.278	0.369
	$\{3,4\}$	0.424	1.278	0.298
	$\{3,5\}$	0.449	1.278	0.273

Inference Based on Means: Two-Sample models, Ozturk and MacEachern (2005,EES)

- Consider testing $H_0 : \Delta = 0$ against $H_A : \Delta \neq 0$.
- A natural test statistic for this hypothesis is

$$T_n = \frac{\bar{Y}_{1[.]..} - \bar{Y}_{2[.]..}}{\sqrt{var(\hat{\Delta}_{ORR})}}.$$

- For large n, T_n has normal distribution.
- For small sample sizes, Student's t-distribution provides a reasonable approximation to type I error.
- By using Student's t-distribution approximation, we can construct $(1-\gamma)100\%$ confidence interval for Δ

$$\bar{Y}_{1[.]..} - \bar{Y}_{2[.]..} \pm t_{2n-2}^{1-\gamma/2} \sqrt{var(\hat{\Delta}_{ORR})}$$

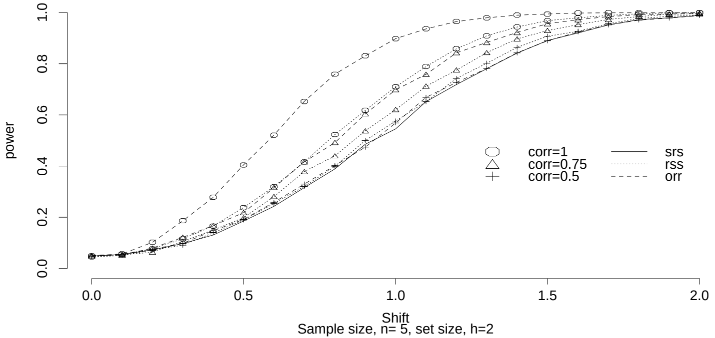
Estimated Type I error rates of T_N

Model	ρ	n	$P_N = 0.05$	$P_t = 0.05$
Normal	1.0	3	0.124	0.051
		5	0.081	0.052
		10	0.066	0.049
		30	0.059	0.054
Normal	0.5	3	0.127	0.054
		5	0.086	0.052
		10	0.064	0.050
		30	0.056	0.052
t_5	1.0	3	0.116	0.047
		5	0.083	0.045
		10	0.060	0.042
		30	0.053	0.048
t_5	$\rho = 0.5$	3	0.116	0.044
		5	0.091	0.050
		10	0.065	0.050
		30	0.054	0.050

Estimated type I error rates of T_N

Model	ρ	n	$P_N = 0.05$	$P_t = 0.05$
t_3	1.0	3	0.099	0.036
		5	0.070	0.036
		10	0.055	0.040
		30	0.051	0.046
t_3	0.5	3	0.105	0.040
		5	0.073	0.041
		10	0.061	0.047
		30	0.049	0.045
L. Nor	1.0	3	0.086	0.033
		5	0.061	0.030
		10	0.049	0.035
		30	0.047	0.043
L. Nor	0.5	3	0.088	0.030
		5	0.062	0.029
		10	0.047	0.030
		30	0.055	0.050

Empirical powers of two sample t-test of ORR design



Inference Based on Means: Linear models

- We express the response vector in a linear model structure

$$\boldsymbol{Y}_i = \boldsymbol{C}_i\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}_i, \text{ for } i = 1, \cdots, n, \tag{1}$$

  where $\boldsymbol{C}_i$ is the $2h \times p$ dimensional design matrix, $\boldsymbol{\beta}$ is the parameter vector and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}'_{i1}, \boldsymbol{\epsilon}'_{i2})'$ is the residual vector. Note that each replication contains responses from two different sets, each of size h.
- Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)$. Then the model 1 can be written as

$$\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon} = \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}, \boldsymbol{\mu} \in V, \tag{2}$$

  where V is a p-dimensional subspace of R^N, $N = 2hn$

- The model 2 can be rewritten to highlight differences between the parameters that are the property of the model and the ranking mechanism

$$\boldsymbol{Y} = \boldsymbol{C}_1\beta_1 + \boldsymbol{C}_2\beta_2 + \sigma\epsilon, \tag{3}$$

where $\boldsymbol{C}_2$ is the design matrix that could account the differences in means of ranking classes and interaction effects between the ranking classes and other factors in the experiment.

- Under some mild assumptions, we can write

$$E\boldsymbol{Y} = \mu, \quad var(\boldsymbol{Y}) = \boldsymbol{\Gamma},$$

where $\boldsymbol{\Gamma}$ is a $2hn \times 2hn$ block-diagonal matrix, each block of which contains $\boldsymbol{\Sigma}$ that corresponds to $var(\epsilon_{i1})$.

Generalized Linear Model

- We transform the model in the following fashion.

$$\boldsymbol{Y}^* = \boldsymbol{\mu}^* + \sigma\boldsymbol{\epsilon}^*,$$

where

$$\boldsymbol{Y}^* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Y}, \quad \boldsymbol{\mu}^* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\mu}, \quad \boldsymbol{\epsilon}^* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\epsilon}. \tag{4}$$

- Let

$$V^* = \left\{\boldsymbol{v}^* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{v}, \boldsymbol{v} \in V\right\}.$$

Since $\boldsymbol{\Gamma}$ is an invertible matrix and V is a p-dimensional subspace V^* is a p-dimensional subspace. Under the transformation 4

$$E\boldsymbol{Y}^* = \boldsymbol{\mu}^*, \quad \boldsymbol{\mu}^* \in V^*, \qquad cov(\boldsymbol{Y}^*) = \sigma^2\boldsymbol{I}.$$

Least Square Estimator

- We estimate μ^* with least square estimators from the transformed model 4 which follows from the projection of $\boldsymbol{Y}^*$ onto the space $\boldsymbol{V}^*$,

$$\hat{\mu}^* = \boldsymbol{P}_{V^*}\boldsymbol{Y}^*,$$

  where $\boldsymbol{P}_{V^*}\boldsymbol{Y}^*$ is the projection of $\boldsymbol{Y}^*$ onto V^*.
- Let $\boldsymbol{C}$ be a basis matrix for the subspace V. Then $\boldsymbol{X} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{C}$ is a basis matrix for the subspace V^*.
- By using the uniqueness of the projection matrix, we write

$$\hat{\mu}^* = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^*.$$

- Note that $\hat{\mu} = \boldsymbol{\Gamma}^{1/2}\hat{\mu}^*$ and then

$$\hat{\mu} = \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{\Gamma}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{\Gamma}^{-1}\boldsymbol{Y} \quad (5)$$
$$\hat{\beta} = (\boldsymbol{C}'\boldsymbol{\Gamma}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{\Gamma}^{-1}\boldsymbol{Y} \quad (6)$$

- Under model (1), both $\hat{\mu}$ and $\hat{\beta}$ are unbiased for μ and β, respectively.

Estimate of the Variance

$$
\begin{aligned}
(N-p)\hat{\sigma}^2 &= ||\boldsymbol{P}_{V_\perp^*}\boldsymbol{Y}^*||^2 \\
&= ||\boldsymbol{P}_{V_\perp^*}(\boldsymbol{\mu}^* + \boldsymbol{\epsilon}^*)||^2 = ||\boldsymbol{P}_{V_\perp^*}\boldsymbol{\epsilon}^*||^2 \\
&= ||\boldsymbol{Y}^*||^2 - ||\hat{\boldsymbol{\mu}}^*||^2
\end{aligned}
$$

In terms of the original model, this reduces to

$$(N-p)\hat{\sigma}^2 = (\boldsymbol{Y} - \hat{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y} - \hat{\mu})$$

Hypothesis Testing

- Let W be a q-dimensional subspace of V, where $q \leq p$.
- Then the general linear hypothesis considers testing

$$H_0 : \boldsymbol{\mu} \in W \quad \text{against } H_A : \boldsymbol{\mu} \notin W. \tag{7}$$

- Let $W^* = \{\boldsymbol{w}^* = \Gamma^{-1/2}\boldsymbol{w}; \boldsymbol{w} \in W\}$. Then $\boldsymbol{\mu}^* \in W^*$ if and only $\boldsymbol{\mu} \in W$. Therefore, testing hypotheses (7) is equivalent to testing

$$H_0 : \boldsymbol{\mu} \in W^* \quad H_A : \boldsymbol{\mu} \notin W^*. \tag{8}$$

- Then the generalized F-test statistic for the hypothesis (8) has the form

$$F_N = \frac{||\boldsymbol{P}_{V^*|W^*}\boldsymbol{Y}^*||^2}{||\boldsymbol{P}_{V^*_\perp}\boldsymbol{Y}^*||^2} \times \frac{N-p}{p-q}. \tag{9}$$

Two treatments

- Let

$$\begin{aligned} Y_{i[j]tl} &= \theta + \alpha_i + \gamma_j + (\alpha\gamma)_{i[j]} + \sigma\epsilon_{i[j]tl} \\ &= \mu_{i[j]} + \sigma\epsilon_{i[j]tl}, \end{aligned}$$

  where $Y_{i[j]tl}$ is the response measurement from the i-th treatment, j-th judgment class, t-th set and l-th replication.
- In this model, we use the usual constraints

$$\sum_{i=1}^{2} \alpha_i = 0, \quad \sum_{j=1}^{2} \gamma_j = 0,$$

$$\sum_{i=1}^{2} (\alpha\gamma)_{i[j]} = 0 \quad \sum_{j=1}^{2} (\alpha\gamma)_{i[j]} = 0.$$

- Under these constraints, the subspace V that spans $\mu_{i[j]}$ has dimension 4.

Generalized Least Squares Estimators

- Assume that F is symmetric. Then under some regularity conditions

$$\begin{aligned}\hat{\theta} &= \bar{Y}_{.[.]..}, \quad \hat{\alpha}_i = (\bar{Y}_{i[.]..} - \bar{Y}_{.[.]..}); i = 1, 2, \\ \hat{\gamma}_j &= (\bar{Y}_{.[j]..} - \bar{Y}_{.[.]..}); j = 1, 2, \; \hat{\alpha\gamma}_{1[1]} = (\bar{Y}_{.[.]1.} - \bar{Y}_{.[.]2.})\end{aligned}$$

- The total variation is partitioned as

Source	df	Sum of Square
Treat	1	$SSTR = b\sum_{i=1}^{2}(\bar{Y}_{i[.]..} - \bar{Y}_{.[.]..})^2$
Jud	1	$SSJ = b\sum_{j=1}^{2}(\bar{Y}_{.[j]..} - \bar{Y}_{.[.]..})^2$
Int	1	$SSI = c\sum_{j=1}^{2}(\bar{Y}_{.[.]j.} - \bar{Y}_{.[.]..})^2$
Res	$N-3$	$SSR = SST - SSTR - SSJ$

$SST = \frac{b}{4n}\sum_{i=1}^{n}\{(Y_{1[1]1i} - Y_{2[2]1i})^2 + (Y_{2[1]2i} - Y_{1[2]2i})^2\} +$
$c\sum_{i=1}^{n}\{Y_{1[1]1i}Y_{2[2]1i} + Y_{2[1]2i}Y_{1[2]2i}\}$
$b = 2n/(\sigma_\epsilon^2(1-\rho))$, $c = 2n/(\sigma_\epsilon^2(1+\rho))$

Empirical Power

- Model

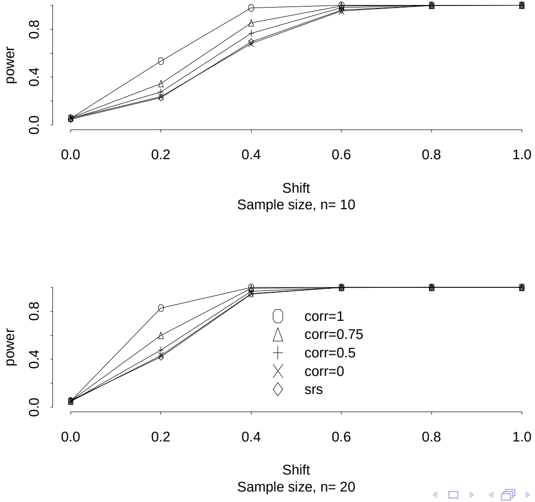
$$Y_{i[j]tl} = \theta + \alpha_i + \gamma_j + (\alpha\gamma)_{i[j]} + \sigma\epsilon_{i[j]tl}, i, j, t = 1, 2; l = 1, ..., n$$

- Dell and Clutter (1972) model: Units are ranked based on estimates of the error terms on each unit,

$$u_i = \epsilon_i + w_i, \epsilon_i \sim F, w_i \sim N(0, \tau^2)$$

- For our model, a vector of H independent error terms, $\epsilon = (\epsilon_1, \cdots, \epsilon_H)$ is generated from F. Another vector of H independent random variates, $\mathbf{w} = (w_1, \cdots, w_H)$, is generated from $N(0, \tau^2)$.
- We add ϵ and $\mathbf{w}$, $\mathbf{u} = \epsilon + \mathbf{w}$, and sort the vector $\mathbf{u}$. Corresponding ϵ values are taken as judgment order statistics.
- The information content of judgment ranking in this model is controlled by the variance of w, or equivalently by the correlation between u and ϵ, $cor = 1/(1 + \tau^2)^{1/2}$.

Empirical power ORR design



Concluding Remarks

- Use of EU related subjective information improves statistical inference in ORR and RSS designs.
- ORR designs unlike RSS designs uses all experimental units available for the study.
- For the inference on the contrasting features of factor levels, ORR design provides improved efficiency over RSS and SRS designs.
- Linear model analysis provides a unified theory that can be used for SRS, RSS and ORR as long as you determine appropriate covariance structure.