Contents

4 Paired Comparisons & Block Designs 3

4.1 Paired Comparisons 3

4.1.1 Paired Data 3

4.1.2 Existing Approaches 6

4.1.3 Paired-comparison Permutation Test 9

4.1.4 A Test for Median of Symmetric Distribution . . 14

4.1.5 Wilcoxon Signed-Rank Test for Paired Comparison 16

4.2 Permutation Test for a RCB Design 20

4.2.1 RCB Design and Model Setup 20

4.2.2 Classic Method–F-test 22

4.2.3 Permutation Test for a RCBD 24

4.3 Friedman's Test for a RCB Design 28

4.4 Cochran's Q Test 32

4.5 Kendall's W Test 33

4.6 Ordered Alternatives for RCBD 37

4 Paired Comparisons & Block Designs

4.1 Paired Comparisons

4.1.1 Paired Data

Recall the **Depression Example** 2.3.2: The data gives the depression scale factors of 9 patients, measured before the therapy X and after the therapy Y.

Patient i	1	2	3	4	5	6	7	8	9
Before X_i	1.83	0.5	1.62	2.48	1.68	1.88	1.55	3.06	1.30
After Y_i	0.88	0.65	0.59	2.05	1.06	1.29	1.06	3.14	1.29

Blood cell count example: suppose we have paired observations of red blood cell counts on 5 subjects using two different reagents

(chemical compounds used to facilitate the measurement process:)

Patient i	1	2	3	4	5
Reagent B	3.93	5.35	5.39	5.16	5.13
Reagent C	3.86	5.10	5.39	5.01	5.05

- The above two examples are **paired data**, where the measurements from the same subject are obtained under two different treatments.
- Therefore, for each i, X_i and Y_i are dependent, and two-sample tests discussed in Chapter 3 are not valid (either gives inflated Type I error or loses power depending on the direction of the correlation).
- Other examples of paired data:
  - before/after scores, twin data, data on 2 hands, 2 eyes etc of some subjects, each person experiences both treatments, or

"matched pairs" (e.g. teammates or competitors).

Reason for pairing: to reduce variability

- Variability within X's and within Y's (between-subject variation) may be large compared to the treatment effect.
- If subjects are highly variable, this may mask the treatment effect.
- The idea is to divide subjects into more homogeneous pairs (or blocks), so that the within-subject variation is small. The treatments should be randomly assigned to members of each pair (e.g. by flipping a fair coin)

Paired data analysis:

- The main interest is to test if there is any difference between two treatments (e.g. "before" and "after").
- To account for the within-subject correlation, look at the difference data $D_i = Y_i - X_i$, where D_i's are independent of each

other (as subjects are assumed to be independent).

4.1.2 Existing Approaches

1. Paired t-test

If we are willing to assume D_i are $i.i.d.$ normal, then we can carry out a paired t-test, i.e. one sample t-test based on D_i.

For the **Blood Cell Count** data:

Patient i	1	2	3	4	5
Reagent B (X_i)	3.93	5.35	5.39	5.16	5.13
Reagent C (Y_i)	3.86	5.10	5.39	5.01	5.05
$D_i = Y_i - X_i$	-0.07	-0.25	0	-0.15	-0.08

$\bar{D} = -0.11$, $S_d = 0.0946$, $t = \frac{\bar{D}}{S_d/\sqrt{5}} = -2.6$. The $df = n - 1 = 4$.
For 2-sided test, p-value $= 2P(t_4 > |-2.6|) = 0.06005$, so we reject H_0 at significance level $\alpha = 0.1$.

However, a two-sample t-test with equal variance leads to $t = 0.2911$ with $df = n + m - 2 = 8$. For two-sided test, p-value=0.7784. So the large between-subject variation masks the group effect.

2. Nonparametric methods for testing for median of symmetric distribution

We can use the nonparametric methods discussed in Chapter 2.

- Let θ denote the median of D_i.
- We assume the distribution of D_i is symmetric around θ (i.e. observing a value k units larger than the median is as likely as observing a value k units smaller than the median).
- Test

$$H_0 : \theta = \theta_0, \quad v.s.\ H_a : \theta \neq \theta_0 \text{ (or } < \theta_0 \text{ or } > \theta_0),$$

  where θ_0 is the claimed median.

Methods discussed in Chapter 2 (review):

- Binomial (sign) test: test statistic

$$S = \#\{D_i > \theta_0\}.$$

  Under $H_0, S \sim Binomial(n, p = 0.5)$.

- Wilcoxon signed rank test: test statistic

$$SR_+ = \sum_{i=1}^{n} I(D_i - \theta_0 > 0) rank(|D_i - \theta_0|),$$

  the sum of ranks of the positive $D_i - \theta_0$. The critical values of SR_+ are tabulated in Table A9 or we can find the null distribution of SR_+ by permutation.

4.1.3 Paired-comparison Permutation Test

Let $F(\cdot)$ be the CDF of the differences D_i.

- Null hypothesis

  $$H_0 : F(\cdot) \text{ is symmetric about 0, i.e. } F(x) = 1 - F(-x),$$

  that is, D and $-D$ have the same distribution, thus $\theta = 0$.

- Possible alternative hypotheses:
  - The upper-tail alternative (D_i tends to fall more towards the positive side):

    $$H_a : F(x) < 1 - F(-x)$$

  - The lower-tail alternative (D_i tends to fall more towards the negative side):

    $$H_a : F(x) > 1 - F(-x)$$

  - The shift alternative:

    $$H_a : F(x) = G(x - \theta),$$

where $G(x)$ is symmetric about 0, θ is the median of D_i. We may test $H_a : \theta \neq 0$ or < 0 or < 0.

For paired data, e.g. the Blood Cell Count data

- If there is no difference in treatments for subject 1, it is equally likely that 3.93 comes from Reagent B as Reagent C. Or, difference D_i could be $+0.07$ or -0.07 with equal chance.
- For n subjects, there are total 2^n possible ways of assigning the two values within each subject (or assigning plus/minus signs for the difference of each subject). If $n = 5$, there are total $2^5 = 32$ ways.

Paired-Comparison Permutation Test:

- Get difference D_i for each pair of data and calculate $\bar{D}_{obs} = n^{-1} \sum_{i=1}^{n} D_i$.
- Obtain the 2^n possible assignments of plus and minus signs to the

$|D_i|$'s (or obtain a sample of R permutations).

- Calculate $\bar{D}^*$ for each permutation.
- Calculate the p-value
  - for upper-tailed test, $p\text{-value}_{\text{upper}} = \#\{\bar{D}^*\text{'s} \geq \bar{D}_{obs}\}/R$
  - for lower-tailed test, $p\text{-value}_{\text{lower}} = \#\{\bar{D}^*\text{'s} \leq \bar{D}_{obs}\}/R$
  - for two-tailed test,

$$p\text{-value} = \frac{\#\text{of } |\bar{D}^*|\text{'s} \geq |\bar{D}_{obs}|}{R} \approx 2 \times p\text{-value}_{\text{upper}},$$

since the distribution of D_i is symmetric under H_0.

Example 4.1.1 *Blood Cell Count Data: compare the red blood cell counts using Reagent B and those using Reagent C with* $n = 5$ *subjects.*

Patient i	*1*	*2*	*3*	*4*	*5*
Reagent B (X_i)	*3.93*	*5.35*	*5.39*	*5.16*	*5.13*
Reagent C (Y_i)	*3.86*	*5.10*	*5.39*	*5.01*	*5.05*
$D_i = Y_i - X_i$	*-0.07*	*-0.25*	*0*	*-0.15*	*-0.08*

Example 4.1.2 *Analysis of the Depression data set.*

4.1.4 A Test for Median of Symmetric Distribution

We can use the same procedure as "Paired-Comparison Permutation Test" to test the hypotheses on median of symmetric distribution, where D_i are not necessarily paired differences.

- Suppose X_i are symmetrically distributed around the median θ.
- We want to test

$$H_0 : \theta = \theta_0 \quad v.s. \; H_a : \theta \neq (\text{or } < \text{ or } >)\theta_0,$$

  where θ_0 is a claimed median.
- Define $D_i = X_i - \theta_0$.
- Then under H_0, D_i is symmetric around 0, i.e. D_i and $-D_i$ have the same distribution.
- Thus the paired-comparison permutation test can be applied.

Example 4.1.3 *The following data are the rainfall totals (inches) for Scranton, PA from 1968-1984:*

26.8, 18.9, 36.3, 28, 17.9, 25, 27.5, 27.7, 32.1, 28, 30.9, 20, 20.2, 33.5, 26.4, 30.9, 33.2. Test if the median rainfall is greater than 23.

4.1.5 Wilcoxon Signed-Rank Test for Paired Comparison

- A nonparametric test for paired data based on ranks.
- The rankings are done differently for paired experiments than for two-sample experiments.
- Recall for paired data, the signed ranks are assigned by
  - ranking the absolute values of the differences D_i
  - re-attach the signs ("+", "0", "-") to the ranks
- Carry out the paired-comparison permutation test on the signed ranks SR_i
  - Calculate the observed test statistic, e.g.

$$\overline{SR} = n^{-1} \sum_{i=1}^{n} SR_i,$$

or the sum of positive ranks,

$$SR_+ = \sum_{i:D_i>0} SR_i.$$

- Compute SR_+^* for all 2^n possible assignments of plus and minus signs to the ranks of $|D_i|$ (or a sample R permutations).
- Compute the p-value:
  - upper-tail: the proportion of SR_+^* that are $\geq SR_+$
  - lower-tail: the proportion of SR_+^* that are $\leq SR_+$
  - two-tailed: twice the one-tail p-value

- The above procedure is essentially the Wilcoxon's signed-rank test.
- Table A9 gives critical values for SR_+ assuming no zero differences and no ties among differences. See also Chapter 2 for the approximate distribution of SR_+.

Example 4.1.4 *For the* **Blood Cell Count** *data:*

Patient i	1	2	3	4	5
Reagent B (X_i)	3.93	5.35	5.39	5.16	5.13
Reagent C (Y_i)	3.86	5.10	5.39	5.01	5.05
$D_i = Y_i - X_i$	-0.07	-0.25	0	-0.15	-0.08
rank of $\lvert D_i \rvert$	2	5	1	4	3
signed rank SR	-2	-5	0	-4	-3
alternative SR	-1	-4	/	-3	-2

- The average signed-rank $\overline{SR} = (-2-5+0-4-3)/5 = -2.8$ or $\overline{SR} = (-1-4-3-2)/4 = -2.5$ (alternative SR).
- Or the signed-rank statistic $SR_+ = 0$ with $n = 4$.
- For $n = 4$, Table A9 gives $P(SR_+ \leq 0) = P(SR_+ \geq n(n+1)/2 - 0 = 10)$, so for

lower-tailed test, p-value$=P(SR_+ \leq 0) = 0.063$.

```
> wilcox.test(d)

Wilcoxon signed rank test with continuity correction
data:  d
V = 0, p-value = 0.1003
alternative hypothesis: true location is not equal to 0

Warning message:
In wilcox.test.default(d) : cannot compute exact p-value with zeroes

> wilcox.test(x, y, paired=T)

Wilcoxon signed rank test with continuity correction
data:  x and y
V = 10, p-value = 0.1003
alternative hypothesis: true location shift is not equal to 0
```

4.2 Permutation Test for a RCB Design

4.2.1 RCB Design and Model Setup

A Randomized Complete Block (RCB) design:

Field 1	Field 2	Field 3	Field 4	Field 5
B	C	D	A	B
A	B	C	D	A
D	A	B	C	D
C	D	A	B	C

- Fields may be heterogeneous (have different characteristics), and this may mask the treatment effect.
- **Features of RCBD**:
  - divide experimental units into blocks (fields here), such that

plots within a field/block are roughly homogeneous;

- in each block, the number of experimental units is equal to the number of treatments;
- within each block, treatments are assigned to the experimental units within blocks.

- An extension of paired experiments to multiple (≥ 2) treatments.
- Other possible blocking factors:
  - location, laboratory, age, medical profiles, weights of animals, soils etc.

Notation and additive model:

$$X_{ij} = \mu + t_i + b_j + \epsilon_{ij},\ i = 1, \cdots, K, j = 1, \cdots, J,$$

where

- X_{ij}: observation for treatment i in block j;

- μ: overall median/mean;
- t_i: effect for treatment i;
- b_j: effect for block j;
- ϵ_{ij}: $i.i.d.$ random error with median/mean 0.

Main interest: test the treatment effect, i.e. test

$$H_0 : t_1 = t_2 = \cdots = t_K = 0.$$

4.2.2 Classic Method–F-test

Denote

- grand mean: $\bar{X} = N^{-1} \sum_{i=1}^{K} \sum_{j=1}^{J} X_{ij}$, $N = JK$;
- mean for treatment i: $\bar{X}_{i\cdot} = J^{-1} \sum_{j=1}^{J} X_{ij}$;
- mean for block j: $\bar{X}_{\cdot j} = K^{-1} \sum_{i=1}^{K} X_{ij}$.

If we are willing to assume $\epsilon_{ij} \sim N(0, \sigma^2)$, then we can carry out the F-test. Under H_0,

$$F = \frac{SST/(K-1)}{SSE/[(K-1)(J-1)]} \sim F_{K-1,(K-1)(J-1)}, \tag{4.1}$$

where

- $SST = J\sum_{i=1}^{K}(\bar{X}_{i.} - \bar{X})^2$: sum of squares for treatment;
- $SSE = \sum_{i=1}^{K}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2$: sum of squares for error.

For testing

$$H_0 : t_1 = t_2 = \cdots = t_K = 0 \text{ versus } H_a : \text{ not all } t_i\text{'s are zero},$$

we can compute the p-value$=P(F_{K-1,(K-1)(J-1)} \geq F_{obs})$.

4.2.3 Permutation Test for a RCBD

If we are NOT willing to make distributional assumption about ϵ_{ij}, we can form the permutation distribution for the F statistic defined in (4.1) under H_0.

Procedure of Permutation Test for a RCBD:

- Compute F_{obs} using the original data.
- Permute the observations **within each block**. Do not mix observations across blocks. There are total $(K!)^J$ permutations.
- For each permutation (or a sample R of them), calculate F^*.
- Permutation p-value is

$$p\text{-value} = \frac{\#\text{of } F^*\text{'s} \geq F_{obs}}{R}.$$

Note: we can calculate something less complicated than the F-statistic in (4.1).

$$\begin{aligned} SSTotal &= \sum_{i=1}^{K}\sum_{j=1}^{J}(X_{ij}-\bar{X})^2 \\ &= SST + SSB + SSE, \end{aligned}$$

where

$$SSB = K\sum_{j=1}^{J}(\bar{X}_{.j}-\bar{X})^2.$$

Therefore

$$F = \frac{SST/(K-1)}{(SSTotal - SST - SSB)/[(K-1)(J-1)]},$$

where $SSTotal$ and SSB are both the same for every within-block permutation. So F is monotonically increasing in

$$SST = J\sum_{i=1}^{K}(\bar{X}_{i.}-\bar{X})^2$$

and in

$$\begin{aligned} SST/J &= \sum_{i=1}^{K}(\bar{X}_{i.} - \bar{X})^2 \\ &= \sum_{i=1}^{K}\left(\bar{X}_{i.}^2 - 2\bar{X}\bar{X}_{i.} + \bar{X}^2\right) \\ &= \sum_{i=1}^{K}\bar{X}_{i.}^2 - 2\bar{X}\sum_{i=1}^{K}\bar{X}_{i.} + K\bar{X}^2 \\ &= \sum_{i=1}^{K}\bar{X}_{i.}^2 - K\bar{X}^2 \end{aligned}$$

and thus in

$$SSTM = \sum_{i=1}^{K}\bar{X}_{i.}^2.$$

Thus, in the above permutation test procedure, we can replace the F_{obs} and F^* by $SSTM_{obs}$ and $SSTM^*$ without affecting the

significance results.

Example 4.2.1 *Soybean data set. See R code*

Response: Number of failures out of 100 planted soybean seeds
Treatments on seeds to enhance germination: 1=none, 2=Arasan, 3=Spergon, 4=Semesan, 5=Fermate

X_{ij}	Block					
Trt	1	2	3	4	5	$\bar{X}_{i.}$
1	8	10	12	13	11	10.8
2	2	6	7	11	5	6.2
3	4	10	9	8	10	8.2
4	3	5	9	10	6	6.6
5	9	7	5	5	3	5.8
$\bar{X}_{.j}$	5.2	7.6	8.4	9.4	7.0	7.5

4.3 Friedman's Test for a RCB Design

Friedman's Test is just the RCBD permutation test applied to ranks.

Steps for the Friedman's test:

- Rank the observations within each block to obtain R_{ij}. Note that within the same block j, the ranks $R_{1j}, \cdots, R_{Kj}$ range from 1 to K, and the mean rank is $(K+1)/2$.
- Define the Friedmans statistic as

$$FM = \frac{12}{K(K+1)} J \sum_{i=1}^{K} \left\{ \bar{R}_{i\cdot} - (K+1)/2 \right\}^2,$$

  where note that $J \sum_{i=1}^{K} \left\{ \bar{R}_{i\cdot} - (K+1)/2 \right\}^2$ is just the SST for ranks.
- The null distribution of FM
  - the critical values are partially tabulated in Table A11;

– for large samples, FM is approximately χ^2_{K-1} **under** H_0.

A few points:

- The χ^2_{K-1} approximation is shaky for small sample sizes.
- Tables are very limited.
- Distribution of FM is complicated when ties exist.
- Alternatively, we can use the permutation approach to conduct a Friedman-like test:
  – find the permutation distribution of FM statistic or equivalently that of $\sum_{i=1}^{K} \bar{R}_{i.}^2$
- For the two-treatment case, Friedman is equivalent to the two-sided sign test—**low power** unless the distribution is very highly heavy-tailed.
- When there are ties, use midranks, and adjust the Friedman's

statistic

$$FM_{ties} = \frac{1}{\frac{1}{J}\sum_{j=1}^{J} S_{Bj}^2} J \sum_{i=1}^{K} \left(\bar{R}_{i.} - \frac{K+1}{2} \right)^2,$$

where S_{Bj}^2 is the sample variance of ranks within block j.

Example 4.3.1 *Data and ranks for the number of failures out of 100 planted soybean seeds:*

Trt i	*block1*	*block2*	*block3*	*block4*	*block5*	$\bar{R}_{i.}$
1	*8 (4)*	*10 (4.5)*	*12 (5)*	*13 (5)*	*11 (5)*	*4.7*
2	*2 (1)*	*6 (2)*	*7 (2)*	*11 (4)*	*5 (2)*	*2.2*
3	*4 (3)*	*10 (4.5)*	*9 (3.5)*	*8 (2)*	*10 (4)*	*3.4*
4	*3 (2)*	*5 (1)*	*9 (3.5)*	*10 (3)*	*6 (3)*	*2.5*
5	*9 (5)*	*7 (3)*	*5 (1)*	*5 (1)*	*3 (1)*	*2.2*
S_{Bj}^2	*2.5*	*2.375*	*2.375*	*2.5*	*2.5*	

$FM_{ties} = \frac{J^2}{\sum_{j=1}^{J} S_{Bj}^2} \sum_{i=1}^{K} \left(\bar{R}_{i.} - \frac{K+1}{2}\right)^2 = 9.35$. *Using* χ^2_4 *approximation,* p*-value*$= P\left(\chi^2{}_4 \geq 9.35\right) = 0.053$.

4.4 Cochran's Q Test

In some applications, we may have a RCBD where the response is either 1 (success) or 0 (failure).

Example 4.4.1 *7 subjects with multiple plantar warts were obtained. Four warts on the foot of each subject were randomly assigned to 4 treatments. The response 1: wart is successfully removed; 0: wart returned.*

	Subject						
Treatment	1	2	3	4	5	6	7
freezing	0	1	0	0	0	1	0
surgical removal	1	1	0	1	0	1	1
salicylic acid	0	0	0	1	0	1	0
duct tape	1	1	0	0	1	0	1

Cochran's Q test (for 0/1 responses):

- Applying Friedman's test (i.e. permutation test for RCBD on ranks) with ties is equivalent to Cochran's Q test.

The ranks of the Wart data:

	1	2	3	4	5	6	7	$Ri.$
freezing	1.5	3.0	2.5	1.5	2.0	3.0	1.5	2.14
surgical	3.5	3.0	2.5	3.5	2.0	3.0	3.5	3.00
salicylic acid	1.5	1.0	2.5	3.5	2.0	3.0	1.5	2.14
duct tape	3.5	3.0	2.5	1.5	4.0	1.0	3.5	2.71
S^2_{Bj}	1.33	1.0	0.0	1.33	1.0	1.0	1.33	

```
> friedman.test(x, grps, blocks)  ### based on chi-square approximation
Friedman rank sum test
data:  x, grps and blocks
Friedman chi-squared = 3.8571, df = 3, p-value = 0.2773
```

4.5 Kendall's W Test

For example, five employees ranked on the basis of leadership ability by each of three supervisors.

Table 1: Leadership ratings

	Supervisor(block)		
Employee	1	2	3
1	3	2	2
2	2	1	1
3	4	4	3
4	1	3	2
5	5	5	4

Goal: to assess the agreement/concordance of the judges/supervisors.

Equivalent hypotheses:

- H_0: supervisors (judges) do not agree (ranks could just as well have been random, i.e. no group (employee) differences)
- H_a: supervisors (judges) agree (tend to rate some individuals higher or lower)

Kendall's W test:

- A test for agreement or concordance among judges, e.g. in a gymnastics or figure skating competition.
- Want to ignore the possible fact that judge A always gives lower scores, or judge B always gives scores with large variability (either very high or very low).
- Other examples
  - food taste tests
  - ranking of students in different areas: mechanical, artistic,

literacy, musical, mathematical, clerical etc

– teaching evaluation scores for different questions

- **How to carry out the test?** Applying Friedman's test (i.e. comparing ranks using the permutation approach for RCBD) is equivalent to Kendall's W test for concordance.

Example The *analyses of "Leadership ability" and "Aptitude scores" data sets.*

4.6 Ordered Alternatives for RCBD

Example: yield data from a RCBD in which 4 different types of tractors were used in tilling the soil.

Table 2: Yield data from a RCBD.

	block						
trt	1	2	3	4	5	6	ranksum
1	120 (1)	208 (4)	199 (4)	194 (4)	177 (4)	195 (4)	21
2	207 (4)	188 (3)	181 (3)	164 (2)	155 (1)	175 (2)	15
3	122 (2)	137 (2)	177 (2)	177 (3)	160 (3)	138 (1)	13
4	128 (3)	128 (1)	160 (1)	142 (1)	157 (2)	179 (3)	11

We anticipate that yields decrease from treatment 1 to 4. That is, we would like to test

$$H_0 : t_1 = t_2 = t_3 = t_4 \quad \text{v.s.} \quad H_a : t_1 \geq t_2 \geq t_3 \geq t_4,$$

corresponding to an **ordered alternative**. In some other applications, we may want to test $H_a : t_1 \leq t_2 \leq t_3 \leq t_4$.

Page's Test:

- Let R_i denote the sum of ranks (within-block-ranks) for the ith treatment.
- Page's statistic is

$$PG = \sum_{i=1}^{K} iR_i$$

  - $H_a : t_1 \leq t_2 \leq t_3 \leq t_4$: treatment response tends to increase as i goes from 1 to K, then PG is large and p-value is obtained by examining the upper tail of the permutation distribution, i.e. p-value=proportion of PG^* that are $\geq PG_{obs}$, where PG^* are the PG statistics obtained with permutation samples (permuting the ranks within each block).

- $H_a : t_1 \geq t_2 \geq t_3 \geq t_4$: responses decrease as i goes from 1 to K, PG is small, p-value=proportion of PG^* that are $\leq PG_{obs}$.

- Procedure is identical to Friedman test, but we calculate PG statistic instead of F or SSTM.

Example 4.6.1 *Analyze the Yield data set and test* $H_0 : t_1 = t_2 = t_3 = t_4$ *v.s.* $H_a : t_1 \geq t_2 \geq t_3 \geq t_4$. *Compare to the results of Friedman's test.*