Advanced Methods in Biometry

- Longitudinal Data -

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Outline

Introduction to Longitudinal Data

- Longitudinal data: Data collected from the same subjects at multiple time points.
- Objectives:
  - Understand within-subject changes over time.
  - Compare changes between groups.
- Common statistical methods:
  - Paired t-test
  - Wilcoxon signed-rank test
  - Repeated Measures ANOVA
  - Split Plot Plan

Paired t-test

Introduction

- Compare means of two related groups.
- Accounts for dependency between pairs.

Statistical Model

$$\mathbf{X}_k = (X_{1k}, X_{2k})' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)' \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Test Statistic for $H_0 : \mu_1 = \mu_2$

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \quad d_i = X_{1i} - X_{2i}, \quad \overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$t \sim T(n-1)$$

Example in R

```
# Sample data
before <- c(85, 90, 76, 92, 88)
after <- c(87, 88, 77, 90, 86)

# Perform paired t-test
result <- t.test(before, after,
                 paired = TRUE)
```

R Output

```
Paired t-test
data: before and after
t = 2.4495, df = 4, p-value = 0.07102
95% CI: -0.5798435 4.3798435
mean of the differences: 1.9
```

Wilcoxon Signed Rank Test

Introduction

- Does not assume normal but a symmetrical distribution of differences

Test Procedure

- Calculate differences $d_i = X_{1i} - X_{2i}$.
- Rank the absolute differences $|d_i|$.
- Assign ranks R_i to $|d_i|$.
- Sum the ranks for positive (T^+) and negative (T^-) differences separately.

Test Statistic

$$W = \min(T^+, T^-)$$

Example in R

```
# Sample data
before <- c(85, 90, 76, 92, 88)
after <- c(87, 88, 77, 90, 86)

# Perform Wilcoxon signed rank test
result <- wilcox.test(before, after,
                      paired = TRUE)
```

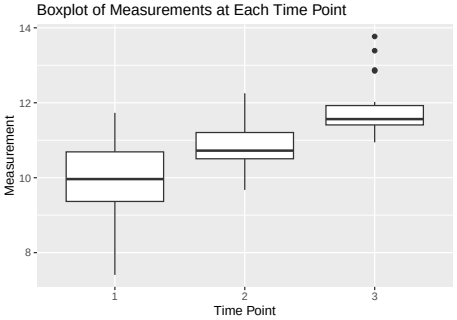
R Output

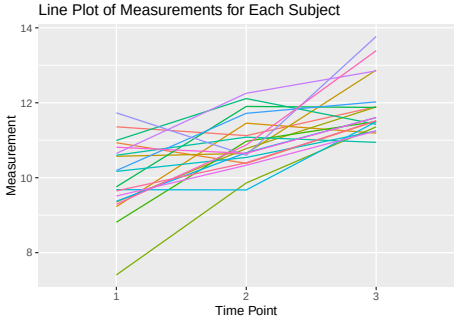
```
Wilcoxon signed-rank test

data:  before and after
V = 8, p-value = 0.3455
alternative hypothesis:
true location shift is not equal to 0
```

Repeated Measures: Example

- Researchers study the effect of a cognitive training program on memory performance.
- 20 participants measured at three time points: before training, immediately after, and one month after.
- Memory performance assessed using a standardized memory test.
- Hypothesis: Memory scores will increase over time.





Longitudinal Data Analysis

Repeated Measures ANOVA

- **Model:**

$$Y = \mu + \text{Time} + \text{Subject} + \epsilon$$

- **Assumptions:**
  - Sphericity
  - Normality of residuals.
- **Limitations:**
  - Sensitive to sphericity violations.
  - Cannot handle missing data or unbalanced designs well.

Mixed Models

- **Model:**

$$Y = X\beta + Zb + \epsilon$$

- **Assumptions:**
  - Normality of residuals.
  - Random effects are normally distributed.
- **Advantages:**
  - Handles missing and unbalanced data.
  - Can model complex covariance structures.

General Longitudinal Data

- **Model:**

$$\mathbf{Y} \sim \mathbf{F}(\boldsymbol{\Sigma})$$

- **Methods:**
  - RM ANOVA for simple cases.
  - General covariance modelling
- **Considerations:**
  - Choice of method depends on data structure and research questions.

Repeated Measures ANOVA Model

Introduction

- Compares means across multiple time points or conditions for the same subjects.
- Accounts for within-subject correlation.

Statistical Model

$$Y_{ij} = \mu + \alpha_i + \Gamma_j + \epsilon_{ij}$$

where:

- Y_{ij} is the response for subject j at time i.
- α_i: fixed time effect
- $\Gamma_j \sim N(0, \sigma^2_\Gamma)$
- $\epsilon_{ij} \sim N(0, \sigma^2)$

Assumptions

- **Normality**: The residuals ϵ_{ij} are normally distributed.
- **Homogeneity of variances**: The variances of ϵ_{ij} are equal across groups.
- **Sphericity**: The variances of the differences between all combinations of related groups (levels) are equal.

Covariance Matrix

- Compound Symmetry: Equal variances and covariances.
- Sphericity: More general form; variances of differences between all pairs are equal.

Sum of Squares in Repeated Measures ANOVA

Suppose we have $j = 1, \ldots, n$ subjects observed at $i = 1, \ldots, d$ time points in Y_{ij}:

Total Sum of Squares (SS_{Total}):

$$\begin{aligned} SS_{Total} &= \sum_{i=1}^{d}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{..})^2 = \underbrace{n\sum_{i=1}^{d}(\overline{Y}_{i.} - \overline{Y}_{..})^2}_{SS_{Between}} + \underbrace{\sum_{i=1}^{d}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i.})^2}_{Within(Error)} \\ &= SS_{Between} + \underbrace{d\sum_{j=1}^{n}(\overline{Y}_{.j} - \overline{Y}_{..})^2}_{SS_{Subject}} + \widetilde{SS}_{Error} \end{aligned}$$

- $\overline{Y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ (means per time point)
- $\overline{Y}_{.j} = \frac{1}{d}\sum_{i=1}^{d} Y_{ij}$ (means per subject)
- $\overline{Y}_{..} = \frac{1}{dn}\sum_{i=1}^{d}\sum_{j=1}^{n} Y_{ij}$ (overall mean)

Repeated Measures ANOVA

- F-test for independent data

$$F = MS_{between} / MS_{Within}$$

- Repeated Measures ANOVA
  - Decompose

$$SS_{total} = SS_{Between} + SS_{Subject} + SS_{Error}$$

  - Degrees of Freedom
    - Between: $d - 1$
    - Subject: $n - 1$
    - Error: $nd - (d - 1) - (n - 1) - 1$
  - F-test

$$F = MS_{between} / MS_{within}$$

Data Evaluations in R

```
library(lme4); library(lmerTest)
fit <- aov(Measurement ~ Time + Error(Subject/Time), data = data)
summary(fit)

fit2 <- lmer(Measurement~Time+(1|Subject),data)
summary(fit2)
anova(fit2)

fit3 <- lm(Measurement~Time,data) #Independent: Explain the differences
anova(fit3)

# See the markdown script
```

Modeling Longitudinal Data with a RM ANOVA / Mixed Model

- Modeling repeated measures within a mixed model implies strong assumptions
- $Y_{ij} = \mu + \alpha_i + \Gamma_j + \epsilon_{ij}$; $\Gamma_i \sim N(0, \sigma^2_\Gamma)$
- The covariance matrix can be of a certain structure only

$$Cov(Y_{ij}, Y_{i'j}) = Cov(\Gamma_j, \Gamma_j) = \sigma^2_\Gamma$$

Covariance Matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_\Gamma + \sigma^2 & \sigma^2_\Gamma & \cdots & \sigma^2_\Gamma \\ \sigma^2_\Gamma & \sigma^2_\Gamma + \sigma^2 & \cdots & \sigma^2_\Gamma \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_\Gamma & \sigma^2_\Gamma & \cdots & \sigma^2_\Gamma + \sigma^2 \end{pmatrix}$$

- All Variances are identical
- All covariances are identical
- **Compound Symmetry** structure

Prominent Covariance Structures: AR(1)

Introduction

- Autoregressive structure assumes that the correlation between observations decreases as the distance between them increases.
- For d equally spaced time points, the AR(1) covariance matrix is given by:

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{d-1} \\ \rho & 1 & \rho & \cdots & \rho^{d-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{d-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{d-1} & \rho^{d-2} & \rho^{d-3} & \cdots & 1 \end{pmatrix}$$

- $(\rho^{|i-j|})$

Example

- Suppose we have measurements at four equally spaced time points ($d = 4$) with autocorrelation $\rho = 0.5$ and variance $\sigma^2 = 8$.

Covariance Matrix

$$\mathbf{\Sigma} = 8 \begin{pmatrix} 1 & 0.5 & 0.25 & 0.125 \\ 0.5 & 1 & 0.5 & 0.25 \\ 0.25 & 0.5 & 1 & 0.5 \\ 0.125 & 0.25 & 0.5 & 1 \end{pmatrix}$$

Prominent Covariance Structures: AR(2)

Introduction

- AR (2): correlation between observations decreases as the distance between them increases with min ρ.
- For d equally spaced time points, the AR(2) covariance matrix is given by:

$$\boldsymbol{\Sigma} = \sigma^2 \left(\rho^{|i-j|/(d-1)}\right)$$

Example

- Suppose we have measurements at four equally spaced time points ($d = 4$) with autocorrelation $\rho = 0.5$ and variance $\sigma^2 = 8$.

Covariance Matrix

$$\boldsymbol{\Sigma} = 8 \begin{pmatrix} 1 & 0.5^{1/3} & 0.5^{2/3} & 0.5 \\ 0.5^{1/3} & 1 & 0.5^{1/3} & 0.5^{2/3} \\ 0.5^{2/3} & 0.5^{1/3} & 1 & 0.5^{1/3} \\ 0.5 & 0.5^{2/3} & 0.5^{1/3} & 1 \end{pmatrix}$$

Prominent Covariance Structures: Toeplitz

Introduction

- Toeplitz: correlation between observations linearly decreases.
- For d equally spaced time points, the Toeplitz covariance matrix is given by:

$$\Sigma = \sigma^2 \left(1 - abs(i - j)/d\right)$$

Example

- Suppose we have measurements at four equally spaced time points ($d = 4$) with variance $\sigma^2 = 8$.

Covariance Matrix

$$\Sigma = 8 \begin{pmatrix} 1 & 0.75 & 0.5 & 0.25 \\ 0.75 & 1 & 0.75 & 0.5 \\ 0.5 & 0.75 & 1 & 0.75 \\ 0.25 & 0.5 & 0.75 & 1 \end{pmatrix}$$

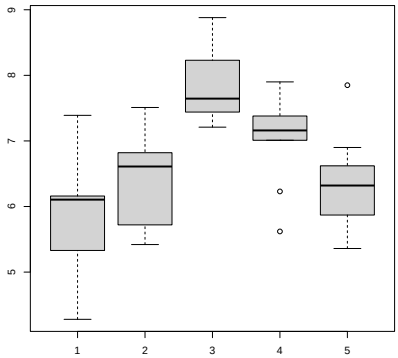
Prominent Covariance Structures: Unstructured

- Unstructured: No specific structure
- For d equally spaced time points, the US covariance matrix is given by:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk}^2 \end{pmatrix}$$

Example: Yeast Study

- $n = 10$ medical yeast samples
- Each sample grown on different growth mediums (increasing fertility)
- Response: protein concentration



ID	M1	M2	M3	M4	M5
1	6.16	6.04	7.21	7.23	6.22
2	4.28	5.42	7.44	6.23	6.03
3	5.26	5.72	7.40	7.02	5.87
4	6.11	6.65	7.44	7.09	6.62
5	6.15	6.67	7.79	5.62	5.80
6	5.33	7.50	8.23	7.38	6.42
7	5.47	6.82	7.94	7.01	6.57
8	6.10	6.57	8.73	7.90	6.90
9	7.39	5.44	7.50	7.32	5.36
10	7.07	7.51	8.88	7.70	7.85

Statistical Model

- Statistical Model
  - $\mathbf{X}_k = (X_{1k}, \ldots, X_{dk})' \sim \mathbf{F}$
  - $E(\mathbf{X}_k) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$
  - Covariance matrix $\boldsymbol{V} = Cov(\mathbf{X}_1)$ (unstructured)
  - No normal distribution
- Hypotheses $H_0 : \mu_1 = \ldots = \mu_d$

  $$H_0 : \boldsymbol{C\mu} = \mathbf{0}$$

- $\boldsymbol{C}$: Contrast matrix

Point Estimators

- Means and covariance matrix
  - $\overline{\mathbf{X}}_{\cdot} = (\overline{X}_{1\cdot}, \ldots, \overline{X}_{d\cdot})'$: Means
  - Empirical covariance matrix

$$\widehat{\boldsymbol{V}} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{X}_k - \overline{\mathbf{X}}_{\cdot})(\mathbf{X}_k - \overline{\mathbf{X}}_{\cdot})'$$

```
x=matrix(c(6.16, 6.04, 7.21, 7.23, 6.22,
4.28, 5.42, 7.44, 6.23, 6.03,
5.26, 5.72, 7.40, 7.02, 5.87,
6.11, 6.65, 7.44, 7.09, 6.62,
6.15, 6.67, 7.79, 5.62, 5.80,
5.33, 7.50, 8.23, 7.38, 6.42,
5.47, 6.82, 7.94, 7.01, 6.57,
6.10, 6.57, 8.73, 7.90, 6.90,
7.39, 5.44, 7.50, 7.32, 5.36,
7.07, 7.51, 8.88, 7.70, 7.85),ncol=5,byrow=T)
```

```
Xbar=colMeans(x)
```

```
Vhat=var(x)
```

Test Statistics

- If $H_0 : \mu_1 = \mu_2 = \ldots = \mu_d$ holds, then each $\mu_\ell = \frac{1}{d}\sum_{j=1}^{d} \mu_j$
- So, contrast of interest is the *GrandMean* contrast

$$\mathbf{C} = \begin{pmatrix} \frac{d-1}{d} & -\frac{1}{d} & \cdots & -\frac{1}{d} \\ -\frac{1}{d} & \frac{d-1}{d} & \cdots & -\frac{1}{d} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{d} & -\frac{1}{d} & \cdots & \frac{d-1}{d} \end{pmatrix} = \mathbf{I} - \frac{1}{d}\mathbf{J}$$

- $\mathbf{I} = diag(d)$ and $\mathbf{J}$: $d \times d$ matrix of 1's
- Note that $\mathbf{c}'_\ell \boldsymbol{\mu} = \mu_\ell - \frac{1}{d}\sum_{j=1}^{d} \mu_j$
- $\boldsymbol{C}$ is also known as centering matrix

Test Statistics: Wald-Type

- Wald-Type Statistics
  - $W = n(\boldsymbol{C}\overline{\boldsymbol{X}}_{\cdot})' \left[\boldsymbol{C}\widehat{\boldsymbol{V}}\boldsymbol{C}'\right]^{+} \boldsymbol{C}\overline{\boldsymbol{X}}_{\cdot}$
  - $[\mathbf{A}]^{+}$: generalized inverse of a matrix
  - Under H_0, WTS has a $\chi^2_{rank(\boldsymbol{C})}$ distribution
  - When do we reject the hypothesis?

```
library(multcomp)
library(MASS)
C=contrMat(n=rep(10,5),"GrandMean")
CX=C%*%Xbar
CVhat = C%*%Vhat%*%t(C)
W=n*t(CX)%*%ginv(CVhat)%*%CX
pvalue= 1-pchisq(W, d-1)
```

Wald-Type Test: Properties

Advantages

-
-
-
-

Disadvantages

-
-
-
-

ANOVA-Type Statistic

- $\widehat{\boldsymbol{V}}$ causes issues in the WTS, let us remove it

$$\begin{aligned} A_1 &= n(\boldsymbol{C}\overline{\mathbf{X}}_{\cdot})' \left[\boldsymbol{C}\boldsymbol{C}'\right]^{+} \boldsymbol{C}\overline{\mathbf{X}}_{\cdot} \\ &= n\overline{\mathbf{X}}_{\cdot}' \boldsymbol{C}' \left[\boldsymbol{C}\boldsymbol{C}'\right]^{+} \boldsymbol{C}\overline{\mathbf{X}}_{\cdot} \\ &= n\overline{\mathbf{X}}_{\cdot}' \boldsymbol{T}\overline{\mathbf{X}}_{\cdot}. \end{aligned}$$

- The final test is given by

$$\begin{aligned} A &= n\overline{\mathbf{X}}_{\cdot}' \boldsymbol{T}\overline{\mathbf{X}}_{\cdot} / \mathit{Trace}(\boldsymbol{T}\widehat{\boldsymbol{V}}), \\ f &= \frac{\mathit{Trace}(\boldsymbol{T}\widehat{\boldsymbol{V}})^2}{\mathit{Trace}(\boldsymbol{T}\widehat{\boldsymbol{V}}\boldsymbol{T}\widehat{\boldsymbol{V}})} \end{aligned}$$

- Under H_0, ATS has a $F(f, \infty)$ distribution
- When do we reject the hypothesis?

```
TT <- t(C)%*%ginv(C%*%t(C))%*%C
TrTV <-sum(c(diag(TT%*%Vhat)))
A <- n*t(Xbar)%*%TT%*%Xbar/TrTV

TVTV<-TT%*%Vhat%*%TT%*%Vhat
TrTVTV <-sum(c(diag(TVTV)))
f <- TrTV^2/TrTVTV

pvalue <- 1-pf(A,f,10^10)

#10^10=infty, arbitrary high nr#
```

ANOVA-Type Test: Properties

Advantages

-
-
-
-

Disadvantages

-
-
-
-

R Package MANOVA.RM

- R-package *MANOVA.RM*

```
data=data.frame(x=c(x),ID=rep(1:10,5),
dose=sort(rep(1:5,10)))
library(MANOVA.RM)

fit<-RM(x~dose, subject="ID", data=data,no.subf=1)

summary(fit)
```

Split Plot Plan: Introduction

What is a Split Plot Plan?

- A type of experimental design commonly used in life sciences, agricultural and industrial experiments.
- Combines elements of both factorial and randomized block designs.
- Multiple independent groups of subjects are observed over time

Structure:

- **Whole Plot Factors:** Factors that split the subjects in independent groups
- **Subplot Factors:** Factors that stratify the repeated measures

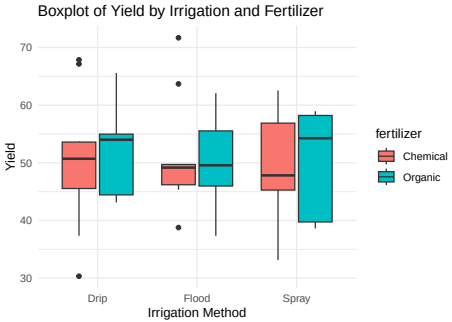
Example:

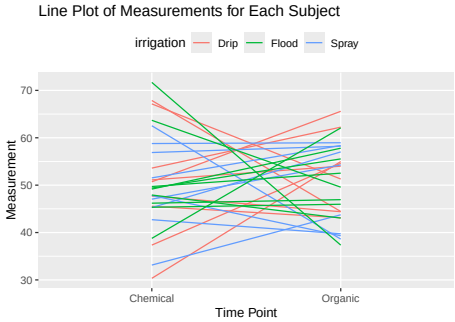
- In agriculture: Testing different irrigation methods (whole plot factor) and different types of fertilizers (subplot factor) on crop yield.

Split Plot Plan: Example

Example:

- Whole Plot Factor: Irrigation Method (3 levels: Drip, Spray, Flood)
- Subplot Factor: Fertilizer Type (2 levels: Organic, Chemical)
- Response: Crop Yield





Split Plot Plan: Interpretation and Output (Mixed Model fit)

```
# Example summary output
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest'
Formula: yield ~ irrigation * fertilizer + (1 | whole_plot)
Random effects:
 Groups     Name        Variance Std.Dev.
 ID (Intercept)  20.04    4.478
 Residual                79.96    8.944
Number of obs: 18, groups:  whole_plot, 9
Fixed effects:
                        Estimate Std. Error       df t value Pr(>|t|)
(Intercept)               48.333      3.162  10.588  15.288 3.35e-08 ***
irrigationSpray           -1.667      4.474   6.000  -0.373    0.719
irrigationFlood           -2.333      4.474   6.000  -0.522    0.618
fertilizerOrganic          1.667      4.474   6.000   0.373    0.719
irrigationSpray:fertilizerOrganic  0.833      6.328   6.000   0.132    0.899
irrigationFlood:fertilizerOrganic  1.833      6.328   6.000   0.290    0.781
```

Split Plot Plan: Interpretation and Output (Mixed Model fit))

```
## Analysis of Variance Table
##                          npar Sum Sq Mean Sq F value
## irrigation                  2 85.622  42.811  0.3325
## fertilizer                  1 30.238  30.238  0.2349
## irrigation:fertilizer       2 67.607  33.804  0.2626
```

Split Plot Plan: Interpretation and Output (Longitudinal Model fit))

```
## ANOVA-Type Statistic (ATS):
##                       Teststatistic  df1     df2     p-value
## irrigation            0.408         1.998  6.446    0.681
## fertilizer            0.198         1.000 Inf       0.656
## irrigation:fertilizer 0.222         1.640 Inf       0.757
```