Experimental Design and Biostatistics (TAMS38)
Lecture 7– Two–level (2^k) factorial designs

Martin Singull

Department of Mathematics
Mathematical Statistics
Linköping University, Sweden

LINKÖPING UNIVERSITY

Content

- Random vectors
  - Expectation
  - Covariance matrix
  - Multivariat Normal distribution
- 2^k-factorial design
  - 2^2-factorial design
  - 2^3-factorial design
  - General, 2^k-factorial design
- Example

Random vector

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)' : n \times 1$ be a random vector (where $\boldsymbol{X}'$ is transpose of $\boldsymbol{X}$), with expectation vector

$$\mathrm{E}(\boldsymbol{X}) = (\mu_1, \mu_2, \ldots, \mu_n)',$$

where $\mu_i = \mathrm{E}(X_i)$ and covariance matrix $\boldsymbol{\Sigma_X} = (\sigma_{ij})_{i,j=1}^n$, where

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

Note that the diagonal elements

$$\sigma_{ii} = \text{cov}(X_i, X_i) = \text{var}(X_i) = \sigma_i^2.$$

The main diagonal of a covariance matrix is thus the variances and outside the main diagonal, we have covariances.

If X_i and X_j are independent then

$$\text{cov}(X_i, X_j) = 0.$$

The reverse is not in general true. It is possible that $\text{cov}(X_i, X_j) = 0$ but X_i and X_j are dependent.

However, under normality it its true in both directions, see Theorem later.

Example

If $X_1, \ldots, X_n$ are independent, $\mathrm{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$ then

$$\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \boldsymbol{I}_n,$$

where $\boldsymbol{I}_n : n \times n$ is the identity matrix.

Linear Transformation

Theorem. If $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ then

$$E(\boldsymbol{Y}) = \boldsymbol{A}\,\mathrm{E}(\boldsymbol{X}) + \boldsymbol{b}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{A}'.$$

Multivariate Normal distribution

Definition. A random vector $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$ is normally distributed if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, where $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ and components $X_1, \ldots, X_n$ are independent and $N(0, 1)$.

The parameters of the multivariate normal distribution is the mean vector and the covariance matrix.

Normally distributed random vectors

Theorem. If $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{Y} + \boldsymbol{g}$ for some matrix $\boldsymbol{F}$ and vector $\boldsymbol{g}$ and for $\boldsymbol{Y}$ being normally distributed vector, then also $\boldsymbol{Z}$ is normally distributed.

Theorem. Components of **normally distributed** vector are independent if and only if the covariance matrix is a diagonal matrix.

Facorial design of type 2^k

With the help of a factorial design we can highlight the impact of different factors of a rather complex phenomenon.

Simultaneously varying several factors we have the possibility to find the optimum level combination.

When you want to investigate whether a response variable is affected by a number of factors begins often with an experiment in which each factor has only two levels.

If we have k factors then we have 2^k level combinations.

Via such an experiment we can determine which factors are interesting and for these factors we make it then a more detailed design, in which each factor can have several levels.

Two factors: 2^2-factorial design

Factor A have two levels coded as -1 and 1.
Factor B have two levels coded as -1 and 1.

-1= "low" level; 1= "high" level.

Model. Level combination A_iB_j gives observations y_{ijk} of the random variable

$$Y_{ijk} = \underbrace{\mu + \tau_i + \beta_j + (\tau\beta)_{ij}}_{=\mu_{ij}} + \varepsilon_{ijk}$$

where $i = -1, 1,\ j = -1, 1,\ k = 1, \ldots, n,\ \varepsilon_{ijk} \sim N(0,\ \sigma)$ and we have constrains

$$\sum_i \tau_i = 0, \sum_j \beta_j = 0, \sum_i (\tau\beta)_{ij} = 0\ \forall j, \sum_j (\tau\beta)_{ij} = 0\ \forall i.$$

Then, we have the matrix of the expectations ($\mathrm{E}(Y_{ijk})$)

	$B = -1$	$B = 1$
$A = -1$	$\mu_{-1,-1} = \mu + \tau_{-1} + \beta_{-1} + (\tau\beta)_{-1,-1}$	$\mu_{-1,1} = \mu + \tau_{-1} + \beta_{1} + (\tau\beta)_{-1,1}$
$A = 1$	$\mu_{1,-1} = \mu + \tau_{1} + \beta_{-1} + (\tau\beta)_{1,-1}$	$\mu_{1,1} = \mu + \tau_{1} + \beta_{1} + (\tau\beta)_{11}$

Given the constraints, we can reduce the number of parameters:

$$\tau_{-1} + \tau_1 = 0 \Rightarrow \tau_{-1} = -\tau_1$$

$$\beta_{-1} + \beta_1 = 0 \Rightarrow \beta_{-1} = -\beta_1$$

$$(\tau\beta)_{-1,1} + (\tau\beta)_{11} = 0 \Rightarrow (\tau\beta)_{-1,1} = -(\tau\beta)_{11}$$

$$(\tau\beta)_{1,-1} + (\tau\beta)_{11} = 0 \Rightarrow (\tau\beta)_{1,-1} = -(\tau\beta)_{11}$$

$$(\tau\beta)_{-1,-1} + (\tau\beta)_{-1,1} = 0 \Rightarrow (\tau\beta)_{-1,-1} = -(\tau\beta)_{-1,1} = (\tau\beta)_{11}$$

We rewrite the expected value matrix using the remaining parameters $\mu, \tau_1, \beta_1, (\tau\beta)_{11}$

	$B = -1$	$B = 1$
$A = -1$	$\mu_{-1,-1} = \mu - \tau_1 - \beta_1 + (\tau\beta)_{11}$	$\mu_{-1,1} = \mu - \tau_1 + \beta_1 - (\tau\beta)_{11}$
$A = 1$	$\mu_{1,-1} = \mu + \tau_1 - \beta_1 - (\tau\beta)_{11}$	$\mu_{1,1} = \mu + \tau_1 + \beta_1 + (\tau\beta)_{11}$

Differences between the expected values in the same column shows the importance of interaction:

$$2\tau_1 - 2(\tau\beta)_{11} \quad \text{or} \quad 2\tau_1 + 2(\tau\beta)_{11}$$

In the additive model $(\tau\beta)_{11} = 0$, in the complete model the interaction is $(\tau\beta)_{11} \neq 0$.

We change the notations for the observations:

A	B	Observations	Average
-1	-1	$y_{(1)1}, \ldots, y_{(1)n}$	$\bar{y}_{(1)\cdot}$
1	-1	$y_{a1}, \ldots, y_{an}$	$\bar{y}_{a\cdot}$
-1	1	$y_{b1}, \ldots, y_{bn}$	$\bar{y}_{b\cdot}$
1	1	$y_{ab1}, \ldots, y_{abn}$	$\bar{y}_{ab\cdot}$

The the complete two level model μ_{ij} is estimated with average of observations in the cell. Hence

$$\begin{aligned}
\bar{y}_{(1)\cdot} &= \hat{\mu} - \hat{\tau}_1 - \hat{\beta}_1 + \widehat{(\tau\beta)}_{11} && (= \hat{\mu}_{-1,-1}) \\
\bar{y}_{a\cdot} &= \hat{\mu} + \hat{\tau}_1 - \hat{\beta}_1 - \widehat{(\tau\beta)}_{11} && (= \hat{\mu}_{1,-1}) \\
\bar{y}_{b\cdot} &= \hat{\mu} - \hat{\tau}_1 + \hat{\beta}_1 - \widehat{(\tau\beta)}_{11} && (= \hat{\mu}_{-1,1}) \\
\bar{y}_{ab\cdot} &= \hat{\mu} + \hat{\tau}_1 + \hat{\beta}_1 + \widehat{(\tau\beta)}_{11} && (= \hat{\mu}_{1,1})
\end{aligned}$$

We can write it as

$$\underbrace{\begin{pmatrix} \bar{y}_{(1)\cdot} \\ \bar{y}_{a\cdot} \\ \bar{y}_{b\cdot} \\ \bar{y}_{ab\cdot} \end{pmatrix}}_{=\bar{\boldsymbol{y}}} = \underbrace{\begin{array}{c} \begin{matrix} I & A & B & AB \end{matrix} \\ \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}}_{=\boldsymbol{F}} \underbrace{\begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\beta}_1 \\ \widehat{(\tau\beta)}_{11} \end{pmatrix}}_{=\hat{\boldsymbol{\xi}}}$$

The matrix $\boldsymbol{F}$ is called a design matrix as values in the columns corresponds for the factor levels of each of the observations.

Note that AB-column $=$ A-column $\times$ B-column where the multiplication is done element-wise.

Matrix $\boldsymbol{F}$ has orthogonal columns. By multiplication of matrix equation above from left by $\boldsymbol{F}'$ we obtain

$$\boldsymbol{F}'\bar{\boldsymbol{y}} = 4\hat{\boldsymbol{\xi}},$$

since $\boldsymbol{F}'\boldsymbol{F} = 4\boldsymbol{I}$ (exercise).

For a 2^2-factorial design we have the following estimator of the parameter vector

$$\hat{\boldsymbol{\xi}} = \frac{1}{4}\boldsymbol{F}'\bar{\boldsymbol{y}},$$

i.e.,

$$\hat{\boldsymbol{\xi}} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\beta}_1 \\ \widehat{(\tau\beta)}_{11} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} \bar{y}_{(1)\cdot} \\ \bar{y}_{a\cdot} \\ \bar{y}_{b\cdot} \\ \bar{y}_{ab\cdot} \end{pmatrix} \quad (1)$$

We have for example

$$\hat{\mu} = \frac{1}{4}(\bar{y}_{(1)\cdot} + \bar{y}_{a\cdot} + \bar{y}_{b\cdot} + \bar{y}_{ab\cdot}) = \bar{y}_{\cdot\cdot}$$

and

$$\hat{\tau}_1 = \frac{1}{4}(-\bar{y}_{(1)\cdot} + \bar{y}_{a\cdot} - \bar{y}_{b\cdot} + \bar{y}_{ab\cdot}).$$

The book chooses to produce effects rather than the estimation of the parameters.

The effect is defined as the difference between high and low levels of factors for example

$$A\text{-effekt} = \tau_1 - \tau_{-1} = 2\tau_1$$
$$AB\text{-effekt} = 2(\tau\beta)_{11}$$

etc.

Now we should show that the estimators given by (1) are unbiased, independent and normally distributed. Let

$$\overline{\boldsymbol{Y}} = \left(\overline{Y}_{(1)\cdot}, \overline{Y}_{a\cdot}, \overline{Y}_{b\cdot}, \overline{Y}_{ab\cdot}\right)'.$$

We know that the random vector $\overline{\boldsymbol{Y}}$ is normally distributed with

$$\mathrm{E}\left(\overline{\boldsymbol{Y}}\right) = \begin{pmatrix} \mu_{-1,-1} \\ \mu_{1,-1} \\ \mu_{-1,1} \\ \mu_{1,1} \end{pmatrix} = \boldsymbol{F} \begin{pmatrix} \mu \\ \tau_1 \\ \beta_1 \\ (\tau\beta)_{11} \end{pmatrix}$$

and that

$$\boldsymbol{\Sigma}_{\overline{\boldsymbol{Y}}} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 & 0 & 0 \\ 0 & \frac{\sigma^2}{n} & 0 & 0 \\ 0 & 0 & \frac{\sigma^2}{n} & 0 \\ 0 & 0 & 0 & \frac{\sigma^2}{n} \end{pmatrix} = \frac{\sigma^2}{n}\boldsymbol{I}.$$

Furthermore, we have the random vectors for the estimator of parameters

$$\hat{\boldsymbol{\xi}} = \frac{1}{4}\boldsymbol{F}'\,\overline{\boldsymbol{Y}}.$$

So the vector $\hat{\boldsymbol{\xi}}$ is normally distributed as it is a linear combination of normally distributed random vector (see theorem above).

Furthermore, the estimator of the parameters are unbiased

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\xi}}) &= \mathrm{E}\left(\frac{1}{4}\boldsymbol{F}'\,\overline{\boldsymbol{Y}}\right) = \frac{1}{4}\boldsymbol{F}'\,\mathrm{E}\left(\overline{\boldsymbol{Y}}\right) = \frac{1}{4}\boldsymbol{F}'\boldsymbol{F}\boldsymbol{\xi} \\ &= \frac{1}{4}4\boldsymbol{I}\boldsymbol{\xi} = \boldsymbol{\xi}\end{aligned}$$

with the covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}} &= \frac{1}{4}\boldsymbol{F}'\boldsymbol{\Sigma}_{\overline{\boldsymbol{Y}}}\frac{1}{4}\boldsymbol{F} = \frac{1}{16}\boldsymbol{F}'\frac{\sigma^2}{n}\boldsymbol{I}\boldsymbol{F} \\ &= \frac{\sigma^2}{16n}\boldsymbol{F}'\boldsymbol{F} = \frac{\sigma^2}{16n}4\boldsymbol{I} = \frac{\sigma^2}{4n}\boldsymbol{I}.\end{aligned}$$

As $\hat{\boldsymbol{\xi}}$ is normally distributed and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}}$ is a diagonal matrix, then the components of $\hat{\boldsymbol{\xi}}$ are independent.

For 2^2-factorial design the random variables $\hat{\mu}$, $\hat{\tau}_1$, $\hat{\beta}_1$, $\widehat{(\tau\beta)}_{11}$ are independent, normally distributed, unbiased with standard deviation $\sigma/\sqrt{4n}$.

Example. So, for example, it holds that

$$\hat{\tau}_1 \sim N\left(\tau_1, \frac{\sigma}{\sqrt{4n}}\right)$$

(and the effect $2\hat{\tau}_1 \sim N\left(2\tau_1, \frac{\sigma}{\sqrt{n}}\right)$).

The sums of squares are given in the same way as in the Two-Way ANOVA, see Lecture 5. THe total sum of squares

$$SS_T = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$

is divided to

$$SS_A = nb \sum \hat{\tau}_i^2 = n \cdot 2 \cdot 2\hat{\tau}_1^2 \quad (df = 1)$$

$$SS_B = na \sum \hat{\beta}_j^2 = n \cdot 2 \cdot 2\hat{\beta}_1^2 \quad (df = 1)$$

$$SS_{AB} = n \sum_i \sum_j \widehat{(\tau\beta)}_{ij}^2 = n \cdot 4 \cdot \widehat{(\tau\beta)}_{11}^2 \quad (df = 1)$$

$$SS_E = (n-1)[s_{(1)}^2 + s_a^2 + s_b^2 + s_{ab}^2] \quad (df = 4(n-1)),$$

where $s_{(1)}^2$ is the sample standard deviation for observations $y_{(1)1}, \ldots, y_{(1)n}$, etc.

Furthermore, the residuals are defined in the usual manner

$$
\begin{aligned}
&y_{(1)k} - \bar{y}_{(1)\cdot} \quad k = 1, \ldots, n \\
&\vdots \\
&y_{abk} - \bar{y}_{ab\cdot} \quad k = 1, \ldots, n
\end{aligned}
$$

Three factors: 2^3-factorial design

Level combination $A_i B_j C_k$ gives observations y_{ijkl} of the random variable

$$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + \\ + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl},$$

where $\varepsilon_{ijkl} \sim N(0, \sigma)$.

Here, $i = -1, 1$, $j = -1, 1$, $k = -1, 1$, $l = 1, \ldots, n$.

We have constrains

$$\sum \tau_i = 0, \quad \sum \beta_j = 0, \quad \sum \gamma_k = 0,$$

$$\sum_i (\tau\beta)_{ij} = 0 \ \forall j, \sum_j (\tau\beta)_{ij} = 0 \ \forall i, \sum_i (\tau\gamma)_{ik} = 0, \ \forall k$$

$$\sum_k (\tau\gamma)_{ik} = 0 \ \forall i, \sum_j (\beta\gamma)_{jk} = 0 \ \forall k, \sum_k (\beta\gamma)_{jk} = 0, \ \forall j$$

$$\sum_i (\tau\beta\gamma)_{ijk} = 0 \ \forall j, k, \sum_j (\tau\beta\gamma)_{ijk} = 0 \ \forall i, k,$$

$$\sum_k (\tau\beta\gamma)_{ijk} = 0 \ \forall i, j.$$

In all of the sums above, we have only two terms.

In the same manner as for 2^2-factorial design we can reduce the amount of the parameters with the help of the constrains.

We choose to keep the parameters

$$\mu, \tau_1, \beta_1, (\tau\beta)_{11}, \gamma_1, (\tau\gamma)_{11}, (\beta\gamma)_{11}, (\tau\beta\gamma)_{111}.$$

We obtain

$$\begin{aligned} \tau_{-1} &= -\tau_1, \ldots, \\ (\tau\beta)_{-1,1} &= -(\tau\beta)_{11},\ (\tau\beta)_{-1,-1} = (\tau\beta)_{11}, \ldots, \\ (\tau\beta\gamma)_{-1,1,1} &= -(\tau\beta\gamma)_{111},\ (\tau\beta\gamma)_{-1,-1,1} = (\tau\beta\gamma)_{111} \end{aligned}$$

etc.

We change the notations for the observations in the same way as done for 2^2-factorial design.

A	B	C	Observations	Average	Expectation
-1	-1	-1	$y_{(1)1}, \ldots, y_{(1)n}$	$\bar{y}_{(1)\cdot}$	$\mu_{-1,-1,-1}$
1	-1	-1	$y_{a1}, \ldots, y_{an}$	$\bar{y}_{a\cdot}$	$\mu_{1,-1,-1}$
-1	1	-1	$y_{b1}, \ldots, y_{bn}$	$\bar{y}_{b\cdot}$	$\mu_{-1,1,-1}$
1	1	-1	$y_{ab1}, \ldots, y_{abn}$	$\bar{y}_{ab\cdot}$	$\mu_{1,1,-1}$
-1	-1	1	$y_{c1}, \ldots, y_{cn}$	$\bar{y}_{c\cdot}$	$\mu_{-1,-1,1}$
1	-1	1	$y_{ac1}, \ldots, y_{acn}$	$\bar{y}_{ac\cdot}$	$\mu_{1,-1,1}$
-1	1	1	$y_{bc1}, \ldots, y_{bcn}$	$\bar{y}_{bc\cdot}$	$\mu_{-1,1,1}$
1	1	1	$y_{abc1}, \ldots, y_{abcn}$	$\bar{y}_{abc\cdot}$	$\mu_{1,1,1}$

Each line represents a level combination, i.e., a cell. For the complete three factor model we estimate the expected value of each cell with the sample mean (average) for the cell, i.e.,

$$\begin{aligned}\bar{y}_{ac\cdot} = \hat{\mu}_{1,-1,1} = \hat{\mu} + \hat{\tau}_1 - \hat{\beta}_1 - \widehat{(\tau\beta)}_{11} + \hat{\gamma}_1 \\ + \widehat{(\tau\gamma)}_{11} - \widehat{(\beta\gamma)}_{11} - \widehat{(\tau\beta\gamma)}_{111}.\end{aligned}$$

Hence,

$$\underbrace{\begin{pmatrix} \bar{y}_{(1)\cdot} \\ \bar{y}_{a\cdot} \\ \bar{y}_{b\cdot} \\ \bar{y}_{ab\cdot} \\ \bar{y}_{c\cdot} \\ \bar{y}_{ac\cdot} \\ \bar{y}_{bc\cdot} \\ \bar{y}_{abc\cdot} \end{pmatrix}}_{=\bar{\boldsymbol{y}}} = \underbrace{\begin{array}{c} \begin{matrix} I & A & B & AB & C & AC & BC & ABC \end{matrix} \\ \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}}_{=\boldsymbol{F}} \underbrace{\begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\beta}_1 \\ \widehat{(\tau\beta)}_{11} \\ \hat{\gamma}_1 \\ \widehat{(\tau\gamma)}_{11} \\ \widehat{(\beta\gamma)}_{11} \\ \widehat{(\tau\beta\gamma)}_{111} \end{pmatrix}}_{=\hat{\boldsymbol{\xi}}}$$

Note that AB-column is the product of A- and B-column etc.

As before, by multiplication from left by $\boldsymbol{F}'$ and dividing with 8 we obtain estimator of the parameter vector

$$\hat{\boldsymbol{\xi}} = \frac{1}{8}\boldsymbol{F}'\bar{\boldsymbol{y}}.$$

In the same way as for 2^2-factorial design we can show that the random vector

$$\hat{\boldsymbol{\xi}} \sim N\left(\boldsymbol{\xi}, \frac{\sigma^2}{8n}\boldsymbol{I}\right)$$

i.e., $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\xi}}} = \frac{\sigma^2}{8n}\boldsymbol{I}$, what implies that all the components $\hat{\boldsymbol{\xi}}$ are independent r.v. with expectation ξ_i and variance $\frac{\sigma^2}{8n}$.

For 2^3-factorial design it holds that the random variables $\hat{\mu}$, $\hat{\tau}_1$, $\hat{\beta}_1$, $\widehat{(\tau\beta)}_{11}$, $\hat{\gamma}_1, \ldots,$ $\widehat{(\tau\beta\gamma)}_{111}$ are independent and normally distributed with the corresponding expectations and the standard deviation $\sigma/\sqrt{8n}$.

Example.

$$\hat{\tau}_1 \sim N\left(\tau_1, \frac{\sigma}{\sqrt{8n}}\right)$$

and the effect is $2\hat{\tau}_1 \sim N\left(2\tau_1, \frac{2\sigma}{\sqrt{8n}}\right)$.

Genreal 2^k-design

For 2^k-factorial design with n observations per cell it holds that the estimator of the vector of parameters

$$\hat{\boldsymbol{\xi}} = \frac{1}{2^k} \boldsymbol{F}' \bar{\boldsymbol{y}}$$

and the estimators of the different parameters are independent, for example the r.v.

$$\hat{\tau}_1 \sim N\left(\tau_1, \frac{\sigma}{\sqrt{2^k n}}\right)$$

and the corresponding holds for all the other estimators.

We consider again the 2^3-factorial design.

The total sum of squares is divided into SS_A, SS_B, SS_C, SS_{AB}, SS_{AC}, SS_{BC} and SS_{ABC} each with one degree of freedom.

$$SS_A = nbc \sum \hat{\tau}_i^2 = n \cdot 2 \cdot 2 \cdot 2\hat{\tau}_1^2 = n \cdot 8\hat{\tau}_1^2$$

$$\vdots$$

$$SS_{AB} = nc \sum_i \sum_j \widehat{(\tau\beta)}_{ij}^2 = n \cdot 8 \cdot \widehat{(\tau\beta)}_{11}^2$$

$$\vdots$$

$$SS_{ABC} = n \sum_i \sum_j \sum_k \widehat{(\tau\beta\gamma)}_{ijk}^2 = n \cdot 8 \cdot \widehat{(\tau\beta\gamma)}_{111}^2$$

and $SS_E = (n-1)[s_{(1)}^2 + s_a^2 + \ldots + s_{bc}^2 + s_{abc}^2]$
with $df = 8(n-1)$.

Testing hypothesis

$H_{0A} : \tau_1 = 0$ vs. $H_{1A} : \tau_1 \neq 0$ is tested with

$v_A = \ldots$ or by doing confidence interval for τ_1.

$H_{0AB} : (\tau\beta)_{11} = 0$ vs. $H_{1AB} : (\tau\beta)_{11} \neq 0$ is tested with

$v_{AB} = \ldots$ etc.

Residuals: $y_{(1)\nu} - \bar{y}_{(1)\cdot} \ldots, y_{abc\nu} - \bar{y}_{abc\cdot}, \nu = 1, \ldots, n.$

If one has, in the 2^k-factorial design, two or more observations for each level combination, then we can do analysis simply by using the Minitab.

In the case of one observation per cell some additional preliminary analysis is needed.

One observation per level combination. Now we have the advantage of the new method to write the parameter estimates.

It is possible to estimate all the parameters in the complete model, but we have no σ^2-estimator.

If a number of the parameters is 0, one can use their corresponding sums of square for the variance estimation.

If it is possible then we can do a "simpler design", i.e., a design where only some part of the factors is included.

To find the important effects we make a normal probability plot for $\hat{\tau}_1, \hat{\beta}_1, \ldots, \widehat{(\tau\beta\gamma)}_{111}$ (for 2^3-factorial design).

The r.v. $\hat{\tau}_1, \hat{\beta}_1, \ldots, \widehat{(\tau\beta\gamma)}_{111}$ are normally distributed with the same variance and varians and should therefore if their expected value are 0 lie on a straight line, see analysis in the example below.

The parameter estimates that fall clearly outside the line are those that have the greatest significance.

Look at the examples 6.3 and 6.4 in the course book. Sometimes we need to transform the data to obtain stability of the variance.

It is good to have a system in which order observations and parameters (effects) listed. Montgomery et al. use the following:

No.	Observations	Parameters	Effects
1	(1)	μ	
2	a	τ_1	A
3	b	β_1	B
4	ab	$(\tau\beta)_{11}$	AB
5	c	γ_1	C
6	ac	$(\tau\gamma)_{11}$	AC
7	bc	$(\beta\gamma)_{11}$	BC
8	abc	$(\tau\beta\gamma)_{111}$	ABC
9	d	δ_1	D
10	ad	$(\tau\delta)_{11}$	AD
11	bd	$(\beta\delta)_{11}$	BD
12	abd	etc	etc
13	cd		
14	acd		
15	bcd		
16	$abcd$		
17	e		
18	ae		
etc	etc		

Remark It is important to distinguish between capital and small letters.

The small letters $a,\ b,\ ab, \ldots$ we use to describe the different levels of the factors that are used for the given observation. For example, level combination ad implies that we have observations for the combination of high level of A and D and low level of the remaining factors.

We use the capital letters to describe the effect and hence to represent columns in the design matrix. The effect AD is the interaction effect between A and D.

Note! A observation that is denoted by ad do **not** measure the interaction effect between A and D. However, the parameter $(\tau\delta)_{11}$ is a measure of the interaction between A and D. Estimate $\widehat{(\tau\delta)}_{11}$ is a linear combination of all measurements.

Example - A Single Replicate of the 2^4 Design

(Example 6.2 from the course book, page 257-260)

A chemical product is produced in a pressure vessel. A factorial experiment is carried out in the pilot plant to study the factors thought to influence the filtration rate of this product.

The four factors are temperature (A), pressure (B), concentration of formaldehyde (C) and stirring rate (D). Each factor is present at two levels.

The design matrix and the response data obtained from a single replicate of the 2^4 experiment are shown below and the 16 runs are made in random order. The process engineer is interested in maximizing the filtration rate.

Current process conditions give filtration rates of around 75 hal/h. The process also currently uses the concentration of formaldehyde, factor C, at the high level.

The engineer would like to reduce the formaldehyde concentration as much as possible but has been unable to do so because it always results in lower filtration rates.

Run Number	Factor				Run Label	Filtration Rate
	A	B	C	D		(gal/h)
1	-	-	-	-	(1)	45
2	+	-	-	-	a	71
3	-	+	-	-	b	48
4	+	+	-	-	ab	65
5	-	-	+	-	c	68
6	+	-	+	-	ac	60
7	-	+	+	-	bc	80
8	+	+	+	-	abc	65
9	-	-	-	+	d	43
10	+	-	-	+	ad	100
11	-	+	-	+	bd	45
12	+	+	-	+	abd	104
13	-	-	+	+	cd	75
14	+	-	+	+	acd	86
15	-	+	+	+	bcd	70
16	+	+	+	+	abcd	96

Design matrix F is placed in columns c1-c16.

```
MTB > Read c1-c16;
SUBC>   File "..\design4.dat";
SUBC>   Decimal ".".
Entering data from file: H: ..\DESIGN4.DAT
16 rows read.
```

y-values are placed in column c17

```
MTB > set c17
DATA> 45 71 48 65 68 60 80 65 43 100 45 104 75 86 70 96
DATA> end
```

Analysis is done using complete model.

```
MTB > copy c1-c16 m1
MTB > copy c17 m2
MTB > trans m1 m3
MTB > mult m3 m2 m4
MTB > copy m4 c18
MTB > let c19 = c18/16
```

Estimates of parameters are placed in column c19

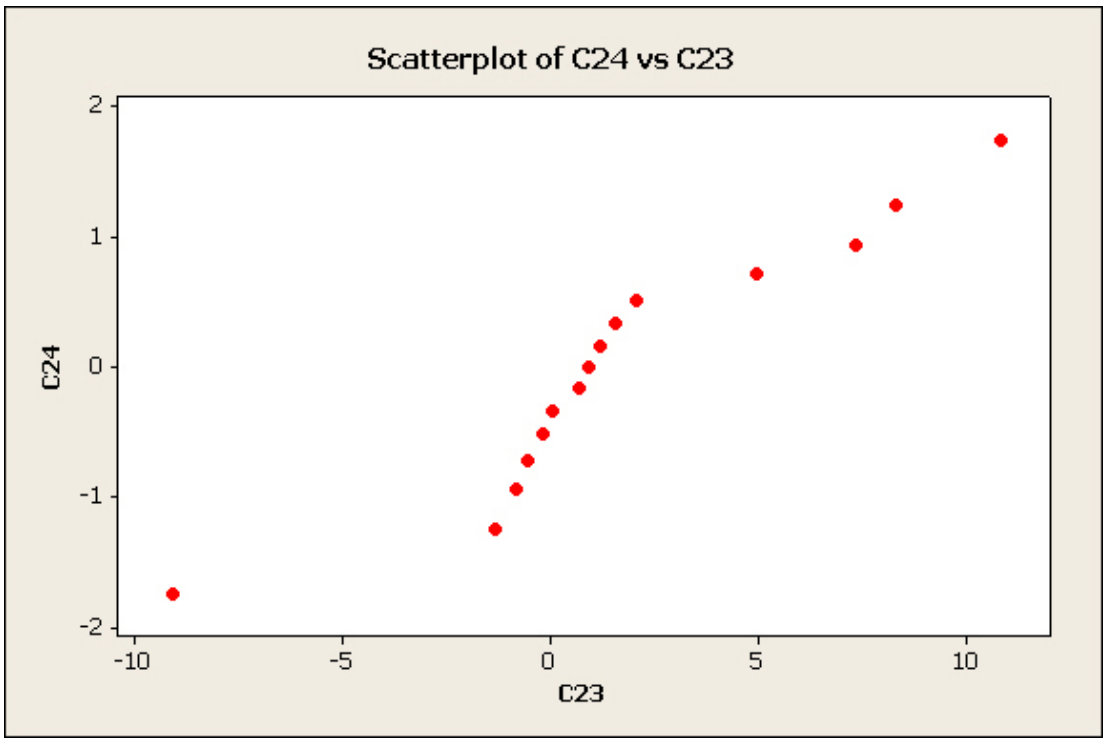
```
MTB > set c20
DATA> 1:16
DATA> end
MTB > sort c19 c20 c21 c22
```

Estimates of the parameters are sorted in the increasing order in column c21

```
MTB > print c21-c22

Data Display

Row       C21  C22
  1  -9.0625    6
  2  -1.3125   15
  3  -0.8125   14
  4  -0.5625   13
  5  -0.1875   11
  6   0.0625    4
  7   0.6875   16
  8   0.9375    8
  9   1.1875    7
 10   1.5625    3
 11   2.0625   12
 12   4.9375    5
 13   7.3125    9
 14   8.3125   10
 15  10.8125    2
 16  70.0625    1
```

Take away $\hat{\mu}$, and do normal probability plot

```
MTB > copy c21 c23;
SUBC> omit 16.
MTB > nscores c23 c24
MTB > plot c24*c23

Scatterplot of C24 vs C23
```



Names A, B, C and D are given for the columns c2, c3, c5 and c9, respectively. We do ANOVA-analysis for the model without B-factor.

```
MTB >  ANOVA 'Y' = A|C|D;
SUBC>   Means A|C|D;
SUBC>   GNormalplot;
SUBC>   NoDGraphs.
```

ANOVA: Y versus A, C, D

Factor	Type	Levels	Values
A	fixed	2	-1, 1
C	fixed	2	-1, 1
D	fixed	2	-1, 1

Analysis of Variance for Y

Source	DF	SS	MS	F	P
A	1	1870.56	1870.56	83.37	0.000
C	1	390.06	390.06	17.38	0.003
D	1	855.56	855.56	38.13	0.000
A*C	1	1314.06	1314.06	58.57	0.000
A*D	1	1105.56	1105.56	49.27	0.000
C*D	1	5.06	5.06	0.23	0.647
A*C*D	1	10.56	10.56	0.47	0.512
Error	8	179.50	22.44		
Total	15	5730.94			

S=4.73682 R-Sq=96.87% R-Sq(adj)=94.13%

Means

A	N	Y
-1	8	59.250
1	8	80.875

C	N	Y
-1	8	65.125
1	8	75.000

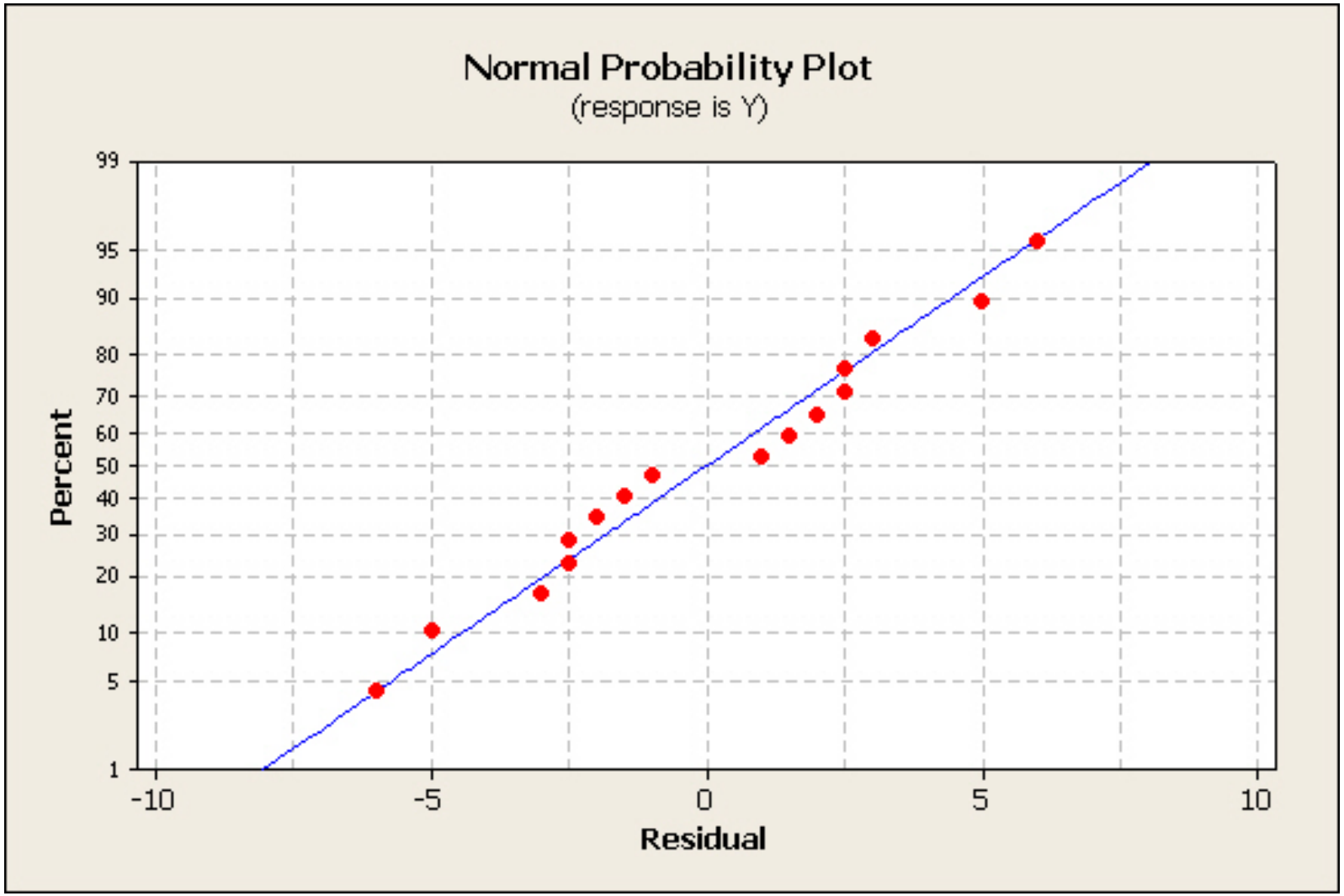
D	N	Y
-1	8	62.750
1	8	77.375

A	C	N	Y
-1	-1	4	45.250
-1	1	4	73.250
1	-1	4	85.000
1	1	4	76.750

A	D	N	Y
-1	-1	4	60.250
-1	1	4	58.250
1	-1	4	65.250
1	1	4	96.500

C	D	N	Y
-1	-1	4	57.250
-1	1	4	73.000
1	-1	4	68.250
1	1	4	81.750

A	C	D	N	Y
-1	-1	-1	2	46.50
-1	-1	1	2	44.00
-1	1	-1	2	74.00
-1	1	1	2	72.50
1	-1	-1	2	68.00
1	-1	1	2	102.00
1	1	-1	2	62.50
1	1	1	2	91.00


Normal Probability Plot
(response is Y)
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-10
-5
0
5
10
Residual

Model in ANOVA-analysis: A-level i, C-level k, D-level l gives

$$Y_{ikl\nu} = \mu_{ikl} + \varepsilon_{ikl\nu},$$

where $\varepsilon_{ikl\nu} \sim N(0, \sigma)$.

Since two different two-factor interaction with A is important, it is best to retain the full three factor model.

To select best combination of levels, we do pairwise comparisons of expected values μ_{ikl}:

$$\begin{aligned} I_{\mu_{ikl}-\mu_{pqr}} &= \left(\bar{y}_{ikl.} - \bar{y}_{pqr.} \mp q_{0.05}(8,8)\frac{s}{\sqrt{2}} \right) \\ &= \left(\bar{y}_{ikl.} - \bar{y}_{pqr.} \mp 5.60\sqrt{\frac{22.44}{2}} \right) \\ &= \left(\bar{y}_{ikl.} - \bar{y}_{pqr.} \mp 18.76 \right), \end{aligned}$$

i.e., $A = 1, C = -1, D = 1$ are significantly better than all apart $A = 1, C = 1, D = 1$, but we prefer $C = -1$.

Hence, we choose $A = 1, C = -1, D = 1$.

Linköping University - Research that makes a difference