Experimental Design and Biostatistics (TAMS38)
Lecture 1 – Introduction and Repetition

Martin Singull

Department of Mathematics
Mathematical Statistics
Linköping University, Sweden

LINKÖPING
UNIVERSITY

Content

- Course - general information
- Example 1-4
- Repetition
  - Models
  - Point estimators
  - Example 5
  - Confidence intervals
  - Hypothesis testing
  - Power of test
  - P-value
- Example 6

Aim TAMS38

The course is intended to give an introduction to the design and analysis of factorial experiments. The emphasis is on

- design of experiments,
- selection of model,
- analysis of observed data,
- ability to interpret the results and
- to draw conclusions.

By the end of the course, the student should be able to:

- design factorial experiments of different types using appropriate randomization;
- choose a suitable model to describe observed data taking into account the design of the experiment that generated the data, then perform an appropriate analysis and draw conclusions by means of hypothesis testing and construction of confidence intervals;
- ...

Organization

- Examinator and lecturer - Martin Singull, martin.singull@liu.se
    - Office: House B, MAI, room: 3A:577
- Course organization,
    - 11 Lectures (22h),
    - 11 Example classes (22h),
        - Group A - Martin Singull
        - Group B - Torkel Erhardsson
    - 6 Computer classes (12h) - **mandatory**.
- Hand in assignments,
    - Three assignments to do in pairs
- Written examination.

`http://courses.mai.liu.se/GU/TAMS38/`

Principles of experimental design

The main principles of experimental design which allows us to collect and then to be able to analyze data properly are:

- **randomization** – the order of individual runs or trials of the experiment is randomly determined (it validates, usually, the assumption about independent distribution of random variables). No or little effect of systematics nuisance factors.
- **replication** – independent repeat of each factor combination
- **blocking** – technique used to improve precision of comparison between interesting us factors. It reduces the variability transmitted from nuisance factors

One should always try to block what is possible and randomize what was not possible to be blocked.

Example 1 – The peanut industry

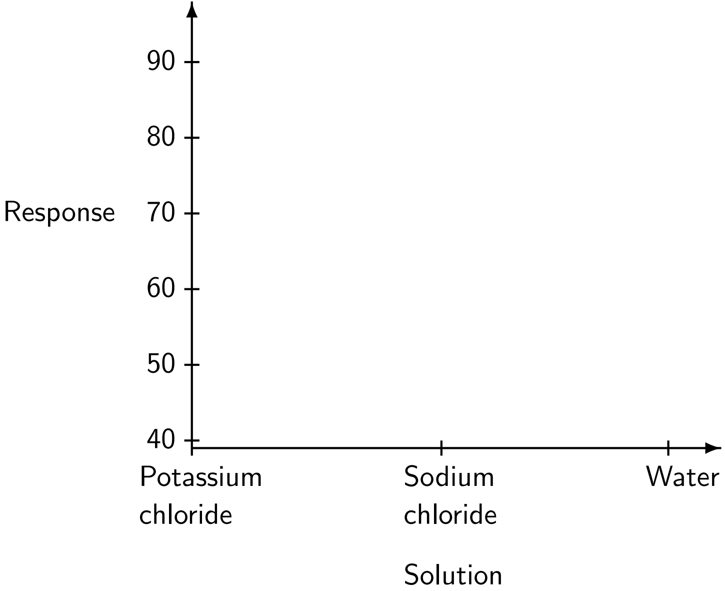
In general one wants to remove the nutshells from the peanuts before use them in the food industry. In many situations it is important that the peanuts are non-effected from the procedure, i.e., that they are whole. To remove the nutshells one heat up the peanuts in an oven and in a solution.

There are three different ovens – convection, microwave and standard. Furthermore, there are three different solutions – potassium chloride, sodium chloride and water. The percentage of non-effected peanuts for each combination are measured.

- How to analyze the data?
- What effect do the different solutions have? The ovens?
- How big are the random errors?

	Oven type		
Solution	Convection	Microwave	Standard
Potassium chloride	64.3	35.6	75.0
	70.0	44.6	76.6
	67.6	40.7	69.0
Cell mean	67.3	40.3	73.5
Sodium chloride	81.4	73.5	86.2
	77.9	64.3	82.1
	83.3	67.4	92.0
Cell mean	80.9	68.4	86.8
Water	71.9	49.6	72.7
	77.7	45.4	76.9
	73.9	53.9	81.5
Cell mean	74.5	49.6	77.7

Interaction plot



Example 2 – Steel industry

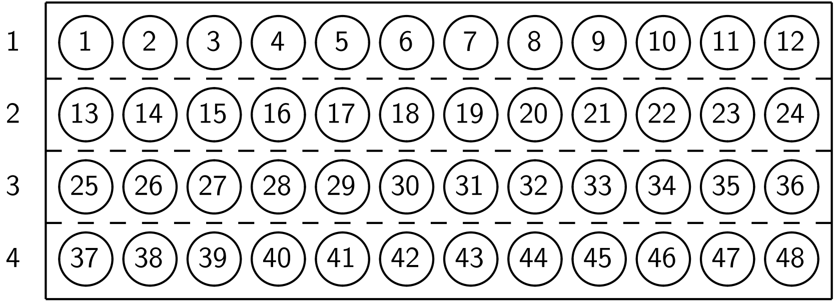
A steel industry make sheet iron with a thin layer of tin for three different can producers. The weight of the tin should be 0.25 pounds for a specified area of the sheet. Both the industry and the producers have laboratories where they test if the tin layer fulfills the weight condition.

Moreover, it is important to test if the four laboratories do comparable measurements.

To test this, one take a sheet and make 48 circular discs out of it and give to the four laboratories. Each laboratory get equal number of discs and each disc can just be analyzed one time.

A sheet with 48 circular discs:

"Band"



1	1	2	3	4	5	6	7	8	9	10	11	12
2	13	14	15	16	17	18	19	20	21	22	23	24
3	25	26	27	28	29	30	31	32	33	34	35	36
4	37	38	39	40	41	42	43	44	45	46	47	48

Complete randomization design

Problem: Which of the 48 discs should be given to the different laboratories?

a) **Complete randomization:** Pick 12 discs at random for Lab A, 12 for Lab B and so on. Result:

A	3,	38,	17,	32,	24,	30,	48,	19,	11,	31,	22,	41
B	44,	20,	15,	25,	45,	4,	14,	5,	39,	7,	40,	34
C	12,	21,	42,	8,	27,	16,	47,	46,	18,	43,	35,	26
D	9,	2,	28,	23,	37,	1,	10,	6,	29,	36,	33,	13

Observations from the four laboratories:

A	B	C	D
0.25	0.18	0.19	0.23
0.27	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28
0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.21	0.29	0.21	0.22
0.28	0.16	0.19	0.21

How does the statistical model look like? How can we analyze the data? Conclusions?

Block design

But, if one suspects that the "bands" on the sheet are different one can use the following design instead.

b) **Block design:** From each block ("band") on the sheet pick three discs for Lab A, three for Lab B, and so on. Result:

	Band			
	1	2	3	4
A	8, 4, 10	23, 24, 19	26, 29, 35	37, 44, 48
B	2, 6, 12	21, 15, 22	34, 33, 32	45, 43, 46
C	1, 5, 11	16, 20, 13	36, 27, 30	41, 38, 47
D	7, 3, 9	17, 18, 14	28, 31,25	39, 40, 42

Statistical model? How to analyze? Conclusions?

Example 3 – Bacteria

One wants to study how seven factors effect the growth of a special bacteria. Each factor has two levels.

If one want to find the best combination of all levels, the one which gives the best growth, one have to do $2^7 = 128$ studies, which sometimes can be to many.

With a reduced *(fractional) factorial design* it can be enough to do $2^3 = 8$ studies if we choose them smart, to analyze the main effects.

- How can we do a design like that?

We will return to this example later.

Example 4 – Effect of music listening during work

The purpose of a study was to see if music during work effected the production in a factory.

Four different music programs A, B, C and D were compared with no music at all E.

Each program was played for a day and one wanted five replicates for each program, i.e., the study lasted for five weeks.

Since there can also be variations between the days and between the weeks the design for the study was chosen to be a *Latin Square*.

Latin Square design

Latin Square design and result:

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	A 133	B 139	C 140	D 140	E 145
2	B 136	C 141	D 143	E 146	A 139
3	C 140	A 138	E 142	B 139	D 139
4	D 129	E 132	A 137	C 136	B 140
5	E 132	D 144	B 143	A 142	C 142

Note that each music program is present one time in each row and in each column.

We will return to that design later.

Remember:

A **design** must always be combined with a **statistical model** for the observations.

In which way the **analysis** should be done depends of course on the statistical model.

Model

We assume that $x_1, \ldots, x_n$ are measures of some physical/ biological/chemical constant.

Model: The measurement x_i is an observation of the random variable

$$X_i = \mu + \varepsilon_i,$$

where

μ =real value of the constant

ε_i =stochastic variable that describes variation (error).

We assume often that $\varepsilon_i \sim N(0, \sigma)$ (some books uses the notation $\varepsilon_i \sim N(0, \sigma^2)$). Hence,

$$X_i \sim N(\mu, \sigma).$$

Sample and random sample

Let $x_1, ..., x_n$ be observations of the r.v. $X_1, ..., X_n$, with probability function $p(k; \theta)$ or density function $f(x; \theta)$ that includes an unknown parameter θ.

We call $x_1, x_2, \ldots, x_n$ a sample that corresponds to the random variables $X_1, X_2, \ldots, X_n$ which is called a random sample.

We want to approximate value of θ, i.e., find a point estimate using $x_1, ..., x_n$.

Point estimator

Definition. A **point estimate** is a function of observations, so

$$\hat{\theta} = g(x_1, ..., x_n).$$

- θ = real thoretical value (fixed number)
- $\hat{\theta}$ = approximative value of θ calculated with observations $x_1, ..., x_n$ (fixed number)
- $\widehat{\Theta}$ = random variable, which describe the variation of $\hat{\theta}$ for different observations.

Some properties:

Definition. A point estimator is an **unbiased** estimator of θ if

$$\mathrm{E}(\widehat{\Theta}) = \mathrm{E}[g(X_1, \dots, X_n)] = \theta.$$

Definition. Suppose that $\widehat{\Theta}$ and $\widetilde{\Theta}$ are two unbiased estimators of θ. If

$$\text{var}(\widehat{\Theta}) \le \text{var}(\widetilde{\Theta})$$

the $\widehat{\Theta}$ is said to be **more efficient** than $\widetilde{\Theta}$.

Definition. Suppose that $\widehat{\Theta}_n$ is an unbiased estimator of θ and that it is defined for all sample sizes n. If for every $\varepsilon > 0$ we have

$$\text{P}(|\widehat{\Theta}_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty,$$

then $\widehat{\Theta}_n$ is a **consistent** estimator of θ.

Theorem. If $\text{E}(\widehat{\Theta}_n) = \theta$ and $\text{var}(\widehat{\Theta}_n) \to 0$ as $n \to \infty$, then $\widehat{\Theta}_n$ is a consistent estimator of θ..

We can often find a proper point estimator utilizing the fact that we can estimate $\mathrm{E}(X_i)$ using the arithmetic mean

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i.$$

There are also theoretical methods to find estimators, e.g., method of moments, maximum likelihood method, least squares, etc.

Example 5

If $x_1, x_2, \ldots, x_n$ are independent observations of r.v. $X_i \sim N(\mu, \sigma)$ then the mean μ and the variance σ^2 can be estimated by

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

What can we say about $\hat{\mu}$?

More about the random variables can be found in *"Egenskaper hos normal-, χ^2-, t- och F-fördelningarna"*.

Confidence intervals

Let $x_1, \ldots, x_n$ be observations of the r.v. following the same distribution, that includes one unknown parameter θ.

Definition. An interval $I_\theta = (a_1(x_1, \ldots, x_n)\ ,\ a_2(x_1, \ldots, x_n))$ such that

$$P(a_1(X_1, \ldots, X_n) \leq \theta \leq a_2(X_1, \ldots, X_n)) = 1 - \alpha$$

is the **confidence interval** for θ with **confidence level** $1 - \alpha$.

Suitable functions a_1 and a_2 can be often found by starting from an estimation variable $\widehat{\Theta}$ and constructing a help variable, that includes θ, but whose distribution is otherwise completely known.

The confidence interval gives those values of θ that are probable with respect to the observed values.

In the case of normally distributed variables the help variable is often t- or χ^2-distributed, see ”*Egenskaper hos normal-, χ^2-, t- och F-fördelningarna*”.

Example 5, cont.

Hypothesis testing

One wants to test **null hypothesis** H_0 (e.g., $H_0 : \theta = 0$) against a **alternative hypothesis** H_1 (e.g., $H_1 : \theta > 0$).

As an alternative hypothesis H_1 one generally selects the hypothesis you want to show is true.

To test H_0 against H_1 we should construct **test statistics** $t(x_1, \dots, x_n)$ en a **critical region** C such that

$$H_0 \text{ is rejected if } t(x_1, \dots, x_n) \in C$$

on significance level α.

The critical region C is chosen so that the test statistics may tend to end up in C, if H_1 is true.

The critical region is related to the **significance level**

$$\begin{aligned}\alpha &= \mathrm{P}(H_0 \text{ is rejected if } H_0 \text{ is true}) \\ &= \mathrm{P}(t(X_1, \dots, X_n) \in C \text{ if } H_0 \text{ is true}).\end{aligned}$$

1) *If $t \in C$, then H_0 should be rejected in favor to H_1.*
   Conclusion: H_1 holds.
2) *If $t \notin C$, then H_0 cannot be rejected.*
   Conclusion: H_0 can be true but does not have to be.

Error of type I: Reject H_0 when it is true (significance level).
Error of type II: Not reject (to accept) H_0 when it is not true.

Example 5, cont.

Power of the test

Sometimes we are interested in the special value of θ_1 in the alternative. One wants to have a high probability to reject H_0 if θ_1 is the true value.

Definition. The **power** of the test for $\theta = \theta_1$ is

$$\begin{aligned} 1 - \beta &= \mathrm{P}(H_0 \text{ is rejected when } \theta = \theta_1) \\ &= \mathrm{P}(t(X_1, \dots, X_n) \in C \text{ then } \theta = \theta_1), \end{aligned}$$

Hence, risk of the error of type II is

$$\begin{aligned} \beta &= \mathrm{P}(H_0 \text{ is not rejected when } \theta = \theta_1) \\ &= \mathrm{P}(t(X_1, \dots, X_n) \notin C \text{ when } \theta = \theta_1). \end{aligned}$$

Example 5, cont.

p-value

Instead of finding the critical region C we utilize, sometimes, so-called p-value, which is a measure of the extreme value of the test statistics.

We assume that the test statistics $t(x_1, \ldots, x_n)$ is a increasing function in $\hat{\theta}$ and has an observed value $t_{obs} = t(x_1, \ldots, x_n)$.

a) If the alternative is H_1': $\theta > \theta_0$, then we reject H_0 for large values of test statistics, i.e., if $t(x_1, \ldots, x_n) > a$ where a is given by

$$\alpha = \mathrm{P}(t(X_1, \ldots, X_n) > a \text{ if } H_0 \text{ is true}).$$

Instead of the critical value *a* the following probability can be calculated

$$p = \mathrm{P}(t(X_1, \ldots, X_n) > t_{obs} \text{ if } H_0 \text{ is true}),$$

where

$$t_{obs} > a \iff p < \alpha.$$

b) If alternative hypothesis is H_1: $\theta \neq \theta_0$, then we reject H_0 if $t(x_1, \dots, x_n) < b_1$ or $t(x_1, \dots, x_n) > b_2$.

"Large" value of t_{obs}:

$$\frac{p}{2} = \mathrm{P}(t(X_1, \dots, X_n) > t_{obs} \text{ if } H_0 \text{ is true}).$$

"Small" value of t_{obs}:

$$\frac{p}{2} = \mathrm{P}(t(X_1, \dots, X_n) < t_{obs} \text{ if } H_0 \text{ is true}).$$

c) The case with H_1'': $\theta < \theta_0$ is analogical to a). (Exercise)

Alternative decision rule:

H_0 is rejected if and only if $p < \alpha$.

Hence, if p-value is small, the result is not consistent with the null hypothesis. A large p-value suggests, however compliance with the null hypothesis.

Example 5, cont.

Example 6 - Comparison of groups

Children can suffer from lead poisoning in the case they, for some reason, eat lead-containing substances. One possible explanation is that children suffering from calcium deficiency.

In a study the 20 male rats were used and randomly divided into two groups. The control group received standard diet with sufficient calcium content, while the treatment group received calcium diet. The rats were staying in separate cages and had in the treatment period, access to lead acetate food. The amount of lead acetate consumed was measured for each rat. Results in one volume are the following:

Control group:
5.4 6.2 3.1 3.8 6.5 5.8 6.4 4.5 4.9 4.0
Treatment group:
8.8 9.5 10.6 9.6 7.5 6.9 7.4 6.5 10.5 8.3

Model: We have observations $x_1, \dots, x_{10}$ and $y_1, \dots, y_{10}$ of independent r.v. $X_i \sim N(\mu_1, \sigma_1)$ and $Y_j \sim N(\mu_2, \sigma_2)$.

Further, we have

$$\bar{x} = 5.06, \quad s_1 = 1.189, \quad \bar{y} = 8.56, \quad s_2 = 1.471.$$

a) Is it reasonable to assume the same variance, i.e., $\sigma_1 = \sigma_2 = \sigma$, with significance level $\alpha = 10\%$?

b) Can one say with 90% certainty that rats given calcium diet consume significantly more lead acetate solution than those given normal diet?

Linköping University - Research that makes a difference

LINKÖPING
UNIVERSITY