ESE

8. Experiment Planning

Credits

Experimentation in Software Engineering: An Introduction
by Claes Wohlin, Per Runeson, Martin Host, Magnus C. Ohlsson, Bjorn Regnell, and Anders Wesslén
Springer-Verlag, 2005 (Formerly printed by Kluwer Academic Press, 2000).

Planning a Controlled Experiment

- Hypothesis formulation
- Variables selection
  - Independent variables
    - (Desired) Factors
      - Treatments, Parameters
    - Noises
  - Dependent Variables
- Experiment design
  - Replication
  - Randomization
  - Blocking
  - Balancing
  - Standard design types
- Instrumentation
- Threats to validity evaluation

Hypotheses formulation

Null hypothesis

H_0 : There is no underlining trend or pattern in the experiment setting.

Usually, H_0 is the statement that experimenters want to reject.

E.g., H_0: $\mu_1 == \mu_2$

"==" is preferred, because it is easier to evaluate than other relation operators.

Serious scientists should make all acts to prove that *H_0* is not to reject.

Hypotheses formulation

Alternative hypothesis

H_1 : There is a significant trend or pattern in the experiment setting.

Usually, the statement in favor of which H_0 is rejected.

E.g.:

1) H_1: $\mu_1 <> \mu_2$

2) H_1: $\mu_1 < \mu_2$

Often but not necessarily, 2) is the H_1 in industry.

Hypotheses formulation: Note

!H_0 **!** $\Leftrightarrow$ H_1

If, with respect to data from a given experiment, you cannot prove that H_0 is reasonable true, this does not mean that H_1 is reasonable true.

Recall that we do not work in an axiomatic system where !(=) $\Leftrightarrow$ <> .

See a next chapter on What is an hypothesis, hypotheses testing, and related risks.

Hypotheses formulation: Example

$H_{0EC} : \mu_{1EC} = \mu 2_{EC}$

$H_{1EC} : \mu_{1EC} <> \mu 2_{EC}$

H_0(Efficiency): CR and FTI perform insignificantly different *in detecting defects.*

H_1(Efficiency): CR and FTI perform significantly different in detecting defects.

Hypotheses formulation: Example

$H0_{ES} : \mu1_{ES} = \mu2_{ES}$

$H1_{ES} : \mu1_{ES} <> \mu2_{ES}$

H0(Effectiveness): CR and FTI perform insignificantly different in detecting defects.

H1(Effectiveness): CR and FTI perform significantly different in detecting defects.

Hypotheses formulation: Refinement.

Examples of Evolution of the Knowledge

1) H_1: μ_1 <> μ_2
2) H_1: $\mu_1 < \mu_2$
3) H_1: $\mu_1 >= 1{,}2*\mu_2$
4) H_1: $\mu_1 >= 1{,}2*\mu_2 + 0{,}5$
5) H_1: $\mu_1 =< 1{,}2*\mu_2^{0,1}$
6) H_1: $\mu_1 = \alpha*\mu_2^{\beta} + \chi$

…

Variables selection (1/2)

Dependent and independent variables must be chosen before designing the experiment.

- ✓ Choice of independent variables
- ✓ Choice of response variables

Usually, response variables come first (from goals)

Once we have chosen those variables, and levels and treatments, experiment strategy and design can be defined.

Variables selection

Usually we cannot afford all the input variables and all their treatments in one time. In order to manage the experiment complexity, we proceed by step-wise-refinement:

- ✓ Identify and verify independent variables
  - ✓ (1st experiment).
- ✓ Reduce the number of significant variables
  - ✓ Choose design factor[s] as that [those] input variable[s] that more than other ones affect[s] outcomes; choose parameters, and identify disturbs and noises.
- ✓ Reduce the number of treatments
  - ✓ For each design factor choose ("*Constant effects model*" or "*Fixed effects model*") or select at random ("*Random effects model*" or "*Variance components model*") a few (e.g. 2) of its alternatives.
- ✓ Reduce the number of blocks
  - ✓ From complete to incomplete "blocked" "factorial" design.

Context selection

Usually we cannot afford to start an experiment by using the most realistic objects and professional subjects, intruding the real processes, and looking for the most general solutions. Anyway, we can characterize the experiment context as in the remaining. Additionally, we could have hybrid situations.

- Offline vs. online
- Student vs. professionals
- Toy vs. real problems
- Specific vs. general

Subject Selection

Generally we select a sample of n subjects ("sample size") from a population.

In order to generalize the results to the desired population, the selection must be representative for that population.

The population should be studied carefully before taking a sample.

Defining the Population

We define our population:

- Experience-based: Based on some given materials in advance, or some certain inputs.
- Environment-based: Based on being affected by a certain environment and conditions.

Subject Selection as Probability Sample

If the probability of selecting each subject is known then the selection is a probability sample.

Probability Samples

Simple random sampling: n subjects are selected from a list of the population at random.

Systematic sampling: The first subject is selected from the list at random, hence the following n-1 ones are selected.

Convenience sampling: The nearest and most convenient people are selected as subjects.

Stratified random sampling: The population is divided in a number of groups or strata with a known distribution between the groups. Random selection is then applied within the strata.

Quota sampling: This type of sampling is used to get subjects from various strata of a population. Hence, Convenient sampling is normally used for each strata.

Experimental Design: Basic Principia

1. Replication
2. Randomization
3. Blocking
4. Balancing

Replication

Replication (*of the basic experiment*) means to repeat the basic experiment, i.e. the minimum set of the elementary experiments. [See also *Replication of an experiment.*]

Replication allows us:

1) to estimate the *experimental error* (hence the significance of a result),

2) to estimate more precisely the *impact of a factor on the output*: e.g., by diminishing the variance of the *sample mean* with respect to the variance of the *single observation*.

(See impact on errors in Ch. 10)

Randomization

Randomization is the mile stone that allows using *statistics* in experimentation (errors are assumed independent and random variables).

- It allows to average on factors that may otherwise show their presence.
- It is used in allocating objects, subjects, and test order.

When complete randomization is not possible to enact, we have to adopt specific planning statistic methods.

Blocks

- They allow to improve the precision from comparison between *interesting* factors.
- They also concern disturbing factors, *known and predictable and (quite) controllable factors* that probably have an effect, but we are not interested in that effect.
  - We block with respect to such a factor, when we arrange the experiment in a way that in a block that factor is constant.
  - *We are not expected to study effects between blocks.*

Balancing

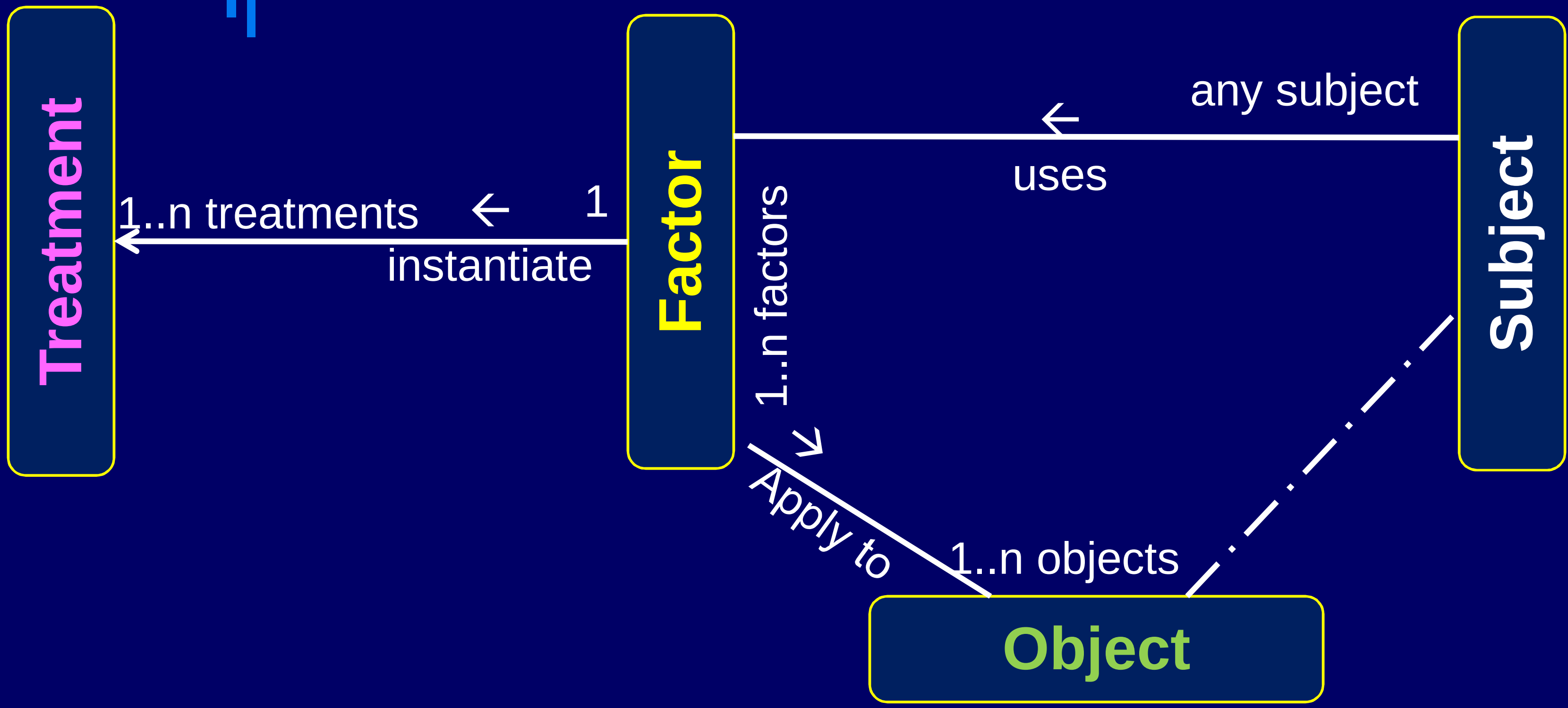
Balancing subjects in an experiment means to assign the treatments so that each treatment has the same number of subjects.

In this case, the statistical analysis is simplified and strengthen.

Usage of Statistical vs. Non-statistical Techniques

- Use also non-statistical knowledge about the problem.
- Adopt experimental plans and data analysis as simple as possible.
- Distinguish between practical and statistical significance.
- In order to improve knowledge, experiments are iteratively realized.

Experiment Design



Q: Does each “n” equal 1, 2 or more?