


Experimentation in Software Engineering

Oporto May, 2003

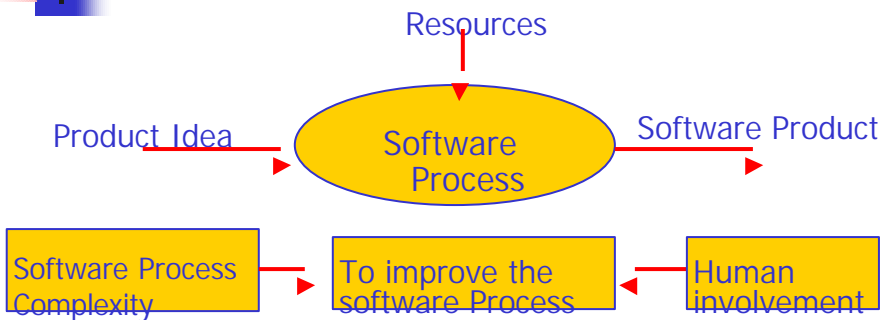
- 
1. Introduction
 2. Type of Studies
 - 2.1 Experimental Studies
 3. Experimentation Process
 - 3.1 Definition
 - 3.2 Planing
 - 3.3 Operation
 - 3.4 Analysis and Interpretation
 - 3.5 Conclusions
 4. Conclusions
-

1. Introduction

- "Software Engineering means application of systematic, disciplined, quantifiable approach to development, operation and maintenance of software" [IEEE90]

- Software Process
- Systematic and disciplined approach
- Quantification

1. Introduction



"Experimentation provides a systematic, disciplined, quantifiable and controlled way of evaluating human-based activities" Wholin 2000

Experimentation in Software Engineering

- Zelkowitz (1997) conclusions over 612 papers:
 - The 30% of papers did not include experimentation and they needed it (20% in other sciences)
 - Only the 10% of papers that include experimentation have controlled experimentation methods
- Tichy (1995) conclusions over 400 paper:
 - The 40% of papers did not include experimentation and they required empirical validation

Experimentation in Software Engineering

- ¿Why in software engineering a lot of asserts aren't validated?
 - It is a new science
 - They need to obtain good quantitative data to make validations, but it often is hard

The way that can convert software engineering claims into validated facts it is the experimental method

¿ Why Software Engineer don't use Experimentation?

Scientific method is not suitable	The software engineers have to observe the phenomenon, to formulate hypothesis and to contrast them
The level of experimentation is enough	The software engineers don't contrast their claims as much as other scientist
The experiments are expensive	It is possible to do a significant experiment that is not expensive
The shows are enough	The shows don't prove nothing
The technology changes speedily	If yesterday you said an important claim that today is not important, that is because it does not well defined

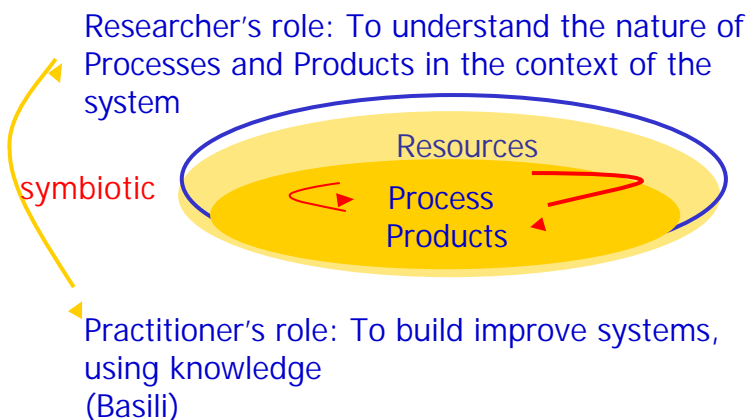
¿ Why Software Engineer don't use Experimentation?

- The software engineers think that
 - The scientific method is not necessary in software engineering
 - ¿How testing the ideas against real world?
- There is not a background of statistical knowledge, so it is very difficult to design an experiment or to analyse the experimental results
- There is not enough culture and bibliography about empirical software engineering

¿Why Software Engineer don't use Experimentation?

- The experimentation in Software Engineering is more difficult than in other sciences, because it is necessary a lot of variables
 - ¿It is a valid raison?
- To publish a experimental study of Software Engineering is more difficult than in other sciences. Furthermore, the empirical studies that are replications era not as important as new studies.
 - But other sciences have two sides: Theory and Practice and both are related

Software Engineering A Laboratory Science



In Conclusion

- We know that Software Nature:
 - It is **development** not production
 - The discipline technologies are **human-based**
 - There are a large number of **variables** that cause differences →
 ¿How measure their effects?

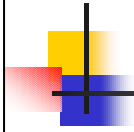
Software Engineering needs more experimentation:

- **To Confirm Theories** and "Conventional Wisdom"
 - ¿To limit McCabe's cyclomatic measure assure quality?
- **To Explore Relationships**
 - ¿How does the design complexity affect the productivity of the designers?
- **To Evaluate the accuracy of Models**
 - ¿Does the PF predict how large the code may be?
- **To Validate Measures**
 - ¿Is the number of methods a valide measure of class complexity?

2. Research Methods

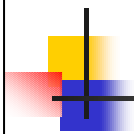
- **The analytic method**
 - Propose a formal theory or set of axioms
 - Develop and derive results
 - If possible, verify the results empirically
- **The engineering and empirical methods (scientific method)**
 - Observing the world
 - Proposing a model or other solutions
 - Measuring and analysing
 - Validating or invalidating the proposed model

Chemistry
Phisique
Mathematiques
Software E.



Research Paradigms

- Quantitative Research
 - Controlled measurement
 - Objective
 - Verification oriented
- Qualitative Research
 - Naturalistic and incontrolled observations
 - Subjective
 - Discovery oriented
- Study
 - An act to test a hypothesis or discover something
 - Can include quantitative and qualitative research



Research Paradigms

Qualitative Research

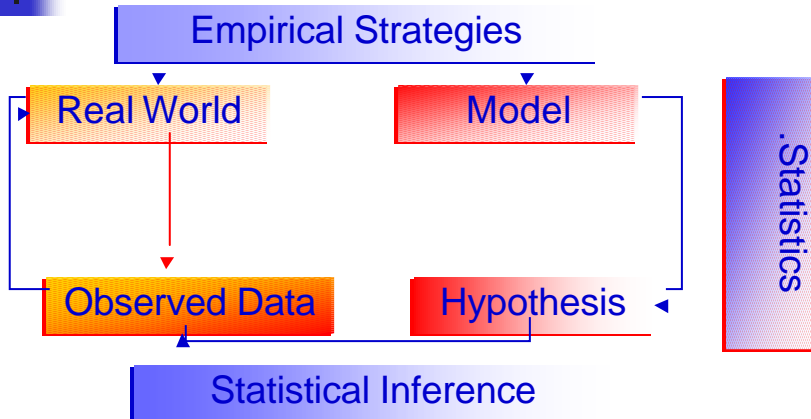
- ❖ We study, using a meeting, the reason because increase the productivity when a team have used a new language.
- ❖ This would be a Qualitative study about thinks like programs logic and human reasoning.
- ❖ The analysis will be about the words which can be organized in order to the researcher can test, compare, analyze and identify patrons.

Research Paradigms

Quantitative Research

- ❖ We study, using a quantitative study, the reason because increase the productivity when a team have used a new language.
- ❖ I must to define the hypothesis, to plan the process, to select the independent and dependent variables, and to control extraneous factors.
- ❖ The analysis will be about the numeric values observed as result of experiment execution, using statistical techniques to test the hypothesis.

Empirical Strategies



Empirical Strategies

First Level: Hypothesis (Model) → Controlled experiments in laboratory, with replication possibility

Second Level: Hypothesis (Model) → in a real environment, using observational studies (case studies)

Third Level: Model applied in all real process, we must made a historical file (surveys). In the futur we have to test the Model with this file.

Empirical Strategies

Depending of the degree of control over data

■ Survey

■ Interviews

■ Questionnaires

■ Retrospective

■ The team are

■ Descriptive

■ Explicative

■ Exploratory

■ Case Study

■ Data collection

■ To avoid confounded factors

■ Statistical Analysis

■ Conclusions

■ Generalization is difficult

■ Observational

■ With little control

■ The team are

■ To compare

■ To establish relationship

■ In a specific time-space

Empirical Strategies

Depending of the degree of control over data

■ Experiment

■ Is a Process

- Statistical Analysis
- It is possible replication
 - To confirm
 - To generalize

■ Controlled Process

■ The team are

- To Confirm Theories and "Conventional Wisdom"
- To Explore Relationships
- To Evaluate the accuracy of Models
- To Validate Measures

Empirical Strategies

Exploratory Survey

Why the developers think that a technique A is better than other B?

Case study (Relationship)

We want to build a model to predict the number of faults in testing, in a enterprise

Experiment

We want to compare two inspection methods, in a laboratory environment, that is, selecting variables and controlling extraneous factors,

Empirical Strategies Factors

- **Execution Control**
How much the researcher control the studie?
- **Measurement Control**
The degree to wich the researcher can decide upon wich measures to be collected
¿In a survey?
- **Investigation Cost**
related with the factors above
- **Easy Replication**
involves repeating the investigation under identical conditions, in another population

Empirical Strategies Comparison

Factor	Survey	Case Study	Experiment
Execution Control	No	No	Yes
Measurement Control	No	Yes	Yes
Investigation Cost	Low	Medium	High
Easy Application	High	Low	High

Experimental Studies (Another Classification)

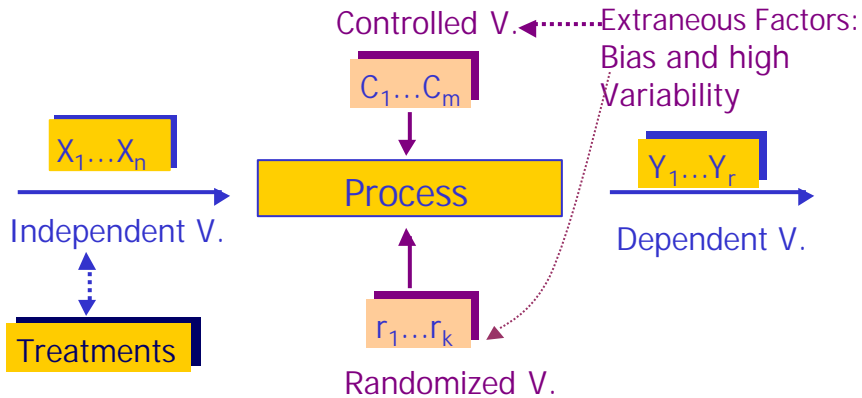
Driven by hypothesis

- Controlled experiment
 - To demonstrate feasibility in small
- Quasi-experiments
 - To simulate the effects of the treatment variables in a realistic environment

Observational Studies (Another Classification)

Driven by understanding	Variable Scopes	
# of sites	A priori defined Deductions: Mathematical formal logic	No a priori defined Deduction: verbal propositions
One	Case study	Case qualitative study
More than one	Field study	Field qualitative study

Controlled Experiment



Basic Concepts

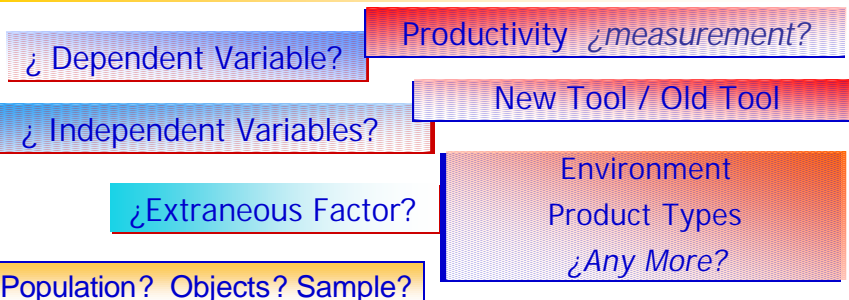
- Independent variables (factor, state, predictand)
 - Which we can control and change in the experiment
 - Dependent variables (response, predictor)
 - They measure the effect of the treatments and appear in the Hypothesis test
 - Controlled variables
 - They can be controlled by the design
 - Randomized variables
 - They are considered as random error in the design
 - Confounded variables
 - They aren't controlled and change together with a independent variable
- To convert in

Basic Concepts

- Treatment: each combination of the levels of different independent variables. If there is only one, each level will be a treatment.
- Population of subjects: we can generalize the results over the population
- Sample: subjects selected from the population (¿subjects selection?→ planning)
- Objects: objects of the study: products, process, resources, models, etc. (Is a part of the Goal definition template)
- Experiment: set of trials (treatment + subject + object)

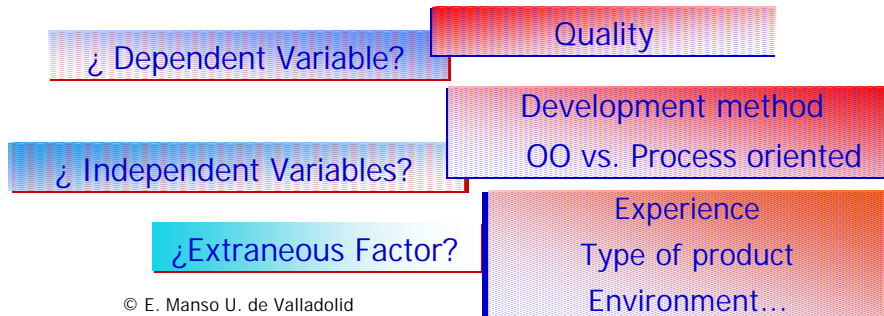
Experiment Example

Analyze a new design tool and a old design tool, for the purpose of to compare their impact with respect to productivity, from the point of view of developers, in the context of the university students.



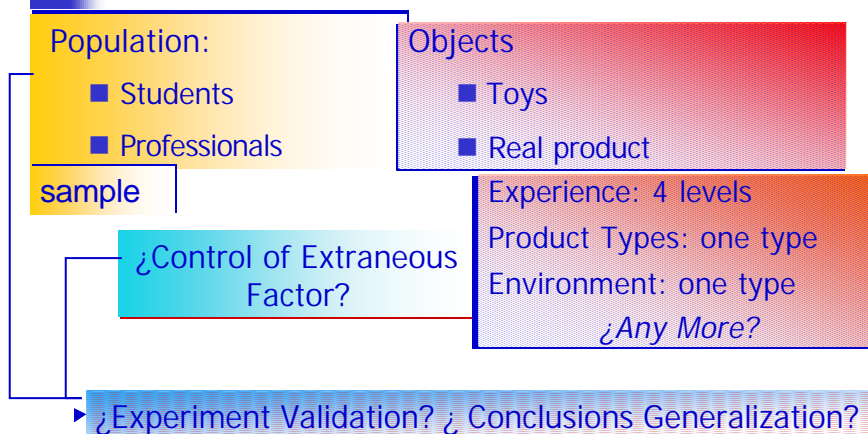
Experiment Example

Analyze the object oriented design method vs. process method, for the purpose of to evaluate with respect to quality, from the point of view of developers, in the context of the university students.



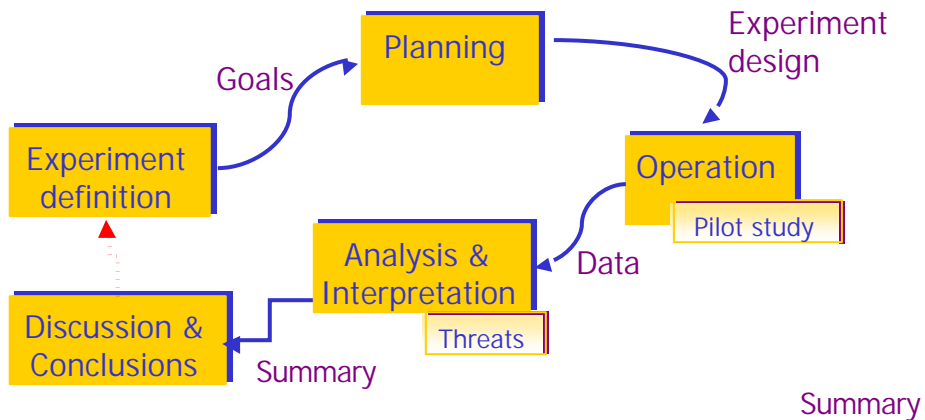
© E. Manso U. de Valladolid

Experiment



© E. Manso U. de Valladolid

3. Experimentation Process



3.1 Experiment Definition ¿Why?

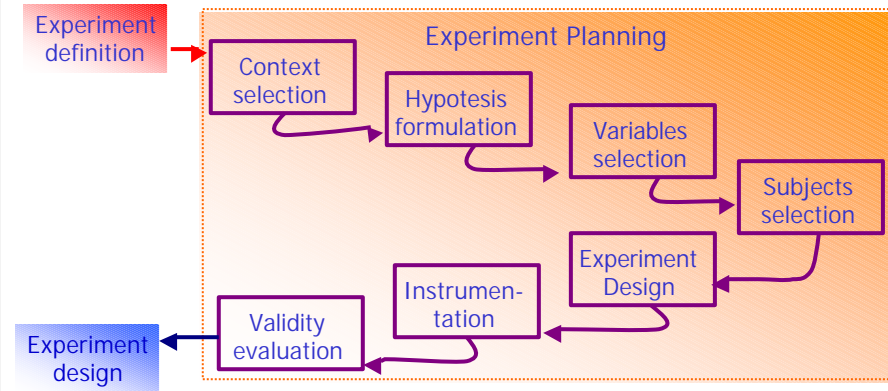
Goal definition template

- Analyze <object of the study>
- For the purpose of <purpose>
- With respect to <quality focus>
- From the point of view of <perspective>
- In the context of <context>

Experiment definition:

- The PBR and checklist techniques
- Evaluation
- Effectiveness and efficiency
- The researcher
- M. Sc and Ph. D students

3.2 Experiment Planning ¿How?



Context Selection

Off-line vs. On-line

Reduces the risk
Produces extra costs

Students vs.
Professional

Reduces the costs
Easier to control
Context generalization?

Toy vs. Real
problem

Reduces the costs & time
Context generalization?

Specific vs.
General

Reduces the costs & time
Context generalization?...

Experiment Context Characterization

Characterization (Basili)		#Objects	
		One	More than one
# subjects per object	One	Single object study	Multi-object variation study
	More than one	Multi-test within object study	Blocked subject - object study

Experiment Context Conclusion

C1. *"Be sure to specify as much of the industrial context as possible. In particular, clearly define the entities, attributes, and measures that are capturing the contextual information"*

It is necessary in

- Observational and
- Experimental studies

C2. *"If a specific hypothesis is being tested, state it clearly prior to performing the study and discuss the theory from which it is derived, so that its implication are apparent"*

Experiment Context Conclusion

C3. "If the research is exploratory, state clearly and, prior to data analysis what questions the investigation is intended to address and how it will address them"

C4. "Describe research that is similar to, or has a bearing on, the current research and how current work relates it"

Hypothesis Formulation

Derived from Experiment definition: one or more H_0

Goal definition template

- Analyze The PBR and checklist techniques(CKL)
- For the purpose of Evaluation With respect to efficiency and effectiveness
- From the point of view of The researcher
- In the context of M. Sc and Ph. D students

▶ H_{01} : PBR efficiency = CKL efficiency

▶ H_{02} : PBR effectiveness = CKL effectiveness

Hypothesis Formulation

H_0 : The observed vehicle is a car

H_1 : The observed vehicle is not a car →

Critical Area (C.A.) = {#wheels ≥ 5 or #wheels ≤ 3 }

If we observe 3 or less wheels or 5 or more wheels we reject H_0 → ¿error?

$\alpha = p(\text{number of wheels} \neq 4 / \text{car})$

If we observe 4 wheels we don't reject H_0 → ¿error?

$\beta = p(\text{number of wheels} = 4 / \text{not car})$

Hypothesis Testing

Derived from Experiment definition: one or more H_0

H_0 : Null Hypothesis (Conservative, there is no treatment effect)

H_1 : Alternative Hypothesis → Critical Area (C.A.)

We decide...	Really H_0 is true	H_1 is true
	$1 - \alpha$	β Error = $P(\emptyset \text{C.A.} / H_1)$
Non reject H_0 (Non significant result)		
Reject H_0 (Significant result)	α Error (significance level) = $P(\text{C.A.} / H_0)$	Test Power = $P(\text{C.A.} / H_1)$

Hypothesis Testing

H_0 : Null Hypothesis

We need to select a "random measure" (m) of the effect of treatment: the estimate

- Time to understand a document
- Percentage of defects detected in a document

Parametric Test → the distribution pattern of m is known

- Time is $N(\mu, \sigma)$
- Percentage is $B(n, p)$ (aprox. $N(p, (p \cdot (1-p))^{1/2})$)

■ Non Parametric Test → the distribution pattern of m is acknowledged

Hypothesis Test: Performance

1. To define Hypothesis H_0 and H_1
2. To select the suitable estimate
3. To determine the error α (usually 0,05 or 0,01)
4. Using 1, 2 and 3 to determine the Critical Area (C.A.)
5. Using n , H_0 and H_1 , and the C.A. to determine b
6. To reject H_0 or not from the observed (estimation)
7. value of estimate

$\alpha_1 = 0.05$	→	$\beta = \beta_1$
$\alpha_2 = 0.10$	→	$\beta = \beta_2 < \beta_1$
$\alpha_3 = 0.01$	→	$\beta = \beta_3 > \beta_1$

Hypothesis Testing

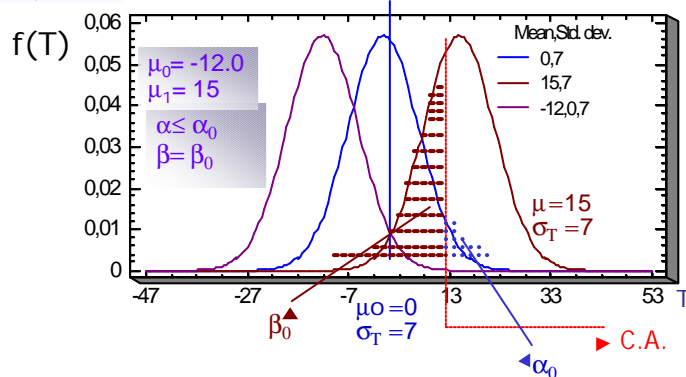
- $1 - \beta$ (Test Power): probability that the test will reveal a true pattern if H_0 is false
 - The pattern when H_0 is false can be unknown \rightarrow ¿ $1 - \beta$?
- We should choose a test with as high power as possible (increasing n , for example)
- $1 - \beta$ depends on α , sample size (n) and effect size
- $1 - \beta$ is better when we have test parametric

Hypothesis Testing

$T = \bar{X}$ = estimate with known pattern $N(\mu, \sigma_T)$ when H_0 is true

$H_0: \mu = 0 \quad H_1: \mu > \mu_0$

$H_0: \mu \leq 0$



Hypothesis Testing

Conclusions

D1. "Identify the population from which the subjects and objects are drawn"

D2. "Define the process by which the subjects and objects were selected"

- The conclusion may be useful if the sample are representative
- We must to exclude the students with a lot of experience in the experiment. They are not representative.

Hypothesis Testing

Conclusions

D3. "Define the process by subjects and objects are assigned to treatments"

D4. "Restrict yourself to simple study designs or, at least, to designs that are fully analyses in the statistical literature. If you are not using a well-documented design and analysis method, you should consult a statistician to see whether yours is the most effective design for what you want to accomplish"

Hypothesis Testing

Conclusions

D5. "Define the experiment unit"

- If you are evaluating teams but you get measures from each team member ¿what it is the experimental unit? → team

D6. "For formal experiments, perform a pre-experiment or precalculation to identify or estimate the minimum required sample size"

- The sample size determine the test power

Variables Selection

- Independent variables
 - Which we can control and change in the experiment
- Dependent variables
 - They measure the effect of the treatments and appear in the Hypothesis test
- Controlled variables
 - They can be controlled by the design
- Randomized variables
 - They are considered as random error in the design
- Confounded variables
 - They aren't controlled and change together with a independent variable

To convert
in

Subjects Selection

- ¿How to select the subjects?
 - Can be probability or non-probability
 - Simple random sampling, systematic sampling ...
 - Convenience sampling, quota sampling ...
- ¿Size of the sample?
 - If there is a large variability, a larger size we need

The Sample from the Population must be representative

Experiment Design Choice

- The experiment design define trials organization
- Is related with the analysis, interpretation and conclusions of the experiment

Relevant
Questions

- ¿How many independent variables are there?
 - Only one → Simple experiments
 - More than one → Factorial experiments

Repeated measures

- ¿How many treatments per subject?

Blocking Randomization

- ¿How “to control” extraneous factors?

Crossed design
Nested design

- ¿How “to combine” the independent variables levels? → # treatments

The answers will depend on the validity **Threats** we want to control

General Design Principles

Randomization Blocking Balancing

Randomization is used to

- Assure the observations are from independent random variables
- Allocate objects, subjects and in which order the test are performed
- Average out the effect of a extraneous factor

Blocking

- Blocking subjects is used to eliminate the undesired effect in the comparison among the treatments of a extraneous factor that we are not interested in
 - Within a block the **undesired effect** is the same, and we can study the effect of treatments on that block
- Blocking increases the precision of the experiment
- Blocking treatments is used to reduce de amount of treatments for subject

General Design Principles

Randomization Blocking Balancing

Balancing

- The number of subjects per treatment is the same
- It is not necessary, but is desirable from the point of view of statistical analysis of the data.

Treatment ₁	Treatment ₂	Treatment ₃	Treatment ₄
Subject ₃	Subject ₅	Subject ₄	Subject ₆
Subject ₈	Subject ₂	Subject ₁	Subject ₁₁
Subject ₁₀	Subject ₇	Subject ₁₃	Subject ₁₄
Subject ₁₅	Subject ₁₆	Subject ₉	Subject ₁₂

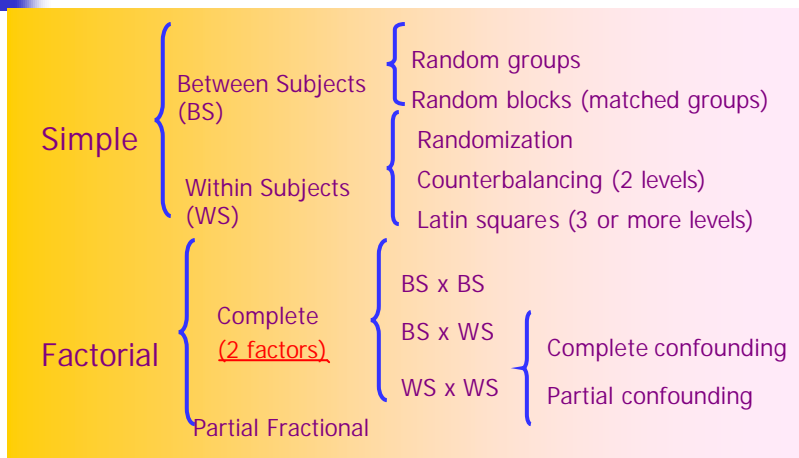
Randomized and Balanced

General Design Principles

The principal Claims of the experiment design are:

- To reduce the variability
- To control extraneous factors
- To reduce the different threats to experiment validity as much as possible

Experiment Design A Taxonomy



Simple Design Between Subjects Characteristics

- Each subject has only one Treatment
- Threats to internal validity: **Selection is the principal threat**, it is the effect of natural variation in human performance.
- To avoid this threat:
 - **Randomization**: the subjects are assigned to the treatment randomly
 - **Blocking**: We have subject in each block with the same value in the blocked variable. We assign randomly all treatments in each block

Simple Design Between Subjects Statistical Hypothesis

The most common is to compare the means of the dependent variable for each treatment

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

Notation:

\bar{a} the grand mean

\bar{m}_i the mean of the dependent variable for treatment i (the effect of treatment i)

y_{ij} the j th measure of the dependent variable for treatment i

Model: $y_{ij} = \bar{a} + \bar{m}_i + e_{ij}$

parameters

Error: Random variable

Simple Between Subjects Statistical Hypothesis

With two treatments

Example of hypothesis

$$H_0: \mu_A = \mu_B \quad H_1: \mu_A \neq \mu_B \quad (\text{or } H_1: \mu_A > \mu_B)$$

Example of Analysis:

$\bar{X}_A - \bar{X}_B$ estimate with known pattern $N(0, \sigma)$ / $H_0 \rightarrow$ t-test

If the estimate has unknown pattern \rightarrow Mann Whitney-test

With k treatments

Example of hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad H_1: \neg H_0$

Example of Analysis:

ANOVA (ANalysis Of VAriance) if the variables pattern is $N(\mu, \sigma)$

Kruskall Wallis (non-parametric test)

Simple Between Subjects Example: Blocking

Subjects Experience	Subjects (32) assigned to groups A and B Matched groups							
[6.4, 28.4)	10B	14B	19A	15B	37A	27A	74B	11A
[28.4, 45)	24B	64B	4B	1A	55A	80B	33A	83A
[45, 62)	42B	45A	71A	13A	77A	17B	66B	50B
[62, 95.8)	2B	49A	25A	84B	29B	18B	69A	75A

Experiment about documentation and maintainability relation
(Tryggeseth, 1997)

If we have small size a randomized design is not adequate

¿Balanced design? $\rightarrow \# A = \# B$

Simple B.S. Blocking Design

Example Cartwright, 1998

- The experiment was a replication of an experiment previously conducted at other university:
 - To investigate the impact of class inheritance upon the maintenance of C++ software
- The subjects had to make the same maintenance change to one of two versions of a C++ program
 - The first version was implemented using inheritance, the second version had no inheritance

Simple B.S. Blocking Design

Example Cartwright, 1998

- Dependent variable: Completion Time, in minutes, to modify a database program
- Treatments: version flat vs. Version with inheritance
 - $E(\text{Time/flat}) = \mu_{\text{flat}}$ $E(\text{Time/inheritance}) = \mu_{\text{inh}}$
- Hypothesis for time:
 - Ho: That 3 levels of inheritance has no impact upon time to make a correct maintenance change as compared with no inheritance
 - H1: \neg Ho $\alpha = 0.05$
 - Ho: $m_{\text{inh}} = m_{\text{flat}}$ $H_1: m_{\text{inh}} \neq m_{\text{flat}} \rightarrow \text{T-statistic}$

Simple B.S. Blocking Design

Example Cartwright, 1998

- Dependent variable: size of maintenance change
- Treatments: version flat vs. Version with inheritance
 - $E(\text{Time/flat}) = \mu_{\text{flat}} \quad E(\text{Time/inheritance}) = \mu_{\text{inh}}$
- Hypothesis for size of maintenance:
 - Ho: That 3 levels of inheritance has no impact upon size of a correct maintenance change as compared with no inheritance
 - H1: $\neg \text{Ho} \quad \alpha = 0.05$
 - Ho: $m_{\text{inh}} = m_{\text{flat}} \quad H_1 : m_{\text{inh}} \neq m_{\text{flat}} \rightarrow \text{T-statistic}$**

Simple B.S. Randomized Design

Example: one factor with more than two levels

- Dependent variable: quality of software
- Treatments: programming languages C, C++ and JAVA
- Hypothesis
 - Ho: These 3 programming languages has no impact upon quality of software
 - H1: $\neg \text{Ho} \quad \alpha = 0.05$
 - Ho: $\mu_C = \mu_{C++} = \mu_{\text{JAVA}} \quad H1: \neg \text{Ho} \rightarrow \text{ANOVA}$

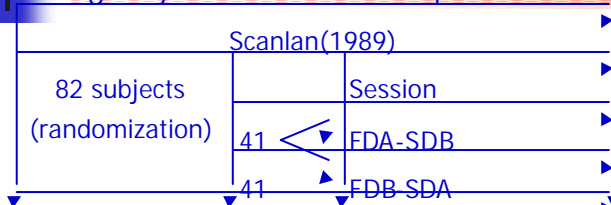
- Independent variable: ratio, interval or absolute scale
- SPSS: the treatments have to have numeric codification
- Analysis and interpretation

Simple Design Within Subjects Characteristics

- Each subject is a block: he uses all treatments, so we have repeated measures
- We need to resolve threats to internal validity:
 - Maturation (boring, learning...)
 - Instrumentation
 - Mortality
- Treatment Order (Practice) is controlled with
 - Randomization
 - Counterbalancing and Latin Square, that permit measure the practice effect, as a independent variable

Simple Within Subjects Randomization: Characteristics

Are you interested in the practice effect? NO



- Levels of independent variable(treatments): pseudocode (SD) and flow diagram (FD) Dependent variable : understandability
- Pattern of objects: A and B → to avoid maturation
Algorithms: simple, medium and complex
- **The sequence of the 6 objects is random** ▶ Practice

So we can control Maturation, Instrumentation and Practice

Simple Within Subjects

Randomization: Statistical Hypothesis

ANOVA

Notation:

m_i the mean of the dependent variable for treatment i (the main effect of treatment i)

b_j the main effect of subject j

y_{ij} the measure of the dependent variable for treatment i on subject j

Model: $y_{ij} = a + m_i + b_j + e_{ij}$ ← Error: Random variable

To compare the means of the dependent variable for each treatment

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

Simple Within Subjects

Counterbalancing: Characteristics

¿Are you interested in the practice effect (order)? YES

- The order of treatments(A,B) to each subject will be ABBA

Incomplete counterbalancing

Group G1 with AB

Group G2, similar to G1, with BA

How can we have two "similar" groups?

1. Thinking about extraneous variables that can influence in the dependent variable
2. Blocking, Randomization

Simple Within Subjects

Counterbalancing: Statistical Hypothesis

ANOVA Notation:

m_i the mean of the dependent variable for treatment i (the effect of treatment i)

b_j the effect of group (order)

$g_{j(k)}$ the main effect of subject k of group j

y_{ijk} the measure of the dependent variable for treatment i on subject k of group j

ANOVA Model: $y_{ij} = a + m_i + b_j + g_{j(k)} + e_{ijk}$

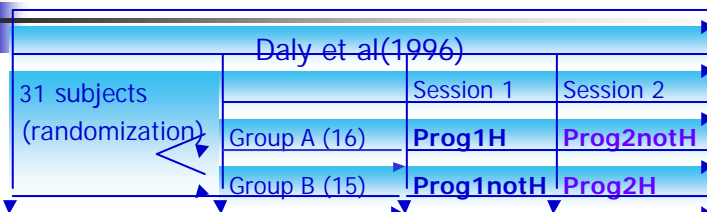
Error: Random variable

Hot: $\mu_1 = \mu_2 = \dots = \mu_k$

Hog: $\beta_1 = \beta_2 = \dots = \beta_k$

Simple W.S Counterbalancing

Example



- Levels of independent variable(treatments): with inherit (H) without inherit (notH)
- The subjects was paired by programming OO skill, and then assigned randomly to Group A and Group B
- **Instrumentation and maturation are confounded with session**, so we can not measure them
- **We can measure the practice effect (order)**

Simple Within Subjects

Latin Square: Characteristics

If we have more than 2 treatments (K) ¿How many "sequences" will have in a counterbalancing design? **K!**

Latin Square design reduce the effort selecting a sequences subgroup of the K!

- We have to select as sequences as treatments number (K)
- Each treatment has a different position per sequence

A possibility with 3 treatments X Y Z

Session 1	Session 2	Session 3	
X	Y	Z	Group A
Y	Z	X	Group B
Z	X	Y	Group C

Simple Within Subjects

Latin Square: Statistical Hypothesis

ANOVA Notation:

m_i the mean of the dependent variable for treatment i
(the main effect of treatment i)

b_j the main effect of group (order)

$g_{j(k)}$ the main effect of subject k of group j

y_{ijk} the measure of the dependent variable for treatment i on subject k of group j

ANOVA Model: $y_{ijk} = a + m_i + b_j + g_{j(k)} + e_{ijk}$

Error: Random variable

Hot: $\mu_1 = \mu_2 = \dots = \mu_k$

Hog: $\beta_1 = \beta_2 = \dots = \beta_k$

Factorial Experiment Characteristics

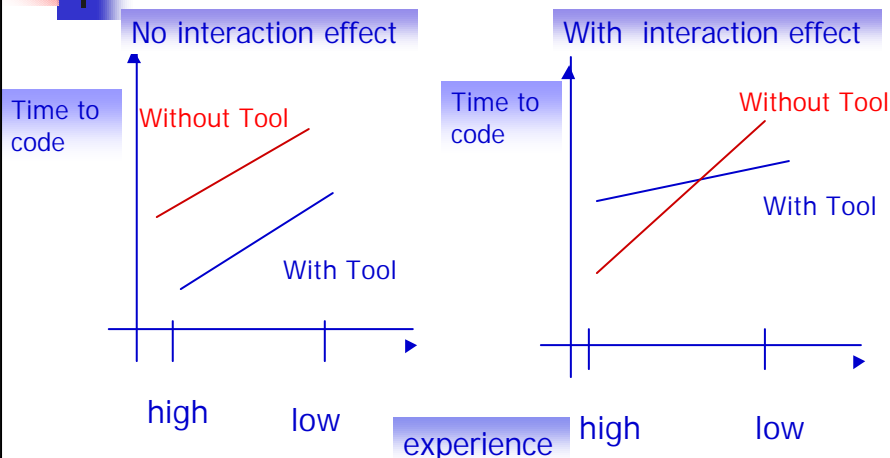
Why we have to choose Factorial experiment?

- If the absence of a second (Or third or...) variable can affect performance in the first variable (in the others variables)

Example: You are interested in the effects of a new design tool on productivity. This tool may be used differently by designers who are experts in object-oriented design from those who are new to o-o design.

- If you design a simple experiment randomized or blocking, you would get an incomplete or incorrect view of the tool effects.

Factorial Experiment Characteristics



Factorial Experiment Characteristics

There are 2 independent variables (i.v.) A and B

- We have $k_1 * k_2$ treatment, if A has k_1 levels and B has k_2 levels

There are r independent variables (i.v.) $A_1 .. A_r$

- We have $k_1 * k_2 * ... * k_r$ treatments, if each A_i has K_i levels

I need to include 3 factors which have 4, 3 and 2 levels

¿How many treatments do we have?

¿How many subjects?

Factorial Experiment Characteristics

Problems that we need to resolve:

1. What factors should be included?
 - Do we include experience as a factor?
2. How many levels of each factor?
 - The percentiles can be a guide
3. How should the levels of the factor be spaced?
 - Time, exam results, age...
4. How should the experimental units (subjects) be selected?
 - Randomization? Blocking?

Factorial Experiment Characteristics

Problems that we need to resolve:

5. How many subjects should be selected for each treatment?
 - This is related with the test power
6. What steps should be taken to control experimental error?
 - Control of extraneous factors
7. What criterion measures should be used to evaluate the effects of the treatment factor?
 - Do we consider interaction effects?
 - Do we consider higher-order interaction effects?

Factorial Experiment Two Factors: Characteristics

Do we want to consider interaction effects?

Crossed design		Factor B	
		b1	b2
Factor A	a1	a1 b1	a1 b2
	a2	a2 b1	a2 b2

We can study A*B interaction

Nested Design			
Factor A			
a1		a2	
Factor B		Factor B	
b1	b2	b1	b2
a1 b1	a1 b2	a2 b1	a2 b2

We can not study A*B interaction

Factorial Experiment

Two Factors: Statistical Hypothesis

ANOVA Notation crossed design:

m_i the mean of the dependent variable A for treatment i (the main effect of treatment A)

b_j the main effect of treatment B_j

g_{ij} the interaction effect of treatments A_i B_j

y_{ijk} the measure of the dependent variable for subject k on A_i B_j treatment

ANOVA Model: $y_{ijk} = a + m_i + b_j + g_{ij} + e_{ijk}$

Error: Random variable

$$H_{0A}: \mu_1 = \mu_2 = \dots = \mu_k \quad H_{0B}: \beta_1 = \beta_2 = \dots = \beta_k$$

$$H_{0AB}: \gamma_{ij} = \gamma \quad \forall i, j$$

Complete Factorial Experiment

Two Factors: Example

Random Block x Random Block (Finney et al 1998)

Dependent variable: comprehension of the written specifications (z-formal language)

Independent variables: comments (yes or no) and significant names (yes or no)

Two factorial design		Significant names	
		0	1
Comments	0	A	B
	1	C	D

Four specification versions

Control of the extraneous factor: Blocks C1..C6 (147 students from 6 study types)

C_i → random assigned to A, B, C and D

Factorial Experiment

Two Factors: Example

repeated measures x repeated measures in blocks

(*Complete confusion*) Basili et al 1997

Dependent variable: defect detection rate

Independent variables: types of documents (ATM, PG) and reading techniques (USUAL, Perspective-Based Reading)

Two factorial design	Session 1	Session 2	12 subjects: random assignation to two blocks of treatments
Group 1	USUAL/ATM	PBR/PG	
Group 2	USUAL/PG	PBR/ATM	<i>Complete confusion</i> of interaction effect with group The main effects are within-block effects

Factorial Experiment

Two Factors in Blocks: Statistical Hypothesis

ANOVA Notation

m_i the mean of the dependent variable A for treatment i (the main effect of treatment A_i)

b_j the main effect of treatment B_j

g_{ij} the interaction effect of treatments $A_i B_j$ totally confounded with the group main effect I_k

$p_{m(k)}$ the subject main effect, nested in group k

y_{ijk} the measure of the dependent variable for subject k on $A_i B_j$ treatment

ANOVA Model: $y_{ijk} = a + m_i + b_j + p_{m(k)} + g_{ij} + e_{ijk}$

Error:
Random variable

$H_{0A}: \mu_1 = \mu_2 = \dots = \mu_k$ $H_{0B}: \beta_1 = \beta_2 = \dots = \beta_k$

$H_{0AB}: \gamma_{ij} = \gamma \forall i, j$

Factorial Experiment

Two Factors: Example

repeated measures x repeated measures in blocks

Solution 2

Dependent variable: defect detection rate

Independent variables: types of documents (ATM, PG) and reading techniques (USUAL, Perspective-Based Reading)

Two factorial design	Session 1	Session 2
Group 1	USUAL/ATM	USUAL/PG
Group 2	PBR/PG	PBR/ATM

12 subjects: random assignation to two blocks of treatments

Complete confusion of ϵ ?
effect with group?

Instrumentation

The Instrumentation provides means for

- Performing the experiment
- To monitor it

The experiment results shall be the same independently of the instrumentation

Objects	To choose appropriated objects (specifications, code documents...)
Guidelines	The participants need to be guided in the experiment (process description, checklist...) Additionally training
Measurement instruments	Data collection via manual forms, interviews etc. that must be validated

Validity Evaluation

- Internal Validity
 - ¿Does the treatment cause the effect?
- Conclusion validity
 - If you measure a phenomenon twice, the outcome shall be the same
- Construct validity
 - ¿The selected variables reflect the construct of the cause and the effect well?
- External validity
 - ¿Can the results be generalized outside of our scope?

Threats to Internal Validity

History	Different treatments applied to the same object at different times... ¿are the circumstances the same?
Maturation	The subjects react differently as time passes (tired, bored, learning)
Selection	¿Is the sample representative for the whole population? It is the effect of natural variation in human performance. Volunteers are more motivated...
Instrumentation	¿Are the artefacts used for experiment execution designed correctly? Documents to be inspected ...
Mortality	Persons who drop out from the experiment
Treatment Order	¿How much know the subject about the treatment?

Threats to Conclusion Validity

Low statistical power

The ability of the test to reveal a true pattern if H_0 is false ¿ $1-\beta = 0,3?$

Violated assumptions of statistical test

Some statistical test are more robust than others ¿ $Y \rightarrow N(\mu, \sigma)?$

Fishing and error rate

Three investigations with $\alpha = 0,05 \rightarrow (1-0,05)^3 = 0,14$

Reliability of measures

Objective measures are better than subjective ¿LDC or PFA?

Reliability of treatment implementation

The application of treatments to subjects must be standardized

Random irrelevancies in experimental setting

Random heterogeneity of subjects

Threats to Construct Validity

Inadequate preoperational explication of constructs

The method A is better than B ¿Does it mean?

~~Mono-operation bias: The experiment under-represent the construct~~

Confounding constructs and levels of constructs: The difference depends on if the subjects have 1,3 or 5 years of language experience

Interaction of different treatments: We cannot conclude whether the effect is due to either of treatments or of a combination of treatments

Interaction of testing and treatment: the application of treatments can make the subjects more receptive to the treatment

Restricted generalizability across constructs: The treatment improved the productivity, but ¿What about the maintainability?

Hypothesis guessing: The subjects base their behaviour on their guesses about the hypothesis

Experiment expectancies of the subjects can bias the results. ¿Solution?

Threats to External Validity

Interaction of selection and treatment

Sample not representative of the population

We select only programmers in an inspection experiment

Interaction of setting and treatment

Material not representative

Toy problems, methods old-fashioned

Interaction of history and treatment

The experiment is conducted in a special time which affects the results

Priority of Validity Threats

There is a conflict between some of the types of validity threats

- The subjects measures several factors, which increase the construct validity but there is a risk about conclusion validity → tedious measurements affect the reliability of measures

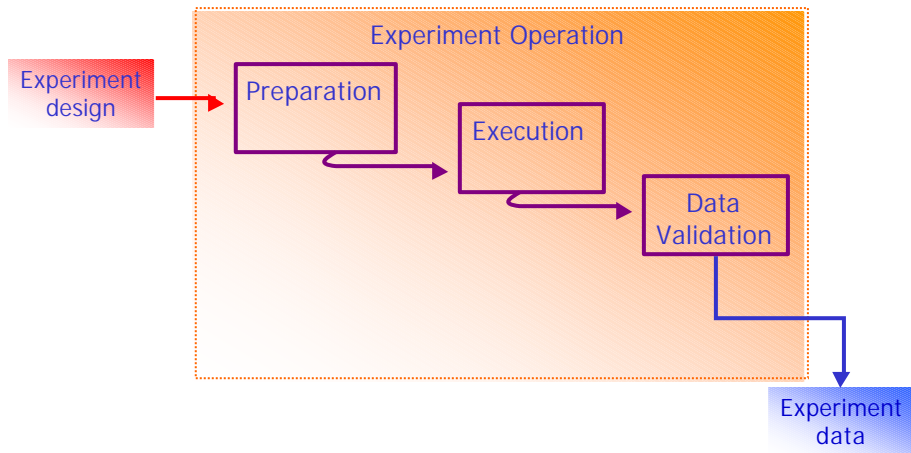
■ Theory Testing:

- Internal Construct Conclusion External

■ Applied Research

- Internal External Construct Conclusion

3.3 Experiment Operation



Experiment Operation Preparation

Obtain consent

The participants have to agree to the research objectives

Sensitive results

To assure the personal results confidentiality

Inducements

To offer some kind of inducements in order to attract people

Deception

If it is necessary, it should be explained to the participants as soon as possible

Experiment Operation

Pilot Studies

Pilot studies are conducted

- To find mistakes in the experimental procedure
- To test that the instructions are clear
- To check tasks have reasonable complexity, but they can be completed within the allotted time
- To ensure performance of any automatic data collection techniques
- To attempt to identify other unforeseen circumstances

Experiment Operation

Execution & Data Validation

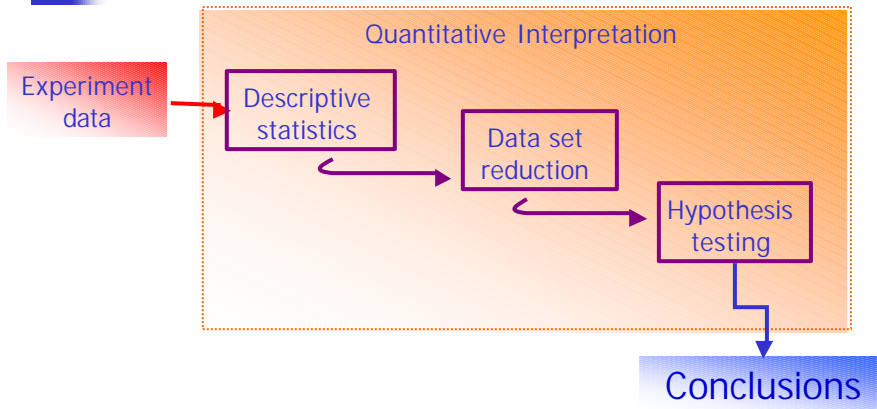
Data collection

- Manually by the participants that fill out forms
- Manually supported by tools in interviews
- Automatically by tools

Data validation

- The participants have understood the forms
- Someone may not have worked seriously (outliers)
- The experiment has been applied in the correct way

3.4 Analysis and Interpretation



Descriptive Statistics

- To describe and graphically present relevant aspects of the data set
- The scale of measurement restricts the type of statistics

Measures of			
Scale Type	Central Tendency	Dispersion	Dependency
Nominal	Mode	Frequency	
Ordinal	Median	Interval of variation	Spearman coef.
Interval	Percentiles Mean	Standard deviation	Kendall coet. Pearson coef.
Ratio	Geometric mean	Range Variation Coefficient	

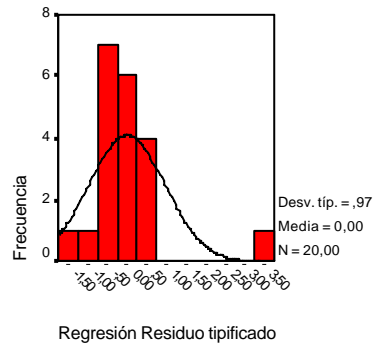
Descriptive Statistics

▪ (Cartwright 1998)

- box-plot
- Histograms
- Measures of central tendency etc.
- Scatter plot

Histogram residuals (Wholin)

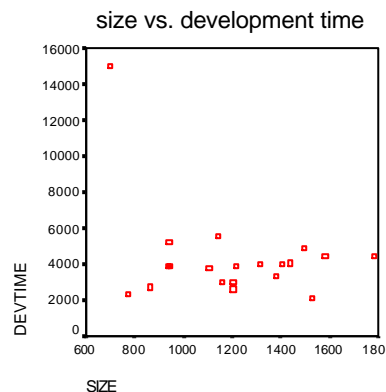
Dependent variable: DEVTIME



Data Set Reduction

To identify outliers

- Outlier: the data point is much smaller than one could expect looking at the other data points
 - 3 or more standard deviations over the mean
- Related with data validation



Data Set Reduction

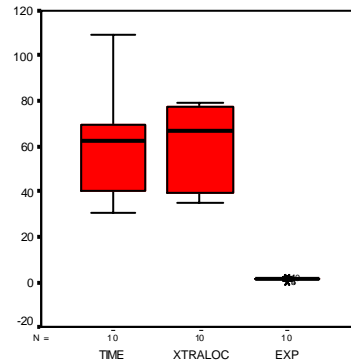
To identify outliers

■ Using a Box -Plot

$$d = P_{75} - P_{25}$$

Lower Tail: $P_{25} - 1,5d$ Upper Tail: $P_{75} + 1,5d$

Values outside the lower and upper tails are called **outliers**



XTRALOC	P ₂₅	38,25
	P ₅₀	66,50
	P ₇₅	77,50

Parametric and no Parametric Test

Applicability: What are the assumptions made by different test?

- About variables distribution (Normality, independency etc)
- About scales

Power

- The power of parametric tests is generally higher than for non parametric tests. That is parametric test require smaller experiments

Some parametric test are robust, this means that permit some deviations from requirements

Parametric and no Parametric Test Model Adequacy Cheking

■ **Normal Distribution:** The Chi-2 test can be made to asses to which degree the assumption about the data normally distributed is fulfilled

■ **Independence:** when the test assumes that the data is a sample from several independent stochastic variables, we need to check that there is not correlation between the sample sets (Pearson coefficient, Spearman coefficient, etc)

Residuals: In many statistical models, as ANOVA or Lineal models, there is a term that represent Residual (statistical error). Usually the residuals are normally distributed. We can check this property using a normal plot, or a chi-2 test

Parametric and no Parametric Test

Design	Parametric	Non Parametric
One Factor 2 treatments. Completely randomized	<u>T-test</u> F-test	Mann-Whitney Chi-2
One Factor 2 treatments. Matched	Paired t-test	Wilcoxon Sign Test
One factor more than 2 treatments	ANOVA	Kruskall-WallisChi-2
More tha one factor	ANOVA	Chi-2

Prediction Models: Lineal Models, Logit, Logistic

Hypothesis Testing

One factor with 2 levels (Cartwright, 1998)

- Dependent variable: Completion Time, in minutes, to modify a database program
- Treatments: version flat vs. Version with inheritance
 - $E(\text{Time/flat}) = \mu_{\text{flat}}$ $E(\text{Time/inheritance}) = \mu_{\text{inh}}$
- Hypothesis
 - Ho: That 3 levels of inheritance has no impact upon time to make a correct maintenance change as compared with no inheritance
 - H1: \neg Ho $\alpha = 0.05$
 - Ho: $m_{\text{inh}} = m_{\text{flat}}$ $H_1: m_{\text{inh}} \neq m_{\text{flat}}$ \rightarrow T-statistic**

Hypothesis Testing

ANOVA. One factor with K levels

Dependent variable Y

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} = N(0, \sigma_\varepsilon) \text{ independents}$$

$$\sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y})^2 = \sum_{j=1}^k \sum_{i=1}^n (y_{.j} - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - y_{.j})^2$$

Total variation

Treatment variation

Residual variation

Hypothesis Testing

ANOVA : One factor with K levels

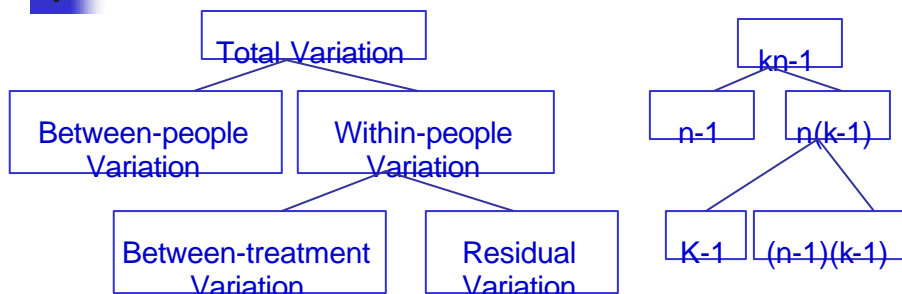
Source of variation	Sum of squares	Degrees of freedom	Mean square	$F_{k-1, n-k, \alpha}$ $H_0: \alpha_i = 0$
Between treatments	$SS_{\text{Treatment}}$	$(k-1)$	$MS_{\text{Treatment}}$	$F_{\text{observed}} = MS_{\text{Treatment}} / MS_{\text{Error}}$
Residual (error)	SS_{Error}	$N-k$	MS_{Error}	
Total	SCT	$N-1$		

Conclusions: $F_{\text{observed}} > F_{k-1, n-k, \alpha} ? \rightarrow$ to reject H_0

[Results-SPSS](#) [Example-Languages](#)

Hypothesis Testing

ANOVA: One factor within subjects



Schematic representation of the analysis

Partition of degree of freedom

Dependent variable Y

$$Y_{ij} = \mu + \alpha_i + \beta_i + \varepsilon_{ij} \quad \varepsilon_{ij} = N(0, \sigma_\varepsilon) \text{ independents}$$



Hypothesis Testing

ANOVA: One factor within subjects

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F_{k-1, (n-1)(k-1), \alpha}$ $H_0: \alpha_i = 0$
Between people	$SS_{b, \text{people}}$	$(n-1)$	$MS_{b, \text{people}}$	$F_{\text{observed}} = MS_{\text{Treat}} / MS_{\text{res}}$
Within people	$SS_{w, \text{people}}$	$n(k-1)$	$MS_{w, \text{people}}$	
Treatments	SS_{treat}	$K-1$	MS_{treat}	
Residual (error)	SS_{res}	$(n-1)(k-1)$	MS_{res}	

Conclusions: $F_{\text{observed}} > F_{k-1, (n-1)(k-1), \alpha} ? \rightarrow$ to reject H_0



Hypothesis Testing

Daly 1995

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	TIME	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
TIME	Lineal	680,625	1	680,625	2,597	,124
TIME * GROUP	Lineal	50,625	1	50,625	,193	,666
Error(TIME)	Lineal	4717,250	18	262,069		

ANOVA within subjects



3.5 Drawing Conclusions

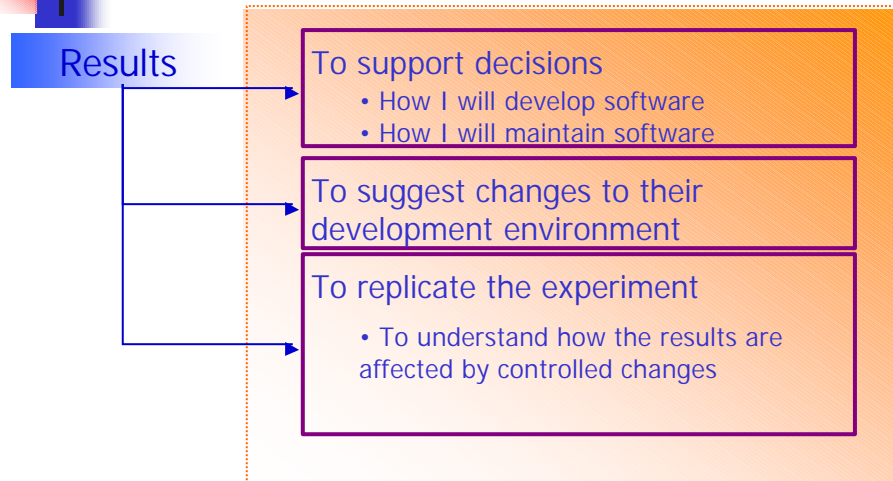
- If the null hypothesis is rejected we can conclude that the results were significant. Then we can to make general conclusions about independent and dependent variables
 - The conclusions can be generalized to contexts that are similar to experimental setting (External validity).
- Conclusions practical importance
 - Although the result may be statistically significant, it is not necessarily that the result is of practical importance. And vice versa, the lesson learned from a non-significant experiment may be of practical importance



4. Conclusions

- It is necessary more Replication
- To study the concepts concerning object oriented
 - inheritance
 - agregation
- To show the results, including non significant results
- To elaborate a more specific guide to experiment in software engineering

Dissemination and Decision-Making (Pfleeeger)

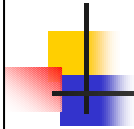


Signs of Maturity

- Level of sophistication of the goals of an experiment
- Understanding interesting things about the discipline

For software Engineering mean:

- Can we build models to measure and differentiate process and products?
- Can we measure the effect of a change in a particular process or product variable?
- Can we predict product variables based upon a process model in a context?
- Can we control for product effects, given a particular set of context variables?



Signs of Maturity

- A pattern of knowledge built from a series of experiments

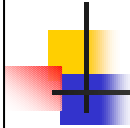
- Does the discipline build in prior (Models, experiments, knowledge)?
- Was the study an isolated event? or
- Did it lead to other studies that used its information?
- Have studies been replicated?
- Does the building of knowledge exist in one research group or has it spread to others?

Family of Experiments and Replication



Bibliography

- [IEEE90] IEEE. "Standard Glossary of Software Engineering Terminology" IEEE Std 610.12-1990, 1990
- Basili V.R. "The role of experimentation in software engineering: Past, Present and Future" T.R. 1996
- Basili V.R., Scott G, Laitenberger O., Lanubile F, Shull F, Sorumgard S, Zelkowitz M. "The empirical investigation of perspective-based reading. Empirical software engineering, vol 1, n° 2, 1996
- Cartwright M, Shepperd M. ESERG:TR98-02
- Daly J., Brooks A., Miller A. Roper J., Wood M. "Evaluation inheritance depth on the maintainability of object-oriented software" Empirical software engineering, vol 1 no 2, 1996
- Dolado Cosín, J.J. & Fernández Sanz L. "Medición para la gestión en la Ingeniería del Software". Ra-Ma 2000

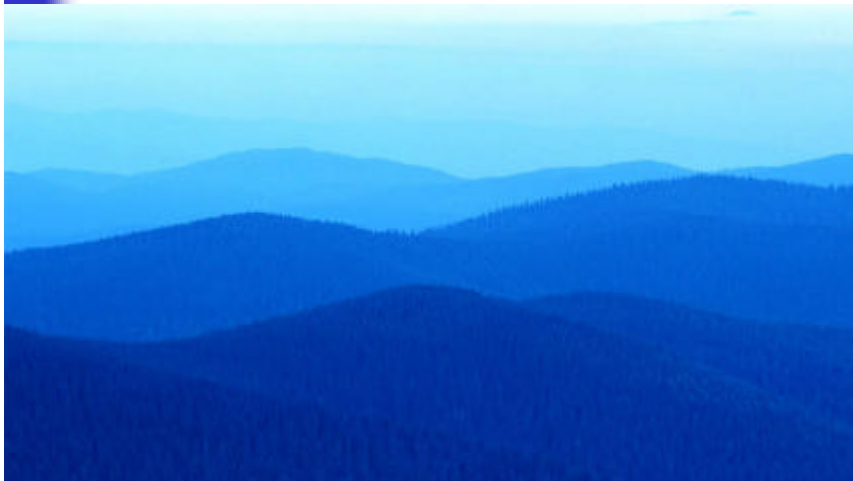
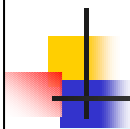


Bibliografía

- Finney, K., Rennolls, K., Fedorec, A., " Measuring the comprehensibility of z specifications", Journal of systems and software, vol. 42, nº 1, 1998.
- Pflegger S.L. 2Experimental design and analysis in software engineering" Annals of Software Engineering 1(1995)219-253
- [Scanlan D.A. "Structured flowcharts outperform pseudocode: an experimental comparison" IEEE Software, vol 6, no 5 1989
- Tichy W.F., Lucowicz L., Prechelt L., Heinz E.A. "Experimental evaluationin computer science: a quantitative study", Journal of Systems and Software, 28(1), 1995.
- Wohlin C. Et al. 2Experimentation in software engineering. An Introduction". Kluwer Ac. P. 2000.
- Zelkowitz M.V. & Wallace D." Experimental validation in software engineering". Information and Software Technology39(11), November, 1997.

© E. Manso U. de Valladolid

113



© E. Manso U. de Valladolid

114