**Experimentation**

**Measure**  
Mapping from attribute of entity to measurement value.  
**Valid** measure doesn't violate any attribute properties

**Effect size**  
Quantifying the difference between two groups.

**Scale** of a measure  
Inches, meters, etc

**Admissible transformation, Rescaling**  
Transform measure into different scale, preserves relationship among objects.

**Meaningful/Meaningless statements**  
Statements are true/false when measures are rescaled.

**Scale types**- Nominal: Maps attribute to name/symbol (classification, labelling)  
 Transformations preserve one-to-one mapping  
- Ordinal: Ranks entities after ordering criterion (>, better, more complex)  
 Transformations preserve order of entities (grades, complexity)  
- Interval: Difference btw two measures is meaningful, ordering like Ordinal + relative distance  
 Transformations are lin. comb. M’ = aM+b (Celsius, Fahrenheit)  
- Ratio: Meaningful zero value + ratio btw measures (length, Kelvin)  
 Transformations only differ by factor M’ = aM

**Objective Measure**No judgement in measurement value (LOC, delivery date).

**Subjective Measure**Person making measurement contributes with judgement, mostly nominal/ordinal (skill, usability).

**Direct Measure**Directly measurable (LOC, #defects in test)

**Indirect Measure**Involves/Is derived from measurement of other attributes (defect density = #defects/LOC, productivity = LOC/effort)

**Internal Attributes**Can be measured purely in terms of the object (i.e. Effort, Size, Age)

**External Attributes**Can only be measured with respect to how object relates to other objects (i.e. Cost, Reliability, Productivity)

**Experiment Process**

Starting point: Idea of cause and effect relationship, relationship btw cause construct & effect construct. (Theory, hypothesis).

Experiment (to test hypothesis): Input, observe output, draw conclusion about relation construct

**Dependent variable**Studied to see effect of changes in independent variables.

**Independent variable**All variables that are manipulated and controlled**.**

**Factor**Independent variable that is changed to study effect (as opposed to fixed)

**Treatment**One particular value of a factor

**Object**Characterized by independent variables

**Subject**People who apply treatments

Experiment consists of a set of **tests/trails**Combination of treatment, subject, object

**Experimental Error**Affected by number of tests

**Scoping**- Scope experiment in terms of problem, objective and goals  
- State hypothesis clearly, define objective & goals  
- Object of study (what is studied?)  
- Purpose/Intention  
- Quality focus (which effect is studied?)  
- Perspective (whose view?)  
- Context/Environment (where? #Subjects/Objects, One of each: Single object study)

**Goal template**Analyse <Object(s) of study>  
for the purpose of <Purpose>  
with respect to their <Quality focus>  
from the point of view of the <Perspective>  
in the context of <Context>.

**Goal definition framework**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Object of study** | **Purpose** | **Quality focus** | **Perspective** | **Context** |
| Product Process Model Metric Theory | Characterize Monitor  Evaluate  Predict  Control  Change | Effectiveness  Cost  Reliability  Maintainability  Portability  Security | Developer  Modifier  Maintainer  Customer  User  Researcher | Subjects  Objects |

**Planning  
-** Design, instrumentation and threats  
- Determine Context in detail, environment  
- State hypothesis formally (Null & alternative hypothesis)  
- Determine variables and their values + scale  
- Consider validity of expected Results

**Context Selection**- Offline vs. Online  
- Student vs. Professional  
- Toy vs. real Problems  
- Specific vs. General

**Hypothesis Formulation  
Null Hypothesis H0**- There are no real underlying trends/patterns  
- Needs to be rejected with high significance  
  
**Alternative Hypothesis Ha/H1**  
- Hypothesis in favour of which H0 is rejected

**Type-I-error**  
- Statistical test indicates pattern when there is none  
- P(Type-I-error) = P(reject H0 | H0 true)

**Type-II-error**- Statistical test hasn’t indicated a pattern when there actually is one  
- P(Type-II-error) = P(not reject H0 | H0 false)

**Power** of a statistical test  
- Probability that test will reveal true pattern if H0 is false (-> choose test with max power)  
- Power = P(reject H0 | H0´ false) = 1 – P(type-II-error)

**Variables Selection**  
- Choose (in)dependent variables + measurement scales + range + levels

**Selection of Subjects**- Selection must be representative so results can be generalised  
- probability (known subjects) vs. non-probability (unknown subjects) sample

**Experiment Design  
Randomization**- in allocation of objects, subjects and order of tests

**Blocking**- eliminate undesired effect in comparison among treatments  
- within one block, undesired effect is the same (i.e. experienced vs unexperienced)

**Balancing**- equal number of subjects per treatment/block

**One factor, two treatments** µi = Mean of dependent variable for treatment i  
yij = jth measure of dependent variable for treatment i  
Ex. Hypothesis  
H0 : µ1 = µ2, H1 : µ1 != µ2, µ1 < µ2 or µ1 > µ2 (t-test, Mann-Whitney)

**One factor, more than two treatments**H0 : µi = µj for all pairs (I, j), H1 : µi != µj  for at least one pair (i, j) (ANOVA, Kruskal-Wallis)

**Two factors**- 3 Hypotheses: one for each factor + interaction between factors  
 **Instrumentation**- Experiment Objects, Guidelines, Measurements

**Validity Evaluation**- Results have **adequate validity** if they can be generalised for the targeted population

**Internal validity**  
- validity within environment & reliability of results  
- relationship is not caused by a factor which isn’t measured or controlled   
- treatment causes outcome  
Threats: Behaviour caused by disturbing factor

**External validity**- How general are findings? (outside the actual context) Threats: wrong environment/Objetcs

**Construct validity**- judging if treatment reflects cause construct, outcome provides true picture of effect construct  
Threats: Experiment setting doesn’t reflect construct under study (i.e. number of courses -/-> skill)

**Conclusion** **validity**- relationship between treatment and outcome of experiment  
Threats: choice of stat. tests, sample sizes, implementation, measurement

**Operation  
-** Collect measurements  
- Preparation: Needed material, inform subjects  
- Execution: Ensure experiment conducted accordingly & data collection  
- Data validation: Ensure correctness of data, provide valid picture of experiment

**Analysis & Interpretation**- Analyse, evaluate measurements  
- Understand data using descriptive statistics, provide visualization  
- Data reduction: reduce number of data points, variables (provide same Information?)  
- Hypothesis test  
- Interpretation: was hypothesis possible to reject?

**Descriptive Statistics**Measures of Central Tendency:  
- (arithmetic) mean (interval, ratio)  
- median (ordinal, interval, ratio)  
- mode (nominal, ordinal, interval, ratio)

Measures of Dispersion  
- (sample) variance (interval, ratio)  
- standard deviation (interval, ratio)  
- range (interval, ratio)  
- variation interval (ordinal, interval, ratio)  
- coefficient of variation (ratio)  
- (relative) frequency (all)

Measures of Dependency  
- linear regression  
- transformation -> linear regression  
- covariance  
- (Pearson) correlation coefficient

Graphical Visualisation  
- scatter plot  
- box plot  
- outliers  
- (cumulative) histogram  
- pie chart

**Data Set Reduction**- Outliers can be excluded if caused by rare event, that will never happen again

**Hypothesis Testing**

**Presentation & package**- Documentation of results, lessons learned