Carnegie Mellon University

Effective User Survey Design and Data Analysis

Hanan Hibshi and Travis Breaux

Monday, September 12, 2016

Full Day Tutorial
23rd IEEE International Requirements Engineering Conference

institute for SOFTWARE RESEARCH

Tutorial Outline

- Effective User Survey Design
  - Session 1: Introduction to experimental research and user surveys
  - Session 2: Building user surveys
- Analyzing and Reporting User Survey Data
  - Session 3: Quantitative analysis of survey data
  - Session 4: Qualitative analysis of survey data

Session 3:

QUANTITATIVE ANALYSIS OF SURVEY DATA

Topics of Third Session

- Statistical significance
- Data preparation
- Statistical tests
- Role of assumptions in statistical tests
- Effect of study design on data analysis
- Effect of conditions (levels of a variable)
- Threats to validity
- Power analysis
- Reporting results and descriptive statistics

Identifying the Variables and Metrics

- We need to think of independent/dependent variables
  - **Independent variable:** that is being changed or controlled
  - **Dependent variable:** that is being tested
  - **Control variable:** holding a dependent variable constant (e.g. age)

Is developers' productivity affected by type of programming language?

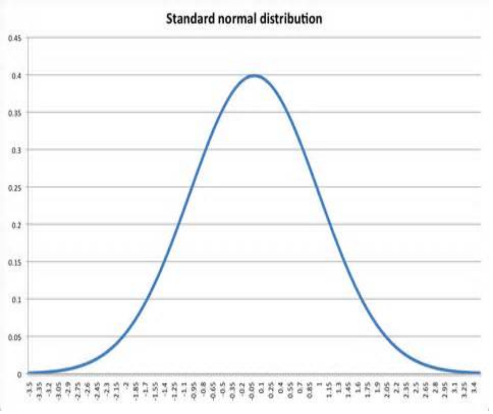
- How to measure productivity:
  - Time!
  - Number of lines
  - Number of hours
  - Number of tasks, submissions, commits, results etc.

Without Statistical Significance

- Mary scored 90% on PL assessment test, John scored 75%,
  - Mary is a better programmer then John
- The average score of three male students on PL test: 70%, the average of three female students is 90%
  - Females are better programmers then males
- Arguments:
  - Size of compared groups is very small
  - The individuals in the groups are not representative of all males and females
  - I can go and find another 6 students and prove the opposite!

Statistical Significance

- We cant collect data from every male and female
- The full population follow a normal distribution
- We can select a small sample that represent the population
- Significance tests can indicate if results from our sample can generalize



- Lazar, J., Feng, J.H. and Hochheiser, H., 2010. Research methods in human-computer interaction. John Wiley & Sons.

How Significance Testing Work

- **Null hypothesis: $\boldsymbol{H_0}$** (No difference between groups)
- **Alternative hypothesis: $\boldsymbol{H_1}$** (Group A is better then B
- The significance tests tell us to reject or accept H_0
- Example:
  - $\boldsymbol{H_0}$ There is no difference between the transaction time of an ATM with touch screens and ATM with buttons
  - $\boldsymbol{H_1}$ ATM with touch screen has a shorter transaction time than an ATM with buttons
- Significance testing is subject to Type I and Type II errors

- Lazar, J., Feng, J.H. and Hochheiser, H., 2010. Research methods in human-computer interaction. John Wiley & Sons.

Type I and Type II Errors

- Type I (false +ve) is worse than Type II (false –ve)
- α: Type 1 error rate or significance level
- β: Type 2 error rate. (1- β) is the power of the test

		Reality	
		H_0 is True	H_0 is False
Decision	Accept H_0	Accurate $(1-\alpha)$	Type II Error β
	Reject H_0	Type I Error α	Accurate $(1-\beta)$

Power: The probability of successfully rejecting a null hypothesis when its false and should be rejected [1]

[1] Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd Edition.

[2] Rosenthal, R. and Rosnow, R.L., 1991. Essentials of behavioral research: Methods and data analysis. McGraw-Hill Humanities Social.

P-value and Degrees of Freedom

- P-value
  - The probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true
  - We want this to be as small as possible
  - $p< \alpha$
  - If the P is higher it does not imply that the null hypothesis is true! It means we CANNOT REJECT the null hypothesis given the level α
  - Refers to one null hypotheses
- Degrees of Freedom
  - The degree of freedom is the number of values in a calculation that we can vary
  - $(20 + 10 + x + y)/4 = 25$

Data Preparation

- Data need to be organized, cleaned-up, and coded

Pno	Age	Gender	Degree
P0001	31	Male	College
P0002	20	Female	Graduate
P0003	27	Female	High School

Pno	Age	Gender	Degree
P0001	31	1	2
P0002	20	0	3
P0003	27	0	1

Likert-scale Coding Example

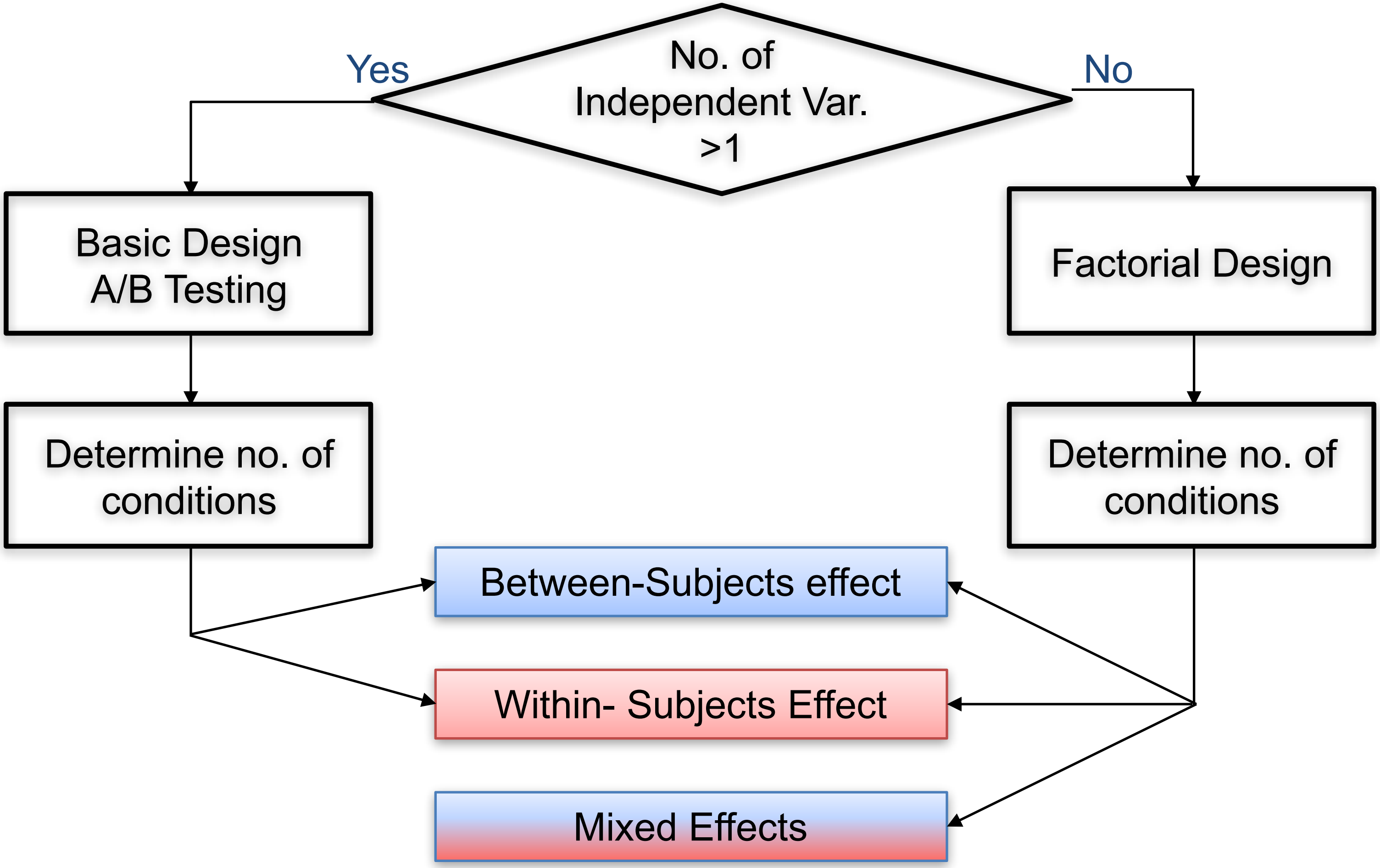
- Very unlikely = 1
- Unlikely = 2
- Natural = 3
- Likely = 4
- Very likely =5

Pno	Q1	Q2	Q3
P0001	3	1	1
P0002	2	1	0
P0003	5	3	1

Types of Variables

- Categorical, discrete
  - Ordinal
  - Nominal
- Continuous

Experiment Design: Structure



- Lazar, J., Feng, J.H. and Hochheiser, H., 2010. Research methods in human-computer interaction. John Wiley & Sons.

Common Statistical Tests for Experiments

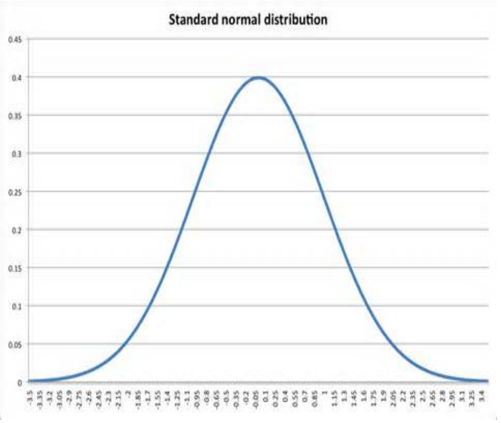
- Parametric tests
  - T-test
  - ANOVA (Analysis of Variance)
- Linear regression
- Non-parametric tests
  - Chi-square: significance of Frequency (contingency tables)
  - Fisher’s
  - Wilcoxon signed ranks test
  - Man-Whitney U test

T-Test

- One IV, two groups/conditions
- Independent sample T-test
  - Between Subjects design
  - Assumption: observations are independent
- Paired Sample T-Test
  - Within-Subjects design
- Example:
  - There was a significant difference in the scores for [C1] (M=4.2, SD=1.3) and no [C2] (M=2.2, SD=0.84) conditions; t (8)=2.89, p = 0.020

T-Test Assumptions

- DV scores are normally distributed
    - Violated: use non-parametric test
- Homogeneity of variance
    - Violated: use transformation (e.g. logs)
- Independent errors



ANOVA

- Analyses of Variance
- F statistic
- Between subjects
  - One Way ANOVA: One IV, 3 or more groups
  - Factorial ANOVA: 2 or more IV, 2 or more groups
- Within subjects
  - Repeated Measures
- Mixed design
  - Split-plot ANOVA

Linear Regression

- You can have a number of IV
- Purposes:
  - Model construction
    - Find the mathematical relationship (find the equation)
  - Predication
    - Predict the DV based on known factors
- Enter variables one at a time vs. all together
- Multi-level regression
  - Good for mixed methods
  - Can analyze hierarchal data

Power

- β: Type 2 error rate, (1- β) is the power of the test
- $Power = P(reject\ H_0|H_1\ is\ true)$
- Power analysis: calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size
  - Specify your effect size
  - Your α, β
  - G power is a popular tool
  - Priori or post-hoc

Threats to Validity

- External validity
  - The degree to which results can be generalized
  - Example threat: convenience sampling
- Internal validity
  - the extent to which a causal conclusion based on a study is warranted, which is determined by the degree to which a study minimizes systematic error or 'bias‘
  - Example Threats: learning effect, fatigue, confounding variables…
  - Randomization, reduction of fatigue, attention checks
- Construct validity
  - Are we measuring what is claimed to be measured
  - Example threats: Errors in measuring instrument, lower control over measurement