User-Centric Computing for Human-Computer Interaction NPTEL-MOOCS L28

Dr Samit Bhattacharya

Computer Science and Engineering

IIT Guwahati

Empirical Research Stages

- Broadly, four (or five) stages
  - Identification of research question(s)
  - Determination of variables
  - Design of experiment
  - Analysis of empirical data
  - There is also a fifth stage: building of a model, if that is what we want

Understanding the Stages

- We discussed first THREE stages
- **Today, we shall discuss remaining stages (data analysis and model building)**

Basic Idea

- Consider aesthetic judgment behavior study (for RQ4)
- We decided to make use of twelve participants
- We designed twelve tasks; one task for each test condition
- We performed a repeated-measure experiment and used Latin Square method for counterbalancing
- We have 144 data items ($12 \times 12 = 144$)

Basic Idea

- A portion of data shown in Table

Interfaces / Participants	I_1	I_2	I_3	...	I_{10}	I_{11}	I_{12}
P_1	3	1	2	...	5	4	2
P_2	4	2	3	...	4	4	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P_{12}	2	2	2	...	3	5	2

Basic Idea

- What to do with this data?
- We can use this for regression analysis or training set for the learning-based modeling effort
- **However, the data might be misleading**

Basic Idea

- We made use of twelve interfaces - represent only a very small fraction of all interfaces
- Twelve participants are also a very small fraction of all the potential users (even if we consider specific demographic profiles)

Basic Idea

- We are dealing with *samples* rather than the actual *population*
- Where is the guarantee the observations was not due to *chance*

Basic Idea

- If we **conduct another study** with a completely different set of participants and interfaces, we **may end up with a completely different data set** leading to a different model altogether

Basic Idea

- How likely is that possibility - we need to answer this question first
- Points to "**statistical significance**" of the data

Fundamental Concern

- We work with **samples BUT wants to draw conclusion** on larger **population**
- Thus, we wish to find relationship that holds between rating of *any user* for *any interface*
  - Rather than one that is applicable for only the twelve participants and the twelve interfaces

Fundamental Concern

- Necessary to determine *nature* of data
  - Is the data occurred *by chance*
  - Or due to *specially designed test conditions* (the interfaces) - often termed as "treatment"

Basic Idea

- Statistical significance tests answer the question
- If we perform significance test on data and find the *statistic* is *significant* with $p<0.05$ (to learn soon), we can say with *confidence* that the data is due to the "treatment" and not by chance in 95% of the times

Basic Idea

- Our starting point is a *statistic*
- In the context of user-centric research, is the *mean* (or *average*) of a group of data items (the sample)
- Typically, we are interested to test the significance of the *difference* between the means of several groups of data (in most situations)

Basic Idea

- Ex - let us assume there is only one factor (N or the number of objects) for the aesthetic rating. The factor has two levels: 4 and 8
- Thus, we have only two interfaces (one each for the two levels)
- There are twelve participants as before

Basic Idea

Participants		P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	Group Mean
Interfaces	N = 4	3	2	3	4	3	3	2	5	5	3	2	4	3.25
	N = 8	2	1	1	3	3	4	2	2	1	1	3	1	2.00

Basic Idea

- The *mean*, along with other measures such as *median*, *mode*, *variance* and *standard deviations*, are *descriptive statistics*
- Can reveal many things in themselves
  - E.g., the two means reveal that aesthetic of interfaces with lesser number of objects tend to be better (as per user judgment)

Basic Idea

- Can't have such general conclusion
- Same reason - where is the guarantee that the difference shall remain the way it is (i.e., group mean for N=4 is **greater** than other group mean)?

Basic Idea

- We are **not concerned about absolute values** (may change)
- Rather about **relative difference** between them
- The only way to be confidant of the reliability of the observations is to go or the significance tests

Basic Idea

- We are **not concerned about absolute values** (may change)
- Rather about **relative difference** between them
- The only way to be confidant of the reliability of the observations is to go or the significance tests

Basic Idea

- Ex – consider the research question

  How the aesthetic score (in a scale of 1-5) depends on the number of objects an interface has?

- Corresponding hypotheses (null and alternative)

  H_0: *The aesthetic score (in a scale of 1-5) does not depend on the number of objects an interface has.*

  H_1: *The aesthetic score (in a scale of 1-5) depends on the number of objects an interface has.*

Basic Idea

- With the significance test, we try to refute the null hypothesis
- Let us try to understand the process with one simple (probably the simplest) statistical significance test

Paired Sample t-test

- If you perform the test, you are likely to get something like Table

The difference in group means = 1.25
The two-tailed p value = 0.020583
t = 2.702015
Degrees of freedom (df) = 11
The difference is statistically significant

Paired Sample t-test

- Although everything in Table is important, we do not report all these information
- Difference of the means is found to be statistically significant - we simply report this fact

Paired Sample t-test

- With a specific format

  $t(11) = 2.70$, $p < 0.05$, statistically significant

Paired Sample t-test

- In the format, we use the lowercase 't'
- After it, we put degrees of freedom (df) value (11) within parenthesis (without any space in between)
- This is followed by an '=' symbol
- Followed by the t-statistic (2.70)
- Then we put a comma and a space character
- Finally, we write the 'p' value as "$p < 0.05$"

Paired Sample t-test

- The ‘p’ value roughly indicates the probability that the data occurred by chance
- The value ‘0.05’ is a pre-defined value
- It indicates that the probability of getting the results by chance is about 5%

Paired Sample t-test

- Since the test result indicates difference between means is statistically significant, we reject null hypothesis

Paired Sample t-test

- It may be noted that the tests run the risk **of TWO types of errors**

Paired Sample t-test

- **Type I error** (also known as the α error or "false positive")
  - Occurs when we reject a null hypothesis, which is true and should not be rejected
  - To avoid Type I errors, we typically use a very low value of p (e.g., $p < 0.05$)

Paired Sample t-test

- **Type II error** (also known as the β error or "false negative)
  - Indicates a situation where we do not reject a null hypothesis although it is false and should have been rejected
  - To avoid, it is generally recommended to go for larger sample sizes

Note

- Many important terminologies and issues (e.g., *p-value*, *two-tailed distribution*) we mentioned in the passing
  - Our aim is to introduce the idea - for more details, reference may be consulted

Note

- We also did not explain actual calculations - not necessary to perform by yourself at all
  - Instead, you can utilize any statistical package such as the SPSS™

Techniques for Tests

- Broadly TWO categories
  - Parametric
  - Non-parametric

Parametric Tests

- t-test is one of the many methods collectively known as the *parametric* methods
- Applicable subject to the fulfilment of THREE conditions

Parametric Tests - Conditions

- Data should come from a "normally distributed" population
- We should use at least an interval scale (with equally-spaced intervals) of measurement for the dependent variable (a ratio scale is even better)
- Variance in the groups of data should be approximately equal

Parametric Tests

- Did we do the right thing by applying the t-test on our data as explained earlier ?

Parametric Tests

- We used a Likert-type scale to record ratings
- Likert scales are ordinal scales of measurement
- Thus, we are violating the second condition for a parametric test - **ideally we should not have performed the t-test on our data**

Parametric Tests

- Let us now make TWO assumptions

1st Assumption

- The rating scale is an interval scale
- Ex - 1 indicates good aesthetics, 2 indicates *doubly as good* aesthetics, 3 indicates *triply as good* aesthetics and so on till the rating 5, which indicates *five times as good* aesthetics
  - Now each rating not only indicates a judgment, but does so relative to the other judgments (with respect to an absolute starting point at 1 = good)
  - Table records the interval scale data rather than the Likert rating

2nd Assumption

- Rating scores follow a "normal" (Gaussian) distribution
- **Distribution refers to entire population** – we plot the ratings by *all* the users (if that is possible)
  - Our sample data (ratings of twelve participants) drawn from a normally distributed population
  - Sample data in itself need not be normally distributed

Parametric Tests

- With these assumptions, we can perform *t-test* on our data
- Among the three conditions, the 2nd and 3rd are easy to verify
- It may be difficult to know the nature of the population data (the 1st condition)

Parametric Tests

- If you are not sure, you can go for additional tests such as the Shapiro-Wilk or the Kolmogorov-Smirnov test
  - Can reveal if the sample data is taken from a normally distributed population

Non-Parametric Tests

- If your data do not support any one of the above three condition, you should go for the *non-parametric* tests of significance

Parametric test	Experiment design	Non-parametric test
	Between-subject design with one factor having two levels. Nominal (categorical) scale of measurement	Chi-square test
	Within-subject design with one factor having two levels. Nominal (categorical) scale of measurement	McNemar's test
Independent-samples t-test	Between-subject design with one factor having two levels.	Man-Whitney U test
Paired-sample t-test	Within-subject design with one factor having two levels.	Wilcoxon signed ranks test
One-way ANOVA	Between-subject design with one factor having more than two levels.	Kruskal-Wallis test
Factorial ANOVA	Between-subject design with two or more factors, each having two or more levels.	
Repeated measure ANOVA	Within-subject design with one factor having three or more levels. Also applicable in within-subject design with two or more factors, each having two or more levels.	Friedman test

Model-Building from Data

- With the statistical significance test, we can ascertain the reliability of the data
- The test in itself does not help us to build the computational user model
- To get the model, we have to do something more

Mathematical Model - Example

- Recollect the research question RQ3 - we are interested to know relationship between aesthetic rating and number of objects
- t-test revealed that the data is reliable; it did not happen by chance (with 95% probability)
- Thus, we can use the data to build our model

Mathematical Model - Example

- Recollect the research question RQ3 - we are interested to know relationship between aesthetic rating and number of objects
- t-test revealed that the data is reliable; it did not happen by chance (with 95% probability)
- Thus, we can use the data to build our model

Data - Recap

Participants		P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	Group Mean
Interfaces	N = 4	3	2	3	4	3	3	2	5	5	3	2	4	3.25
	N = 8	2	1	1	3	3	4	2	2	1	1	3	1	2.00

Mathematical Model - Example

- First, we group the data into pairs $<AS_m, N>$ where AS is mean rating and N is the number of objects
  - We have two pairs: <3.25,4> and <2.00,8>

Mathematical Model - Example

- We can use these values to come up with an equation of a straight line in the slope-intercept form

$$AS = m.N + c$$

Mathematical Model - Example

- We replace the AS_m and N values in Eq. to get two equations

$$\left.\begin{aligned} 3.25 &= m.4 + c \\ 2.00 &= m.8 + c \end{aligned}\right\}$$

Mathematical Model - Example

- Solving, we get $m = -0.3125$ and $c = 4.5$
- Replacing these values in Eq., we obtain the **line equation, which is our model**

$$AS = -0.3125N + 4.5$$

- We followed **linear regression technique**

Mathematical Model - Example

- The model is predictive - if we are given the number of objects, we can predict the aesthetic rating
- We should keep in mind that the predicted rating would be **average** of the ratings likely to be given by a group of users

Mathematical Model

- In this example, we made use of only two data points (the two pairs) - may be an *over fit*
- In practice, we should collect more data points empirically (three or more) to obtain a line equation

Mathematical Model

- In that case, we should also check for the *goodness of fit* typically expressed with the value of R^2
  - Lies between 0 and 1, with higher values indicating better fit

Mathematical Model

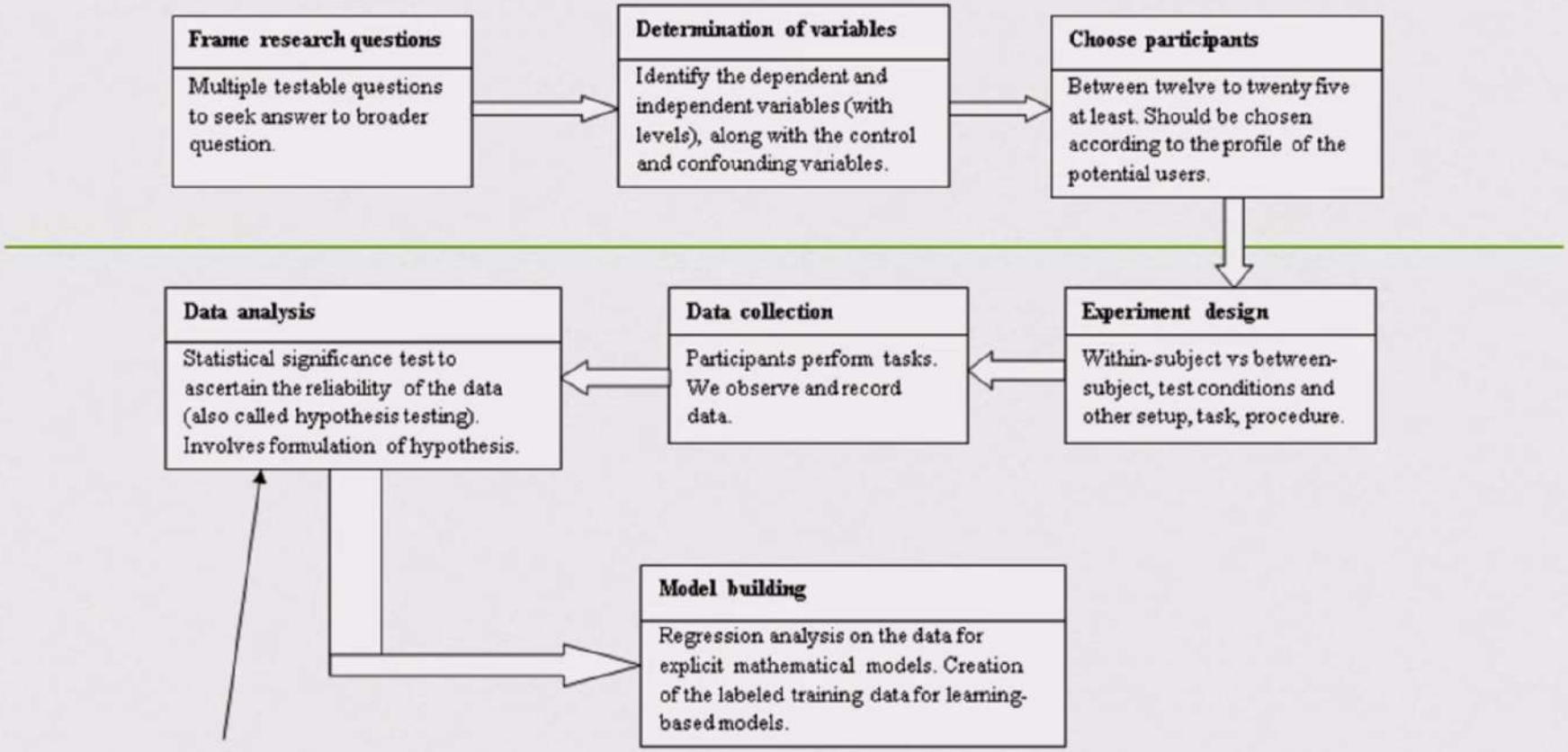
- We can also go for *non-linear regression* if the linear regression is not yielding a good model (indicated by a low R^2 value)

Mathematical Model

- We can use any statistical package to perform all these computations (determination of the model through regression along with the R^2 value)

Machine Learning Approach

- We can go for a machine-learning approach with our data as well
- In that case, the data we collect empirically will serve as the labelled training data
  - We can employ the data in supervised learning-based model building processes


Frame research questions
Multiple testable questions to seek answer to broader question.
Determination of variables
Identify the dependent and independent variables (with levels), along with the control and confounding variables.
Choose participants
Between twelve to twenty five at least. Should be chosen according to the profile of the potential users.
Experiment design
Within-subject vs between-subject, test conditions and other setup, task, procedure.
Data collection
Participants perform tasks. We observe and record data.
Data analysis
Statistical significance test to ascertain the reliability of the data (also called hypothesis testing). Involves formulation of hypothesis.
Model building
Regression analysis on the data for explicit mathematical models. Creation of the labeled training data for learning-based models.
We may stop at this stage if we are only interested in evaluating our design.