User Modeling and Personalization
9: User Evaluation of Adaptive Systems

Eelco Herder

L3S Research Center / Leibniz University of Hanover
Hannover, Germany

13 June 2016

Outline I

Outline II

User Evaluation

Often, evaluation is seen as the final mandatory stage of a project.

Instead of in-between *formative* evaluation whether a theory or approach holds, in many projects only a *summative* evaluation is planned to show 'that the system works'.

However, when constructing a new adaptive system, the whole development cycle should be covered by various evaluation studies

- from the gathering of requirements to the testing of the system under development

Formative evaluation

Formative evaluations are aimed at checking the first design choices before actual implementation and getting clues for revising the design in an iterative design-re-design process.

Phases of Evaluation - 1: The Requirement Phase

The requirement phase is usually the first phase in the system design process.

It can be defined as a process of finding out what a client (or a customer) requires from a software system.

During this phase, it can be useful to gather data about typical users (features, behavior, actions, needs, environment, etc), the application domain, the system features and goals, etc.

Techniques for gathering requirements

Task analysis
Task analysis methods are based on breaking down the tasks of potential users into users' actions and users' cognitive processes.

In most cases, the tasks to be analyzed are broken down into in sub-tasks

Cognitive and Socio-technical Models
The purpose of cognitive task models is the understanding of the internal cognitive process as a person performs a task, and the representation of knowledge that she needs to do that.

Contextual design

Contextual design is usually organized as a semi-structured interview, covering the interesting aspects of a system, while users are working in their natural work environment on their own work

Focus Group

Focus group is an informal technique that can be used to collect user opinions. It is structured as a discussion about specific topics moderated by a trained group leader.

Systematic Observation

Systematic observation can be defined as a *particular approach to quantifying behavior*. This approach is typically concerned with naturally occurring behavior observed in a real context.

Systematic observation is typically carried out in the context of observational studies.

Preliminary evaluation phase

The preliminary evaluation phase occurs during the system development.

It is very important to carry out one or more evaluations during this phase, to avoid expensive and complex re-design of the system once it is finished.

Techniques for preliminary evaluation

Heuristic evaluation

A heuristic is a general principle or a rule of thumb that can guide a design decision or be used to critique existing decisions.

Heuristic evaluation describes a method in which a small set of evaluators examine a user interface and look for problems that violate some of the general principles of good interface design.

Domain expert review

In the first implementation phases of an adaptive web site, the presence of domain experts and human designers can be beneficial.

A domain expert can help defining the dimensions of the user model and domain-relevant features.

They can also contribute towards the evaluation of correctness of the inference mechanism and interface adaptations.

Card sorting

A generative method for exploring how people group items and it is particularly useful for defining web site structures.
It can be used to discover the latent structure of an unsorted list of categories or ideas.

Cognitive walkthrough

An evaluation method wherein experts play the role of users in order to identify usability problems.

Wizard of Oz prototyping

A form of prototyping in which the user appears to be interacting with the software when, in fact, the input is transmitted to the wizard (the experimenter) who is responding to user's actions.

Prototyping

Prototypes are artifacts that simulate or animate some but not all features of the intended system.

They can be divided in two main categories: static, paper-based prototypes and interactive, software-based prototypes.

Participative evaluation

Another qualitative technique useful in the former evaluation phases is the participative evaluation, wherein final users are involved with the design team and participate in design decisions.

3. Final evaluation phase

The final evaluation phase occurs at the end of the system development and it is aimed at evaluating the overall quality of a system with users performing real tasks.

Usability testing

According to the ISO definition ISO 9241-11:1998, usability is "the extent to which a product can be used by specified users, to achieve specified goals, with effectiveness, efficiency and satisfaction, in a specified context of use"

Based on this definition, the usability of a web site could be measured by how easily and effectively a specific user can browse the web site, to carry out a fixed set of tasks, in a defined set of environments.

In particular, the usability test has four necessary features:

- participants represent real users;
- participants do real tasks;
- users' performances are observed and sometimes recorded
- users' opinions are collected by means of interviews or questionnaires

It is important to notice that '*observational*' usability tests of adaptive web sites can only be applied to evaluate general usability problems at the interface.

If one would test the usability of one adaptation technique compared to another one, a *controlled* experiment should be carried out.

Observational Methods

One way to find out about a phenomenon is simply to look at it in a systematic and scientifically rigorous way, without manipulating anything. The advantage is that you get an unbiased picture on how people behave.

An example observational study would be the analysis of the interaction and purchase history of Amazon users. This would help in building hypotheses or finding out some phenomenon, such as (the examples are made up):

- there are clusters of users that have similar purchasing behavior
- items with more positive ratings are bought more often
- special offers for cd's are more successful if the user already owns a cd by this artist or band

The downside to observational methods is that they are generally much more time-consuming to perform than controlled experiments.

They also do not allow the identification of cause and effect (the best you can get are *descriptive statistics* and *correlations*).

Findings from observational studies can be used as a basis for recommender algorithms (e.g. content-based versus collaborative filtering) or for certain adaptation decisions (e.g. hiding menu items that have not been used for a while).

The effects of these algorithms or adaptation decisions can best be measured in controlled experiments.

Controlled Experiments

Controlled experiments are one of the most relevant evaluation techniques for the development of the adaptive web.

The general idea underlying a controlled experiment is that by changing one element (the *independent variable*) in a controlled environment its effects on user's behavior can be measured (on the *dependent variable*).

The aim of a controlled experiment is to empirically support a hypothesis and to verify cause-effect relationships by controlling the experimental variables.

The most important criteria to follow in every experiment are:

- participants have to be credible: they have to be real users of the application under evaluation;
- experimental tasks have to be credible: users have to perform tasks usually performed when they are using the application.

The schematic process of a controlled experiment can be summarized in the following steps

1. Develop research hypothesis.
In statistics, usually two hypotheses are considered: the null hypothesis and the alternative hypothesis.

The null hypothesis foresees no dependencies between independent and dependent variables and therefore no relationships in the population of interest

(e.g., the adaptivity does not cause any effect on user performance).

2. Identify the experimental variables.

The hypothesis can be verified by manipulating and measuring variables in a controlled situation.

In a controlled experiment two kinds of variables can be identified:

- *independent variables* (e.g., the presence of adaptive behavior in a web site)
- *dependent variables*
  - the task completion time
  - the number of errors
  - proportion/qualities of tasks achieved
  - interaction patterns,
  - learning time/rate
  - user satisfaction
  - number of clicks
  - back button usage
  - home page visit
  - cognitive load measured through blood pressure, pupil dilatation, eye-tracking, number of fixations and fixation times)

3. Select the experimental methods and conduct the experiment.

The selection of an experimental method consists primarily of collecting the data using a particular experimental design.

In an ideal experiment, only the independent variable should vary from condition to condition.

In reality, other factors are found to vary along with the treatment differences.
These unwanted factors are called *confounding variables* (or nuisance variables) and they usually pose serious problems if they influence the behavior under study.

4. Data analysis and report writing.
In controlled experiments, data are usually analyzed by means of descriptive and inferential statistics.

Descriptive statistics, such as mean, variance, standard deviation, are designed to describe or summarize a set of data.

Inferential statistics are used to evaluate the statistical hypotheses. These statistics are designed to make inferences about larger populations.

Step 1. Develop research hypothesis

Within (HCI) research and academia, researchers employ (black-box) testing and evaluation to validate novel design ideas and systems

- usually by showing that human performance or work practices are somehow improved when compared to some baseline set of metrics (e.g., other competing ideas)
- or that people can achieve a stated goal when using this system (e.g., performance measures, task completions)
- or that their processes and outcomes improve.

Karl Popper coined the concept of *falsification*

- It is easier to prove that something is not true than that something is true.
- For this reason, evaluation aims to falsify the null hypothesis (which is the exact opposite of your hypothesis)
- It is common to accept the experimental outcome if the probability of the same outcome in another experiment is 95% ($p < .05$)

Step 2. Identify the experimental variables

The dependent variables are the outcome variables, in other words: the effects that you want to measure.
In many cases, objective measures are rather straightforward, such as:

- for elearning: decrease of learning time, increased retention time of learned material
- for online stores: number of products purchased (and not returned)
- for contextual help: number of suggestions followed, reduction of error rates
- for personalized portals: number of returning users, click-through rate

In other cases, general measures or theory-assessment measures (improvement with respect to theories on user behavior) can be used, such as:

- user satisfaction (measured using a standardized questionnaire)
- interaction time with the system
- reduction of *behavioral complexity* in Web navigation

The independent variables are the variables that are manipulated in the evaluation.

- A personalized system versus a non-personalized system
- Recommender A versus Recommender B
- Men versus women
- Beginners versus experts

More complicated designs are possible as well, in which, for example, systems A and B are evaluated by beginnrs and experts.

In all these cases, care should be taken that improvements are measurable and that any comparison with other systems is fair.

Step 3. Select the experimental methods

There are many different possible experimental designs and methods. We consider the two most common approaches:

- *Between-groups* designs use separate groups of participants for each of the different conditions in the experiment.
- *Repeated measures* designs expose each participant to all of the conditions of the experiment (in our case there typically two conditions)

Between-groups designs

Between-groups designs have several advantages over repeated-measures designs:

- *Simplicity:* you only need to make sure that participants are randomly allocated to the different conditions
- *Less chance of practice and fatigue effects:* participant performance will spontaneously vary from trial to trial. Carry-over effects are likely to happen (e.g. participant becomes tired, or already expects what is to come)

A disadvantage of between-groups designs is that any difference between groups may be due to your own manipulation or just due to unforeseen differences between both groups.

- unforeseen differences can be minimized by random allocation, making sure that both groups are similar in terms of age, gender, education, etcetera
- the effect of unforeseen differences (unsystematic variation) can be measured and compensated for

Repeated-measures designs

Repeated-measures designs (a.k.a. within-group design) have several advantages too:

- *Economy:* you can use each participant several times
- *Sensitivity:* you don't need to take unforeseen differences between groups into account

A big disadvantage of repeated-measures designs is the carry-over effect: participants become fatigue, bored, better practiced at doing the set tasks, and so on. To avoid this, you should *counterbalance* the order:

- if you have two conditions, half the participants get the conditions in the order A then B, the others get B first and then A.

Many other things to consider

- Can all participants participate at the same time or should each one be invited individually? (More time-consuming, but prevents cheating other undesirable effects and allows for personal interviews)
- Who will be the participants? A common problem of most evaluations in adaptive systems is that often the sample is too narrow (and often composed of students or the researcher's colleagues)
- Will participants perform differently in the morning than in the afternoon?
- Does the experimenter always have to be the same person?
- ...

How manipulation can go wrong

- I once conducted an experiment on measuring user performance in finding information on various Web sites
- One measure was user satisfaction: how challenging/exciting were the tasks (actually boring financial stuff)
- The participants from Utrecht University seemed to like the tasks far better than the participants from Twente University
- It turned out that my colleague has had to invest quite some effort in convincing people to participate; for me this went easier
- Most likely the Utrecht participants just wanted to please the experimenter

Conduct preliminary studies

Just one or two test rounds with fellow students, colleagues or friends can prevent a lot of damage:

- Are the task descriptions clear and unambiguous enough (you don't want your participants to do completely different things)
- Is there any influence of location or time of the day (typically, people are tired at the end of the day). If you can't conduct all experiments in the morning, uniformly spread them
- Does the material work (prototype Website, logging material)
- Have a first check at the results
- Try to standardize as much as possible.

Step 4. Data analysis and report writing.

What to report about a study

Hypotheses or Research Questions

- What are you going to test (and why)

Experimental Setup

- Participant Pool
- Material and Procedure
- Independent Variables (what you're going to manipulate)
- Dependent Variables (what you're going to measure)

Results

- Descriptive statistics (means, standard deviations, questionnaire outcomes, other important remarks)
- Inferential statistics (test outcomes, only in experimental settings)

Discussion

- Leave any interpretation to the discussion section. The result section is *objective*, the discussion section tells you *what to do with it*.

Data Collection Methods - 1: The Collection of User's Opinion

The collection of user's opinion, also known as query technique, is a method that can be used to elicit details about the user's point of view of a system

Questionnaires
Questionnaires have pre-defined questions and a set of closed or open answers. The styles of questions can be general, open-ended, scalar, multi-choice, ranked.

Questionnaires are less flexible than interviews, but can be administered more easily

Types of questionnaires

On-line questionnaires

To collect general user data and preferences in order to generate recommendations.

They can be used to acquire a user interest profile in collaborative and feature-based recommender systems.

Pre-test questionnaires

To establish the user's background

- to place her within the population of interest
- to classify the user before the experiment (e.g. in a stereotype)
- or to use this information to find possible correlations after the experiment (e.g., computer skilled users could perform better, etc).

Post-test questionnaires
To collect structured information after the experimental session, or after having tried a system for a while.

Besides, post-test questionnaires can be exploited to compare the assumption in the user model to an external test.

Pre and post-test questionnaires
Exploited together to collect changes due to real or experimental user-system interaction.

For instance, in adaptive elearning systems, pre and post-test questionnaires can be exploited to register improvements in the student's knowledge.

Interviews

Interviews are used to collect self-reported opinions and experiences, preferences and behavioral motivations.

Interviews are more flexible than questionnaires and they are well suited for exploratory studies. Interviews can be structured, semistructured, and unstructured.

Interviews may be time-consuming and the interpretation of the answers may be subjective and hard to write down in a structured manner.

In both cases it is important to state or ask questions in as neutral a way as possible, in order not to introduce some form of bias. Also the rating scale (yes/no, five-point Likert scale, . . .) influences the answers.

User Observation Methods

This family of methods is based on direct or indirect user's observation. They can be carried out with or without predetermined tasks.

Think aloud protocols
Methods that make use of the user's thought throughout the experimental session, or simply while the user is performing a task.

The user is explicitly asked to think out loud when she is performing a task in order to record her spontaneous reactions.

The main disadvantage of this method is that it disturbs performance measurements. Another possible protocol is *constructive interaction*, where more users work collaboratively to solve problems at the interface.

User observation

A data collection method wherein the user's behavior is observed during an experimental session or in her real environment when she interacts with the system.

In the former case, the user's actions are usually quantitatively analyzed and measurements are taken, while in the latter case the user's performance is typically studied from a qualitative point of view.

Logging use

Can be considered a kind of indirect observation and consists in the analysis of log files that register all the actions of the users.

The log files analysis shows the real behavior of users and is one of the most reliable ways to demonstrate the real effectiveness of user modeling and adaptive solutions

Analyzing and Interpreting Data

After having conducted the evaluation, you probably have data derived from questionnaires, interviews, observations or logging mechanisms. This data is commonly analyzed in two different ways:

Descriptive statistics

Descriptive statistics describe the main features of a collection of data quantitatively. Descriptive statistics aim to summarize a data set, rather than use the data to learn about the population that the data are thought to represent.

Descriptive statistics are important, as they summarize how well users appreciated a system, how many tasks they completed, how much time they needed, etcetera. This forms a base for interpretation of the inferential statistics.

Inferential statistics

Inferential statistics is the process of drawing conclusions from data that are subject to random variation, for example, observational errors or sampling variation. Inferential statistics are used to prove hypotheses.

Inferential statistics are used for finding out differences between two (or more) groups of users. These groups may be explicitly created in a controlled experiment or based on independent variables in a observational study (e.g. it may turn out that females seem to perform consistently better than males; this can be verified by comparing both gender groups).

Descriptive Statistics

A first step in data analysis is *exploring* the data and summarizing it in descriptive statistics. We are interested in:

- the distribution of the data
- the mean and standard deviation

For an overview, a plot of the frequency distribution is useful.

The normal distribution

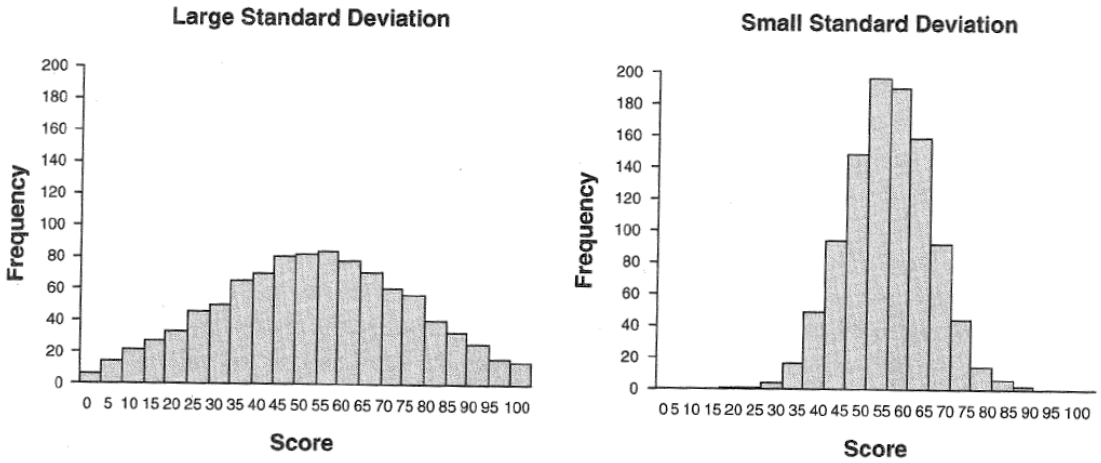


Figure 1.7 Two distributions with the same mean, but large and small standard deviations

The data in the graph above follows a *normal* distribution, which is characterized by a bell-shaped curve that you can plot on top of the data.

When is a distribution normal?

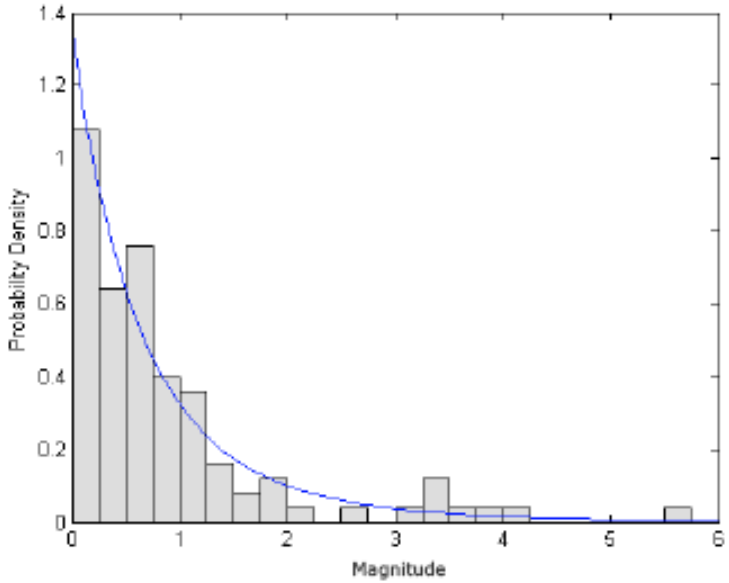
Normal distributions may be (a bit) positively or negatively skewed.

In computer science, we typically 'decide' whether a distribution can be considered normal or not, based on inspecting the plot.

A more objective approach would be to use a test that compares the scores in the sample to a normally distributed set of scores.

- The most common method is the *Kolmogorov-Smirnov* test (available in any statistical package, not further explained in this lecture)

The Power Law Distribution - far from normal



A famous power-law distribution is the Pareto distribution, also called the '80-20 rule'.

- 20% of active forum users are responsible for 80% of the activity
- 20% of the population controls 80% of the wealth

Descriptive statistics for normal distributions

For normal or normal-like distribution it is common to report the mean $\overline{x}$ and the standard deviation σ.

The mean is the sum of all scores divided by the number of scores.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The standard deviation is a measure for variance in the data (how far is the spread from the mean).

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Descriptive statistics for non-normal distributions

If the data does not follow a normal distribution, you might want to report other statistics instead:

- the *median* is the middle score of a distribution of scores when they are ranked in order of magnitude
- the *mode* is the single most common score

More importantly, if you have normally distributed data, you can use *parametric* inferential statistics. Otherwise, you need to use *non-parametric* inferential statistics.

Inferential statistics

Correlations

In observational studies, one is often interested in finding *correlations* between observed variables.

The example below is an evaluation of the PivotBar, a dynamic toolbar that provides contextual recommendations for Web pages or sites to be revisited. We tested the assumption that better recommendations go hand-in-hand with better take-up.

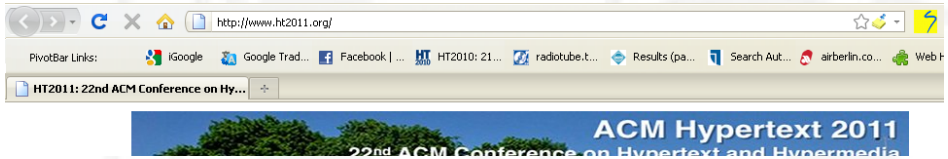


Figure: The PivotBar - a Firefox extension with contextual recommendations for revisitation

On average, 12.1% (σ=7.3) of all revisits resulted from a click on the PivotBar.

The average percentage of blind hits was 18.1% (σ=12.0), meaning that these revisits were suggested in the PivotBar but not triggered by it.

The strong correlation between the PivotBar clicks and blind hits (r=0.92, $p < 0.01$) suggest a direct connection between the quality of recommendations and the take-up of the tool.

User	Total Visits	Revisit (%)	PivotBar (%)	BlindHits (%)
1	603	50.1	30.8	22.8
2	535	45.0	19.5	51.0
3	445	39.6	15.9	8.5
4	578	51.2	15.9	15.9
5	1,111	36.1	13.0	20.7
6	716	45.5	12.3	28.8
7	1,219	49.1	8.8	18.0
8	899	41.7	8.8	8.5
9	379	56.2	7.0	11.7
10	1,047	39.6	5.8	16.1
11	1089	43.3	4.7	7.6
12	674	29.4	11.1	6.6
13	896	34.6	3.9	19.0

Table: Click data during the evaluation period.

Caution

- Correlational research does not allow causal statements to be made
- If you have many measures to compare, chances are odd that two or more measures correlate *by chance*

Pearson correlation

The most common parametric correlation measure for ratings is the Pearson correlation measure:

$$r(x,y) = \frac{\sum_1^n (x-\overline{x})(y-\overline{y})}{n * \sigma(x)\sigma(y)}$$

The standard deviation σ is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Spearman's rank correlation

The non-parametric **Spearman's rank correlation coefficient** ρ is used if the data does not follow a normal distribution. The definition is similar to the Pearson correlation. The only difference is that the original values are transformed into ranks and the correlations are computed on the ranks.

$$\rho(x,y) = \frac{\sum_1^n (rank(x) - \overline{rank(x)})(rank(y) - \overline{rank(y)})}{n * \sigma(rank(x))\sigma(rank(y))}$$

Kendall's tau

An alternative for Spearman's coefficient is **Kendall's tau** τ, which can be used if you have a small data set with a large number of tied ranks.

Let N be the number of joint observations x and y. Let C be the number of *concordant pairs*, pairs of any two of the N items (x,y) for which yields that both $x_i > x_j$ and $y_i > y_j$ (or $x_i < x_j$ and $y_i < y_j$).

And let D be the number of *discordant pairs*, pairs of items for which the above does not yield. Kendal's tau is defined as the difference between the concordant and discordant pairs, divided by all possible item pairs.

$$\tau = \frac{C - D}{\frac{1}{2}N(N-1)}$$

Comparing Two Means: the T-Test

In controlled experiments - for example when you compare two recommender algorithms - two types of variation can be observed:

- **Unsystematic variation** is the natural variation that occurs due to natural differences between participants that you haven't controlled for (e.g. their attitude with respect to recommendations in general)
- **Systematic variation** is due to the assignment of the participants in one condition and not in the other (e.g. one of the two algorithms to be compared)

The **t-test** is a parametric test that checks how likely it is that the differences between two groups are caused by systematic variation (and not due to natural variance in the data).

The *Dependent t-test* is used when you used the same participants in both conditions (repeated-measures design).

$$t_{dep} = \frac{(\overline{x} - \overline{y}) - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

where μ_0 is the expected difference between both groups (usually 0!).

($\frac{\sigma}{\sqrt{n}}$ is called the *standard error* - a measure for how likely it is that the mean will change if you would conduct the experiment with more participants)

The *Independent t-test* is used when you have different participants assigned to each condition.

The main difference is that the standard error is calculated from the variance in each condition independently (still, the variance is assumed to be /emphroughly equal in both conditions).

$$t_{indep} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{{\sigma_x}^2}{n_x} + \frac{{\sigma_y}^2}{n_y}}}$$

(as μ_0 usually is 0, this is left out of the equation for simplicity)

However, the above formula assumes that the sample sizes are equal ($n_x = n_y$). Often, this is not the case. Therefore, a better approach is to weight the variance by the size of the sample on which it is based:

$$\sigma_p{}^2 = \frac{(n_x - 1)\sigma_x{}^2 + (n_y - 1)\sigma_y{}^2}{n_x + n_y - 2}$$

The resulting weighted average variance is then just replaced in the equation for the independent t-test:

$$t_{indep-corr} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_p{}^2}{n_x} + \frac{\sigma_p{}^2}{n_y}}}$$

Degrees of freedom and significance

The degree of freedom relates to the number of observations that are free to vary in order to keep the mean constant.

If we have a mean based on N samples, we can exchange $N-1$ of these samples, but these $N-1$ samples determine the value that the N^{th} sample should have in order to keep the mean constant.

- For the dependent t-test, the degree of freedom is $n-1$
- For the independent t-test, the degree of freedom is $n_x + n_y - 2$

With the t-value and the degrees of freedom, the significance p of the test can be calculated or looked up in a distribution table. $p < 0.05$ is considered weakly significant, $p < 0.01$ as significant.

Distribution tables are old-fashioned. Use a statistical package instead.

What if the data is not normally distributed?

Use non-parametric tests instead (which work about the same as the t-test):

- instead of the dependent t-test: use the Wilcoxon signed-rank test
- instead of the independent t-test: use the Wilcoxon rank-sum test or the Mann-Whitney test

The logic behind both tests is: first rank the data *ignoring the condition to which a person belonged* from lowest to highest. If there is no difference between the conditions, you would expect to find a similar number of high and low ranks in each condition: the summed total of ranks in each group will be (about) the same.

Some Key Issues in the Evaluation of Adaptive Systems

Specification of Control Conditions

A problem that is inherent in the evaluation of adaptive systems, occurs when the control conditions of experimental settings are defined.

In many studies, the adaptive system is compared to a *non-adaptive version* of the system.

However, adaptation is often an essential feature of these systems and switching the adaptivity off might result in an absurd or useless system

A preferred strategy might be to compare a set of different adaptation decisions (as far as applicable).

Based on the same inferred user characteristics the system can be adapted in different ways.

- For instance, an adaptive learning system that adapts to the current knowledge of the learner might use a variety of adaptation strategies, including link annotation, link hiding, or curriculum sequencing.

However, the variants should be *as similar as possible* in terms of functionality and layout (often referred to as *ceteris paribus*, all things being equal) in order to be able to trace back the effects to the adaptivity itself.

Sampling

A proper experimental design requires not only to specify control conditions but also to select adequate samples.

On the one hand the sample should be very *heterogeneous* in order to maximize the effects of the system's adaptivity:

- the more differences between users, the higher the chances that the system is able to detect these differences and react accordingly.

On the other hand, from a statistical point of view, the sample should be very *homogeneous* in order to minimize the secondary variance and to emphasize the variance of the treatment.

A common strategy to reduce undesired variance is using *repeated measurement* (within-group design).
The main advantages of this kind of experimental design include:

- fewer participants are required
- statistical analysis is based on differences between treatments rather than between groups that are assigned to different treatments.

However, this strategy is often not adequate for the evaluation of adaptive systems, because of *carr-over effects*.

Definition of criteria

The criteria usually taken in consideration for evaluation (e.g., task completion time, number of errors, number of viewed pages) sometimes do not fit the aims of the system.

- For the evaluation of a recommender system, the relevance of the information provided is more important than the time spent to find it.
- lots of applications are designed for long-time interaction and therefore it is hard to correctly evaluate them in a short and controlled test.