Module IN 2111

3D User Interfaces
- Dreidimensionale Nutzerschnittstellen -

Prof. Gudrun Klinker

Evaluation of User Interfaces
SS 2023

...

Literature

Related Work:

- Ben Shneiderman and Catherine Plaisant: *Designing the User Interface, Strategies for Effective Human-Computer Interaction*, 4th edition, Addison Wesley, 2005, (http://wps.aw.com/aw_shneider_dtui_4).
- D. Bowman, E. Kruijff, J. LaViola Jr., I. Poupyrev: *3D User Interfaces, Theory and Practice*, Addison Wesley, 2004.
- J. Bortz: Statistik für Human- und Sozialwissenschaftler, 6. Auflage, Springer, 2004.
- J. McClave, F. Dietrich II: Statistics, Dellen Publishing Company, 1985.
- J.E. Swan II, S.R. Ellis, B.D. Adelstein, Conducting Human-Subject Experiments with Virtual and Augmented Reality, VR 2007 Tutorial, (http://www.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2007-Tutorial.pdf).

Agenda

1. Introduction
2. Evaluation Design
3. Usability Testing
4. Statistics Tools

1 Introduction

Purposes of Evaluations

- Definition
  Analysis, assessment, and testing of an artifact
- Iterative approach
  Design – evaluation – redesign – …
- Goals
  - Problem identification
  - Redesign
  - Understanding of usability
    (to obtain design guidelines)
  - Development of performance models
    (to predict user performance)

1 Introduction

Terminology

- *Usability*
  Encompasses everything about an artifact and everything that affects the person‘s use of the artifact
- *Evaluation*
  Measures some aspects of the usability of an interface:
  - System performance, task performance, user preference

Agenda

1. Introduction
2. Evaluation Design
3. Usability Testing
4. Statistics Tools

2. Evaluation Design

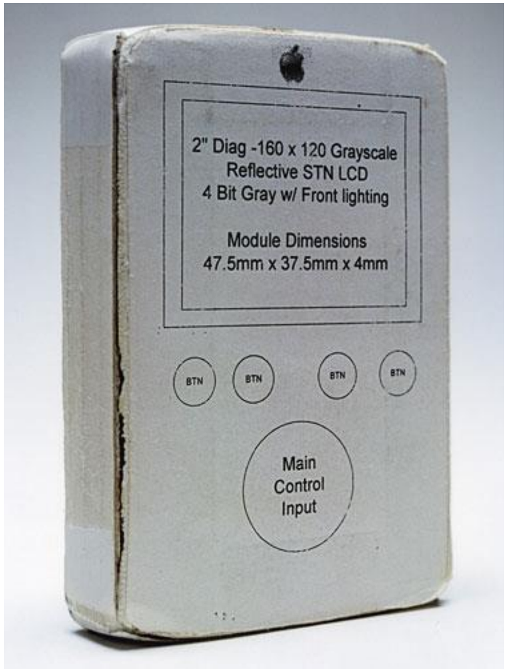
2.1 Planning an Evaluation

- User task analysis
  - Generate lists of detailed task descriptions, sequences, relationships, user work information flow
- Representative scenarios
  - Must be accurate and complete
  - More than simple atomic, mechanical or physical-level tasks
  - Should include high-level, cognitive, problem-solving tasks

2.1 Planning an Evaluation

- Prototyping
  - Paper-based sketch
  - Storyboard
  - Static mockup

  - Wizard of Oz (WOZ) technique:
    Human as a substitute for missing functionality
    (e.g. speech recognition)



Apple via Wired Magazine

2. Evaluation Design

2.1 Planning an evaluation

→ 2.2 Evaluation approaches

2.3 Evaluation metrics

2.2 Evaluation Approaches

Phases and Strategies

- P1: before development
  - Expert reviews
  - User surveys
- P2: during development
  - Usability testing and laboratories
  - User surveys
- P3: after development
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- Phase 1
  - Cognitive walkthrough
  - Heuristic evaluation
- Phase 2
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- Phase 3
  - Questionnaires
  - Interviews and demos

2.2 Evaluation Approaches

Phases and Strategies

- **P1: before development**
  - Expert reviews
  - User surveys
- P2: during development
  - Usability testing and laboratories
  - User surveys
- P3: after development
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- **Phase 1**
  - Cognitive walkthrough
  - Heuristic evaluation
- Phase 2
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- Phase 3
  - Questionnaires
  - Interviews and demos

2.2.1 Phase 1: Before Development

Expert Reviews

- Half day to one week effort
  - A lengthy training period may sometimes be required to explain the task domain or operational procedures.
- Can be scheduled at several points in the development process
  - When experts are available
  - When the design team is ready for feedback.

2.2.1 Phase 1: Before Development

Expert Reviews

- Different experts tend to find different problems in an interface. 2-3 expert reviewers can be highly productive.
- Danger: Experts may not have an adequate understanding of the task domain or the user communities.
- For successful expert reviews: Choose knowledgeable experts who are familiar with the project situation and who have a longer term relationship with the organization.
- Problem: Even experienced expert reviewers know little about how typical users, especially first-time users will really behave.

2.2.1 Phase 1: Before Development

Cognitive Walkthrough

- Step through common tasks that a user would perform
- Evaluate the interface‘s ability to support each step

2.2.1 Phase 1: Before Development

Heuristic Evaluation

Guidelines-based expert evaluation

- Experts apply a set of heuristics or design guidelines
- NO representative users

2.2 Evaluation Approaches

Phases and Strategies

- **P1: before development**
  - Expert reviews
  - User surveys
- P2: during development
  - Usability testing and laboratories
  - User surveys
- P3: after development
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- **Phase 1**
  - Cognitive walkthrough
  - Heuristic evaluation
- Phase 2
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- Phase 3
  - Questionnaires
  - Interviews and demos

2.2.1 Phase 1, 2 and 3

User Surveys

- Familiar, inexpensive and general
  - Complementary to usability tests and expert reviews
- Important:
  - Clear goals in advance
  - Development of focused items that help attain the goals
- Users could be asked for their subjective impressions about specific aspects of the interface.
- ! Many people prefer answering a brief survey displayed on a screen, instead of filling in and returning a printed form !

2.2 Evaluation Approaches

Phases and Strategies

- P1: before development
  - Expert reviews
  - User surveys
- **P2: during development**
  - Usability testing and laboratories
  - User surveys
- P3: after development
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- Phase 1
  - Cognitive walkthrough
  - Heuristic evaluation
- **Phase 2**
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- Phase 3
  - Interviews and demos

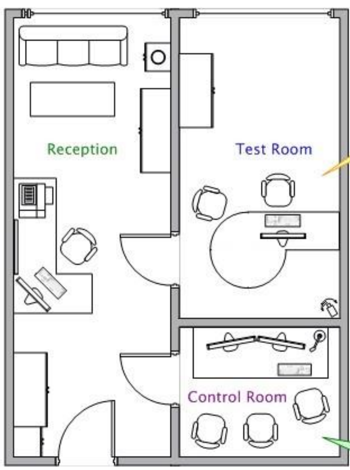
2.2.2 Phase 2: During Development

Usability Testing and Laboratories

- Increasingly important since the early 1980s
- Speed up in many projects plus dramatic cost savings
- Specially constructed usability laboratories
  - Typical setup: two 10 by 10 foot areas, one for the participants to do their work and another, separated by a half-silvered mirror, for the testers and observers (designers, managers, and customers).

2.2.2 Phase 2: During Development

Usability Testing and Laboratories

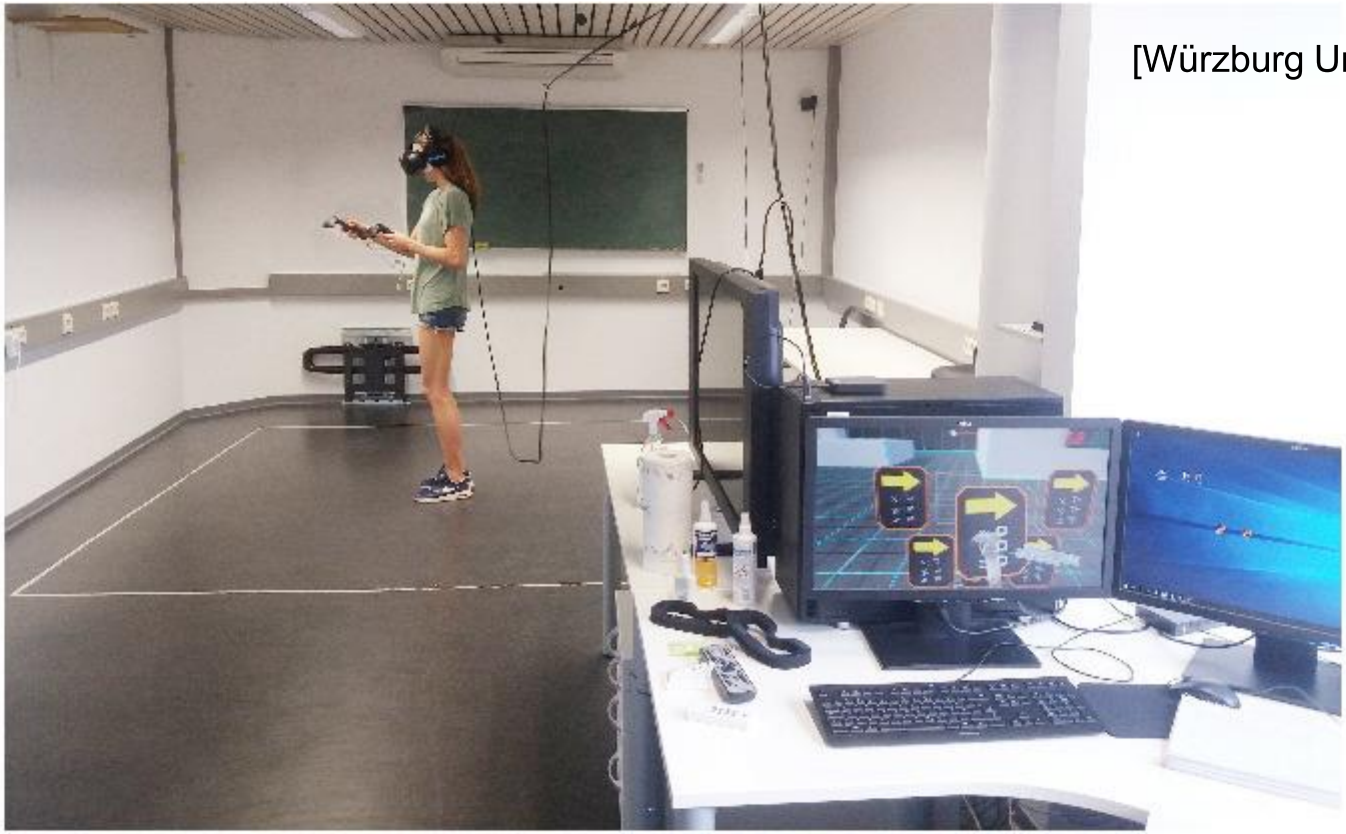


[interUX]

[fio hotJar]

2.2.2 Phase 2: During Development

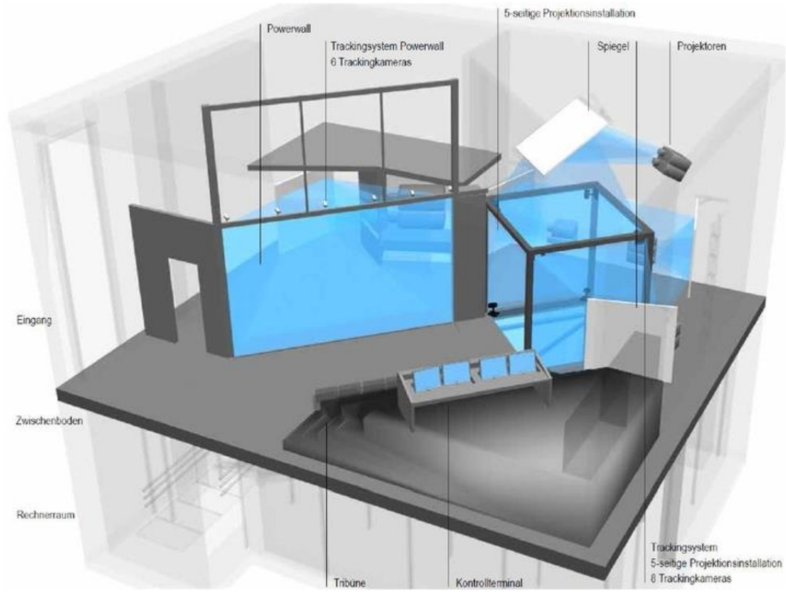
Usability Testing and Laboratories



[Würzburg University]

2.2.2 Phase 2: During Development

Usability Testing and Laboratories



[VR Cave,
Leibniz Supercomputing Centre LRZ]

2.2.2 Phase 2: During Development

Usability Testing and Laboratories

- Participants should represent the intended user communities
  - Background in computing
  - Experience with the task
  - Motivation
  - Education
  - Knowledge of the natural language used in the interface
- Limitations
  - Emphasizes first-time usage
  - Limited coverage of the interface features
- Important
  - Detailed logging/videotaping during user tests

2.2.2 Phase 2: During Development

Formative Evaluation (*„Discount (cheap) usability testing“***)**

- Observational, empirical method
- Applied during evolving stages of design
- Assess user interaction by interactively placing representative users in task-based scenarios
- Goals
  - Identify problems
  - Assess design‘s ability to support user exploration, learning and task performance
- Informal .. Very formal and extensive
  - Qualitative results: critical incidents, user comments, general reactions
  - Quantitative results: task timing, errors

2.2.2 Phase 2: During Development

Summative Evaluation (Competitive usability testing)

- Statistical comparison of two or more configurations of UI designs, UI components, and/or UI techniques
- Representative users perform task scenarios
- Formal or informal

- Generally performed after UI designs are complete
- Factorial experimental design with multiple independent variables
- Helps evaluators compare the productivity and cost benefits associated with different UI designs
- Requires consistent set of task scenarios that compare a design‘s support for specific user task performance

2.2.2 Phase 2: During Development

Questionnaires

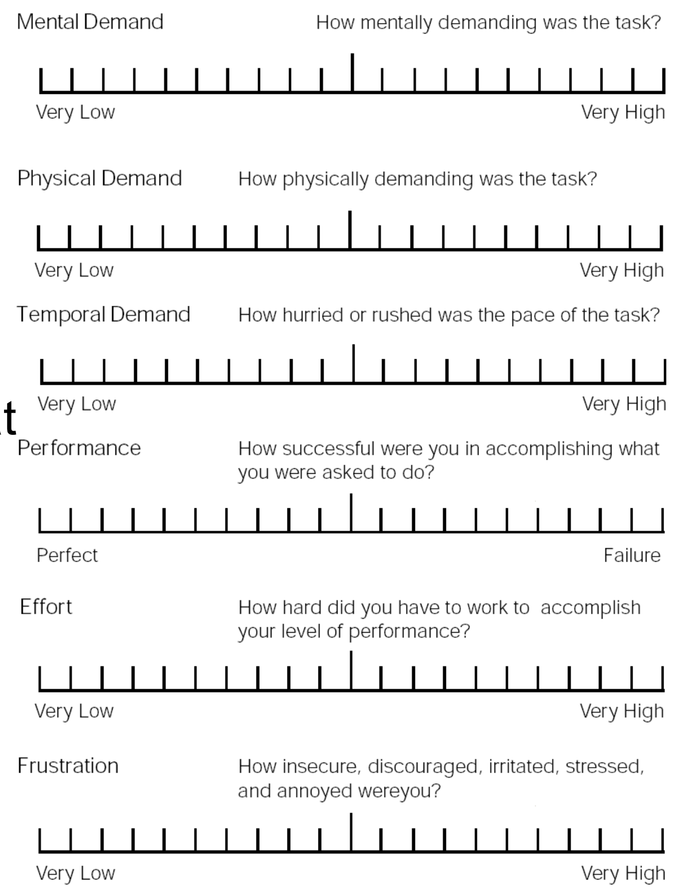
- Written set of questions
- Given to users before, in between or after they have participated in a usability evaluation session
- Demographic information, subjective data
- Examples
  - SUS
  - NASA-TLX

- Used frequently
  - Help to find out about degree of presence, cyber sickness

2.2.2 Phase 2: During Development

Mental workload

- NASA-TLX (Task Load Index)
  - Subjective workload assessment tool
  - For various human-machine systems
  - Multi-dimensional rating procedure
  - Score based on a weighted average of rat
    - Mental Demands
    - Physical Demands
    - Temporal Demands
    - Own Performance
    - Effort
    - Frustration



SG Hart, LE Staveland. **Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,** Human Mental Workload, 1988
NASA TLX Homepage: http://humansystems.arc.nasa.gov/groups/TLX/

2.2.2 Phase 2: During Development

System Usability Scale (SUS)

- John Brooke, 1986
- Subjective determination of usability (estimation based on 10 questions) to determine:
  - Effectiveness (users can achieve their goals)
  - Efficiency (Little effort is required for the achievement of these goals)
  - Satisfaction (the experience was satisfactory)
- Score in range 0..100 (raw score * 2.5)

	Strongly disagree				Strongly agree
I think that I would like to use this system frequently	○	○	○	○	○
I found the system unnecessarily complex	○	○	○	○	○
I thought the system was easy to use	○	○	○	○	○
I think that I would need the support of a technical person to be able to use this system	○	○	○	○	○
I found the various functions in this system were well integrated	○	○	○	○	○
I thought there was too much inconsistency in this system	○	○	○	○	○
I would imagine that most people would learn to use this system very quickly	○	○	○	○	○
I found the system very cumbersome to use	○	○	○	○	○
I felt very confident using the system	○	○	○	○	○
I needed to learn a lot of things before I could get going with this system	○	○	○	○	○

J Brooke. **System Usability Scale (SUS): A Quick-and-Dirty Method of System Evaluation User Information**, Digital Equipment Co Ltd, Reading, UK, 1986

2.2 Evaluation Approaches

Phases and Strategies

- P1: before development
  - Expert reviews
  - User surveys
- **P2: during development**
  - Usability testing and laboratories
  - User surveys
- P3: after development
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- Phase 1
  - Cognitive walkthrough
  - Heuristic evaluation
- **Phase 2**
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- Phase 3
  - Questionnaires
  - Interviews and demos

2.2.2 Phase 2: During Development

- **Interviews**
  - Oral conversation with users
    - Can gather more information than a questionnaire
    - Useful for getting subjective reactions, opinions, insights about users‘ reasoning
  - Structured interviews
    - Predefined set of questions and responses
  - Open-ended interviews
    - Interviewees can provide additional information
    - Interviewer can ask broad questions
- **Demos**
  - Often shown in conjunction with user interviews

Interviews and demos often used at the end of formative or summative usability evaluations

2.2 Evaluation Approaches

Phases and Strategies

- P1: before development
  - Expert reviews
  - User surveys
- P2: during development
  - Usability testing and laboratories
  - User surveys
- **P3: after development**
  - Acceptance tests
  - Evaluation during active use
  - User surveys

Methods

- Phase 1
  - Cognitive walkthrough
  - Heuristic evaluation
- Phase 2
  - Formative evaluation
  - Summative evaluation
  - Questionnaires
  - Interviews and demos
- **Phase 3**
  - Interviews and demos

2.2.3 Phase 3: After Development

Acceptance Tests

- For large implementation projects
  - In large systems: 8-10 such tests should be carried out on different components of the interface and with different user communities.
- Test after product completion
  - Further field testing before national or international distribution
- Goal: Force as much of the evolutionary development as possible into the prerelease phase, when change is relatively easy and inexpensive to accomplish.

2.2.3 Phase 3: After Development

Evaluation during active use

- Interviews and focus group discussions
  - Interviews with individual users to pursue specific issues
  - Group discussions to ascertain the universality of comments
- Continuous user-performance data logging
  - Guidance to system maintainers in optimizing performance and reducing costs for all participants
- Online or telephone consultants
- Online suggestion box or trouble reporting
- Online bulletin board or newsgroup
- User newsletters and conferences

Overview

2. Evaluation Design

2.1 Planning an evaluation

2.2 Evaluation strategies and phases (approaches)

→ 2.3 Evaluation metrics

2.3 Evaluation Metrics

→ 2.3.1 System Performance Metrics

2.3.2 Task Performance Metrics

2.3.3 User Preference Metrics

2.3.1 System Performance Metrics

- Average frame rate
- Average latency
- Network delay
- Optical distortion
- …

- Only important insofar as they affect users‘ experience or task performance

2.3.2 Task Performance Metrics

- Time to navigate to a specific location
- Accuracy of object placement
- Number of errors a user makes in selecting an object from a set
- Speed of learning a concept
- Spatial awareness
- …

- Problem: users cannot optimize simultaneously for both speed and accuracy

2.3.3 User Preference Metrics

- Subjective perception of the interface by the user
  - Perceived ease of use
  - Ease of learning
  - Satisfaction
- Obtained by questionnaires or interviews

2.3.3 User Preference Metrics

- Presence (the „feeling of being there“) (3D UIs)
  - User rating on a given scale
  - Physiological measurements
  - User‘s reactions to events
  - Test of memory for environment and objects
- User comfort
  - Simulator sickness
  - Physical aftereffects of being exposed to 3D systems
  - Subjective measures (rating scales)

Agenda

1. Introduction
2. Evaluation Design
3. Usability Testing
4. Statistics Tools

3. Usability Testing

→ 3.1 Definition

3.2 Testing process

3.3 Experimental structure

3.1 Definition

Definition

Usability is the measure of the quality of the user experience when interacting with something
– whether a Web site, a traditional software application, or any other device the user can operate in some way or another [J. Nielson].

Ergonomic requirements (DIN ISO EN 9241):

- **Pragmatic quality** (PQ)
  - **Effectiveness** (accuracy, completeness of task performance)
  - **Efficiency** (operating expense, speed)
- **Hedonic quality** (HQ): emotional, satisfactory experience
  - **Satisfaction** (freedom of interference, positive attitude of using a product)

3.2 Testing Process

- Quantitative procedure to test against predefined goals
  e.g.: Initial performance, long-term performance, learnability, memorability, most-used feature set, first impression, long-term satisfaction)
  - Reference values
  - Actual values (test data)
- Process of testing a UI (randomized tests)
  - Preparation
  - Introduction of test procedure to subjects
  - Test
  - Final discussion with subjects
  - Analysis
    - Quantitative, qualitative or subjective data
  - Report

3.2 Testing Process

Measurement function: $t = f(x, y, z, \ldots)$

- Independent variables (*x*, *y*, *z*,…) “**factors**”, parameters
  - Objects / systems under investigation (e.g., UI alternatives)
  - Variables can assume any value („**level**“) within the defined range
- **Dependent variables** (*t*, …)
  - Measured attributes / properties of the system (e.g., task completion time, error counts, survey answers, scores, …)
  - Functions of the independent variables
- **Confounding factors**
  - Additional factors (e.g., fatigue, learning) that can have an unintended influence on the dependent variables

3.2 Testing Process

Factorial design

- One-factor tests:
  - Only one independent variable *x* with l_x levels
    $\rightarrow l_x$ UI alternatives

Example:
Input device ε {mouse, arrows on keyboard, tangible 3D-object}
→ 3 UI alternatives

3.2 Testing Process

Factorial design

- Two-factor tests:
  - Two independent variables, *x* with l_x levels and *y* with l_y levels
    $\rightarrow l_x * l_y$ UI alternatives

Example:
Input device ε {mouse, arrows on keyboard, tangible 3D-object}
Output ε {sound, 3D graphics}
→ 3*2 = 6 UI alternatives

- N-factor tests: N independent variables …

3.3 Experiment Design

Test designs (comparing several UI alternatives)

- ***Between-subject design***
  Users are divided into several groups, each working with one UI alternative (i.e., with a different level of the independent variables)

Sample process of testing a UI (per test person)

- Demographic questionnaire
- Explain test scenario to test person
- Explain UI to test person
  - If necessary, let them play with the UI
- Test
  - Quantitative test
  - Qualitative test (questionnaire)
  - Interview
- Final discussion

Note: not all steps are executed in every experiment

3.3 Experiment Design

Test designs (comparing several UI alternatives)

- ***Within-subject design***
  All users are exposed to all UI alternatives
  - Less subjects required
  - Larger statistical strength: no bias due to different users
  - But: carry-over effects (learning, fatigue)
  - Permutations of test sequences required (for counterbalancing)
    → requires n! subjects to test n UI alternatives
  - common simplification: (**Latin Squares** requiring only n subjects)

A	B	C
B	C	A
C	A	B

A	B	C
C	A	B
B	C	A

Example:
Input device ε {mouse, arrows on keyboard, tangible 3D-object}
Output ε {sound, 3D graphics}
→ 6 UI alternatives
→ 720 test subjects

Example 2:
Input device ε {mouse, tangible 3D-object}
Output ε {sound, 3D graphics}
→ 4 UI alternatives
→ 24 test subjects

3.3 Experiment Design

Test designs (comparing several UI alternatives)

- *Within-subject design*

Sample process of testing a UI (per test person)
- Demographic questionnaire
- Explain test scenario to test person
- For every UI alternative (in randomized order):
  - Explain UI alternative to test person
    - If necessary, let them play with the UI
  - Test
    - Quantitative test
    - Qualitative test (questionnaire)
    - Interview
- Final discussion

Note: not all steps are executed in every experiment

Agenda

1. Introduction
2. Evaluation Design
3. Usability Testing
4. Statistics Tools

4.1 General Approach

How can we decide whether two UI alternatives are „different“ with respect to some criterion x?

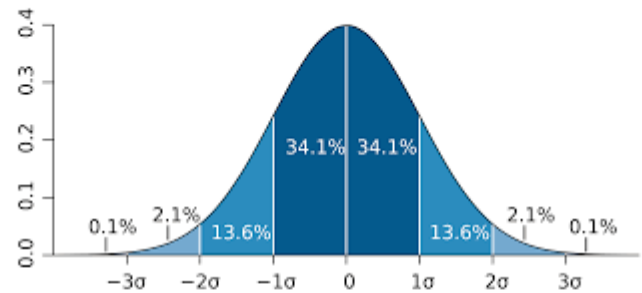
Hypothesis testing

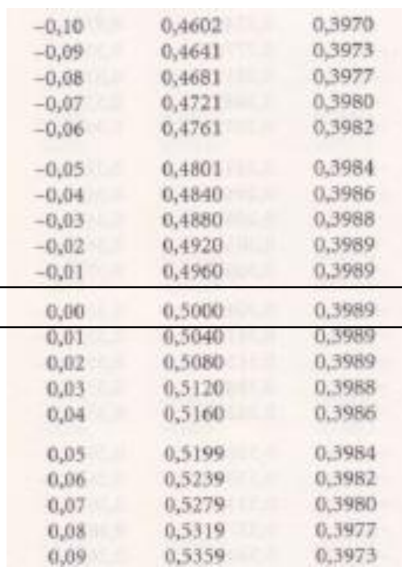
- Is UI *a* different from UI *b*?
- Is the difference significant?

4.2. Basics

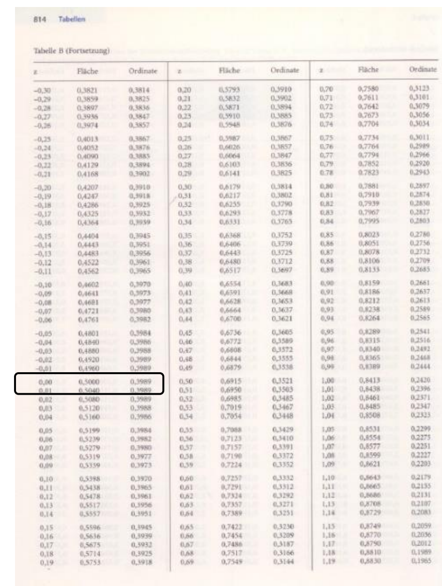
Simulations

- (Normal) population
  - Mean μ
  - Variance σ^2
  - Standard deviation σ





Z	area	ordinate
−0,10	0,4602	0,3970
−0,09	0,4641	0,3973
−0,08	0,4681	0,3977
−0,07	0,4721	0,3980
−0,06	0,4761	0,3982
−0,05	0,4801	0,3984
−0,04	0,4840	0,3986
−0,03	0,4880	0,3988
−0,02	0,4920	0,3989
−0,01	0,4960	0,3989
0,00	0,5000	0,3989
0,01	0,5040	0,3989
0,02	0,5080	0,3989
0,03	0,5120	0,3988
0,04	0,5160	0,3986
0,05	0,5199	0,3984
0,06	0,5239	0,3982
0,07	0,5279	0,3980
0,08	0,5319	0,3977
0,09	0,5359	0,3973

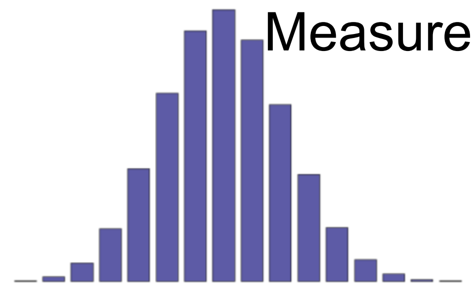


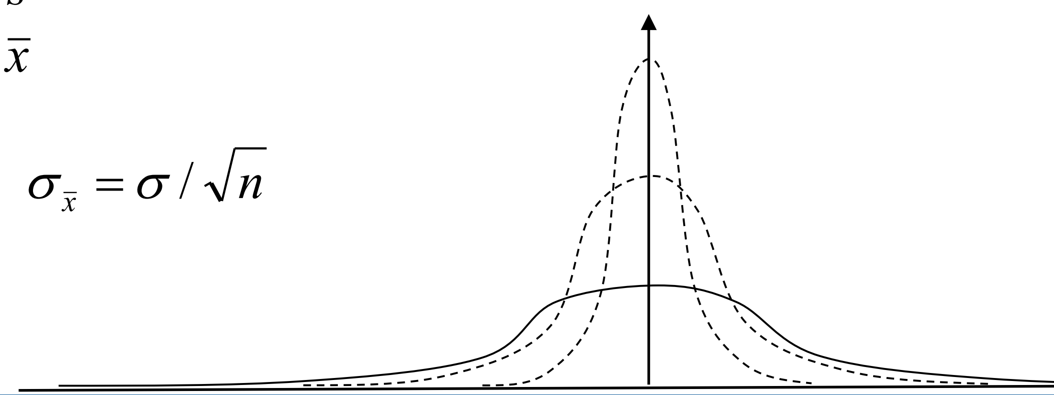
4.2 Basics

Simulations

- (Normal) population
  - Mean μ
  - Variance σ^2
  - Standard deviation σ
- Random samples
  - Observations n
  - Sample mean $\bar{x} = \sum x_i / n$
  - Sample variance $s^2 = \sum (x_i - \bar{x})^2 / (n-1)$
  - Sample standard deviation s
- Sampling distribution (s. d.) of $\bar{x}$
  - Mean of s. d. $\mu_{\bar{x}} = \mu$
  - Standard deviation of s. d. $\sigma_{\bar{x}} = \sigma / \sqrt{n}$
- Sample statistic
  - z, t, χ^2, F

Measurements





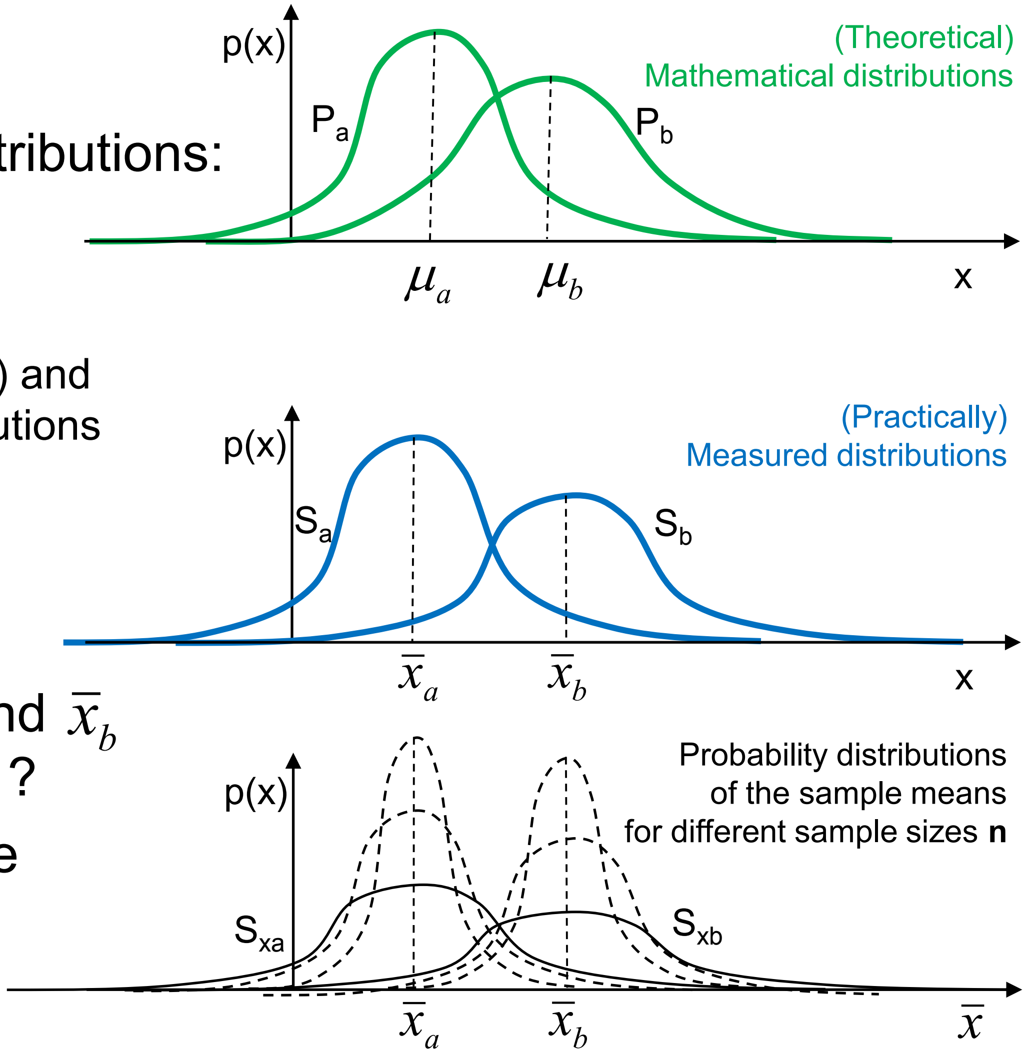
$\bar{x}$

4.2 Basics

- Two unknown probability distributions:

  $P_a : (\mu_a, \sigma_a^2) \qquad P_b : (\mu_b, \sigma_b^2)$
- Are they different?
  - Depends on shapes (variances) and positions (means) of the distributions
- Must be estimated from two sampling distributions:

  $S_a : (\bar{x}_a, s_a^2) \qquad S_b : (\bar{x}_b, s_b^2)$
- Are the sample means $\bar{x}_a$ and $\bar{x}_b$ good estimates of μ_a and μ_b?
- Probability distributions of the sample means



4.3 Test Procedure

- Two sampling distributions:

  $S_a : (\bar{x}_a, s_a^2), \; S_b : (\bar{x}_b, s_b^2)$

- Null hypothesis H_0:
  There is only one distribution

  $P_a : (\mu_a, \sigma_a^2) \quad \mu_a = \mu_b$

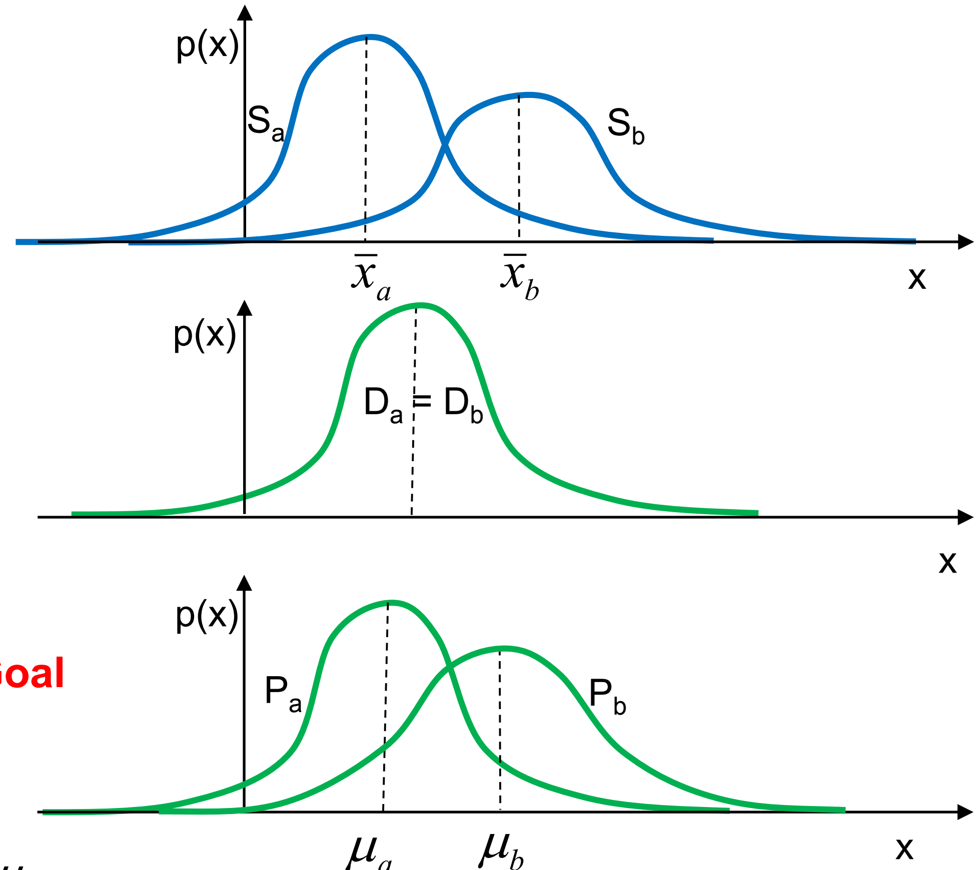
- Alternate hypothesis H_1:

  Distributions are different:

  $P_a : (\mu_a, \sigma_a^2) \quad P_b : (\mu_b, \sigma_b^2)$

  - Two-sided $\mu_a \neq \mu_b$
  - One-sided $\mu_a < \mu_b \quad or \quad \mu_a > \mu_b$

Goal



Indirect proof:
Prove H_1 by showing that it is extremely unlikely that H_0 is true.

4.3 Test Procedure

- Two sampling distributions:

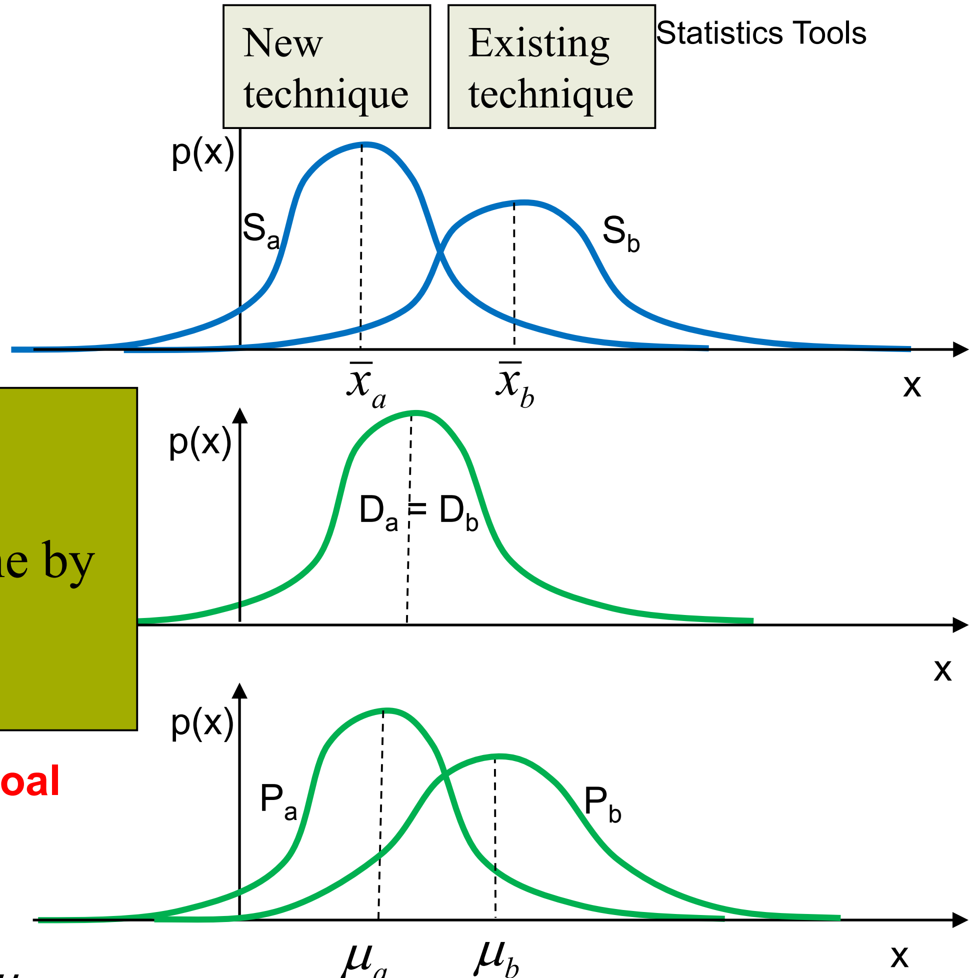
  $S_a : (\bar{x}_a, s_a^2),\ S_b : (\bar{x}_b, s_b^2)$

Example:
Prove that a new interaction technique is faster (or slower) than an existing one by showing that it is extremely unlikely that both are equally fast.

Distributions are different: **Goal**

$P_a : (\mu_a, \sigma_a^2) \qquad P_b : (\mu_b, \sigma_b^2)$

- Two-sided $\mu_a \neq \mu_b$
- One-sided $\mu_a < \mu_b \quad or \quad \mu_a > \mu_b$

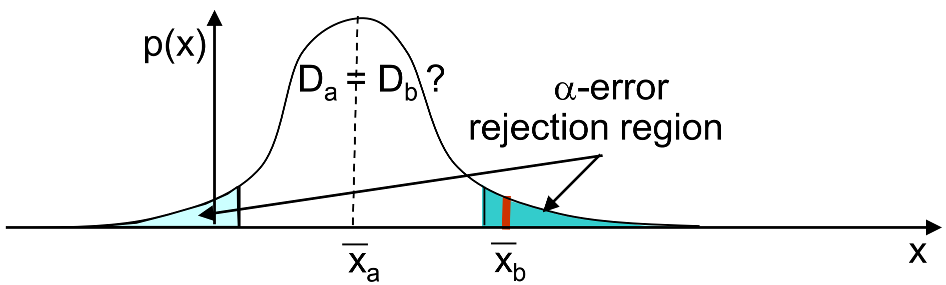


Indirect proof:
Prove H_1 by showing that it is extremely unlikely that H_0 is true.

4.3 Test Procedure

Indirect proof:
Prove H_1 by showing that it is extremely unlikely that H_0 is true.

- Evaluation of the null hypothesis H_0
  There is only one distribution: $\mu_a = \mu_b$

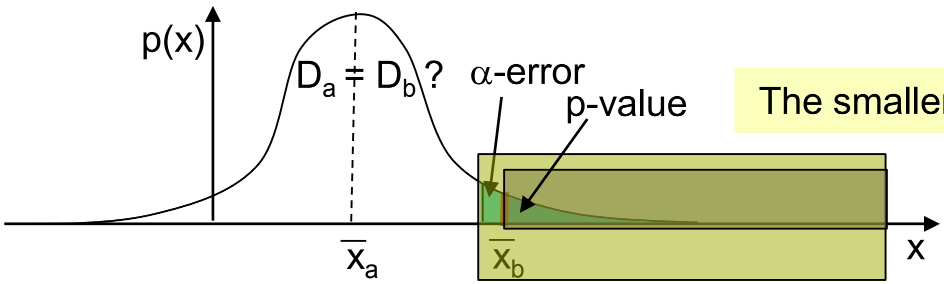


- Define a “rejection region” for H_0
- If x_b lies in the “rejection region”, it is highly unlikely that $\mu_a=\mu_b$
- → **Reject hypothesis H_0**: $\mu_a = \mu_b$ and thus conclude that $H_1 : \mu_a \neq \mu_b$ (or $\mu_a < \mu_b$ *or* $\mu_a > \mu_b$, resp.) is true
- Risk that we make a wrong decision (reject H_0 even though it is true): α-error (type I error)

Decision at significance level α

4.3 Test Procedure

α-error versus p-value



α-error:

- Fixed percentage (5% = 0.05) of the area under the curve
- Preselected value to indicate what kind of risk will be taken

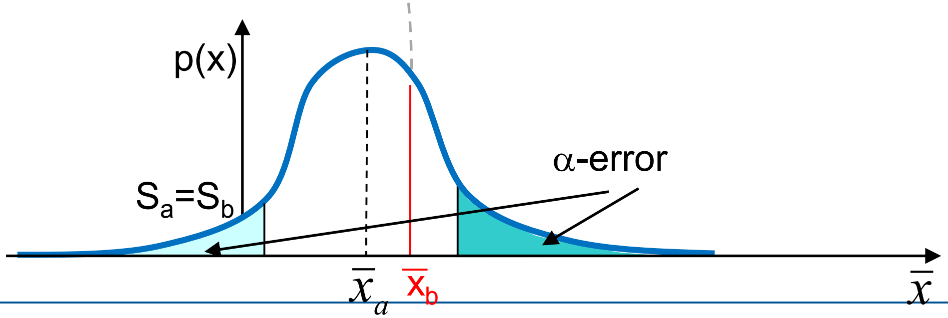
p-value:

- Computed percentage (depends on x_b) of the area under the curve
- If $p<\alpha$: reject Nullhypothesis

4.3 Test Procedure

α-error

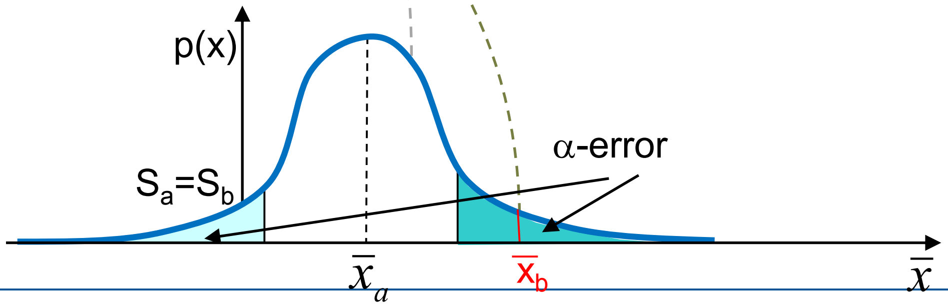
	"Real" H_0	"Real" H_1
Estimated H_0	Correct (wasted time)	?
Estimated H_1		



4.3 Test Procedure

α-error

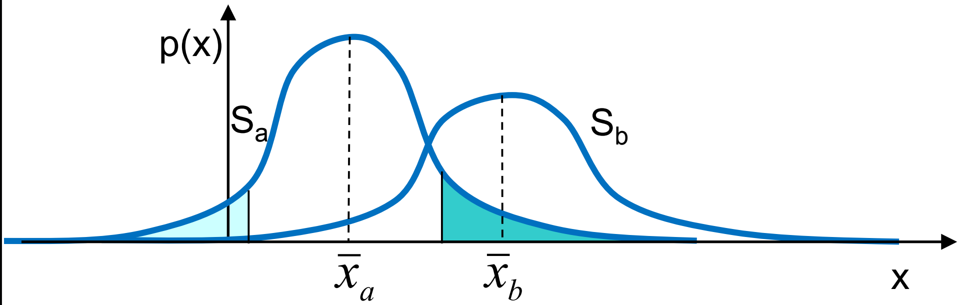
	"Real" H_0	"Real" H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct



4.3 Test Procedure

New technique | Existing technique | Statistics Tools

Example:
Prove that a new interaction technique is faster than an existing one (H_1) by showing that it is extremely unlikely that both are equally fast (H_0).



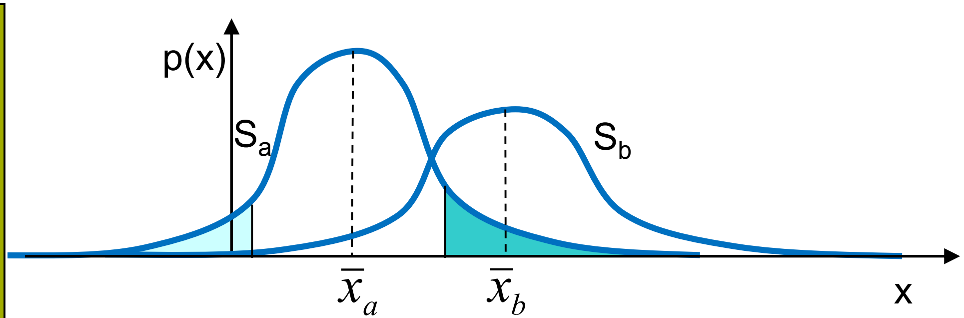
4.3 Test Procedure

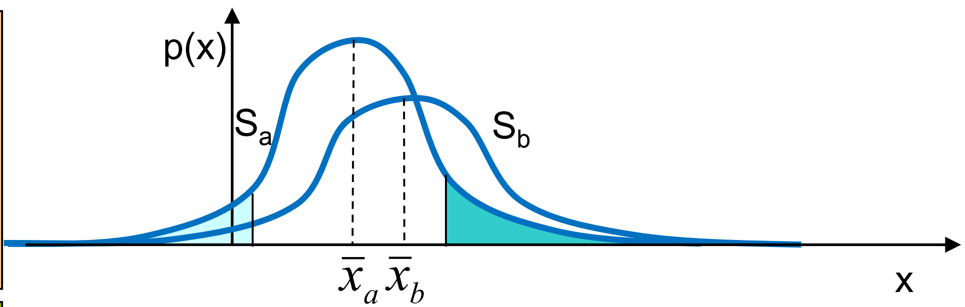
New technique | Existing technique | Statistics Tools

Example:
Prove that a new interaction technique is faster than an existing one (H_1) by showing that it is extremely unlikely that both are equally fast (H_0)

NOTE: WE CANNOT CONCLUDE THE INVERSE.
I.e.: If we cannot disprove H_0, we cannot conclude that H_1 is not true

Example:
If D_b is NOT in the rejection region, we cannot conclude that the new technique is NOT faster





New technique | Existing technique

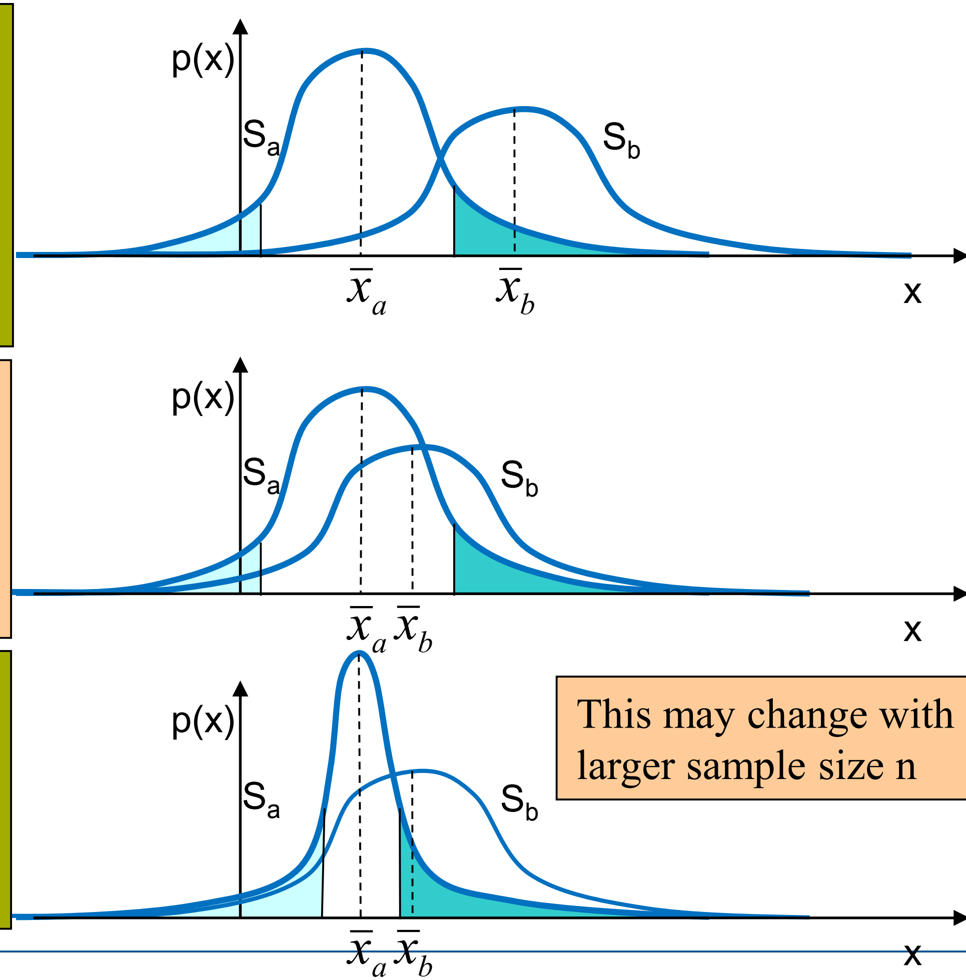
Statistics Tools

4.3 Test Procedure

Example:
Prove that a new interaction technique is faster than an existing one (H_1) by showing that it is extremely unlikely that both are equally fast (H_0)

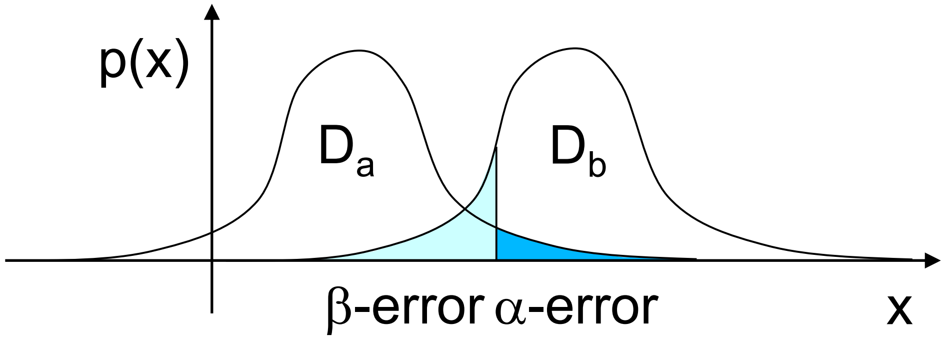
NOTE: WE CANNOT CONCLUDE THE INVERSE.
I.e.: If we cannot disprove H_0, we cannot conclude that H_1 is not true

Example:
If D_b is NOT in the rejection region, we cannot conclude that the new technique is NOT faster



4.3 Test Procedure

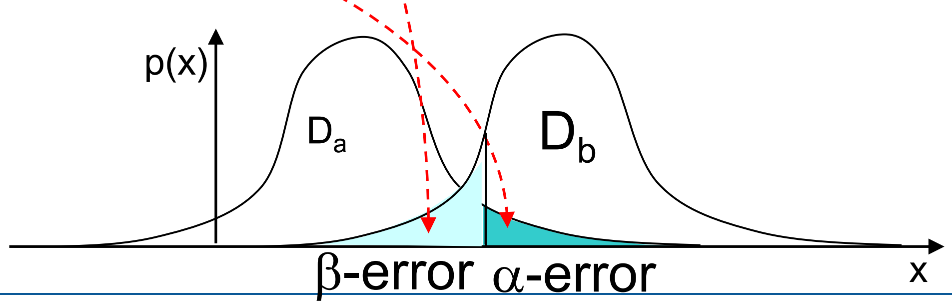
- Problem: You can avoid making mistakes by always accepting H_0 (by reducing α-error → 0) “no risk no fun”
- What if H_0 is wrongly accepted (should have been rejected)?



4.3 Test Procedure

β-error

	“Real” H_0	“Real” H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct



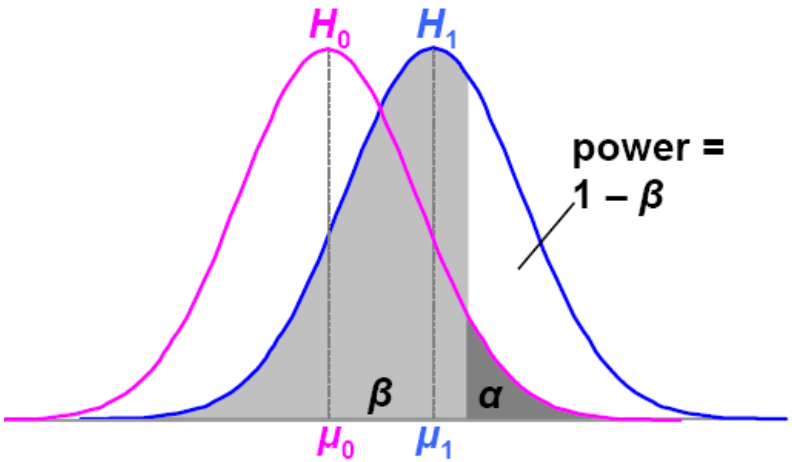
4.4 Interpretation of Test Results

	"Real" H_0	"Real" H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct

Power

Probability that a test will reject a false null hypothesis H_0

I.e.: that it will accept a correct alternate hypothesis H_1



Figures from Slides: J. Swan, S. Ellis, B. Adelstein, "Conducting Human-Subject Experiments with Virtual and Augmented Reality", IEEE VR 2007.

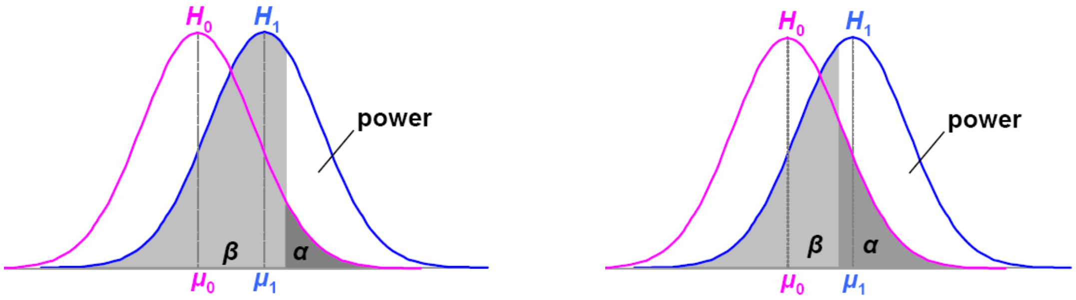
4.4 Interpretation of Test Results

	“Real” H_0	“Real” H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct

Dependence of power on **α-level**

Increasing $\alpha \rightarrow$

- Increasing power
- Decreasing type II error
- Increasing type I error



Figures from Slides: J. Swan, S. Ellis, B. Adelstein, “Conducting Human-Subject Experiments with Virtual and Augmented Reality”, IEEE VR 2007.

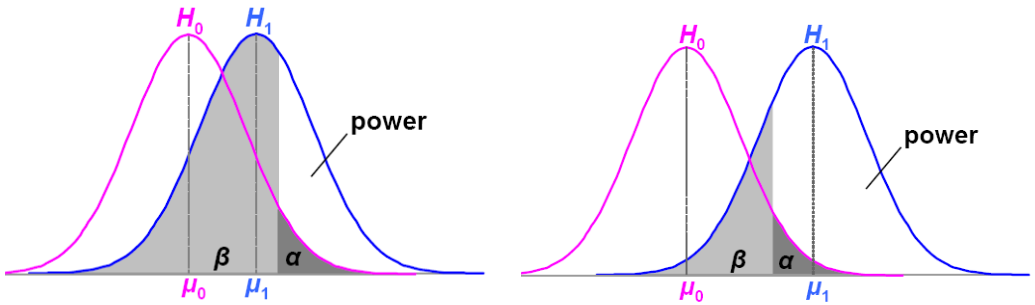
4.4 Interpretation of Test Results

	“Real” H_0	“Real” H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct

Dependence of power on **effect** (difference between μ_0 and μ_1)

Large effect →

- Increasing power
- Decreasing type II error
- Unaffected α and type I error



Figures from Slides: J. Swan, S. Ellis, B. Adelstein, “Conducting Human-Subject Experiments with Virtual and Augmented Reality”, IEEE VR 2007.

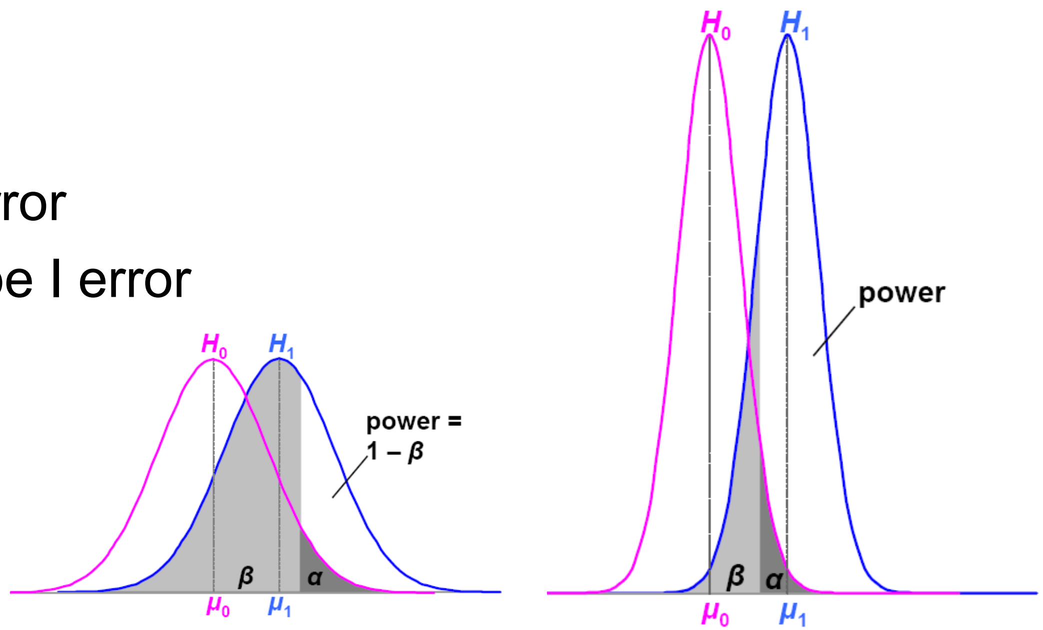
4.4 Interpretation of Test Results

	“Real” H_0	“Real” H_1
Estimated H_0	Correct (wasted time)	Type II error β-error
Estimated H_1	Type I error α-error	Correct

Dependence of power on **sample size n**

Increasing sample size n →

- Decreasing variance
- Increasing power
- Decreasing type II error
- Unaffected α and type I error



Figures from Slides: J. Swan, S. Ellis, B. Adelstein, “Conducting Human-Subject Experiments with Virtual and Augmented Reality”, IEEE VR 2007.

Thank you!