Das Usability-Experiment als Ergänzung zu typischen Usability- und A/B-Tests

Inferenzstatistisch abgesicherte Ergebnisse in kleinen Stichproben

Heinrich R. Liesefeld
Centigrade GmbH
Science Park 2
66123 Saarbrücken
rene.liesefeld@centigrade.de

Abstract
A/B-Tests ermöglichen es, Design-Varianten zu vergleichen und die Ergebnisse inferenzstatistisch abzusichern. Usability-Tests helfen zwar, effizient die gröbsten Usability-Probleme eines Interfaces aufzudecken, ermöglichen aber normalerweise keine solche inferenzstatistische Absicherung. Dieser Artikel macht deutlich, dass Inferenzstatistik für Usability-Engineers von größerer Bedeutung ist als gemeinhin angenommen: Es geht um nichts Geringeres als die Vermeidung berechtigten Misstrauens, also um Glaubwürdigkeit. Dafür benötigen A/B-Tests eine große Teilnehmerzahl und sind daher meist auf eine Testung über das Web angewiesen. Da dies in vielen Projekten (z. B. weil besondere Hardware benötigt wird) nicht möglich ist, soll mit dem Usability-Experiment eine Lücke im Methoden-Portfolio von Usability-Engineers geschlossen werden. Das Usability-Experiment wendet die über 100-jährige Erfahrung experimenteller Psychologen mit der Messung menschlichen Verhaltens auf Usability-Fragestellungen an. Es erlaubt (im Gegensatz zu Usability-Tests) den inferenzstatistisch abgesicherten Vergleich von Design-Alternativen mit (im Vergleich zu A/B-Tests) relativ kleinen Stichproben (ab ca. 10 Teilnehmern). Zudem eröffnet es eine Reihe interessanter neuer Einblicke ins Nutzerverhalten.

Keywords:
/// Usability-Testing
/// A/B-Tests
/// Inferenzstatistik
/// Experimentaldesign
/// kognitive Psychologie

1. Misstrauen in die Ergebnisse typischer Usability-Tests

Lassen Sie mich mit einem typischen Ergebnis aus einem typischen Usability-Test beginnen: Angenommen, Sie wollen entscheiden, ob Checkbox A in Abbildung 1 funktioniert, d. h. ob Nutzer die einzelnen Zustände verstehen und die Checkbox korrekt bedienen. Es stellt sich heraus, dass 4 der 5 Test-Teilnehmer direkt und ohne Ihre Hilfe mit Checkbox A umgehen können. Im nächsten Teammeeting präsentieren Sie stolz dieses eindeutige Ergebnis. Womit Sie nicht gerechnet haben, ist, dass Peter Meier, ein motivierter Kollege, seine eigenen Tests durchgeführt hat. Peter war maßgeblich an der Entwicklung von Checkbox B beteiligt. Sein Test hat nun ergeben, dass nur 2 von 5 Teilnehmern Checkbox A auf Anhieb verstehen, während 4 von 5 Teilnehmer Checkbox B direkt verstehen. Wer hat nun Recht? Welche Checkbox ist besser? Hat Peter vielleicht die Daten manipuliert, um seine präferierte Checkbox durchzusetzen? **[Abb. 1]**

Abb. 1.
Checkboxen A und B. Links im „Unchecked"- und rechts im „Checked"-Zustand.

Besonders die Vermutung, dass Daten manipuliert werden, schürt das Misstrauen in Tests und Umfragen und führt zu der Empfehlung: „Traue keiner Statistik, die du nicht selbst gefälscht hast!" – offensichtlich ein ernstzunehmendes Problem für Usability-Engineers; diese sind ja schließlich darauf angewiesen, dass Entscheidungsträger ihren Empfehlungen vertrauen. Ist dieses Misstrauen gerechtfertigt? Hier offensichtlich ja! Das Schließen von einer **Stichprobe** (einige Nutzer) auf die **Population** (alle Nutzer) birgt Gefahren. Muss man sich damit abfinden? Nein! Diese Gefahren bestehen zwar für typische **Usability-Tests** (s. Kasten 1) können aber durch andere Methoden abgefangen werden[1].

2. Sich gegen den Zufall absichern: Inferenzstatistik

Glücklicherweise sind Usability-Engineers nicht die ersten, die allgemeingültige Aussagen auf Grund begrenzter Stichproben treffen müssen. Fast jeder, der empirisch arbeitet, steht vor diesem Problem. Die angestrebte Gültigkeit kann z. B. alle Wirbeltiere umfassen oder sich auf die Nutzer einer Website für professionelle Unterwasser-Fotografen beschränken. Leider kann – selbst in den sehr speziellen Fällen – selten jedes einzelne Individuum der Ziel-Population an dem Test teilnehmen.

Nähern wir uns den sich daraus ergebenden Problemen mit einer allgemeingültigen Aussage über eine mittelgroße Population von aktuell in etwa sieben Milliarden

Kasten 1: Abgrenzung des Usability-Experiments zu anderen Verfahren

Bei typischen **Usability-Tests** interagieren repräsentative Nutzer unter der Anleitung eines Usability-Engineers mit einem Interface. Sie bekommen bestimmte Aufgaben und werden bei deren Bearbeitung beobachtet und aufgefordert zu verbalisieren, was sie tun. Häufig werden ihre Interaktionen mit dem Interface und Äußerungen aufgezeichnet (teilweise ergänzt durch z. B. Mimik oder Augenbewegungen). Stellen, an denen die Interaktion nicht reibungslos funktioniert, werden als Usability-Probleme identifiziert (s. Shneiderman & Plaisant, 2005, S. 144-147).

Typische **A/B-Tests,** für die unter anderem Amazon und Google bekannt sind, vergleichen das Verhalten von Nutzern, die mit unterschiedlichen Designs konfrontiert werden. Um den Unterschied in der Usability zweier Checkboxen zu untersuchen, wird jeder Nutzer zufällig Gruppe A oder Gruppe B zugeordnet. Nutzer der Gruppe A bekommen Checkbox A und Nutzer der Gruppe B bekommen Checkbox B. Gemessen wird, bei welcher der Checkboxen mehr Nutzer ein gewünschtes Verhalten zeigen, z. B. den Zustand der Checkbox verändern, um die AGBs zu akzeptieren (s. Tullis & Albert, 2008, S. 216-217).

Ein weiteres, dem Usability-Experiment verwandtes, Verfahren ist die **experimentelle Evaluation.** Auch hier kommen experimentelle und inferenzstatistische Techniken zum Einsatz. Bei experimentellen Evaluationen wird aber meistens das Verhalten von Programmen oder Maschinen untersucht. Beim Usability-Experiment hingegen soll das Verhalten von Nutzern Aufschluss über die Usability eines Interfaces geben. Die Messung menschlichen Verhaltens birgt besondere Schwierigkeiten, die andere Methoden erfordern als die Messung maschinellen Verhaltens.

Individuen (die Menschheit): Männer sind größer als Frauen. Natürlich sind die Damen der Basketball-Nationalmannschaft größer als die Herren Bodenturner. Wenn Sie allerdings die Körpergröße der nächsten Frau und des nächsten Mannes, die/der Ihnen über den Weg läuft, messen, werden Sie mit einiger Wahrscheinlichkeit feststellen, dass der Mann größer ist als die Frau. Der Vorteil einer solchen **Zufallsziehung** ist, dass das Ergebnis nicht durch die Wahl der Stichprobe verfälscht wird. Mit einiger Wahrscheinlichkeit haben Sie allerdings zufälligerweise eine relativ große Frau oder einen relativ kleinen Mann getroffen. Messen Sie die Körpergröße von noch 5 Frauen und 5 Männern und mitteln Sie diese jeweils sechs Werte pro Gruppe! Sie sind jetzt bei der Stichprobengröße eines typischen Usability-Tests. Dass der Mittelwert der Männer größer ist als der der Frauen ist schon wesentlich wahrscheinlicher. Das ist der Vorteil einer **Mittelung.** Was hat es zu bedeuten, wenn das in Ihrer Stichprobe nicht der Fall ist? Haben Sie bewiesen, dass die Annahme, Männer wären größer als Frauen, ein dummes Vorurteil ist? Natürlich nicht. Es bleibt eine Restwahrscheinlichkeit, dass Sie ausgerechnet besonders große Frauen oder besonders kleine Männer getroffen haben.

Eine gewisse Irrtumswahrscheinlichkeit ist unvermeidbar, sie kann aber zumindest berechnet und den Anforderungen entsprechend verringert werden. Für jede derartige Fragestellung gibt es Tests zur Bestimmung der Wahrscheinlichkeit, dass ein in der Stichprobe beobachteter Unterschied auch in der Population vorhanden ist. Im aktuellen Beispiel ist dies ein *t*-Test für unabhängige Stichproben (s. Bortz & Schuster, 2010, S. 120; Tullis & Albert, 2008, S. 28-29). Üblicherweise akzeptiert man eine Irrtumswahrscheinlichkeit von etwa fünf Prozent (p (probability) < 0,05). Es geht um die Wahrscheinlichkeit, den beobachteten Unterschied zu beobachten, obwohl sich die beiden Gruppen eigentlich **nicht** unterscheiden. Wenn diese Irrtumswahrscheinlichkeit unter fünf Prozent liegt, spricht man davon, dass die beiden Gruppen sich **signifikant** unterscheiden. Dieses Vorgehen ermöglicht es, von einer Stichprobe (mit einer gewissen Wahrscheinlichkeit) auf die Population zu schließen, also allgemeingültige Aussagen zu treffen. Einige Variablen nehmen Einfluss auf diese Wahrscheinlichkeit: Je größer die Stichprobe, je genauer die Messung und je größer der tatsächliche Unterschied (der **Effekt**) ist, desto kleiner wird die Wahrscheinlichkeit daneben zu liegen. Die ersten beiden Variablen können Sie (bedingt) beeinflussen. Auf den Effekt haben Sie keinen Einfluss. Wenn der Effekt, wie beim Körpergrößenunterschied zwischen Männern und Frauen, extrem groß ist, haben Sie leichtes Spiel: Sie brauchen weder eine besonders große Stichprobe, noch eine besonders genaue Messung. Solche klaren Effekte sind aber leider eher spärlich gesät. Um Ergebnisse interferenzstatistisch abzusichern, brauchen Usability-Engineers also große Stichproben oder sehr genaue Messungen.

3. Große Stichproben: Typische A/B-Tests

A/B-Tests (s. Kasten 1) nehmen den Weg über die Stichprobengröße: Angenommen, bei insgesamt 50 Nutzern pro Gruppe benutzen 45 Nutzer der Gruppe A und 40 Nutzer der Gruppe B ihre jeweilige Checkbox richtig. Man könnte hier stehen bleiben und feststellen, dass Checkbox A besser funktioniert; immerhin war die Erfolgsquote um 10% höher als für Checkbox B. Diese naive Art der Dateninterpretation, wie man sie auch aus den Medien gewohnt ist (z. B. vom ARD-Deutschlandtrend oder vom Politbarometer im ZDF), verschließt ganz einfach die Augen vor dem Problem, dass man nicht ohne weiteres allgemeingültige Aussagen auf Grund einer Stichprobe machen kann. Diese Naivität führt zu einem berechtigten Misstrauen gegenüber Statistiken. Vereinfacht und plakativ ausgedrückt: Nur inferenzstatistische Laien behaupten, dass 40 weniger sei als 45, dass 3% weniger sei als 5% oder dass 2,1 Punkte weniger seien als 2,3 Punkte…; und sie schaden damit dem Ansehen und Einfluss seriöser Empiriker. Es kann nämlich auch sein, dass diese Ergebnisse rein zufällig sind und die Stichprobenunterschiede keine Populationsunterschiede widerspiegeln. Wären andere Nutzer getestet worden, hätte vielleicht Checkbox B die Nase vorn.

Um dem Rechnung zu tragen, muss man die Irrtumswahrscheinlichkeit kennen. Dazu berechnet man zunächst eine Teststatistik und vergleicht diese mit der entsprechenden Verteilung. Die in diesem Fall

Gruppe	Richtig	Falsch	Summe	χ^2	p
		Peters Stichprobe			
A	a = ***2***	b = ***3***	5		
B	c = ***4***	d = ***1***	5	1,67	0,197
Summe	6	4	*N* = ***10***		
		Mittelgroße Stichprobe			
A	a = ***45***	b = ***5***	50		
B	c = ***40***	d = ***10***	50	1,96	0,161
Summe	85	15	*N* = ***100***		
		Große Stichprobe			
A	a = ***90***	b = ***10***	100		
B	c = ***80***	d = ***20***	100	3,92	0,048
Summe	170	30	*N* = ***200***		

Anm. In den jeweils inneren Zellen ist angegeben, wie viele Nutzer einer Gruppe mit der jeweiligen Checkbox richtig (a, c) bzw. falsch (b, d) interagiert haben. Die zur Berechnung von χ^2 relevanten Werte (a, b, c, d und ***N***) sind kursiv und fett gedruckt.

Tab. 1.
A/B-Test von Checkbox A gegen Checkbox B mit unterschiedlichen Stichprobengrößen

angemessene Teststatistik heißt χ^2 (chi-Quadrat, vgl. Bortz & Schuster, 2010, S. 138; Tullis & Albert, 2008, S. 33-35). Tabelle 1 zeigt die Ergebnisse aus drei fiktiven A/B-Tests der Checkboxen aus Abbildung 1 in unterschiedlich großen Stichproben. **[Tab. 1]**

Für die mittelgroße Stichprobe ergibt sich:

$$\chi^2 = \frac{N * (ad - bc)^2}{((a + b) * (c + d) * (a + c) * (b + d)} =$$

$$\frac{100 * (45 * 10 - 5 * 40)^2}{(50 * 50 * 85 * 15)} = 1{,}96$$

Ein Vergleich mit der χ^2-Verteilung liefert $p = 0{,}161$, also eine Irrtumswahrscheinlichkeit von 16%; diese ist zu hoch, um aus den Daten eine zuverlässige Aussage darüber abzuleiten, welche Checkbox besser ist. Angenommen, der hier beobachtete Trend würde sich genauso fortsetzen (eine eher gewagte Annahme, die ich hier nur zur Illustration treffe!), bräuchte man eine in etwa doppelt so große Stichprobe, um eine statistisch abgesicherte Aussage machen zu können.

Für diese große Stichprobe (s. Tabelle 1) ist $p < 0{,}05$. Man kann also auf Grund der großen Stichprobe mit einer vertretbaren Irrtumswahrscheinlichkeit von unter 5% sagen, dass Checkbox A besser funktioniert als Checkbox B.

Zum Vergleich enthält Tabelle 1 noch das oben erwähnte „deutliche" Ergebnis Ihres Kollegen Peter Meier, der schon der Datenmanipulation verdächtigt wurde. Obwohl mehr als die Hälfte von Peters Teilnehmern Checkbox A falsch und fast alle Checkbox B richtig benutzt haben, nimmt er eine 19,7% Irrtumswahrscheinlichkeit in Kauf, wenn er dieses Ergebnis zugunsten von Checkbox B interpretiert. Weder Sie noch Peter haben Ihre Statistiken gefälscht und kommen doch zu unterschiedlichen Ergebnissen; Sie beide haben einfach unterschiedliche Stichproben aus derselben Population gezogen und die Irrtumswahrscheinlichkeit ignoriert.

4. Genaue Messungen: Das Usability-Experiment

Hier tritt ein praktisches Problem zu Tage: Je kleiner ein Usability-Unterschied (ein Effekt) ist, desto mehr Nutzer müssen getestet werden, um ein Ergebnis inferenzstatistisch abzusichern. Da man in der (guten) Praxis aber schlechte Konzepte möglichst von vornherein aussortiert und nur vielversprechende Bedienkonzepte gegeneinander testet, sind potentiell vorhandene Usability-Unterschiede üblicherweise eher klein. Um mit typischen A/B-Tests Ergebnisse inferenzstatistisch abzusichern, muss man also Zugang zu einer enorm großen Anzahl an Nutzern haben. Das bedeutet in der Praxis, dass diese Art der Testung nur über das Web möglich ist. Interfaces, für die das Web nicht in Frage kommt – z. B. aus Gründen der Geheimhaltung oder weil besondere Hardware benötigt wird –, wären praktisch von den Segnungen der Inferenzstatistik ausgeschlossen. Es kann außerdem häufig nur eine sehr geringe Anzahl repräsentativer Nutzer rekrutiert werden. Dies ist problematisch für Tests von Designs, bei denen Eigenschaften der Zielpopulation (z. B. Domänenwissen) eine Rolle spielen.

Glücklicherweise bleibt aber die oben erwähnte zweite Stellschraube der Messgenauigkeit. Wenn die Messgenauigkeit nur hoch genug ist, lassen sich auch in kleinen Stichproben Ergebnisse inferenzstatistisch absichern. Eine Möglichkeit, das Verhalten eines Menschen genau zu messen, ist, ihn dieses Verhalten so oft wie möglich wiederholen zu lassen. Der Mittelwert aus diesen wiederholten Beobachtungen hat ein gutes Signal-Rausch-Verhältnis, d. h. dieser Mittelwert ist eine genaue Messung. Hier kommt das Usability-Experiment ins Spiel. Durch die Verwendung von **Experimentaldesigns,** wie sie in der experimentellen Psychologie entwickelt wurden, ist es möglich solch genaue Daten zu erheben und somit auch kleine Effekte in kleinen Stichproben inferenzstatistisch abzusichern. Es handelt sich hierbei um eine Reihe von Techniken, denen eine über mehr als 100 Jahre gewachsene Menge an theoretisch-methodischen Überlegungen und praktischer Erfahrungen zu Grunde liegen. Da eine angemessene Darstellung den Umfang dieses Artikels sprengen würde, beschränke ich mich im Folgenden auf ein Beispiel aus meiner Arbeit bei Centigrade.

5. Ein Usability-Experiment: Vergleich zweier Navigationsprototypen

Für die Hauptnavigation der neuen Software-Generation eines Kunden hatten meine Kollegen und ich zwei vielversprechende Prototypen entwickelt, die in Abbildung 2 dargestellt sind. Beim ersten Prototyp, dem **Swipe-Navigator,** erfolgte die Navigation entweder über eine Wischgeste **(Swipe)** mit zwei Fingern oder über

eine Navigationslandkarte **(Navi-Map),** die durch Berührung des Bildschirms mit drei Fingern geöffnet wurde und in der der gewünschte Screen dann mittels eines Taps auf das entsprechende Icon ausgewählt wurde. Beim zweiten Prototyp, dem **Two-Way-Slider,** führte das Auflegen von zwei Fingern zum Öffnen einer Art Pie-Menü, in dem entweder durch einen **Slide** oder einen **Tap** auf das entsprechende Icon zum gewünschten Screen navigiert wurde. Die Frage war, ob der Swipe-Navigator oder der Two-Way-Slider effizienter ist, d. h. mit welchem der beiden Prototypen die Nutzer schneller zum Ziel gelangen würden. **[Abb. 2]**

Abb. 2.
Die beiden Navigationsprototypen mit ihren jeweils zwei Methoden. Bei Navi-Map und Tap wird der Ziel-Screen durch einen Tap ausgewählt. Die Pfeile beim ersten Schritt von Swipe und Slide zeigen die Bewegungsrichtung der jeweiligen Geste an und waren während des Experiments natürlich nicht zu sehen.

5.1. Methoden

5.1.1. Teilnehmer

Da die untersuchte Navigation kein Domänenwissen erfordert, sondern eher nur relativ grundlegende kognitive und motorische Prozesse eine Rolle spielen, war es nicht notwendig, repräsentative Teilnehmer mit Domänenexpertise zu rekrutieren. Unsere Stichprobe bestand daher aus pragmatischen Gründen aus 14 Angehörigen der Universität des Saarlandes (9 Frauen, Median Alter = 23,5 Jahre), die jeweils 8 EUR für ihre Teilnahme erhielten. Alle waren Rechtshänder mit normaler oder einer auf Normalniveau korrigierten Sehstärke und ohne Farbenblindheit und gaben ihre Einwilligung zur Teilnahme nach erfolgter Aufklärung.

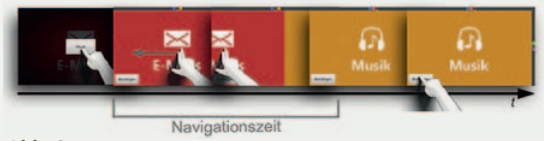

Abb. 3.
Ein Durchgang (Navigationsweg) mit der Methode „Swipe" des Prototypen „Swipe-Navigator". Teilnehmer bestätigten die Aufgabe (Navigation zum Screen „Musik") durch einen Tap auf den Button in der Mitte des Bildschirms, navigierten und bestätigten dann das Erreichen des Ziel-Screens durch einen Tap auf den Button „Bestätigen". Die interessierende Navigationszeit ist die Zeit von der Bestätigung der Aufgabe bis zum Erreichen des Zielscreens.

5.1.2. Studiendesign

Die Prototypen und das Experimentalprogramm wurden mit Expression Blend 4 (Microsoft Inc.) erstellt und von den Teilnehmern über einen Touch-Monitor (M2256PW, 3M Touch Systems Inc.) bearbeitet. Der Touch-Monitor war so angebracht, dass er bequem aus dem Stand erreicht werden konnte. Teilnehmer bearbeiteten insgesamt 12 Blöcke à 20 Durchgänge, jeweils 6 Blöcke mit einem Navigationsprotypen. Innerhalb eines Blockes navigierten sie von jedem der fünf Screens zu jedem anderen Screen, d. h. sie bewältigten jeden der möglichen 20 Navigationswege. Alle Teilnehmer bearbeiteten die Navigationswege in der gleichen pseudorandomisierten Reihenfolge. Die Pseudorandomisierung unterlag der Einschränkung, dass der Start-Screen eines Durchgangs immer der Ziel-Screen des vorhergehenden Durchgangs war. Die Reihenfolge der Navigationsprototypen wurde über die Teilnehmer hinweg balanciert: Die Hälfte der Teilnehmer begann mit dem Swipe-Navigator und die andere Hälfte begann mit dem Two-Way-Slider. Direkt nach der Bearbeitung eines jeden Prototyps füllten die Teilnehmer einen Fragebogen zur User-Experience aus. Am Ende des Experiments machten sie einige Angaben zu ihrer Person (Alter, Geschlecht, Touchscreen-Erfahrung etc.). Eine Sitzung dauerte inklusive Vor- und Nachgespräch, Bearbeitung beider Prototypen und der Fragebögen in etwa eine Stunde.

5.1.3. Material

Screens enthielten nur die zur Navigation notwendigen Informationen. Die Bezeichnungen der Screens waren so gewählt, dass sie jedem Teilnehmer geläufig waren. Das Programm konnte als eine Art Medienverwaltung interpretiert werden (vgl. Abb. 2).

5.1.4. Prozedur

Zu Beginn eines Durchgangs erschien in der Mitte des Bildschirms der Name des Ziel-Screens. Teilnehmer bestätigten diese Anweisung (indem sie darauf tappten) und navigierten dann so schnell wie möglich zu dem in der Anweisung gezeigten Ziel-Screen. Die Teilnehmer zeigten durch einen Tap auf einen Button unten links im Bildschirm an, dass die Navigation beendet ist. Die im Folgenden analysierte **Navigationszeit** ist die Zeit der Bestätigung der Anweisung bis zum Erreichen des Ziel-Screens. **[Abb. 3]**

Teilnehmer	Swipe-Navigator	Two-Way-Slider	d_i	$(d_i - m_d)^2$
1	3,58	1,45	2,13	2,04
2	2,03	2,49	-0,46	1,35
3	3,09	1,48	1,61	0,83
4	1,95	1,61	0,35	0,12
5	3,35	2,88	0,47	0,05
6	2,14	1,45	0,70	0,00
7	1,76	0,67	1,09	0,15
8	1,59	0,93	0,66	0,00
9	2,06	0,68	1,38	0,46
10	1,98	2,48	-0,50	1,44
11	2,02	0,83	1,19	0,24
12	1,69	1,33	0,36	0,12
13	1,38	0,79	0,59	0,01
14	1,47	1,20	0,27	0,18
Mittelwert	2,15	1,45	***0,70***	0,50
Summe	30,11	20,27	9,83	***7,00***

Tab. 2.
Mittlere Navigationszeiten (in s pro Navigationsweg) pro Teilnehmer und Prototyp. Zur einfacheren Berechnung der Teststatistik t ist zusätzlich pro Teilnehmer i die Differenz zwischen den beiden Prototypen (d_i) sowie das Quadrat der Abweichung dieser Differenz von der mittleren Differenz $(d_i - m_d)^2$ angegeben.

Anm.
Die beiden für die Berechnung von t relevanten Werte (m_d und $\sum_{i=1}^{n}(d_i\text{-}m_d)^2$)sind kursiv und fett gedruckt.

Aus Tabelle 2 ergibt sich:

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - m_d)^2}{n-1}} = \sqrt{\frac{7{,}00s}{14-1}} = 0{,}73s \text{ und } t(n-1) = \sqrt{n}\frac{m_d}{s_d} = \sqrt{14} * \frac{0{,}70s}{0{,}73s} = 3{,}59$$

mit n = Stichprobengröße, d_i = Differenz in der Navigationszeit zwischen den beiden Prototypen für Teilnehmer i, m_d = Mittelwert der Differenzen, s_d = Standardabweichung der Differenzen

5.2. Ergebnisse und Diskussion

5.2.1. Inferenzstatistische Absicherung des Navigationszeit-Unterschieds

Analysiert wurden Navigationszeiten bei erfolgreicher Navigation. Wie zu erwarten, gab es kaum Navigationsfehler (98% korrekt für beide Navigationsprototypen). Es standen also für jeden der in Tabelle 2 aufgelisteten Mittelwerte (pro Teilnehmer und Prototyp) durchschnittlich 20*6*0,98 ≈ 118 Messwerte zur Verfügung. Die inferenzstatistischen Absicherung des Geschwindigkeitsvorteils des Two-Way-Sliders (1,45s/Weg) gegenüber dem Swipe-Navigator (2,15s/Weg) erlaubt ein t-Test für abhängige Stichproben (vgl. Bortz & Schuster, 2010, S. 125; Tullis & Albert, 2008, S. 29-30): **[Tab. 2]**

Ein Blick auf die t-Verteilung zeigt, dass dieses $t(13) = 3{,}59$ einem $p = 0{,}003$ entspricht. Das bedeutet, mit einer sehr geringen Irrtumswahrscheinlichkeit von 0,3% ist die Navigation mit dem Swipe-Navigator tatsächlich langsamer als die Navigation mit dem Two-Way-Slider.

5.2.2. Bestimmung des Gewinns

Um anzudeuten, was mit derart erhobenen Daten außerdem möglich ist, berechne ich im Folgenden noch exemplarisch den Gewinn, der sich daraus ergeben würde den Two-Way-Slider anstatt des Swipe-Navigators zu verwenden.

Genauso falsch und irreführend wie einen Mittelwerts-Unterschied in einer Stichprobe mit einem Unterschied in der Population gleichzusetzen, wäre es, auf Basis eines Mittelwert-Unterschieds den zu erwartenden Gewinn zu bestimmen. Man kann allerdings aus den Daten einer Stichprobe berechnen, in welchem Bereich der Gewinn mit einer gewissen Wahrscheinlichkeit liegt.

Dieses **Konfidenzintervall** ist hier (vgl. Bortz & Schuster, 2010, S. 119):

$$[\text{untere Grenze; obere Grenze}] = \left[m_d - t(n-1)_{1-\alpha/2} * \frac{s_d}{\sqrt{n}}; m_d + t(n-1)_{1-\alpha/2} * \frac{s_d}{\sqrt{n}}\right] =$$
$$\left[0{,}70s - 2{,}16 * \frac{0{,}73s}{\sqrt{14}}; 0{,}70s + 2{,}16 * \frac{0{,}73s}{\sqrt{14}}\right] = [0{,}28s;\ 1{,}12s]$$

mit α = gewünschte Irrtumswahrscheinlichkeit = 0,05

Unter der Annahme, dass alle Navigationswege gleich häufig sind, werden bei Verwendung des Two-Way-Sliders anstatt des Swipe-Navigators pro Navigationsweg zwischen 0,28s und 1,12s eingespart. Dies macht bei angenommenen 50 Navigationswegen pro Arbeitsstunde in etwa zwischen 13,90s und 56,10s aus. Bei einer täglichen Bedienung des Interfaces von 6h und 220 Arbeitstagen pro Jahr werden dann, bei einem Stundenlohn von 20 EUR und 100 Mitarbeitern, in etwa zwischen 10.200 EUR und 41.100 EUR pro Jahr eingespart. Dies gilt mit einer vertretbaren Irrtumswahrscheinlichkeit von 5%. Wer auf Grund der mittleren Differenz von 0,7s pro Navigationsschritt behauptet, es würden 25.700 EUR pro Jahr eingespart, ist im besten Fall naiv. Zugegebenermaßen ist das Konfidenzintervall in dem aktuellen Beispiel relativ breit, also die Schätzung relativ ungenau. Für eine nahezu beliebig genaue Schätzung muss einfach die Teilnehmerzahl oder die Menge an Messwerten pro Teilnehmer erhöht werden.[2]

5.2.3. Weitere Ergebnisse

Die erhobenen Daten erlauben außerdem eine Reihe weiterer interessanter Analysen. Diese führten unter anderem zu folgenden Aussagen:

- Wenn man nur die häufigsten Navigationswege (zwischen Texte und E-Mails, s. Abb. 2) betrachtet, ist der Swipe-Navigator genauso schnell wie der Two-Way-Slider (beide 1,34s; $p = 0{,}997$).
- Teilnehmer werden (auch noch nachdem sie sich an das Experiment und die Architektur des Interfaces gewöhnt haben – also bei der Bearbeitung des zweiten Prototypen) mit etwas Übung schneller (2,13s in Block 1 gegen 1,28s in Block 6; $p = 0{,}001$). Sie erreichen aber nach einigen Blöcken ein stabiles Niveau (1,42s in Block 5 gegen 1,28s in Block 6; $p = 0{,}30$).
- Die allgemeine User-Experience unterscheidet sich laut den eingesetzten Fragebögen nicht zwischen den Prototypen ($p = 0{,}21$). Die beiden Navigationsmethoden des Swipe-Navigators werden allerdings als besser integriert empfunden ($p = 0{,}008$).
- In Einklang damit wurden beim Swipe-Navigator beide Methoden gleich häufig benutzt (Swipe: 52% vs. Navi-Map, 47%, $p = 0{,}78$), während beim Two-Way-Slider hauptsächlich getappt wurde (Tap: 62% vs. Slide: 6%, $p < 0{,}001$). Entsprechend der Design-Intention wurde nämlich beim Swipe-Navigator das Swipe für kurze Wege (ein Swipe; $p = 0{,}06$) und die Navi-Map für weite Wege (wenn mehr als drei Swipes notwendig gewesen wären; $p < 0{,}001$) bevorzugt.

6. Schlussfolgerungen und Ausblick

Wie unser Navigationsexperiment zeigt, ist es mit Hilfe des Usability-Experiments möglich, Ergebnisse auch in kleinen Stichproben inferenzstatistisch abzusichern. Inferenzstatistik ist notwendig, da es beträchtliche Gefahren birgt, aufgrund einer Stichprobe (einige Nutzer) allgemeingültige Aussagen (über alle Nutzer) zu treffen. Auf lange Sicht erzeugt das Ignorieren der Irrtumswahrscheinlichkeit berechtigtes Misstrauen und schadet damit dem Ansehen von Usability-Engineers.

Natürlich soll das Usability-Experiment kein Ersatz für Usability-Tests sein. Usability-Tests sind äußerst effizient beim Aufdecken grober Usability-Probleme, wohingegen sich das Usability-Experiment auf die Beantwortung einiger weniger konkreter Fragen beschränken muss. Diese Methoden sind also eher ergänzend als konkurrierend gedacht. Es bietet sich zum Beispiel bei wichtigen Entscheidungen an, Ergebnisse eines Usability-Tests in einem Usability-Experiment abzusichern. Welches der beiden Verfahren besser geeignet ist hängt auch von der Fragestellung ab. Um zu untersuchen, wie intuitiv ein Interface bedienbar ist, sind Usability-Experimente mit ihren vielen Wiederholungen eher ungeeignet. Allerdings werden viele Interfaces über einen längeren Zeitraum und sehr intensiv benutzt; Nutzer erlernen also den Umgang damit. Der Lernfortschritt über die vielen Wiederholungen lässt sich in Usability-Experimenten gut erfassen wird aber in Usability-Tests normalerweise nicht abgebildet. Fragebögen oder Befragungen zur User-Experience lassen sich ergänzend zu beiden Methoden einsetzen. Hier kommt es sicherlich auf die Zielsetzung des Produktes an, ob die Experience nach der ersten Nutzung oder die Experience nach der hundertsten Nutzung das interessantere Maß ist.

Literatur

1. Bortz, J. & Schuster, C. (2010). Statistik für Human- und Sozialwissenschaftler (7. Auflage). Berlin, Deutschland: Springer.
2. Tullis, T. & Albert, B. (2008). Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Burlington, MA: Morgan Kaufmann.
3. Shneiderman, B. & Plaisant, C. (2005). Designing the User Interface: Strategies for Effective Human-Computer Interaction (4. Ausg.). Boston, MA u. a.: Pearson Education.

[1] Diese Verfahren sollen Usability-Tests ergänzen und nicht ersetzen (s. hierzu auch die Diskussion im letzten Absatz dieses Artikels).

[2] Die Annahmen zur (relativen) Häufigkeit der Navigationswege und Nutzungsdauer des Interfaces sind relativ willkürlich und sollten durch empirische Daten ersetzt werden. In der Praxis verlängern sicherlich noch zusätzliche Einflussfaktoren die Navigationszeiten. Wenn diese Faktoren nicht unverhältnismäßig stark auf die effizientere Variante wirken (dies wäre eine Interaktion, s. z. B. Bortz & Schuster, 2010), bleibt der Nettovorteil jedoch bestehen. Die nicht-perfekte Präzision jeder empirischen Messung wird bei der Berechnung des Konfidenzintervalls automatisch berücksichtigt.