DIPLOMARBEIT

Wie man mit Statistik schlecht informiert:

Praktiken der Ergebnisdarstellung bei Experimenten im Software Engineering

Autor:
Marcus Michalsky
(Matrikelnummer.: 3786045)

Betreuer / Gutachter:
Prof. Dr. Lutz Prechelt

Zweitgutachter:
Prof. Dr. Volker Roth

Freie Universität Berlin

16. Dezember 2009

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe. Diese Arbeit wurde noch keiner Prüfungsbehörde in gleicher oder ähnlicher Form vorgelegt.

Berlin, den 15. Dezember 2009
Marcus Michalsky

Zusammenfassung

Im Software Engineering geht es um die Entwicklung von kostengünstiger, performanter und sicherer Software. Der Bereich des Empirical Software Engineerings vergleicht Technologien und Werkzeuge und untersucht dessen Effekte auf die Softwareentwicklung. Eine empirische Methode ist die des kontrollierten Experiments.
In dieser Arbeit wurden die Praktiken der Ergebnisdarstellung für eine Menge von kontrollierten Experimenten aus Artiklen zweier Software Engineering Zeitschriften und einer Konferenz ermittelt. Dabei wurden drei Aspekte berücksichtigt. Es wurden (1) die Verwendung beschreibender Statistiken und grafischer Darstellungen, (2) Art und Umfang der Darstellung statistischer Auswertungsverfahren sowie (3) Präsenz und Zustand von Rohdaten untersucht.
Insgesamt werden mit 7% der untersuchten Artikel wenig kontrollierte Experimente berichtet. Beschreibende Statistiken sind auf ein Minimum beschränkt (Mittelwerte und Standardabweichungen werden zu ca. 50% berichtet). Grafiken zur Beschreibung der aus den Experimenten gewonnenen Daten sind in ein von zwei Fällen vorhanden. Die Ergebnisse statistischer Auswertungsverfahren sind von gerigem Detail. Oft wird lediglich berichtet, ob der Test ein signifikantes Ergebnis liefert. Ansonsten werden p-Werte und teilweise Signifikanzniveaus berichtet. Effektgrößen und Konfidenzintervalle sind häufig nicht vorhanden. Daher kann auch nur schwer der praktische Nutzen diskutiert werden. Die gemessenen Daten der Experimente werden nur selten berichtet. Anfragen waren zur Hälfte erfolgreich. Die Qualität der gesendeten Rohdaten ist sehr unterschiedlich.
Anschauungsbeispiele sollen zeigen, wie bestimmte Aspekte zum Verständnis beitragen, bzw. die Schlußfolgerungen beeinflussen können.
Die durchwachsenen Ergebnisse im noch jungen Bereich des Software Engineerings zeigen, dass weiterhin Bedarf an Richtlinien und Empfehlungen besteht. Außerdem sollten regelmässige Analysen (wie diese Arbeit) den derzeitigen Stand analysieren und aktuelle Defizite berichten, um diese in der Zukunft zu vermeiden.

Inhaltsverzeichnis

KAPITEL 1 Einleitung **1**
1.1. Ziele der Arbeit 2
1.2. Gliederung 2

KAPITEL 2 Grundlagen **3**
2.1. Experimente 3
2.2. Statistik 6
2.2.1. beschreibende Statistik 6
2.2.2. Signifikanztests (allgemein) 12
2.2.3. Signifikanztests (konkret) 15
2.2.4. Kritik an Hypothesentests 20

KAPITEL 3 Verwandte Arbeiten **23**

KAPITEL 4 Methodik **27**
4.1. Identifizierungsprozedur 27
4.1.1. Suchstrategie 27
4.1.2. Literaturverwaltung 30
4.1.3. Fortschrittserfassung 30
4.2. Untersuchte Merkmale 30
4.2.1. Beschreibende Statistiken und grafische Darstellungen . . . 31
4.2.2. Ergebnisdarstellung benutzter statistischer Methoden 31
4.2.3. Präsenz und Zustand von Rohdaten 32

KAPITEL 5 Ergebnisse **33**
5.1. Identifizierte Artikel 33
5.2. Untersuchte Merkmale 35
5.2.1. Beschreibende Statistiken und grafische Darstellungen . . . 35

5.2.2. Ergebnisdarstellung benutzter statistischer Methoden 37
5.2.3. Präsenz und Zustand von Rohdaten 41

KAPITEL6 Diskussion **45**
6.1. Identifizierte Artikel 45
6.2. Untersuchte Merkmale 46
6.2.1. Beschreibende Statistiken und grafische Darstellungen . . . 46
6.2.2. Ergebnisdarstellung benutzter statistischer Methoden 47
6.2.3. Präsenz und Zustand von Rohdaten 50

KAPITEL7 Anschauungsbeispiele **53**
7.1. Artikel: EMSE09-57 53
7.2. Artikel: EMSE06-303 57
7.3. Artikel: EMSE08-693 59

KAPITEL8 Schlussfolgerung und Ausblick **63**

Abbildungsverzeichnis **65**

Tabellenverzeichnis **67**

Identifizierte Artikel von 2006–2009 **68**

Literaturverzeichnis **73**

Anhang **81**
.1. Tabellarische Auflistung der Variablen 81

Kapitel 1

Einleitung

Im Software Engineering geht es um Techniken zur Entwicklung von Software, die je nach Domäne wenig kosten soll, schnell sein muss, oder andere Anforderungen wie z.B. Sicherheitaspekte erfüllen soll. (vgl. Zelkowitz and Wallace, 1997)
Der Schwerpunkt im Empirical Software Engineering liegt in den Fragen welche Technologien unter welchen Bedingungen von wem angewendet werden sollen. Von Interesse ist der Vergleich verschiedener Technologien und dessen Effekte auf die Entwicklung von Software. (vgl. Sjøberg et al., 2005; Hannay et al., 2007)
Die Antwort auf diese Fragen soll mit Hilfe von Informationen, die aus Beobachtungen resultieren, gefunden werden. Eine mögliche Methode ist dabei die des kontrollierten Experiments. (vgl. Zelkowitz and Wallace, 1997)
Damit die Antworten auch von wissenschaftlichem Wert sind, genügt es nicht festzustellen, eine Technik sei gut oder schlecht. Mit Hilfe von statistischen Signifikanztests können Hypothesen akzeptiert oder wiederlegt werden. Es wird entschieden, ob mehrere Stichproben aus der selben Grundgesamtheit stammen (oder anders ausgedrückt ob die verglichenen Techniken gleiche Resultate liefern). (vgl. Kruskal and Wallis, 1952; Dybå et al., 2006)
Die in der Medizin schon lange angewandte empirische Forschung ist laut Kitchenham et al. (2002) ungenügend. Daher sei es auch kaum verwunderlich, dass die Situation im noch jungen Bereich des Software Engineerings ähnlich aussehe. Fukuda and Ohashi (1997) und Wilkinson (1999) geben Richtlinien an wie die Verwendung statistischer Methoden in Medizin und Psychologie berichtet werden sollen. Kitchenham et al. (2002), Jedlitschka and Pfahl (2005) und Kitchenham et al. (2008) geben Richtlinien fürs Software Engineering, teilweise generell für empirische Studien, teilweise speziell für Experimente.

1.1 Ziele der Arbeit

In dieser Arbeit sollen die Praktiken der Ergebnisdarstellung von (kontrollierten) Experimenten im Software Engineering unter drei Gesichtspunkten analysiert werden:

1. Verwendung von beschreibenden Statistiken und grafischen Darstellungen
2. Art und Umfang der Ergebnisdarstellung statistischer Methoden
3. Präsenz und Zustand von Rohdaten

Der erste Aspekt soll untersuchen, ob und in welchem Detail die aus den Experimenten gewonnenen Daten numerisch und grafisch beschrieben werden. Zweitens soll ermittelt werden, ob und in welcher Form die Ergebnisse von statistischen Methoden zur Auswertung der Daten (Signifikanztests) berichtet werden. Außerdem soll drittens aufzeigen, ob Rohdaten verfügbar sind und dessen Zustand untersuchen.

1.2 Gliederung

Kapitel 2 stellt die für das Verständnis relevanten Grundlagen dar. Diese gliedern sich in die Beschreibung von Experimenten und der Darstellung und Auswertung der aus den Experimenten ermittelnden Daten mit Hilfe der Statistik. In Kapitel 3 werden verwandte Arbeiten, die ähnliche Aspekte an Experimenten im Software Engineering untersuchen vorgestellt. Die Methodik, die zur Identifizierung der Experimente und Erfassung der in 1.1 beschriebenen Merkmale verwendet wurde, wird in Kapitel 4 beschrieben. Die Ergebnisse der Untersuchung zeigt Kapitel 5. Kapitel 6 diskutiert die Ergebnisse. Anschauungsbeispiele, die die Wichtigkeit der Angabe bestimmter Merkmale verdeutlichen sollen, werden in Kapitel 7 gegeben. Kapitel 8 schließt mit einem Fazit und gibt einen Ausblick.

Kapitel 2

Grundlagen

Dieses Kapitel beschreibt die für die in Kapitel 4 beschriebene Untersuchung relevanten Grundlagen. Abschnitt 2.1 behandelt kontrollierte Experimente. Im Einzelnen werden ihre Definition und die Gültigkeit der Ergebnisse beschrieben. Abschnitt 2.2 stellt Größen und grafische Darstellungen der beschreibenden Statistik sowie Verfahren zur Auswertung der Daten der schließenden Statistik vor.

2.1 Experimente

Neben älteren Definitionen von Basili (1996) und Zelkowitz and Wallace (1997) geben Shadish et al. (2002, S. 12) folgende „moderne“ Definition von Experimenten:

> "*Experiment:* A study in which an intervention is deliberately introduced to observe its effects."
>
> "*Randomized Experiment:* An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers."
>
> "*Quasi-Experiment:* An experiment in which units are not assigned to conditions randomly."

Ein Experiment ist also eine Studie, in der Auswirkungen eines bewussten Eingriffs (verändern eines Zustands) beobachtet werden. Alle anderen Einflussfaktoren werden konstant gehalten. Im Bereich des Software Engineerings handelt es sich dabei z.B. um die Auswirkungen einer Methode oder eines Werkzeugs auf die Arbeitszeit oder Qualität der Ergebnisse. (vgl. Zelkowitz and Wallace, 1997)

Wenn verschiedene Methoden/Werkzeuge (*treatments*) verglichen werden, werden diese den Subjekten (*subjects*) zugewiesen und bilden somit Gruppen (*groups*). Subjekte können z.B. Personen oder Programme sein. Erfolgt diese Zuweisung per Zufall, z.B. durch den Wurf einer Münze oder generierte Zufallszahlen, so spricht man von einem randomisierten Experiment. (vgl. Shadish et al., 2002, S. 13)
Durch die randomisierte Zuweisung ist es wahrscheinlich, dass sich eventuelle Unterschiede in den Subjekten gegenseitig aufheben. Die beobachtete Wirkung, gemessen in den abhängigen Variablen (*dependent variables*), kann dadurch auf die Veränderung der unabhängigen Variablen (*independent variables*), also die verschiedenen Eingriffe, zurückgeführt werden. Die verschiedenen Eingriffe sind also die Ursache des beobachteten Effekts. (vgl. Basili, 1996; Shadish et al., 2002, S. 13)
Man spricht von einem Quasi-Experiment, wenn eine randomisierte Zuweisung nicht möglich ist. Dies ist z.B. der Fall, wenn verschiedene Abteilungen, Institute oder Dienststellen eines Unternehmens die Versuchsgruppen bilden. Diese wird man nicht mit angemessenem Aufwand mischen können. Werden neue mit alten Technolgien verglichen, wäre eine randomisierte Zuweisung aus ethischen Gründen nicht durchführbar. (vgl. Kampenes et al., 2009)
Bleibt die randomisierte Zuweisung aus hat dies zur Folge, dass sich Unterschiede nicht mehr gleichmäßig auf die Gruppen verteilen. Für die beobachteten Unterschiede in den Ausgabevariablen (abhängigen Variablen) könnten die Unterschiede in den Gruppen die Ursache sein und nicht unbedingt die Veränderung der Eingabevariablen (unabhängigen Variablen). (vgl. Kampenes et al., 2009)
Um solche eventuellen Unterschiede aufzudecken, werden z.B. Vortests durchgeführt, um einen Leistungswert der Versuchspersonen zu erhalten und damit die Unterschiede auszugleichen. Außerdem kann eine solche Messung in der statistischen Auswertung (Paaren) benutzt werden. (vgl. Kampenes et al., 2009; Shadish et al., 2002, S. 14)
Für die Gültigkeit (*validity*) der aus den Ergebnissen des Experiments getroffenen Schlüsse geben Shadish et al. (2002, S. 38) folgende vierstufige Typologie an:

> "*Statistical Conclusion Validity:* The validity of inferences about the correlation (covariation) between treatment and outcome."

> "*Internal Validity:* The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured."

> "*Construct Validity:* The validity of inferences about the higher order constructs that represent sampling particulars."

> "*External Validity:* The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables."

Die interne Gültigkeit (*internal validity*) bezieht sich darauf, ob die Veränderung der unabhängigen Variablen und die Beobachtungen der abhängigen Variablen unter den experimentellen Bedingungen einen Kausalzusammenhang bilden. Dies ist laut Shadish et al. (2002) der Fall, wenn die Veränderung der unabhängigen Variablen zeitlich vor den Beobachtungen geschieht, wenn veränderte und beobachtete Werte kovariieren und wenn keine anderen Erklärungen für den Zusammenhang plausibel sind. Der zeitliche Aspekt ist bei Experimenten automatisch erfüllt. Die Kovariation wird mit Hilfe von statistischen Tests ermittelt.
Handelt es sich um ein randomisiertes Experiment, so sind andere Ursachen unwahrscheinlich. Bei Quasi-Experimenten, bei denen die Randomisierung z.B. unangebracht, unpraktisch oder zu kostenintensiv ist, können systematische Unterschiede zwischen den Gruppen bestehen, die die Ergebnisse eventuell beeinflussen. Werden diese Ursachen im Entwurf und der Analyse erkannt und berücksichtigt, so kann dieser Effekt reduziert werden oder gar verschwinden. (vgl. Kampenes et al., 2009)
Die Gültigkeit der statistischen Schlüsse (*statistical conclusion validity*) behandelt den Aspekt der Kovariation genauer. Dabei geht es um die Fragen, ob Ursache und Wirkung zusammenhängen und wenn ja wie stark. Die getroffenen Schlüsse können z.B. auf Grund von zu geringer Trennschärfe (siehe Abschnitt 2.2.2) oder Verletzung der Voraussetzungen der verwendeten statistischen Tests falsch sein. (vgl. Shadish et al., 2002, S. 42-45)
Ob die Folgerungen auch unter anderen Bedingungen gelten, also z.B. mit anderen Versuchspersonen, in anderen Umgebungen oder gar anderen Variablen, beschreiben Shadish et al. (2002) mit der externen Gültigkeit (*external validity*). Ein Beispiel dafür ist die Verwendung von Studenten als Versuchspersonen. Damit ist nicht gewährleistet, dass für Fachleute in der Industrie gleiche Ergebnisse reproduzierbar sind.
Die Gültigkeit des Konstrukts (*construct validity*) behandelt Aspekte der Durchführung des Experiments. Diese kann eventuell nicht objektiv sein, wenn z.B. der Experimentator (Leiter des Experiments) zuviel Informationen ausgibt. So könnten eher gewünschte Ergebnisse bekannt werden und das Ergebnis beeinflussen. Ein anderes Beispiel ist die Nutzung von alten bzw. innovativen Werkzeugen als Instrumente. Diese könnten besonders negative bzw. positive Ergebnisse aufgrund ihrer Popularität hervorrufen. (vgl. Shadish et al., 2002, S. 73)

2.2 Statistik

In diesem Abschnitt werden die für die in Kapitel 4 beschriebene Untersuchung relevanten Grundlagen beschrieben. Abschnitt 2.2.1 behandelt Größen und grafische Darstellungen der beschreibenden Statistik. Die Abschnitte 2.2.2, 2.2.3 und 2.2.4 beschreiben die schließende Statistik gegliedert in Hypothesentests und dessen Kritik sowie konkrete statistische Auswertungsverfahren.

2.2.1 beschreibende Statistik

Die beschreibende Statistik dient dazu, einen Überblick über einen Datensatz zu bekommen. Sie gibt Größen an, die die Lokation und Spreizung der Daten beschreibt und verwendet Grafiken zur Visualisierung.

Größen

Der *Mittelwert* (arithmetisches Mittel) eines Datensatzes gibt den Durchschnittswert der Daten an. Seien $x_1, x_2, \ldots, x_n$ die n Datenwerte. Dann ist der Mittelwert definiert durch:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

Die durch Messungen unvermeidbaren Fehler können durch wiederholtes Messen und Bilden des Mittelwertes reduziert werden. (vgl. Meschede, 2006, S. 6)
Da der Mittelwert anfällig gegen Ausreißer (siehe Seite 11) ist, wird häufig der sogenannte *Median* verwendet. Der Median der aufsteigend nach Größe sortierten Zahlen $x_1, x_2, \ldots, x_n$ ($x_i < x_{i+1}\ \forall i \in \{1, \ldots, n-1\}$) ist definiert durch:

$$M = \begin{cases} \frac{n+1}{2} \text{ - tes Element} & \text{, falls } n \text{ ungerade} \\ \left\lfloor \frac{n+1}{2} \right\rfloor \text{ - tes und } \left\lceil \frac{n+1}{2} \right\rceil \text{ - tes Element} & \text{, falls } n \text{ gerade} \end{cases} \tag{2.2}$$

Besitzt der Datensatz also eine gerade Anzahl von Elementen, so existieren zwei Mediane. (vgl. Cormen et al., 2001, S. 183)
Der durchschnittliche quadratische Abstand der Datenwerte zum Mittelwert wird als *Varianz* (Streuung) bezeichnet und ist definiert durch:

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{(n-1)} \tag{2.3}$$

Die Quadratwurzel der Varianz heißt *Standardabweichung* σ.

In Experimenten existieren Mittelwert, Varianz und Standardabweichung sowohl für die Grundgesamtheit als auch für Stichproben. Um eine einheitliche Notation zu schaffen sei Tabelle 2.1 gegeben.

Bedeutung	Grundgesamtheit	Stichprobe
Mittelwert	μ	$\overline{x}$
Varianz	σ^2	s^2
Standardabweichung	σ	s

Tabelle 2.1: Formelzeichen statistischer Begriffe

Weitere Maßzahlen, um einen Datensatz zu beschreiben, sind die sogenannten Quartile bzw. Perzentile. Ein *Perzentil* ist derjenige Wert, der den Datensatz derart teilt, dass ein bestimmter Prozentsatz der Daten kleiner ist als er selbst. So ist z.B. das 50-Perzentil der Datenmenge $(1, 5, 3, 8, 2, 4, 7)$ der Wert 4, denn 50% der Datenwerte ist kleiner als 4, nämlich 1,2 und 3. (vgl. Ross, 2006, S. 23)
Die 25-,50- und 75-Perzentile sind besonders ausgezeichnet. Sie werden erstes, zweites und drittes bzw. unteres, zentrales und oberes *Quartil* genannt. Wie der Name schon sagt, teilen sie die Datenwerte in vier Viertel. Das 50-Perzentil bzw. zweites/zentrales Quartil ist der Median. (vgl. Ross, 2006, S. 24)
Zusätzlich zu Mittelwerten von Stichproben, die Punktschätzungen von Mittelwerten der Grundgesamtheit darstellen ($E[\overline{x}] = \mu$) (vgl. Ross, 2006, S. 185), existieren Konfidenzintervalle. Das $100(1 - \alpha)\%$-Konfidenzintervall mit einem Signifikanzniveau $\alpha = 0.05$ ist ein 95%-Konfidenzintervall, d.h. werden mehrmals Stichproben aus der Grundgesamtheit gewählt und Mittelwerte berechnet, würde der Mittelwert in 95% der Fälle in dem Intervall liegen.[1] (vgl. NIST, 2009a; Bower, 2009a; Deacon, 2009; Ross, 2006, S. 220)

Skalen

Variablen, die u.A. in kontrollierten Experimenten gemessen werden, teilt Stevens (1946) in vier verschiedene Klassen (siehe Tabelle 2.2, Seite 8).
Variablen auf nominalen (oder auch kategorischen) Skalen können Werte aus einer bestimmten Menge annehmen. Die Werte dieser Menge besitzen keine Ordnung (Hierarchie). Sie können lediglich auf Gleichheit geprüft werden. Zulässige Statistiken für Variablen auf nominalen Skalen sind Häufigkeiten sowie der am häufigsten vorhandene Wert (Modus).

[1] Konfidenzintervalle können auch für andere Parameter der Grundgesamtheit angegeben werden. Der Mittelwert ist ein Beispiel dafür.

Skala	zulässige Statistiken	Operationen
Nominal	Häufigkeit, Modus	$=$
Ordinal	", Median, Perzentile	$=, <, >$
Interval	", Mittelwert, Standardabweichung	$=, <, >, =_{\text{int}}$
Rational	alle Statistiken	$=, <, >, =_{\text{int}}, =_{\text{rat}}$

Tabelle 2.2: Variablenskalen

Ordinale Skalen besitzen alle Eigenschaften von Nominalskalen. Zusätzlich lassen sich die Werte ordnen, d.h. es ist für zwei beliebige Werte bestimmbar, ob der erste größer bzw. kleiner dem zweiten ist. Hier ist es außerdem zulässig (und durch die Ordnung möglich) einen Median zu bestimmen. Stevens (1946) stellt fest, dass es zwar Fälle gebe, in denen die Angabe von Perzentilen (und u.U. auch Mittelwerte und Standardabweichungen) für ordinale Skalen sinnvoll ist, jedoch könne man generell nicht davon ausgehen, dass die aufeinanderfolgenden Intervalle zwischen den Werten gleich groß sind. Daher sei Vorsicht geboten. Betrachtet man beispielsweise eine mehrstufige Bewertungsskala von gut bis schlecht, so kann man dabei eher von äquidistanten Abständen ausgehen als bei Schulnoten. Dort könnte man sich fragen, ob der Abstand von vier nach fünf der gleiche ist wie z.B. von drei nach vier.
Auf Intervallskalen lassen sich alle statistischen Größen berechnen, die keine „echte" Null voraussetzen. Die Null auf einer Intervallskala ist willkürlich festgelegt. Ein Beispiel hierfür sind Temperaturskalen. Der Nullpunkt auf einer Temperaturskala gemessen in Grad Celsius ist festgelegt als der Schmelzpunkt von Wasser. Eine in Kelvin gemessene Temperaturskala legt den Nullpunkt als den Punkt fest, bei dem alle Moleküle ruhen. Null Grad Kelvin entsprechen -273,15 Grad Celsius. (vgl. Meschede, 2006, S. 209,211)
Anhand des Beispiels ist erkennbar, dass sich Werte aus verschiedenen Temperaturskalen nicht sinnvoll vergleichen lassen. Ihre Werte sind um einen konstanten Wert verschoben. Zusätzlich zu den Operationen welche auch bei nominalen und ordinalen Skalen vorhanden sind lassen sich auf Intervallskalen Intervalle bzw. Differenzen vergleichen.
Es handelt sich um eine rationale Skala, wenn die Null absolut ist, d.h. Null bedeutet hier „nicht vorhanden". Beispiele für rationale Skalen sind Zeitmessungen und Zählungen. Ein Wert von null bedeutet dabei es ist keine Zeit vergangen bzw. es wurde nichts (null) gezählt. Eine Temperatur von null hingegen heißt nicht, dass es keine Temperatur gibt.

(a) Datenpaare

x	y	x	y
1	5	11	1
2	8	12	3
3	3	13	14
4	12	14	5
5	13	15	16
6	7	16	14
7	18	17	3
8	6	18	10
9	9	19	2
10	11	20	14

(b) Streudiagramm

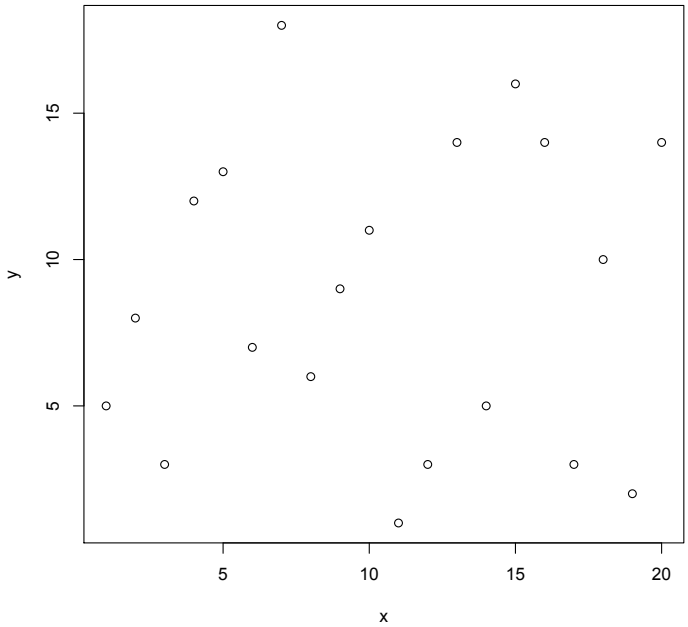


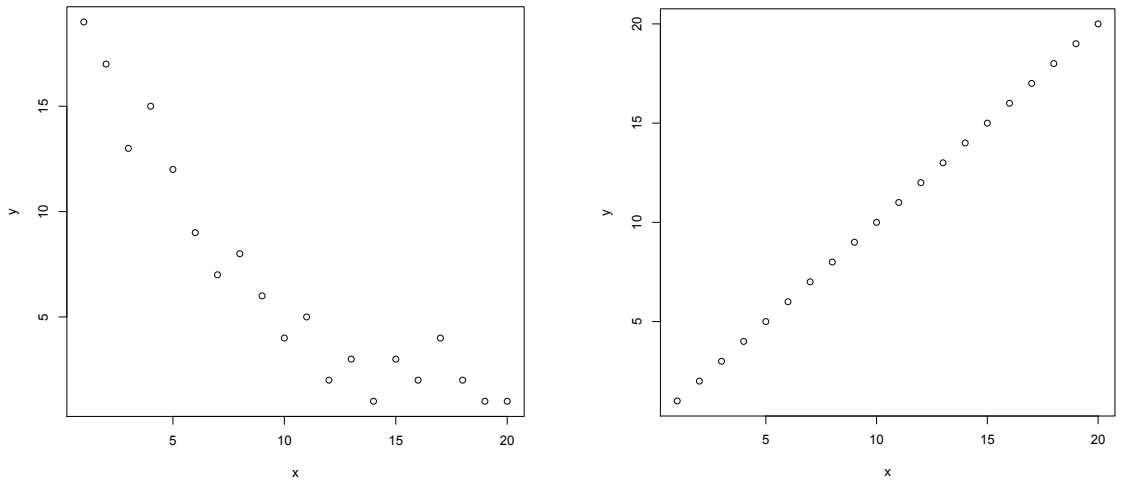
Abbildung 2.1: Streudiagramm und Datenpaare (Beispiel)

Grafiken

Das sogenannte *Streudiagramm (scatterplot)* eignet sich zur Visualisierung von Wertepaaren. In Experimenten sind Wertepaare in Form von unabhängigen (manipulierten) Variablen und abhängigen (beobachteten) Variablen zu finden. Sie werden anhand ihrer x- und y-Koordinaten in ein zweidimensionales Diagramm eingetragen (vgl. Ross, 2006, S. 32). Ein fiktives Beispiel ist in Abbildung 2.1 dargestellt.

Ein Streudiagramm kann auch dazu dienen, eine eventuelle Korrelation der Daten aufzuzeigen, d.h. ob generell größere x-Werte mit größeren y-Werten oder größere x-Werte mit kleineren y-Werten zusammenhängen. Ein Maß dafür gibt der *Korrelationskoeffizient*. Seien (x_i, y_i) mit $i = 1, \ldots, n$ Wertepaare und s_x, s_y die Standardabweichungen der x- bzw- y-Werte, dann ist der Korrelationskoeffizient definiert durch:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{x})}{(n-1)s_x s_y} \tag{2.4}$$



(a) $r = -0.91$ (b) $r = 1$

Abbildung 2.2: Korrelationskoeffizient (Beispiel)

Für r gilt $-1 \leq r \leq 1$. Hängen große x-Werte mit großen y-Werten zusammen, so gilt $r > 0$ und umgekehrt große x-Werte mit kleinen y-Werten $r < 0$. Liegt nur ein schwacher oder gar kein Zusammenhang vor so ist $r \approx 0$. (vgl. Ross, 2006, S. 34)
Für die Datenpaare aus Abbildung 2.1a ist $r = 0.06$. In Abbildung 2.2a ist ein starker negativer, in Abbildung 2.2b ein maximaler positiver Zusammenhang dargestellt.
Versucht man mit Hilfe des Korrelationskoeffizienten einen Zusammenhang von unabhängiger und abhängiger Variable nachzuweisen sei darauf hingewiesen, dass es sich nicht um einen Kausalzusammenhang handelt. Es kann mit Hilfe des Korrelationskoeffizienten nicht entschieden werden, in welcher Richtung die Kausalität gilt. Unklar ist, ob die Veränderung der unabhängigen Variablen die Ursache und die beobachtete Veränderung der abhängigen Variablen die Wirkung ist oder umgekehrt. (vgl. Ross, 2006, S. 36)
Handelt es sich bei dem darzustellenden Datensatz nicht um Wertepaare sondern um einzelne Werte, können diese mit Hilfe von *Boxplots* dargestellt werden. Ein Beispiel für ein Boxplot basierend auf den y-Werten aus Abbildung 2.1a ist in Abbildung 2.3 dargestellt.
Es handelt sich um ein eindimensionales Diagramm. Die Box erstreckt sich vom unteren/ersten Quartil bis zum oberen/dritten Quartil. Die Strecke dazwischen (Q3 - Q1) wird *Quartilabstand (interquartile range)* genannt. In der Box ist der Median (schwarzer Punkt) und der Mittelwert (M) sowie deren Standardfehler eingezeichnet. Außerdem enthalten sind die einzelnen Datenpunkte (Kreise). Die linienförmigen Ausläufer werden *whiskers* genannt. Der linke whisker endet dort

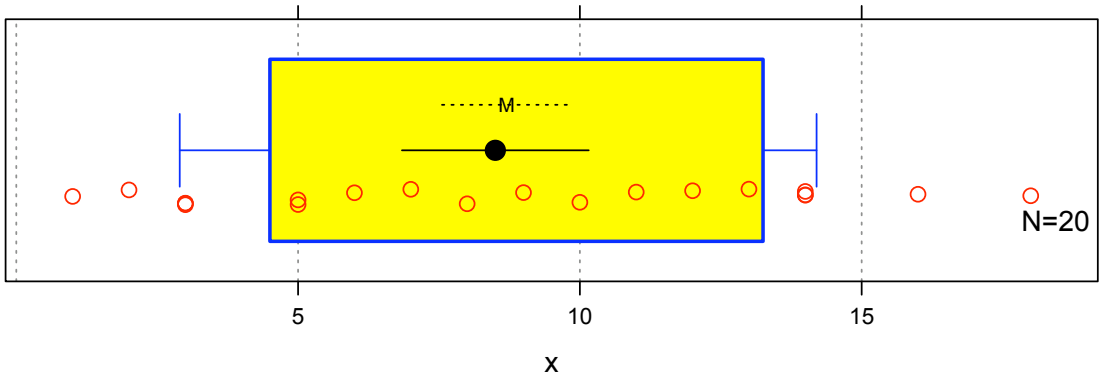


Abbildung 2.3: Boxplot (Beispiel 1)

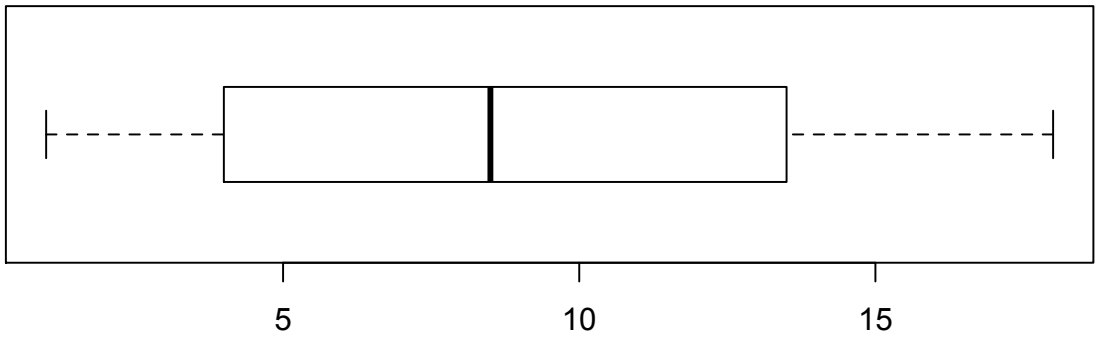


Abbildung 2.4: Boxplot (Beispiel 2)

wo genau 10% der Datenwerte kleiner sind, der rechte bei 90%. Boxplots eignen sich z.B. besonders gut, um eine abhängige Variable eines Experiments für die verschiedenen Gruppen untereinander in einer Skala zu visualisieren. Dort sind Unterschiede wesentlich besser zu erkennen als in Tabellen mit numerischen Werten.
Boxplots werden allerdings teilweise unterschiedlich dargestellt. So existieren z.B. auch Diagramme, in denen keine einzelnen Datenwerte (*strips*) dargestellt werden (siehe Abbildung 2.4). (vgl. Ross, 2006, S. 25)
In einigen Fällen sind außerhalb der whiskers Datenpunkte eingezeichnet. Es handelt sich dabei um sogenannte Ausreißer. Bower (2009b) definiert eine Größe

$$k = 1.5 \cdot (Q3 - Q1) \qquad . \tag{2.5}$$

Alle Datenwerte außerhalb des Intervalls $[Q1 - k, Q3 + k]$ werden als Ausreißer behandelt. Im Gegensatz zu Mittelwert und Standardabweichung, welche durch Ausreißer beeinflusst werden, sind Boxplots dahingegen resistent, dass die inter-

essanten Aspekte (die Box, in der 50% der Daten liegen) nicht betroffen sind.
In kontrollierten Experimenten sollten die gewonnenen Daten auf Ausreißer hin überprüft werden, da einige statistische Auswertungsmethoden auf Vergleichen von Mittelwerten oder Standardabweichungen beruhen.

2.2.2 Signifikanztests (allgemein)

Mit den aus kontrollierten Experimenten gewonnenen Daten der Stichproben können mit Hilfe von statistischen Methoden Schlüsse über die Grundgesamtheit getroffen werden. Es wird eine im Vorfeld aufgestellte *Hypothese*, also eine Vermutung über einen oder mehrere Parameter der Verteilung der Grundgesamtheit (Mittelwert, Standardabweichung, Varianz), getestet. (vgl. Ross, 2006)
Eine Hypothese kann z.B. die Vermutung sein, dass mit Hilfe eines bestimmten Werkzeuges eine Programmieraufgabe effizienter (also in geringerer Zeit) gelöst werden kann. Ein kontrolliertes Experiment, das diese Hypothese untersucht, würde dann z.B. aus zwei Gruppen bestehen, in denen die eine Gruppe die Aufgabe mit Hilfe des Werkzeuges, die andere aber ohne lösen muss. Für beide Gruppen würde die Zeit (und eventuell noch in geeigneter Form die Korrektheit) gemessen werden. Die Hypothese in Form von Parametern besagt dann, dass der Mittelwert der benötigten Zeit zum Lösen der Aufgabe mit Werkzeug geringer ist als ohne. Statistische Signifikanztests können diese Hypothese anhand der Stichprobendaten untersuchen.
In einem Hypothesentest wird die sogenannte *Nullhypothese* gegen die *Alternativhypothese* getestet. Handelt es sich um ein Problem mit nur einer Stichprobe so wäre die Nullhypothese z.B. $\mu = \mu_0$, also der Mittelwert der Grundgesamtheit ist gleich einem konstanten Wert μ_0. Die zugehörige Alternativhypothese ist $\mu \neq \mu_0$. Der Mittelwert ist also nicht gleich dem konstanten Wert [2]. Hierbei handelt es sich um einen zweiseitigen Test. Ist im Vorfeld schon eine Tendenz bekannt, so kann auch ein einseitiger Test durchgeführt werden. Dann wäre die Alternativhypothese entweder $\mu < \mu_0$ (Nullhypothese $\mu \geq \mu_0$) oder $\mu > \mu_0$ (Nullhypothese $\mu \leq \mu_0$) . (vgl. Ross, 2006, S. 273)
Ein oft vorkommendes Problem ist der Vergleich zweier Verfahren, Methoden oder Werkzeuge (Zweistichprobenproblem). Hier lautet die Nullhypothese dann $\mu_x = \mu_y$, also die Gleichheit der beiden Mittelwerte der jeweiligen Grundgesamtheiten, und die Alternativhypothese $\mu_x \neq \mu_y$ bzw. für einseitige Tests $\mu_x > \mu_y$ (Nullhypothese $\mu \leq \mu_0$) oder $\mu_x < \mu_y$ (Nullhypothese $\mu \geq \mu_0$) . (vgl. Ross, 2006, S. 284)
Werden mehr als zwei Stichproben verglichen, so besagt die Nullhypothese, dass

[2] Anstelle der Mittelwerte können auch andere Parameter von Verteilungen wie Varianz oder Standardabweichung getestet werden. Die Mittelwerte sind also nur ein Beispiel.

	H_0 **wahr** (H_1 falsch)	H_0 **falsch** (H_1 wahr)
H_0 **akzeptiert**	korrekt	Fehler zweiter Art
H_0 **verworfen**	Fehler erster Art	korrekt

Tabelle 2.3: Fehler erster und zweiter Art bei Hypothesentests

all deren Mittelwerte gleich sind, die Alternativhypothese, dass mindestens einer ungleich ist. (vgl. Ross, 2006, S. 400)
Das Ergebnis von Hypothesentests ist das Akzeptieren oder Verwerfen der Nullhypothese. Tabelle 2.3 zeigt alle möglichen Ausgänge von Hypothesentests. H_0 ist die Nullhypothese, H_1 die Alternativhypothese.
Ist H_0 in Wirklichkeit wahr (damit H_1 falsch) und wird vom Test akzeptiert, so ist die Entscheidung korrekt. Wird sie allerdings verworfen, so ist dies falsch und wird als Fehler erster Art bezeichnet. Die Wahrscheinlichkeit einen Fehler erster Art zu begehen wird als Signifikanzniveau α bezeichnet und wird als Eingabeparameter eines statistischen Tests festgelegt. (vgl. Dybå et al., 2006; Bower, 2009c; Ross, 2006, S. 266-267)
Wenn stattdessen H_1 wahr ist (H_0 damit falsch) und (fälschlicherweise) H_0 doch akzeptiert wird, ist dies ein Fehler zweiter Art. Die Wahrscheinlichkeit einen solchen zu begehen wird Operationscharakteristik β genannt. Die Gegenwahrscheinlichkeit $1-\beta$, also die Wahrscheinlichkeit die Nullhypothese korrekterweise zu verwerfen, wird als Macht bzw. Trennschärfe (*statistical power*) bezeichnet. (vgl. Dybå et al., 2006; Ross, 2006, S. 271-273)
Die Trennschärfe kann entweder vor Durchführung des Experiments (*a priori*) oder anhand der Ergebnisse (*post hoc*) durchgeführt werden. (vgl. Dybå et al., 2006)
Allerdings ist es empfehlenswert eine a priori Berechnung durchzuführen, denn ist die Trennschärfe (Wahrscheinlichkeit einen signifikanten Unterschied zu entdecken) niedrig, so hat ein Experiment keinen Nutzen. Eine geringe Trennschärfe sollte zu einer Änderung der Pläne führen. (vgl. Dybå et al., 2006; Cohen, 1988)
Wird überhaupt keine Berechnung der Trennschärfe durchgeführt und der Test ergibt, dass die Alternativhypothese nicht verworfen werden kann, so kann die Ursache dafür nicht benannt werden. Es könnte einerseits sein, dass das Phänomen tatsächlich nicht existiert, andererseits könnte die Trennschärfe so gering gewesen sein, dass ein eventuell exisitierender Unterschied nur zu sehr geringer Wahrscheinlichkeit hätte gefunden werden können. (vgl. Dybå et al., 2006; Bower, 2009d)
In der Berechnung der Trennschärfe sind drei Parameter involviert: Das Signifikanzniveau α, die Stichprobengröße N und die Effektgröße (siehe unten) der

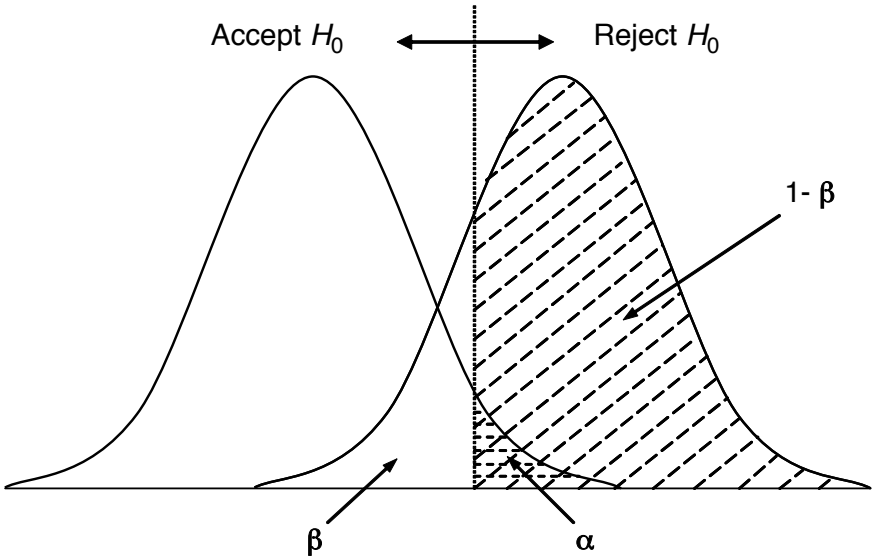


Abbildung 2.5: Veranschaulichung der Fehlertypen aus Dybå et al. (2006)

Grundgesamtheiten. (vgl. Dybå et al., 2006)
Somit kann z.B. auch die benötigte Stichprobengröße berechnet werden, um eine gewünschte Trennschärfe (oft 0.80 (vgl. Bower, 2009d)) zu erreichen. Außerdem kann mit der geplanten Stichprobengröße, mit der das Experiment durchgeführt werden soll, sowie mit einer Schätzung der Effektgröße aus Erfahrungswerten oder früheren Experimenten die Trennschärfe, also die Wahrscheinlichkeit einen möglichen signifikanten Unterschied zu entdecken, berechnet werden. Für bereits durchgeführte Experimente, in denen keine Trennschärfeanalyse durchgeführt wurde, kann dies im Nachhinein noch erfolgen.
Eine Veranschaulichung des Sachverhalts vom Akzeptieren oder Verwerfen der Nullhypothese und den involvierten Wahrscheinlichkeiten bietet die aus Dybå et al. (2006) entnommene Abbildung 2.5. Diese zeigt einen einseitigen Test mit den Wahrscheinlichkeitsverteilungen für Null- bzw. Alternativhypothese. Die gestrichelte vertikale Linie stellt die Grenze zwischen Akzeptieren und Verwerfen der Nullhypothese dar. Die Fläche, die von der jeweiligen Verteilung über diese hinweg geht, beschreibt die jeweiligen Fehlergrößen α und β.
Ob die Nullhypothese verworfen wird, wird anhand des aus dem statistischen Test resultierenden p-Wertes entschieden. Er gibt die Wahrscheinlichkeit an, die auf den Daten des Experiments basierenden Ergebnisse (abhängige Variablen) zu erhalten unter der Annahme, dass H_0 wahr ist. (vgl. Bower, 2009c)
Als Merkregel empfielt Bower (2009c):

> "If the p-value is low, the null hypothesis must go."

Resultiert ein niedriger p-Wert, so ist es unwahrscheinlich solche Ergebnisse (wie sie im Experiment gemessen wurden) zu erhalten, wenn die Nullhypothese wahr ist. Daher wird die Nullhypothese verworfen.
Quantitativ bedeutet ein niedriger p-Wert $p < \alpha$. Das Signifikanzniveau α wird meist auf 0.05 festgelegt. (vgl. Dybå et al., 2006)
Verwirft man die Nullhypothese zugunsten der Alternativhypothese, so ist noch keine Aussage darüber getroffen, wie groß der Unterschied ist. Um diesen zu bestimmen berechnet man *Effektgrößen (effect sizes)*. Die Effektgröße beschreibt die Größenordnung des Unterschiedes der verglichenen Methoden oder Werkzeuge. Da allerdings keine Daten der Grundgesamtheit zur Verfügung stehen, lässt sich nur eine geschätzte Effektgröße aus den Daten der Stichprobe berechnen (*estimated effect size*) und nicht die tatsächliche (*population effect size*). Die Effektgröße der Grundgesamtheit lässt sich allerdings durch Aggregation mehrerer geschätzer Effektgrößen bestimmen (Meta-Analyse). (vgl. Kampenes et al., 2007)
Kampenes et al. (2007) unterscheiden zwischen standardisierten und nichtstandardisierten Effektgrößen. Standardisierte Effektgrößen sind skalenfrei und berechnen sich durch den Quotienten aus Mittelwertdifferenz und Standardabweichung:

$$d_{pop} = \frac{\mu_1 - \mu_2}{\sigma} \tag{2.6}$$

Da die Standardabweichung der Grundgesamtheit σ nicht bekannt ist, wird diese geschätzt. Kampenes et al. (2007) nennt mit Hedges' g, Cohen's d und Glass' Δ drei häufig benutzte Schätzer. Nichtstandardisierte Effektgrößen werden entweder prozentual oder direkt in den Einheiten in denen die Variablen gemessen wurden (Mittelwertdifferenz) angegeben. Dadurch sind sie laut Kampenes et al. (2007) einfacher zu interpretieren aber nicht skalenfrei.

2.2.3 Signifikanztests (konkret)

Signifikanztests lassen sich generell in parametrische und nichtparametrische Tests aufteilen. Parametrisch bedeutet, dass die Verteilung der Daten der Grundgesamtheit bis auf einen Satz von Parametern wie z.B. Mittelwert und Standardabweichung bekannt ist. (vgl. Ross, 2006, S. 183)
Eine bekannte Verteilung ist die sogenannte *Normalverteilung* (vgl. Meschede, 2006, S. 7). Sie ist definiert durch

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-(x-\overline{x})^2/(2\sigma^2)} \tag{2.7}$$

Die Normalverteilung für den Mittelwert $\mu = 1$ einer Datenmenge mit Standardabweichung $\sigma = 1$, kurz $\mathcal{N}(\mu, \sigma) = \mathcal{N}(1, 1)$, ist in Abbildung 2.6 dargestellt.

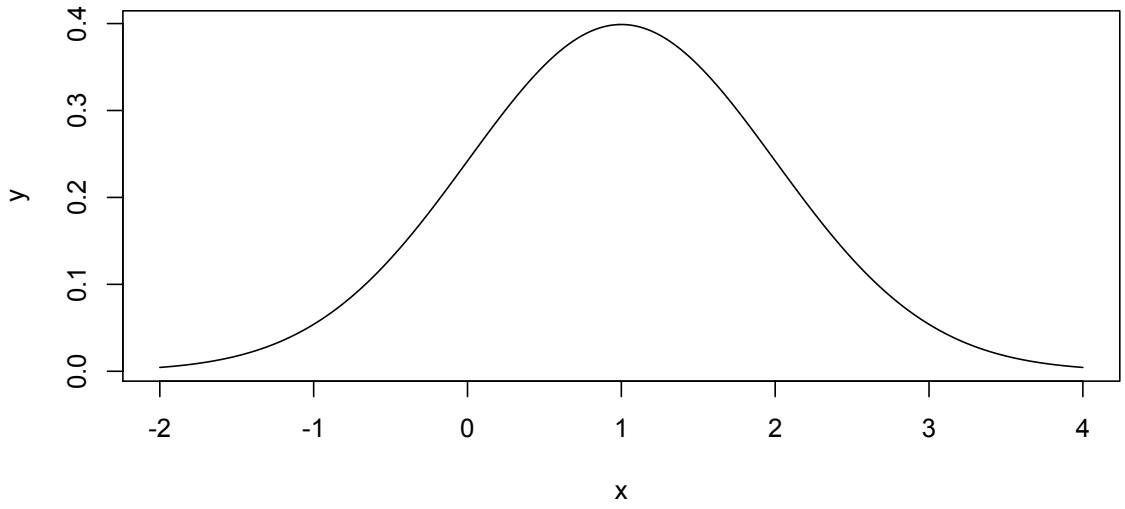


Abbildung 2.6: Normalverteilung $\mathcal{N}(1,1)$

Das Maximum der Normalverteilung liegt beim Mittelwert. Im Intervall $[\overline{x}-\sigma, \overline{x}+\sigma]$ befinden sich ca. 68% der Messwerte, ca. 95% im Intervall $[\overline{x}-2\sigma, \overline{x}+2\sigma]$ und 99,7% im Intervall $[\overline{x}-3\sigma, \overline{x}+3\sigma]$. Dieser Sachverhalt ist auch unter dem Namen empirische Regel bekannt. (vgl. Ross, 2006, S. 30-31)
Um zu testen, ob eine Stichprobe aus einer normalverteilten Grundgesamtheit stammt, werden Anpassungstests durchgeführt. Hier lautet die Nullhypothese, dass der Stichprobe eine Normalverteilung zu Grunde liegt. (vgl. Ross, 2006, S. 437)
Zwei konkrete Verfahren, um dies zu testen, sind der Shapiro-Wilk Test und der Kolmogorov-Smirnov Test. (vgl. Shapiro and Wilk, 1965; NIST, 2009b)
Allerdings lässt sich mit diesen Tests nur nachweisen, dass keine Normalverteilung vorliegt, denn die Nullhypothese kann nicht bewiesen werden (siehe dazu auch Abschnitt 2.2.4).
Eine grafische Methode, um zu überprüfen ob eine Datenmenge einer Normalverteilung genügt, ist der Quantil-Quantil-Plot (bzw. "Normal-Probability-Plot"). Die Daten werden dabei in einem zweidimensionalen Diagramm gegen eine theoretische Normalverteilung gezeichnet. Liegen die Punkte annähernd auf einer Geraden, so spricht dies für eine Normalverteilung. (vgl. NIST, 2009c)
Ein Beispiel für ein Quantil-Quantil-Plot für eine zufällig generierte Menge von Punkten aus einer Normalverteilung[3] zeigt Abbildung 2.7.

[3]siehe *rnorm* in R (R Documentation, 2009)

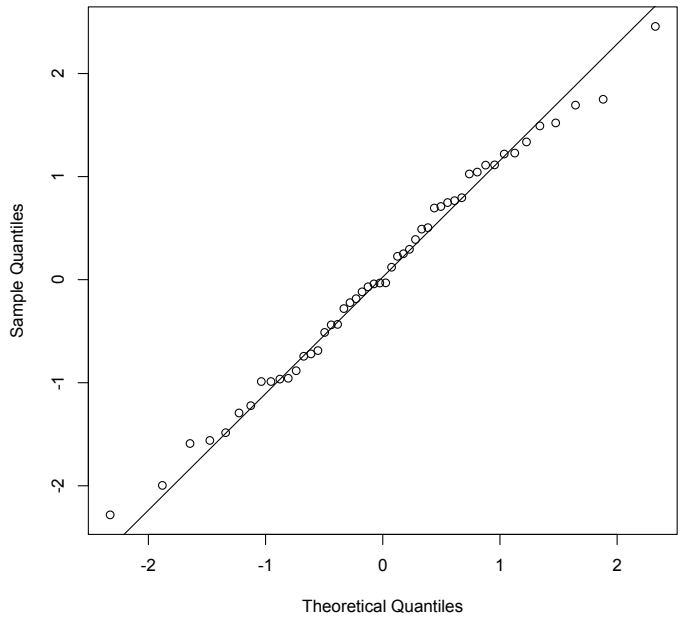


Abbildung 2.7: Quantil-Quantil-Plot (Beispiel)

Parametrische Tests

Parametrische Tests sind z.B. der t-Test und die Varianzanalyse (ANOVA). Der t-Test gliedert sich in eine Reihe verschiedener Tests. Diese sind der Einstichproben t-Test, der Zweistichproben t-Test und der t-Test für gepaarte Stichproben. Der Zweistichproben t-Test vergleicht die Mittelwerte zweier Stichproben und lehnt die Nullhypothese ab, sofern die Differenz ungleich null ist (sofern es sich um einen zweiseitigen Test wie in Abschnitt 2.2.2 beschrieben handelt) bzw. größer/kleiner eines bestimmten Wertes (bei einseitigen Tests) ist. Als Annahmen setzt er voraus, dass die Stichproben aus normalverteilten Grundgesamtheiten stammen und ihre Varianzen gleich sind (Ross, 2006, S. 286). Werden die Annahmen nicht eingehalten, so kann dies das Ergebnis (p-Wert) beeinflussen. (vgl. Miller, 1997, S.5 und S.41)
Wurde bei einem Experiment ein Vortest durchgeführt oder es wurde ein "within-participants-design"[4] gewählt, so kann der t-Test für gepaarte Stichproben verwendet werden. Die Information, dass bestimmte Werte von gleichen Subjekten entstanden sind, kann somit in den Test mit einbezogen werden. Der t-Test für gepaarte Stichproben entspricht dem Einstichproben t-Test der Differenzen der Stichprobenpaare. (vgl. Ross, 2006, S. 293)

[4]Im Gegensatz zum "between-participants-design" werden hier die unterschiedlichen Konditionen auf alle Subjekte angewendet. (vgl. Shadish et al., 2002, S.505,513)

Sollen zwei oder mehr Stichproben verglichen werden, so kann die Varianzanalyse (ANOVA) verwendet werden. So wie beim t-Test gibt es auch hier verschiedene Analysen. Die einfache Varianzanalyse (*one-way ANalysis Of VAriance*) testet den Einfluss eines sogenannten Faktors auf einen Datensatz. Also zum Beispiel den Einfluss einer verwendeten Methode auf die benötigte Zeit. Es wird überprüft, ob alle Mittelwerte gleich sind (Nullhypothese) oder mindestens einer ungleich ist (Alternativhypothese). Genauso wie beim t-Test setzt die Varianzanalyse voraus, dass die Stichproben aus normalverteilten Grundgesamtheiten stammen und die gleiche Varianz besitzen. Außerdem erlaubt die Varianzanalyse den Einfluss mehrerer Faktoren zu untersuchen. Je nachdem wieviele Faktoren untersucht werden, wird sie dann n-fache Varianzanalyse (n-*way ANOVA*) genannt, wobei n für die Anzahl der Faktoren steht. (vgl. Ross, 2006, S. 410)
Somit kann z.B. der Einfluss eines Verfahrens und der Erfahrung der Versuchspersonen (dies wären in einem Experiment die unabhängigen Variablen oder auch die Faktoren der Varianzanalyse) auf die abhängige Variable der benötigten Zeit zur Lösung der Aufgaben bestimmt werden. Für jeden Faktor sowie deren Interaktionen ergibt sich ein eigener p-Wert, der angibt, ob der Faktor bzw. die Interaktion mehrerer Faktoren signifikant sind (einen Einfluss auf die abhängige Variable haben).
Die Berechnung für die zweifache Varianzanalyse kann in Ross (2006, S. 410-417) nachgelesen werden. Konzeptionell werden die Messdaten in Matrizen (bei zweifacher Analyse bilden die Spalte und die Zeile die beiden Faktoren) angeordnet und diese auf einen eventuellen Zeilen- bzw. Spalteneffekt hin untersucht (eben ob es einen Einfluss des Faktors gibt oder nicht).

Nichtparametrische Tests

Nichtparametrische Tests lassen sich anwenden ohne eine Annahme über die Form der Verteilung zu treffen. Sie lassen sich auch dann anwenden, wenn Daten nur auf ordinalen Skalen vorliegen. Sie basieren auf der Berechnung von Rangsummen. Dabei werden alle gemessenen Werte nach ihrer Ordnung aufsteigend sortiert und ihnen Ränge zugeordnet. Der Rang ist dabei deren Position in der geordneten Liste. Sind doppelte Werte vorhanden, so wird der Gruppe von Doppelten der Mittelwert ihrer Ränge zugeordnet. (vgl. Kruskal and Wallis, 1952; Ross, 2006)
Der Mann-Whitney-U bzw. Wilcoxon-Test[5] testet die Nullhypothese „H_0: Die Mediane der Grundgesamtheiten sind gleich.“ gegen die Alternativhypothese „H_A: Die Mediane sind nicht gleich“. Der Mann-Whitney-U Test ist die nichtparame-

[5]Der Wilcoxon-Test funktioniert nur für zwei Stichproben gleicher Größe, der Mann-Whitney-U Test erlaubt auch unterschiedlich große Stichproben. (vgl. Kruskal and Wallis, 1952)

trische Alternative zum Zwei-Stichproben t-Test. (vgl. NIST, 2009d)
Für gepaarte Daten lässt die der Wilcoxon-Signed-Rank Test anwenden. Dieser ordnet die Beträge der Differenzen der Datenpaare ebenfalls nach ihrer Ordnung. Analog zum t-Test kann auch hier der Wilcoxon-Test für das Einstichprobenproblem auf die Differenzen angewendet werden. Die Nullhypothese sagt aus, dass der Median der Differenzen gleich einem bestimmten Wert (meist 0) ist.
Sollen mehr als zwei Stichproben verglichen werden, so kann der Kruskal-Wallis Test benutzt werden. Er stellt die Alternative zur parametrischen einfachen Varianzanalyse dar. (vgl. Kruskal and Wallis, 1952)
Eine weitere nichtparametrische Methode, die allerdings nicht auf der Berechnung von Rangsummen und Medianvergleichen beruht, ist der sogenannte Bootstrap. Wie andere parametrische Tests, kann die Methode angewendet werden, ohne dass Annahmen über die Verteilung der Daten getroffen werden müssen. Mit der Bootstrap-Technik ist es u.A. möglich Standardfehler und Konfidenzintervalle sowie p-Werte zum Testen von Hypothesen zu berechnen. Dabei wird die Stichprobe so behandelt, als wäre sie die Grundgesamtheit. Dann wird eine festgelegte Anzahl an Stichproben aus der eigentlichen Stichprobe gezogen und für diese die Teststatistik (hier: die Mittelwertdifferenz) berechnet. Als Ergebnis erhält man den Mittelwert der Mittelwertdifferenzen sowie dessen Standardabweichung (Standardfehler). (vgl. UCLA: Academic Technology Services, Statistical Consulting Group, 2009)
Die Verwendung der Bootstrap-Technik in R beschreibt Quick-R (2009). Ist man am Konfidenzintervall der Mittelwertdifferenz interessiert, sollte das (von R ausgegebene) BCa-Konfidenzintervall verwendet werden. BCa steht für "Bias-correction and acceleration". Eventuelle Verzerrungen der Teststatistik sowie des Standardfehlers sind hier mit einberechnet.(vgl. Thompson, 2009)
Die Bootstrap-Technik berechnet eine vorgegebene Anzahl an Mittelwertdifferenzen aus der gezogenen Stichprobe. Für einseitige Hypothesentests mit der Nullhypothese $\mu_x - \mu_y < 0$ berechnet sich der p-Wert als

$$p = \frac{|\text{Mittelwerte} \geq 0|}{|\text{berechnete Mittelwerte}|} \quad . \tag{2.8}$$

Sind die voraussetzenden Annahmen für parametrische Tests verletzt, so ist die Chance einen signifikanten Unterschied zu entdecken bei nicht-parametrischen Tests größer. (vgl. Kruskal and Wallis, 1952)
Sind die Annahmen erfüllt, sollten parametrische Tests genutzt werden, da diese aussagekräftigere Ergebnisse liefern. (vgl. Ross, 2006, S. 465)

Regressionsanalyse

Existiert nur eine Ergebnisvariable Y (abhängige Variable), die von mehreren Einflußvariablen $x_1, \ldots, x_n$ (unabhängigen Variablen) abhängt, so lässt sich mit Hilfe der Regressionsanalyse prüfen, ob zwischen ihnen ein linearer Zusammenhang der Form

$$Y = k_0 + k_1 x_1 + \ldots + k_n x_n \qquad k_0, \ldots, k_n \text{ konstant} \tag{2.9}$$

besteht. Hypothesen lassen sich für jeden Regressionsparameter k_i aufstellen. Die Nullhypothese besagt dann, dass die Ergebnisvariable nicht von der Einflussvariablen abhängt, also $H_0 : k_i = 0$ und $H_A : k_i \neq 0$. Ob die Anwendung der Regressionsanalyse sinnvoll ist, lässt sich mit Hilfe des Streudiagramms testen. Dort sollten die Datenpunkte annähernd auf einer Geraden liegen. Ist ein exponentieller Zusammenhang im Diagramm erkennbar, so ist es möglich die logarithmierten Daten zu testen. (vgl. Ross, 2006, S. 319, 330, 346)

Korrelationsanalyse

Wie schon in 2.2.1 beschrieben, ist der Korrelationskoeffizient ein Maß für den Zusammenhang zweier Variablen. Oft verwendete Koeffizienten sind der Spearman und der Pearson Korrelationskoeffizient. Für ordinale Daten sollte Spearman, für rationale Pearson verwendet werden. Hier lässt sich die Null-Hypothese $H_0 : r = 0$ gegen die Alternativhypothese $H_A : r \neq 0$ testen. Der resultierende p-Wert gibt die Wahrscheinlichkeit an, einen Koeffizient wie den beobachteten zu bekommen, wenn die Null-Hypothese wahr ist (also der wirkliche Koeffizient gleich null). (vgl. StatPac Inc., 2009)

2.2.4 Kritik an Hypothesentests

Obwohl statistische Signifikanztests in der Praxis oft verwendet werden, existieren auch diverse Kritiken. Laut Gliner et al. (2002) existieren zwei häufige Fehlinterpretationen bei Hypothesentests. Zum einen werde oft fälschlicherweise angenommen, dass ein kleinerer p-Wert eine stärkere Beziehung bedeute. Zum anderen bedeute statistische Signifikanz (verwerfen der Nullhypothese zugunsten der Alternativhypothese) gleichzeitig auch einen praktischen Nutzen.
Cohen (1994) sagt, dass Hypothesentests nicht das aussagen was wir wissen wollen. Uns interessiere: Was ist die Wahrscheinlichkeit dafür, dass H_0 wahr ist, wenn bestimmte Daten gegeben sind ($P(H_0|D)$). Hypothesentests würden allerdings die Wahrscheinlichkeit angeben, solche Daten zu bekommen, sofern H_0 wahr ist ($P(D|H_0)$). Der p-Wert würde deshalb oft fehlinterpretiert werden als die Wahrscheinlichkeit dafür, dass H_0 falsch ist.
Greenwald et al. (1996) stellen drei generelle Mängel bei Hypothesentests fest:

1. Da die Nullhypothese quasi immer falsch ist, ist deren Test uninformativ.
2. Nullhypothesentests sagen nicht das aus was wir wissen wollen.
3. Nullhypothesentests sind parteiisch gegenüber der Nullhypothese.

Dass Nullhypothesensignifikanztests (NHST) nicht das aussagen was wir wissen wollen, lege daran, dass sie keine Schätzung über das Ausmaß der kausalen Beziehung der Variablen angeben (Effektgrößen).
Da NHST lediglich Schlüsse in eine Richtung (gegen die Nullhypothese) zulassen, seien sie parteiisch. Die Nullhypothese könne nicht bewiesen werden. Wird sie verworfen, so wird beschlossen, sie ist falsch. Kann sie allerdings auf Grund eines hohen p-Wertes nicht verworfen werden, so könne nichts ausgesagt werden.
Warum NHST trotzdem ein gängiges Verfahren bliebe, erklären Greenwald et al. (1996) damit, dass NHST eine Entweder-Oder Antwort liefern. Entweder wird die Nullhypothese verworfen oder eben nicht. Sie seien nützlich um Fragen zu klären, die eine Vorhersage über eine Tendenz treffen, weniger allerdings, wenn es darum ginge, bestimmte Werte von Parametern zu bestimmen.
Um die Schwächen von NHST wie Fehlinterpretationen von p-Werten zu umgehen, empfehlen Greenwald et al. (1996) p-Werte in numerischer Form anzugeben, anstelle lediglich zu berichten, die Ergebnisse seien signifikant ($p < 0.05$) oder nicht signifikant ($p > 0.05$).

Kapitel 3

Verwandte Arbeiten

Sjøberg et al. (2005) erfassten kontrollierte Experimente im Zeitraum von 1993 bis 2002 aus zwölf Zeitschriften und Konferenzen des Software Engineerings. Sie untersuchten wie die Experimente durchgeführt und in wie weit relevante Informationen berichtet wurden. Insgesamt wurden über 5.400 Artikel untersucht. In 103 der Artikel fanden sie 113 kontrollierte Experimente (1.9%). Für diese wurde folgendes qualitativ festgehalten:

- Themenbereich
- Anzahl, Bildungsgrad, Rekrutierung von Versuchspersonen
- Typ, Dauer, Kategorie und Umfang der zu erledigenden Aufgaben
- Ort der Durchführung und verwendete Hilfsmittel
- Behandlung von interner und externer Validität
- Replizierung

Die Themen wurden nach Glass et al. (2002) kategorisiert, die Artikel anhand von IEEE (2009) klassifiziert. Das am häufigsten vorkommende Thema (49%) ist „Software Life Cycle/engineering“, gefolgt von „Methods/Techniques“ (32%). Das prominenteste Schlüsselwört nach der IEEE Taxonomie ist „Code Inspections and Walkthroughs“ (35%).
Von über 5.400 Versuchspersonen sind 9% Fachleute, 87% Studenten. Die erledigten Aufgaben sind zur Hälfte Aufgaben aus dem Bereich Analyse (neben Planung, Entwicklung und Modifikation). Drei Viertel der verwendeten Applikationen waren entweder für das Experiment konstruiiert bzw. Teil von Studentenprojekten. Ein Experiment wurde in einer Büroumgebung mit Fachleuten durchgeführt.

Gefahren zur internen/externen Gültigkeit sind explizit in 71/78 Experimenten angegeben. 20 Experimente sind als Replizierung vorheriger Experimente beschrieben.
Sjøberg et al. (2005) schließen, dass die Darstellung der untersuchten Eigenschaften der Experimente nicht systematisch ist. Es wird keine einheitliche Terminologie verwendet. Sie fordern Richtlinien für Methodik und geben Vorschläge dafür.

Basierend auf den extrahierten kontrollierten Experimenten von Sjøberg et al. (2005) untersuchten Dybå et al. (2006) die Trennschärfe (*statistical power*) von statistischen Tests. Sie fanden zwölf Artikel, die Trennschärfe beinhalten, neun davon beschreiben Verfahren, um sie zu bestimmen (drei „a priori“ und sechs „post hoc“). 92 der 113 identifizierten Experimente (78 von 103 Artikeln) beinhalten insgesamt 459 statistische Tests, für die eine „post hoc“ Berechnung der Trennschärfe möglich ist. Diese wurde für die nach Cohen (1988, S. 26) definierten kleinen, mittleren und großen Effektgrößen durchgeführt. Die Ermittlung der Trennschärfe ergab 0.11, 0.36 bzw. 0.63. Für kleine Effektgrößen besitzen 97% der verwendeten Tests (78% für mittlere, 69% für große) eine Chance von weniger als 50% einen möglichen signifikanten Unterschied zu erkennen.
Aus den Beobachtungen schließen die Autoren, dass die Ergebnisse verglichen mit anderen Disziplinen wie „Information Systems“ wesentlich schlechter sind. Die explizite Berücksichtigung von Trennschärfe im Empirical Software Engineering sei sehr begrenzt. Sie geben außerdem Hinweise, wie die Schärfe verbessert werden kann. Forscher sollten nicht angeben, dass ein Phänomen nicht existiert, sondern, dass sie keinen signifikanten Unterschied beobachten konnten (evtl. wegen zu geringer Trennschärfe). Sie stimmen mit der Empfehlung von Wilkinson (1999) und Kitchenham et al. (2002) überein, ausreichend statistische Daten wie Stichprobengröße, Signifikanzniveau, Mittelwerte und Standardabweichungen anzugeben, um Meta-Analysen zu ermöglichen. So lange nicht bewiesen sei, dass mittlere Effektgrößen im Software Engineering irrelevant sind, sollten die Ergebnisse als Besorgnis erregend gelten.

Kampenes et al. (2007) verwenden ebenfalls die Menge der identifizierten Experimente von Sjøberg et al. (2005). Von diesen wurden die 92 Experimente aus 78 Artikeln, in denen Hypothesentests durchgeführt werden, auf das Ausmaß der Beschreibung und Interpreation von Effektgrößen untersucht. 29% der Experimente geben mindestens eine Effektgröße an. Standardisierte Effektgrößen werden für 55, unstandardisierte für 46 der 429 Tests angegeben. Der praktische Nutzen wird nicht problematisiert.
Für 284 Tests, in denen ausreichend Informationen berichtet werden, wurde eine geschätzte standardisierte Effektgröße (*estimated standardized mean difference effect*

size) berechnet. Es ergaben sich zu jeweils einem drittel kleine, mittlere und große Effektgrößen.
Dass so wenige Artikel Effektgrößen behandeln könnte laut den Autoren daran liegen, dass wenige davon Kenntnis besitzen. Empirical Software Engineering sei außerdem eine junge Disziplin, daher könne man solche Ergebnisse erwarten. Sie empfehlen Effektgrößen, sowohl standardisiert als auch nicht standardisiert, immer anzugeben und deren praktischen Nutzen zu diskutieren.

Weiterhin verwendeten Hannay et al. (2007) die von Sjøberg et al. (2005) extrahierte Menge von Experimenten um zu untersuchen, welche Artikel explizit Theorien verwenden. Als Theorie bezeichnen sie eine Erklärung, warum und in welcher Weise ein Phänomen auftritt. Theorien seien eine wichtige Alternative zu Erkenntnissen, das sich durch entwickelnde Tendenzen gefestigt hat. Die Autoren geben einen Überblick darüber, welche Theorien für welche Zwecke verwendet wurden. Festgehalten sind die Identität und Rolle der Theorien. Zur Kategorisierung der Identität wurde die Taxonomie von Gregor (2006) adaptiert. Diese teilt die Theorien in fünf Typen (u.A. Analyse, Vorhersage) auf, gibt also Aufschluss darüber, was die Theorie zu erklären versucht. Für die Rollen wurden Kategorien entwickelt, die die Positionen wiederspiegeln, die Theorien einnehmen (z.B. Entwurf oder post hoc Erklärung).
In den Artikeln wurden insgesamt 40 Theorien aus 24 Artikeln extrahiert, 15 davon verwendeten explizit den Begriff Theorie. Da nur ein Viertel der durchsuchten Artikel Theorien nutzt, um Ursache-Wirkungsbeziehungen zu beschreiben, folgern die Autoren, dass sie keine Beweise für Forschung, die von Theorien angetrieben ist, gefunden haben. Vor allem würden die Theorien nicht benutzt werden, um als Rahmen für Diskussionen zu dienen, und in verschiedenen Artikeln vor allem von verschiedenen Autoren untersucht zu werden (zwei Theorien wurden in mehreren Artikeln verwendet, eine davon von verschiedenen Autoren). In anderen Wissenschaften sei die Nutzung von Theorien selbstverständlich, dort existiere lediglich die Frage wie man sie verwenden soll. Laut den Autoren beschreiben Theorien Phänomene in einem begrifflich einheitlichen Rahmen und lassen Forschungsfragen einfach erkennbar erscheinen. Sie erachten die Nutzung von Theorien vor allem als Basis für weitere Forschung für sinnvoll.

Die Untersuchung von Sjøberg et al. (2005) wurde von Kampenes et al. (2009) verfeinert, in dem randomisierte und Quasi-Experimente seperat betrachtet wurden. Vor allem wurden Zuweisungsmethoden/-prozeduren, Techniken zum Vermeiden von systematischen Fehlern bei der Auswahl von Versuchspersonen, Durchführung von Vortests und Berücksichtigung der Ergebnisse in der statistischen Berechnung betrachtet.

40 der 113 Experimente (35%) wurden als Quasi-Experimente indentifiziert, 10% verwendeten den Begriff explizit. Unter der Hälfte dieser verwenden Ergebnisse von Vortests zur Gruppenverteilung, 2 Experimente verwenden die Informationen auch in statistischen Berechnungen.
Aufgrund der geringen Verwendung von Vortestergebnissen und die geringe Berücksichtigung der möglichen Gefahren der internen Gültigkeit bzgl. systematischer Auswahlfehler schließen Kampenes et al. (2009), dass wenig Erkenntnis darüber besteht, wie Quasi-Experimente durchgeführt werden. Forscher haben Unkenntnis über den speziellen Entwurf und dessen statistische Auswertung. Daher müsse man daran arbeiten, das zu verbessern, um sicherzustellen, dass die aus den Quasi-Experimenten gezogenen Schlußfolgerungen gültig sind.

Die bisherigen Arbeiten aus dem Simula Research Laboratory[1] extrahieren, wie diese Arbeit, Informationen aus kontrollierten Experimenten, allerdings mit unterschiedlichem Schwerpunkt. Die statistische Auswertung wird, bis auf Trennschärfe, Effektgröße und verwendete statistische Verfahren, nicht betrachtet. Außerdem basieren die Untersuchungen auf einer Menge von älteren Artikeln aus den Jahren 1993-2002.

[1] http://www.simula.no

Kapitel 4

Methodik

In diesem Kapitel wird die Vorgehensweise beschrieben, mit der die Untersuchung durchgeführt wurde. Abschnitt 4.1 beschreibt die gewählte Prozedur zur Identifizierung von Artikeln mit kontrollierten Experimenten. Diese Grundmenge wurde auf die in Abschnitt 4.2 aufgestellten Merkmale hin untersucht.

4.1 Identifizierungsprozedur

Kitchenham (2004) beschreibt, dass wenn eine Recherche nicht genau und fair ist, besäße sie keinen wissenschaftlichen Wert. Aus diesem Grund solle eine genau definierte Suchstrategie sowie eine Methode zur Verwaltung der Literatur und Fortschrittserfassung erstellt werden.

4.1.1 Suchstrategie

Als Grundmenge der Untersuchung sollten alle kontrollierten Experimente aus Artikeln der Zeitschriften

- „Transactions on Software Engineering“ (TSE)
- „Empirical Software Engineering“ (EMSE)

sowie aus den Artikeln der Konferenz

- „International Conference on Software Engineering“ (ICSE)

im Zeitraum von Januar 2003 bis Dezember 2009 identifiziert werden. Diese waren aus dem internen Netz der Universität im Volltext verfügbar. Ausgenommen

von der Suche wurden "Editorial", "In this Issue", "Announcements", "Short Notes", "Responses", "Comments", "Guest Editor's Introduction", "State of Journal Address", "Call for Papers", "Reviewer's List", "Erratum", "Annual Index", "Keynote Talks", "Experience Reports", "Research Demonstrations", "Demonstration Sessions", "Workshop Sessions", "Tutorial Sessions", "Panel Sessions", "Education & Training Track", "Doctoral Symposium", sofern diese vorhanden waren[1]. Außerdem ausgenommen wurde eine Ausgabe der Zeitschrift TSE ("Issue No. 6 - November/December 2009"), da diese zum Zeitpunkt der Untersuchung noch nicht verfügbar war.
Alle anderen Artikel wurden nach der in Abbildung 4.1 skizzierten Methode (Seite 29) untersucht. Wurde im Titel oder in den Schlagwörtern ("Keywords", "General Terms", "Index Terms") direkt erwähnt, dass es sich um ein randomisiertes oder Quasi-Experiment (bzw. allgemein kontrolliertes Experiment) handelt, wurde es in die Menge der Experimente M aufgenommen (Zeilen 1–4 und 5–8). War dies nicht der Fall, es traten aber potentielle Wörter in Titel oder Schlagwörtern auf, so wurde die Zusammenfassung (*abstract*) gelesen und bei relevanten Begriffen (SetOfPotentialKeywords) der Artikel in die Menge M aufgenommen. Relevante Begriffe sind alle mit Experimenten und statistischen Verfahren verbundenen Begriffe, die auch in Kapitel 2 erwähnt wurden. Dazu zählen u.A. "controlled experiment", "randomized experiment", "quasi experiment", "groups", "subjects", "tasks", "treatments", "hypothesis" oder direkt erwähnte statistische Auswertungsverfahren wie t-Test oder ANOVA. Bei weiterhin bestehenden Zweifeln wurde der Artikel mit aufgenommen. Dies war z.B. bei unklaren Formulierungen wie "We conducted an empirical study…" oder "Experimentation shows that our approach is promising" der Fall. Artikel, die so fälschlicherweise als Experiment identifiziert wurden, sind bei der in Abschnitt 4.2 beschriebenen Analyse wieder entfernt worden. Außerdem wurde bei Titeln wie

- "Evaluate/Evaluating … with Respect to …"
- "Are … more optimistic than …"
- "Influence of … on …"
- "Is … more … than/better than … regarding/in context of …"
- "… in the Presence of …"
- "Effect/Impact of … on …"
- "Do … outperform …"

[1] Die Artikel der Konferenz ICSE sind in einigen Jahrgängen (2004,2007,2009) nicht in Kategorien aufgeteilt.

```
M = ∅ ▷ set of discovered papers

for each Paper
    do REVIEW METHOD(Paper, Paper.Title, Paper.Abstract, Paper.Keywords)
```

```
REVIEW METHOD(Paper, Title, Abstract, Keywords)

    ▷ if title states explicitly, add and return

    for each t in {"controlled experiment",
                   "quasi experiment",
                   "randomized experiment"}
        do if SUBSTRING(t, Title)
              then M = M ∪ Paper
           return

    ▷ if Keywords/General Terms/Index Terms state explicitly add and return

    for each t in {"controlled experiment",
                   "quasi experiment",
                   "randomized experiment"}
        do if SUBSTRING(t, Keywords)
              then M = M ∪ Paper
           return

    ▷ if title or keywords contain certain words, read abstract and decide

    for each t in {"experiment", "study", "empirical"}
        do if [SUBSTRING(t, Title) || SUBSTRING(t, Keywords)]
              then for each s in SetOfPotentialKeywords
                       do if SUBSTRING(s, Abstract)
                             then M = M ∪ Paper
```

Abbildung 4.1: Identifizierungsprozedur

- "Compare/Comparing ... and ..."

in jedem Fall die Zusammenfassung gelesen und wie in Zeile 11–13 behandelt. Es wurden sowohl randomisierte als auch Quasi-Experimente mit aufgenommen, da beide die gewonnenen Daten mit Hilfe statistischer Auswertungsverfahren untersuchen.

4.1.2 Literaturverwaltung

Jeder identifizierte Artikel wurde sofort (1) als Referenz in einer BIBTEX-Bibliothek und (2) in Textform (PDF) gespeichert. Für Referenzschlüssel und Dateinamen wurde folgende Konvention getroffen:

```
<Zeitschrift><Jahr>-<Seitenzahl>
```

`<Zeitschrift>` bezeichnet die Abkürzung der Zeitschrift bzw. der Konferenz, also TSE, EMSE oder ICSE, `<Jahr>` die letzten beiden Stellen des Jahres und `<Seitenzahl>` die Startseite des Artikels.
`TSE09-1000` würde also den Artikel der Zeitschrift „Transactions on Software Engineering" (TSE) aus dem Jahre 2009 auf Seite 1000 und folgende beschreiben. Der Artikel würde unter dem Namen `TSE09-1000.pdf` gespeichert werden.

4.1.3 Fortschrittserfassung

Zur Fortschrittserfassung diente eine Checkliste, in der für beide Zeitschriften jede Ausgabe jeden Jahres und für die Konferenz jedes Jahr als erledigt markiert werden konnte. Somit wurde auch nach Unterbrechungen eine Durchsicht ohne Lücken sichergestellt.

4.2 Untersuchte Merkmale

Die durch die in Abschnitt 4.1 beschriebene Prozedur identifizierten Experimente sollten auf bestimmte Merkmale hin untersucht werden. Zur Aufstellung der relevanten Merkmale wurden Empfehlungen zum Berichten von Ergebnissen in empirischen Studien bzw. kontrollierten Experimenten von Kitchenham et al. (2002) und Jedlitschka and Pfahl (2005) und Empfehlungen zum Berichten von Ergebnissen statistischer Analysen von Fukuda and Ohashi (1997), Wilkinson (1999) und Gliner et al. (2002) in Betracht gezogen. Als Vorlage für relevante Merkmale von Trennschärfen und Effektgrößen dienten die Artikel von Dybå et al. (2006) und Kampenes et al. (2007).
Für die Erfassung der Merkmale wurde eine Tabelle für relevante Merkmale von

Artikeln und eine Tabelle für Merkmale von abhängigen Variablen angelegt. Wurden in einem Artikel mehrere Experimente beschrieben, so wurden Subartikel eingeführt (und mit a,b,c usw. gekennzeichnet). Die Daten wurden als CSV-Daten exportiert und mit der Statistiksoftware R[2] analysiert. Die folgende Beschreibung der Variablen ist gegliedert in die drei Bereiche, die in Abschnitt 1.1 beschrieben wurden. Eine Auflistung der Variablen und deren Datentypen ist in Anhang .1 gegeben.

4.2.1 Beschreibende Statistiken und grafische Darstellungen

Zur Erfassung beschreibender Statistiken wurde für jede abhängige Variable festgehalten, ob Maße für zentrale Tendenzen (Mittelwert, Median), Spreizung der Daten (Standardabweichung) sowie Quartile, Minimum und Maximum numerisch angegeben wurden. Außerdem wurde festgehalten, ob tabellarische Daten und Boxplots berichtet wurden.

4.2.2 Ergebnisdarstellung benutzter statistischer Methoden

Für jede abhängige Variable wurden die durchgeführten statistischen Tests identifiziert. Dabei wurden für jeden Test folgende Aspekte dokumentiert:

- Name des Tests
- Einhaltung der Annahmen bei parametrischen Tests
- p-Wert des Tests[3]
- Signifikanzniveau α
- Ergebnis ist signifikant bzw. nicht signifikant

Für Regressionsanalysen und Korrelationsanalysen wurde außerdem die Angabe und der Wert der Regressionsparameter und Korrelationskoeffizienten protokolliert. Weiterhin wurde festgehalten, ob Konfidenzintervalle angegeben wurden. Die Experimente wurden ebenfalls auf die Präsenz von Trennschärfenanalysen hin untersucht. Es sollte ermittelt werden, ob eine Analyse durchgeführt wurde und ob dies im Vorfeld des Experiments (a priori) oder hinterher (post hoc) geschah. Für Effektgrößen wurde protokolliert, ob standardisierte, nichtstandardisierte oder beide Typen angegeben wurden. Außerdem wurde ermittelt, ob die Autoren die verwendete Software zur Berechnung der Ergebnisse und dessen Version angeben.

[2] http://www.r-project.org

[3] Bei der Varianzanalyse und Regressionsanalyse wurde dies für jeden Faktor und deren Interaktion bzw. für jeden Parameter festgehalten.

4.2.3 Präsenz und Zustand von Rohdaten

Für jedes Experiment wurde erfasst, ob Rohdaten im Artikel verfügbar sind. Waren keine Rohdaten angegeben, wurden die Autoren des Artikels nacheinander per e-Mail kontaktiert. Die Anfrage beinhaltete eine kurze Beschreibung des Anliegens mit der Bitte die Rohdaten der Experimente zur Verfügung zu stellen. Antwortete ein Autor nach wenigen Tagen nicht, so wurde der nächste kontaktiert. Es wurden bis zu drei Autoren pro Artikel angeschrieben. Antwortete keiner, so wurde der Artikel mit „unbeantwortet“ markiert. Aus erhaltenen Antworten konnten verschiedene Gründe für das Ausbleiben der Daten abgelesen werden. Außerdem konnten zugesendete Rohdaten auf Format, Dokumentation und Zustand in Bezug auf Maschinenlesbarkeit hin untersucht wurden.
Zur Fortschrittserfassung wurde eine Tabelle benutzt. Dort wurde für jeden Artikel das Anfragedatum und die Antwort vom Erst-, Zweit- und Drittautor (sofern vorhanden) festgehalten, sowie der Gesamtstatus pro Artikel.

Kapitel 5

Ergebnisse

In diesem Kapitel werden die aus der Untersuchung erhaltenen Ergebnisse präsentiert. Abschnitt 5.1 stellt die identifizierten Experimente, die kontrollierte Experimente beschreiben, vor. In Abschnitt 5.2 werden die Ergebnisse der Untersuchung der identifizierten Experimente gegliedert nach den in Abschnitt 1.1 genannten Aspekten dargestellt.

5.1 Identifizierte Artikel

Die Anzahl der durchsuchten Artikel der verschiedenen Zeitschriften bzw. der Konferenz ist in Tabelle 5.1 (Seite 34) dargestellt. Untersucht wurden insgesamt 907 Artikel, davon 144 der Zeitschrift EMSE, 428 der Zeitschrift TSE und 335 der Konferenz ICSE. Außerdem ist die Anzahl jeweils seperat für die Jahre 2003–2009 aufgelistet. Die Anzahl der Artikel in den Zeitschriften konnte über die Webseiten von EMSE[1] und TSE[2] ermittelt werden. Bzgl. der Konferenz ICSE wurde die Anzahl der „Technical Papers“ aus Clarke et al. (2003), Finkelstein et al. (2004), Griswold and Nuseibeh (2005), Osterweil et al. (2006, Message from the Program Chairs), Emmerich and Rothermel (2007), Schäfer et al. (2008, Foreword) und Fickas et al. (2009) ermittelt.

Die Anzahl der 907 Artikel, die kontrollierte Experimente beschreiben, ist in Tabelle 5.2 (Seite 34) und Abbildung 5.1 (Seite 35) dargestellt. Tabelle 5.2 beinhaltet die Anzahl der Artikel mit kontrollierten Experimenten pro Jahr und Zeitschrift bzw. Konferenz sowie die Gesamtanzahl als absoluten und relativen Wert. Die Werte in den Klammern geben die absolute und relative Gesamtanzahl für die Jahre 2006–2009 an, die vor den Klammern die der Jahre 2003–2009. Die genauere

[1] http://www.springerlink.com/content/100262/
[2] http://www.computer.org/portal/web/csdl/transactions/tse#1

	2003	2004	2005	2006	2007	2008	2009	Gesamt
EMSE	15	15	19	21	24	26	24	144 (95)
ICSE	42	58	44	36	49	56	50	335 (191)
TSE	85	72	67	56	52	52	44	428 (204)
Gesamt	142	145	130	113	125	134	118	**907 (490)**

Tabelle 5.1: Anzahl untersuchter Artikel

	2003*	2004*	2005*	2006	2007	2008	2009	Gesamt	%
EMSE	1	3	5	7	5	4	6	31 (22)	22 (23)
ICSE	3	3	4	2	3	1	0	16 (6)	5 (3)
TSE	6	2	4	2	1	4	1	20 (8)	5 (4)
Gesamt	10	8	13	11	9	9	7	**67 (36)**	**7 (7)**

Tabelle 5.2: Anzahl identifizierter Artikel mit kontrollierten Experimenten

Untersuchung der Artikel auf die in Abschnitt 4.2 beschriebenen Merkmale wurde für die Jahre 2006-2009 durchgeführt, die Identifizierungsprozedur zusätzlich für die Jahre 2003-2005.
Die Zeitschrift EMSE stellt sowohl insgesamt als auch bis auf eine Ausnahme (2003) in den einzelnen Jahren die meisten Artikel mit kontrollierten Experimenten. ICSE und TSE unterscheiden sich untereinander nur gering. Relativ gesehen berichten 22% der in EMSE herausgegebenen Artikel über kontrollierte Experimente, ICSE und TSE jeweils 5%.
Ein Vergleich mit einer älteren Arbeit von Sjøberg et al. (2005) (siehe auch Kapitel 3) ist in Tabelle 5.3 dargestellt. Sie gibt die Mittelwerte der Artikel mit kontrollierten Experimenten pro Jahr jeweils aus der Arbeit von Sjøberg et al. (2005) und dieser Arbeit (aufgerundet) an. Die Zeitschrift EMSE berichtet nach den Ergebnissen dieser Arbeit durchschnittlich zwei Experimente mehr im Jahr, als nach den Ergebnissen von Sjøberg et al. (2005). TSE und ICSE weisen bei beiden Arbeiten gleiche Durchschnittswerte auf.

	Sjøberg et al. (2005) (Tabelle 19)	diese Arbeit
EMSE	4	6
ICSE	2	2
TSE	2	2

Tabelle 5.3: Mittelwert der identifizierten Artikel pro Jahr (aufgerundet)

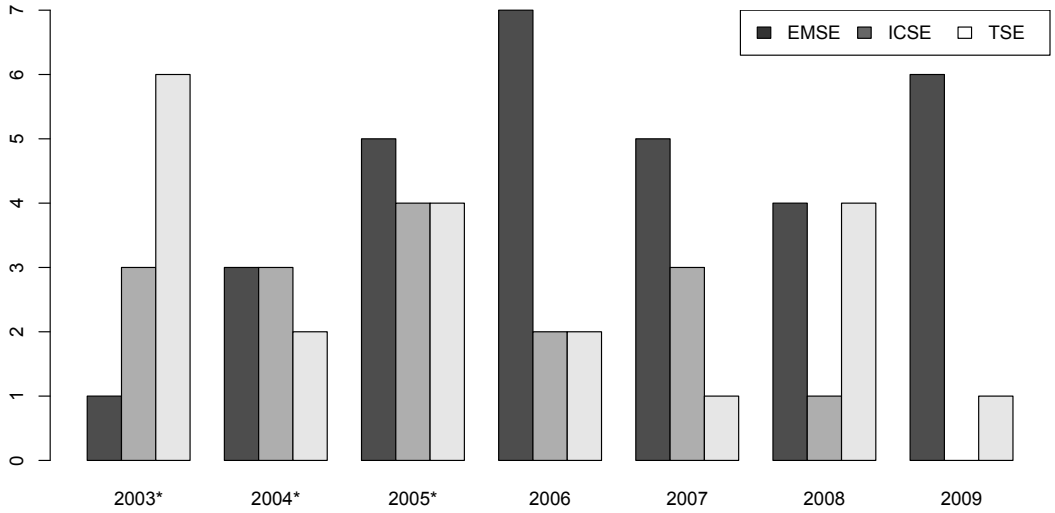


Abbildung 5.1: Anzahl identifizierter Artikel mit kontrollierten Experimenten

5.2 Untersuchte Merkmale

Die genauere Untersuchung der identifizierten Artikel, die kontrollierte Experimente berichten, erfolgte für die Jahre 2006–2009. Die Anzahl der berichteten Experimente ist in Tabelle 5.4 (Seite 36) und Abbildung 5.2 (Seite 36) dargestellt. In den 36 Artikeln werden insgesamt 53 kontrollierte Experimente beschrieben. Bis auf eine Ausnahme (2008) werden die meisten in Artikeln der Zeitschrift EMSE berichtet.
33 (62%) der Experimente sind randomisierte und 7 (13%) Quasi-Experimente. Für 13 (25%) konnte nicht ermittelt werden, ob eine randomisierte Zuweisung der Subjekte auf die Gruppen stattfand oder nicht. Zwölf der Experimente wurden als Replikation beschrieben. Demnach wurden also 41 verschiedenartige Experimente berichtet.
In den 53 Experimenten wurden insgesamt 153 abhängige Variablen gemessen, davon 99 (65%) in EMSE, 18 (12%) in ICSE und 36 (23%) in TSE. Dies sind durchschnittlich drei Variablen pro Experiment.
In den meisten Fällen (136 bzw. 89%) wurden Variablen auf rationalen Skalen gemessen. Weitere waren ordinale (11 bzw. 7%), kategorische (4 bzw. 3%) und Differenzskalen (2 bzw. 1%).

5.2.1 Beschreibende Statistiken und grafische Darstellungen

Die Anzahl an numerisch berichteten statistischen Größen und Grafiken ist in Tabelle 5.5 (Seite 37) dargestellt. Die Tabelle zeigt, wie oft die jeweilige Größe bzw.

	2006	2007	2008	2009	Gesamt
EMSE	7	7	4	12	30
ICSE	3	5	3	0	11
TSE	4	1	5	2	12
Gesamt	14	13	12	14	**53**

Tabelle 5.4: Anzahl an berichteten kontrollierten Experimenten

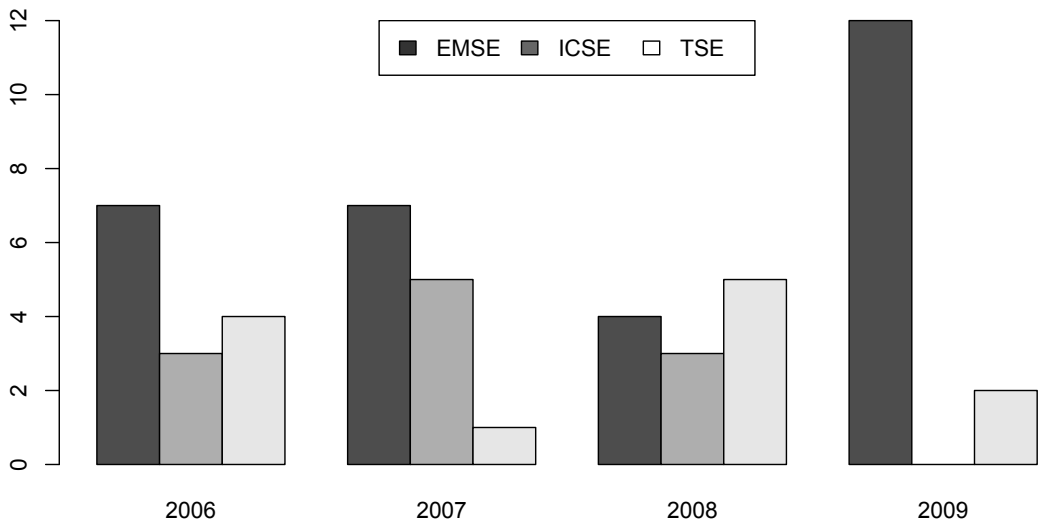


Abbildung 5.2: Anzahl an berichteten kontrollierten Experimenten

	Berichtet	Möglich	%
Mittelwert	98	138	71%
Standardabweichung	77	138	56%
Median	60	149	40%
Quartile (Q1 und Q3)	7	149	5%
Extrema (Min und Max)	30	149	20%
Mittelwert und Standardabweichung	76	138	55%
Fünf-Punkte Statistik	5	149	3%
Tabellen	108	153	71%
Boxplots	72	149	48%

Tabelle 5.5: Angabe von statistischen Größen und Grafiken

Grafik berichtet wurde und wie oft sie hätte berichtet werden können[3], sowie den relativen Anteil. Am Häufigsten wurden Mittelwerte berichtet. Für etwas mehr als die Hälfte der Variablen wurden Standardabweichungen angegeben. 76 mal (55%) sind sowohl Mittelwert als auch Standardabweichung berichtet worden. Extremwerte (Minimum und Maximum) waren für 30 der Variablen (20%) vorhanden. Quartile wurden in 7 Fällen (5%) numerisch bzw. in 77 Fällen (52%) entweder numerisch oder innerhalb von Boxplots berichtet. Eine Fünf-Punkte Statistik (*five-point summary*) in numerischer Form beinhaltete nur ein Artikel für fünf Variablen (3%).
72 mal (48%) wurden statistische Größen mit Hilfe von Boxplots dargestellt, 108 mal (71%) in Tabellen. Andere berichteten sie im Textfluss oder gaben sie gar nicht an.
Für insgesamt sieben Variablen auf nominellen oder ordinalen Skalen aus zwei Artikeln wurden Mittelwerte oder Standardabweichungen angegeben.

5.2.2 Ergebnisdarstellung benutzter statistischer Methoden

Eine Übersicht über die Anzahl durchgeführter Tests gibt Tabelle 5.6 sowie Abbildung 5.3. In den 53 Experimenten wurden insgesamt 519 statistische Tests durchgeführt. Die Tests, die am Häufigsten durchgeführten wurden, sind der Mann-Whitney Test (137), der t-Test (81), die Varianzanalyse (71) und die Korrelationsanalyse (75). Am Wenigsten verwendet wurden der Permutationstest (1), die Kovarianzanalyse (3), der Chi-Quadrat-Test (10) und die Regressionsanalyse (13).
Die Anzahl durchgeführter Tests pro Experiment zeigt Abbildung 5.4. Es wur-

[3]Bestimmte Statistiken anzugeben ist nur auf bestimmten Skalen sinnvoll (siehe Abschnitt 2.2 unter Skalen).

Name	Anzahl
t-Test	81
ANOVA	71
ANCOVA	3
Mann-Whitney Test	137
Wilcoxon-Signed-Rank Test	34
Kruskal-Wallis Test	26
Chi-Square Test	10
Fisher's Exact Test	18
McNemar's Test	50
Permutation Test	1
Regression	13
Correlation	75
Gesamt	**519**

Tabelle 5.6: Anzahl durchgeführter statistischer Tests

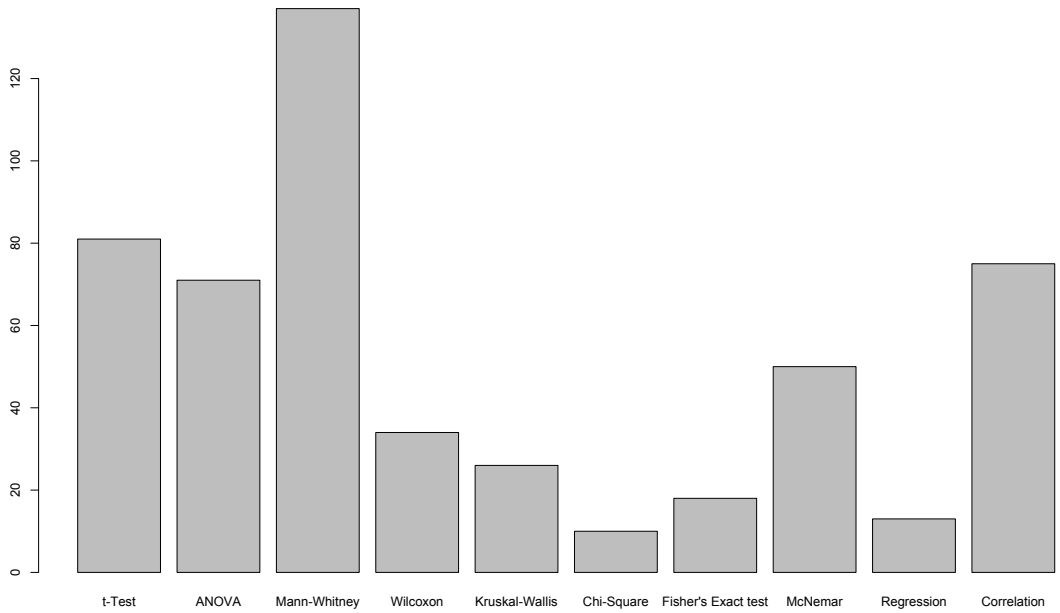


Abbildung 5.3: Anzahl durchgeführter statistischer Tests

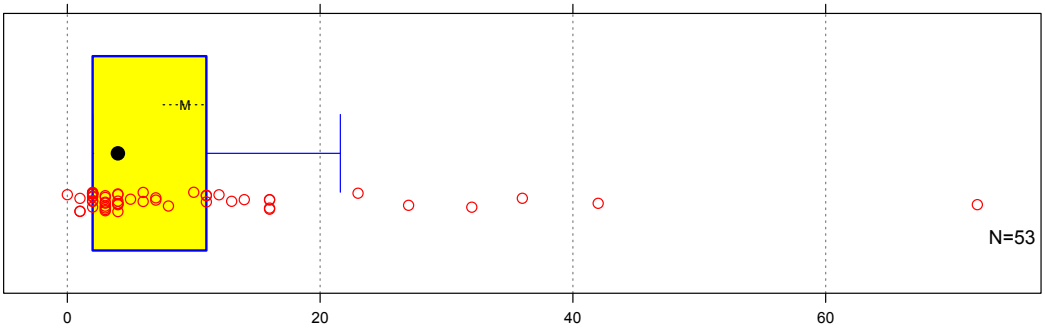


Abbildung 5.4: Anzahl durchgeführter Tests pro Experiment

den aufgerundet durchschnittlich 4 (Median) bzw. 10 (Mittelwert) Tests pro Experiment durchgeführt. In einem Experiment wurde kein Test durchgeführt. Maximal wurden 72 Tests in einem Experiment durchgeführt.

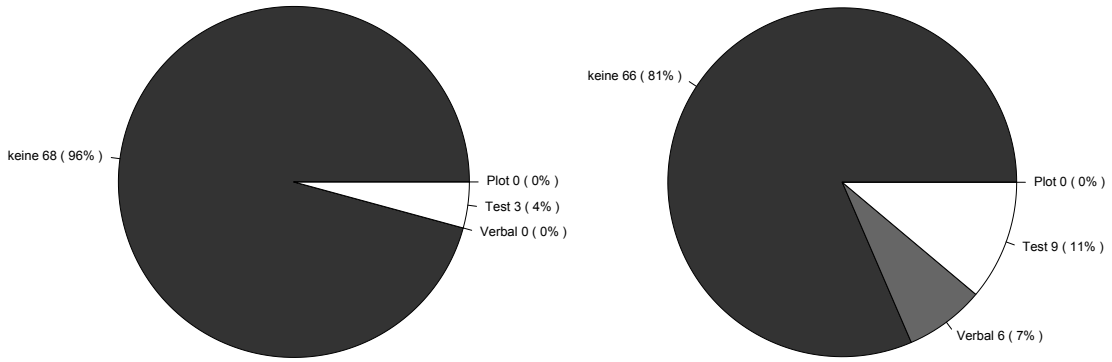
Konfidenzintervalle für die Mittelwertdifferenz der gemessenen abhängigen Variablen wurden in sechs von 36 Artikeln (17%) angegeben. Für etwas mehr als der Hälfte der möglichen Fälle (402 von 695[4] bzw. 58%) wurden p-Werte berichtet. Der Rest berichtete lediglich ob die Ergebnisse signifikant oder nicht signifikant sind. Das Signifikanzniveau des Tests wurde in 80% der Fälle angegeben. Das am Häufigsten benutzte Signifikanzniveau ist $\alpha = 0.05$ (378 mal). Die Parameter der Regressionsgleichungen wurden in 85% der Fälle mit angegeben, die Koeffizienten der Korrelationen in 83%.

Für die Berechnung der Ergebnisse gaben 42% der Experimente den Namen der benutzen Softwaren an. In 14 Fällen wurde SPSS benutzt. Weitere waren SPlus, Statview und Excel.

Die Problematisierung der Annahmen von parametrischen Tests (siehe Abschnitt 2.2.3) zeigt Abbildung 5.5a für die Varianzanalyse bzw. Abbildung 5.5b für den t-Test. Für drei von insgesamt 71 durchgeführten Varianzanalysen wurden Tests durchgeführt, um die Annahmen zu überprüfen. Bzgl. des t-Tests wurde neun Mal ein Test auf Normalität durchgeführt. Weitere sechs Mal wurde eine verbale Begründung angegeben. In 66 Fällen sind die Annahmen nicht problematisiert worden. Eine grafische Darstellung wurde sowohl für den Test der Annahmen der Varianzanalyse als auch den t-Test nicht verwendet.

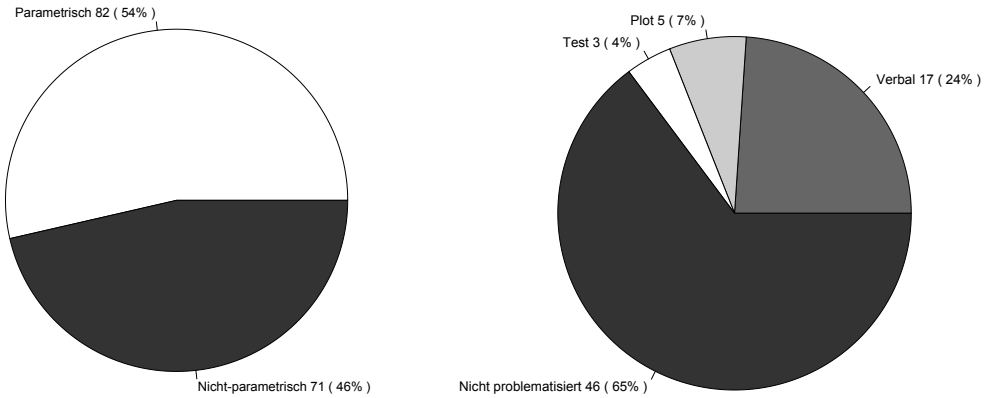
Parametrische Tests wurden für etwa die Hälfte der abhängigen Variablen durchgeführt. Für 71 Variablen (46%) wurden ausschließlich nicht-parametrische Signifikanztests verwendet (siehe Abbildung 5.6a). In einem Drittel dieser Fälle wurde die Wahl begründet. Abbildung 5.6b zeigt, dass überwiegend verbal (24%), teil-

[4]Es wurden insgesamt 519 Tests durchgeführt. Varianzanalysen bzw. Regressionsanalysen geben für jeden Faktor und für dessen Interaktion bzw. für jeden Parameter p-Werte an.



(a) Varianzanalyse (b) t-Test

Abbildung 5.5: Problematisierung der Annahmen bei parametrischen Tests



(a) Verteilung (b) Begründungen

Abbildung 5.6: Begründung für die Verwendung nicht-parametrischer Verfahren

	Dybå et al. (2006)	diese Arbeit
a priori	3 von 78 (4%)	4 von 36 (11%)
post hoc	6 von 78 (8%)	2 von 36 (6%)
Gesamt	9 von 78 (12%)	6 von 36 (17%)

Tabelle 5.7: Duchführung von Trennschärfeanalysen (nach Artikel)

	Kampenes et al. (2007)	diese Arbeit
Standardisiert	8 von 92 (9%)	7 von 53 (13%)
Nicht-Standardisiert	17 von 92 (18%)	5 von 53 (9%)
Beides	2 von 92 (2%)	9 von 53 (17%)
Keine Effektgröße	65 von 92 (71%)	32 von 53 (61%)

Tabelle 5.8: Angabe von Effektgrößen (nach Experimenten)

weise aber auch statistische Tests durchgeführt (4%) oder Grafiken angegeben wurden (7%).

Die Anzahl der Artikel, in denen Trennschärfeanalysen berichtet wurden, ist in Tabelle 5.7 aufgeführt. Zum Vergleich sind außerdem die Ergebnisse der Analyse von Dybå et al. (2006) gegenübergestellt. Insgesamt wurde in sechs von 36 Artikeln (17%) Trennschärfe diskutiert (bei Dybå et al. (2006) waren es neun von 79 bzw. 12%). Vier der Artikel berichteten eine im Vorfeld (a priori) durchgeführte Analyse. Die anderen zwei wurden im Nachhinein (post hoc) berechnet.

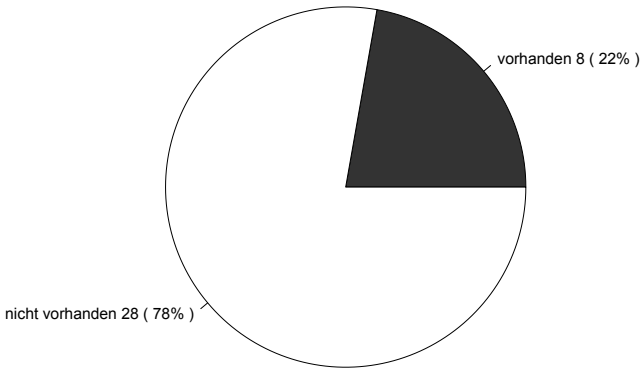
Tabelle 5.8 zeigt die Anzahl der Experimente, bei denen Effektgrößen angegeben wurden. Die Ergebnisse der Analyse von Kampenes et al. (2007) sind ebenfalls aufgeführt. Bei ca. zwei Drittel der Experimente wurden keine Effektgrößen berechnet (32 von 53 bzw. 61%). In 17% der Fälle wurden sowohl standardisierte als auch nicht standardisierte Effektgrößen berichtet. Sieben mal wurden ausschließlich standardisierte, fünf mal ausschließlich nichtstandardisierte Effektgrößen angegeben. Cohen's d und Hedge's g waren die häufigsten Vertreter von standardisierten Effektgrößen der d-Familie, der Pearson Korrelationskoeffizient der der r-Familie. Nichtstandardisierte Effektgrößen wurden fast ausnahmslos prozentual angegeben.

5.2.3 Präsenz und Zustand von Rohdaten

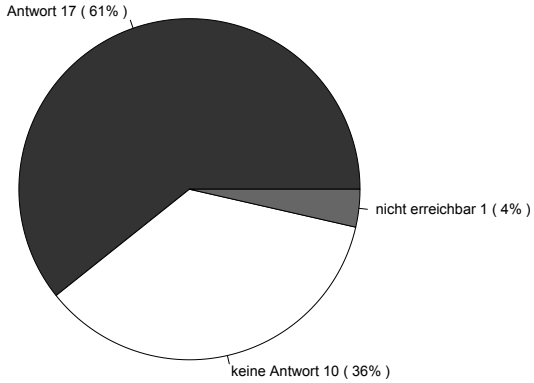
Die Verfügbarkeit von Rohdaten der Experimente in den Artikeln ist in Abbildung 5.7a dargestellt. In acht von 36 Artikeln sind die Rohdaten verfügbar. Das Ergebnis der Anfrage für die 28 Artikel, in denen keine Rohdaten verfügbar waren, zeigt Abbildung 5.7b. Für 17 der Artikel wurden die Anfragen beantwortet,

für zehn nicht. Bei einem Artikel konnte keine Anfrage gestellt werden, da keine gültige e-Mail Adresse zur Verfügung stand. Abbildung 5.7c zeigt die Anzahl der positiven und negativen Antworten. Von 17 Antworten waren 7 positiv (Rohdaten gesendet). Die anderen zehn enthielten Vorwände, weshalb die Rohdaten nicht zur Verfügung gestellt werden können. Als Vorwand wurde u.A. genannt, dass die Autoren nicht mehr dort wo sie die Experimente durchgeführt haben arbeiten würden und deshalb keinen Zugriff mehr auf die Daten hätten. Ein weiterer Grund wäre die andauernde Datenanalyse und Nutzung der Daten in künftigen Publikationen. Andere gaben auch an, sie stünden unter der Aufsicht bestimmter Gremien, die ihnen verbieten die Daten herauszugeben. Teilweise würden die Daten nicht verfügbar gemacht werden können, weil im Vorfeld Vertraulichkeitsvereinbarungen mit den Firmen getroffen wurden. Die Daten müssten dann erst anonymisiert und teilweise ins englische übersetzt werden. Dafür fehle die Zeit.

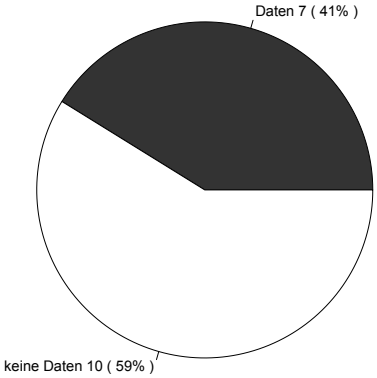
Als Formate für Rohdaten wurden Excel-Tabellen (4), CSV-Dateien (1) und Word-Dokumente (1) verwendet. Die CSV-Dateien sowie zwei der Excel-Tabellen sind maschinenlesbar, d.h. können ohne Änderungen in die Statistiksoftware R importiert und verarbeitet werden. Dokumentiert wurde nur ein Datensatz. Bei zwei weiteren wurden die Spalten konform zu den Variablennamen im veröffentlichten Artikel benannt.



(a) Verfügbarkeit



(b) Ergebnis der Anfrage



(c) Erhalt von Daten

Abbildung 5.7: Präsenz und Verfügbarkeit von Rohdaten

Kapitel 6

Diskussion

Dieses Kapitel diskutiert die in Kapitel 5 vorgestellten Ergebnisse. Die Unterteilung der Abschnitte ist konform zu Kapitel 4 und 5 in denen die Methodik der Untersuchung und dessen Ergebnisse beschrieben wurden. Abschnitt 6.1 diskutiert die Ergebnisse der Artikeldurchsicht. In Abschnitt 6.2 werden die Ergebnisse der untersuchten Merkmale aus den identifizierten Experimenten in den bereits genannten Aspekten bewertet.

6.1 Identifizierte Artikel

Die Aussage von Dybå et al. (2006) und Dieste et al. (2009), der Begriff "experiment" würde im Bereich des Software Engineerings bzw. der Informatik uneinheitlich benutzt werden, konnte (für die untersuchte Menge von Artikeln) bestätigt werden. Die Beobachtung von Dieste et al. (2009), dass viele Artikel fälschlicherweise identifiziert und hinterher wieder verworfen werden, trifft auch in dieser Untersuchung zu. Der Grund dafür könnte darin liegen, dass viele Empirische Studien (nicht nur kontrollierte Experimente) den Begriff "experiment" im Text oder den Schlüsselwörtern verwenden. (vgl. Dybå et al., 2006)
Um eine gute Abdeckung (*recall*) zu erhalten, sollten die Artikel auf allgemeinere Begriffe (zusätzlich zu "experiment", "empirical study" und "evaluation") hin untersucht werden. Dieste et al. (2009) konnten so die Anzahl an identifizierten Artikeln mit kontrollierten Experimenten auf Basis der Menge von Sjøberg et al. (2005) von 75% auf 92% steigern. Dies bestätigt die Wahl auch bei dieser Untersuchung die Titel und Zusammenfassungen auf mehrere Schlüsselwörter hin zu untersuchen. Die Identifizierung erfolgte für die Jahre 2003–2009, schließt sich also an die Untersuchung von Sjøberg et al. (2005), die Artikel von 1993–2002 untersuchte, an.

6.2 Untersuchte Merkmale

Wie auch Kitchenham et al. (2008) basierend auf den Daten von Sjøberg et al. (2005) feststellten, zeigt auch diese Untersuchung, dass nur wenig kontrollierte Experimente berichtet werden (Tabelle 5.1 und 5.2). Die geringe Anzahl könnte mit dem zur Durchführung verbundenem Aufwand und dessen Kosten zusammenhängen. Der im Vergleich zu ICSE und TSE hohe Anteil an Experimenten in der Zeitschrift EMSE ist dadurch erklärbar, dass es sich um eine Zeitschrift handelt, die speziell den Bereich "Empirical Software Engineering" behandelt.
Das Maximum an Experimenten (12) in der Zeitschrift EMSE im Jahr 2009 (Abbildung 5.4) ist damit zu erklären, dass ein Artikel fünf Experimente berichtete. Der in Tabelle 5.3 gezogene Vergleich der durchschnittlichen Anzahl an berichteten Experimenten pro Jahr zeigt lediglich für die Zeitschrift EMSE einen höheren Anteil in dieser Arbeit. Dies könnte daran liegen, dass Sjøberg et al. (2005) und Kampenes et al. (2007) eine etwas eingeschränktere Definition von kontrollierten Experimenten benutzen als die in dieser Untersuchung verwendete von Shadish et al. (2002). Sie definierten Versuchsobjekte (*experimental units*) nur als Personen.
Die hohe Anzahl an Variablen, die auf rationalen Skalen gemessen wurden, ergibt sich dadurch, dass hauptsächlich die benötigte Arbeitszeit zur Lösung der Aufgaben bzw. die Anzahl an Defekten gemessen wurden. Außerdem wurden weitere Anzahlen (*counts*) gemessen. Fragebögen wurden häufig als Anzahl richtiger Antworten erfasst. Dies erklärt auch die geringe Anzahl an Variablen die auf nominalen und ordinalen Skalen gemessen wurden.
Die Erfassung einzelner Aspekte, die im Folgenden diskutiert werden, erwies sich als sehr aufwändig, da keine einheitliche Gliederung zum Berichten von empirischen Studien (bzw. speziell kontrollierten Experimenten) erkennbar ist. Somit ist es schwierig relevante Information zu finden. Dies stellten auch Jedlitschka and Pfahl (2005) fest.
Die für die weitere Untersuchung vorliegende Menge an kontrollierten Experimenten kann als repräsentativ angesehen werden, da sie nach einer festen Vorgehensweise bestimmt wurde. Es wurden verschiedene Quellen zur Identifikation gewählt. Dabei handelt es sich sowohl um spezielle (empirische) Quellen als auch allgemeine Quellen des Software Engineerings.

6.2.1 Beschreibende Statistiken und grafische Darstellungen

Kitchenham et al. (2002) und Jedlitschka and Pfahl (2005) geben in ihren Richtlinien zum Berichten von Empirischen Studien bzw. Experimenten an, dass angemessene beschreibende Statistiken wie Lokation (Mittelwert, Median) und Spreizung (Standardabweichung, Varianz) angegeben werden sollen. Dies dient dazu eine Übersicht über die gewonnenen Daten zu erhalten. Mittelwerte und Stan-

dardabweichungen wurden allerdings lediglich in etwas mehr als der Hälfte der Fälle angegeben. Der gegen Ausreisser robustere Median ist nicht einmal für die Hälfte der Variablen berichtet worden. Extremwerte erreichen mit 20% ein noch schlechteres Resultat.
Die grafische Präsentation von beschreibenden Statistiken wird überwiegend in Tabellen berichtet. Sobald aber mehrere Experimente mit mehreren Gruppen und Variablen vorkommen werden auffällige Werte in Tabellen immer unscheinbarer. Laut Wilkinson (1999) und Kitchenham et al. (2002) fallen Grafiken ins Auge und sind für den Leser besser zu verstehen. Dies wird allerdings nur für die Hälfte der gemessenen Variablen auch befolgt, indem Boxplots berichtet werden.
Wenig beschreibende Statistiken anzugeben erschwert es dem Leser einen Überblick über die aus dem Experiment gewonnenen Daten zu bekommen. Führt man ein Experiment durch und erstellt z.B. Boxplots der Messdaten, so kann dort eventuell schon eine Tendenz sichtbar werden. Außerdem lassen sich extreme Werte (oder auch Ausreißer) identifizieren.
Angesichts dieser Ergebnisse entsteht die Vermutung, dass der Fokus eher in der Auswertung der Daten mit Hilfe der schließenden Statistik liegt. Eventuell liegt der Grund für die geringe Berichterstattung von beschreibenden Statistiken vor allem mit Grafiken aber auch an der Seitenbeschränkung der Zeitschriften, in denen die Artikel veröffentlicht werden.

6.2.2 Ergebnisdarstellung benutzter statistischer Methoden

Mit vier (Median) bzw. 10 (Mittelwert) durchgeführten Tests pro Experiment und durchschnittlich drei gemessenen Variablen werden also mehrere Tests pro Variable durchgefürt. Dies liegt daran, dass einige Autoren sowohl parametrische als auch nichtparametrische Tests pro Variable durchführen. Außerdem werden mit Hilfe von Varianzanalysen einzelne Faktoren analysiert. Abbildung 5.4 zeigt, dass bis zu 72 Tests pro Experiment durchgeführt wurden. In einigen Fällen ist dies sinnvoll, in anderen allerdings weniger.
Betrachtet man den Artikel, in dem 72 Tests durchgeführt wurden, so stellt man fest, dass ein Experiment mit zwei Gruppen, fünft gestellten Aufgaben und 6 gemessenen Variablen beschrieben wird. Für jede einzelne Aufgabe wurde jeweils ein parametrischer Test (t-Test) und ein nichtparametrischer Test (Mann-Whitney Test) durchgeführt. Dies begründen die Autoren damit, dass aufgrund von eventuell verletzen Annahmen der parametrischen Tests zusätzlich nichtparametrische Tests durchgeführt wurden. Die große Anzahl an Tests ist also begründet und vertretbar.
Für den Artikel mit der zweithöchsten Anzahl an durchgeführten Tests ist die Entscheidung allerdings weniger vertretbar. Hierbei handelt es sich um ein Ex-

periment mit einer abhängigen Variable und zwei unabhängigen Variablen. Die beiden unabhängigen Variablen besitzen neun bzw. vier veschiedene Werte. Für diverse Kombinationen dieser Werte wurden insgesamt 42 Rangsummentests für die gemessenen Werte der einen abhängigen Variable durchgeführt. Hier stellt sich die Frage, ob nicht eine Varianzanalyse sinnvoller gewesen wäre, bzw. ob die für kontrollierte Experimente eher untypisch große Anzahl an Werten der unabhängigen Variablen nötig ist, oder ob man diese eventuell hätte gruppieren können.
Schließt man eher seltener durchgeführte Tests wie McNemars Test und den Permutationstest sowie Regressions- und Korrelationsanalysen aus, so ergibt sich ein ausgeglichenes Verhältnis von 48% zu 52% zwischen parametrischen Tests (t-Test, Anova, Chi-Quadrat) und nichtparametrischen Tests (Mann-Whitney Test, Wilcoxon test, Kruskal-Wallis Test).
Die große Anzahl an verwendeten nichtparametrischen Verfahren lässt vermuten, dass viele Autoren eher defensiv agieren. Abbildung 5.6a bestätigt diese Vermutung. Für etwas weniger als der Hälfte der Variablen wurden für die Auswertung ausschließlich nichtparametrische Verfahren verwendet. Allerdings bleibt die Vorgehensweise in zwei Dritteln der Fälle unbegründet (siehe Abbildung 5.6b). Ein Drittel begründet ihr Vorgehen größtenteils verbal. Lediglich 11% geben Ergebnisse von statistischen Tests oder Grafiken als Begründung an. Es entsteht der Eindruck, dass größtenteils nicht überprüft wird, ob parametrische Verfahren angewendet werden können. Die Einschränkung der Seitenanzahl in den Zeitschriften könnte aber auch ein Grund sein, weshalb dies so selten berichtet wird. Gegen eine zumindest verbale Begründung spricht allerdings nichts. Die Empfehlung von Fukuda and Ohashi (1997), Wilkinson (1999) und Kitchenham et al. (2002), die Annahmen der verwendeten Tests einzuhalten, wird dem zufolge selten befolgt.
Für die andere Hälfte der Variablen wurden (auch) parametrische Tests durchgeführt. Der Anteil für die die Einhaltung der Annahmen im Artikel problematisiert wird ist für die Varianzanalyse bzw. den t-Test ind Abbildung 5.5 dargestellt. Hierbei wird deutlich, dass dies nur für einen sehr geringen Teil der durchgeführten Tests geschieht. Für diesen gerigen Anteil wurden überwiegend Tests wie Shapiro-Wilk und Kolmogorov-Smirnov verwendet, welche, wie in Abschnitt 2.2.3 beschrieben, nicht das aussagen was von Interesse ist. Ob die Autoren im Vorfeld die Annahmen überprüft, die Ergebnisse aber nicht in den Artikel mit aufgenommen haben, oder ob sie die Verfahren ohne Überprüfung anwenden, kann nicht überprüft werden.
Das Ergebnis der Hypothesentests wird in unterschiedlichem Detail berichtet. Laut Fukuda and Ohashi (1997), Kitchenham et al. (2002) und Jedlitschka and Pfahl (2005) soll sowohl das Signifikanzniveau α als auch der resultierende p-

Wert angegeben werden. Allerdings wird nur für etwas mehr als die Hälfte der Tests ein numerischer p-Wert berichtet. Lediglich anzugeben, das Ergebnis ist signifikant bzw. nicht signifikant ist nicht ausreichend. So z.B. kann nicht unterschieden werden, ob die Nullhypothese auf Grund eines p-Wertes von 0.051 oder 0.85 abgelehnt wurde. Wie in Abschnitt 2.2.4 erwähnt bedeutet ein kleinerer p-Wert zwar keinen stärkeren Zusammenhang, man sollte allerdings bei einem Ergebnis von $p = 0.051$ zu einer vorsichtigeren Schlußfolgerung kommen. Die Angabe des Signifikanzniveaus ist mit 80% erfreulich hoch. Da der von Cohen festgelegte Wert von 0.05 auch hier hauptsächlich verwendet wurde, könnte man bei den restlichen 20% davon ausgehen, dass sie diesen ebenfalls verwendeten.
Die von Wilkinson (1999) und Kitchenham et al. (2002) geforderte und mit insgesamt 17% geringe a priori Berechnung der Trennschärfe ist dadurch erklärbar, dass eine sinnvolle Berechnung im Vorfeld des Experiments nicht möglich ist. Es kann meistens keine genaue Effektgröße berechnet werden, da zu wenig Erfahrung vorhanden ist, um eine Effektgröße zu schätzen bzw. zu wenig ältere Ergebnisse vorliegen, um sie zu berechnen. Dybå et al. (2006) berechnen die Trennschärfe für die von Cohen (1988) aufgestellten Werte für kleine, mittlere und große standardisierte Effektgrößen, denn es ist schwierig im Vorfeld des Experiments zu entscheiden, ob der zu beobachtende Effekt z.B. mit dem bloßen Auge erkennbar ist (vgl. Cohen, 1988, S.26). Die Stichprobengröße ist oft dadurch schon vorgegeben, dass die Versuche im Rahmen einer Vorlesung oder ähnlichem durchgeführt werden, daher wäre eine a priori Berechnung nur nützlich um die Trennschärfe zu berechnen, nicht aber um die benötigte Stichprobengröße zu bestimmen, damit eine genügend große Trennschärfe erreicht wird. Allerdings könnte eine Berechnung im Nachhinein stattfinden, indem eine geschätze Effektgröße anhand der Mittelwertdifferenz berechnet wird. Somit stellt man fest, ob mit diesem Experimententwurf eine Erkennung eines eventuell signifikanten Unterschieds überhaupt wahrscheinlich war. Das Resultat kann dann in die Bewertung der Ergebnisse mit einfließen.
Die u.A. von Wilkinson (1999) und Gliner et al. (2002) aufgestellte Forderung Effektgrößen zu berechnen, wird überwiegend nicht eingehalten. Für ca. zwei Drittel der Experimente (bei Kampenes et al. (2007) fast drei Viertel) werden keine Effektgrößen berichtet. Erfreulich ist, dass von den fast 40% der Experimente, die Effektgrößen angeben, ca. die Hälfte sowohl standardisierte als auch nichtstandardisierte Effektgrößen angeben. Wenn die abhängigen Variablen in anschaulichen Einheiten gemessen wurden, dann empfiehlt es sich beide Typen anzugeben, denn standardisierte Effektgrößen eignen sich besser zum vergleichen während nichtstandardisierte meist anschaulicher sind.
Anhand der Angabe von Effektgrößen kann abgelesen werden, ob die Forschung auch eine praktische Relevanz hat, d.h. ob sich z.B. die Schulung der Mitarbei-

ter zur Verwendeung des getesteten Werkzeugs finaziell lohnt (weil sie dann ihre Aufgaben schneller und eventuell besser erledigen können). Da die Berechnung allerdings auf Mittelwerten basiert, sind die Werte auch anfällig gegen Ausreißer. Die Werte können also sehr stark schwanken.
Besser geeignet für die Entscheidung der praktischen Relevanz sind Konfidenzintervalle. Da sie ein Intervall angeben, in dem mit gewisser Sicherheit ein Mittelwert (oder auch eine Mittelwertdifferenz[1]) liegt. Somit erhält man eine untere und obere Grenze. Bedauerlicherweise sind Konfidenzintervalle in nur 17% der Artikel berichtet worden. Cohen (1994) vermutet, der Grund dafür sei ihre Größe. Ein großes Konfidenzintervall bedeutet außerdem eine geringe Trennschärfe. Die Intervalle lassen sich aber auch nicht mit nichtparametrischen Tests wie den Mann-Whitney Test berechnen, der zu einem Großteil zur Auswertung der untersuchten Artikel verwendet wurde. Die Feststellung von Kampenes et al. (2007), dass der praktische Nutzen nicht diskutiert werden würde, bestätigt sich.

6.2.3 Präsenz und Zustand von Rohdaten

Obwohl Fukuda and Ohashi (1997), Kitchenham et al. (2002) und Jedlitschka and Pfahl (2005) ausdrücklich empfehlen Rohdaten zur Verfügung zu stellen, wird dies nur in einem Viertel der Artikel auch praktiziert. Dies macht weitere Analysen unmöglich. (vgl. Jedlitschka and Pfahl, 2005)
Ein relativ großer Teil der Autoren antwortet auf Anfragen bzgl. der Rohdaten. Leider sind nicht einmal die Hälfte der Antworten positiv (enthalten Rohdaten). Die Vertraulichkeitsvereinbarungen erweisen sich hierbei als das Hauptproblem. Es wäre allerdings möglich die Daten zu anonymisieren, sodass keine Namen von Versuchspersonen, sondern lediglich anonyme Namen festgehalten werden, die die Integrität der Daten nicht verletzen. Verwendet man diese Methode direkt in der Planung und Durchführung des Experiments erfordert dies nicht viel Zeitaufwand. Mehr Zeit hingegen erfordert die Transformation der Daten[2], sodass z.B. keine direkten Werte abgelesen und somit eventuell Rückschlüsse auf die beteiligten Institutionen geschlossen werden können. Dies müsste dann wohl eher außerhalb der regulären Arbeitszeit geschehen, ist daher also eher unwahrscheinlich.
Dass für ein Drittel der Artikel die Autoren überhaupt nicht auf die Anfragen antworten ist bedauerlich. Vielleicht wäre eine auf dem Forschungsgebiet bekannte Institution (sichtbar in der e-Mail Adresse) bei solchen Anfragen erfolgreicher.

[1] Eine Mittelwertdifferenz entspricht z.B. dem Unterschied der benötigten Zeit der Gruppen mit bzw. ohne Hilfe eines Werkzeugs/Methode.

[2] Eine Möglichkeit wäre hier die Messwerte alle mit einem Faktor zu multiplizieren und diesen hinterher zu verwerfen. Somit bleiben Verhältnisse erhalten und es können weiterhin statistische Methoden verwendet werden.

Die Formate der erhaltenen Rohdaten sind mit Tabellen und CSV-Dateien gut gewählt. Allerdings sind nur wenige davon auch maschinenlesbar. Die teilweise aufwändig gestalteten Tabellen sind zwar optisch eventuell besser zu erfassen, für die Weiterverarbeitung jedoch nicht geeignet. Dort sind im Vorfeld teilweise umfangreiche Änderungen notwendig. Dies wird außerdem dadurch erschwert, dass Spalten mit kurzen unverständlichen Namen versehen sind und keine Dokumentationen vorhanden sind.

Kapitel 7

Anschauungsbeispiele

In diesem Kapitel werden Anschauungsbeispiele gegeben, die zeigen sollen, warum die Angabe von bestimmten Aspekten wie grafische Darstellungen oder Konfidenzintervallen sinnvoll ist. Dabei wurden Artikel bevorzugt, bei denen die Rohdaten in maschinell lesbarer Form vorlagen und verständlich waren, da diese ohne größere Konvertierungen genutzt werden konnten. Das Kapitel ist gegliedert nach Artikeln. Da es nicht darum geht einzelne Autoren zu kritisieren sondern Anschauungsbeispiele zu geben, werden die Autoren nicht direkt referenziert sondern die Artikel wie in 4.1.2 beschrieben benannt.

7.1 Artikel: EMSE09-57

Der Artikel EMSE09-57 beschreibt ein kontrolliertes Experiment und eine Replikation. Das Experiment sollte den Nutzen eines bestimmten Werkzeuges zur Identifizierung von Verknüpfungen in Programmcode und Dokumentation analysieren. Die Versuchspersonen wurden zufällig in vier Gruppen eingeteilt. Jede Person sollte eine der beiden Aufgaben mit Hilfe des Werkzeuges und eine ohne lösen (siehe Abbildung 7.1).
Veränderte unabhängige Variablen sind also die Methode (manuell/Werkzeug), die Aufgabe und der Labortermin. Beobachtet als abhängige Variablen wurden

	Group			
	A	B	C	D
Lab1	T1 / AT	T1 / MT	T2 / AT	T2 / MT
Lab2	T2 / MT	T2 / AT	T1 / MT	T1 / AT

Abbildung 7.1: Experimententwurf aus EMSE09-57

Table 5 Descriptive statistics

	Task/method		Time (min)			F-Measure (%)			Recall (%)			Precision (%)		
			Median	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median	Mean	Std. Dev.
Exp I	All	MT	120.00	125.20	27.22	53.41	52.38	16.06	58.44	59.72	20.33	54.55	55.64	26.06
		AT	75.00	83.75	23.78	65.47	61.36	17.16	60.84	58.81	21.77	69.81	68.54	15.97
	T1	MT	115.00	115.50	22.54	49.47	47.22	17.79	74.19	67.74	22.19	39.63	38.77	16.93
		AT	72.50	80.00	22.24	52.66	53.61	14.62	51.62	52.69	20.60	61.22	59.07	12.22
	T2	MT	122.50	135.00	29.06	59.91	57.52	13.01	50.00	51.69	11.39	71.97	72.51	22.70
		AT	87.50	87.50	25.85	71.93	69.10	16.58	62.66	64.94	22.20	82.45	78.00	13.78
Exp II	All	MT	142.00	138.90	14.58	75.46	71.59	14.50	64.19	67.70	16.98	77.99	78.66	16.76
		AT	87.50	85.25	13.40	75.35	75.39	13.46	65.06	65.22	11.89	74.84	73.35	19.51
	T1	MT	140.00	134.50	19.01	79.44	73.75	14.42	80.96	71.20	18.63	77.99	80.07	15.69
		AT	92.50	92.17	12.36	65.13	64.98	9.76	75.35	72.24	12.40	61.02	64.19	22.49
	T2	MT	142.00	143.30	7.71	66.76	69.43	15.60	62.80	64.19	16.06	77.56	77.24	19.15
		AT	78.50	78.33	11.33	78.07	79.25	7.91	77.46	78.54	14.86	84.48	82.50	11.44

Task: T1 (tracing use cases onto code classes) or T2 (tracing interaction diagrams onto test cases)
Method: MT (manual tracing) or AT (ADAMS Re-Trace)

Abbildung 7.2: Berichtete beschreibende Statistiken aus EMSE09-57

die benötigte Zeit zum Lösen der Aufgaben und ein Maß zusammengesetzt aus Abdeckung (*recall*) und Treffgenauigkeit (*precision*).

Abbildung 7.2 zeigt die Tabelle der beschreibenden Statistiken, die im Artikel angegeben wurde. Es sind zwar für beide Experimente für die jeweiligen Gruppen der Median, der Mittelwert und die Standardabweichung der vier abhängigen Variablen angegeben, allerdings könnten die interessanten Resultate in der Menge der Zahlen untergehen. Um z.B. zu erkennen, ob die Gruppe, die mit dem Werkzeug gearbeitet hat, schneller die Aufgaben gelöst hat, muss der Leser die richtige Spalte suchen und dann verschiedene Zeilen vergleichen. Selbst dann wird ihm aber lediglich das Minimum an Maßzahlen wie Mittelwert und Standardabweichung (und der gegen Ausreißer robustere Median) präsentiert. Betrachtet man die Mittelwertspalte der Zeit, so stellt man fest, die AT Gruppe (arbeitete mit dem Werkzeug) benötigt in beiden Experimenten für beide Aufgaben weniger Zeit als die MT Gruppe (arbeitete ohne Werkzeug). Außerdem ist die Standardabweichung der Zeit in Experiment I stets höher als im zweiten Experiment.

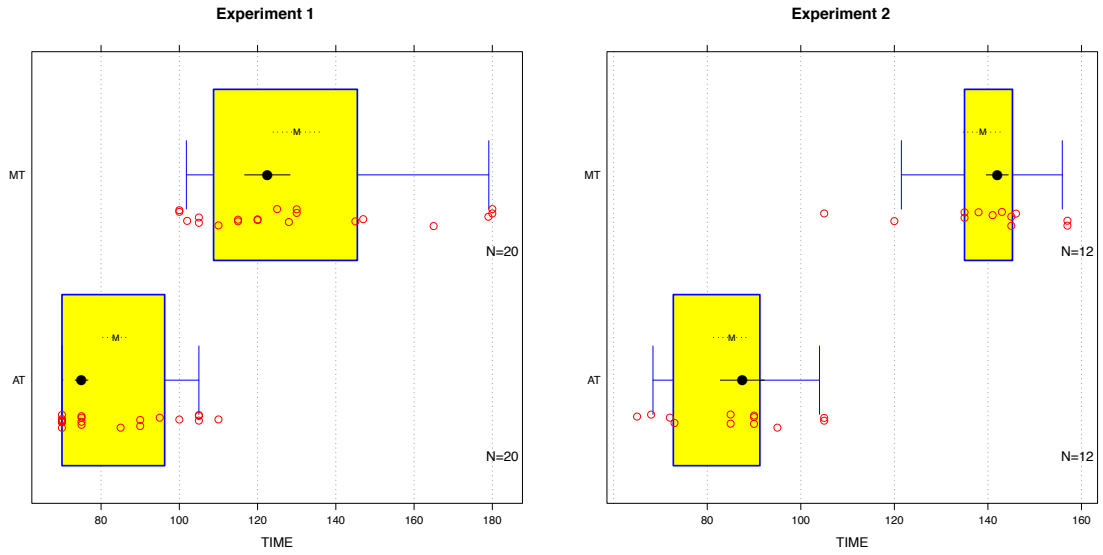
Für zusätzliche Analysen und Erstellung von Grafiken sind die zugesendeten Rohdaten im bestmöglichem Zustand (siehe Abbildung 7.3). Die Daten sind in einer Tabelle zusammengefasst, wobei die Spaltennamen gleich den Namen der Variablen im Artikel sind.

Abbildung 7.4 zeigt Boxplots der abhängigen Variable Zeit. Hierbei wird durch die Position der Boxen sofort der Unterschied in der benötigten Zeit sichtbar. Außerdem ist anhand der Breite der Boxen der Unterschied der Variabilitäten zu erkennen.

Zum Testen auf einen signifikanten Unterschied wurde der nichtparametrische Wilcoxon-Test für gepaarte Stichproben verwendet. Die berichteten Ergebnisse

EXPERIMENT	SUBJECT ID	LABORATORY	TASK	METHOD	RECALL	PRECISION	F-MEASURE	TIME
1	1	1	T1	AT	70.97	56.41	62.86	70
1	1	2	T2	MT	48.19	53.33	50.63	120
1	2	1	T1	AT	67.74	41.72	51.64	70
1	2	2	T2	MT	80.72	48.91	60.91	130
1	3	1	T1	AT	44.09	64.06	52.23	70
1	3	2	T2	MT	57.83	50.00	53.63	102
1	4	1	T1	AT	43.01	70.18	53.33	105
1	4	2	T2	MT	59.04	100.00	74.24	105
1	5	1	T1	AT	41.94	41.05	41.49	75

Abbildung 7.3: Ausschnitt der Rohdaten von EMSE09-57



(a) (b)

Abbildung 7.4: Boxplots für EMSE09-57 (Variable: Zeit)

sind in Abbildung 7.5 dargestellt. Sowohl für beide Experimente als auch für dessen Kombination ergeben sich signifikante Unterschiede in der Zeit. Außerdem wurden Mittelwert- und Mediandifferenzen sowie der prozentuale Unterschied (nichtstandardisierte Effektgrößen) angegeben. Die Mittelwerte sind allerdings anfällig gegen Ausreißer. Nach Formel 2.5 ist der Datenpunkt x=105 in Experiment 2 der MT Gruppe ein Ausreißer. Weiterhin wäre ein Vertrauensbereich informativer als ein einzelner Wert. Geht man davon aus, dass die Daten nicht normalverteilt sind, so lässt sich mit Hilfe der Bootstrap-Methode ein Vertrauensbereich für die Mittelwertdifferenz berechnen. Die Ergebnisse der Bootstrapberechnungen in R (basierend auf 10000 Berechnungen) sind in Tabelle 7.1 dargestellt. Der Standardfehler zeigt die Schwankungen der Mittelwertdifferenz an. Anhand der Konfidenzintervalle ist erkennbar, dass sowohl in beiden Expe-

Table 6 Wilcoxon paired test p-values and descriptive statistics of differences (by subject)

	Variable	Median	Mean	Std. Dev.	p-value	% of Positive effect
Exp I	Time	−36.00	−46.30	33.78	**7.1e-05**	95.00
	F-measure	11.50	8.92	20.37	**0.048**	65.00
	Recall	−7.95	−0.90	26.47	0.820	40.00
	Precision	18.57	13.10	32.17	**0.045**	65.00
Exp II	Time	−60.25	−49.00	−19.06	**0.001**	100.00
	F-measure	−5.09	0.53	14.26	0.484	33.33
	Recall	12.34	7.69	20.05	0.098	58.33
	Precision	−5.62	−5.31	19.50	0.867	41.67
All	Time	−44.00	−49.06	29.01	**6.1e-07**	96.88
	F-measure	5.69	5.74	18.54	0.064	53.13
	Recall	−3.30	2.32	24.29	0.430	46.88
	Precision	4.62	6.20	29.18	0.129	56.25

The data presented in bold indicated that the p-value is less than 0.05, i.e., the results are intended as statistically significant

Abbildung 7.5: Wilcoxon-Tests und Effektgrößen aus EMSE09-57

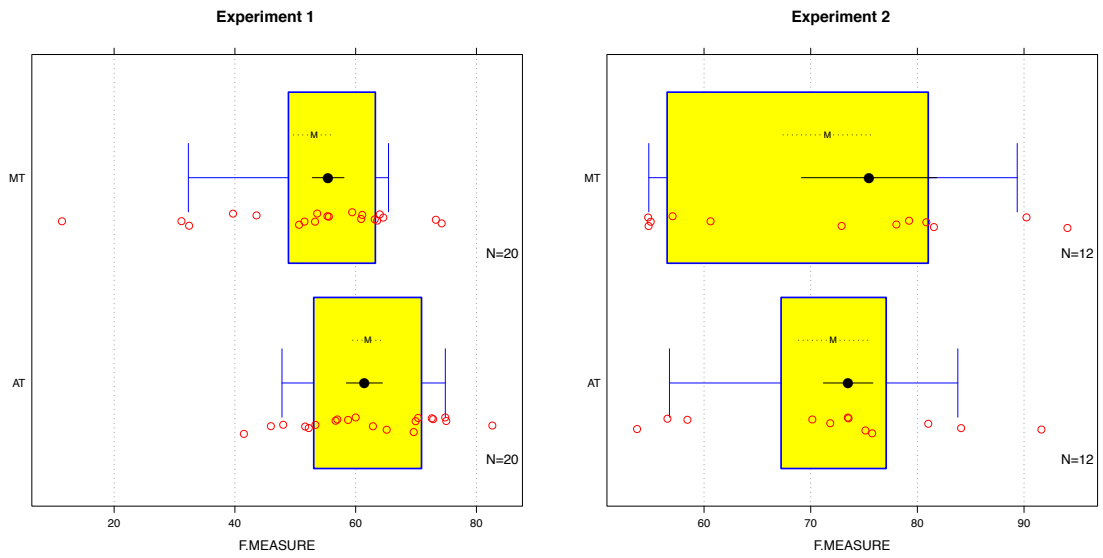
	Mittelwertdifferenz	Standardfehler	Konfidenzintervall (BCa)
Experiment 1	-46.3	6.86	(-61.0, -34.0)
Experiment 2	-53.7	5.58	(-63.9, -41.9)
Beide	-49.1	4.82	(-58.6, -39.7)

Tabelle 7.1: Bootstrap Ergebnisse für EMSE09-57 (Zeit)

rimenten als auch in der Analyse der kombinierten Daten die Versuchspersonen mit Hilfe des Werkzeugs mehr als 30 Minuten weniger Zeit benötigen. Anhand der Bootstrapberechnungen der Zeit unter zusätzlicher Berücksichtigung des F-Maßes kann der praktische Nutzen diskutiert werden. Tabelle 7.2 zeigt Ergebnisse des Bootstraps in R (10000 Replikationen) für das F-Maß. Betrachtet man die Konfidenzintervalle für das F-Maß, so stellt man fest, dass zwei von drei die Null beinhalten. Ist die Null im Intervall enthalten, so kann der Effekt sowohl negativ als auch positiv sein. Ein Unternehmer stünde also vor der Entscheidung seine Mitarbeiter für die Nutzung des Werkzeugs schulen zu lassen, damit sie ihre Ar-

	Mittelwertdifferenz	Standardfehler	Konfidenzintervall (BCa)
Experiment 1	8.92	4.17	(1.42, 17.8)
Experiment 2	0.52	5.19	(-9.59, 10.7)
Beide	5.77	3.73	(-1.20,13.4)

Tabelle 7.2: Bootstrap Ergebnisse für EMSE09-57 (F-Maß)



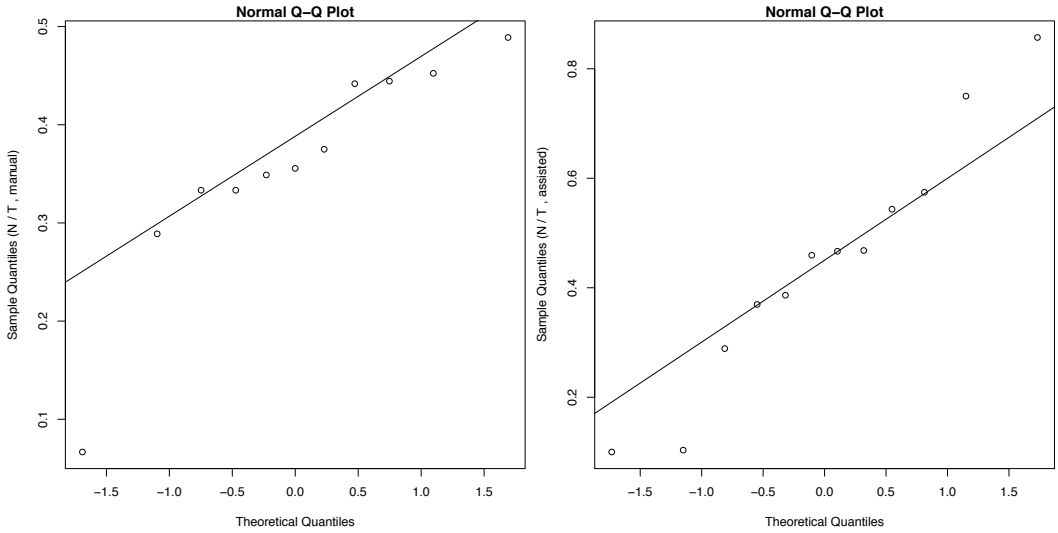
(a) (b)

Abbildung 7.6: Boxplots für EMSE09-57 (Variable: F-Maß)

beit schneller erledigen können. Da sich die Qualität der Ergebnisse aber auch negativ auswirken kann (Null im Konfidenzintervall des F-Maßes), wird deutlich, dass eine Zeitersparnis allein nicht ausreicht. Abbildung 7.6 stellt die Daten des F-Maßes anhand von Boxplots dar. Dabei ist eine große Variabilität (große Box) der MT Gruppe in Experiment 2 zu erkennen. Diese sollte weiter untersucht werden. Führt man den Wilcoxon-Test für das F-Maß mit den Daten aus Experiment 1 ohne den Ausreißer der MT Gruppe durch, so ergibt sich ein nichtsignifikantes Ergebnis von $p = 0.08$.
Außerdem sei an dieser Stelle noch einmal auf die externe Gültigkeit hingewiesen, denn das Experiment wurde im Rahmen einer Vorlesung mit Studenten als Versuchspersonen durchgeführt.

7.2 Artikel: EMSE06-303

Dieser Artikel beschreibt ein Experiment, in dem die Effizienz eines Werkzeugs zur Unterstützung der Konsolidierung von Anforderungen untersucht wird. Die studentischen Versuchsteilnehmer wurden zufällig auf zwei Gruppen aufgeteilt, die jeweils manuell bzw. mit Hilfe des Werkzeugs arbeiteten. Gemessene abhängige Variablen sind die benötigte Zeit, die Anzahl analysierter Anforderungen und weitere Anzahlen.



(a) Gruppe ohne Werkzeug (b) Gruppe mit Werkzeug

Abbildung 7.7: Quantil-Quantil-Plots für N/T (EMSE06-303)

Die Hypothese wurde einseitig getestet, da die Autoren nur daran interessiert waren, herauszufinden, ob mit Hilfe ihrers Werkzeugs effizientere Ergebnisse resultieren. Eine der getesteten Hypothesen lautet:

> H_0^1: Die Anzahl an analysierten Anforderungen pro Minute für beide Gruppen (mit und ohne Werkzeug) sind gleich.
>
> H_A^1: Die Anzahl an analysierten Anforderungen pro Minute der Gruppe, die mit dem Werkzeug arbeitet, ist größer als die der Gruppe die manuell arbeitet.

In den Rohdaten ist leider nicht vermerkt, welche der drei im Artikel erwähnten Studenten nicht mit in die Analyse mit einbezogen wurden, weil sie nicht alle Teile des Experiments mitgemacht haben. Im Artikel wird beschrieben, dass für beide Gruppen jeweils 10 Versuchspersonen mit einbezogen wurden, die Rohdaten beinhalten allerdings Werte für 12 Teilnehmer der einen und 11 Teilnehmer der anderen Gruppe.

Für die Überprüfung der Annahmen des t-Tests wurde beschrieben, dass "normal probability plots" durchgeführt wurden. Leider wurde lediglich erwähnt, dass die Autoren zu dem Schluss kommen, dass parametrische Tests anwendbar sind. Verwendet man alle Datenpunkte für die Anzahl an analysierten Anforderungen pro Minute (N/T), so entstehen die in Abbildung 7.7 gezeigten Quantil-Quantil-Plots. Die Datenpunkte weichen sichtbar von einer Geraden ab, was wie

	Mittelwertdifferenz	Standardfehler	Konfidenzintervall (BCa)
N/T	0.090	0.072	(-0.039, 0.24)

Tabelle 7.3: Bootstrap Ergebnisse für EMSE06-303 (N/T)

in Abschnitt 2.2.3 beschrieben, Abweichungen von der Normalverteilung bedeutet.
Man könnte also auch defensiv agieren und nichtparametrische Tests durchführen. Ein einseitiger Mann-Whitney-Test mit einem Signifikanzniveau von $\alpha = 0.05$ ergibt einen p-Wert von 0.07. Eine zusätzliche Berechnung mit Hilfe der Bootstrap-Methode ist in Tabelle 7.3 dargestellt.
Der im Artikel durchgeführte t-Test ergab ein signifikantes Ergebnis mit einem p-Wert kleiner als 0.004. Die Autoren schließen daraus, dass mit Hilfe des Werkzeugs eine höhere Leistung erziehlt werden kann. Weitere Analysen seien aber erforderlich, um endgültige Schlußfolgerungen zu ziehen.
Die Ergebnisse des Bootstraps zeigen allerdings, dass die Schwankungen der Mittelwertdifferenz fast so groß sind wie die Differenz selbst. Außerdem enthält der Vertrauensbereich die Null, d.h. der Effekt kann auch leicht negativ werden, bzw. verschwinden. Die obere Grenze des Vertrauensbereichs von 0.24 bedeutet, dass in ca. 3 Minuten maximal eine Anforderung mehr analysiert werden kann. Das wären 20 Anforderungen mehr pro Stunde. Zusammen mit der Qualität der Ergebnisse kann dann der praktische Nutzen diskutiert werden.

7.3 Artikel: EMSE08-693

Ein weiteres Beispiel soll aufzeigen, dass es teils schwierig ist Rohdaten zu verwenden, wenn diese nicht hinreichend aufbereitet sind. Es soll nicht dazu dienen die Autoren zu kritisieren, denn es ist lobenswert, dass sie auf die Rohdatenanfrage geantwortet haben und die Daten des Experiments aufbewahren.
Abbildung 7.9 (Seite 61) zeigt die zugesendeten Rohdaten des Artikels EMSE08-693. Diese sind nicht ohne zusätzlichen Aufwand als CSV-Daten exportierbar und maschniell verarbeitbar. Teilweise sind die Bedeutungen auch unklar.
Konsultiert man den zugehörigen Artikel, so stellt man fest, es handelt sich um ein Experiment mit zwei Gruppen und zehn Aufgaben. Unabhängige Variablen sind der Typ (Wissenschaftler, Student) und die Konfiguration des verwendeten Tools (Gruppe). Gemessen wurden die erzielten Punkte der einzelnen Aufgaben. Die Gruppen sind in Spalte C benannt. Der dritte Buchstabe der ID (Spalte B) scheint ebenfalls die Gruppe zu beschreiben. Spalte E bis N, Zeile 4 sind mit 1–10 numeriert, sind also die Punkte der zehn Aufgaben. Rechnet man nach, dann

Table 3 Number of participants per person group and configuration

Configuration	A		B	
Person Group	Researcher	Student	Researcher	Student
Participants	7	6	5	9

Abbildung 7.8: Experimententwurf aus EMSE08-693

stellt man fest, dass Spalte P die Summe der Spalten E bis N ist, also die Gesamtpunktzahl.
Abbildung 7.8 zeigt den Experimententwurf. Dort ist erkennbar, dass für beide Gruppen sowohl Wissenschaftler als auch Studenten eingeteilt wurden. Der zweite Buchstabe der ID in Spalte B von Abbildung 7.9 könnte demnach "Researcher" bzw. "Student" bedeuten. In dieser Form ist es allerdings schwierig diese Information zu verwenden. Bei der Nummerierung der ID ist außerdem zu erkennen, dass Gruppe A nur ungerade und Gruppe B nur gerade Ziffern besitzt. Der erste Buchstabe (K bzw. M) beschreibt das Institut von dem die Testperson stammt. Eine Empfehlung, wie die Rohdaten verständlicher werden und vor allem maschinenlesbar zeigt Abbildung 7.10 (Seite 62).
Für die Auswertung des Experiments wurden laut dem Artikel nur 27 von 29 Datenpunkte benutzt:

> "Two of the 29 participants failed with respect to the exclusionary criteria (filter tasks) and therefore the description and evaluation of the experiment addresses only 27 participants."

Da allerdings nicht vermerkt ist, welche Datenpunkte nicht mit einbezogen wurden, können die Ergebnisse nur schwierig nachvollzogen bzw. überprüft werden. Außerdem wurden im Anhang A des Artikels die maximal zu erzielenden Punkte der einzelnen Aufgaben angegeben. So soll es z.B. für Aufgabe 1 maximal null Punkte und für Aufgabe 3 maximal einen Punkt geben. In den Rohdaten allerdings sind fast alle Werte für Aufgabe 1 größer als null. Für Aufgabe 3 gibt es einen Versuchsteilnehmer der mehr als die maximale Punktzahl erhalten hat. Die insgesamt maximale Punktzahl aller Aufgaben von 15 wurde auch mehrmals überschritten.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Experiment Evaluation**															
2	**Task Results**															
3																
4		*ID*	*Group*		1	2	3	4	5	6	7	8	9	10		
5		K_R_A_1	A		1.5	2	0	0	1	0	1	1.5	1	0		8
6		K_R_A_3	A		1.5	2	0	1	1	2	1	3	1	3		15.5
7		K_R_A_5	A		0.5	2	0	1	1	0.5	0	0	0	0		5
8		K_R_A_7	A		2	2	0	1	1	0	1	2.5	1	0		10.5
9		K_R_A_9	A		1.5	1	0	1	1	1.5	0.5	2	1	0		9.5
10		M_R_A_1	A		1	2	0	0	0	0	0	0	0	0		3
11		M_R_A_3	A		0	0	0	1	1	0	0	1	1	0		4
12																
13																
14		K_R_B_2	B		2	0	1	1	1	2	1	3	1	0		12
15		K_R_B_4	B		2	2	1	1	1	2	1	3	1	3		17
16		K_R_B_6	B		2	2	1	0	0	1	1	3	1	3		14
17		K_R_B_8	B		2	2	1	1	1	0	1	4	1	0		13
18		K_R_B_10	B		2	2	1	1	1	1	1	3	1	3		16
19																
20																
21		K_S_A_1	A		2	2	0	1	1	1	0	0	0	0		7
22		K_S_A_11	A		2	2	1	1	1	0	1	0	1	0		9
23		K_S_A_13	A		1.5	2	0	0	0	0.5	1	1	1	0		7
24		K_S_A_15	A		1.5	2	0	1	0	2	1	1.5	1	0		10
25		K_S_A_3	A		1.5	2	0	1	0	1	0	0	0	0		5.5
26		K_S_A_5	A		0.5	2	0	1	1	2	1	2.5	1	3		14
27		K_S_A_7	A		0	0	0	0	0	0	0	0	0	0		0
28		K_S_A_9	A		0	0	0	0	0	0	0	0	0	0		0
29																
30																
31		K_S_B_10	B		2	2	1	1	1	2	1	3	1	0		14
32		K_S_B_12	B		2	2	1	1	1	1.5	1	3	1	2		15.5
33		K_S_B_14	B		2	0	1	1	0	1.5	1	3	1	0		10.5
34		K_S_B_16	B		2	0	1	1	1	0.5	1	3	1	0		10.5
35		K_S_B_2	B		2	2	1	1	1	1	1	3	1	0		13
36		K_S_B_4	B		2	1.5	1	1	0	2	1	3	1	0		12.5
37		K_S_B_6	B		2	2	2	1	1	0	1	2.5	1	0		12.5
38		K_S_B_8	B		2	2	1	1	1	0.5	1	2	1	0		11.5
39		M_S_B_2	B		2	2	1	1	1	2	1	0	1	3		14

Abbildung 7.9: Unveränderte Rohdaten von EMSE08-693

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Id	Institute	Type	Group	Task1	Task2	Task3	Task4	Task5	Task6	Task7	Task8	Task9	Task10	Sum
2	K_R_A_1	K	researcher	A	1.5	2	0	0	1	0	1	1.5	1	0	8
3	K_R_A_3	K	researcher	A	1.5	2	0	1	1	2	1	3	1	3	15.5
4	K_R_A_5	K	researcher	A	0.5	2	0	1	1	0.5	0	0	0	0	5
5	K_R_A_7	K	researcher	A	2	2	0	1	1	0	1	2.5	1	0	10.5
6	K_R_A_9	K	researcher	A	1.5	1	0	1	1	1.5	0.5	2	1	0	9.5
7	M_R_A_1	M	researcher	A	1	2	0	0	0	0	0	0	0	0	3
8	M_R_A_3	M	researcher	A	0	0	0	1	1	0	0	1	1	0	4
9	K_R_B_2	K	researcher	B	2	0	1	1	1	2	1	3	1	0	12
10	K_R_B_4	K	researcher	B	2	2	1	1	1	2	1	3	1	3	17
11	K_R_B_6	K	researcher	B	2	2	1	0	0	1	1	3	1	3	14
12	K_R_B_8	K	researcher	B	2	2	1	1	1	0	1	4	1	0	13
13	K_R_B_10	K	researcher	B	2	2	1	1	1	1	1	3	1	3	16
14	K_S_A_1	K	student	A	2	2	0	1	1	1	0	0	0	0	7
15	K_S_A_11	K	student	A	2	2	1	1	1	0	1	0	1	0	9
16	K_S_A_13	K	student	A	1.5	2	0	0	0	0.5	1	1	1	0	7
17	K_S_A_15	K	student	A	1.5	2	0	1	0	2	1	1.5	1	0	10
18	K_S_A_3	K	student	A	1.5	2	0	1	0	1	0	0	0	0	5.5
19	K_S_A_5	K	student	A	0.5	2	0	1	1	2	1	2.5	1	3	14
20	K_S_A_7	K	student	A	0	0	0	0	0	0	0	0	0	0	0
21	K_S_A_9	K	student	A	0	0	0	0	0	0	0	0	0	0	0
22	K_S_B_10	K	student	B	2	2	1	1	1	2	1	3	1	0	14
23	K_S_B_12	K	student	B	2	2	1	1	1	1.5	1	3	1	2	15.5
24	K_S_B_14	K	student	B	2	0	1	1	0	1.5	1	3	1	0	10.5
25	K_S_B_16	K	student	B	2	0	1	1	1	0.5	1	3	1	0	10.5
26	K_S_B_2	K	student	B	2	2	1	1	1	1	1	3	1	0	13
27	K_S_B_4	K	student	B	2	1.5	1	1	0	2	1	3	1	0	12.5
28	K_S_B_6	K	student	B	2	2	2	1	1	0	1	2.5	1	0	12.5
29	K_S_B_8	K	student	B	2	2	1	1	1	0.5	1	2	1	0	11.5
30	M_S_B_2	M	student	B	2	2	1	1	1	2	1	0	1	3	14

Abbildung 7.10: Überarbeitete Empfehlung für die Rohdaten von EMSE08-693

Kapitel 8

Schlussfolgerung und Ausblick

Wie auch schon Sjøberg et al. (2005) feststellten, ist die Anzahl der durchgeführten (bzw. in Artiklen berichteten) Experimenten sehr gering. Die Ermittlung der Art und Weise, wie in diesen Artikeln die Ergebnisse von kontrollierten Experimenten präsentiert werden, wurde unter drei Gesichtspunkten durchgeführt.
Die Angabe von beschreibenden Statistiken ist auf ein Minimum beschränkt, indem meist Mittelwerte und Standardabweichungen in tabellarischer Form angegeben werden. Boxplots zur Visualisierung der gewonnenen Daten werden nur mässig verwendet.
Für statistische Tests zur Auswertung der Daten werden Ergebnisse mit geringem Detail berichtet, indem nur die Entscheidung des Tests bzw. der resultierende p-Wert und/oder das Signifikanzniveau berichtet werden. Die Diskussion der praktischen Relevanz in Form von Effektgrößen und Konfidenzintervallen ist selten vorhanden.
Rohdatenanfragen zur weiteren Untersuchung werden zwar häufig beantwortet, allerdings können nur in einzelnen Fällen diese auch verfügbar gemacht werden. Ein weiteres Problem ist der teils schlechte Zustand der Rohdaten.
Aufgrund der erhaltenen Ergebnisse bleibt die Aussage von Kitchenham et al. (2002)

> "In our view, the standard of empirical software engineering is poor."

> "…the majority of problems result from lack of statistical expertise in the empirical research community."

bestehen. Der noch junge Bereich des Empirical Software Engineerings sollte mit weiteren Richtlinien und Empfehlungen ergänzt werden. Bestehende Probleme sollten aufgezeigt werden, um sie künftig zu vermeiden.

Abbildungsverzeichnis

2.1. Streudiagramm und Datenpaare (Beispiel) 9
2.2. Korrelationskoeffizient (Beispiel) 10
2.3. Boxplot (Beispiel 1) 11
2.4. Boxplot (Beispiel 2) 11
2.5. Veranschaulichung der Fehlertypen aus Dybå et al. (2006) 14
2.6. Normalverteilung $\mathcal{N}(1,1)$ 16
2.7. Quantil-Quantil-Plot (Beispiel) 17

4.1. Identifizierungsprozedur 29

5.1. Anzahl identifizierter Artikel mit kontrollierten Experimenten . . . 35
5.2. Anzahl an berichteten kontrollierten Experimenten 36
5.3. Anzahl durchgeführter statistischer Tests 38
5.4. Anzahl durchgeführter Tests pro Experiment 39
5.5. Problematisierung der Annahmen bei parametrischen Tests 40
5.6. Begründung für die Verwendung nicht-parametrischer Verfahren . 40
5.7. Präsenz und Verfügbarkeit von Rohdaten 43

7.1. Experimententwurf aus EMSE09-57 53
7.2. Berichtete beschreibende Statistiken aus EMSE09-57 54
7.3. Ausschnitt der Rohdaten von EMSE09-57 55
7.4. Boxplots für EMSE09-57 (Variable: Zeit) 55
7.5. Wilcoxon-Tests und Effektgrößen aus EMSE09-57 56
7.6. Boxplots für EMSE09-57 (Variable: F-Maß) 57
7.7. Quantil-Quantil-Plots für N/T (EMSE06-303) 58
7.8. Experimententwurf aus EMSE08-693 60
7.9. Unveränderte Rohdaten von EMSE08-693 61
7.10. Überarbeitete Empfehlung für die Rohdaten von EMSE08-693 . . . 62

Tabellenverzeichnis

2.1. Formelzeichen statistischer Begriffe 7
2.2. Variablenskalen 8
2.3. Fehler erster und zweiter Art bei Hypothesentests 13

5.1. Anzahl untersuchter Artikel 34
5.2. Anzahl identifizierter Artikel mit kontrollierten Experimenten . . . 34
5.3. Mittelwert der identifizierten Artikel pro Jahr (aufgerundet) 34
5.4. Anzahl an berichteten kontrollierten Experimenten 36
5.5. Angabe von statistischen Größen und Grafiken 37
5.6. Anzahl durchgeführter statistischer Tests 38
5.7. Duchführung von Trennschärfeanalysen (nach Artikel) 41
5.8. Angabe von Effektgrößen (nach Experimenten) 41

7.1. Bootstrap Ergebnisse für EMSE09-57 (Zeit) 56
7.2. Bootstrap Ergebnisse für EMSE09-57 (F-Maß) 56
7.3. Bootstrap Ergebnisse für EMSE06-303 (N/T) 59

1. Erfasste Variablen bzgl. der Artikel 81
2. Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 1) 82
3. Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 2) 83
4. Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 3) 84

Identifizierte Artikel (2006–2009)

H. Do, G. Rothermel, and A. Kinneer, "Prioritizing junit test cases: An empirical assessment and cost-benefits analysis," *Empirical Software Engineering*, vol. 11, no. 1, pp. 33–70, 03 2006/03/16/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-5965-8

J. Maldonado, J. Carver, F. Shull, S. Fabbri, E. Dória, L. Martimiano, M. Mendonça, and V. Basili, "Perspective-based reading: A replicated experiment focused on individual reviewer effectiveness," *Empirical Software Engineering*, vol. 11, no. 1, pp. 119–142, 03 2006/03/16/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-5967-6

S. Syed-Abdullah, M. Holcombe, and M. Gheorge, "The impact of an agile methodology on the well being of development teams," *Empirical Software Engineering*, vol. 11, no. 1, pp. 143–167, 03 2006/03/16/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-5968-5

A. Sinha and C. Smidts, "An experimental evaluation of a higher-ordered-typed-functional specification-based test-generation technique," *Empirical Software Engineering*, vol. 11, no. 2, pp. 173–202, 06 2006/06/11/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-6401-9

C. Bunse, "Using patterns for the refinement and translationof uml models: A controlled experiment," *Empirical Software Engineering*, vol. 11, no. 2, pp. 227–267, 06 2006/06/11/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-6403-7

J. Natt och Dag, T. Thelin, and B. Regnell, "An experiment on linguistic tool support for consolidation of requirements from multiple sources in market-driven product development," *Empirical Software Engineering*, vol. 11, no. 2, pp. 303–329, 06 2006/06/11/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-6405-5

J. Carver, F. Shull, and V. Basili, "Can observational techniques help novices overcome the software inspection learning curve? an empirical investigation," *Empirical Software Engineering*, vol. 11, no. 4, pp. 523–539, 12 2006/12/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-9021-5

L. Karlsson, T. Thelin, B. Regnell, P. Berander, and C. Wohlin, "Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques," *Empirical Software Engineering*, vol. 12, no. 1, pp. 3–33, 02 2007/02/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-006-7240-4

M. Genero, E. Manso, A. Visaggio, G. Canfora, and M. Piattini, "Building measure-based prediction models for uml class diagram maintainability," *Empirical Software Engineering*, vol. 12, no. 5, pp. 517–549, 10 2007/10/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9038-4

D. Yeh, P.-c. Sun, W. Chu, C.-L. Lin, and H. Yang, "An empirical study of a reverse engineering method for the aggregation relationship based on operation propagation," *Empirical Software Engineering*, vol. 12, no. 6, pp. 575–592, 12 2007/12/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9043-7

M. Müller and A. Höfer, "The effect of experience on the test-driven development process," *Empirical Software Engineering*, vol. 12, no. 6, pp. 593–615, 12 2007/12/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9048-2

M. Wojcicki and P. Strooper, "Maximising the information gained from a study of static analysis technologies for concurrent software," *Empirical Software Engineering*, vol. 12, no. 6, pp. 617–645, 12 2007/12/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9044-6

M. Babar, B. Kitchenham, and R. Jeffery, "Comparing distributed and face-to-face meetings for software architecture evaluation: A controlled experiment," *Empirical Software Engineering*, vol. 13, no. 1, pp. 39–62, 02 2008/02/28/. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9052-6

S. Acuña, M. Gómez, and N. Juristo, "Towards understanding the relationship between team climate and software quality—a quasi-experimental study," *Empirical Software Engineering*, vol. 13, no. 4, pp. 401–434, 08 2008/08/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9074-8

D. Budgen, B. Kitchenham, S. Charters, M. Turner, P. Brereton, and S. Linkman, "Presenting software engineering results using structured abstracts: a

randomised experiment," *Empirical Software Engineering*, vol. 13, no. 4, pp. 435–468, 08 2008/08/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9075-7

J. Knodel, D. Muthig, and M. Naab, "An experiment on the role of graphical elements in architecture visualization," *Empirical Software Engineering*, vol. 13, no. 6, pp. 693–726, 12 2008/12/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9069-5

A. De Lucia, R. Oliveto, and G. Tortora, "Assessing ir-based traceability recovery tools through controlled experiments," *Empirical Software Engineering*, vol. 14, no. 1, pp. 57–92, 02 2009/02/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9090-8

P. Sfetsos, I. Stamelos, L. Angelis, and I. Deligiannis, "An experimental investigation of personality types impact on pair effectiveness in pair programming," *Empirical Software Engineering*, vol. 14, no. 2, pp. 187–226, 04 2009/04/01/. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9093-5

C. Gresse von Wangenheim, M. Thiry, and D. Kochanski, "Empirical evaluation of an educational game on software measurement," *Empirical Software Engineering*, vol. 14, no. 4, pp. 418–452, 08 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9092-6

M. El-Attar and J. Miller, "A subject-based empirical evaluation of ssucd's performance in reducing inconsistencies in use case models," *Empirical Software Engineering*, vol. 14, no. 5, pp. 477–512, 10 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9101-9

J. Cruz-Lemus, M. Genero, M. Manso, S. Morasca, and M. Piattini, "Assessing the understandability of uml statechart diagrams with composite states—a family of empirical studies," *Empirical Software Engineering*, vol. 14, no. 6, pp. 685–719, 12 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-009-9106-z

G. Güleşir, K. van den Berg, L. Bergmans, and M. Akşit, "Experimental evaluation of a tool for the verification and transformation of source code in event-driven systems," *Empirical Software Engineering*, vol. 14, no. 6, pp. 720–777, 12 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-009-9107-y

R. Abraham and M. Erwig, "Inferring templates from spreadsheets," in *ICSE '06: Proceedings of the 28th international conference on Software engineering*. New York, NY, USA: ACM, 2006, pp. 182–191. [Online]. Available: http://doi.acm.org/10.1145/1134285.1134312

C. F. J. Lange and M. R. V. Chaudron, "Effects of defects in uml models: an experimental investigation," in *ICSE '06: Proceedings of the 28th international conference on Software engineering*. New York, NY, USA: ACM, 2006, pp. 401–411. [Online]. Available: http://doi.acm.org/10.1145/1134285.1134341

T. H. Ng, S. C. Cheung, W. K. Chan, and Y. T. Yu, "Do maintainers utilize deployed design patterns effectively?" in *ICSE '07: Proceedings of the 29th international conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 168–177. [Online]. Available: http://dx.doi.org/10.1109/ICSE.2007.33

B. Ellis, J. Stylos, and B. Myers, "The factory pattern in api design: A usability evaluation," in *ICSE '07: Proceedings of the 29th international conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 302–312. [Online]. Available: http://dx.doi.org/10.1109/ICSE.2007.85

F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, and M. Ceccato, "The role of experience and ability in comprehension tasks supported by uml stereotypes," in *ICSE '07: Proceedings of the 29th international conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 375–384. [Online]. Available: http://dx.doi.org/10.1109/ICSE.2007.86

F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, M. Ceccato, and C. A. Visaggio, "Are fit tables really talking?: a series of experiments to understand whether fit tables are useful during evolution tasks," in *ICSE '08: Proceedings of the 30th international conference on Software engineering*. New York, NY, USA: ACM, 2008, pp. 361–370. [Online]. Available: http://doi.acm.org/10.1145/1368088.1368138

E. Arisholm, L. C. Briand, S. E. Hove, and Y. Labiche, "The impact of uml documentation on software maintenance: An experimental evaluation," *IEEE Transactions on Software Engineering*, vol. 32, no. 6, pp. 365–381, 2006. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2006.59

H. Do and G. Rothermel, "On the use of mutation faults in empirical assessments of test case prioritization techniques," *IEEE Transactions on Software Engineering*, vol. 32, no. 9, pp. 733–752, 2006. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2006.92

T. Menzies, J. Greenwald, and A. Frank, "Data mining static code attributes to learn defect predictors," *IEEE Transactions on Software Engineering*, vol. 33, no. 1, pp. 2–13, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.10

E. Arisholm, H. Gallis, T. Dybå, and D. I. Sjøberg, "Evaluating pair programming with respect to system complexity and programmer expertise," *IEEE Transactions on Software Engineering*, vol. 33, no. 2, pp. 65–86, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.17

M. Agrawal and K. Chari, "Software effort, quality, and cycle time: A study of cmm level 5 projects," *IEEE Transactions on Software Engineering*, vol. 33, no. 3, pp. 145–156, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.29

Z. Li, M. Harman, and R. M. Hierons, "Search algorithms for regression test case prioritization," *IEEE Transactions on Software Engineering*, vol. 33, no. 4, pp. 225–237, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.38

W. Masri, A. Podgurski, and D. Leon, "An empirical study of test case filtering techniques based on exercising information flows," *IEEE Transactions on Software Engineering*, vol. 33, no. 7, pp. 454–477, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.1020

K. M. Lui, K. C. Chan, and J. Nosek, "The effect of pairs in program design tasks," *IEEE Transactions on Software Engineering*, vol. 34, no. 2, pp. 197–211, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.70755

D. Dig, K. Manzoor, R. E. Johnson, and T. N. Nguyen, "Effective software merging in the presence of object-oriented refactorings," *IEEE Transactions on Software Engineering*, vol. 34, no. 3, pp. 321–335, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2008.29

W. J. Dzidek, E. Arisholm, and L. C. Briand, "A realistic empirical evaluation of the costs and benefits of uml in software maintenance," *IEEE Transactions on Software Engineering*, vol. 34, no. 3, pp. 407–432, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2008.15

J. C. Carver, N. Nagappan, and A. Page, "The impact of educational background on the effectiveness of requirements inspections: An empirical study," *IEEE Transactions on Software Engineering*, vol. 34, no. 6, pp. 800–812, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2008.49

M. Jørgensen and T. M. Gruschke, "The impact of lessons-learned sessions on effort estimation and uncertainty assessments," *IEEE Transactions on Software Engineering*, vol. 35, no. 3, pp. 368–383, 2009. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2009.2

Literaturverzeichnis

V. R. Basili, "The role of experimentation in software engineering: past, current, and future," in *ICSE '96: Proceedings of the 18th international conference on Software engineering*. Washington, DC, USA: IEEE Computer Society, 1996, pp. 442–449.

K. M. Bower. (2009, September) What is a p-value? www. [Online]. Available: http://www.keithbower.com/Basic%20Stats/What%20is%20a%20p-value.htm

K. M. Bower. (2009, September) What are confidence intervals? www. [Online]. Available: http://www.keithbower.com/Basic%20Stats/WhatAreCI.htm

K. M. Bower. (2009, September) Outliers in boxplots. www. [Online]. Available: http://www.keithbower.com/Basic%20Stats/Outliers%20in%20Boxplots.htm

K. M. Bower. (2009, September) What is statistical power? www. [Online]. Available: http://www.keithbower.com/Basic%20Stats/What%20is%20power.htm

L. A. Clarke, L. Dillon, and W. Tichy, "Message from the chairs," in *ICSE '03: Proceedings of the 25th International Conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2003, p. .15.

J. Cohen, *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*, 2nd ed. Lawrence Erlbaum, January 1988. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0805802835

J. Cohen, "The earth is round (p < .05)." *American Psychologist*, vol. 49, no. 12, pp. 997–1003, 1994.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Second Edition*. The MIT Press, September 2001.

J. Deacon. (2009, November) Descriptive statistics. www. [Online]. Available: http://www.biology.ed.ac.uk/research/groups/jdeacon/statistics/tress3.html#DESCRIPTIVE%20STATISTICS

O. Dieste, A. Grimán, and N. Juristo, "Developing search strategies for detecting relevant experiments," *Empirical Software Engineering*, vol. 14, no. 5, pp. 513–539, 10 2009. [Online]. Available: http://dx.doi.org/10.1007/s10664-008-9091-7

T. Dybå, V. B. Kampenes, and D. I. Sjøberg, "A systematic review of statistical power in software engineering experiments," *Information and Software Technology*, vol. 48, no. 8, pp. 745 – 755, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6V0B-4HGD75T-3/2/f0670d7c4c3c1d0da5b66e3f50786742

W. Emmerich and G. Rothermel, "Message from the program committee co-chairs," in *ICSE '07: Proceedings of the 29th international conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2007, p. xvi.

S. Fickas, J. Atlee, and P. Inverardi, "Foreword," in *ICSE '09: Proceedings of the 2009 IEEE 31st International Conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2009, p. iii.

A. Finkelstein, J. Estublier, and D. Rosenblum, "Letter from the chairs," in *ICSE '04: Proceedings of the 26th International Conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2004, p. .15.

H. Fukuda and Y. Ohashi, "A guideline for reporting results of statistical analysis in Japanese Journal of Clinical Oncology," *Jpn. J. Clin. Oncol.*, vol. 27, no. 3, pp. 121–127, 1997. [Online]. Available: http://jjco.oxfordjournals.org/cgi/content/abstract/27/3/121

R. L. Glass, I. Vessey, and V. Ramesh, "Research in software engineering: an analysis of the literature," *Information and Software Technology*, vol. 44, no. 8, pp. 491 – 506, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/B6V0B-45KT0DK-1/2/6fd665b80e974e965c9665720209d8d6

J. A. Gliner, N. L. Leech, and G. A. Morgan, "Problems with null hypothesis significance testing (nhst): What do the textbooks say ?" *Journal of Experimental Education, The*, vol. 71, no. 1, pp. 83–92, Sep. 2002.

A. Greenwald, A. R. Gonzalez, R. J. Harris, and D. Guthre, "Effect sizes and p values: What should be reported . . . ?" *Psychophysiology*, vol. 33, 1996.

S. Gregor, "The nature of theory in information systems," *MIS Quarterly*, vol. 30, no. 3, pp. 491–506, September 2006, referenced in Hannay et al. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.6996&rep=rep1&type=pdf

W. Griswold and B. Nuseibeh, "Introduction to research papers," in *ICSE '05: Proceedings of the 27th international conference on Software engineering*. New York, NY, USA: ACM, 2005, pp. 48–48.

J. E. Hannay, D. I. Sjøberg, and T. Dybå, "A systematic review of theory use in software engineering experiments," *IEEE Transactions on Software Engineering*, vol. 33, no. 2, pp. 87–107, 2007. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2007.12

IEEE. (2009, 06) IEEE keyword taxonomy. www. [Online]. Available: http://www.computer.org/mc/keywords/software.htm

A. Jedlitschka and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," *Empirical Software Engineering, International Symposium on*, vol. 0, p. 10 pp., 2005. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ISESE.2005.1541818

V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "Systematic review: A systematic review of effect size in software engineering experiments," *Inf. Softw. Technol.*, vol. 49, no. 11-12, pp. 1073–1086, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2007.02.015

V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg, "A systematic review of quasi-experiments in software engineering," *Information and Software Technology*, vol. 51, no. 1, pp. 71 – 82, 2009, special Section - Most Cited Articles in 2002 and Regular Research Papers. [Online]. Available: http://www.sciencedirect.com/science/article/B6V0B-4SD6SHR-2/2/98755e0c4a5ca1df95545f862b159214

B. Kitchenham, "Procedures for performing systematic reviews," Keele University and NICTA, Tech. Rep., 2004.

B. Kitchenham, H. Al-Khilidar, M. Babar, M. Berry, K. Cox, J. Keung, F. Kurniawati, M. Staples, H. Zhang, and L. Zhu, "Evaluating guidelines for reporting empirical software engineering studies," *Empirical Software Engineering*, vol. 13, no. 1, pp. 97–121, 02 2008. [Online]. Available: http://dx.doi.org/10.1007/s10664-007-9053-5

B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, 2002. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2002.1027796

W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, dec 1952.

D. Meschede, *Gerthsen Physik*, 23rd ed. Springer, Berlin, Mar. 2006.

R. G. Miller, *Beyond ANOVA: Basics of Applied Statistics*. Chapman & Hall/CRC, 1997. [Online]. Available: http://www.amazon.com/Beyond-ANOVA-Applied-Statistics-Statistical/dp/0412070111

NIST. (2009, December) Normal probability plot. www. [Online]. Available: http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

NIST. (2009, September) What are confidence intervals? www. [Online]. Available: http://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm

NIST. (2009, November) Kolmogorov-smirnov goodness-of-fit test. www. [Online]. Available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

NIST. (2009, September) Do two arbitrary processes have the same mean ? www. [Online]. Available: http://www.itl.nist.gov/div898/handbook/prc/section3/prc35.htm

L. J. Osterweil, D. Rombach, and M. L. Soffa, "ICSE '06: Proceedings of the 28th international conference on software engineering," in *ICSE '06: Proceedings of the 28th international conference on Software engineering*. New York, NY, USA: ACM, 2006.

Quick-R. (2009, December) Nonparametric bootstrapping. www. [Online]. Available: http://www.statmethods.net/advstats/bootstrapping.html

R Documentation. (2009, December) The normal distribution. www. [Online]. Available: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/Normal.html

S. M. Ross, *Statistik für Ingenieure und Naturwissenschaftler*, 3rd ed. Spektrum Akademischer Verlag, Mar. 2006.

W. Schäfer, M. B. Dwyer, and V. Gruhn, "ICSE '08: Proceedings of the 30th international conference on software engineering," in *ICSE '08: Proceedings of the 30th international conference on Software engineering*. New York, NY, USA: ACM, 2008.

W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Houghton Mifflin, 2002.

S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965. [Online]. Available: http://www.jstor.org/stable/2333709

D. I. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal, "A survey of controlled experiments in software engineering," *IEEE Transactions on Software Engineering*, vol. 31, no. 9, pp. 733–753, 2005. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TSE.2005.97

StatPac Inc. (2009, November) The statistics calculator - statistical analysis tests at your fingertips. www. [Online]. Available: http://www.statpac.com/statistics-calculator/correlation-regression.htm

S. S. Stevens, "On the theory of scales of measurement," *Science,New Series*, vol. 103, no. 2684, pp. 677–680, jun 1946.

L. Thompson. (2009, December) Bootstrap confidence intervals. www. [Online]. Available: https://home.comcast.net/~lthompson221/5537/Bootstrap_CIs.html

UCLA: Academic Technology Services, Statistical Consulting Group. (2009, December) Introduction to sas. www. [Online]. Available: http://www.ats.ucla.edu/stat/R/library/bootstrap.htm

L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations." *American Psychologist*, vol. 54, no. 8, pp. 594–604, August 1999. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=amp-54-8-594&site=ehost-live

M. V. Zelkowitz and D. Wallace, "Experimental validation in software engineering," *Information and Software Technology*, vol. 39, pp. 735–743, 1997.

Anhang

.1 Tabellarische Auflistung der Variablen

Variablenname	*Typ*
id	string
article	string
journal	enum(“EMSE”, “ICSE”, “TSE”)
year	integer
expType	enum(“randomized”,“quasi”,“unknown”)
replication	boolean
variablesN	integer
rawDataAvailable	enum(“article”, “reference”, “unavailable”)
rawDataResponse	enum(“data”, “excuse”, “none”)
rawDataFormat	enum(“xls”, “doc”, “csv”)
rawDataMachineReadable	boolean
rawDataDoc	string
softwareName	string
softwareVersion	string

Tabelle 1: Erfasste Variablen bzgl. der Artikel

Variablenname	Typ
id	string
journal	enum(“EMSE”, “ICSE”, “TSE”)
year	integer
meta	boolean
var	string
scale	enum(“nominal”, “ordinal”, “difference”, “rational”)
levelsN	integer
descriptiveMean	boolean
descriptiveStandardDev	boolean
descriptiveMedian	boolean
descriptiveQuartiles	boolean
descriptiveMin	boolean
descriptiveMax	boolean
graphicsTable	boolean
graphicsBox	boolean
Ci	enum(67, 95)
powerAnalysis	enum(“priori”, “post”, “none”)
effectSizeCat	enum(“standardized”, “unstandardized”, “both”, “none”)
effectSizeStdFamily	enum(“r-family”, “d-family”)
effectSizeStdName	string
effectSizeNstdType	enum(“percentage”, “mean difference”)
ttestUsed	boolean
ttestAssumptions	enum(“plot”, “test”, “explanation”,“none”)
ttestPvalue	float
ttestSignificant	boolean
ttestAlpha	float
AnovaUsed	boolean
AnovaAssumptions	enum(“plot”, “test”, “explanation”,“none”)
AnovaNway	integer
AnovaInteractionPvalue	float
AnovaFactorsSignificant	boolean
AnovaFactor1	string
AnovaPvalue1	float
AnovaFactor1Significant	boolean
AnovaFactor2	string
AnovaPvalue2	float
AnovaFactor2Significant	boolean
AnovaFactor3	string
AnovaPvalue3	float
AnovaFactor3Significant	boolean
AnovaAlpha	float

Tabelle 2: Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 1)

Variablenname	Typ
AncovaUsed	boolean
AncovaFactorsN	integer
AncovaPvalue1	float
AncovaFactor1Significant	boolean
AncovaPvalue2	float
AncovaFactor2Significant	boolean
AncovaPvalue3	float
AncovaFactor3Significant	boolean
AncovaAlpha	float
MannWhitneyUsed	boolean
MannWhitneyPvalue	float
MannWhitneySignificant	boolean
MannWhitneyAlpha	float
WilcoxonUsed	boolean
WilcoxonPvalue	float
WilcoxonSignificant	boolean
WilcoxonAlpha	float
KruskalWallisUsed	boolean
KruskalWallisPvalue	float
KruskalWalliesSignificant	boolean
KruskalWallisAlpha	float
RegAnalysisUsed	boolean
RegAnalysisVars	integer
RegAnalysisCoefficients	boolean
RegAnalysisPvalue1	float
RegAnalysisParam1Significant	boolean
RegAnalysisPvalue2	float
RegAnalysisParam2Significant	boolean
RegAnalysisPvalue3	float
RegAnalysisParam3Significant	boolean
RegAnalysisAlpha	float
SpearmanCorrUsed	boolean
SpearmanCorrCoefficient	float
SpearmanCorrPvalue	float
SpearmanCorrSignificant	boolean
SpearmanCorrAlpha	float

Tabelle 3: Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 2)

Variablenname	Typ
KendallCorrUsed	boolean
KendallCorrCoefficient	float
KendallCorrPvalue	float
KendallCorrSignificant	boolean
KendallCorrAlpha	float
ChiSquaredUsed	boolean
ChiSquaredPvalue	float
ChiSquaredSignificant	boolean
ChiSquaredAlpha	float
McNemarUsed	boolean
McNemarPvalue	float
McNemarSignificant	boolean
McNemarAlpha	float
FisherExactUsed	boolean
FisherExactPvalue	float
FisherExactSignificant	boolean
FisherExactAlpha	float
PermutationUsed	boolean
PermutationPvalue	float
PermutationSignificant	boolean
PermutationAlpha	float

Tabelle 4: Erfasste Variablen bzgl. der abhängigen Experimentvariablen (Teil 3)