

Claes Wohlin · Per Runeson
Martin Höst · Magnus C. Ohlsson
Björn Regnell · Anders Wesslén

Experimentation in Software Engineering

Sperimentazione nell'ingegneria del software

Claes Wohlin Per Runeson

Martin Host: Magnus C. Ohlsson

Bjorn Regnell " Anders Wesslen'

Sperimentazione dentro Ingegneria del software

123

Claes Wohlin
Scuola di Informatica
Istituto di tecnologia Blekinge
Karlskrona, Svezia

Per Runeson
Dipartimento di Informatica
Università di Lund
Lund, Svezia

Martin Host
Dipartimento di Informatica
Università di Lund
Lund, Svezia

Magnus C. Ohlsson
Verifica del sistema Svezia AB
Malmo, Svezia "

Bjorn Regnell
Dipartimento di Informatica
Università di Lund
Lund, Svezia

Anders Wesslen
ST-Ericsson AB
Lund, Svezia

ISBN 978-3-642-29043-5 DOI
10.1007/978-3-642-29044-2 Springer
Heidelberg New York Dordrecht Londra

ISBN 978-3-642-29044-2 (eBook)

Numerico di controllo della Biblioteca del Congresso: 2012940660

Codici ACM: D.2

© Springer science+business media (successore nell'interesse di Kluwer Academic Publishers, Boston)
2000, Springer-Verlag Berlino Heidelberg 2012 Quest'opera
è soggetta a copyright. Tutti i diritti sono riservati all'editore, sia che si tratti di tutto o di parte del materiale, in particolare i diritti di traduzione, ristampa, riutilizzo di illustrazioni, recitazione, trasmissione, riproduzione su microfilm o in qualsiasi altro modo fisico,
nonché trasmissione o archiviazione di informazioni e recupero, adattamento elettronico, software per computer o mediante
metodologia simile o dissimile ora nota o sviluppata in seguito. Sono esentati da questa riserva legale brevi estratti relativi a
recensioni o analisi accademiche o materiale fornito appositamente allo scopo di essere inserito ed eseguito su un sistema
informatico, ad uso esclusivo dell'acquirente dell'opera. La duplicazione di questa pubblicazione o di parti di essa è consentita solo
in base alle disposizioni della legge sul copyright del luogo in cui si trova l'editore, nella sua versione attuale, e l'autorizzazione
all'uso deve essere sempre ottenuta da Springer.

Le autorizzazioni per l'utilizzo possono essere ottenute tramite RightsLink presso il Copyright Clearance Center. Le violazioni sono
perseguibili ai sensi della rispettiva legge sul copyright.
L'utilizzo nella presente pubblicazione di nomi descrittivi generali, nomi registrati, marchi, marchi di servizio, ecc. non implica, anche
in assenza di specifica indicazione, che tali nomi siano esenti dalle leggi e dai regolamenti di tutela in materia e quindi liberi da usi
generalii. utilizzo.
Sebbene i consigli e le informazioni contenute in questo libro siano ritenute vere e accurate alla data di pubblicazione, né gli
autori, né i redattori, né l'editore possono accettare alcuna responsabilità legale per eventuali errori o omissioni che potrebbero
essere commessi. L'editore non fornisce alcuna garanzia, espresa o implicita, rispetto al materiale contenuto nel presente
documento.

Stampato su carta priva di acidi

Springer fa parte di Springer Science+Business Media (www.springer.com)

Prefazione

La sperimentazione è fondamentale per qualsiasi attività scientifica e ingegneristica.

Comprendere una disciplina implica costruire modelli dei vari elementi della disciplina, ad esempio, gli oggetti nel dominio, i processi utilizzati per manipolare tali oggetti, la relazione tra i processi e gli oggetti. L'evoluzione della conoscenza del dominio implica l'evoluzione di tali modelli testandoli tramite esperimenti di varie forme. L'analisi dei risultati dell'esperimento implica l'apprendimento, l'incapsulamento della conoscenza e la capacità di cambiare e affinare i nostri modelli nel tempo. Pertanto, la nostra comprensione di una disciplina si evolve nel tempo.

Questo è il paradigma che è stato utilizzato in molti campi, ad esempio fisica, medicina, industria manifatturiera. Questi campi si sono evoluti come discipline quando hanno iniziato ad applicare il ciclo di costruzione di modelli, sperimentazione e apprendimento. Ogni campo è iniziato con la registrazione delle osservazioni e si è evoluto fino alla manipolazione delle variabili del modello e allo studio degli effetti dei cambiamenti in tali variabili. I campi differiscono nella loro natura, in ciò che costituisce gli oggetti fondamentali del campo, nelle proprietà di quegli oggetti, nelle proprietà del sistema che li contiene, nella relazione degli oggetti con il sistema e nella cultura della disciplina. Queste differenze influiscono sul modo in cui vengono costruiti i modelli e su come viene eseguita la sperimentazione.

Come altre discipline scientifiche e ingegneristiche, l'ingegneria del software richiede il ciclo di costruzione di modelli, sperimentazione e apprendimento. Lo studio dell'ingegneria del software è una scienza di laboratorio. Gli attori della disciplina sono i ricercatori e i professionisti. Il ruolo del ricercatore è comprendere la natura dell'oggetto (prodotti), i processi che li creano e li manipolano e la relazione tra i due nel contesto del sistema. Il ruolo del professionista è quello di costruire sistemi "migliorati", utilizzando le conoscenze disponibili fino ad oggi. Questi ruoli sono simbiotici. Il ricercatore ha bisogno di laboratori per studiare i problemi affrontati dai professionisti e sviluppare ed evolvere soluzioni basate sulla sperimentazione. Il professionista deve capire come costruire sistemi migliori e il ricercatore può fornire modelli per aiutarlo.

Nello sviluppo di modelli e nella sperimentazione, sia il ricercatore che il professionista devono comprendere la natura della disciplina dell'ingegneria del software. Non tutti i software sono uguali: ci sono un gran numero di variabili che causano differenze e loro

gli effetti devono essere compresi. Come la medicina, dove la variazione nella genetica umana e nella storia medica è spesso un fattore importante nello sviluppo dei modelli e nell'interpretazione dei risultati dell'esperimento, l'ingegneria del software si occupa di contesti diversi che influenzano l'input e i risultati. Nell'ingegneria del software le tecnologie sono per lo più ad alta intensità umana piuttosto che automatizzate. Come nel settore manifatturiero, il problema principale è comprendere e migliorare la relazione tra i processi e i prodotti che creano. Ma a differenza della produzione, il processo nell'ingegneria del software è lo sviluppo e non la produzione. Quindi non possiamo raccogliere dati da ripetizioni esatte dello stesso processo. Dobbiamo costruire i nostri modelli a un livello di astrazione più elevato, ma facendo comunque attenzione a identificare le variabili di contesto.

Attualmente, esiste un insieme insufficiente di modelli che ci consentono di ragionare sulla disciplina, una mancanza di riconoscimento dei limiti delle tecnologie per determinati contesti e un'analisi e una sperimentazione insufficienti in corso, ma quest'ultima situazione sta migliorando, come evidenziato da questo libro di testo.

Questo libro rappresenta una pietra miliare nel consentirci di formare sia il ricercatore che il professionista nella sperimentazione dell'ingegneria del software. È un contributo importante al campo. Gli autori hanno accumulato un'incredibile raccolta di conoscenze e le hanno confezionate in modo eccellente, fornendo un processo per la definizione dell'ambito, la pianificazione, l'esecuzione, l'analisi, l'interpretazione e il confezionamento degli esperimenti. Coprono tutti gli argomenti necessari, dalle minacce alla validità alle procedure statistiche.

È ben scritto e copre un'ampia gamma di informazioni necessarie per eseguire esperimenti nell'ingegneria del software. Quando ho iniziato a fare esperimenti ho dovuto reperire varie fonti di informazione, quasi sempre provenienti da altre discipline, e adattarle come meglio potevo alle mie esigenze. Se avessi avuto questo libro ad aiutarmi, mi avrebbe risparmiato un'enorme quantità di tempo e fatica e i miei esperimenti probabilmente sarebbero stati migliori.

Professore Vittorio R. Basili

Prefazione

Sono onorato di essere stato invitato a scrivere una prefazione per questa revisione del libro degli autori con lo stesso titolo pubblicato nel 2000. Ho utilizzato l'edizione originale sin dalla sua pubblicazione come insegnante e ricercatore. Gli studenti dei miei corsi alla Colorado State University, alla Washington State University, all'Università di Denver e all'Universitaet Wuerzburg hanno utilizzato il libro nel corso degli anni. Alcuni erano dipendenti a tempo pieno presso importanti aziende che lavoravano a lauree in Ingegneria dei Sistemi, altri a tempo pieno Master e Ph.D. studenti. Il libro ha funzionato bene per loro. Oltre alla trattazione dei metodi sperimentali di ingegneria del software, ne hanno apprezzato la concisione. Sono lieto di vedere che la versione rivista è compatta e facile da lavorare come la prima.

Le aggiunte e le modifiche in questa versione rivista riflettono molto bene la maturazione del campo dell'ingegneria del software empirica da quando il libro è stato originariamente pubblicato: la crescente importanza della replicazione e della sintesi degli esperimenti, e la necessità di accademici e professionisti di trasferire con successo nuove tecnologie basate su prove quantitative convincenti. Un altro importante miglioramento riguarda il trattamento ampliato delle questioni etiche nella sperimentazione dell'ingegneria del software. Soprattutto perché in questo campo non esiste un codice etico formale, è di vitale importanza che gli studenti siano consapevoli di tali problemi e abbiano accesso a linee guida su come affrontarli.

L'edizione originale di questo libro enfatizzava gli esperimenti. Nell'industria, tuttavia, i casi di studio tendono ad essere più comuni per valutare la tecnologia, i processi di ingegneria del software o gli artefatti. Pertanto l'aggiunta di un capitolo sugli studi di casi è assolutamente necessaria e accolta con favore. Lo stesso vale per il capitolo sulle revisioni sistematiche della letteratura.

Avendo tenuto per una dozzina d'anni un popolare corso di ingegneria del software quantitativo con l'edizione originale, questa versione rivista con le sue aggiunte e aggiornamenti fornisce molti dei materiali che ho aggiunto separatamente nel corso degli anni. Anzi, lo fa senza perdere la compattezza e la concisione dell'edizione originale. Io, per esempio, sono entusiasta di questa versione rivista e continuerò a usarla come testo nei miei corsi e come risorsa per i miei studenti ricercatori.

La professoressa Anneliese Amschler Andrews

Prefazione dall'edizione originale

Sono convinto che gli ingegneri del software non solo debbano conoscere i metodi e i processi dell'ingegneria del software, ma che debbano anche sapere come valutarli.

Di conseguenza, ho insegnato i principi della sperimentazione e degli studi empirici come parte del curriculum di ingegneria del software. Fino ad ora, ciò significava selezionare un testo da un'altra disciplina, solitamente la psicologia, e integrarlo con articoli di giornale o conferenze che fornissero agli studenti esempi di esperimenti e studi empirici di ingegneria del software.

Questo libro colma un'importante lacuna nella letteratura sull'ingegneria del software: fornisce uno sguardo conciso e completo su un aspetto importante dell'ingegneria del software: l'analisi sperimentale del funzionamento dei metodi, delle metodologie e dei processi dell'ingegneria del software. Poiché tutti questi cambiamenti cambiano così rapidamente nel nostro campo, è importante sapere come valutarne di nuovi. Questo libro insegna come procedere e quindi è prezioso non solo per lo studente di ingegneria del software, ma anche per il professionista praticante dell'ingegneria del software che sarà in grado di farlo.

• Valutare le tecniche di ingegneria del software. •

Determinare il valore (o la mancanza di esso) delle affermazioni fatte nei confronti di un ingegnere del software-metodo o processo negli studi pubblicati.

Infine, questo libro costituisce una risorsa preziosa per il ricercatore di ingegneria del software.

Professoressa Anneliese Amschler Andrews (ex von Mayrhauser)

Prefazione

Hai mai avuto la necessità di confrontare metodi o tecniche di ingegneria del software l'uno con l'altro? Questo libro presenta la sperimentazione come un modo per valutare nuovi metodi e tecniche nell'ingegneria del software. Gli esperimenti sono strumenti preziosi per tutti gli ingegneri del software coinvolti nella valutazione e nella scelta tra diversi metodi, tecniche, linguaggi e strumenti.

Può darsi che tu sia un professionista del software, che desidera valutare metodi e tecniche prima di introdurli nella tua organizzazione. Potresti anche essere un ricercatore che desidera valutare i nuovi risultati della ricerca rispetto a qualcosa di esistente, al fine di ottenere un fondamento scientifico per le tue nuove idee. Potresti essere un insegnante che crede che la conoscenza degli studi empirici nell'ingegneria del software sia essenziale per i tuoi studenti. Infine, potresti essere uno studente di ingegneria del software che desidera apprendere alcuni metodi per trasformare l'ingegneria del software in una disciplina scientifica e ottenere dati quantitativi confrontando diversi metodi e tecniche. Questo libro fornisce linee guida ed esempi di come dovrresti procedere per avere successo nella tua missione.

Ingegneria e scienza del software

Il termine “ingegneria del software” è stato coniato nel 1968 e l’area è ancora in fase di maturazione. Nel corso degli anni l’ingegneria del software è stata guidata dallo sviluppo tecnologico e dalla ricerca di advocacy. Quest’ultimo si riferisce al fatto che negli anni abbiamo inventato e introdotto nuovi metodi e tecniche basati sul marketing e sulla convinzione piuttosto che sui risultati scientifici. In una certa misura, ciò è comprensibile considerando il ritmo con cui la società dell’informazione si è affermata negli ultimi vent’anni. Tuttavia, a lungo termine non è accettabile se vogliamo avere il controllo del software che sviluppiamo. Il controllo deriva dalla capacità di valutare nuovi metodi, tecniche, linguaggi e strumenti prima di utilizzarli. Inoltre, questo ci aiuterebbe a trasformare l’ingegneria del software in una disciplina scientifica. Prima di esaminare le questioni che dobbiamo affrontare per trasformare l’ingegneria del software in scienza, esaminiamo il modo in cui la scienza viene vista in altri settori.

In "L'ultimo teorema di Fermat" del Dr. Simon Singh, [160], si discute di scienza. Il succo della discussione può essere riassunto come segue. Nella scienza, i fenomeni fisici vengono affrontati avanzando ipotesi. Il fenomeno viene osservato e se le osservazioni sono in linea con l'ipotesi, questa diventa prova dell'ipotesi. L'intenzione è anche che l'ipotesi consenta di prevedere altri fenomeni. Gli esperimenti sono importanti per verificare l'ipotesi e in particolare la capacità predittiva dell'ipotesi. Se i nuovi esperimenti supportano l'ipotesi, allora abbiamo più prove a favore dell'ipotesi.

Man mano che le prove crescono e diventano forti, l'ipotesi può essere accettata come teoria scientifica.

La sintesi mira fondamentalmente a verificare le ipotesi attraverso la ricerca empirica.

Questo potrebbe non essere il modo in cui oggi viene condotta la maggior parte della ricerca nell'ingegneria del software. Tuttavia, la necessità di valutare e validare nuove proposte di ricerca conducendo studi empirici è riconosciuta in misura maggiore oggi rispetto a 10 anni fa.

Gli studi empirici comprendono indagini, esperimenti e studi di casi. Pertanto, l'obiettivo di questo libro è introdurre e promuovere l'uso di studi empirici nell'ingegneria del software con un'enfasi particolare sulla sperimentazione.

Scopo

Lo scopo del libro è introdurre studenti, insegnanti, ricercatori e professionisti alla sperimentazione e alla valutazione empirica con particolare attenzione all'ingegneria del software. L'obiettivo è in particolare quello di fornire linee guida su come eseguire esperimenti per valutare metodi, tecniche e strumenti nell'ingegneria del software, sebbene siano fornite brevi introduzioni anche per altri approcci empirici.

L'introduzione alla sperimentazione avviene attraverso una prospettiva di processo.

L'attenzione si concentra sui passaggi che dobbiamo compiere per eseguire un esperimento. Il processo può essere generalizzato ad altri tipi di studi empirici, ma in questo caso l'attenzione principale è rivolta agli esperimenti e ai quasi-esperimenti.

La motivazione per questo libro deriva dal bisogno di supporto che abbiamo sperimentato quando abbiamo reso la nostra ricerca sull'ingegneria del software più sperimentale. Sono disponibili diversi libri che trattano l'argomento in termini molto generali o si concentrano su alcune parti specifiche della sperimentazione; la maggior parte di essi si concentra sui metodi statistici nella sperimentazione. Questi sono importanti, ma mancano libri che approfondiscano la sperimentazione da una prospettiva di processo. Inoltre, ci sono pochi libri che affrontano la sperimentazione nell'ingegneria del software in particolare, e in realtà nessun libro quando è stata pubblicata l'edizione originale di questo libro.

Ambito

Lo scopo del libro sono principalmente gli esperimenti di ingegneria del software come mezzo per valutare metodi, tecniche, ecc. Il libro fornisce alcune informazioni

riguardanti studi empirici in generale, inclusi studi di casi, revisioni sistematiche della letteratura e indagini. L'intenzione è quella di fornire una breve comprensione di queste strategie e in particolare di metterle in relazione con la sperimentazione.

I capitoli del libro coprono diversi passaggi da seguire per eseguire esperimenti di ingegneria del software. Inoltre, nel libro vengono forniti esempi di studi empirici relativi all'ingegneria del software. È di particolare importanza illustrare agli ingegneri del software che gli studi empirici e la sperimentazione possono essere praticati con successo nell'ingegneria del software. Nel libro sono inclusi due esempi di esperimenti. Questi vengono introdotti per illustrare il processo dell'esperimento e per esemplificare come possono essere riportati gli esperimenti di ingegneria del software. L'intenzione è che questi studi funzionino come buoni esempi e fonti di ispirazione per ulteriore lavoro empirico nell'ingegneria del software. Il libro si concentra principalmente sugli esperimenti, ma va ricordato che sono disponibili anche altre strategie, ad esempio casi di studio e sondaggi. In altre parole, non dobbiamo ricorrere alla ricerca di advocacy e al marketing senza dati quantitativi quando sono disponibili strategie di ricerca come, ad esempio, gli esperimenti.

Pubblico target

Il pubblico a cui è rivolto il libro può essere suddiviso in quattro categorie.

Gli studenti possono utilizzare il libro come introduzione alla sperimentazione nell'ingegneria del software con particolare attenzione alla valutazione. Il libro è adatto come libro di testo per studi universitari o universitari in cui viene sottolineata la necessità di studi empirici nell'ingegneria del software. Nel libro sono inclusi esercizi e compiti di progetto per combinare il materiale più teorico con alcuni aspetti pratici.

Gli insegnanti possono utilizzare il libro nelle loro classi se credono nella necessità di rendere l'ingegneria del software più empirica. Il libro è adatto come introduzione alla zona. Dovrebbe essere abbastanza autonomo, anche se è consigliato un corso introduttivo di statistica.

I ricercatori possono utilizzare il libro per saperne di più su come condurre studi empirici e utilizzarli come un ingrediente importante nella loro ricerca. Inoltre, l'obiettivo è che possa essere fruttuoso ritornare al libro e usarlo come lista di controllo quando si effettuano ricerche empiriche.

I professionisti possono utilizzare il libro come un "libro di ricette" quando valutano alcuni nuovi metodi o tecniche prima di introdurli nella loro organizzazione. Ci si aspetta che i professionisti imparino come utilizzare gli studi empirici nel loro lavoro quotidiano quando cambiano, ad esempio, il processo di sviluppo nell'organizzazione in cui lavorano.

Contorno

Il libro è diviso in tre parti principali. La struttura del libro è riassunta nella Tabella 1, che mostra anche una mappatura all'edizione originale di questo libro.

La prima parte fornisce un'introduzione generale all'area degli studi empirici nel Cap. 1. Colloca gli studi empirici in generale e gli esperimenti in particolare nel contesto dell'ingegneria del software. Nel cap. 2, le strategie empiriche (ricerche, studi di casi ed esperimenti) vengono discusse in generale e viene elaborato il contesto degli studi empirici, in particolare dal punto di vista dell'ingegneria del software. Il capitolo 3 fornisce una breve introduzione alla teoria e alla pratica della misurazione. Nel cap. 4 forniamo una panoramica su come condurre revisioni sistematiche della letteratura, per sintetizzare i risultati di diversi studi empirici. Il capitolo 5 fornisce una panoramica dei casi di studio come tipo correlato di studi empirici. Nel cap. 6, l'attenzione è posta sulla sperimentazione introducendo il processo sperimentale generale.

La Parte II ha un capitolo per ogni fase dell'esperimento. Il Capitolo 7 discute come definire l'ambito di un esperimento, mentre il Cap. 8 si concentra sulla fase di pianificazione. Il funzionamento dell'esperimento è discusso nei capp. 9 e 10 presentano alcuni metodi per analizzare e interpretare i risultati. Il capitolo 11 discute la presentazione e il confezionamento dell'esperimento.

La Parte III contiene due esperimenti di esempio. Nel cap. 12, viene presentato un esempio in cui l'obiettivo principale è illustrare il processo dell'esperimento, e l'esempio nel cap. 13 viene utilizzato per illustrare come un esperimento di ingegneria del software può essere riportato in un articolo.

Alcuni esercizi e dati sono presentati nell'Appendice A. Infine, il libro mostra alcune tabelle statistiche nell'Appendice B. Le tabelle sono incluse principalmente per fornire supporto ad alcuni degli esempi contenuti nel libro. Tabelle più complete sono disponibili nella maggior parte dei libri di statistica.

Esercizi

Gli esercizi sono divisi in quattro categorie, la prima presentata alla fine di ogni capitolo nelle Parti I e II del libro (Capitoli 1–11), e le altre tre nell'Appendice A:

Comprensione. Alla fine di ogni capitolo vengono fornite cinque domande che catturano i punti più importanti. L'obiettivo è garantire che il lettore abbia compreso i concetti più importanti.

Formazione. Questi esercizi offrono l'opportunità di praticare la sperimentazione. Gli esercizi sono particolarmente mirati all'analisi dei dati e alla risposta a domande relative a un esperimento.

Revisione. Questo esercizio è mirato agli esempi di esperimenti presentati nei capp. 12–13. L'obiettivo è quello di dare l'opportunità di rivedere alcuni esperimenti presentati. Dopo aver letto diversi esperimenti presentati in letteratura, lo è

Tabella 1 Struttura del libro

Soggetto	Aggiornamenti principali originali rivisti edizione della versione		
Parte I. Contesto			
Introduzione	1	1	
Strategie empiriche	2	2	Nuove sezioni sulla replica, sintesi, tecnologia trasferimento ed etica
Misurazione	3	3	Nuova sezione sulla misurazione in pratica
Revisioni sistematiche della letteratura	4	10a	Nuovo capitolo
Casi di studio	5		Nuovo capitolo
Processo dell'esperimento	6	4	
Parte II. Fasi del processo sperimentale			
Definizione dell'ambito	7	5b	Nuovo esempio funzionante
	8	6	Terminologia adattata
Pianificazione	9	7	
Operazione Analisi e interpretazione	10	8	
Presentazione e pacchetto	11	9	Revisione importante
Parte III. Esperimenti di esempio			
Illustrazione del processo dell'esperimento	12	11	
Le prospettive sono davvero diverse? 13			Nuovo capitolo
Appendici			
Esercizi	UN	13	Esercizi di comprensione spostati a ciascun capitolo
Tabelle statistiche	B	UN	

^{UN} Intitolato Sondaggio e con una portata diversa^B Definizione dal titolo

chiaro che la maggior parte degli esperimenti soffre di alcuni problemi. Ciò è dovuto principalmente a ereditare i problemi legati all'esecuzione della sperimentazione nell'ingegneria del software. Invece di promuovendo la critica del lavoro altrui, abbiamo fornito alcuni esempi di studi che abbiano condotto noi stessi. Sono, a nostro avviso, rappresentativi del tipo di esperimenti pubblicati in letteratura. Ciò include quello che hanno i loro punti di forza e di debolezza.

Incarichi. L'obiettivo di questi esercizi è illustrare come possono farlo gli esperimenti essere utilizzati nella valutazione. Questi incarichi sono esempi di studi che possono essere svolti all'interno di un corso, presso un'università o nell'industria. Sono deliberatamente mirati a problemi che possono essere risolti con esperimenti abbastanza semplici. Gli incarichi può essere svolto dopo aver letto il libro oppure è possibile svolgere uno dei compiti fuori mentre il libro viene letto. Quest'ultimo offre l'opportunità di esercitarsi durante la lettura i capitoli. In alternativa, vorremmo raccomandare agli insegnanti di formulare un compito, nell'ambito della loro area di competenza, che può essere utilizzato in tutto il libro per esemplificare i concetti presentati in ogni capitolo.

Ringraziamenti

Questo libro è basato su "Experimentation in Software Engineering: An Introduction", pubblicato nel 2000. Questa nuova versione è sia una revisione che un'estensione del libro precedente. Abbiamo rivisto parti del libro, ma abbiamo anche aggiunto nuovo materiale, ad esempio, riguardante revisioni sistematiche della letteratura e ricerche di casi di studio.

Un libro non è quasi mai solo una conquista degli autori. Il supporto e l'aiuto di diverse persone tra cui famiglie, amici, colleghi, ricercatori internazionali sul campo e organizzazioni finanziarie sono spesso un prerequisito per un nuovo libro.

Questo libro non fa certamente eccezione. In particolare, vorremmo esprimere la nostra sincera gratitudine ai lettori di "Experimentation in Software Engineering: An Introduction". Il tuo utilizzo del libro è stato una grande fonte di ispirazione e una motivazione per pubblicare la versione attuale. In particolare, vorremmo ringraziare il signor Alan Kelon Oliveira de Moraes, Universidade Federal de Pernambuco, Brasile per aver inviato l'e-mail che ha effettivamente dato il via alla revisione del libro. Inoltre, vorremmo ringraziare le seguenti persone per aver contribuito alla realizzazione del libro.

Innanzitutto desideriamo ringraziare il primo principale utilizzatore esterno del libro, il Prof. Giuseppe Visaggio, Università di Bari in Italia, per aver adottato la bozza di questo libro in uno dei suoi corsi e per aver fornito preziosi feedback. Vorremmo anche esprimere la nostra gratitudine alla Prof.ssa Anneliese Andrews, Università di Denver, USA e al Dr. Khaled El Emam, Università di Ottawa, Canada per averci incoraggiato a pubblicare il libro e per i preziosi commenti. Desideriamo inoltre ringraziare il Dott. Lionel Briand, Università del Lussemburgo, Lussemburgo; Il Dr. Christian Bunse, Università di Mannheim, Germania e il Dr. John Daly già presso l'Istituto Fraunhofer per l'ingegneria del software sperimentale, Kaiserslautern, Germania, per aver fornito i dati per l'esempio sulla progettazione orientata agli oggetti. I nostri ringraziamenti vanno anche al Dr. Thomas Thelin per averci permesso di includere un esperimento che ha fatto insieme a due degli autori del libro.

Le prime bozze del libro sono state utilizzate e valutate internamente all'interno del Gruppo di ricerca sull'ingegneria del software presso l'Università di Lund. Pertanto, vorremmo ringraziare i membri del gruppo per aver fornito feedback sulle diverse bozze del libro.

In particolare, vorremmo ringraziare il Dr. Lars Brathall per aver dedicato del tempo alla revisione

il manoscritto in modo molto approfondito e fornendo preziosi commenti. Desideriamo inoltre ringraziare i revisori anonimi per il loro contributo al libro.

Per la versione attuale del libro, abbiamo ricevuto preziosi input e proposte di miglioramento, e pertanto vorremmo ringraziare le seguenti persone per il loro prezioso contributo: Prof. Anneliese Andrews, Denver University, USA; il prof.

David Budgen, Università di Durham, Regno Unito; Prof. Barbara Kitchenham, Keele University, Regno Unito; Prof. Dieter Rombach e Moinul Islam, Università di Kaiserslautern, Germania; Prof. J'urgen Borstler, Dr. Samuel Fricker e Dr. Richard Torkar, Blekinge Institute of Technology, Svezia. Grazie anche al Sig. Jesper Runeson per il lavoro sulla trasformazione LATEX del libro.

Oltre alle persone di cui sopra, vorremmo anche ringraziare tutti i membri dell'ISERN (International Software Engineering Research Network) per la discussione interessante e illuminante riguardante la ricerca empirica sull'ingegneria del software in generale.

Per il capitolo sui casi di studio, siamo grati per il feedback alle liste di controllo da parte dei membri ISERN e dei partecipanti della Scuola Internazionale Avanzata di Ingegneria del Software Empirico nel settembre 2007. Un ringraziamento speciale al Dr. Kim Weyns e al Dr. Andreas Jedlitschka per il loro lavoro revisione di una prima bozza del capitolo.

Nel corso degli anni hanno contribuito al libro numerosi progetti di ricerca presso l'Università di Lund e il Blekinge Institute of Technology. Diverse sovvenzioni hanno finanziato progetti di ricerca in cui gli studi empirici hanno rappresentato una pietra miliare, e quindi hanno contribuito a plasmare la nostra esperienza che abbiamo cercato di documentare attraverso il libro. Questo libro è in una certa misura il risultato di tutti questi progetti di ricerca.

Prof. Claes Wohlin, Prof.

Per Runeson, Prof.

Martin Host, " Dr.

Magnus C. Ohlsson, Prof.

Bjorn Regnell e Dr. Anders
Wesslen'

Contenuto

Parte I Contesto

1 Introduzione	3
1.1 Contesto dell'ingegneria del software	3
1.2 Scienza e ingegneria del software	5
1.3 Esercizi	8
2 Strategie empiriche.....	9
2.1 Panoramica delle strategie empiriche	10
2.2 Sondaggi	12
2.2.1 Caratteristiche del sondaggio	12
2.2.2 Scopi dell'indagine	13
2.2.3 Raccolta dati	13
2.3 Casi di studio.....	14
2.3.1 Disposizioni di casi di studio	15
2.3.2 Fattori confondenti e altri aspetti	15
2.4 Esperimenti.....	16
2.4.1 Caratteristiche	17
2.4.2 Processo dell'esperimento.....	18
2.5 Confronto di strategie empiriche	18
2.6 Repliche	19
2.7 Teoria nell'ingegneria del software	21
2.8 Evidenza aggregata da studi empirici.....	22
2.9 Empirismo nel contesto dell'ingegneria del software	24
2.9.1 Valutazione empirica dei cambiamenti di processo	24
2.9.2 Paradigma del miglioramento della qualità	26
2.9.3 Fabbrica dell'esperienza	27
2.9.4 Obiettivo/Domanda/Metodo metrico	29
2.10 Trasferimento tecnologico su base empirica	30
2.11 Etica nella sperimentazione	33
2.12 Esercizi	36

3 Misurazione.....	37
3.1 Concetti fondamentali	38
3.1.1 Tipi di bilancia.....	39
3.1.2 Misure oggettive e soggettive	40
3.1.3 Misure dirette o indirette 41
3.2 Misure nell'ingegneria del software	41
3.3 Misurazioni in pratica 42
3.4 Esercizi	43
4 Revisioni sistematiche della letteratura	45
4.1 Pianificazione della revisione	45
4.2 Conduzione della revisione	46
4.3 Reporting della revisione	51
4.4 Studi di mappatura	52
4.5 Esempi di recensioni	52
4.6 Esercizi	54
5 Casi di studio	55
5.1 Casi di studio nel suo contesto	56
5.1.1 Perché casi di studio nell'ingegneria del software?....	57
5.1.2 Processo di ricerca sul caso di studio	58
5.2 Progettazione e Pianificazione	58
5.2.1 Pianificazione del caso di studio	58
5.2.2 Protocollo del caso di studio	60
5.3 Preparazione e raccolta dei dati	61
5.3.1 Interviste	62
5.3.2 Osservazioni	64
5.3.3 Dati di archivio	65
5.3.4 Metriche	65
5.4 Analisi dei dati	65
5.4.1 Analisi quantitativa dei dati	65
5.4.2 Analisi qualitativa dei dati	66
5.4.3 Validità	68
5.5 Segnalazione	69
5.5.1 Caratteristiche	69
5.5.2 Struttura	71
5.6 Esercizi 72
6 Processo dell'esperimento	73
6.1 Variabili, Trattamenti, Oggetti e Soggetti	74
6.2 Processo.....	76
6.3 Panoramica	81
6.4 Esercizi	81

Parte II Fasi del processo sperimentale

7 Ambito	85
7.1 Esperimento nell'ambito.....	85
7.2 Esempio di esperimento.....	87
7.3 Esercizi	88
8 Pianificazione	89
8.1 Selezione del contesto	89
8.2 Formulazione di ipotesi	91
8.3 Selezione delle variabili	92
8.4 Selezione dei soggetti.....	92
8.5 Progettazione dell'esperimento	93
8.5.1 Scelta del disegno dell'esperimento	93
8.5.2 Principi generali di progettazione	94
8.5.3 Tipi di progettazione standard	95
8.6 Strumentazione	101
8.7 Valutazione della validità	102
8.8 Descrizione dettagliata delle minacce alla validità.....	104
8.8.1 Validità delle conclusioni	104
8.8.2 Validità interna	106
8.8.3 Validità di costrutto.....	108
8.8.4 Validità esterna	110
8.9 Priorità tra i tipi di minacce alla validità.....	111
8.10 Esempio di esperimento.....	112
8.11 Esercizi	116
9 Funzionamento	117
9.1 Preparazione	117
9.1.1 Impegnare i partecipanti	118
9.1.2 Preoccupazioni relative alla strumentazione	119
9.2 Esecuzione	120
9.2.1 Raccolta dati	120
9.2.2 Ambiente sperimentale	120
9.3 Convalida dei dati	121
9.4 Esempio di operazione	121
9.5 Esercizi	121
	121
	122
10 Analisi e interpretazione	123
10.1 Statistiche descrittive.....	123
10.1.1 Misure di Tendenza Centrale	124
10.1.2 Misure di dispersione.....	126
10.1.3 Misure di dipendenza	127
10.1.4 Visualizzazione grafica	128
10.2 Riduzione del set di dati	131

10.3 Verifica di ipotesi	132
10.3.1 Concetti di base	132
10.3.2 Test parametrici e non parametrici	135
10.3.3 Panoramica dei test.....	136
10.3.4 Prova t.....	138
10.3.5 Mann-Whitney	139
10.3.6 Test F	140
10.3.7 T-Test accoppiato	140
10.3.8 Wilcoxon	141
10.3.9 Prova dei segni.....	142
10.3.10 ANOVA (Analisi della varianza)	143
10.3.11 Kruskal-Wallis	144
10.3.12 Chi-2.....	145
10.3.13 Verifica Adeguatezza Modello	148
10.3.14 Trarre conclusioni.....	149
10.4 Esempio di analisi.....	150
10.5 Esercizi	151
11 Presentazione e pacchetto	153
11.1 Struttura del rapporto sull'esperimento	153
11.2 Esercizi	157
Parte III Esperimenti di esempio	
12 Illustrazione del processo dell'esperimento	161
12.1 Ambito	161
12.1.1 Definizione dell'obiettivo.....	161
12.1.2 Riepilogo dell'ambito	163
12.2 Pianificazione	163
12.2.1 Selezione del contesto	163
12.2.2 Formulazione di ipotesi	163
12.2.3 Selezione delle variabili.....	165
12.2.4 Selezione dei soggetti.....	165
12.2.5 Progettazione dell'esperimento	165
12.2.6 Strumentazione	166
12.2.7 Valutazione di validità	166
12.3 Funzionamento	167
12.3.1 Preparazione	167
12.3.2 Esecuzione	168
12.3.3 Convalida dei dati.....	168
12.4 Analisi e Interpretazione	169
12.4.1 Statistiche descrittive.....	169
12.4.2 Riduzione dei dati.....	172
12.4.3 Verifica di ipotesi.....	172
12.5 Riepilogo	173
12.6 Conclusione	174

13 Le prospettive sono davvero diverse? Ulteriore	
Sperimentazione sulla lettura dei requisiti basata su scenari.....	175
13.1 Introduzione	176
13.2 Lavori correlati	177
13.3 Domande di ricerca	181
13.4 Pianificazione dell'esperimento.....	182
13.4.1 Variabili.....	182
13.4.2 Ipotesi	182
13.4.3 Progettazione	184
13.4.4 Minacce alla validità	184
13.5 Funzionamento dell'esperimento	186
13.6 Analisi dei dati	187
13.6.1 Prestazioni individuali per prospettive diverse....	187
13.6.2 Difetti rilevati da diverse prospettive	189
13.6.3 La dimensione del campione è sufficientemente grande?	192
13.6.4 Esperienza dei soggetti	194
13.7 Interpretazioni dei risultati	194
13.8 Riepilogo e Conclusioni	195
13.9 Dati sulla performance individuale	197
13.10 Dati sui difetti riscontrati da Perspectives.....	198
13.10.1 Documento PG	198
13.10.2 Documento ATM	199

Appendici

A Esercizi	203
A.1 Formazione.....	203
A.1.1 Dati normalmente distribuiti A.1.2	204
Esperienza	204
A.1.3 Programmazione	205
A.1.4 Progettazione	209
A.1.5 Ispezioni	212
A.2 Revisione	212
A.3 Compiti.....	213
A.3.1 Unit test e revisioni del codice	214
A.3.2 Metodi di ispezione A.3.3	214
Notazione dei requisiti	215
B Tavole statistiche	217
Riferimenti.....	223
Indice	233

Parte I
Sfondo

Capitolo 1

Introduzione

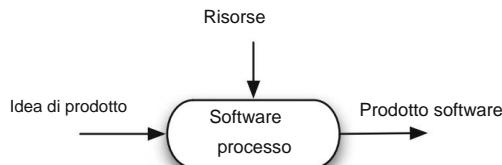
La rivoluzione tecnologica dell'informazione ha fatto sì, tra l'altro, che il software diventi parte integrante di un numero sempre maggiore di prodotti. Il software si trova in prodotti che vanno dai tostapane alle navette spaziali. Ciò significa che una grande quantità di software è stata e viene sviluppata. Lo sviluppo del software non è affatto facile; è un processo altamente creativo. La rapida crescita del settore ha fatto sì che numerosi progetti software incontrassero problemi in termini di funzionalità mancanti, superamento dei costi, mancato rispetto delle scadenze e scarsa qualità. Questi problemi o sfide furono identificati già negli anni '60 e nel 1968 fu coniato il termine "ingegneria del software" con l'intento di creare una disciplina ingegneristica incentrata sullo sviluppo di sistemi ad alta intensità di software.

L'ingegneria del software è formalmente definita dall'IEEE [84] come "*l'ingegneria del software significa l'applicazione di un approccio sistematico, disciplinato e quantificabile allo sviluppo, al funzionamento e alla manutenzione del software*". L'ingegneria del software in generale è presentata e discussa in libri come, ad esempio, da Sommerville [163] e Pfeleger e Atlee [134]. L'obiettivo qui è presentare come gli studi empirici e gli esperimenti in particolare si inseriscono nel contesto dell'ingegneria del software. Tre aspetti della definizione di cui sopra sono di particolare importanza qui. Innanzitutto, implica un processo software che punta a diverse fasi del ciclo di vita; in secondo luogo, sottolinea la necessità di un approccio sistematico e disciplinato; infine, si evidenzia l'importanza della quantificazione. L'uso di studi empirici è correlato a tutti e tre. Il contesto dell'ingegneria del software è ulteriormente discusso nella Sez. 1.1. La necessità di rendere l'ingegneria del software più scientifica e il ruolo importante svolto dagli studi empirici in questo contesto sono discussi nella sez. 1.2.

1.1 Contesto dell'ingegneria del software

Un modello di processo software viene utilizzato per descrivere i passaggi da eseguire e le attività da svolgere durante lo sviluppo del software. Esempi di modelli di processo software sono il modello a cascata, lo sviluppo incrementale, lo sviluppo evolutivo, la spirale

Fig. 1.1 Un'illustrazione del processo software



modello e diversi approcci agili allo sviluppo del software. Questi e altri modelli sono discussi nella letteratura generale sull'ingegneria del software. Una vista semplicistica del processo software è mostrata in Fig. 1.1. Va notato che il processo è cruciale sia che lavoriamo con lo sviluppo di un nuovo prodotto o con la manutenzione di un prodotto esistente.

Nella Figura 1.1, un'idea e risorse, principalmente sotto forma di persone, sono input per il processo software e le persone sviluppano un prodotto software attraversando le diverse fasi ed eseguendo le diverse attività nel processo software.

Lo sviluppo di prodotti software è molte volte un compito complesso. I progetti software possono durare un lungo periodo di tempo e coinvolgere molte persone (anche se si utilizzano metodi agili), a causa della complessità dei prodotti software sviluppati.

Ciò implica che spesso anche il processo software diventa molto complesso. Consiste in molte attività diverse e molti documenti vengono scritti prima che il prodotto finale possa essere consegnato. La complessità del processo software significa che è difficile ottimizzarlo o addirittura trovare un processo sufficientemente buono. Pertanto, è importante che le aziende si impegnino a migliorare il proprio modo di fare business se intendono rimanere competitive. Ciò significa che la maggior parte delle aziende cerca continuamente di migliorare i propri processi software al fine di migliorare i prodotti, ridurre i costi e così via. Il processo software sottolinea la necessità di un approccio sistematico e disciplinato al lavoro.

Essere agili non fa eccezione, c'è ancora la necessità di avere un approccio strutturato, anche se i metodi agili sottolineano la necessità di non documentare troppo ed enfatizzano la necessità di avere codice in esecuzione in modo continuo anziché "solo" alla fine di un grande progetto. È necessario anche un approccio sistematico e disciplinato quando si migliora il processo software, e quindi è necessario un modo per migliorare il processo.

Un esempio di processo di miglioramento su misura per lo sviluppo di software è il Quality Improvement Paradigm (QIP), definito da Basili [7]. Consiste in diverse fasi per supportare un approccio sistematico e disciplinato al miglioramento. Il QIP è presentato brevemente nella Sez. 2.9.2. Un processo di miglioramento più generale è il noto ciclo Pianifica/Fai/Studio/Agisci [23, 42]. I processi di miglioramento comprendono due attività, anche se non sempre viene utilizzata la stessa terminologia, che desideriamo evidenziare:

- Valutazione del processo software. •

Valutazione di una proposta di miglioramento del processo software.

La valutazione è condotta per individuare aree di miglioramento idonee. Esistono diversi modelli per valutare il processo software. Il più noto è probabilmente il Capability Maturity Model (CMM) del Software Engineering Institute

alla Carnegie-Mellon University, USA [33, 130]. I modelli di valutazione aiutano a individuare dove sono necessari miglioramenti. CMM ha cinque livelli di maturità con le cosiddette aree di processo chiave su ciascun livello. Si raccomanda alle aziende di concentrarsi sulle aree di miglioramento in base al loro livello di maturità.

Supponendo che sia possibile identificare le aree di miglioramento attraverso una qualche forma di valutazione, il passo successivo è determinare come queste aree di miglioramento possano essere affrontate per far fronte ai problemi identificati. Ad esempio, se vengono rilevati troppi difetti nei test di sistema, potrebbe essere possibile migliorare i test precedenti, le ispezioni o anche parti specifiche dello sviluppo, ad esempio la progettazione del software. L'obiettivo è che la valutazione della situazione attuale e la conoscenza dello stato dell'arte portino all'identificazione di proposte concrete di miglioramento del processo.

Una volta individuate le proposte di miglioramento è necessario stabilire quali introdurre, se presenti. Spesso non è possibile modificare semplicemente il processo software esistente senza avere maggiori informazioni sull'effetto reale della proposta di miglioramento. In altre parole, è necessario valutare le proposte prima di apportare modifiche importanti.

Un problema che si pone è che una proposta di miglioramento del processo è molto difficile da valutare senza il diretto coinvolgimento umano. Per un prodotto è possibile costruire prima un prototipo per valutare se è qualcosa su cui lavorare ulteriormente. Per un processo non è possibile costruire un prototipo. È possibile effettuare simulazioni e confrontare diversi processi, ma è bene ricordare che si tratta pur sempre di una valutazione basata su un modello. L'unica vera valutazione di un processo o di una proposta di miglioramento del processo è che le persone lo utilizzino, poiché il processo è solo una descrizione finché non viene utilizzato dalle persone. Gli studi empirici sono cruciali per la valutazione dei processi e delle attività umane. È inoltre utile utilizzare studi empirici quando è necessario valutare l'uso di prodotti o strumenti software. La sperimentazione fornisce un modo sistematico, disciplinato, quantificabile e controllato per valutare le attività basate sull'uomo. Questo è uno dei motivi principali per cui la ricerca empirica è comune nelle scienze sociali e comportamentali, si veda ad esempio Robson [144].

Inoltre, per i ricercatori nel campo dell'ingegneria del software sono importanti soprattutto gli studi empirici e gli esperimenti. Nuovi metodi, tecniche, linguaggi e strumenti non dovrebbero semplicemente essere suggeriti, pubblicati e commercializzati. È fondamentale valutare le nuove invenzioni e proposte rispetto a quelle esistenti. La sperimentazione offre questa opportunità e dovrebbe essere utilizzata di conseguenza. In altre parole, dovremmo utilizzare i metodi e le strategie disponibili quando conduciamo ricerche nell'ingegneria del software. Questo verrà ulteriormente discusso in seguito.

1.2 Scienza e ingegneria del software

L'ingegneria del software è una materia interdisciplinare. Si estende da questioni tecniche come database e sistemi operativi, attraverso questioni linguistiche, ad esempio sintassi e semantica, fino a questioni sociali e psicologia. Lo sviluppo del software è ad alta intensità umana; almeno oggi non siamo in grado di produrre nuovo software. È un

disciplina basata sulla creatività e sull'ingegno delle persone che lavorano nel settore.

Tuttavia, quando studiamo e facciamo ricerca nell'ingegneria del software, dovremmo mirare a trattarla come una disciplina scientifica. Ciò implica l'utilizzo di metodi scientifici per fare ricerca e quando si prendono decisioni riguardanti i cambiamenti nel modo in cui sviluppiamo il software.

Per svolgere ricerca scientifica nell'ingegneria del software, dobbiamo comprendere i metodi a nostra disposizione, i loro limiti e quando possono essere applicati. L'ingegneria del software deriva dalla comunità tecnica. Pertanto, è naturale esaminare i metodi utilizzati per la ricerca, ad esempio, nella progettazione dell'hardware e nella teoria della codifica, ma in base alla natura dell'ingegneria del software dovremmo esaminare anche altre discipline. Glass ha riassunto quattro metodi di ricerca nel campo dell'ingegneria del software [62]. Inizialmente furono presentati in un contesto di ingegneria del software da Basili [9]. I metodi sono:

Scientifico Il mondo viene osservato e sulla base dell'osservazione viene costruito un modello, ad esempio un modello di simulazione.

Ingegneria Si studiano le soluzioni attuali e si propongono modifiche, per poi valutarle.

Empirico Un modello viene proposto e valutato attraverso studi empirici, per esempio, casi di studio o esperimenti.

Analitica Viene proposta una teoria formale che viene poi confrontata con osservazioni empiriche.

Il metodo ingegneristico e il metodo empirico possono essere visti come variazioni del metodo scientifico [9].

Tradizionalmente, il metodo analitico viene utilizzato nelle aree più formali dell'ingegneria elettrica e dell'informatica, ad esempio nella teoria e negli algoritmi elettromagnetici. Il metodo scientifico viene utilizzato in ambiti applicativi, come la simulazione di una rete di telecomunicazioni per valutarne le prestazioni. Va tuttavia notato che la simulazione in quanto tale non viene applicata solo nel metodo scientifico. La simulazione può essere utilizzata anche come mezzo per condurre un esperimento. Probabilmente il metodo ingegneristico è dominante nell'industria.

Gli studi empirici sono stati tradizionalmente utilizzati nelle scienze sociali e in psicologia, dove non siamo in grado di stabilire alcuna legge della natura, come in fisica.¹ Nelle scienze sociali e in psicologia, si occupano del comportamento umano. L'osservazione importante, in questo contesto, è quindi che l'ingegneria del software è in gran parte governata dal comportamento umano attraverso le persone che sviluppano il software. Pertanto, non possiamo aspettarci di trovare regole o leggi formali nell'ingegneria del software, tranne forse quando ci concentriamo su aspetti tecnici specifici. Il focus di questo libro è sull'applicazione e l'utilizzo di studi empirici nell'ingegneria del software. L'obiettivo è in particolare quello di enfatizzare il processo che sta alla base dell'esecuzione di studi empirici in generale e di sperimentazione in particolare. Viene presentato un processo sperimentale,

¹Lehman [110] ha fatto riferimento alle leggi dell'evoluzione del software, ma questa nozione non è stata diffusa nei successivi lavori sulla teoria, vedere oltre la Sez. 2.7.

che evidenzia i passaggi fondamentali per eseguire esperimenti, fornisce linee guida su cosa fare ed esemplifica i passaggi utilizzando esempi di ingegneria del software.

Va notato che non si sostiene che i metodi analitici, scientifici e ingegneristici siano inappropriati per l'ingegneria del software. Sono necessari anche per l'ingegneria del software, ad esempio possiamo costruire modelli matematici per la crescita dell'affidabilità del software [116]. Inoltre, i metodi di ricerca non sono ortogonali e quindi potrebbe essere opportuno condurre uno studio empirico all'interno, ad esempio, del metodo ingegneristico. Il punto importante è che dovremmo fare un uso migliore dei metodi disponibili nell'ambito della ricerca empirica. Sono spesso utilizzati in altre discipline, ad esempio nelle scienze comportamentali, e la natura dell'ingegneria del software ha molto in comune con le discipline esterne alle parti tecniche dell'ingegneria.

Secondo Zendler [182], i primissimi esperimenti di ingegneria del software furono condotti alla fine degli anni '60 da Grant e Sackmann [69] sul lavoro di testing online e offline. Negli anni '70, alcuni pionieri condussero esperimenti sulla programmazione strutturata [115], sui diagrammi di flusso [151] e sul test del software [126].

La necessità di una sperimentazione sistematica nell'ingegneria del software fu sottolineata a metà degli anni '80 da Basili et al. [15]. Da allora sono stati pubblicati altri articoli che sottolineano la necessità dell'empirismo nell'ingegneria del software, si veda ad esempio il lavoro di Basili, Fenton, Glass, Kitchenham, Pfleeger, Pickard, Potts e Tichy [9,57,62,97,140,169]. La mancanza di prove empiriche nella ricerca sull'ingegneria del software è sottolineata da Tichy et al. [170], Zelowitz e Wallace [181] e Glass et al. [63]. Le ultime pubblicazioni indicano che la ricerca nell'ingegneria del software è ancora troppo una ricerca di advocacy [140]. È necessario un approccio più scientifico all'ingegneria del software. Il focus di questo libro è sull'ingegneria del software e sull'applicazione e l'uso di studi empirici, in particolare sulla sperimentazione, nell'ingegneria del software. Il numero di esperimenti pubblicati nell'ingegneria del software è aumentato ed è stato pubblicato un numero considerevole di esperimenti, come esaminato da Sjøberg et al. [161].

Le strategie empiriche nell'ingegneria del software includono:

- Impostazione di esperimenti formali,
- Studio di progetti reali nell'industria, ovvero esecuzione di un caso di studio, e
- Esecuzione di sondaggi attraverso, ad esempio, interviste.

Queste strategie sono descritte più dettagliatamente nei capitoli. 2 e 5 prima di concentrare il resto del libro sulla sperimentazione. Un'introduzione più generale a queste strategie di ricerca è presentata, ad esempio, da Robson [144]. I casi di studio in generale sono elaborati da Yin [180] e i casi di studio specificatamente nell'ingegneria del software sono elaborati da Runeson et al. [146]. Le strategie di ricerca non sono né completamente ortogonali né concorrenti.

Forniscono una classificazione conveniente, ma alcuni studi possono essere visti come una combinazione di essi o come una via di mezzo tra due di essi. Pertanto, ci sono sia somiglianze che differenze tra le strategie.

Il motivo principale per utilizzare la sperimentazione nell'ingegneria del software è consentire la comprensione e l'identificazione delle relazioni tra diversi fattori o variabili. Esistono numerose idee preconcette, ma sono vere? Fa

l'orientamento agli oggetti migliora il riutilizzo? Le ispezioni sono convenienti? Dovremmo fare delle riunioni ispettive o è sufficiente consegnare le osservazioni ad un moderatore?

Questo tipo di domande possono essere studiate per migliorare la nostra comprensione dell'ingegneria del software. Una migliore comprensione è la base per cambiare e migliorare il modo in cui lavoriamo, quindi gli studi empirici in generale e la sperimentazione in particolare sono importanti.

L'introduzione all'area si basa sull'introduzione di un processo di sperimentazione. I passaggi fondamentali del processo possono essere utilizzati anche per altri tipi di studi empirici. L'obiettivo, tuttavia, è fornire linee guida e supporto per l'esecuzione di esperimenti nell'ingegneria del software. Inoltre, va notato che gli esperimenti 'veri', cioè gli esperimenti con randomizzazione completa, sono difficili da eseguire nell'ingegneria del software. Gli esperimenti di ingegneria del software sono spesso quasi-esperimenti, cioè esperimenti in cui, ad esempio, non è stato possibile assegnare i partecipanti agli esperimenti a gruppi in modo casuale [37]. I quasi-esperimenti sono importanti e possono fornire risultati preziosi. Il processo presentato in questo libro mira sia a esperimenti "veri" che a quasi-esperimenti. Quest'ultimo è particolarmente supportato da una discussione approfondita sulle minacce agli esperimenti.

Pertanto, l'intento di questo libro è quello di fornire un'introduzione agli studi empirici e alla sperimentazione, al fine di evidenziare le opportunità e i vantaggi derivanti dalla sperimentazione nel campo dell'ingegneria del software. Il metodo di ricerca empirica può e deve essere utilizzato maggiormente nell'ingegneria del software. Gli argomenti contro gli studi empirici nell'ingegneria del software sono confutati da Tichy et al. [169]. Si spera che questa guida pratica alla sperimentazione nell'ingegneria del software faciliti l'uso di studi empirici e sperimentazioni sia nella ricerca che nella pratica dell'ingegneria del software.

1.3 Esercizi

- 1.1.** Perché gli esperimenti possono essere visti come prototipi per cambiamenti di processo?
- 1.2.** Come possono essere utilizzati gli esperimenti nelle attività di miglioramento?
- 1.3.** Perché gli studi empirici sono importanti nell'ingegneria del software?
- 1.4.** Quando il metodo di ricerca empirica è più adatto all'ingegneria del software rispetto rispettivamente ai metodi scientifico, ingegneristico e analitico?
- 1.5.** In quali tre strategie si dividono i metodi empirici?

Capitolo 2

Strategie empiriche

Esistono due tipi di paradigmi di ricerca che hanno approcci diversi agli studi empirici. *La ricerca esplorativa* si occupa di studiare gli oggetti nel loro ambiente naturale e di lasciare che i risultati emergano dalle osservazioni. Ciò implica che è necessario un *disegno di ricerca flessibile* [1] per adattarsi ai cambiamenti nel fenomeno osservato. La ricerca sulla progettazione flessibile viene anche definita *ricerca qualitativa*, poiché è informata principalmente da dati qualitativi. La ricerca induttiva tenta di interpretare un fenomeno sulla base delle spiegazioni fornite dalle persone. Si occupa di scoprire le cause noteate dai soggetti nello studio e di comprendere la loro visione del problema in questione. Il soggetto è la persona che partecipa ad uno studio empirico per valutare un oggetto.

La ricerca esplicativa si occupa principalmente di quantificare una relazione o di confrontare due o più gruppi con l'obiettivo di identificare una relazione causa-effetto.

La ricerca viene spesso condotta attraverso l'impostazione di esperimenti controllati. Questo tipo di studio è uno studio a *disegno fisso* [1], il che implica che i fattori vengono fissati prima del lancio dello studio. La ricerca sulla progettazione fissa viene anche definita *ricerca quantitativa*, poiché si basa principalmente su dati quantitativi. Le indagini quantitative sono appropriate quando si testano gli effetti di alcune manipolazioni o attività. Un vantaggio è che i dati quantitativi promuovono confronti e analisi statistiche. È possibile che la ricerca qualitativa e quella quantitativa indaghino sugli stessi argomenti, ma ciascuna di esse affronterà un diverso tipo di domanda. Ad esempio, si potrebbe avviare un'indagine quantitativa per verificare in che misura un nuovo metodo di ispezione riduce il numero di difetti riscontrati durante il test. Per rispondere alle domande sulle fonti delle variazioni tra i diversi gruppi di ispezione, si potrebbe avviare un'indagine qualitativa.

Come accennato in precedenza, strategie di progettazione fisse, come gli esperimenti controllati, sono appropriate quando si testano gli effetti di un trattamento, mentre uno studio di progettazione flessibile di convinzioni, comprensioni e prospettive multiple è appropriato per scoprire perché i risultati di un'indagine quantitativa sono così come sono. Sono. I due approcci dovrebbero essere considerati complementari piuttosto che competitivi.

Gli obiettivi di questo capitolo sono: (1) introdurre strategie di ricerca empirica, (2) evidenziare alcuni aspetti importanti in relazione alle strategie empiriche e (3) illustrare come le strategie possono essere utilizzate nel contesto del trasferimento tecnologico e miglioramento. Per raggiungere il primo obiettivo, viene fornita una panoramica delle strategie empiriche, vedere Sez. 2.1, quindi indagini, casi di studio ed esperimenti verranno discussi più in dettaglio. Le diverse strategie empiriche sono presentate brevemente nelle Sez. 2.2–2.4, e un loro confronto è fornito nella Sez. 2.5. Il secondo obiettivo viene affrontato affrontando le repliche degli esperimenti nella Sez. 2.6, le teorie in relazione agli studi empirici sono brevemente discusse nella Sez. 2.7, e l'aggregazione degli studi empirici sono elaborati nella Sez. 2.8. Infine, l'uso delle strategie di ricerca all'interno di un processo di trasferimento tecnologico e come parte di un programma di miglioramento è discusso nella Sez. 2.9.

2.1 Panoramica delle strategie empiriche

A seconda dello scopo della valutazione, che si tratti di tecniche, metodi o strumenti, e a seconda delle condizioni per l'indagine empirica, possono essere condotti tre principali tipi diversi di indagini (strategie): *indagine, studio di caso ed esperimento* [144].

Definizione 2.1. Un **sondaggio** è un sistema per raccogliere informazioni da o su persone per descrivere, confrontare o spiegare le loro conoscenze, atteggiamenti e comportamenti [58].

Un'indagine è spesso un'indagine condotta in retrospettiva, quando, ad esempio, uno strumento o una tecnica è in uso da un po' di tempo [133]. I mezzi principali per raccogliere dati qualitativi o quantitativi sono interviste o questionari. Queste vengono effettuate prelevando un campione rappresentativo della popolazione da studiare.

I risultati dell'indagine vengono poi analizzati per trarre conclusioni descrittive ed esplicative. Vengono poi generalizzati alla popolazione da cui è stato prelevato il campione. Le indagini sono discusse ulteriormente da Fink [58] e Robson [144].

Definizione 2.2. Il caso di studio nell'ingegneria del software è: un'indagine empirica che attinge a molteplici fonti di prova per indagare un caso (o un piccolo numero di casi) di un fenomeno contemporaneo di ingegneria del software nel suo contesto di vita reale, specialmente quando il confine tra fenomeno e contesto non può essere chiaramente specificato [146].

I casi di studio vengono utilizzati per ricercare progetti, attività o incarichi. I dati vengono raccolti per uno scopo specifico durante lo studio. Sulla base della raccolta dei dati è possibile effettuare analisi statistiche. Il caso di studio è normalmente finalizzato a tracciare un attributo specifico o a stabilire relazioni tra attributi diversi. Il livello di controllo è inferiore in un caso di studio che in un esperimento. Un caso di studio è uno studio osservazionale mentre l'esperimento è uno studio controllato [181]. Un caso di studio può, ad esempio, mirare a costruire un modello per prevedere il numero di errori nei test [2]. L'analisi statistica multivariata viene spesso applicata in questo tipo di

studi. I metodi di analisi includono la regressione lineare e l'analisi delle componenti principali [118]. La ricerca sui casi di studio è ulteriormente discussa in generale, ad esempio, da Robson [144], Stake [165] e Yin [180], e specificamente per l'ingegneria del software da Pfleeger [133], Kitchenham et al. [97], Verner et al. [173], Runeson e Host " [145], e Runeson et al. [146].

Per la strategia di indagine empirica al centro dell'attenzione di questo libro, l'esperimento, definiamo:

Definizione 2.3. L'esperimento (o esperimento controllato) nell'ingegneria del software è un'indagine empirica che manipola un fattore o una variabile dell'ambiente studiato. In base alla randomizzazione, trattamenti diversi vengono applicati a o da soggetti diversi, mantenendo costanti le altre variabili e misurando gli effetti sulle variabili di risultato. Negli esperimenti orientati all'uomo, gli esseri umani applicano trattamenti diversi agli oggetti, mentre negli esperimenti orientati alla tecnologia vengono applicati trattamenti tecnici diversi a oggetti diversi.

Gli esperimenti vengono condotti principalmente in un ambiente di laboratorio, che fornisce un elevato livello di controllo. Durante la sperimentazione, i soggetti vengono assegnati a trattamenti diversi in modo casuale. L'obiettivo è manipolare una o più variabili e controllare tutte le altre variabili a livelli fissi. Viene misurato l'effetto della manipolazione e in base a ciò è possibile eseguire un'analisi statistica. Nei casi in cui è impossibile assegnare casualmente i trattamenti ai soggetti, possiamo utilizzare quasi-esperimenti.

Definizione 2.4. Il quasi-esperimento è un'indagine empirica simile a un esperimento, in cui l'assegnazione dei trattamenti ai soggetti non può basarsi sulla randomizzazione, ma emerge dalle caratteristiche dei soggetti o degli oggetti stessi.

Negli studi sperimentali, i metodi di inferenza statistica vengono applicati con lo scopo di mostrare con significatività statistica che un metodo è migliore dell'altro [125, 144, 157]. I metodi statistici sono ulteriormente discussi nel Cap. 10.

I sondaggi sono molto comuni nell'ambito delle scienze sociali dove, ad esempio, vengono sondati gli atteggiamenti per determinare come voterà una popolazione alle prossime elezioni. Un rilievo non prevede alcun controllo sull'esecuzione o sulla misurazione, sebbene sia possibile confrontarlo con altri simili, ma non è possibile manipolare le variabili come negli altri metodi di indagine [6].

La ricerca sul caso di studio è una tecnica in cui vengono identificati i fattori chiave che possono avere qualche effetto sul risultato e quindi l'attività viene documentata [165, 180]. La ricerca sui casi di studio è un metodo osservativo, ovvero viene effettuata osservando un progetto o un'attività in corso.

Un esperimento è un'indagine formale, rigorosa e controllata. In un esperimento i fattori chiave vengono identificati e manipolati, mentre gli altri fattori nel contesto vengono mantenuti invariati, vedere sez. 6.1. La separazione tra casi di studio ed esperimento può essere rappresentata dal livello di controllo del contesto [132].

In un esperimento vengono deliberatamente applicate diverse situazioni e l'obiettivo è normalmente quello di distinguere tra due situazioni, ad esempio una situazione di controllo e una situazione sotto indagine. Esempi di fattori manipolati potrebbero essere, ad esempio, il metodo di ispezione o l'esperienza degli sviluppatori di software. In un caso di studio, il contesto è governato dal progetto reale oggetto di studio.

Tabella 2.1 Tipologia di progettazione e dati qualitativi e quantitativi nelle strategie empiriche

Strategia	Tipo di progettazione	Qualitativo/quantitativo
Sondaggio	Fisso	Entrambi
Caso di studio	Flessibile	Entrambi
Sperimentare	Fisso	Quantitativo

Alcune delle strategie di ricerca potrebbero essere basate su dati qualitativi o quantitativi, a seconda della progettazione dell'indagine, vedere Tabella 2.1. La classificazione di un'indagine dipende dalla struttura dei questionari, cioè da quali dati vengono raccolti e se è possibile applicare metodi statistici. Ciò vale anche per i casi di studio, ma la differenza è che un'indagine viene effettuata in retrospettiva mentre un caso di studio viene svolto durante l'esecuzione di un progetto. Un'indagine potrebbe anche essere avviata prima dell'esecuzione di un progetto. In quest'ultimo caso, l'indagine si basa su esperienze precedenti e quindi condotta in retrospettiva a queste esperienze, sebbene l'obiettivo sia quello di ottenere alcune idee sui risultati del prossimo progetto.

Gli esperimenti sono quasi puramente quantitativi poiché si concentrano sulla misurazione di diverse variabili, modificarle e misurarle nuovamente. Durante queste indagini vengono raccolti dati quantitativi e quindi vengono applicati metodi statistici.

Tuttavia, possono essere raccolti dati qualitativi per facilitare l'interpretazione dei dati [93].

Le sezioni seguenti forniscono un'introduzione a ciascuna strategia empirica.

2.2 Sondaggi

Le indagini vengono condotte quando l'uso di una tecnica o di uno strumento ha già avuto luogo [133] o prima che venga introdotto. Potrebbe essere visto come un'istantanea della situazione per catturare lo stato attuale. I sondaggi potrebbero, ad esempio, essere utilizzati per sondaggi di opinione e ricerche di mercato.

Quando si eseguono ricerche di indagine, l'interesse potrebbe essere, ad esempio, nello studio di come un nuovo processo di sviluppo abbia migliorato l'atteggiamento degli sviluppatori nei confronti della garanzia della qualità o nel dare priorità agli attributi di qualità [94]. Quindi viene selezionato un campione di sviluppatori tra tutti gli sviluppatori dell'azienda. Viene costruito un questionario per ottenere le informazioni necessarie per la ricerca. Ai questionari risponde il campione di sviluppatori. Le informazioni raccolte vengono quindi organizzate in una forma che può essere gestita in modo quantitativo o qualitativo.

2.2.1 Caratteristiche dell'indagine

Le indagini campionarie non vengono quasi mai condotte per creare una comprensione del particolare campione. Lo scopo è invece quello di comprendere la popolazione da cui è stato estratto il campione [6]. Ad esempio, intervistando 25 sviluppatori su cosa pensano di un nuovo processo, è possibile valutare l'opinione della popolazione più ampia di 100 sviluppatori presenti nell'azienda. Le indagini mirano allo sviluppo di conclusioni generalizzate.

Le indagini hanno la capacità di fornire un gran numero di variabili da valutare, ma è necessario puntare ad ottenere la massima comprensione dal minor numero di variabili poiché questa riduzione facilita anche il lavoro di raccolta e analisi dei dati. I sondaggi con molte domande sono noiosi da compilare per gli intervistati e di conseguenza la qualità dei dati potrebbe peggiorare. D'altro canto, le indagini mirano a fornire ampie panoramiche, che possono richiedere domande in diversi campi.

2.2.2 Scopi del sondaggio

Gli obiettivi generali per condurre un sondaggio sono i seguenti [6]:

- Descrittivo •
- Esplicativo •
- Esplorativo

È possibile condurre indagini descrittive per consentire asserzioni su alcune popolazioni.

Ciò potrebbe determinare la distribuzione di determinate caratteristiche o attributi.

La preoccupazione non riguarda il motivo per cui esiste la distribuzione osservata, ma piuttosto quale sia quella distribuzione.

Le indagini esplicative mirano a fornire affermazioni esplicative sulla popolazione.

Ad esempio, quando studiamo come gli sviluppatori utilizzano una determinata tecnica di ispezione, potremmo voler spiegare perché alcuni sviluppatori preferiscono una tecnica mentre altri ne preferiscono un'altra. Esaminando le relazioni tra le diverse tecniche candidate e diverse variabili esplicative, potremmo provare a spiegare perché gli sviluppatori scelgono una delle tecniche.

Infine, le indagini esplorative vengono utilizzate come studio preliminare a un'indagine più approfondita per garantire che non siano previste questioni importanti. Ciò potrebbe essere possibile creando un questionario strutturato in modo approssimativo e lasciando che sia un campione della popolazione a rispondere. Le informazioni vengono raccolte e analizzate e i risultati vengono utilizzati per migliorare l'indagine completa. In altre parole, l'indagine esplorativa non risponde alla domanda di ricerca di base, ma può fornire nuove possibilità che potrebbero essere analizzate e dovrebbero quindi essere seguite nell'indagine più mirata o approfondita.

2.2.3 Raccolta dati

I due mezzi più comuni per la raccolta dei dati sono i questionari e le interviste [58]. I questionari potrebbero essere forniti sia in formato cartaceo che in formato elettronico, ad esempio tramite posta elettronica o pagine web. Il metodo base per la raccolta dei dati tramite questionari è l'invio del questionario insieme alle istruzioni su come compilarlo. La persona che risponde risponde al questionario e poi lo restituisce al ricercatore.

Lasciare che gli intervistatori gestiscano i questionari (per telefono o faccia a faccia) invece che agli intervistati stessi, offre una serie di vantaggi:

- I sondaggi basati su interviste in genere ottengono tassi di risposta più elevati rispetto, ad esempio, alla posta sondaggi.
- Un intervistatore generalmente diminuisce il numero di “non so” e “nessuna risposta”, perché l'intervistatore può rispondere a domande sul questionario. • È possibile per l'intervistatore osservare e porre domande. Lo svantaggio sono i costi e i tempi, che dipendono dalla dimensione del campione e sono legati anche alle intenzioni dell'indagine.

2.3 Casi di studio

Un caso di studio viene condotto per indagare una singola entità o fenomeno nel suo contesto di vita reale, in uno spazio temporale specifico. In genere, il fenomeno può essere difficile da distinguere chiaramente dal suo ambiente. Il ricercatore raccoglie informazioni dettagliate, ad esempio, su un singolo progetto durante un periodo di tempo prolungato.

Durante l'esecuzione di un caso di studio, dovrebbero essere applicate una varietà di diverse procedure di raccolta dati e prospettive di analisi [146]. In questo capitolo viene fornita una breve introduzione per impostare il contesto per i diversi tipi di strategie empiriche, mentre un'introduzione più approfondita è fornita nel cap. 5.

Se, ad esempio, volessimo confrontare due metodi, lo studio può essere definito come un caso di studio o un esperimento, a seconda della scala della valutazione, della capacità di isolare i fattori e della fattibilità della randomizzazione. Un esempio di approccio al caso di studio potrebbe essere quello di utilizzare un progetto pilota per valutare gli effetti di un cambiamento rispetto ad alcuni valori di riferimento [97].

I casi di studio sono molto adatti per la valutazione industriale di metodi e strumenti di ingegneria del software perché possono evitare problemi di scale-up. La differenza tra casi di studio ed esperimenti è che gli esperimenti campionano le variabili che vengono manipolate, mentre i casi di studio selezionano dalle variabili che rappresentano la situazione tipica. Un vantaggio dei casi di studio è che sono più facili da pianificare e sono più realistici, ma lo svantaggio è che i risultati sono difficili da generalizzare e da interpretare, cioè è possibile mostrare gli effetti in una situazione tipica, ma richiede più tempo. analisi per generalizzare ad altre situazioni [180].

Se l'effetto di un cambiamento di processo è molto diffuso, uno studio di caso è più adatto. L'effetto del cambiamento può essere valutato solo ad un alto livello di astrazione perché il cambiamento del processo include cambiamenti più piccoli e più dettagliati durante tutto il processo di sviluppo [97]. Inoltre, gli effetti del cambiamento non possono essere identificati immediatamente. Ad esempio, se volessimo sapere se un nuovo strumento di progettazione aumenta l'affidabilità, potrebbe essere necessario attendere fino a dopo la consegna del prodotto sviluppato per valutare gli effetti sui guasti operativi.

La ricerca sui casi di studio è un metodo standard utilizzato per studi empirici in varie scienze come la sociologia, la medicina e la psicologia. Nell'ambito dell'ingegneria del software, i casi di studio non dovrebbero essere utilizzati solo per valutare come o perché si verificano determinati fenomeni, ma anche per valutare le differenze tra, ad esempio, due metodi di progettazione. Ciò significa in altre parole valutare quale dei due metodi sia più adatto in una determinata situazione [180]. Un esempio di caso di studio nell'ingegneria del software è un'indagine se l'uso della lettura basata sulla prospettiva aumenta la qualità delle specifiche dei requisiti. Uno studio come questo non può verificare che la lettura prospettica riduca il numero di difetti che raggiungono il test, poiché ciò richiede un gruppo di riferimento che non utilizzi tecniche basate sulla prospettiva, ma può portare luce sui meccanismi in gioco in un contesto ispettivo.

2.3.1 Disposizioni di casi di studio

Uno studio di caso può essere applicato come strategia di ricerca comparativa, confrontando i risultati dell'utilizzo di un metodo o di qualche forma di manipolazione con i risultati dell'utilizzo di un altro approccio. Per evitare bias e garantire la validità interna, è necessario creare una solida base per valutare i risultati del caso di studio. Kitchenham et al. proporre tre modi per organizzare lo studio per facilitare questo [97]:

- Una soluzione è un confronto dei risultati dell'utilizzo del nuovo metodo rispetto a uno scenario di riferimento aziendale. L'azienda dovrebbe raccogliere dati da progetti standard e calcolare caratteristiche come la produttività media e il tasso di difetti. Successivamente è possibile confrontare i risultati del caso studio con i dati di riferimento.
- È possibile scegliere un progetto gemello come base di partenza. Il progetto in studio utilizza il nuovo metodo e il progetto gemello quello attuale. Entrambi i progetti dovrebbero avere le stesse caratteristiche, vale a dire i progetti devono essere comparabili.
- Se il metodo si applica a singoli componenti del prodotto, potrebbe essere applicato in modo casuale ad alcuni componenti e non ad altri. Questo è molto simile a un esperimento, ma poiché i progetti non vengono estratti a caso dalla popolazione di tutti i progetti, non è un esperimento.

2.3.2 Fattori confondenti e altri aspetti

Quando si eseguono studi di casi è necessario minimizzare gli effetti dei fattori confondenti. Un fattore di confusione è un fattore che rende impossibile distinguere gli effetti di due fattori l'uno dall'altro. Questo è importante poiché non abbiamo lo stesso controllo su un caso di studio come su un esperimento. Ad esempio, potrebbe essere difficile stabilire se un risultato migliore dipenda dallo strumento o dall'esperienza dell'utente dello strumento. Gli effetti confondenti potrebbero comportare problemi nell'imparare a utilizzare uno strumento o un metodo quando si cerca di valutarne i benefici, o nell'utilizzare personale molto entusiasta o scettico.

Ci sono sia pro che contro con i casi di studio. I casi di studio sono preziosi perché incorporano qualità che un esperimento non può visualizzare, ad esempio scala, complessità, imprevedibilità e dinamismo. Alcuni potenziali problemi con i casi di studio sono:

- Uno studio di caso piccolo o semplificato raramente è un buon strumento per scoprire i principi e le tecniche dell'ingegneria del software. Gli aumenti di scala portano a cambiamenti nel tipo di problemi che diventano più indicativi. In altre parole, il problema può essere diverso in un caso di studio piccolo e in un caso di studio ampio, sebbene l'obiettivo sia studiare gli stessi problemi. Ad esempio, in un caso di studio di piccole dimensioni il problema principale potrebbe essere la tecnica effettivamente studiata, mentre in un caso di studio di grandi dimensioni il problema principale potrebbe essere il numero di persone coinvolte e quindi anche la comunicazione tra le persone.
- I ricercatori non hanno il pieno controllo della situazione di un caso di studio. Questo è positivo, da un certo punto di vista, perché i cambiamenti imprevedibili spesso dicono molto sui problemi studiati. Il problema è che non possiamo essere sicuri degli effetti a causa di fattori confondenti.

I casi di studio sono ulteriormente elaborati nel cap. 5.

2.4 Esperimenti

Gli esperimenti vengono lanciati quando vogliamo il controllo sulla situazione e vogliamo manipolare il comportamento in modo diretto, preciso e sistematico. Inoltre, gli esperimenti coinvolgono più di un trattamento per confrontare i risultati. Ad esempio, se è possibile controllare chi utilizza un metodo e chi utilizza un altro metodo, e quando e dove vengono utilizzati, è possibile eseguire un esperimento. Questo tipo di manipolazione può essere effettuata in una situazione off-line, ad esempio in un laboratorio in condizioni controllate, dove gli eventi sono organizzati per simulare la loro apparizione nel mondo reale. In alternativa, gli esperimenti possono essere condotti on-line, il che significa che l'indagine viene eseguita sul campo in un contesto di vita reale [6]. Il livello di controllo è più difficile in una situazione online, ma alcuni fattori potrebbero essere controllabili mentre altri potrebbero essere impossibili.

Gli esperimenti possono essere *orientati all'uomo* o *alla tecnologia*. Negli esperimenti orientati all'uomo, gli esseri umani applicano trattamenti diversi agli oggetti, ad esempio due metodi di ispezione vengono applicati a due pezzi di codice. Negli esperimenti orientati alla tecnologia, in genere vengono applicati strumenti diversi a oggetti diversi, ad esempio due strumenti di generazione di casi di test vengono applicati agli stessi programmi. L'esperimento orientato all'uomo ha meno controllo di quello orientato alla tecnologia, poiché gli esseri umani si comportano diversamente in diverse occasioni, mentre gli strumenti (per lo più) sono deterministici. Inoltre, a causa degli effetti di apprendimento, un soggetto umano non può applicare due metodi allo stesso pezzo di codice, cosa che due strumenti possono fare senza pregiudizi.

Come accennato in precedenza, considerare la nozione di contesto consente di stabilire in modo più rigoroso la differenza tra casi di studio ed esperimenti. Esempi di contesti diversi potrebbero essere l'area di applicazione e il tipo di sistema [132]. In un esperimento, identifichiamo i contesti di interesse, le sue variabili e li campioniamo.

Ciò significa che selezioniamo oggetti che rappresentano una varietà di caratteristiche tipiche dell'organizzazione in cui viene condotto l'esperimento e progettiamo la ricerca in modo che venga misurato più di un valore per ciascuna caratteristica. Un esempio potrebbe essere quello di indagare l'effetto di un metodo di ispezione rispetto ai difetti riscontrati durante il test in due sistemi diversi, utilizzando due linguaggi di programmazione diversi, ad esempio, in una situazione in cui un'organizzazione è passata da un linguaggio di programmazione a un altro. Quindi i diversi sistemi costituiscono il contesto per valutare il metodo di ispezione e quindi nell'esperimento sono necessari oggetti simili. Il metodo di ispezione diventa la variabile indipendente e un esperimento coinvolgerà oggetti in cui vengono utilizzati i diversi linguaggi di programmazione.

La progettazione dell'esperimento dovrebbe essere fatta in modo tale che gli oggetti coinvolti rappresentino tutti i metodi che ci interessano. Inoltre, è possibile considerare la situazione attuale come la linea di base (controllo), il che significa che la linea di base rappresenta un livello (o valore) della variabile indipendente e la nuova situazione sarà quella che vogliamo valutare. Quindi il livello della variabile indipendente per la nuova situazione descrive come la situazione valutata differisce dal controllo. Tuttavia, i valori di tutte le altre variabili dovrebbero rimanere gli stessi, ad esempio il dominio dell'applicazione e l'ambiente di programmazione.

2.4.1 Caratteristiche

Gli esperimenti sono appropriati per indagare diversi aspetti [72, 162], tra cui:

- Confermare le teorie, cioè testare le teorie esistenti.
- Confermare la saggezza convenzionale, cioè testare le concezioni delle persone.
- Esplorare le relazioni, cioè verificare che una certa relazione sia valida.
- Valutare l'accuratezza dei modelli, ovvero testare l'accuratezza di determinati modelli è come previsto.
- Validare le misure, cioè garantire che una misura misuri effettivamente quello che è dovrebbe.

La forza di un esperimento è che può indagare in quali situazioni le affermazioni sono vere e possono fornire un contesto in cui si consiglia l'uso di determinati standard, metodi e strumenti.

2.4.2 Processo dell'esperimento

La realizzazione di un esperimento prevede diverse fasi. I diversi passaggi sono:

1. Scoping
- 2.
- Pianificazione
3. Operazione 4. Analisi e interpretazione 5. Presentazione e pacchetto

Il processo dell'esperimento è presentato nel cap. 6, e le diverse fasi sono discusse più dettagliatamente nei capp. [7–11](#).

2.5 Confronto di strategie empiriche

I prerequisiti per un'indagine limitano la scelta della strategia di ricerca. Un confronto tra strategie può essere basato su una serie di fattori diversi. La tabella [2.2](#) è un'estensione dei diversi fattori discussi da Pfleeger [133]. I fattori sono ulteriormente descritti di seguito.

Il controllo di esecuzione descrive quanto controllo ha il ricercatore sullo studio.

Ad esempio, in un caso di studio, i dati vengono raccolti durante l'esecuzione di un progetto. Se la direzione decide di interrompere il progetto studiato, ad esempio per motivi economici, il ricercatore non può continuare a portare avanti il caso di studio. L'opposto è l'esperimento in cui il ricercatore ha il controllo dell'esecuzione.

Il controllo della misurazione è il grado in cui il ricercatore può decidere quali misure raccogliere e includere o escludere durante l'esecuzione dello studio.

Un esempio è come raccogliere dati sulla volatilità dei requisiti. Durante l'esecuzione di un'indagine non possiamo includere questo tipo di misure, ma in un caso di studio o in un esperimento è possibile includerle. In un sondaggio, possiamo raccogliere solo dati relativi all'opinione delle persone sulla volatilità dei requisiti.

Strettamente correlato ai fattori di cui sopra è il *costo dell'indagine*. A seconda della strategia scelta, il costo varia. Ciò è legato, ad esempio, alla dimensione dell'indagine e alla necessità di risorse. La strategia con il costo più basso è il sondaggio, poiché non richiede una grande quantità di risorse. La differenza tra casi di studio ed esperimenti è che se scegliamo di indagare un progetto in un caso di studio, il risultato del progetto è una qualche forma di prodotto che può essere venduto al dettaglio, cioè è un'indagine on-line. In un esperimento offline il risultato è una qualche forma di esperienza o conoscenza che non è direttamente redditizia allo stesso modo di un prodotto.

Un altro aspetto importante da considerare è la possibilità di *replicare* l'indagine. Lo scopo di una replica è dimostrare che il risultato dell'esperimento originale è valido per una popolazione più ampia. Una replica diventa una replica "vera" se è possibile replicare sia il progetto che i risultati. Non lo è

Tabella 2.2 Fattori della strategia di ricerca

Fattore	Sondaggio	Caso di studio	Sperimentare
Controllo dell'esecuzione	NO	NO	sì
Controllo della misurazione	NO	sì	sì
Costo dell'indagine	Basso	Medio	Alto
Facilità di replica	Alto	Basso	Alto

raro che l'obiettivo sia eseguire una replica, ma i risultati, per alcuni misura, si rivelano diversi rispetto ai risultati dello studio originale.

Un altro aspetto riguarda la replica, nel senso che riguarda gli studi nel tempo, si tratta di studi longitudinali [141]. La differenza principale tra un longitudinale studio e una replica è che uno studio longitudinale viene condotto principalmente con il stessi soggetti e una replica è per lo più uno studio condotto con nuovi soggetti. In in altre parole, replica significa diversi studi e uno studio longitudinale è uno solo studio. Lo studio longitudinale viene condotto su un periodo di tempo, ad esempio, a l'indagine può essere effettuata in diverse occasioni, gli esperimenti possono essere ripetuti e il caso lo studio può anche essere longitudinale se condotto su un periodo di tempo. Una longitudinale lo studio è normalmente condotto per comprendere, descrivere o valutare qualcosa che cambiamenti nel tempo [144].

La scelta della strategia empirica dipende dai prerequisiti per l'indagine, dallo scopo della stessa, dalle risorse disponibili e da come vorremmo analizzare la situazione. dati raccolti. Easterbrook et al. [50] forniscono ulteriori consigli sulla selezione della ricerca strategie. Inoltre, il confine tra diversi tipi di studio non è sempre presente taglio netto. Ad esempio, un caso di studio comparativo può anche essere definito un quasi-esperimento in un contesto industriale e uno studio osservazionale post-hoc del software risultati del corso di ingegneria, può anche essere definito un esperimento studentesco.

2.6 Repliche

La replica di un esperimento comporta la ripetizione dell'indagine in condizioni simili condizioni, variando ad esempio la popolazione soggetta. Questo aiuta a trovare quanta fiducia è possibile riporre nei risultati dell'esperimento. Se il presupposto della randomizzazione è corretto, cioè i soggetti sono rappresentativi di a popolazione più ampia, le repliche all'interno di questa popolazione mostrano gli stessi risultati di esperimento eseguito in precedenza. Se non otteniamo gli stessi risultati, lo siamo stati incapace di catturare tutti gli aspetti della progettazione dell'esperimento che influenzano il risultato. Anche se è possibile misurare una determinata variabile o replicare un esperimento, potrebbe essere difficile e troppo costoso.

Le repliche possono essere di diversi tipi [89, 155]:

- Le repliche *ravvicinate* seguono il più fedelmente possibile le procedure originali. Questo tipo viene talvolta definito repliche *esatte* [155]. • Le repliche *differenziate* studiano le stesse domande di ricerca, utilizzando diverse procedure sperimentali. Possono anche variare deliberatamente una o più condizioni principali dell'esperimento.

Basili et al. [20], hanno proposto una classificazione a grana più fine:

- Repliche rigorose (sinonimo di chiudere ed esatto) • Repliche che variano variabili intrinseche allo studio • Repliche che variano variabili intrinseche al focus dello studio • Repliche che variano le variabili di contesto nell'ambiente in cui il si valuta la soluzione
- Repliche che variano il modo in cui viene eseguito l'esperimento. • Repliche che estendono la teoria

In altri campi di ricerca vengono utilizzati molti schemi di classificazione diversi [64] e non esiste una terminologia standardizzata tra i campi di ricerca. Nemmeno la terminologia nel campo dell'ingegneria del software è consolidata. La distinzione sopra presentata tra repliche *vicine* e *differenziate* è un punto di partenza per specificare le repliche nell'ingegneria del software.

Il vantaggio delle repliche ravvicinate è che i fattori noti vengono tenuti sotto controllo, creando fiducia nel risultato. Tuttavia, le repliche ravvicinate a volte richiedono che gli stessi ricercatori conducano lo studio, poiché hanno una conoscenza tacita delle procedure dell'esperimento che difficilmente può essere documentata [153, 154]. D'altra parte, esiste un rischio sostanziale di bias dello sperimentatore negli studi di replica ravvicinata [95]. Inoltre, viene messo in dubbio che qualsiasi replica nell'ingegneria del software possa essere classificata come simile, poiché tanti fattori possono variare nel complesso contesto di un esperimento di ingegneria del software [89].

Le repliche differenziate possono invece essere utilizzate per studi più esplorativi. Se le differenze nei fattori e nei contesti sono ben documentate e analizzate, è possibile acquisire maggiori conoscenze da studi replicati. I fattori da considerare e segnalare per gli studi di replica differenziata includono [89]:

- *Sito* in cui viene condotto l'esperimento • *Sperimentatori* che conducono l'esperimento • *Progetto* scelto per l'esperimento • *Strumentazione*, ovvero moduli e altro materiale • *Variabili misurate* • *Soggetti* che conducono l'esperimento

Questi fattori sono discussi in dettaglio nel cap. 8. Vengono sollevate argomentazioni a favore della replica delle ipotesi originali, piuttosto che di specifici disegni sperimentali [123], vale a dire a favore di repliche differenziate piuttosto che di repliche ravvicinate.

2.7 Teoria nell'ingegneria del software

"Una teoria fornisce spiegazioni e comprensione in termini di concetti di base e meccanismi sottostanti, che costituiscono un'importante controparte alla conoscenza delle tendenze passeggiere e della loro manifestazione" [72]. Gli esperimenti possono essere condotti per generare, confermare ed estendere le teorie, come menzionato sopra. Tuttavia, l'uso della teoria è scarso nell'ingegneria del software, come concluso da Hannay et al. nella loro revisione sistematica della letteratura sugli esperimenti di ingegneria del software, 1993-2002 [72].

Hanno trovato 40 teorie in 23 articoli, dei 113 articoli della revisione. Solo due teorie sono state utilizzate in più di un articolo!

Endres e Rombach [53] identificano un elenco di 50 risultati a cui si riferiscono come "leggi", che è una nozione per la descrizione di un fenomeno ripetibile nel contesto delle scienze naturali. Endres e Rombach applicano questa nozione all'ingegneria del software.

Molte delle "leggi" elencate sono più generali dell'ingegneria del software, ad esempio "ci vogliono 5.000 ore per trasformare un principiante in un esperto". Nella loro concezione, *le teorie* spiegano le "leggi", *le ipotesi* propongono una spiegazione provvisoria del perché il fenomeno si comporta come osservato, mentre una *congettura* è un'ipotesi sul fenomeno. Endres e Rombach elencano 25 ipotesi e 12 congetture presenti nella letteratura sull'ingegneria del software.

Zendler [182] adotta un altro approccio e definisce una "teoria preliminare dell'ingegneria del software", composta da tre ipotesi fondamentali, sei ipotesi centrali e quattro ipotesi elementari. Esiste un rapporto gerarchico tra le ipotesi, essendo quelle fondamentali quelle più astratte, e quelle elementari quelle più concrete, derivanti dai risultati di studi sperimentali.

Gregor [70] descrive cinque tipi generali di teoria, che possono essere adattati al contesto dell'ingegneria del software [72]:

1. *Analisi*: teorie di questo tipo descrivono l'oggetto di studio e includono, ad esempio, tassonomie, classificazioni e ontologie.
2. *Spiegazione*: questo tipo di teorie spiega qualcosa, ad esempio il perché succede qualcosa.
3. *Previsione*: queste teorie mirano a prevedere cosa accadrà, ad esempio, in termini di modelli matematici o probabilistici.
4. *Spiegazione e previsione*: queste teorie combinano i tipi 2 e 3, ed è tipicamente ciò che viene indicato come una "teoria basata empiricamente".
5. *Progettazione e azione*: teorie che descrivono come fare le cose, tipicamente prescriptive sotto forma di scienza del design [76]. Si discute se questa categoria debba essere denotata come teoria.

Sjoberg et al. [162] propongono un quadro per le teorie dell'ingegneria del software, composto da quattro parti principali:

- Costrutti •
- Proposizioni •
- Spiegazioni • Ambito

Tabella 2.3 Quadro per le teorie dell'ingegneria del software, come proposto da Sjøberg et al. [162]

Classe Archetipo	Sottoclassi
Attore	Individuo, team, progetto, organizzazione o settore
Tecnologia	Modello di processo, metodo, tecnica, strumento o linguaggio
Attività	Pianificare, creare, modificare o analizzare (un sistema software)
Sistema software	I sistemi software possono essere classificati in base a molte dimensioni, come dimensione, complessità, dominio applicativo, progetto aziendale/scientifico/studentesco o amministrativo/embedded/tempo reale, ecc.

I *costrutti* sono le entità in cui si esprime la teoria, e alle quali la teoria offre una descrizione, spiegazione o previsione, a seconda del tipo di teoria come sopra definita. Le *proposizioni* sono costituite dalle relazioni proposte tra i costrutti. Le *spiegazioni* provengono da ragionamenti logici o osservazioni empiriche delle proposizioni, cioè dalla relazione tra i costrutti.

L' *ambito* della teoria definisce le circostanze in cui si presume che la teoria sia applicabile. Sjøberg et al. [162] suggeriscono che l'ambito venga espresso in termini di quattro classi di archetipi: attore, tecnologia, attività e sistema software, vedere Tabella 2.3.

Nonostante siano attraenti da un punto di vista teorico, nessuno dei due sistemi teorici proposti ha avuto finora un impatto significativo nel campo dell'ingegneria del software. Le teorie sono importanti per la concettualizzazione e la comunicazione della conoscenza all'interno di un campo di ricerca e sono utili quando si aggregano ricerche esistenti e si impostano studi di replica. Le teorie possono anche essere utilizzate per la comunicazione con i professionisti nel processo decisionale, che si tratti di scelte strategiche di tecnologia o di decisioni di progetto basate su modelli di previsione. Pertanto, la costruzione della teoria nell'ingegneria del software dovrebbe essere sviluppata, affinché il campo si sviluppi in un campo scientifico maturo.

2.8 Evidenza aggregata da studi empirici

Man mano che il numero di studi empirici cresce, appare la necessità di aggregare prove da più studi empirici, ad esempio studi di replica. In primo luogo, le ricerche dovrebbero basarsi l'una sull'altra, in modo che la nuova ricerca debba sempre prendere in considerazione le conoscenze esistenti come punto di partenza. In secondo luogo, diversi studi empirici insieme possono fornire risposte a domande a cui non trovano sufficiente risposta i singoli studi presi isolatamente. La raccolta e la sintesi delle prove empiriche devono soddisfare di per sé gli standard scientifici.

Le revisioni sistematiche della letteratura sono mezzi per raccogliere e sintetizzare prove empiriche da diverse fonti. Kitchenham e Charters definiscono le revisioni sistematiche della letteratura come “[una] forma di studio secondario che utilizza una metodologia ben definita per

identificare, analizzare e interpretare tutte le prove disponibili relative a una specifica domanda di ricerca in modo imparziale e (in una certa misura) ripetibile" [96]. Gli studi empirici ricercati vengono definiti *studi primari* mentre la revisione sistematica della letteratura in quanto tale viene definita *studio secondario*. Kitchenham e Charters forniscono linee guida per tali revisioni, riassunte nel Cap. 4.

Una revisione sistematica della letteratura ha una domanda di ricerca specifica, simile a una domanda di ricerca per un singolo studio empirico. La domanda di ricerca è correlata ai *risultati* degli studi empirici esaminati ed è tipicamente nella forma: "La tecnologia/metodo A è migliore o no di B?" [106].

La ricerca di studi empirici viene effettuata utilizzando query di database, nonché ricercando riviste, atti di conferenze e letteratura grigia, come rapporti tecnici, sulla base di argomenti chiave [96]. Vengono proposte anche procedure "snowballing", ovvero seguire i riferimenti da o verso un articolo per trovare altri articoli rilevanti [145]. Va notato che l'effetto valanga può essere sia all'indietro che in avanti.

L'effetto valanga all'indietro significa seguire l'elenco di riferimento e l'effetto valanga in avanti si riferisce al guardare i documenti che citano il documento che è stato ritenuto pertinente.

Se la domanda di ricerca è più generale, o se il campo di ricerca è meno esplorato, può essere invece avviato uno *studio di mappatura* (noto anche come *studio di scoping*).

Gli studi di mappatura hanno domande di ricerca più ampie, mirando a identificare lo stato della pratica o della ricerca su un argomento e in genere identificare le tendenze della ricerca [106]. Dato il loro campo di applicazione più ampio, le procedure di ricerca e classificazione sono meno rigorose e presentano caratteristiche più qualitative.

Sia le revisioni sistematiche della letteratura che gli studi di mappatura devono avere criteri chiari per l'inclusione e l'esclusione degli studi, nonché tassonomie per la loro classificazione.

Per le revisioni sistematiche della letteratura, un criterio naturale è che gli studi siano empirici, mentre gli studi di mappatura possono includere anche lavoro non empirico.

Quando viene raccolto un insieme di studi empirici su un argomento, avviene la sintesi o aggregazione. Le sintesi basate su metodi statistici sono chiamate *meta-analisi*. Esempi di meta-analisi nell'ingegneria del software includono metodi di rilevamento dei difetti [74, 121], metodi agili [46] e programmazione in coppia [73].

Se le procedure di meta-analisi non sono applicabili, deve essere utilizzata la sintesi descrittiva.

Questi includono la visualizzazione e la tabulazione dei dati e le statistiche descrittive dei dati [96]. Più è ampia la domanda di ricerca per una revisione della letteratura, più metodi qualitativi sono necessari per la sua sintesi. Cruzes e Dyba presentano una panoramica dei metodi di sintesi qualitativa [39].

L'interesse e la conduzione di revisioni sistematiche della letteratura sull'ingegneria del software sono cresciuti sostanzialmente durante il primo decennio del ventunesimo secolo.

Kitchenham et al. riportano 53 revisioni sistematiche uniche della letteratura pubblicate tra il 2004 e il 2008 [103,104]. Oltre alla sintesi dei risultati empirici, la conduzione delle revisioni porta all'identificazione di proposte di miglioramento, sia nella rendicontazione degli studi empirici in quanto tali, sia nei database in cui sono archiviati.

Le revisioni sistematiche della letteratura sono elaborate in modo più approfondito nel Cap. 4.

2.9 Empirismo nel contesto dell'ingegneria del software

Perché dovremmo eseguire esperimenti e altri studi empirici nell'ingegneria del software? La ragione principale per condurre studi empirici quantitativi è l'opportunità di ottenere risultati oggettivi e statisticamente significativi riguardanti la comprensione, il controllo, la previsione e il miglioramento dello sviluppo del software.

Gli studi empirici sono un input importante per il processo decisionale in un'organizzazione che cerca il miglioramento.

Prima di introdurre nuove tecniche, metodi o altri modi di lavorare, è preferibile una valutazione empirica delle virtù di tali cambiamenti. In questa sezione viene presentato un quadro per la valutazione dei cambiamenti dei processi software, in cui vengono suggerite diverse strategie empiriche in tre diversi contesti: desktop, laboratorio e progetti di sviluppo.

Per avere successo nello sviluppo del software ci sono alcuni requisiti di base [7, 8, 42]:

1. Comprensione del processo e del prodotto software.
2. Definizione delle qualità del processo e del prodotto.
3. Valutazione dei successi e dei fallimenti.
4. Feedback informativo per il controllo del progetto.
5. Imparare dall'esperienza.
6. Imballaggio e riutilizzo dell'esperienza rilevante.

Gli studi empirici sono importanti per supportare il raggiungimento di questi requisiti e si inseriscono nel contesto della ricerca industriale e accademica sull'ingegneria del software, nonché in un'organizzazione che apprende, alla ricerca del miglioramento continuo. Un esempio di organizzazione che apprende, chiamato Experience Factory, è proposto da Basili in concomitanza con il Paradigma del Miglioramento della Qualità [7], come ulteriormente descritto nel seguito di questa sezione. Questo approccio include anche un meccanismo per definire e valutare una serie di obiettivi operativi utilizzando la misurazione. Questo meccanismo è chiamato metodo Goal/Question/Metric (GQM) [17], che è ulteriormente descritto di seguito. Il metodo GQM è descritto più dettagliatamente da van Solingen e Berghout [172].

2.9.1 Valutazione Empirica dei Cambiamenti di Processo

Un'organizzazione in cerca di miglioramento desidera valutare l'impatto dei cambiamenti di processo (ad esempio, un nuovo metodo o strumento) prima di introdurli per migliorare il modo di lavorare. Gli studi empirici sono importanti per ottenere informazioni oggettive e quantificabili sull'impatto dei cambiamenti. Nelle sez. 2.2–2.4, vengono descritte tre strategie empiriche: indagini, studi di casi ed esperimenti, e vengono confrontate nella Sez. 2.5. Questa sezione descrive come le strategie possono essere utilizzate quando vengono valutate le modifiche del processo software [177]. L'obiettivo è discutere le strategie in termini di un modo adeguato di gestire il trasferimento tecnologico dalla ricerca all'industria

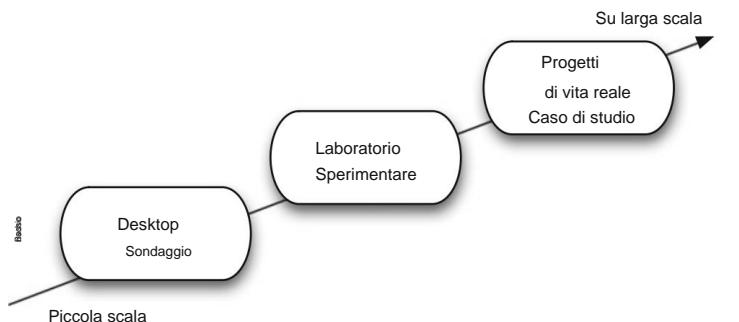


Fig. 2.1 Indagini, esperimenti e casi di studio

utilizzo. Il trasferimento tecnologico e alcune diverse fasi di tale processo in relazione all'utilizzo di strategie empiriche sono discussi nella Sez. 2.10.

Nella Fig. 2.1, le strategie sono collocate in ambienti di ricerca appropriati. L'ordine delle strategie si basa sulla dimensione "normale" dello studio. L'obiettivo è ordinare gli studi in base a come normalmente possono essere condotti per consentire un modo controllato di trasferire i risultati della ricerca nella pratica. Poiché un sondaggio non interviene in larga misura nello sviluppo del software, il rischio è minimo.

Un esperimento è per lo più piuttosto limitato rispetto a un progetto reale e il caso di studio è tipicamente mirato a un progetto specifico. Inoltre, un esperimento può essere condotto in un ambiente universitario prima di effettuare uno studio nell'industria, riducendo così i costi e i rischi, vedere anche Linkman e Rombach [113].

Gli ambienti di ricerca sono:

Desktop La proposta di modifica viene valutata offline senza eseguire il processo modificato. Pertanto, questo tipo di valutazione non coinvolge persone che applicano il metodo, lo strumento, ecc. Nell'ambiente desktop, è opportuno condurre sondaggi, ad esempio, attraverso valutazioni basate su interviste e studi della letteratura.

Laboratorio La proposta di modifica viene valutata in un ambiente di laboratorio offline (*in vitro*¹), dove viene condotto un esperimento e una parte limitata del processo viene eseguita in modo controllato.

Vita reale La proposta di cambiamento viene valutata in una situazione di sviluppo della vita reale, cioè viene osservata on-line (*in vivo*²). Si tratta, ad esempio, di progetti pilota. In questo ambiente è spesso troppo costoso condurre esperimenti controllati. Invece, i casi di studio sono spesso più appropriati.

Nella Fig. 2.1, la collocazione dei diversi ambienti di ricerca indica un aumento della scala e del rischio. Per provare, ad esempio, un nuovo metodo di progettazione

¹Latino significa "nel bicchiere" e si riferisce agli esperimenti chimici nella provetta.

²Latino significa "nella vita" e si riferisce a esperimenti in un ambiente reale.

in un progetto di design su larga scala e in un ambiente realistico, possiamo applicarlo in un progetto di sviluppo come studio pilota. Questo è, ovviamente, più rischioso rispetto a uno studio di laboratorio o desktop, poiché il fallimento della modifica del processo può mettere a repentaglio la qualità del prodotto consegnato. Inoltre, spesso è più costoso realizzare esperimenti e casi di studio, rispetto alla valutazione desktop, poiché uno studio desktop non comporta l'esecuzione di un processo di sviluppo. Si precisa che i costi si riferiscono al costo per l'accertamento della stessa cosa. Ad esempio, è probabilmente meno costoso intervistare prima le persone sull'impatto atteso di un nuovo metodo di revisione piuttosto che eseguire un esperimento controllato, che a sua volta è meno costoso che utilizzare effettivamente il nuovo metodo in un progetto con i rischi connessi all'adozione di una nuova tecnologia. .

Prima di effettuare uno studio di caso in un progetto di sviluppo, è necessario effettuare studi limitati in uno o entrambi gli ambienti desktop e di laboratorio per ridurre i rischi. Tuttavia, non esiste una conclusione generale su ordine e costi; per ogni proposta di cambiamento, dovrebbe essere fatta un'attenta valutazione di quali strategie empiriche siano più efficaci per la situazione specifica. La questione fondamentale è scegliere la migliore strategia in base a costi e rischi e in molti casi si consiglia di iniziare su piccola scala e poi, man mano che le conoscenze aumentano e il rischio diminuisce, lo studio viene ampliato.

Indipendentemente dalla strategia di ricerca che utilizziamo, è necessario un supporto metodologico in termini di come lavorare con il miglioramento, come raccogliere dati e archiviare le informazioni. Tali questioni verranno ulteriormente discusse successivamente.

2.9.2 Paradigma del miglioramento della qualità

Il Quality Improvement Paradigm (QIP) [7] è uno schema di miglioramento generale su misura per il business del software. QIP è simile al ciclo Pianifica/Fai/Studio/Agisci [23, 42] e comprende sei passaggi, come illustrato nella Fig. 2.2.

Questi passaggi sono spiegati di seguito [16].

1. *Caratterizzare*. Comprendere l'ambiente in base ai modelli disponibili, ai dati, all'intuizione, ecc. Stabilire linee di base con i processi aziendali esistenti nell'organizzazione e caratterizzarne la criticità.
2. *Stabilisci obiettivi*. Sulla base della caratterizzazione iniziale e delle capacità che hanno una rilevanza strategica per l'organizzazione, stabilire obiettivi quantificabili per il successo e il miglioramento delle prestazioni e del miglioramento del progetto e dell'organizzazione. Le aspettative ragionevoli sono definite in base alla linea di base fornita dalla fase di caratterizzazione.
3. *Scegli il processo*. Sulla base della caratterizzazione dell'ambiente e degli obiettivi prefissati, scegliere i processi di miglioramento adeguati, i metodi e gli strumenti di supporto, assicurandosi che siano coerenti con gli obiettivi prefissati.
4. *Esegui*. Eseguire lo sviluppo del prodotto e fornire feedback sul progetto in base ai dati sui risultati degli obiettivi raccolti.

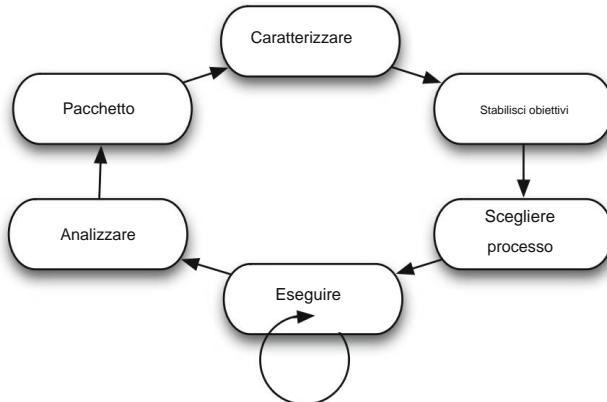


Fig. 2.2 I sei passi del paradigma del miglioramento della qualità [7]

5. **Analizza.** Alla fine di ogni progetto specifico, analizzare i dati e le informazioni raccolte per valutare le pratiche attuali, determinare i problemi, registrare i risultati e formulare raccomandazioni per futuri miglioramenti del progetto.
6. **Pacchetto.** Consolidare l'esperienza acquisita sotto forma di modelli nuovi, o aggiornati e perfezionati, e altre forme di conoscenza strutturata acquisita da questo e da progetti precedenti.

Il QIP implementa due cicli di feedback [16], vedere anche Fig. 2.2:

- Il *ciclo di feedback del progetto (ciclo di controllo)* è il feedback fornito al progetto durante la fase di esecuzione. Qualunque siano gli obiettivi dell'organizzazione, il progetto utilizzato come pilota dovrebbe utilizzare le sue risorse nel miglior modo possibile; pertanto gli indicatori quantitativi a livello di progetto e di compito sono utili per prevenire e risolvere i problemi.
- Il *ciclo di feedback aziendale (ciclo di capitalizzazione)* è il ciclo di feedback fornito all'organizzazione. Ha il duplice scopo di fornire informazioni analitiche sulle prestazioni del progetto al momento del completamento del progetto confrontando i dati del progetto con l'intervallo nominale nell'organizzazione e analizzando concordanza e discrepanza. L'esperienza riutilizzabile viene accumulata in una forma utile e applicabile ad altri progetti.

2.9.3 Fabbrica dell'esperienza

Il QIP si basa sul fatto che il miglioramento dello sviluppo del software richiede un apprendimento continuo. L'esperienza dovrebbe essere racchiusa in modelli di esperienza che possano essere effettivamente compresi e modificati. Tali modelli di esperienza sono archiviati in un repository, chiamato *experience base*. I modelli sono accessibili e possono essere modificati per il riutilizzo nei progetti attuali.

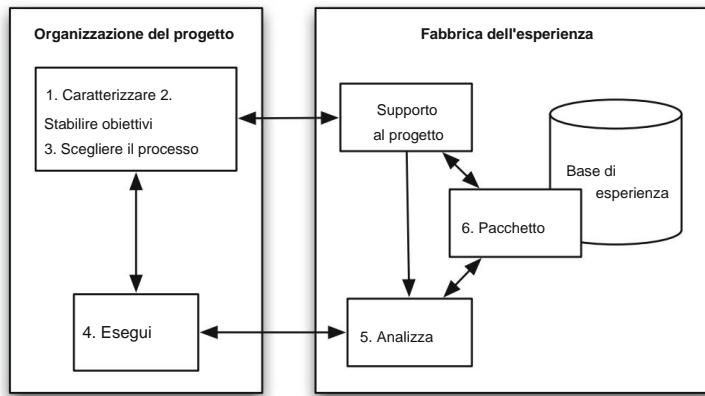


Fig. 2.3 Fabbrica dell'esperienza

QIP si concentra su una separazione logica dello sviluppo del progetto (eseguito dall'Organizzazione del progetto) dall'apprendimento sistematico e dal confezionamento dell'esperienza riutilizzabile (eseguito dalla Experience Factory) [8]. La Experience Factory è quindi un'organizzazione separata che supporta lo sviluppo del prodotto analizzando e sintetizzando tutti i tipi di esperienza, agendo come un archivio di tale esperienza e fornendo quell'esperienza a vari progetti su richiesta, vedere Fig. 2.3.

L'Experience Factory confeziona l'esperienza “costruendo modelli e misure informali, formali o schematizzati di vari processi, prodotti e altre forme di conoscenza tramite persone, documenti e supporto automatizzato” [16].

L'obiettivo dell'Organizzazione del Progetto è produrre e mantenere il software. L'organizzazione del progetto fornisce a Experience Factory le caratteristiche del progetto e dell'ambiente, i dati di sviluppo, le informazioni sull'utilizzo delle risorse, i record di qualità e le informazioni sui processi. Fornisce inoltre feedback sulle prestazioni effettive dei modelli elaborati dalla experience factory e utilizzati dal progetto.

La Experience Factory elabora le informazioni ricevute dall'organizzazione di sviluppo e restituisce un feedback diretto a ciascun progetto, insieme a obiettivi e modelli adattati da progetti simili. Fornisce inoltre linee di base, strumenti, lezioni apprese e dati, adattati al progetto specifico.

Per poter migliorare, un'organizzazione di sviluppo software deve introdurre nuove tecnologie. Ha bisogno di sperimentare e registrare le proprie esperienze derivanti da progetti di sviluppo ed eventualmente modificare l'attuale processo di sviluppo. Quando la tecnologia è sostanzialmente diversa dalla pratica attuale, la valutazione può essere offline per ridurre i rischi. La valutazione del cambiamento, come discusso sopra, può assumere la forma di un esperimento controllato (per una valutazione dettagliata nel piccolo) o di un caso di studio (per studiare gli effetti di scala). In entrambi i casi, il metodo Obiettivo/Domanda/Metrica, come descritto di seguito, fornisce un quadro utile.

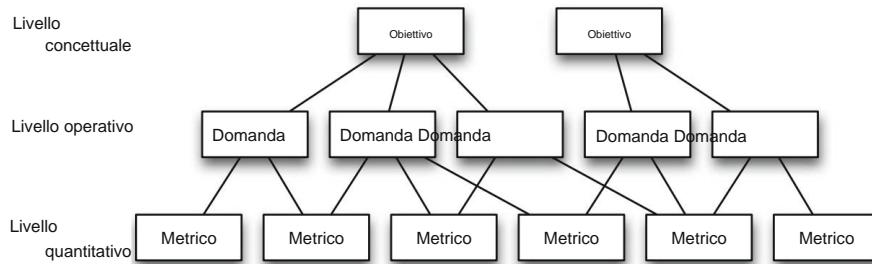


Fig. 2.4 Struttura gerarchica del modello GQM

2.9.4 Obiettivo/ Domanda/ Metodo metrico

Il metodo Obiettivo/Domanda/Metrico (GQM) [17, 26, 172] si basa sul presupposto che affinché un'organizzazione possa misurare in modo mirato deve:

1. Specificare gli obiettivi per sé e per i suoi progetti,
2. Tracciare tali obiettivi nei dati che intendono definire tali obiettivi operativamente, E
3. Fornire un quadro per interpretare i dati rispetto agli obiettivi dichiarati.

Il risultato dell'applicazione del metodo GQM è la specifica di un modello di misurazione mirato a un particolare insieme di questioni e un insieme di regole per l'interpretazione dei dati di misurazione.

Il modello di misurazione risultante ha tre livelli, come illustrato dalla gerarchia struttura chimica in Fig. 2.4:

1. *Livello concettuale* (Obiettivo). Si definisce un obiettivo per un oggetto, per molteplici ragioni, rispetto a diversi modelli di qualità, da diversi punti di vista, relativi ad un particolare ambiente. Oggetti di misurazione sono prodotti, processi e risorse (vedi anche Cap. 3).
2. *Livello operativo* (domanda). Una serie di domande viene utilizzata per caratterizzare il modo in cui verrà eseguita la valutazione/raggiungimento di un obiettivo specifico sulla base di un modello di caratterizzazione. Le domande cercano di caratterizzare gli oggetti di misurazione (prodotto, processo e risorsa) rispetto ad un aspetto di qualità selezionato e di determinarne la qualità dal punto di vista selezionato.
3. *Livello quantitativo* (metrico). Ad ogni domanda viene associato un insieme di dati per poter rispondere in modo quantitativo (oggettivamente o soggettivamente).

Il processo di definizione degli obiettivi è fondamentale per il successo dell'applicazione del metodo GQM. Gli obiettivi sono formulati sulla base di (1) politiche e strategie dell'organizzazione, (2) descrizioni di processi e prodotti e (3) modelli organizzativi. Una volta formulati gli obiettivi, le domande vengono sviluppate sulla base di questi obiettivi. Una volta sviluppate le domande, si procede ad associare le domande alle metriche appropriate.

Linee guida pratiche su come utilizzare il GQM per il miglioramento dei processi basati sulla misurazione sono fornite da Briand et al. [26], e van Solingen e Berghout [172].

Nel cap. 3, vengono ulteriormente descritti gli aspetti generali della misurazione.

2.10 Trasferimento tecnologico su base empirica

Gli studi empirici hanno un valore a sé stante, ma possono anche far parte di uno scambio di conoscenze e di uno sforzo di miglioramento congiunto tra il mondo accademico e l'industria, ad esempio nel trasferimento tecnologico come discusso anche in precedenza. L'ingegneria del software è un'area di ricerca applicata e quindi è prevista la ricerca su problemi di rilevanza industriale. In molti casi non è sufficiente fare semplicemente ricerca accademica, ad esempio, sull'ingegneria dei requisiti o sul test del software con la motivazione che queste aree rappresentano una sfida per l'industria. L'ingegneria del software è preferibilmente condotta congiuntamente dal mondo accademico e dall'industria per consentire il trasferimento di conoscenze in entrambe le direzioni e, infine, il trasferimento di nuovi metodi, tecnologie e strumenti dal mondo accademico all'industria. La ricerca congiunta offre un'eccellente opportunità per migliorare lo sviluppo del software industriale sulla base di prove concrete, e quindi costituisce un buon esempio di ingegneria del software basata sull'evidenza [48, 100].

Sulla base di un'impresa di collaborazione a lungo termine, un modello per il trasferimento tecnologico è stato documentato e presentato da Gorscheck et al. [66]. I sette passaggi del modello sono riassunti di seguito per illustrare come diversi studi empirici e in particolare esperimenti possano essere utilizzati per un miglioramento guidato empiricamente. Il modello è illustrato in Fig. 2.5. Il modello è strettamente correlato alla discussione sul miglioramento dei processi software nella Sez. 2.9. L'obiettivo principale del modello è l'utilizzo di diversi metodi empirici per creare una soluzione a un problema industriale reale e portarla all'applicazione industriale.

Identificazione del problema/problema industriale. Il primo passo è identificare le reali sfide industriali in un contesto industriale specifico, il che implica che il ricercatore sia presente presso il/i partner industriale/i. L'identificazione delle sfide può essere effettuata utilizzando, ad esempio, un sondaggio o interviste, che sono brevemente presentate nella Sez. 2.2. L'obiettivo è catturare le sfide e in particolare le questioni adatte alla ricerca. Qualsiasi sfida identificata deve poter essere formulata come problema di ricerca per evitare che il ricercatore finisca nel ruolo di consulente che affronta problemi a breve termine.

Uno dei principali vantaggi derivanti dall'esecuzione approfondita di questo passaggio è che crea un'opportunità per costruire un trust comune e garantisce che i partner industriali e i loro dipendenti si abituino alla presenza di ricercatori nel loro ambiente. In questa fase, l'impegno sia del management che degli altri professionisti coinvolti nello sforzo congiunto è cruciale per garantire il successo futuro secondo Wohlin et al. [179].

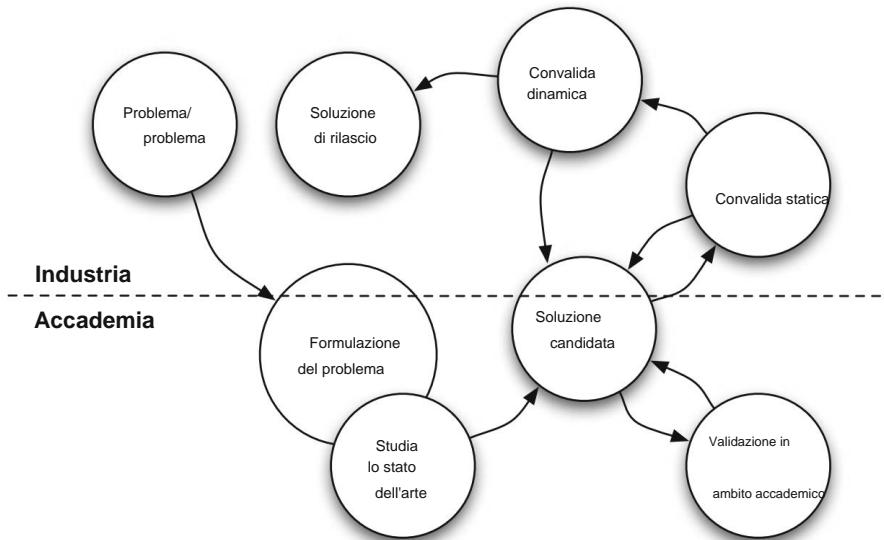


Fig. 2.5 Modello di trasferimento tecnologico (adattato dalla descrizione di Gorscheck et al. [66])

Formulazione del problema. Sulla base delle sfide identificate, la sfida dovrebbe essere formulato come un problema di ricerca e le domande di ricerca dovrebbero essere specificate. Se vengono identificate diverse sfide, è necessario stabilire la priorità indirizzo. Inoltre, dovrebbe esserci una persona di contatto principale per la sfida scelta identificata. La persona dovrebbe preferibilmente non solo essere nominata; dovrebbe essere una persona che vorrebbe essere il protagonista all'interno dell'azienda e fungere da campione per la collaborazione di ricerca. Ciò include aiutare a entrare in contatto con le persone giuste in azienda e contribuire a garantire che i ricercatori abbiano accesso ai sistemi, documentazione e dati quando necessari.

Come parte naturale della formulazione del problema di ricerca, i ricercatori condurre una ricerca bibliografica. Questo può essere fatto come una revisione sistematica della letteratura come presentato nel cap. 4. È necessaria un'indagine della letteratura per conoscere gli approcci esistenti alla sfida industriale identificata. Fornisce una base per la comprensione il rapporto tra gli approcci disponibili e le effettive esigenze industriali.

Soluzione candidata. Sulla base degli approcci disponibili e delle esigenze effettive, viene sviluppata una soluzione adatta, che può includere l'adattamento ai processi attuali, metodi, tecnologie e strumenti utilizzati in azienda. La soluzione viene preferibilmente sviluppata in stretta collaborazione con il/i partner industriale/i in modo che l'applicabilità può essere continuamente assicurato. Sebbene possa esserlo una soluzione specifica per un'azienda derivato, l'intenzione del ricercatore è quella di sviluppare una soluzione generica, che poi viene istanziato in un contesto specifico.

Validazione in ambito accademico. Una prima validazione della soluzione proposta viene preferibilmente condotta in un ambiente accademico per ridurre al minimo il rischio, ovvero una validazione offline. In molti casi questo può essere condotto come esperimento come descritto in diversi capitoli di questo libro o come caso di studio di un progetto studentesco. Una panoramica della ricerca sui casi di studio è fornita nel Cap. 5. La convalida in un ambiente accademico può essere condotta con gli studenti come soggetti o con rappresentanti dei/i partner industriale/i.

L'obiettivo principale in questa fase è individuare eventuali difetti evidenti nella soluzione proposta e identificare proposte di miglioramento della soluzione candidata. Questo viene fatto in un contesto accademico per garantire che la migliore soluzione possibile sia disponibile quando la si porta all'industria.

Convalida statica. Nella convalida statica, i rappresentanti del settore valutano la soluzione candidata offline. Ciò può essere fatto attraverso una presentazione della soluzione candidata seguita da interviste a diversi rappresentanti del settore, preferibilmente nei diversi ruoli interessati, o workshop congiunti. Inoltre, è preferibile fare una presentazione generale all'organizzazione per renderla consapevole della soluzione proposta in una fase iniziale. Ciò dà anche l'opportunità al personale di alzare la voce in una fase iniziale. Si spera che ciò contribuisca a superare qualsiasi resistenza una volta che la nuova soluzione sarà integrata nel modo in cui il software viene sviluppato all'interno dell'organizzazione.

In base alla convalida statica, potrebbe essere necessario modificare la nuova soluzione in base al feedback. I sette passaggi sono tutti iterativi, quindi è più una questione di quale ordine iniziano e non dovrebbero assolutamente essere visti come un approccio a cascata senza cicli di feedback.

Convalida dinamica. Una volta che la nuova soluzione supera la convalida statica e vi è accordo e impegno per implementare la nuova soluzione, è tempo di passare a una convalida dinamica. Ciò è preferibilmente fatto come valutazione pilota. Il modo esatto in cui condurre la validazione dipende dal tipo di soluzione. La nuova soluzione può essere utilizzata in un progetto, un sottoprogetto o per parti di un sistema, oppure per un'attività specifica.

Indipendentemente, si raccomanda di seguire attentamente la validazione dinamica per valutare la soluzione. La soluzione dinamica può essere studiata utilizzando un approccio di studio di caso come descritto nel Cap. 5.

Soluzione di rilascio. Una soluzione generica deve essere adattata a ciascuna situazione specifica. È necessario garantire che qualsiasi soluzione di ricerca sia adeguatamente affidata a un campione industriale e che l'azienda abbia un supporto sufficiente in termini di descrizioni, formazione e potenziale supporto degli strumenti. Quest'ultima non è principalmente responsabilità dei ricercatori, ma essi devono supportare i loro partner collaborativi per garantire che il trasferimento della nuova soluzione sia adeguatamente attuato e integrato nell'organizzazione prima di passare alla successiva sfida industriale.

Preferibilmente, l'utilizzo più ampio viene studiato anche empiricamente attraverso un caso di studio. Ciò aiuterà a ottenere prove empiriche per la nuova soluzione sviluppata nella collaborazione di ricerca.

Osservazione conclusiva. Il modello di trasferimento delineato illustra come possono essere applicate diverse strategie empiriche per supportare il trasferimento di nuovi risultati della ricerca dall'identificazione dei bisogni all'effettivo utilizzo industriale.

Infine, è interessante notare che i rappresentanti dell'industria sono interessati principalmente all'adattamento specifico al loro ambiente, mentre dal punto di vista dei ricercatori diventa un caso per la soluzione generica. Pertanto, i partner che collaborano possono avere obiettivi principali diversi, ma alla fine entrambi beneficiano dello sforzo congiunto. Il partner industriale ottiene una soluzione a una sfida identificata e i ricercatori sono in grado di valutare un risultato della ricerca in un ambiente industriale reale.

Gorschek e Wohlin [65] presentano un esempio di una soluzione generica per l'astrazione dei requisiti e una particolare esemplificazione industriale dell'approccio è presentata da Gorschek et al. [67] separatamente.

2.11 Etica nella sperimentazione

Qualsiasi attività di ricerca empirica che coinvolga soggetti umani deve tenere in considerazione aspetti etici. Alcuni aspetti sono regolati dalle leggi nazionali, altri non lo sono affatto. Andrews e Pradhan hanno identificato le questioni etiche nell'ingegneria del software e hanno ritenuto insufficienti le politiche esistenti [3]. Hall e Flynn hanno esaminato la pratica e la consapevolezza etica nel Regno Unito e hanno riscontrato un'inconsapevolezza allarmante [71], e nulla indica che questo paese sia un'eccezione.

Singer e Vinson hanno avviato una discussione su questioni etiche [158], hanno continuato a discutere casi di questioni etiche [159] e hanno fornito linee guida pratiche per la conduzione di studi empirici [174]. Hanno individuato quattro principi chiave:

- I soggetti devono dare *il consenso informato* alla loro partecipazione, il che implica che dovrebbero avere accesso a tutte le informazioni rilevanti sullo studio, prima di prendere la decisione di partecipare o meno. La loro decisione deve essere esplicita e libera, anche rispetto alle dipendenze implicite da dirigenti, professori, ecc. • Lo studio dovrebbe avere *valenza scientifica* in modo da motivare i soggetti ad esporsi ai rischi dello studio empirico, anche se minimi. • I ricercatori devono adottare tutte le misure possibili per mantenere *la riservatezza* dei dati e delle informazioni sensibili, anche quando ciò sia in conflitto con gli interessi di pubblicazione.

- Sopprimendo rischi, danni e benefici, deve prevalere la *beneficenza*, non solo per i singoli soggetti, ma anche per gruppi di soggetti e organizzazioni.

Questi principi vengono trasformati di seguito in linee guida più pratiche, relative alla pianificazione, conduzione e reporting di uno studio sperimentale. Rimandiamo anche a Sieber [156] per una checklist dei rischi per i soggetti da affrontare nella sperimentazione.

Revisione etica. Nei paesi in cui la legislazione richiede una revisione etica per gli studi che coinvolgono soggetti umani, come Canada, Stati Uniti e Australia, le procedure e la documentazione per tali studi devono essere seguite per consentire lo studio. IL

La revisione implica che una proposta venga sottoposta all'approvazione del Comitato di revisione etica (ERB) presso l'università o l'agenzia governativa. Queste procedure derivano per lo più dalle esigenze della ricerca biomedica e quindi generalmente non sono adattate alle esigenze dell'ingegneria del software. Vinson e Singer menzionano, ad esempio, che in Canada non è chiaro se gli studi che utilizzano il codice sorgente (scritto da esseri umani e rivelando informazioni su di essi) e i suoi dati siano soggetti alle procedure di revisione [174].

La documentazione necessaria nella revisione include tipicamente una descrizione del progetto, comprendente dettagli su argomenti e trattamenti, la documentazione su come viene ottenuto il consenso informato e una revisione degli aspetti etici del progetto.

Consenso informato. La base per uno studio empirico orientato all'uomo (ad esempio un esperimento) è che i soggetti partecipino volontariamente e che abbiano informazioni sufficienti per prendere la decisione di partecipare o meno. Inoltre, ciò include la possibilità di ritirarsi dallo studio in qualsiasi momento, senza alcuna penalità per il soggetto. Al fine di rendere chiaro ed esplicito questo processo decisionale, il consenso dovrà essere prestato per iscritto.

Un modulo di consenso comprende tipicamente i seguenti elementi [174]:

- *Titolo del progetto di ricerca:* a scopo identificativo. •

Informazioni di contatto: contatti sia di ricerca che di etica. •

Consenso e comprensione: i soggetti dichiarano di comprendere le condizioni del progetto e di accettarle. • *Recesso:* prevede la facoltà di recedere senza penalità. • *Riservatezza:* definite le promesse relative al trattamento confidenziale dei dati e partecipazione.

• *Rischi e benefici:* elencare esplicitamente ciò che i soggetti rischiano e guadagnano. • *Chiarimento:* il diritto del soggetto di porre domande per chiarire il proprio ruolo nello studio. •

Firma: prevalentemente sia del soggetto che del ricercatore, una copia per ciascuno, per indicare che si tratta di un accordo reciproco.

In alcuni disegni sperimentali, la completa divulgazione dell'obiettivo e delle procedure della ricerca può compromettere la conduzione dell'esperimento in quanto tale. Ad esempio, conoscendo l'ipotesi in anticipo, i soggetti potrebbero modificare il loro comportamento di conseguenza. Quindi, è possibile utilizzare la divulgazione parziale , nel senso che gli obiettivi e le procedure sperimentali sono presentati a un livello di astrazione più elevato.

Per gli studi empirici in azienda (in vivo), il consenso deve includere sia l'organizzazione che i singoli soggetti. In particolare, i soggetti non possono essere obbligati a partecipare, e sono liberi di ritirarsi senza sanzioni. Inoltre, devono essere prese in considerazione anche le questioni relative alla riservatezza e ai risultati sensibili all'interno dell'azienda.

Il consenso può essere differenziato a seconda che sia prestato per le finalità dell' studio in corso o se i dati possono essere utilizzati per ulteriori studi con obiettivi diversi.

Riservatezza. I soggetti devono essere sicuri che tutte le informazioni che condividono con i ricercatori rimarranno riservate. Tre aspetti della riservatezza sono [174]:

- *Privacy dei dati*, riferendosi all'accesso limitato ai dati, imposto ad esempio attraverso la protezione tramite password e la crittografia. • *Anonimato dei dati*, affrontato mantenendo separate le identità dei soggetti dati.
- *Anonimato della partecipazione*, il che significa che la decisione sul consenso dovrebbe essere mantenuta segreta.

Poiché gli studi empirici (compresi gli esperimenti) mirano a trarre conclusioni generali, non vi è alcun conflitto principale con il mantenimento della riservatezza dei dettagli.

I problemi relativi alla privacy dei dati possono essere risolti anche mediante buone pratiche lavorative. Tuttavia, poiché il numero dei soggetti è spesso ridotto, esiste il rischio che le informazioni possano essere ricondotte a individui, anche se anonimizzati, mettendo così a rischio l'anonymato. Inoltre, per la validità esterna dello studio (vedi par. 8.7), dovrebbero essere riportate informazioni sul contesto dello studio, che potrebbero contrastare con l'anonymato.

L'anonymato della partecipazione è la cosa più difficile da raggiungere. Gli studenti di una classe, che sono iscritti agli esperimenti, possono avere il diritto formale di rifiutare la partecipazione, ma è difficile nascondere al ricercatore quali studenti partecipano o meno.

Allo stesso modo nelle aziende, i manager saprebbero facilmente chi sta partecipando allo studio. Vinson e Singer consigliano che "per gli studi che coinvolgono studenti, i ricercatori dovrebbero evitare di reclutare studenti in classe e dovrebbero evitare di provare a reclutare i propri studenti" [174] – un consiglio seguito da pochi.

Risultati sensibili. I risultati di qualsiasi studio empirico possono essere sensibili sotto diversi aspetti per le diverse parti interessate. La performance individuale di una materia è un esempio che manager o professori vorrebbero vedere. Anche le conclusioni dello studio empirico possono essere delicate, soprattutto se uno sponsor del progetto vi partecipa. I risultati possono anche essere sensibili ai ricercatori, ad esempio, se un esperimento non supporta le loro ipotesi.

Queste situazioni sottolineano gli standard morali delle parti interessate. Le possibili misure da adottare per prepararsi a queste situazioni includono diversi tipi di indipendenza. Per risultati sensibili a:

- *I soggetti*, assicurarsi che si applichino procedure di riservatezza, indipendentemente dai fatti rivelati (reato esente [159]), • *Gli sponsor*, includere dichiarazioni chiare sui diritti per la pubblicazione indipendente dei risultati anonimizzati nel modulo di consenso informato per le aziende e nei contratti di progetto di ricerca , • *I ricercatori* considerano la possibilità di avere colleghi per eseguire analisi statistiche su dati anonimizzati (sia soggetti che scale) indipendentemente dagli sperimentatori, soprattutto quando il trattamento è progettato dagli stessi sperimentatori. Ciò riduce anche la minaccia delle aspettative dello sperimentatore.

Queste azioni riducono il rischio di rimanere bloccati in dilemmi etici e aumentano la validità di tutti gli studi empirici.

Incentivo. Nel reclutare soggetti per un esperimento, devono esserci incentivi per motivare la loro partecipazione. L'esperienza e la conoscenza acquisite con la candidatura

un nuovo metodo potrebbe essere un incentivo sufficiente. Per trattare tutti i partecipanti in modo equo, a tutti i soggetti dovrebbe essere data l'opportunità di conoscere tutti i trattamenti, anche se il disegno sperimentale non lo richiede.

Possono essere previsti anche incentivi monetari, ad esempio sotto forma di pagamento in contanti, di partecipazione a una lotteria o, per i soggetti professionali, della retribuzione ordinaria.

Indipendentemente dalla forma, l'incentivo deve essere bilanciato per garantire che il consenso a partecipare sia realmente volontario e non forzato da incentivi economici o di altro tipo troppo grandi.

Feedback. Per mantenere relazioni a lungo termine e fiducia con i soggetti di uno studio, il feedback dei risultati e dell'analisi è importante. I soggetti non devono essere d'accordo sull'analisi, ma dovrebbe essere data loro l'opportunità di ottenere informazioni sullo studio e sui suoi risultati. Se possibile, da un punto di vista della riservatezza, i dati relativi alle prestazioni individuali possono essere riportati insieme all'analisi complessiva.

Conclusione sull'etica. Singer e Vinson chiedono, nei loro primi lavori, un codice etico per l'ingegneria empirica del software [159]. Eppure, 10 anni dopo, la comunità non ne ha ancora sviluppata una; la più vicina sono le linee guida di Vinson e Singer [174], riassunte sopra. Le agenzie di finanziamento della ricerca iniziano a richiedere l'applicazione di codici etici generali, che potrebbero non essere adatti allo scopo. Linee guida etiche concrete e personalizzate per la ricerca empirica sull'ingegneria del software andrebbero a beneficio sia dei soggetti, che intendono proteggere, sia dello sviluppo del campo di ricerca in quanto tale.

2.12 Esercizi

2.1. Qual è la differenza tra ricerca qualitativa e quantitativa?

2.2. Cos'è un sondaggio? Fornire esempi di diversi tipi di indagini nell'ingegneria del software.

2.3. Quale ruolo giocano le repliche e le revisioni sistematiche della letteratura nella costruzione della conoscenza empirica?

2.4. Come si può combinare la Experience Factory con il metodo Goal/Question/Metrics e studi empirici su un contesto di trasferimento tecnologico?

2.5. Quali sono i principi etici chiave da osservare durante la conduzione degli esperimenti?

Capitolo 3

Misurazione

La misurazione del software è fondamentale per consentire il controllo di progetti, prodotti e processi, o come affermato da DeMarco: “*Non puoi controllare ciò che non puoi misurare*” [41]. Inoltre, la misurazione è una parte centrale negli studi empirici. Gli studi empirici vengono utilizzati per studiare gli effetti di alcuni input sull’oggetto in studio. Per controllare lo studio e vederne gli effetti, dobbiamo essere in grado sia di misurare gli input per descrivere ciò che causa l’effetto sull’output, sia di misurare l’output.

Senza misurazioni non è possibile avere il controllo desiderato e quindi non è possibile condurre uno studio empirico.

Misurazione e misura sono definite come [56]: “*La misurazione è il processo mediante il quale numeri o simboli vengono assegnati ad attributi di entità nel mondo reale in modo tale da descriverli secondo regole chiaramente definite*”. Una misura è il numero o il simbolo assegnato a un’entità da questa relazione per caratterizzare un attributo.

Invece di dare un giudizio direttamente sull’entità reale, studiamo le misure e diamo il giudizio su di esse. La parola metrica o metriche viene spesso utilizzata anche nell’ingegneria del software. Si possono individuare due diversi significati. Innanzitutto, la metrica del software viene utilizzata come termine per denotare il campo di misurazione nell’ingegneria del software. Il libro di Fenton e Pfleeger [56] ne è un esempio. In secondo luogo, la parola metrica viene utilizzata per denotare un’entità misurata, ad esempio, le righe di codice (LOC) sono una metrica del prodotto. Più precisamente, è una misura della dimensione del programma. La misurazione del software è discussa ulteriormente anche da Shepperd [150].

In questo capitolo viene presentata la teoria di base della misurazione. La sezione 3.1 descrive il concetto di base della teoria della misurazione e i diversi tipi di misura su scala.

Esempi di misure nell’ingegneria del software e la relazione con l’analisi statistica sono presentati nella sez. 3.2, mentre gli aspetti pratici delle misurazioni sono discussi nella Sez. 3.3.

3.1 Concetti di base

Una misura è una mappatura dall'attributo di un'entità a un valore di misurazione, solitamente un valore numerico. Le entità sono oggetti che possiamo osservare nel mondo reale.

Lo scopo di mappare gli attributi in un valore di misurazione è caratterizzare e manipolare gli attributi in modo formale. Una delle caratteristiche fondamentali di una misura è quindi quella di preservare le osservazioni empiriche dell'attributo [57]. Cioè, se l'oggetto A è più lungo dell'oggetto B, la misura di A deve essere maggiore della misura di B.

Quando utilizziamo una misura negli studi empirici, dobbiamo essere certi che la misura sia valida. Per essere *valida*, la misura non deve violare alcuna proprietà necessaria dell'attributo che misura e deve essere un'adeguata caratterizzazione matematica dell'attributo.

Una misura valida consente di distinguere tra loro oggetti diversi, ma entro i limiti dell'errore di misura gli oggetti possono avere lo stesso valore di misura. La misura deve anche preservare le nostre nozioni intuitive sull'attributo e sul modo in cui distinguere oggetti diversi [97]. Una misura deve essere valida sia analiticamente che empiricamente. La validità analitica di una misura si riferisce alla sua capacità di catturare in modo accurato e affidabile l'elemento di interesse. La validità empirica (a volte definita capacità statistica o predittiva) descrive quanto bene, ad esempio, un punteggio è correlato a qualcosa misurato in un altro contesto.

La dimensione dell'effetto è un modo semplice per quantificare la differenza tra due gruppi. Ciò è particolarmente importante nella sperimentazione, poiché potrebbe essere possibile mostrare una differenza statisticamente significativa tra due gruppi, ma potrebbe non essere significativa da un punto di vista pratico. Nella maggior parte dei casi, è possibile mostrare differenze statisticamente significative con un numero sufficientemente elevato di soggetti in un esperimento, ma ciò non significa necessariamente che sia significativo da un punto di vista pratico.

Può darsi che la differenza sia troppo piccola o che il costo per sfruttarla sia semplicemente troppo alto.

La mappatura da un attributo a un valore di misurazione può essere effettuata in molti modi diversi e ogni diversa mappatura di un attributo è una *scala*. Se l'attributo è la lunghezza di un oggetto, possiamo misurarlo in metri, centimetri o pollici, ognuno dei quali è una scala diversa della misura della lunghezza.

Poiché la misura di un attributo può essere misurata in scale diverse, a volte desideriamo trasformare la misura in un'altra scala. Se questa trasformazione da una misura all'altra preserva la relazione tra gli oggetti, si parla di trasformazione ammissibile [56]. Una *trasformazione ammissibile* è detta anche *riscalamento*.

Con le misure dell'attributo facciamo affermazioni sull'oggetto o sulla relazione tra diversi oggetti. Se le affermazioni sono vere anche se le misure vengono riscalate, sono dette *significative*, altrimenti sono *prive di significato* [27]. Ad esempio, se misuriamo le lunghezze degli oggetti A e B rispettivamente a 1 m e 2 m, possiamo affermare che B è lungo il doppio di A. Questa affermazione è vera anche

se ridimensioniamo le misure in centimetri o pollici, ed è quindi significativo.

Un altro esempio: misuriamo la temperatura nella stanza A e nella stanza B a **10°C e 20°C** e affermiamo che la stanza B è due volte più calda della stanza A. Se ridimensioniamo le temperature sulla scala Fahrenheit, otteniamo le temperature **50°F e 68°F**. L'affermazione non è più vera e quindi priva di significato.

A seconda della trasformazione ammissibile che può essere effettuata su una scala, si possono definire diversi tipi di scala. Le scale appartenenti a un tipo di scala condividono le stesse proprietà e i tipi di scala sono più o meno potenti, nel senso che si possono fare affermazioni più significative quanto più potente è la scala. I tipi di scala più comunemente utilizzati sono descritti di seguito.

Le misure possono anche essere classificate in altri due modi: (1) se la misura è diretta o indiretta, o (2) se la misura è oggettiva o soggettiva. Queste classificazioni verranno discusse ulteriormente più avanti in questo capitolo.

3.1.1 Tipi di bilancia

I tipi di scala più comuni sono i seguenti¹ [27, 56, 57]:

Nominale La scala nominale è la meno potente tra i tipi di scala. Mappa solo l'attributo dell'entità in un nome o simbolo. Questa mappatura può essere vista come una classificazione delle entità in base all'attributo.

Le trasformazioni possibili per le scale nominali sono quelle che preservano il fatto che le entità possono essere mappate solo uno a uno.

Esempi di scala nominale sono la classificazione, l'etichettatura e la tipizzazione dei difetti.

Ordinale La scala ordinale classifica le entità secondo un criterio di ordinamento, ed è quindi più potente della scala nominale. Esempi di criteri di ordinamento sono "maggiore di", "meglio di" e "più complesso".

Le trasformazioni possibili per la scala ordinale sono quelle che preservano l'ordine delle entità, ovvero **M0 DF .M** / dove **M0** e **M** sono misure diverse sullo stesso attributo, e **F** è una funzione crescente monotona.

Esempi di scala ordinale sono i voti e la complessità del software.

Intervallo La scala dell'intervallo viene utilizzata quando la differenza tra due misure è significativa, ma non il valore stesso. Questo tipo di scala ordina i valori allo stesso modo della scala ordinale ma esiste una nozione di "distanza relativa" tra due entità. La scala è quindi più potente della scala di tipo ordinale.

¹Fenton et al. [56,57] presentano un tipo di quinta scala. Il tipo di scala è la scala assoluta ed è un caso speciale della scala proporzionale. La scala assoluta viene utilizzata quando il valore stesso è l'unica trasformazione significativa. Un esempio di scala assoluta è il conteggio.

Le trasformazioni possibili con questo tipo di scala sono quelle in cui le misure sono una combinazione lineare tra loro, cioè $M_0 D \circ MC$ dove M_0 e M sono misure diverse sullo stesso attributo. Misure su questa scala sono rare nell'ingegneria del software.

Esempi di scala a intervalli sono la temperatura misurata in gradi Celsius o Fahrenheit.

Rapporto	Se esiste un valore zero significativo e il rapporto tra due misure è significativo, è possibile utilizzare una scala di rapporti. Le trasformazioni possibili sono quelle che hanno lo stesso zero e le scale differiscono solo di un fattore, cioè $M_0 D \circ M$ dove M_0 e M sono misure diverse sullo stesso attributo.
----------	--

Esempi di scala proporzionale sono la lunghezza, la temperatura misurata in Kelvin e la durata di una fase di sviluppo.

Le scale di misurazione sono legate alla ricerca qualitativa e quantitativa.

Inoltre, si riferisce a quali statistiche possono essere utilizzate sulle misure. Ciò è ulteriormente discusso nel cap. 10. Secondo Kachigan [90], la ricerca qualitativa riguarda la misurazione sulle scale nominale e ordinale, mentre la ricerca quantitativa tratta la misurazione sulle scale di intervallo e di rapporto.

3.1.2 **Misure oggettive e soggettive**

A volte, la misurazione di un attributo non può essere misurata senza considerare il punto di vista da cui viene preso. Possiamo dividere le misure in due classi:

Obiettivo Una misura oggettiva è una misura in cui non vi è alcun giudizio sul valore di misurazione e dipende quindi solo dall'oggetto da misurare. Una misura oggettiva può essere misurata più volte e da ricercatori diversi e lo stesso valore può essere ottenuto all'interno dell'errore di misurazione. Esempi di misure oggettive sono le righe di codice (LOC) e la data di consegna.

Soggettivo Una misura soggettiva è l'opposto della misura oggettiva. La persona che effettua la misurazione contribuisce esprimendo una sorta di giudizio. La misura dipende sia dall'oggetto che dal punto di vista da cui vengono presi. Una misura soggettiva può essere diversa se l'oggetto viene misurato nuovamente. Una misura soggettiva è per lo più di tipo scala nominale o ordinale. Esempi di misure soggettive sono l'abilità del personale e l'usabilità.

Le misure soggettive sono sempre soggette a potenziali distorsioni. Ciò è ulteriormente discusso nella Sez. 3.3.

3.1.3 Misure dirette o indirette

Gli attributi che ci interessano a volte non sono direttamente misurabili. Queste misure devono essere ricavate attraverso altre misure che siano direttamente misurabili. Per distinguere le misure misurabili dirette dalle misure derivate, dividiamo le misure in misure dirette e misure indirette.

Diretta Una misurazione diretta di un attributo è direttamente misurabile e non implica misurazioni su altri attributi. Esempi di misure dirette sono le righe di codice e il numero di difetti rilevati nel test.

Indiretto Una misurazione indiretta implica la misurazione di altri attributi.

La misura indiretta deriva dalle altre misure. Esempi di misure indirette sono la densità dei difetti (numero di difetti diviso per il numero di righe di codice) e la produttività dei programmatore (righe di codice divise per lo sforzo del programmatore).

3.2 Misure nell'ingegneria del software

Gli oggetti che interessano l'ingegneria del software possono essere suddivisi in tre diverse classi:

Processo Il processo descrive quali attività sono necessarie per produrre il software.

Prodotto I prodotti sono gli artefatti, i risultati finali o documenti che ne risultano da un'attività di processo.

Risorse Le risorse sono gli oggetti, come personale, hardware o software, necessari per un'attività di processo.

In ciascuna delle classi facciamo anche una distinzione tra attributi interni ed esterni [55]. Un attributo interno è un attributo che può essere misurato esclusivamente in termini di oggetto. Gli attributi esterni possono essere misurati solo rispetto al modo in cui l'oggetto si relaziona con altri oggetti. Esempi di diverse misure software sono mostrati nella Tabella 3.1.

Spesso nell'ingegneria del software, gli ingegneri del software vogliono fare dichiarazioni su un attributo esterno di un oggetto. Sfortunatamente, gli attributi esterni sono per lo più misure indirette e devono essere derivati da attributi interni dell'oggetto. Gli attributi interni sono per lo più misure dirette.

Le misure fanno spesso parte di un programma di misurazione. La creazione di programmi di misurazione del software è discussa, ad esempio, da Grady e Caswell [68] e Hetzel [75].

Le misurazioni nell'ingegneria del software sono diverse dalle misurazioni in altri settori, ad esempio la fisica. In questi ambiti è spesso chiaro quali siano gli attributi e come vengono misurati. Nell'ingegneria del software, tuttavia, a volte è difficile definire un attributo in modo misurabile con cui tutti

Tabella 3.1 Esempi di misure nell'ingegneria del software

Classe	Esempi di oggetti	Tipo di attributo	Esempio di misure
Processo	Test	Interno	Sforzo
		Esterno	Costo
Prodotto	Codice	Interno	Misurare
		Esterno	Affidabilità
Risorsa	Personale	Interno	Età
		Esterno	Produttività

è d'accordo [56]. Un'altra differenza è che è difficile dimostrare che le misure lo siano qualsiasi altra cosa tranne i tipi di scala nominale o ordinale nell'ingegneria del software. Validazione delle misure indirette è più difficile sia delle misure dirette che dei modelli per ricavare la misura esterna devono essere convalidati.

Quando conduciamo studi empirici, siamo interessati ai tipi di scala di le misure poiché da esse dipende l'analisi statistica. Formalmente, la statistica i metodi di analisi dipendono dal tipo di scala, ma i metodi sono per lo più piuttosto robusti per quanto riguarda il tipo di scala. La regola di base è che più potenti sono i tipi di scala che abbiamo utilizzare, più potenti saranno i metodi di analisi che potremo utilizzare, vedere il Cap. 10.

Molte misure nell'ingegneria del software sono spesso misurate con valori nominali o scale ordinali, oppure non è dimostrato che sia un tipo di scala più potente. Questo significa che non possiamo utilizzare i metodi di analisi statistica più potenti, che richiedono scale di intervallo o di rapporto, per gli studi empirici che conduciamo.

Briand et al. [27] sostengono che possiamo utilizzare l'analisi statistica più potente anche se non possiamo dimostrare di avere scale di intervalli o di rapporti. Molti di più potenti metodi statistici sono robusti alle distorsioni non lineari della scala degli intervalli se le distorsioni non sono troppo estreme. Se prendiamo cura e consideriamo attentamente il rischi, possiamo utilizzare i metodi statistici più potenti e ottenere risultati simili altrimenti non sarebbe fattibile senza un campione molto ampio di misure.

3.3 Misurazioni nella pratica

In pratica le metriche vengono definite dal ricercatore e poi raccolte durante il fase operativa dello studio empirico. Quando si tratta di come dovrebbero essere le metriche essere raccolti è un vantaggio se non richiede troppo impegno da parte dei soggetti nello studio. In molti esperimenti i soggetti compilano moduli per fornire i dati, ma è anche possibile definire sistemi di strumentazione in cui i dati vengono rilevati automaticamente raccolti, ad esempio, dall'ambiente di sviluppo. Lethbridge et al. [111] discutere diverse tecniche generali per la raccolta.

Poiché le metriche raccolte costituiscono la base per l'ulteriore analisi, la qualità dei parametri raccolti sono importanti per l'analisi continua dello studio. Questo significa che è importante capire veramente che tipo di metriche vengono raccolte e che saranno

certi della tipologia di scala a cui appartengono. È importante anche capire quale distribuzione rappresentano, in particolare se sono distribuiti normalmente oppure no.

Per quanto riguarda la distribuzione, questa potrebbe essere indagata mediante statistiche descrittive. I dati possono, ad esempio, essere rappresentati in un grafico oppure può essere utilizzata un'altra tecnica per analizzare in che misura i dati sono distribuiti normalmente. Ciò è ulteriormente approfondito nel cap. 10. Quando si tratta del tipo di scala, questo si basa su come vengono definite le metriche e deve essere compreso dal ricercatore quando le metriche vengono definite.

Il modo in cui vengono definite le metriche può influenzare notevolmente la loro efficacia nel mostrare ciò a cui il ricercatore è interessato. Ad esempio, Kitchenham et al. [102] confrontano due modi di visualizzare la produttività e mostrano che un grafico a dispersione che mostra lo sforzo rispetto alle dimensioni fornisce informazioni migliori rispetto a un grafico che mostra la produttività nel tempo. Un consiglio generale è quello di non utilizzare parametri costruiti dal rapporto di due misure indipendenti a meno che non si sia sicuri di comprendere le implicazioni della misura.

Durante lo svolgimento dello studio è importante assicurarsi che i dati raccolti siano corretti. Ciò significa che il ricercatore dovrebbe applicare procedure di garanzia della qualità durante l'esperimento, ad esempio rivedendo il modo in cui i soggetti compilano i moduli, controllando la coerenza tra i diversi valori, ecc. La validazione dei dati è ulteriormente discussa nel Cap. 8.

Un fattore correlato a questo riguarda chi è l'inventore o il proprietario degli aspetti che vengono indagati in un esperimento. Idealmente qualcun altro oltre all'inventore dei nuovi metodi dovrebbe valutarli negli esperimenti e in altri approcci di ricerca, come raccomandato da Kitchenham et al. [98]. L'inventore di un metodo desidera naturalmente che il metodo funzioni bene e c'è sempre il rischio che il ricercatore scelga consciamente o inconsciamente metriche favorevoli al metodo indagato. Se i soggetti sanno che il ricercatore è l'inventore del metodo investigato, ciò potrebbe influenzare anche le loro prestazioni. Se vengono condotti esperimenti in cui vengono studiati i propri metodi, la progettazione e la selezione delle metriche potrebbero essere riviste da ricercatori esterni.

3.4 Esercizi

3.1. Cosa sono misura, misurazione e metrica e come sono correlati?

3.2. Quali sono i quattro principali tipi di scale di misurazione?

3.3. Qual è la differenza tra una misura diretta e una indiretta?

3.4. In quali tre classi sono suddivise le misurazioni nell'ingegneria del software?

3.5. Cosa sono gli attributi interni ed esterni e come sono principalmente correlati misure dirette e indirette?

Capitolo 4

Revisioni sistematiche della letteratura

Le revisioni sistematiche della letteratura vengono condotte per *"identificare, analizzare e interpretare tutte le prove disponibili relative a una specifica domanda di ricerca"* [96]. Poiché si mira a fornire un quadro completo, esaustivo e valido delle prove esistenti, sia l'identificazione, l'analisi e l'interpretazione devono essere condotte in modo scientifico e rigoroso. Per raggiungere questo obiettivo, Kitchenham e Charters hanno adattato le linee guida per le revisioni sistematiche della letteratura, principalmente medica, le hanno valutate [24] e aggiornate di conseguenza [96]. Queste linee guida, strutturate secondo un processo in tre fasi per la *pianificazione, la conduzione e il reporting* della revisione, sono riassunte di seguito.

4.1 Pianificazione della Revisione

Pianificare una revisione sistematica della letteratura prevede diverse azioni:

Identificazione della necessità di una revisione. La necessità di una revisione sistematica nasce da un ricercatore che mira a comprendere lo stato dell'arte in un'area, o da professionisti che desiderano utilizzare prove empiriche nelle loro attività decisionali o di miglioramento strategiche. Se nel settore sono disponibili revisioni della letteratura più o meno sistematiche, queste dovrebbero essere valutate in termini di portata e qualità, per valutare se sono sufficienti a soddisfare le attuali esigenze di revisione. Una revisione sistematica della letteratura può essere vista come un metodo di ricerca per effettuare una revisione della letteratura.

Specificare le domande di ricerca. L'area della revisione sistematica e le domande di ricerca specifiche pongono il focus per l'identificazione degli studi primari, l'estrazione dei dati dagli studi e l'analisi. Pertanto, le domande di ricerca devono essere ben pensate e formulate. Gli aspetti da tenere in considerazione nel formulare le domande di ricerca includono [96]:

- La *popolazione* nella quale vengono raccolte le prove, cioè quale gruppo di persone, programmi o imprese sono di interesse per la revisione?

- L' *intervento* applicato nello studio empirico, ovvero quale tecnologia, strumento o procedura è oggetto di studio? • Il *confronto* con cui si confronta l'intervento, ovvero come viene definito il trattamento di controllo? In particolare l'intervento "placebo" è fondamentale, poiché "non utilizzare l'intervento" nella maggior parte dei casi non è un'azione valida nell'ingegneria del software. • I *risultati* dell'esperimento non dovrebbero essere solo statisticamente significativi, ma anche significativi da un punto di vista pratico. Ad esempio, probabilmente non è interessante che un risultato sia migliore del 10% sotto qualche aspetto se richiede il doppio del tempo.

- È necessario definire il *contesto* dello studio, ovvero una visione estesa della popolazione, compreso se è condotto nel mondo accademico o industriale, in quale segmento industriale e anche gli incentivi per i soggetti [78, 132]. • Dovranno esserlo anche i *disegni sperimentali* da includere nella domanda di ricerca definita.

Staples e Niazi raccomandano che la portata di una revisione sistematica della letteratura sia limitata da domande di ricerca chiare e ristrette per evitare studi ingestibili [166].

Sviluppo di un protocollo di revisione. Il protocollo di revisione definisce le procedure per la revisione sistematica della letteratura. Funziona anche come registro per condurre la revisione. Si tratta quindi di un documento "vivo", importante sia per lo svolgimento pratico della revisione, sia per la sua validità. Kitchenham e Charters propongono che i seguenti elementi siano trattati in un protocollo di revisione [96]:

- Contesto e logica
- Domande di ricerca
- Strategia di ricerca per studi primari
- Criteri di selezione degli studi
- Procedure di selezione degli studi
- Liste di controllo e procedure di valutazione della qualità degli studi
- Strategia di estrazione dei dati
- Sintesi dei dati estratti
- Strategia di diffusione
- Calendario del progetto

Il protocollo è preferibilmente rivisto da colleghi per garantirne la coerenza e la validità. L'esperienza derivante dalla revisione sistematica della letteratura sottolinea l'importanza di uno studio pre-revisione per aiutare a definire le domande di ricerca, oltre ad essere aperti a modificare le domande di ricerca durante lo sviluppo del protocollo, man mano che il problema in studio diventa più chiaro [24].

4.2 Conduzione della revisione

Condurre la revisione significa mettere in pratica il protocollo di revisione. Ciò include:

Identificazione della ricerca. L'attività principale in questo passaggio prevede la specifica delle stringhe di ricerca e la loro applicazione ai database. Tuttavia, include anche ricerche manuali in riviste e atti di conferenze, nonché ricerche nei siti Web dei ricercatori o invio di domande ai ricercatori. La ricerca sistematica di studi primari basata su riferimenti da e verso altri studi è chiamata "valanga" [145].

La strategia di ricerca è un compromesso tra il reperimento di tutti gli studi primari rilevanti e il non ottenere un numero schiaccianiente di falsi positivi, che devono essere esclusi manualmente [43]. Un falso positivo è un risultato erroneamente positivo quando non dovrebbe esserlo; in questo caso significa che un documento viene trovato e quindi ritenuto interessante, ma poi si scopre che non lo è e quindi deve essere rimosso. La stringa di ricerca è sviluppata a partire dall'area da coprire e dalle domande di ricerca.

L'utilizzo di più database è necessario per coprire tutta la letteratura pertinente, ma crea anche duplicati, che devono essere identificati e rimossi. Alla fine, si deve accettare che gli articoli trovati siano un campione della popolazione di tutti gli articoli su un argomento specifico. La questione fondamentale è che il campione provenga effettivamente dalla popolazione prevista.

Gli studi primari pubblicati tendono ad avere un *bias di pubblicazione*, il che significa che (in un certo senso) è più probabile che i risultati *positivi* vengano pubblicati rispetto a quelli *negativi*.

Pertanto, dovrebbe essere ricercata anche la letteratura grigia, come relazioni tecniche, tesi, pubblicazioni rifiutate e lavori in corso [96].

È preferibile archiviare i risultati della ricerca e un registro delle azioni intraprese utilizzando un sistema di gestione di riferimento.

Selezione degli studi primari. La base per la selezione degli studi primari sono i criteri di inclusione ed esclusione. I criteri dovrebbero essere sviluppati in anticipo, per evitare pregiudizi. Tuttavia, potrebbero dover essere modificati nel corso della selezione, poiché tutti gli aspetti di inclusione ed esclusione non sono evidenti nella fase di pianificazione.

L'insieme identificato di studi candidati viene elaborato in relazione ai criteri di selezione. Per alcuni studi è sufficiente leggere il titolo o l'abstract per giudicare l'articolo, mentre altri articoli necessitano di un'analisi più approfondita, ad esempio, della metodologia o delle conclusioni per determinarne lo status. Abstract strutturati [30] possono aiutare il processo di selezione.

Poiché il processo di selezione è una questione di giudizi, anche con criteri di selezione ben definiti, è consigliabile che due o più ricercatori valutino ciascun articolo, o almeno un campione casuale di articoli. Quindi l'accordo tra valutatori può essere misurato utilizzando la statistica Cohen Kappa [36] ed essere riportato come parte della valutazione della qualità della revisione sistematica della letteratura. Tuttavia, va notato che è possibile ottenere una statistica Cohen Kappa relativamente elevata poiché molti documenti trovati nella ricerca automatica vengono facilmente esclusi dai ricercatori quando li valutano manualmente. Pertanto, potrebbe essere importante condurre la valutazione in più fasi, ovvero iniziare rimuovendo quei documenti che ovviamente non sono rilevanti sebbene trovati nella ricerca.

Valutazione della qualità dello studio. Valutare la qualità degli studi primari è importante, soprattutto quando gli studi riportano risultati contraddittori. La qualità del

gli studi primari possono essere utilizzati per analizzare la causa di risultati contraddittori o per valutare l'importanza dei singoli studi durante la sintesi dei risultati.

Non esiste una definizione universalmente accettata e applicabile di "qualità dello studio".

I tentativi di mappare i criteri di qualità dalla medicina non sono stati mappati alla gamma di qualità degli studi di ingegneria del software [47].

Lo strumento più utile dal punto di vista pratico per la valutazione della qualità sono le liste di controllo, anche se il loro supporto empirico può essere debole. Uno studio di Kitchenham et al. ha anche dimostrato che sono necessari almeno tre revisori per effettuare una valutazione valida [105].

Le liste di controllo utilizzate nella valutazione della qualità degli studi empirici sono disponibili nella letteratura empirica sull'ingegneria del software [96, 105, 145].

La valutazione della qualità può portare all'esclusione di alcuni studi primari, se la qualità dello studio rientra nei criteri di selezione. Vale anche la pena notare che dovrebbe essere valutata la qualità degli studi primari, non la qualità dei resoconti. Tuttavia, spesso è difficile giudicare la qualità di uno studio se questo viene riportato in modo inadeguato. Potrebbero essere necessari contatti con gli autori per trovare o chiarire informazioni carenti nei rapporti.

Estrazione e monitoraggio dei dati. Una volta deciso l'elenco degli studi primari, vengono estratti i dati degli studi primari. Un modulo di estrazione dei dati è progettato per raccogliere le informazioni necessarie dai rapporti degli studi primari. Se i dati di valutazione della qualità vengono utilizzati per la selezione dello studio, il modulo di estrazione è separato in due parti, una per i dati di qualità, da compilare durante la valutazione della qualità, e una per i dati di studio da compilare durante l'estrazione dei dati.

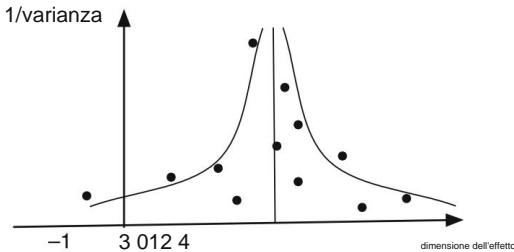
Il modulo di estrazione dei dati è progettato in base alle domande di ricerca. Per la sintesi meta-analitica pura, i dati sono un insieme di valori numerici, che rappresentano il numero di soggetti, le caratteristiche degli oggetti, gli effetti del trattamento, gli intervalli di confidenza, ecc. Per insiemi di studi meno omogenei, devono essere incluse descrizioni più qualitative degli studi primari. Oltre ai dati grezzi, per ciascuno studio primario vengono registrati il nome del revisore, la data di estrazione dei dati e i dettagli di pubblicazione.

Il modulo di estrazione dei dati dovrebbe essere sperimentato prima di essere applicato all'intera serie di studi primari. Se possibile, l'estrazione dei dati dovrebbe essere eseguita in modo indipendente da due ricercatori, almeno per un campione degli studi, al fine di valutare la qualità della procedura di estrazione.

Se uno studio primario viene pubblicato in più di un articolo, ad esempio se un articolo di una conferenza viene esteso alla versione di una rivista, solo un caso dovrebbe essere conteggiato come studio primario. Nella maggior parte dei casi si preferisce la versione journal in quanto è la più completa, ma per l'estrazione dei dati è possibile utilizzare entrambe le versioni. Anche le relazioni tecniche di supporto o la comunicazione con gli autori possono servire come fonti di dati per l'estrazione.

Sintesi dei dati. La forma più avanzata di sintesi dei dati è *la meta-analisi*. Ciò si riferisce ai metodi statistici applicati per analizzare i risultati di diversi studi indipendenti. La meta-analisi presuppone che gli studi sintetizzati siano omogenei o che la causa della disomogeneità sia ben nota [135]. Una meta-analisi confronta *le dimensioni degli effetti* e i valori p per valutare il risultato sintetizzato. Lo è principalmente

Fig. 4.1 Un esempio di grafico a imbuto per 12 studi ipotetici



applicabile a esperimenti replicati, se presenti, a causa del requisito di omogeneità.

In sintesi, gli studi da includere in una meta-analisi devono [135]:

- Essere dello stesso tipo, ad esempio esperimenti formali
- Avere la stessa ipotesi di test
- Avere le stesse misure del trattamento e dei costrutti degli effetti
- Riportare gli stessi fattori esplicativi

Le procedure di meta-analisi prevedono tre fasi principali [135]:

1. Decidere quali studi includere nella meta-analisi.
2. Estrarre la dimensione dell'effetto dal rapporto dello studio primario o stimare se non esiste dimensione dell'effetto pubblicata.
3. Combinare le dimensioni degli effetti degli studi primari per stimare e testare effetto combinato.

Oltre alle procedure primarie di selezione degli studi presentate sopra, la meta-analisi dovrebbe includere un'analisi dei *bias di pubblicazione*. Tali metodi includono il *grafico a imbuto*, come illustrato nella Figura 4.1, dove le dimensioni degli effetti osservati sono tracciate rispetto a una misura della dimensione dello studio, ad esempio l'inverso della varianza o un'altra misura di dispersione (vedere Sezione 10.1.2). Se l'insieme degli studi primari è completo, i dati dovrebbero essere distribuiti secondo uno schema a "imbuto". Le lacune nel funnel indicano che alcuni studi non sono stati pubblicati o trovati [135].

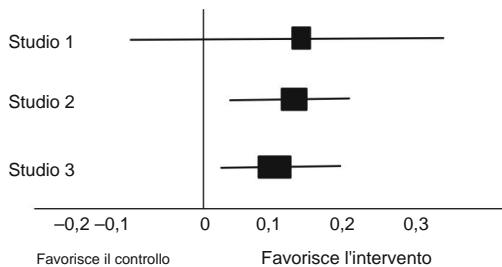
La *dimensione dell'effetto* è un indicatore, indipendente dall'unità o dalla scala utilizzata in ciascuno degli studi primari. Dipende dal tipo di studio, ma in genere potrebbe essere la differenza tra i valori medi di ciascun trattamento. Questa misura deve essere normalizzata per consentire confronti con altre scale, cioè divisa per la deviazione standard combinata [135].

L'analisi presuppone l'omogeneità tra gli studi e viene quindi eseguita con un modello a *effetti fissi*. La meta-analisi stima la reale dimensione dell'effetto calcolando un valore medio delle dimensioni dell'effetto dei singoli studi, che sono esse stesse medie.

Esistono test per identificare l'eterogeneità, come il test Q e il test del rapporto di verosimiglianza, che dovrebbero essere applicati per garantire che le condizioni del modello siano soddisfatte [135].

Per i dati disomogenei, esiste un modello a *effetti casuali*, che consente la variabilità dovuta a un fattore sconosciuto, che influenza le dimensioni dell'effetto per gli studi primari. Questo modello fornisce stime sia per l'errore di campionamento, come il modello a effetti fissi, sia per la variabilità nelle sottopopolazioni disomogenee.

Fig. 4.2 Un esempio di foresta trama per tre ipotetici studi



Metodi meno formali per la sintesi dei dati includono la sintesi *descrittiva* o *narrativa*.

Questi metodi tabulano i dati degli studi primari in un modo che porta luce alla domanda di ricerca. Come requisito minimo sui dati tabulati, Kitchenham e le Carte propongono che siano presentati i seguenti punti [96]:

- Dimensione del campione per ciascun intervento
- Stime della dimensione dell'effetto per ciascun intervento con errori standard per ciascun effetto
- Differenza tra i valori medi per ciascun intervento e la confidenza intervallo per la differenza
- Unità utilizzate per misurare l'effetto

I risultati statistici possono essere visualizzati utilizzando *i forest plot*. Un appezzamento di bosco presenta il medie e varianza della differenza tra i trattamenti per ciascuno studio. Un esempio l'appezzamento forestale è mostrato in Fig. 4.2.

La sintesi di studi disomogenei e di studi con metodi misti richiede approcci qualitativi. Cruzes e Dyba [39] hanno esaminato gli studi secondari sul software ingegneria, che includeva la sintesi di prove empiriche. Ne hanno identificati diversi metodi di sintesi, molti dei quali provenienti dalla medicina, di cui sette metodi sono stati utilizzati ingegneria del software. Questi metodi vengono brevemente introdotti di seguito. Per maggiori dettagli, fare riferimento a Cruzes e Dyba [39] e riferimenti correlati.

- *L'analisi tematica* è un metodo che mira a identificare, analizzare e riportare modelli o temi negli studi primari. Come minimo, organizza e presenta i dati in modo ricco di dettaglio e interpreta vari aspetti dell'argomento in studio.
- *La sintesi narrativa*, di cui sopra, racconta una 'storia' che ha origine dai prove primarie. Le prove grezze e le interpretazioni sono strutturate, utilizzando for esempio di tabulazione di dati, raggruppamenti e clustering o conteggio dei voti come a strumento descrittivo. La sintesi narrativa può essere applicata a studi qualitativi o dati quantitativi, o combinazioni di essi.
- Il metodo *dell'analisi comparativa* è finalizzato all'analisi di connessioni causali complesse. Utilizza la logica booleana per spiegare le relazioni tra causa ed effetto nei studi primari. L'analisi elenca le condizioni necessarie e sufficienti in ciascuno di essi gli studi primari e trae conclusioni dalla presenza/assenza di indipendenti variabili in ciascuno degli studi. Questo è simile al *Line of* di Noblit e Hare [127].
- *sintesi argomentativa*, a cui fa riferimento Kitchenham e Charters [96].

- Il metodo *dell'indagine dei casi* è originariamente definito per i casi di studio, ma può essere applicato anche a esperimenti disomogenei. Aggrega la ricerca esistente applicando uno strumento di indagine di domande specifiche a ciascuno studio primario [114], simile all'estrazione dei dati menzionata sopra. I dati dell'indagine sono quantitativi e quindi l'aggregazione viene eseguita utilizzando metodi statistici [108]. • *La metaetnografia* traduce gli studi l'uno nell'altro e sintetizza le traduzioni in concetti che vanno oltre i singoli studi. Le interpretazioni e le spiegazioni negli studi primari sono trattate come dati nello studio metaetnografico. Questo è simile alla *traduzione reciproca* e alla *sintesi confutativa* di Noblit e Hare [127], a cui fa riferimento Kitchenham e Charters [96]. • *La meta-analisi* è, come accennato in precedenza, basata su metodi statistici per integrare dati quantitativi provenienti da diversi casi. • *L'analisi di scoping* mira a fornire una panoramica della ricerca in un campo, piuttosto che sintetizzare i risultati della ricerca. Lo scoping viene anche chiamato studio di mappatura, ulteriormente discusso nella Sez. 4.4.

Indipendentemente dal metodo di sintesi, dovrebbe essere effettuata *un'analisi di sensibilità* per analizzare se i risultati sono coerenti tra i diversi sottoinsiemi di studi. I sottoinsiemi di studi possono essere, ad esempio, solo studi primari di alta qualità, studi primari di tipo particolare o studi primari con buoni rapporti, che presentano tutti i dettagli necessari.

4.3 Segnalazione della Revisione

Come ogni altro studio empirico, la revisione sistematica della letteratura può essere riferita a pubblici diversi. In particolare, se lo scopo della revisione è influenzare i professionisti, il formato del rapporto deve essere adattato adeguatamente al suo pubblico.

Kitchenham e Charters [96] elencano le seguenti forme di diffusione rivolte ai professionisti:

1. Giornali e riviste rivolte ai professionisti
2. Comunicati stampa alla stampa popolare e specializzata
3. Brevi opuscoli riassuntivi
4. Poster
5. Pagine Web
6. Comunicazione diretta agli enti interessati

Per il pubblico accademico, la rendicontazione dettagliata delle procedure per lo studio è fondamentale per la capacità di valutare e valutare la qualità della revisione sistematica della letteratura. La segnalazione include idealmente le modifiche al protocollo di studio, gli elenchi completi degli studi primari inclusi ed esclusi, i dati sulla loro classificazione, nonché i dati grezzi derivati da ciascuno degli studi primari. Se i vincoli di spazio non consentono la pubblicazione di tutti i dettagli, si consiglia di pubblicare online una relazione tecnica di supporto. Una struttura dettagliata per il rapporto accademico è proposta da Kitchenham e Charters [96].

4.4 Studi di mappatura

Se la domanda di ricerca per la revisione della letteratura è più ampia, o il campo di studio è meno esplorato, è possibile avviare uno *studio di mappatura* invece di una revisione sistematica della letteratura. Uno studio di mappatura [131], a volte indicato come *studio di scoping* [96], ricerca un campo più ampio per qualsiasi tipo di ricerca, al fine di ottenere una panoramica dello stato dell'arte o dello stato della pratica su un argomento.

Uno studio di mappatura segue lo stesso processo di principio delle revisioni sistematiche della letteratura, ma ha criteri diversi per inclusioni/esclusioni e qualità. A causa della sua portata più ampia e della diversa tipologia di studi, i dati raccolti e la sintesi tendono ad essere più qualitativi rispetto alle revisioni sistematiche della letteratura. Tuttavia, per il contributo e la rilevanza di uno studio cartografico è importante che l'analisi vada oltre la pura statistica descrittiva e metta in relazione le tendenze e le osservazioni con le esigenze del mondo reale.

Kitchenham et al. [106] hanno fornito un riassunto delle caratteristiche chiave degli studi di mappatura rispetto alle revisioni sistematiche della letteratura, presentato nella Tabella 4.1.

4.5 Recensioni di esempio

Kitchenham et al. riportano 53 revisioni sistematiche uniche della letteratura sull'ingegneria del software pubblicate tra il 2004 e il 2008 [103,104]. Concludono che vi è una crescita del numero di revisioni sistematiche della letteratura pubblicate e che anche la qualità delle revisioni tende ad aumentare. Tuttavia, esiste ancora una grande differenza tra coloro che sono a conoscenza e utilizzano linee guida sistematiche per la propria condotta e coloro che non fanno riferimento ad alcuna linea guida.

In una di queste revisioni sistematiche della letteratura, Sjøberg et al. [161] esaminano gli studi sperimentali condotti nell'ingegneria del software. Hanno effettuato ricerche su nove riviste e tre atti di conferenze nel decennio dal 1993 al 2002, esaminando 5.453 articoli per identificare 103 esperimenti, ovvero l'1,9% degli articoli presentati come esperimenti. Le due categorie di ricerca più frequenti sono Ciclo di vita/ingegneria del software (49%) e Metodi/Tecniche (32%) classificati secondo lo schema di Glass et al. [63]. Ciò è dovuto al numero relativamente elevato di esperimenti rispettivamente sulle tecniche di ispezione e sulle tecniche di progettazione orientata agli oggetti.

Utilizzando lo stesso insieme di studi primari, Dyba et al. [49] hanno esaminato il potere statistico negli esperimenti di ingegneria del software e Hannay et al. [72] hanno esaminato l'uso della teoria nell'ingegneria del software. Dieste et al. [43] hanno studiato diverse strategie di ricerca sullo stesso insieme di studi, se si dovrebbero cercare titoli, abstract o testi completi, e anche aspetti relativi a quali database cercare.

I primi tentativi di sintetizzare cinque esperimenti sulle tecniche di ispezione da parte di Hayes [74] e Miller [121] indicano che gli esperimenti di ingegneria del software in questo campo non sono sufficientemente omogenei da consentire l'applicazione di metodi statistici

4.5 Recensioni di esempio

53

Passo	Descrizione	Bottone
1. Definisci obiettivi	Obiettivi	Definisci obiettivo
2. Definisci domande	Domande	Definisci domanda
3. Analizza	Ambito	Analisi
4. Analizza	Strategia	Analisi
5. Analizza	Requisiti	Analisi
6. Analizza	Prova	Analisi
7. Analizza	Risultati	Analisi

meta-analisi. Concludono inoltre che i dati grezzi devono essere resi disponibili ai metaanalisti, così come ulteriori informazioni non pubblicate dagli autori primari dello studio.

In una revisione della letteratura più recente sull'efficacia della programmazione in coppia, Hannay et al. [73] hanno condotto una meta-analisi sui dati di 18 studi primari. Riportano analisi separate per tre costrutti di risultato: qualità, durata e impegno.

Visualizzano anche i risultati utilizzando gli appezzamenti forestali.

4.6 Esercizi

4.1. Qual è la differenza tra una revisione sistematica della letteratura e una revisione più generale della letteratura?

4.2. Quali strategie di ricerca esistono per gli studi primari?

4.3. Perché due ricercatori dovrebbero condurre alcuni degli stessi passaggi in modo sistematico articolo di letteratura?

4.4. Quali requisiti sono stabiliti per gli studi primari da includere in una meta-analisi?

4.5. Quali sono le differenze chiave tra uno studio sistematico della letteratura e uno studio di mappatura?

Capitolo 5

Casi di studio

Il termine "caso di studio" appare di tanto in tanto nel titolo o negli abstract degli articoli di ricerca sull'ingegneria del software. Tuttavia, gli studi presentati spaziano da studi sul campo molto ambiziosi e ben organizzati, a piccoli esempi di giocattoli che pretendono di essere casi di studio. Questi ultimi dovrebbero preferibilmente essere definiti esempi o illustrazioni. Inoltre, esistono diverse tassonomie utilizzate per classificare la ricerca.

Il termine studio di caso viene utilizzato parallelamente a termini come studio sul campo e studio osservazionale, ciascuno incentrato su un aspetto particolare della metodologia di ricerca. Ad esempio, Lethbridge et al. utilizzano *studi sul campo* come termine più generale [111], mentre Easterbrook et al. chiamano *i casi di studio* una delle cinque "classi di metodi di ricerca" [50]. Zelkowitz e Wallace propongono una terminologia leggermente diversa da quella utilizzata in altri campi e classificano il monitoraggio del progetto, lo studio di caso e lo studio sul campo come *metodi osservativi* [181]. Questa pletora di termini causa confusione e problemi quando si tenta di aggregare più studi empirici.

La metodologia del caso di studio è adatta a molti tipi di ricerca sull'ingegneria del software, poiché gli oggetti di studio sono fenomeni contemporanei, difficili da studiare isolatamente. I casi di studio non generano gli stessi risultati, ad esempio, sulle relazioni causali degli esperimenti controllati, ma forniscono una comprensione più profonda dei fenomeni studiati nel loro contesto reale. Poiché sono diversi dagli studi empirici analitici e controllati, gli studi di casi sono stati criticati per essere di minor valore, impossibili da generalizzare, influenzati dai ricercatori, ecc. La critica può essere affrontata applicando pratiche di metodologia di ricerca adeguate e accettando che la conoscenza non è solo significatività statistica [59, 109].

L'obiettivo di questo capitolo è fornire alcune indicazioni per il ricercatore che conduce casi di studio. Questo capitolo è basato su Runeson e Host [" 145] e maggiori dettagli sui casi di studio nell'ingegneria del software possono essere ottenuti da Runeson et al. [146]. Nello specifico, le liste di controllo per i ricercatori vengono derivate attraverso un'analisi sistematica delle liste di controllo esistenti [79, 145] e successivamente valutate dal dottorato, studenti e dai membri dell'International Software Engineering Research Network e aggiornati di conseguenza.

Il capitolo non fornisce affermazioni assolute su quello che è considerato un "buon" caso di studio nell'ingegneria del software. Piuttosto si concentra su una serie di questioni che tutto

contribuire alla qualità della ricerca. Il requisito minimo per ciascuna questione deve essere giudicato nel suo contesto e molto probabilmente evolverà nel tempo.

Il capitolo è strutturato come segue. Per prima cosa introduciamo il contesto della ricerca sui casi di studio, discutiamo le motivazioni per i casi di studio sull'ingegneria del software e definiamo un processo di ricerca sui casi di studio nella Sez. 5.1. La sezione 5.2 discute la progettazione di un caso di studio e la pianificazione della raccolta dei dati. La sezione 5.3 descrive il processo di raccolta dei dati. Nella sez. 5.4 vengono trattate le questioni relative all'analisi dei dati e il reporting è discusso nella Sez. 5.5.

5.1 Casi di studio nel suo contesto

Tre definizioni comunemente usate di ricerca di casi di studio sono fornite da Robson [144], Yin [180] e Benbasat et al. [22] rispettivamente. Le tre definizioni concordano nel ritenere che il case study sia un metodo empirico volto a *indagare i fenomeni contemporanei* nel loro contesto. Robson la definisce una strategia di ricerca e sottolinea l'uso di *molteplici fonti di prova*, Yin la definisce un'indagine e osserva che *il confine tra il fenomeno e il suo contesto potrebbe non essere chiaro*, mentre Benbasat et al. rendono le definizioni un po' più specifiche, menzionando *la raccolta di informazioni da poche entità* (persone, gruppi, organizzazioni) e la *mancanza di controllo sperimentale*.

La ricerca-azione è strettamente correlata alla ricerca di casi di studio con lo scopo di "influenzare o modificare alcuni aspetti di qualunque sia il focus della ricerca" [144].

Più strettamente, uno studio di caso è puramente osservativo mentre la ricerca-azione è focalizzata e coinvolta nel processo di cambiamento. Negli studi sul miglioramento dei processi software [44, 85] e sul trasferimento tecnologico [66], il metodo di ricerca potrebbe essere caratterizzato come ricerca-azione se il ricercatore partecipa attivamente ai miglioramenti.

Tuttavia, quando studiamo gli effetti di un cambiamento, ad esempio, negli studi pre e post evento, classifichiamo la metodologia come caso di studio. Nella ricerca sui sistemi informativi, dove la ricerca-azione è ampiamente utilizzata, si discute su come trovare l'equilibrio tra azione e ricerca, vedere ad esempio Baskerville e Wood-Harper [21] o Avison et al. [5]. Per la parte di ricerca della ricerca-azione, possono essere utilizzate anche queste linee guida per studi di casi.

Easterbrook et al. [50] annoverano anche gli studi etnografici tra le principali metodologie di ricerca. Preferiamo considerare gli studi etnografici come un tipo specializzato di casi di studio focalizzati sulle pratiche culturali [50] o studi di lunga durata con grandi quantità di dati partecipanti-osservatori [98]. Zelkowitz e Wallace definiscono quattro diversi "metodi osservativi" nell'ingegneria del software [181]; monitoraggio del progetto, studio di caso, asserzione e studio sul campo. Preferiamo considerare il monitoraggio del progetto come parte di un caso di studio e gli studi sul campo come casi di studio multipli, mentre l'asserzione non è considerata un metodo di ricerca accettato.

Robson riassume il suo punto di vista, che sembra funzionale anche nell'ingegneria del software: "Molti studi di progettazione flessibile, sebbene non esplicitamente etichettati come tali, possono essere utilmente visti come casi di studio" [144].

Un caso di studio può contenere elementi di altri metodi di ricerca, ad esempio, un'indagine può essere condotta all'interno di un caso di studio, una ricerca bibliografica spesso precede un caso di studio e le analisi di archivio possono far parte della sua raccolta di dati. I metodi etnografici, come le interviste e le osservazioni, vengono utilizzati principalmente per la raccolta dei dati nei casi di studio.

Yin aggiunge specificatamente alle caratteristiche di un caso di studio che [180]:

- "Gestisce la situazione tecnicamente particolare in cui ce ne saranno molti più variabili rispetto ai punti dati e come risultato
- Si basa su molteplici fonti di prova, con la necessità di convergere i dati in a triangolare la moda, e come altro risultato
- Beneficia dello sviluppo preliminare di proposte teoriche per guidare la raccolta e l'analisi dei dati.

Pertanto, uno studio di caso non fornirà mai conclusioni con significatività statistica.

Al contrario, molti diversi tipi di prove, cifre, dichiarazioni, documenti, sono collegati insieme per supportare una conclusione forte e rilevante.

In sintesi, le caratteristiche chiave di un caso di studio sono le seguenti [146]:

1. È di tipo flessibile, in grado di far fronte alle caratteristiche complesse e dinamiche del reale fenomeni mondiali, come l'ingegneria del software,
2. Le sue conclusioni si basano su una chiara catena di prove, sia qualitative che quantitative, raccolte da molteplici fonti in modo pianificato e coerente, e
3. Si aggiunge alla conoscenza esistente essendo basato su una teoria precedentemente stabilita, se tali esistono, o costruendo la teoria.

5.1.1 Perché casi di studio nell'ingegneria del software?

L'area dell'ingegneria del software comprende lo sviluppo, il funzionamento e la manutenzione del software e dei relativi artefatti. La ricerca sull'ingegneria del software è in larga misura finalizzata a indagare come lo sviluppo, il funzionamento e la manutenzione vengono condotti dagli ingegneri del software e da altre parti interessate in condizioni diverse.

Gli individui, i gruppi e le organizzazioni realizzano lo sviluppo del software e le questioni sociali e politiche sono importanti per questo sviluppo. Cioè, l'ingegneria del software è una disciplina multidisciplinare che coinvolge aree in cui vengono condotti casi di studio, come la psicologia, la sociologia, le scienze politiche, il servizio sociale, l'economia e la pianificazione comunitaria (ad esempio [180]). Ciò significa che molte domande di ricerca nell'ingegneria del software sono adatte per la ricerca di casi di studio.

La definizione di caso di studio nella Sez. 2.1 si concentra sullo studio dei fenomeni nel loro contesto, soprattutto quando il confine tra il fenomeno e il suo contesto non è chiaro. Ciò è particolarmente vero nell'ingegneria del software. La sperimentazione nell'ingegneria del software ha chiaramente dimostrato che ci sono molti fattori che influenzano il risultato di un'attività di ingegneria del software, ad esempio, quando si tenta di replicare gli studi, vedere Sez. 2.6. I casi di studio offrono un approccio che non necessita di rigore

confine tra l'oggetto studiato e il suo ambiente; forse la chiave per la comprensione sta nell'interazione tra i due?

5.1.2 Processo di ricerca sul caso di studio

Quando si conduce un caso di studio, ci sono cinque fasi principali del processo da seguire:

1. Progettazione del caso di studio: vengono definiti gli obiettivi e pianificato il caso di studio.
2. Preparazione per la raccolta dei dati: procedure e protocolli per la raccolta dei dati sono definiti.
3. Raccolta dati: esecuzione con raccolta dati sul caso studiato.
4. Analisi dei dati raccolti 5. Reporting

Questo processo è quasi lo stesso per qualsiasi tipo di studio empirico; si confronti, ad esempio, con il processo delineato nel cap. 6 e ulteriormente approfondito nei capp. 7–11 per gli esperimenti e Kitchenham et al. [98]. Tuttavia, poiché la metodologia del caso di studio è una strategia di progettazione flessibile, è necessaria una quantità significativa di iterazione nei passaggi [2].

La raccolta e l'analisi dei dati possono essere condotte in modo incrementale. Se vengono raccolti dati insufficienti per l'analisi, è possibile pianificare una maggiore raccolta di dati, ecc. Eisenhardt aggiunge due passaggi tra 4 e 5 sopra nel suo processo di costruzione di teorie dalla ricerca di casi di studio [52] (a) modellare ipotesi e (b) racchiudere la letteratura , mentre il resto, salvo variazioni terminologiche, è identico a quanto sopra.

Le cinque fasi del processo sono presentate nelle Sez. 5.2–5.5, dove preparazione e la raccolta dei dati è presentata in una sezione comune, ovvero la Sez. 5.3.

5.2 Progettazione e Pianificazione

La ricerca sui casi di studio è di tipo flessibile ma ciò non significa che la pianificazione non sia necessaria. Al contrario, una buona pianificazione di un caso di studio è fondamentale per il suo successo. Ci sono diverse questioni che devono essere pianificate, ad esempio quali metodi utilizzare per la raccolta dei dati, quali dipartimenti di un'organizzazione visitare, quali documenti leggere, quali persone intervistare, quanto spesso dovrebbero essere condotte le interviste, ecc.

Questi piani possono essere formulati in un protocollo di studio di caso, vedere Sez. 5.2.2.

5.2.1 Pianificazione del caso di studio

Un piano per uno studio di caso dovrebbe contenere almeno i seguenti elementi [144]:

- *Obiettivo*: cosa ottenere? • *Il caso*: cosa si studia?

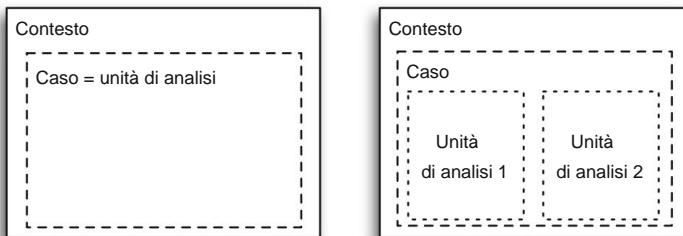


Fig. 5.1 Caso di studio olistico (a sinistra) e caso di studio integrato (a destra)

- *Teoria:* quadro di riferimento •
- Domande di ricerca:* cosa sapere? • *Metodi:* come raccogliere i dati? • *Strategia di selezione:* dove cercare i dati?

L'obiettivo dello studio può essere, ad esempio, esplorativo, descrittivo, esplicativo o migliorativo. L'obiettivo è naturalmente formulato in modo più generale e meno preciso rispetto ai disegni di ricerca fissi. Inizialmente l'obiettivo è più simile a un punto focale che evolve durante lo studio. Le domande di ricerca indicano ciò che è necessario sapere per raggiungere l'obiettivo dello studio. Analogamente all'obiettivo, le domande di ricerca evolvono durante lo studio e vengono ristrette a domande di ricerca specifiche durante le iterazioni dello studio [2].

Nell'ingegneria del software, il caso potrebbe essere un progetto di sviluppo software, che è la scelta più semplice. Può alternativamente essere un individuo, un gruppo di persone, un processo, un prodotto, una politica, un ruolo nell'organizzazione, un evento, una tecnologia, ecc. Può anche costituire un'unità di analisi all'interno di un caso. Gli studi sui "programmi giocattolo" o simili sono ovviamente esclusi a causa della mancanza di un contesto di vita reale.

Yin [180] distingue tra studi di casi olistici, in cui il caso viene studiato nel suo insieme, e studi di casi integrati in cui più unità di analisi vengono studiate all'interno di un caso, vedere Fig. 5.1. Se definire uno studio composto da due casi come olistico o incorporato dipende da ciò che definiamo come contesto e obiettivi della ricerca.

Ad esempio se si studiano due progetti in due aziende diverse e in due domini applicativi diversi, entrambi utilizzando pratiche agili. Da un lato, i progetti possono essere considerati due unità di analisi in un caso studio incorporato se il contesto è quello delle aziende di software in generale e l'obiettivo della ricerca è studiare pratiche agili. D'altra parte, se si considera il contesto come l'azienda specifica o il dominio applicativo, devono essere visti come due casi olistici separati.

L'utilizzo delle teorie per sviluppare la direzione della ricerca non è ben consolidato nel campo dell'ingegneria del software, come discusso nella sez. 2.7. Tuttavia, la definizione del quadro di riferimento dello studio rende chiaro il contesto della ricerca del caso di studio e aiuta sia coloro che conducono la ricerca sia coloro che ne esaminano i risultati. In mancanza di teoria, il quadro di riferimento può in alternativa essere espresso in termini di

punto di vista adottato nella ricerca e il background dei ricercatori. I casi di studio della teoria fondata naturalmente non hanno una teoria specificata [38].

Le principali decisioni sui metodi di raccolta dei dati sono definite in fase di progettazione del caso di studio, sebbene le decisioni dettagliate sulle procedure di raccolta dei dati vengano prese successivamente. Lethbridge et al. [111] definiscono tre categorie di metodi: diretti (ad esempio interviste), indiretti (ad esempio strumenti strumentali) e indipendenti (ad esempio analisi della documentazione). Questi sono ulteriormente elaborati nella Sez. 5.3.

Negli studi di casi, il caso e le unità di analisi dovrebbero essere selezionati intenzionalmente. Ciò è in contrasto con le indagini e gli esperimenti, in cui i soggetti vengono campionati da una popolazione alla quale si intende generalizzare i risultati. Lo scopo della selezione può essere quello di studiare un caso che si prevede sia "tipico", "critico", "rivelatore" o "unico" sotto qualche aspetto [22], e il caso viene selezionato di conseguenza. In un caso di studio comparativo, le unità di analisi devono essere selezionate per avere la variazione delle proprietà che lo studio intende confrontare. Tuttavia, in pratica, molti casi vengono selezionati in base alla disponibilità [22], il che è simile per gli esperimenti [161].

La selezione dei casi è particolarmente importante quando si replicano i casi di studio. Un caso di studio può essere replicato letteralmente, ovvero il caso viene selezionato per prevedere risultati simili, oppure può essere replicato teoricamente, ovvero il caso viene selezionato per prevedere risultati contrastanti per ragioni prevedibili [180].

5.2.2 Protocollo di studio del caso

Il protocollo del caso studio è un contenitore per le decisioni progettuali sul caso studio nonché le procedure sul campo per portare avanti lo studio. Il protocollo è un documento in continua evoluzione che viene aggiornato quando i piani per il caso di studio vengono modificati e serve a diversi scopi:

1. Serve da guida durante la raccolta dei dati e in questo modo impedisce al ricercatore di perdere la raccolta dei dati che era stato pianificato di raccogliere.
2. I processi di formulazione del protocollo rendono la ricerca concreta nella fase di pianificazione, che può aiutare il ricercatore a decidere quali fonti di dati utilizzare e quali domande porre.
3. Altri ricercatori e persone interessate possono esaminarlo per fornire un feedback sui piani. Il feedback sul protocollo da parte di altri ricercatori può, ad esempio, ridurre il rischio di perdere fonti di dati rilevanti, domande di intervista o ruoli da includere nella ricerca e di mettere in discussione la relazione tra domande di ricerca e domande di intervista.
4. Può fungere da registro o diario in cui vengono registrate tutte le raccolte e le analisi dei dati insieme alle decisioni di modifica basate sulla natura flessibile della ricerca.
Questa può essere un'importante fonte di informazioni quando il caso di studio verrà riportato in seguito. Per tenere traccia dei cambiamenti durante il progetto di ricerca, il protocollo dovrebbe essere mantenuto sotto qualche forma di controllo della versione.

Tabella 5.1 Schema del protocollo del caso di studio secondo Brereton et al. [25]

Sezione	Contenuto
Sfondo	Ricerche precedenti, domande di ricerca principali e aggiuntive
Progetto	Design a caso singolo o multiplo, integrato o olistico; oggetto di studio; proposizioni derivate da domande di ricerca
Selezione	Criteri per la selezione dei casi
Procedure e ruoli	Procedure sul campo; Ruoli dei membri del gruppo di ricerca
Raccolta dati	Identificare i dati, definire il piano di raccolta e archiviazione dei dati
Analisi	Criteri di interpretazione, collegamento tra dati e domande di ricerca, spiegazioni alternative
Validità del piano	Tattiche per ridurre le minacce alla validità
Limitazioni dello studio	Specificare i problemi di validità rimanenti
Segnalazione	Destinatari
Programma	Stime per le tappe principali
Appendici	Qualsiasi informazione dettagliata

Brereton et al. [25] propongono uno schema di un protocollo di studio di caso, che è riassunto nella tabella 5.1. Come mostra la proposta, il protocollo è piuttosto dettagliato sostenere un approccio di ricerca ben strutturato.

5.3 Preparazione e raccolta dei dati

Esistono diverse fonti di informazioni che possono essere utilizzate in un caso di studio. È importante utilizzare diverse fonti di dati in un caso di studio al fine di limitare il effetti di un'interpretazione di una singola fonte di dati. Se la stessa conclusione può essere ricavato da diverse fonti di informazione, ad esempio la triangolazione (brevemente descritta in il contesto degli esperimenti nella Sez. 6.2), questa conclusione è più forte di una conclusione sulla base di un'unica fonte. In un caso di studio, è anche importante tenerne conto punti di vista di ruoli diversi e per indagare le differenze, ad esempio tra progetti e prodotti diversi. Comunemente, le conclusioni vengono tratte analizzando differenze tra le fonti dei dati.

Secondo Lethbridge et al. [111], le tecniche di raccolta dei dati possono essere suddivise in tre livelli:

- *Primo grado*: metodi diretti significa che il ricercatore è in contatto diretto con i soggetti e raccogliere dati in tempo reale. Questo è il caso, ad esempio interviste, focus group, sondaggi Delphi [40] e osservazioni con "pensare ad alta voce". protocolli" [129].
- *Secondo grado*: metodi indiretti in cui il ricercatore raccoglie direttamente i dati grezzi senza interagire effettivamente con i soggetti durante la raccolta dei dati. Esempi sono la registrazione dell'utilizzo di strumenti di ingegneria del software e le osservazioni attraverso la registrazione video.

- *Terzo grado:* analisi indipendente degli artefatti di lavoro in cui vengono utilizzati dati già disponibili e talvolta compilati. Questo è ad esempio il caso quando vengono analizzati documenti come le specifiche dei requisiti e le segnalazioni di guasti di un'organizzazione o quando vengono analizzati i dati provenienti da database organizzativi come la contabilità del tempo.

I metodi di primo grado sono per lo più più costosi da applicare rispetto ai metodi di secondo o terzo grado, poiché richiedono uno sforzo significativo sia da parte del ricercatore che dei soggetti. Un vantaggio dei metodi di primo e secondo grado è che il ricercatore può in larga misura controllare esattamente quali dati vengono raccolti, come vengono raccolti, in quale forma vengono raccolti i dati, qual è il contesto, ecc. I metodi di terzo grado sono per lo più meno costosi, ma non offrono lo stesso controllo al ricercatore; quindi neanche la qualità dei dati è sotto controllo, né per quanto riguarda la qualità dei dati originali né il loro utilizzo per lo scopo del caso di studio. In molti casi il ricercatore deve, in una certa misura, basare i dettagli della raccolta dei dati sui dati disponibili. Per i metodi di terzo grado, va anche notato che i dati sono stati raccolti e registrati per uno scopo diverso da quello dello studio di ricerca, contrariamente alle linee guida generali della metrica [172]. Non è certo che i requisiti sulla validità e completezza dei dati fossero gli stessi quando i dati sono stati raccolti come nello studio di ricerca.

Nelle sez. [5.3.1–5.3.4](#), discutiamo metodi specifici di raccolta dati, in cui abbiamo riscontrato interviste, osservazioni, dati di archivio e metriche applicabili a casi di studio di ingegneria del software [22, 146, 180].

5.3.1 Interviste

Nella raccolta dati basata su interviste, il ricercatore pone una serie di domande a una serie di soggetti sulle aree di interesse nel caso di studio. Nella maggior parte dei casi viene condotta un'intervista per ogni singolo soggetto, ma è possibile condurre interviste di gruppo. Il dialogo tra il ricercatore e il/i soggetto/i è guidato da una serie di domande dell'intervista.

Le domande dell'intervista si basano sulle domande di ricerca (sebbene non siano formulate nello stesso modo). Le domande possono essere *aperte*, ovvero consentendo e invitando a un'ampia gamma di risposte e domande da parte del soggetto intervistato, oppure *chiuse* offrendo un insieme limitato di risposte alternative.

Le interviste possono essere suddivise in interviste *non strutturate*, *semi-strutturate* e *completamente strutturate* [144]. In un'intervista non strutturata, le domande dell'intervista sono formulate come preoccupazioni e interessi generali del ricercatore. In questo caso la conversazione dell'intervista si svilupperà in base all'interesse del soggetto e del ricercatore. In un colloquio completamente strutturato, tutte le domande vengono pianificate in anticipo e tutte le domande vengono poste nello stesso ordine del piano. In molti modi, un'intervista completamente strutturata è simile a un sondaggio basato su questionari. In un'intervista semistrutturata, le domande sono pianificate, ma non necessariamente vengono poste nello stesso ordine in cui sono elencate.

Tabella 5.2 Panoramica delle tipologie di intervista

	Non strutturato	Semistrutturato	Completamente strutturato
Focali tipici	Come gli individui sperimentano qualitativamente il fenomeno	Come gli individui vivono qualitativamente e quantitativamente il fenomeno	Il ricercatore cerca di trovare le relazioni tra i costrutti
Domande dell'intervista	Guida all'intervista con le aree su cui concentrarsi	Mix di domande aperte e chiuse	Domande chiuse
Obiettivo	Esplorativo	Descrittivo ed esplicativo	Descrittivo ed esplicativo

Lo sviluppo della conversazione nell'intervista può decidere in quale ordine vengono gestite le diverse domande, e il ricercatore può utilizzare l'elenco delle domande per essere certo che tutte le domande vengano gestite, cioè più o meno come una lista di controllo. Inoltre, le interviste semistruzzurate consentono l'improvvisazione e l'esplorazione degli oggetti studiati. Le interviste semi-strutturate sono comuni nei casi di studio. I tre tipi di interviste sono riassunti nella Tabella 5.2.

Una sessione di intervista può essere suddivisa in più fasi. Innanzitutto il ricercatore presenta gli obiettivi dell'intervista e del caso studio e spiega come verranno utilizzati i dati dell'intervista. Quindi viene posta una serie di domande introduttive sullo sfondo ecc. dell'argomento; rispondere a queste domande è relativamente semplice.

Dopo l'introduzione vengono poste le domande principali dell'intervista, che occupano la maggior parte dell'intervista. Se l'intervista contiene domande personali e magari delicate, ad esempio riguardanti l'economia, le opinioni sui colleghi, il motivo per cui le cose sono andate male, o domande relative alle competenze dell'intervistato [80], è importante che all'intervistato sia garantita la riservatezza e che l'intervistato si fida dell'intervistatore. Si consiglia di iniziare il colloquio con queste domande o di introdurle prima che si sia instaurato un clima di fiducia. Si raccomanda al ricercatore di riassumere i principali risultati verso la fine dell'intervista, al fine di ottenere feedback ed evitare malintesi.

Durante le sessioni di colloquio si consiglia di registrare la discussione in un formato audio o video idoneo. Anche se si prendono appunti, in molti casi è difficile registrare tutti i dettagli ed è impossibile sapere cosa è importante registrare durante il colloquio. Una volta registrata, l'intervista deve essere trascritta in testo prima di essere analizzata. In alcuni casi può essere vantaggioso che le trascrizioni siano revisionate dal soggetto dell'intervista.

Durante la fase di pianificazione di uno studio di intervista si decide chi intervistare.

A causa della natura qualitativa del caso studio si raccomanda di selezionare i soggetti in base alle differenze invece di cercare di replicare le somiglianze, come discusso nella Sez. 5.2. Ciò significa che è bene cercare di coinvolgere nel colloquio ruoli, personalità, ecc. diversi. Il numero degli intervistati dovrà essere deciso durante lo studio.

Un criterio per stabilire quando vengono condotte interviste sufficienti è la "saturazione", vale a dire quando non si ottengono nuove informazioni o punti di vista da nuovi soggetti [38].

Tabella 5.3 Diversi approcci alle osservazioni Elevata

	consapevolezza di essere osservato	Scarsa consapevolezza di essere osservato
Alto grado di interazione da parte del ricercatore	Categoria 1	Categoria 2
Basso grado di interazione da parte del ricercatore	Categoria 3	Categoria 4

5.3.2 Osservazioni

È possibile condurre osservazioni per indagare su come gli ingegneri del software svolgono un determinato compito. Si tratta di un metodo di primo o secondo grado a seconda della classificazione sopra riportata. Esistono molti approcci diversi per l'osservazione. Un approccio consiste nel monitorare un gruppo di ingegneri del software con un videoregistratore e successivamente analizzare la registrazione. Un'altra alternativa è applicare un protocollo "pensa ad alta voce", in cui il ricercatore pone ripetutamente domande come "Qual è la tua strategia?" e "A cosa stai pensando?" per ricordare ai soggetti di pensare ad alta voce. Ciò può essere combinato con la registrazione di audio e sequenze di tasti come proposto, ad esempio, da Wallace et al. [176]. Un altro tipo sono le osservazioni durante le riunioni, in cui i partecipanti alla riunione interagiscono tra loro e quindi generano informazioni sull'oggetto studiato. Karahasanovic et al. [93] presentano un approccio alternativo in cui viene utilizzato uno strumento di campionamento per ottenere dati e feedback dai partecipanti.

Gli approcci per le osservazioni possono essere suddivisi in alta o bassa interazione del ricercatore e alta o bassa consapevolezza dei soggetti da osservare, vedere Tabella 5.3.

Le osservazioni secondo la categoria 1 o la categoria 2 sono tipicamente condotte nella ricerca-azione o negli studi etnografici classici in cui il ricercatore è parte del team e non è visto solo come ricercatore dagli altri membri del team. La differenza tra la categoria 1 e la categoria 2 è che nella categoria 1 il ricercatore è visto come un "partecipante osservatore" dagli altri soggetti, mentre è visto più come un "partecipante normale" nella categoria 2. Nella categoria 3 il ricercatore è visto solo come ricercatore.

Gli approcci per l'osservazione includono tipicamente osservazioni con tecniche di raccolta dati di primo grado, come un protocollo "pensare ad alta voce" come descritto sopra.

Nella categoria 4 i soggetti vengono tipicamente osservati con una tecnica di secondo grado come la registrazione video (a volte chiamata videoetnografia).

Un vantaggio delle osservazioni è che possono fornire una comprensione profonda del fenomeno studiato. Inoltre, è particolarmente rilevante utilizzare le osservazioni, laddove si sospetta che vi sia una deviazione tra una visione "ufficiale" della questione e il caso "reale" [142]. Va tuttavia notato che produce una notevole quantità di dati che rende l'analisi dispendiosa in termini di tempo.

5.3.3 Dati di archivio

I dati di archivio si riferiscono, ad esempio, a verbali di riunioni, documenti di diverse fasi di sviluppo, dati di guasti, organigrammi, registri finanziari e altre misurazioni precedentemente raccolte in un'organizzazione.

I dati di archivio sono un tipo di dati di terzo grado che possono essere raccolti in un caso di studio. Per questo tipo di dati uno strumento di gestione della configurazione è una fonte importante, poiché consente la raccolta di un numero di documenti diversi e di diverse versioni di documenti. Come per altre fonti di dati di terzo grado è importante tenere presente che i documenti non sono stati originariamente sviluppati con l'intento di fornire dati per la ricerca. Naturalmente è difficile per il ricercatore valutare la qualità dei dati, anche se alcune informazioni possono essere ottenute indagando lo scopo della raccolta dati originale e intervistando le persone rilevanti all'interno dell'organizzazione.

5.3.4 Metriche

Le tecniche di raccolta dati sopra menzionate si concentrano principalmente su dati qualitativi. Tuttavia, anche i dati quantitativi sono importanti in un caso di studio. I dati raccolti possono essere definiti o raccolti ai fini del caso di studio, oppure i dati già disponibili possono essere utilizzati in un caso di studio. Il primo caso offre, ovviamente, la massima flessibilità e i dati più adatti alle domande di ricerca in esame. La definizione di quali dati raccogliere dovrebbe essere basata su una tecnica di misurazione orientata agli obiettivi, come il metodo Goal Question Metric (GQM) [11, 172], presentato nel Cap. 3.

Esempi di dati già disponibili sono dati sull'impegno di progetti più vecchi, dati di vendita di prodotti, parametri di qualità del prodotto in termini di guasti, ecc. Questo tipo di dati può, ad esempio, essere disponibile in un database di parametri in un'organizzazione.

Sì noti tuttavia che il ricercatore non può né controllare né valutare la qualità dei dati, poiché sono stati raccolti per un altro scopo e, come per altre forme di analisi archivistica, esiste il rischio di perdere dati importanti.

5.4 Analisi dei dati

5.4.1 Analisi quantitativa dei dati

L'analisi dei dati viene condotta in modo diverso per i dati quantitativi e qualitativi. Per i dati quantitativi, l'analisi include tipicamente l'analisi delle statistiche descrittive, l'analisi delle correlazioni, lo sviluppo di modelli predittivi e la verifica delle ipotesi. Tutte queste attività sono rilevanti nella ricerca di casi di studio. L'analisi quantitativa dei dati, sebbene principalmente in un contesto sperimentale, è ulteriormente descritta nel Cap. 10.

Le statistiche descrittive, come valori medi, deviazioni standard, istogrammi e grafici a dispersione, vengono utilizzate per comprendere i dati raccolti.

Vengono condotti l'analisi di correlazione e lo sviluppo di modelli predittivi per descrivere come una misurazione di un'attività di processo successiva è correlata a una misurazione di processo precedente. Il test delle ipotesi viene condotto per determinare se esiste un effetto significativo di una o più variabili (variabili indipendenti) su una o più altre variabili (variabili dipendenti).

Va notato che i metodi per l'analisi quantitativa presuppongono un disegno di ricerca fisso. Ad esempio, se in una serie di interviste una domanda con risposta quantitativa viene modificata a metà, ciò rende impossibile interpretare il valore medio delle risposte. Inoltre, i set di dati quantitativi provenienti da singoli casi tendono ad essere molto piccoli, a causa del numero di intervistati o di punti di misurazione, il che causa particolari preoccupazioni nell'analisi.

5.4.2 Analisi qualitativa dei dati

L'obiettivo fondamentale dell'analisi qualitativa è trarre conclusioni dai dati, mantenendo una chiara catena di prove. La catena di prove significa che un lettore dovrebbe essere in grado di seguire la derivazione dei risultati e delle conclusioni dai dati raccolti [180]. Ciò significa che devono essere presentate informazioni sufficienti su ogni fase dello studio e su ogni decisione presa dal ricercatore.

Inoltre, l'analisi della ricerca qualitativa è caratterizzata dal fatto che l'analisi viene condotta parallelamente alla raccolta dei dati e dalla necessità di tecniche di analisi sistematiche. L'analisi deve essere effettuata parallelamente alla raccolta dei dati poiché l'approccio è flessibile e durante l'analisi vengono trovate nuove informazioni. Per indagare su queste intuizioni, è spesso necessario raccogliere nuovi dati e aggiornare strumenti come i questionari delle interviste. La necessità di essere sistematici è una conseguenza diretta del fatto che le tecniche di raccolta dei dati possono essere costantemente aggiornate, pur essendo necessario mantenere una catena di prove.

Al fine di ridurre i pregiudizi da parte dei singoli ricercatori, l'analisi trae vantaggio dal fatto di essere condotta da più ricercatori. I risultati preliminari di ogni singolo ricercatore vengono uniti in un risultato di analisi comune in una seconda fase. Tenere traccia e riferire sullo schema di cooperazione aiuta ad aumentare la validità dello studio.

Tecniche generali di analisi. Esistono due diverse parti dell'analisi dei dati qualitativi, tecniche di generazione di ipotesi e tecniche di conferma di ipotesi [148].

La generazione di ipotesi ha lo scopo di trovare ipotesi dai dati. Quando si utilizzano questo tipo di tecniche, il ricercatore dovrebbe cercare di essere imparziale e aperto a qualsiasi ipotesi si possa trovare nei dati. I risultati di queste tecniche sono le ipotesi in quanto tali. Esempi di tecniche di generazione di ipotesi sono i "confronti costanti" e l'"analisi incrociata di casi" [148]. Le tecniche *di conferma* dell'ipotesi denotano tecniche che possono essere utilizzate per confermare che un'ipotesi è realmente vera, per

esempio, attraverso l'analisi di più dati. La triangolazione e la replica sono esempi di approcci per la conferma delle ipotesi [148]. L'*analisi dei casi negativi* cerca di trovare spiegazioni alternative che respingono le ipotesi. Questi tipi base di tecniche vengono utilizzati in modo iterativo e in combinazione. Si generano le prime ipotesi e poi si confermano. La generazione dell'ipotesi può avvenire all'interno di un ciclo di un caso di studio, o con i dati di un'unità di analisi, e la conferma dell'ipotesi può essere effettuata con i dati di un altro ciclo o unità di analisi [2].

Ciò significa che l'analisi dei dati qualitativi viene condotta in una serie di passaggi (sulla base di Robson [144]). Innanzitutto i dati vengono codificati, il che significa che a partì del testo può essere assegnato un codice che rappresenta un determinato tema, area, costrutto, ecc. Un codice viene solitamente assegnato a molte parti di testo e a una parte di testo possono essere assegnate più di un codice. I codici possono formare una gerarchia di codici e sottocodici. Il materiale codificato può essere combinato con commenti e riflessioni del ricercatore (cioè 'memo'). Fatto ciò, il ricercatore può esaminare il materiale per identificare una prima serie di ipotesi. Possono trattarsi, ad esempio, di frasi simili in diverse parti del materiale, di modelli nei dati, di differenze tra sottogruppi di soggetti, ecc. Le ipotesi individuate possono poi essere utilizzate quando viene condotta un'ulteriore raccolta di dati sul campo, ciò si traduce in un approccio iterativo in cui la raccolta e l'analisi dei dati vengono condotte in parallelo come descritto sopra. Durante il processo iterativo è possibile formulare un piccolo insieme di generalizzazioni, che alla fine danno luogo ad un insieme di conoscenze formalizzato, che rappresenta il risultato finale del tentativo di ricerca. Naturalmente non si tratta di una semplice sequenza di passaggi. Vengono invece eseguiti in modo iterativo e si influenzano a vicenda.

Un esempio di tecnica utile per l'analisi è la tabulazione, in cui i dati codificati sono organizzati in tabelle, che consentono di ottenere una panoramica dei dati.

I dati possono, ad esempio, essere organizzati in una tabella in cui le righe rappresentano i codici di interesse e le colonne rappresentano gli argomenti dell'intervista. Tuttavia, come farlo deve essere deciso per ogni caso di studio.

Sono disponibili strumenti software specializzati per supportare l'analisi qualitativa dei dati, ad esempio NVivo¹ e Atlas.² Tuttavia, in alcuni casi strumenti standard come elaboratori di testo e fogli di calcolo sono utili quando si gestiscono i dati testuali.

Livello di formalismo. Un approccio strutturato è, come descritto sopra, importante nell'analisi qualitativa. Tuttavia, l'analisi può essere condotta a diversi livelli di formalismo. Robson [144] menziona i seguenti approcci:

- *Approcci di immersione*: questi sono gli approcci meno strutturati, con un livello di struttura molto basso, più dipendenti dall'intuizione e dalle capacità interpretative del ricercatore. Questi approcci possono essere difficili da combinare con i requisiti relativi alla conservazione e alla comunicazione di una catena di prove.
- *Approcci di editing*: questi approcci includono pochi codici a priori, cioè i codici sono definiti sulla base dei risultati del ricercatore durante l'analisi.

¹<http://www.qsrinternational.com>

² <http://www.atlasti.com>

- *Approcci modello*: questi approcci sono più formali e includono più a priori basati su domande di ricerca.
- *Approcci quasi statistici*: questi approcci sono molto formalizzati e includono, ad esempio, il calcolo delle frequenze di parole e frasi.

Nella nostra esperienza, gli approcci di modifica e gli approcci basati su modelli sono più adatti nei casi di studio dell'ingegneria del software. È difficile presentare e ottenere una chiara catena di prove negli approcci di immersione informale. È anche difficile interpretare il risultato, ad esempio, della frequenza delle parole nei documenti e nelle interviste.

5.4.3 Validità

La validità di uno studio denota l'affidabilità dei risultati e in che misura i risultati sono veri e non influenzati dal punto di vista soggettivo dei ricercatori. Naturalmente è troppo tardi per considerarne la validità durante l'analisi. La validità deve essere affrontata durante tutte le fasi precedenti dello studio del caso.

Esistono diversi modi per classificare gli aspetti della validità e le minacce alla validità in letteratura. Qui abbiamo scelto uno schema di classificazione, utilizzato anche da Yin [180] per i casi di studio, e simile a quello solitamente utilizzato negli esperimenti controllati nell'ingegneria del software, come ulteriormente elaborato nella Sez. 8.7. Alcuni ricercatori hanno sostenuto la necessità di avere uno schema di classificazione diverso per gli studi di progettazione flessibile (credibilità, trasferibilità, affidabilità e confermabilità), mentre noi preferiamo rendere operativo questo schema per gli studi di progettazione flessibile, invece di cambiare i termini [144]. Questo schema distingue quattro aspetti della validità, che possono essere così riassunti:

- *Validità di costrutto*: questo aspetto della validità riflette in che misura le misure operative studiate rappresentano realmente ciò che il ricercatore ha in mente e ciò che viene indagato in base alle domande di ricerca. Se, ad esempio, i costrutti discussi nelle domande dell'intervista non vengono interpretati allo stesso modo dal ricercatore e dalle persone intervistate, esiste una minaccia alla validità del costrutto.
- *Validità interna*: questo aspetto della validità è importante quando si esaminano le relazioni causali. Quando il ricercatore indaga se un fattore influenza un fattore indagato, c'è il rischio che il fattore indagato sia influenzato anche da un terzo fattore. Se il ricercatore non è a conoscenza del terzo fattore e/o non lo sa sapere in che misura influisce sul fattore indagato, esiste una minaccia alla validità interna.
- *Validità esterna*: questo aspetto della validità riguarda la misura in cui è possibile generalizzare i risultati e in che misura i risultati interessano altre persone al di fuori del caso investigato. Durante l'analisi della validità esterna, il ricercatore cerca di analizzare in che misura i risultati sono rilevanti per altri casi. Negli studi di casi, non esiste una popolazione da cui sia stato estratto un campione statisticamente rappresentativo. Tuttavia, per i casi di studio, l'intenzione è

consentire una generalizzazione analitica in cui i risultati vengono estesi a casi che hanno caratteristiche comuni e quindi per i quali i risultati sono rilevanti, ovvero definendo una teoria. • *Affidabilità*:

questo aspetto riguarda la misura in cui i dati e l'analisi dipendono dai ricercatori specifici.

Ipoteticamente, se un altro ricercatore successivamente conducesse lo stesso studio, il risultato dovrebbe essere lo stesso. Le minacce a questo aspetto della validità sono, ad esempio, se non è chiaro come codificare i dati raccolti o se i questionari o le domande delle interviste non sono chiari. Per l'analisi quantitativa, la controparte dell'affidabilità è la validità delle conclusioni, vedere più avanti la sez. 8.7.

Come descritto sopra, è importante considerare la validità del caso di studio fin dall'inizio. Esempi di modi per migliorare la validità sono la triangolazione; sviluppare e mantenere un protocollo dettagliato di studio di casi; far revisionare progetti, protocolli, ecc. da ricercatori paritari; hanno raccolto dati e ottenuto risultati esaminati dai soggetti del caso; dedicare tempo sufficiente al caso e prestare sufficiente attenzione all'analisi dei "casi negativi", ovvero alla ricerca di teorie che contraddicono le tue scoperte.

5.5 Reporting

Uno studio empirico non può essere distinto dalla sua rendicontazione. Il rapporto comunica i risultati dello studio, ma è anche la principale fonte di informazioni per giudicare la qualità dello studio. I report possono avere un pubblico diverso, come ricercatori paritari, policy maker, sponsor della ricerca e professionisti del settore [180].

Ciò può portare alla necessità di scrivere report diversi per pubblici diversi.

Qui ci concentreremo su resoconti che hanno come pubblico principale i ricercatori tra pari, cioè articoli di riviste o conferenze ed eventualmente relazioni tecniche di accompagnamento [22]. Le linee guida per riportare casi di studio di ingegneria del software ad altri pubblici e in altri formati sono fornite da Runeson et al. [146]. Benbasat et al. proporre che, a causa dell'ampia quantità di dati generati negli studi di casi, "libri o monografie potrebbero essere veicoli migliori per pubblicare ricerche di studi di casi" [22].

Per i casi di studio può essere utilizzata la stessa struttura di alto livello, vedere il Cap. 11, ma poiché sono più flessibili e si basano per lo più su dati qualitativi, il dettaglio di basso livello è meno standardizzato e dipende maggiormente dal singolo caso. Di seguito, discutiamo prima le caratteristiche di un rapporto di studio di caso e poi una struttura proposta.

5.5.1 Caratteristiche

Robson definisce un insieme di caratteristiche che dovrebbe avere un rapporto di studio di caso [144], il che in sintesi implica che dovrebbe:

- Spiegare di cosa trattava lo studio. •
- Comunicare un chiaro senso del caso studiato.

- Fornire una “storia dell’indagine” in modo che il lettore possa vedere cosa è stato fatto e da chi e come.
- Fornire i dati di base in forma mirata, in modo che il lettore possa assicurarsi che le conclusioni sono ragionevoli.
- Articolare le conclusioni dei ricercatori e inserirle nel contesto che interessano.

Inoltre, ciò deve avvenire nel rispetto dell’equilibrio tra dovere e responsabilità del ricercatore obiettivo di pubblicare i propri risultati e l’integrità delle aziende e degli individui [3].

Riportare gli obiettivi del caso di studio e le domande di ricerca è abbastanza semplice. Se vengono modificati sostanzialmente nel corso dello studio, ciò dovrebbe essere segnalato per aiutare a comprendere il caso.

Descrivere il caso potrebbe essere più delicato, poiché ciò potrebbe consentire l’identificazione del caso o dei suoi argomenti. Ad esempio, “una grande azienda di telecomunicazioni in Svezia” è molto probabilmente una filiale della Ericsson Corporation. Tuttavia, il caso potrebbe essere meglio caratterizzato con altri mezzi oltre al solo dominio e paese della domanda. Le caratteristiche interne, come la dimensione dell’unità studiata, l’età media del personale, ecc. possono essere più interessanti delle caratteristiche esterne come il dominio e il fatturato. O il caso costituisce una piccola subunità di una grande azienda, e quindi difficilmente può essere identificata tra le tante subunità, oppure è una piccola azienda e quindi è difficile identificarla tra molti candidati. Bisogna però fare attenzione a trovare questo equilibrio.

Fornire una “storia dell’indagine” richiede un livello di dettaglio sostanzialmente maggiore rispetto alla semplice segnalazione delle metodologie utilizzate, ad esempio “abbiamo avviato un caso di studio utilizzando interviste semi-strutturate”. Poiché la validità dello studio è fortemente correlata a cosa viene fatto, da chi e come, è necessario riferire sulla sequenza delle azioni e dei ruoli che agiscono nel processo di studio. D’altro canto, non c’è spazio per ogni singolo dettaglio della condotta del caso di studio, e quindi occorre trovare un equilibrio. I dati vengono raccolti in abbondanza in uno studio qualitativo e l’analisi ha come obiettivo principale quello di ridurre e organizzare i dati per fornire una catena di prove per le conclusioni. Tuttavia, per stabilire fiducia nello studio, il lettore ha bisogno di istantanee pertinenti dei dati che supportino le conclusioni. Queste istantanee possono assumere la forma, ad esempio, di citazioni (dichiarazioni tipiche o speciali), immagini o racconti con soggetti anonimizzati. Inoltre, le categorie utilizzate nella classificazione dei dati, che portano a determinate conclusioni, possono aiutare il lettore a seguire la catena delle prove.

Infine, le conclusioni devono essere riportate e inserite in un contesto di implicazioni, ad esempio formando teorie. Un caso di studio non può essere generalizzato nel senso di essere rappresentativo di una popolazione, ma questo non è l’unico modo per raggiungere e trasferire la conoscenza. Le conclusioni possono essere tratte senza statistiche e possono essere interpretate e collegate ad altri casi. Comunicare i risultati della ricerca in termini di teorie è una pratica sottosviluppata nell’ingegneria del software [72], come discusso nella sez. 2.7.

Tabella 5.4 Struttura di segnalazione proposta per studi di casi basata su Jedlitschka e Pfahl [86] e adattamenti alla segnalazione di studi di casi secondo Runeson et al. [146]

Intestazioni delle sezioni	Sottosezioni
Titolo	
Paternità	
Estratto strutturato	
Introduzione	Dichiarazione del problema Obiettivi della ricerca Contesto
Lavoro correlato	Studi precedenti Teoria
Progettazione di casi di studio	Domande di ricerca Scelta del caso e del soggetto Procedura/i di raccolta dati Procedure di analisi Procedure di validità
Risultati	Descrizioni di casi e argomenti, che coprono questioni di esecuzione, analisi e interpretazione Sottosezioni, che possono essere strutturate ad esempio secondo uno schema di codifica, ciascuna delle quali collega le osservazioni alle conclusioni Valutazione della validità
Conclusioni e lavoro futuro	Riepilogo dei risultati Relazione con le prove esistenti Impatto/implicazioni Limitazioni Lavoro futuro
Ringraziamenti	
Riferimenti	
Appendici	

5.5.2 Struttura

Per la presentazione accademica di casi di studio, la struttura analitica lineare (problema, lavoro correlato, metodi, analisi e conclusioni) è la struttura più accettata.

La struttura di alto livello per la segnalazione di esperimenti di ingegneria del software proposta da Jedlitschka e Pfahl [86] si adatta quindi anche allo scopo di riportare casi di studio. Tuttavia, sono necessari alcuni cambiamenti, basati sulle caratteristiche specifiche dei casi di studio e su altre questioni basate su una valutazione condotta da Kitchenham et al. [101]. La struttura risultante è presentata nella Tabella 5.4.

In un caso di studio, la teoria può costituire una struttura per l'analisi; quindi, ci sono due tipi di lavoro correlato: (a) studi precedenti sull'argomento e (b) teorie su cui si basa lo studio attuale. La sezione di progettazione corrisponde al protocollo del caso studio, ovvero riporta la pianificazione del caso studio comprese le misure adottate per garantire la validità dello studio. Poiché il caso di studio ha una progettazione flessibile e la raccolta e l'analisi dei dati sono più intrecciate, questi argomenti possono essere combinati in un'unica sezione (come è stato fatto nella sezione 5.3).

Di conseguenza, i contenuti al livello inferiore devono essere adeguati, come proposto nella Tabella 5.4. Nello specifico della sezione dati combinata, lo schema di codifica costituisce spesso una struttura di sottosezione naturale. In alternativa, per uno studio di caso comparativo, la sezione dati può essere strutturata in base ai casi confrontati, e per uno studio longitudinale, la scala temporale può costituire la struttura della sezione dati.

Questa sezione dei risultati combinati include anche una valutazione della validità dei risultati finali.

Nel capitolo successivo viene delineata una panoramica del processo di conduzione degli esperimenti e ogni fase del processo viene presentata in maggior dettaglio nei capitoli successivi.

5.6 Esercizi

5.1. Quando lo studio di caso è una metodologia di ricerca fattibile?

5.2. Che ruolo ha la pianificazione nei casi di studio, essendo una metodologia di ricerca flessibile?

5.3. Quali criteri governano la selezione dei casi per uno studio?

5.4. Elenca tre tipi di interviste e spiega quale tipo è adatto ai diversi tipi situazioni.

5.5. Descrivere un tipico processo di analisi qualitativa.

Capitolo 6

Processo dell'esperimento

La sperimentazione non è semplice; dobbiamo preparare, condurre e analizzare adeguatamente gli esperimenti. Uno dei principali vantaggi di un esperimento è il controllo, ad esempio, di soggetti, oggetti e strumentazione. Ciò garantisce che siamo in grado di trarre conclusioni più generali. Altri vantaggi includono la capacità di eseguire analisi statistiche utilizzando metodi di verifica delle ipotesi e opportunità di replica. Per garantire l'utilizzo dei vantaggi, abbiamo bisogno di un processo che ci supporti nei nostri obiettivi nel condurre correttamente gli esperimenti (la nozione di esperimenti include quasi-esperimenti, se non diversamente specificato). I principi di base alla base di un esperimento sono illustrati nella Fig. 6.1.

Il punto di partenza è che abbiamo un'idea di una relazione di causa ed effetto, cioè crediamo che esista una relazione tra un costrutto di causa e un costrutto di effetto.

Abbiamo una teoria o siamo in grado di formulare un'ipotesi. Un'ipotesi significa che abbiamo un'idea, ad esempio, di una relazione, che possiamo enunciare formalmente in un'ipotesi.

Per valutare le nostre convinzioni, possiamo utilizzare un esperimento. L'esperimento viene creato, ad esempio, per verificare una teoria o un'ipotesi. Nella progettazione dell'esperimento, abbiamo una serie di trattamenti (valori che la variabile studiata può assumere, vedi sotto) su cui abbiamo il controllo. L'esperimento viene eseguito e possiamo osservare il risultato. Ciò significa che testiamo la relazione tra il trattamento e il risultato. Se l'esperimento è impostato correttamente, dovremmo essere in grado di trarre conclusioni sulla relazione tra la causa e l'effetto per il quale abbiamo formulato un'ipotesi.

L'obiettivo principale di un esperimento è principalmente quello di valutare un'ipotesi o una relazione, vedere anche la Sez. 2.4.1. La verifica delle ipotesi normalmente si riferisce alla prima, mentre la seconda riguarda soprattutto la costruzione di un modello relazionale basato sui dati raccolti. Il modello può essere derivato utilizzando metodi statistici multivariati, ad esempio tecniche di regressione, e quindi lo valutiamo in un esperimento. Il focus di questo libro è principalmente sulla verifica delle ipotesi. I metodi statistici multivariati sono trattati, ad esempio, da Kachigan [90, 91] e Manly [118].

Il processo dell'esperimento presentato in questo capitolo è formulato per garantire che vengano intraprese le azioni appropriate per garantire un esperimento di successo. Sfortunatamente non lo è

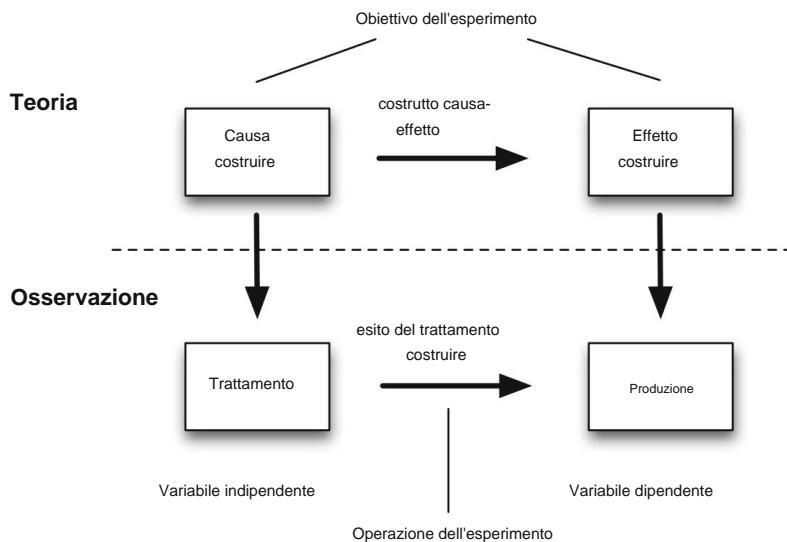


Fig. 6.1 Principi dell'esperimento (adattato da Trochim [171])

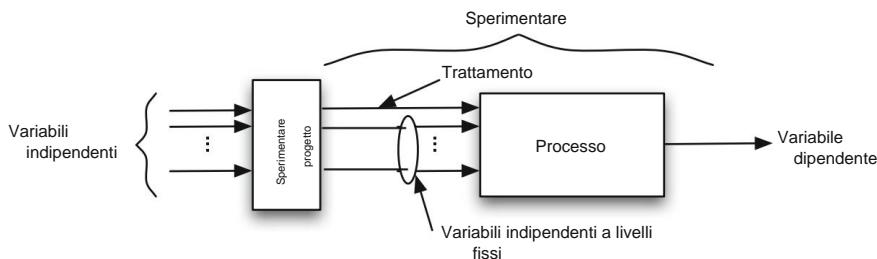
È raro che qualche fattore venga trascurato prima dell'esperimento e la svista ci impedisca di eseguire l'analisi pianificata e quindi non siamo in grado di trarre conclusioni valide. L'obiettivo di un processo è fornire supporto nell'impostazione e nella conduzione di un esperimento. Le attività di un esperimento sono brevemente descritte in questo capitolo e trattate più dettagliatamente nei capitoli successivi, vedere i capp. 7–11.

6.1 Variabili, Trattamenti, Oggetti e Soggetti

Prima di discutere il processo sperimentale, è necessario introdurre alcune definizioni per avere un vocabolario per la sperimentazione. Quando conduciamo un esperimento formale, vogliamo studiare il risultato quando variamo alcune delle variabili di input in un processo. Ci sono due tipi di variabili in un esperimento, variabili indipendenti e dipendenti, vedere Fig. 6.2.

Quelle variabili che vogliamo studiare per vedere l'effetto dei cambiamenti nelle variabili indipendenti sono chiamate *variabili dipendenti* (o variabili di risposta). Spesso in un esperimento è presente una sola variabile dipendente. Tutte le variabili in un processo che vengono manipolate e controllate sono chiamate *variabili indipendenti*.

Esempio. Vogliamo studiare l'effetto di un nuovo metodo di sviluppo sulla produttività del personale. Potremmo aver scelto di introdurre un metodo di progettazione orientato agli oggetti anziché un approccio orientato alle funzioni. La *variabile dipendente* in

**Fig. 6.2** Illustrazione delle variabili indipendenti e dipendenti**Fig. 6.3** Illustrazione di un esperimento

l'esperimento è la produttività. Le *variabili indipendenti* possono essere lo sviluppo metodo, esperienza del personale, supporto degli strumenti e ambiente.

Un esperimento studia l'effetto della modifica di una o più variabili indipendenti. Queste variabili sono chiamate *fattori*. Le altre variabili indipendenti sono controllate a un livello fisso durante l'esperimento, altrimenti non possiamo dire se il fattore o un altro la variabile provoca l'effetto. Un *trattamento* è un valore particolare di un fattore.

Esempio. Il fattore per l'esperimento di esempio riportato sopra è il metodo di sviluppo poiché vogliamo studiare l'effetto del cambiamento del metodo. Usiamo due trattamenti di il fattore: il vecchio e il nuovo metodo di sviluppo.

La scelta del trattamento e a quali livelli deve essere effettuata l'altra variabile indipendente avere, fa parte del disegno dell'esperimento, vedere Fig. 6.3. Viene descritta la progettazione dell'esperimento più dettagliatamente nel cap. 8.

I trattamenti vengono applicati alla combinazione di *oggetti* e *soggetti*. Un oggetto può, ad esempio, essere un documento che deve essere rivisto con diversi tecniche di ispezione. Le persone che applicano il trattamento sono chiamate *soggetti*.¹ IL le caratteristiche sia degli oggetti che dei soggetti possono essere variabili indipendenti in l'esperimento.

¹A volte viene utilizzato il termine *partecipante* al posto del termine soggetto. Il termine soggetto è principalmente utilizzato quando le persone vengono considerate rispetto ai diversi trattamenti e rispetto al analisi e il termine partecipante principalmente quando si tratta di come coinvolgere e motivare le persone in uno studio.

Esempio. Gli oggetti nell'esperimento di esempio sono i programmi da sviluppare e i soggetti sono il personale.

Un esperimento consiste in una serie di *test* (a volte chiamati prove) in cui ogni test è una combinazione di trattamento, soggetto e oggetto. È opportuno osservare che questo tipo di test non deve essere confuso con l'utilizzo dei test statistici, di cui si parlerà più approfonditamente nel Cap. 10. Il numero di test influisce sull'errore sperimentale e offre l'opportunità di stimare l'effetto medio di qualsiasi fattore sperimentale.

L'errore sperimentale ci aiuta a sapere quanta fiducia possiamo riporre nei risultati dell'esperimento.

Esempio. Un test può essere che la persona N (soggetto) utilizzi il nuovo metodo di sviluppo (trattamento) per sviluppare il programma A (oggetto).

Negli esperimenti orientati all'uomo, gli esseri umani sono i soggetti, applicando trattamenti diversi agli oggetti. Ciò implica diverse limitazioni al controllo dell'esperimento. In primo luogo, gli esseri umani hanno competenze e abilità diverse, che di per sé possono rappresentare una variabile indipendente. In secondo luogo, gli esseri umani imparano nel tempo, il che significa che se un soggetto applica due metodi, l'ordine di applicazione dei metodi può avere importanza, e inoltre lo stesso oggetto non può essere utilizzato in entrambe le occasioni. In terzo luogo, gli esperimenti orientati all'uomo sono influenzati da tutti i tipi di influenze e minacce, a causa della capacità del soggetto di indovinare cosa si aspetta lo sperimentatore, della sua motivazione per svolgere i compiti, ecc. Quindi è fondamentale per il risultato dell'esperimento il modo in cui i soggetti vengono selezionati. e trattato.

Gli esperimenti orientati alla tecnologia sono più facili da controllare, poiché la tecnologia può essere resa deterministica. La variabile indipendente fuori controllo in questo tipo di esperimenti potrebbero invece essere gli oggetti selezionati per l'esperimento. Uno strumento o una tecnica possono essere adatti per un tipo di programmi e non per un altro. Quindi è fondamentale per il risultato il modo in cui gli oggetti vengono selezionati.

6.2 Processo

Un processo fornisce passaggi che supportano un'attività, ad esempio lo sviluppo di software. I processi sono importanti in quanto possono essere utilizzati come liste di controllo e linee guida su cosa fare e come farlo. Per eseguire un esperimento, è necessario eseguire diversi passaggi e devono essere in un determinato ordine. Pertanto, è necessario un processo su come eseguire gli esperimenti.

Il processo presentato è incentrato sulla sperimentazione, ma gli stessi passaggi fondamentali devono essere eseguiti in qualsiasi studio empirico, come illustrato per il processo di studio del caso nella Sez. 5.1.2. La differenza principale è il lavoro all'interno di un'attività specifica, ad esempio la progettazione di un sondaggio, un esperimento e un caso di studio differiscono, ma devono essere tutti progettati. Inoltre, poiché i casi di studio sono studi di progettazione flessibili, ci sono diverse iterazioni nelle fasi del processo, mentre esperimenti e sondaggi, come studi di progettazione fissi, eseguono principalmente le fasi una volta. Pertanto, è possibile utilizzare il processo di base

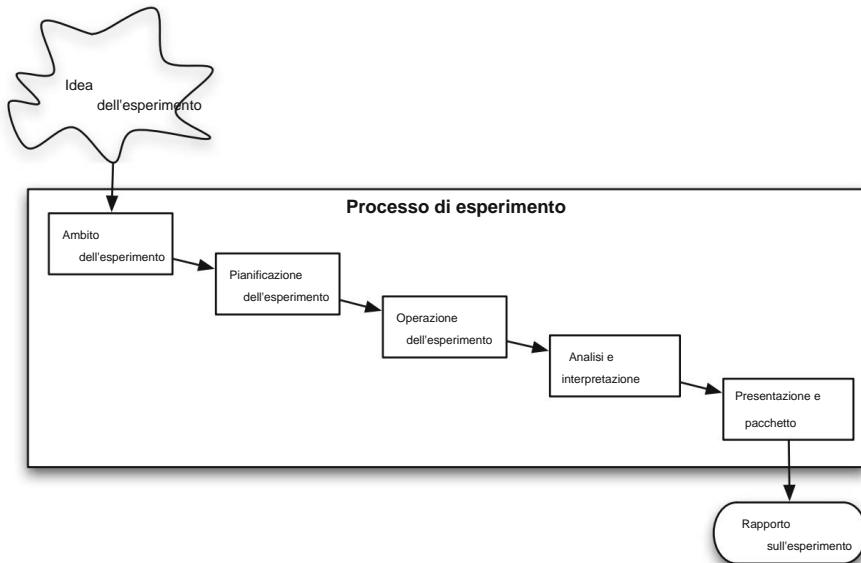


Fig. 6.4 Panoramica del processo dell'esperimento

altri tipi di studi oltre agli esperimenti, ma deve essere adattato allo specifico tipo di studio condotto, ad esempio un sondaggio tramite posta elettronica o un caso di studio di un grande progetto software. Il processo è come viene presentato, tuttavia, adatto a sia esperimenti randomizzati che quasi-esperimenti. Questi ultimi sono spesso utilizzati in ingegneria del software quando campioni casuali, ad esempio, di soggetti (partecipanti) sono irrealizzabili.

Il punto di partenza per un esperimento è l'intuizione e l'idea che un esperimento sarebbe un modo possibile per valutare ciò che ci interessa. In altro In parole povere, dobbiamo renderci conto che un esperimento è appropriato per la domanda che ci poniamo andranno ad indagare. Ciò non è affatto sempre ovvio, soprattutto da allora gli studi empirici non sono utilizzati frequentemente nell'ambito dell'informatica e del software ingegneria [170, 181]. Alcune argomentazioni sul perché informatico dovrebbe sperimentare di più è fornito da Tichy [169]. Se assumiamo di avere realizzato che un esperimento è appropriato, allora è importante pianificarlo attentamente per evitare errori inutili, vedere la Sez. 2.9.

Il processo sperimentale può essere suddiviso nelle seguenti attività principali. *Lo scoping* è il primo passo, in cui definiamo l'esperimento in termini di problema, obiettivo e obiettivi. Poi viene *la pianificazione*, dove viene determinata la progettazione dell'esperimento, viene considerata la strumentazione e vengono valutate le minacce all'esperimento. *Il funzionamento* dell'esperimento segue dalla progettazione. Nell'attività operativa, vengono raccolte le misurazioni che poi vengono analizzate e valutate in *analisi e interpretazione*. Infine, i risultati vengono presentati e confezionati nella *presentazione e pacchetto*. Le attività sono illustrate in Fig. 6.4 e ulteriormente elaborate di seguito,

e poi ciascuna delle attività viene trattata in modo approfondito nei capp. 7–11. Una panoramica del processo dell'esperimento, comprese le attività, è presentata in Fig. 6.5.

Il processo non dovrebbe essere un “vero” modello a cascata; non si presuppone che un'attività sia necessariamente terminata prima che venga avviata l'attività successiva. L'ordine delle attività nel processo indica principalmente l'ordine di inizio delle attività. In altre parole, il processo è in parte iterativo e potrebbe essere necessario tornare indietro e perfezionare un'attività precedente prima di continuare con l'esperimento. L'eccezione principale è quando l'operazione dell'esperimento è iniziata, quindi non è possibile tornare alla definizione e alla pianificazione dell'esperimento. Questo non è possibile perché iniziare l'operazione significa che i soggetti vengono influenzati dall'esperimento, e se si torna indietro c'è il rischio che sia impossibile utilizzare gli stessi soggetti quando si ritorna alla fase operativa del processo sperimentale.

Ambito. La prima attività è l'ambito. L'ipotesi deve essere espressa chiaramente. Non è necessario dichiararlo formalmente in questa fase, ma deve essere chiaro. Inoltre, è necessario definire l'obiettivo e gli scopi dell'esperimento. L'obiettivo è formulato a partire dal problema da risolvere. Per cogliere l'ambito, è stato suggerito un quadro [13]. Il quadro è costituito dai seguenti elementi:

- Oggetto di studio (cosa viene studiato?), • Scopo (qual è l'intenzione?), • Focus sulla qualità (quale effetto viene studiato?), • Prospettiva (di chi è il punto di vista?), e • Contesto (dove è condotto lo studio?).

Questi sono ulteriormente discussi nel cap. 7.

Pianificazione. L'attività di pianificazione è il luogo in cui vengono poste le basi per l'esperimento. Il contesto dell'esperimento è determinato in dettaglio. Ciò include il personale e l'ambiente, ad esempio, se l'esperimento viene condotto in un ambiente universitario con studenti o in un ambiente industriale. Inoltre, l'ipotesi dell'esperimento è dichiarata formalmente, includendo un'ipotesi nulla e un'ipotesi alternativa.

Il passo successivo nell'attività di pianificazione è determinare le variabili (sia variabili indipendenti (input) che variabili dipendenti (output)). Una questione importante riguardante le variabili è determinare i valori che le variabili possono effettivamente assumere. Ciò include anche la determinazione della scala di misurazione, che pone vincoli al metodo che successivamente potremo applicare per l'analisi statistica. I soggetti dello studio vengono identificati.

Inoltre, l'esperimento viene progettato, il che comprende la scelta di un disegno sperimentale adeguato che includa, ad esempio, la randomizzazione dei soggetti. Una questione strettamente legata alla progettazione è la preparazione della strumentazione dell'esperimento. Dobbiamo identificare e preparare oggetti idonei, sviluppare linee guida se necessario e definire procedure di misurazione. Tali questioni vengono ulteriormente discusse nel cap. 8.

Nell'ambito della pianificazione è importante considerare la questione della validità dei risultati che possiamo aspettarci. La validità può essere suddivisa in quattro classi principali: interna,

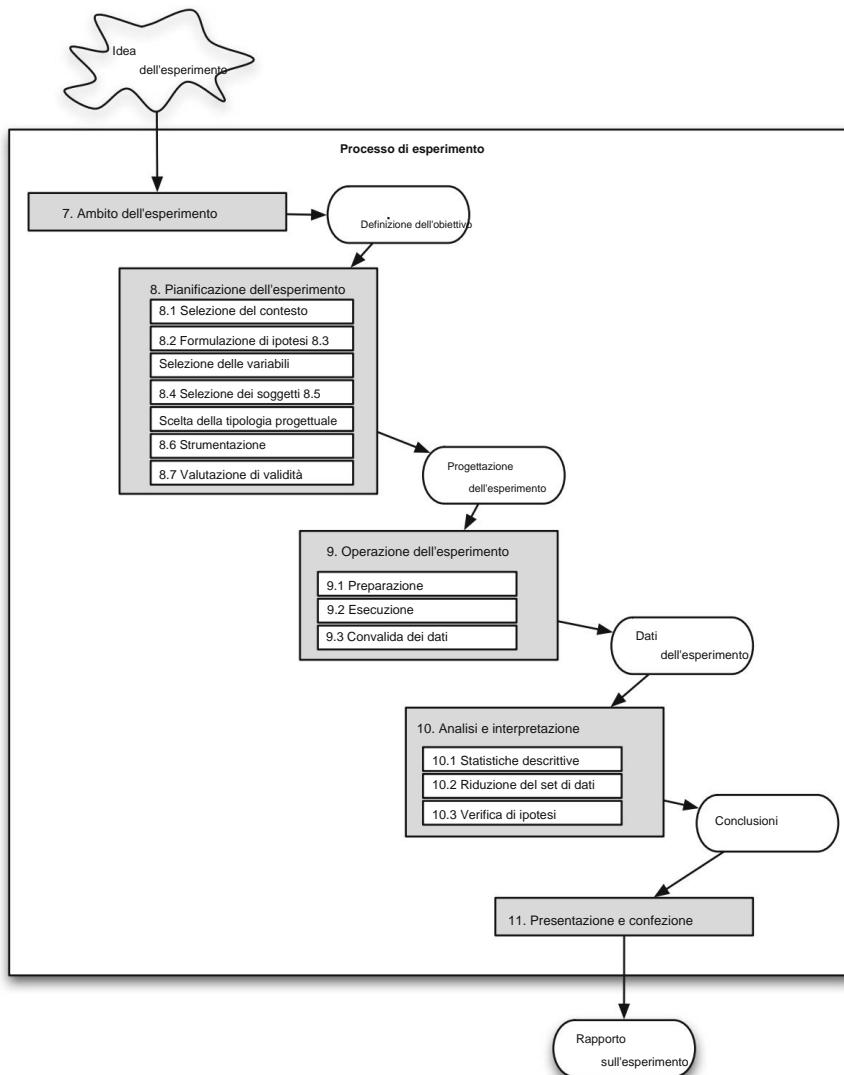


Fig. 6.5 Panoramica del processo dell'esperimento e degli artefatti con riferimenti ai capitoli e alle sezioni di questo libro

validità esterna, di costrutto e di conclusione. La validità interna riguarda la validità all'interno di un dato ambiente e l'affidabilità dei risultati. La validità esterna è una questione di quanto generali siano i risultati. Molte volte vorremmo affermare che i risultati di un esperimento sono validi al di fuori del contesto reale in cui l'esperimento è stato condotto. La validità di costrutto è una questione di giudicare se il trattamento riflette il costrutto di causa e se il risultato fornisce un quadro fedele dell'effetto.

costrutto, vedere Fig. 6.1. La validità della conclusione riguarda la relazione tra il trattamento e il risultato dell'esperimento. Dobbiamo giudicare se esiste una relazione tra il trattamento e il risultato.

La pianificazione è un passaggio cruciale in un esperimento per garantire che i risultati dell'esperimento diventa utile. Una cattiva pianificazione può rovinare qualsiasi studio ben intenzionato.

Operazione. L'operazione consiste in linea di principio in tre fasi: preparazione, esecuzione e convalida dei dati. Nella fase di preparazione ci occupiamo di preparare i soggetti e il materiale necessario, ad esempio i moduli per la raccolta dei dati. I partecipanti devono essere informati dell'intenzione; dobbiamo avere il loro consenso e devono essere impegnati. L'effettiva esecuzione normalmente non è un grosso problema. La preoccupazione principale è garantire che l'esperimento sia condotto secondo il piano e la progettazione dell'esperimento, che include la raccolta dei dati. Infine, dobbiamo cercare di assicurarsi che i dati effettivamente raccolti siano corretti e forniscano un quadro valido dell'esperimento. L'attività operativa è discussa nel Cap. 9.

Analisi e interpretazione. I dati raccolti durante il funzionamento forniscono l'input a questa attività. I dati possono ora essere analizzati e interpretati. Il primo passo nell'analisi è cercare di comprendere i dati utilizzando la statistica descrittiva. Questi forniscono una visualizzazione dei dati. Le statistiche descrittive ci aiutano a comprendere e interpretare i dati in modo informale.

Il passo successivo è considerare se il set di dati debba essere ridotto, rimuovendo i punti dati o riducendo il numero di variabili studiando se alcune variabili forniscono le stesse informazioni. Sono disponibili metodi specifici per la riduzione dei dati.

Dopo aver rimosso i punti dati o ridotto il set di dati, siamo in grado di eseguire un test di ipotesi, in cui il test effettivo viene scelto in base alle scale di misurazione, ai valori dei dati di input e al tipo di risultati che stiamo cercando. I test statistici insieme ad una discussione più dettagliata della statistica descrittiva e delle tecniche di riduzione dei dati si trovano nel Cap.

10.

Un aspetto importante di questa attività è l'interpretazione. Dobbiamo cioè determinare dall'analisi se l'ipotesi era possibile rifiutarla. Ciò costituisce la base per il processo decisionale e le conclusioni su come utilizzare i risultati dell'esperimento, che include la motivazione per ulteriori studi, ad esempio per condurre un esperimento ampliato o uno studio di caso.

Presentazione e confezione. L'ultima attività riguarda la presentazione e il confezionamento dei risultati. Ciò include principalmente la documentazione dei risultati, che può essere effettuata tramite un documento di ricerca per la pubblicazione, un pacchetto di laboratorio a scopo di replica o come parte della base di esperienza di un'azienda. Quest'ultima attività è importante per assicurarsi che le lezioni apprese siano curate in modo adeguato. Inoltre, un esperimento non fornirà mai la risposta definitiva a una domanda, e quindi è importante facilitare la replica dell'esperimento. Una documentazione completa e approfondita è un prerequisito per raggiungere questo obiettivo. Detto questo, l'uso dei pacchetti di laboratorio dovrebbe essere fatto con cautela poiché l'uso dello stesso disegno sperimentale e degli stessi documenti può portare con sé alcuni problemi sistematici e pregiudizi da

l'esperimento originale, come discusso nella Sez. 2.6. Indipendentemente, dobbiamo prenderci del tempo dopo l'esperimento per documentarlo e presentarlo in modo adeguato. La presentazione di un esperimento è ulteriormente elaborata nel Cap. 11.

6.3 Panoramica

Le fasi di questo processo sperimentale vengono descritte più dettagliatamente di seguito e, per facilitare la comprensione del processo, nel cap. 12.

L'obiettivo dell'esempio è seguire da vicino il processo definito per illustrarne l'utilizzo. Una panoramica riassuntiva del processo sperimentale è riportata nella Fig. 6.5.

6.4 Esercizi

6.1. Che cos'è una relazione di causa ed effetto?

6.2. Cos'è un trattamento e perché a volte è necessario applicare i trattamenti in ordine casuale?

6.3. Cosa sono rispettivamente le variabili dipendenti e indipendenti?

6.4. Cosa sono i quasi-esperimenti? Spiegare perché questi sono comuni nell'ingegneria del software.

6.5. Quali sono le fasi principali del processo sperimentale e perché è importante avere fasi distinte?

Parte II

Fasi del processo dell'esperimento

Capitolo 7

Ambito

Condurre un esperimento è un compito ad alta intensità di lavoro. Per utilizzare lo sforzo speso, è importante garantire che l'intenzione con l'esperimento possa essere soddisfatta attraverso l'esperimento. Nella fase di scoping vengono determinate le basi dell'esperimento, come illustrato in Fig. 7.1. Se le fondamenta non vengono gettate correttamente, potrebbe essere necessaria una rielaborazione o, peggio ancora, l'esperimento non può essere utilizzato per studiare ciò che era previsto. Lo scopo della fase di scoping è definire gli obiettivi di un esperimento secondo un quadro definito. Qui seguiamo il modello GQM per la definizione degli obiettivi, originariamente presentato da Basili e Rombach [13].

L'ambito di un esperimento è discusso nella Sez. 7.1. Un esempio di definizione dell'obiettivo dell'esperimento è presentato nella Sez. 7.2.

7.1 Esperimento nell'ambito

L'ambito dell'esperimento viene stabilito definendone gli obiettivi. Lo scopo di un modello di definizione degli obiettivi è garantire che gli aspetti importanti di un esperimento siano definiti prima che abbiano luogo la pianificazione e l'esecuzione. Definendo l'obiettivo dell'esperimento secondo questo modello, le basi vengono gettate correttamente. Il modello di obiettivo è [13]:

Analizzare <Oggetto(i) di studio> ai
fini dello <Scopo> rispetto al loro
<Focus sulla qualità> dal punto di vista della
<Prospettiva> nel contesto del <Contesto>.

L'oggetto di studio è l'entità studiata nell'esperimento. L'oggetto di studio può essere prodotti, processi, risorse, modelli, metriche o teorie. Esempi sono il prodotto finale, il processo di sviluppo o ispezione o un modello di crescita dell'affidabilità. Lo scopo definisce qual è l'intento dell'esperimento. Potrebbe trattarsi di valutare l'impatto di due diverse tecniche o di caratterizzare la curva di apprendimento di un'organizzazione. L'attenzione alla qualità è l'effetto principale oggetto di studio nel

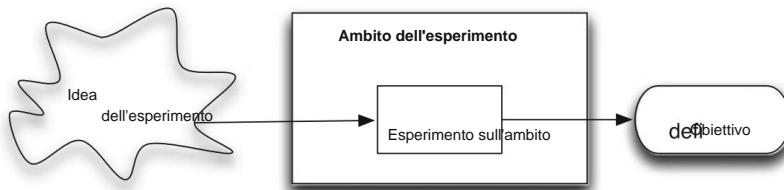


Fig. 7.1 Panoramica della fase di scoping

Tabella 7.1 Classificazione del contesto dell'esperimento

# Oggetti		
# Soggetti	Uno	Più di uno
per oggetto	Studio di un singolo oggetto	Studio della variazione multi-oggetto
Più di uno	Multi-test nello studio dell'oggetto	Studio soggetto-oggetto bloccato

sperimentare. L'attenzione alla qualità può riguardare l'efficacia, i costi, l'affidabilità, ecc. La prospettiva indica il punto di vista da cui vengono interpretati i risultati dell'esperimento. Esempi di le prospettive sono sviluppatore, project manager, cliente e ricercatore. Il contesto è l'"ambiente" in cui viene eseguito l'esperimento. Il contesto definisce brevemente quale il personale è coinvolto nell'esperimento (soggetti) e quali artefatti software (oggetti¹) vengono utilizzati nell'esperimento. I soggetti possono essere caratterizzati da esperienza, dimensione del team, carico di lavoro ecc. Gli oggetti possono essere caratterizzati da dimensione, complessità, priorità, dominio applicativo, ecc.

Il contesto dell'esperimento può essere classificato in termini di numero di soggetti e oggetti coinvolti nello studio [10], vedere Tabella 7.1.

Gli studi su un singolo oggetto vengono condotti su un singolo soggetto e un singolo oggetto. Gli studi sulle variazioni multi-oggetto vengono condotti su un singolo soggetto attraverso una serie di oggetti. Il test multiplo all'interno degli studi sugli oggetti esamina un singolo oggetto attraverso una serie di soggetti. Gli studi soggetto-oggetto bloccati esaminano un insieme di soggetti e un insieme di oggetti. Tutti questi tipi di esperimenti possono essere eseguiti come esperimento o quasi-esperimento. In un quasi-esperimento manca la randomizzazione sia dei soggetti che degli oggetti. Lo studio su un singolo oggetto è un quasi-esperimento se il singolo soggetto e l'oggetto lo sono non selezionato in modo casuale, ma è un esperimento se si scelgono soggetto e oggetto per caso. Viene discussa la differenza tra esperimenti e quasi-esperimenti ulteriormente da Robson [144].

Esempi dei diversi tipi di esperimenti sono forniti dalle serie di esperimenti condotto presso NASA-SEL [10], finalizzato alla valutazione dei principi delle camere bianche e tecniche. Cleanroom è una raccolta di metodi e tecniche di ingegneria assemblati con l'obiettivo di produrre software di alta qualità. Una breve introduzione to Cleanroom è fornita da Linger [112]. La serie di esperimenti è composta da quattro passaggi distinti. Innanzitutto, è stato condotto un esperimento di lettura rispetto a test unitario in un blocco

¹Si noti che gli "oggetti" qui sono generalmente diversi dagli "oggetti di studio" definiti sopra.

Tabella 7.2 Esempio di classificazione del contesto dell'esperimento, da Basili [10]

		# Oggetti	
		Uno	Più di uno
# Soggetti per oggetto	Uno	3. Progetto camera bianca n. 1 a SELEZIONARE [14]	4. Progetti di camere bianche n. 2-4 al SEL [14]
	Più di uno	2. Esperimento in camera bianca a Università del Maryland [149]	1. Lettura e test [12] 5. Lettura basata su scenari vs. lista di controllo [18]

Tabella 7.3 Quadro di definizione degli obiettivi

Oggetto di studio	Scopo	Focus sulla qualità	Prospettiva	Contesto
Prodotto	Caratterizzare	Efficacia	Sviluppatore	Soggetti
Processo	Monitorare	Costo	Modificatore	Oggetti
Modello	Valutare	Affidabilità	Manutentore	
Metrico	Prevedere	Manutenibilità	Responsabile del progetto	
Teoria	Controllare	Portabilità	Direttore aziendale	
	Modifica		Cliente	
			Utente	
			Ricercatore	

studio soggetto-oggetto [12], vedere 1 nella Tabella 7.2. In secondo luogo, un progetto di sviluppo l'applicazione delle tecniche di camera bianca è stata condotta in un ambiente studentesco [149]. IL L'esperimento era un test multiplo all'interno di un esperimento di variazione dell'oggetto, vedere 2 nella Tabella 7.2. In terzo luogo, un progetto che utilizza Cleanroom è stato condotto presso la NASA-SEL [14] come unico esperimento sugli oggetti, vedere 3 nella Tabella 7.2. In quarto luogo, sono stati tre progetti Cleanroom condotto nello stesso ambiente, costituendo uno studio di variazione multi-oggetto [14], vedere 4 nella Tabella 7.2. Il prossimo round è un nuovo esperimento di lettura in cui è diverso vengono analizzate le tecniche [18], vedere 5 nella Tabella 7.2. Anche questa serie di esperimenti lo è discusso da Linkman e Rombach [113].

L'esempio, nella Tabella 7.2, illustra come possono essere gli esperimenti (vedi 1 e 2). condotti come studi preliminari prima dei casi di studio (vedere 3 e 4). Ciò è in linea con la discussione sul trasferimento tecnologico e un'adeguata ordinamento in base ai costi e rischio come discusso nelle Sez. 2.9 e 2.10.

7.2 Esempio di esperimento

Il quadro di definizione degli obiettivi può essere compilato con diversi oggetti di studio, scopi ecc. Nella Tabella 7.3 vengono forniti esempi di elementi.

Un esempio di definizione di studio viene costruito componendo gli elementi di quadro normativo e viene presentato di seguito. L'esempio definisce un esperimento di ispezione dove vengono valutate diverse tecniche di ispezione, ovvero la lettura prospettica rispetto alla lettura basata su liste di controllo. La lettura prospettica è stata introdotta da Basili et al. [18], ed è stato valutato in diversi esperimenti incluso un confronto

della lettura basata sulla prospettiva rispetto a un metodo esistente alla NASA di Maldonado et al. [117] e Laitenberger et al. [107] presentano un confronto tra la lettura basata sulla prospettiva e un approccio basato su una lista di controllo. I ricercatori hanno anche confrontato altre tecniche di lettura come il confronto tra la lettura basata sull'utilizzo e la lettura basata su liste di controllo di Thelin et al. [168].

Gli oggetti studiati sono la tecnica di lettura basata sulla prospettiva (PBR) e una tecnica basata su checklist. Lo scopo è quello di valutare le tecniche di lettura, in particolare rispetto alle differenze tra le prospettive nella PBR. Il focus della qualità è l'efficacia e l'efficienza delle tecniche di lettura. La prospettiva è dal punto di vista del ricercatore. L'esperimento viene eseguito utilizzando M.Sc. e dottorato di ricerca studenti come materie basate su un pacchetto di laboratorio definito con documenti relativi ai requisiti testuali. Lo studio è condotto come uno studio soggetto-oggetto bloccato, vedere la Tabella 7.1, poiché coinvolge molti soggetti e più di un documento sui requisiti.

L'esempio è riassunto come:

Analizzare le tecniche PBR e checklist ai fini della valutazione rispetto all'efficacia e all'efficienza dal punto di vista del ricercatore nel contesto della M.Sc. e dottorato di ricerca studenti che leggono i documenti sui requisiti.

Questo esempio è utilizzato nei capp. 8-10 per illustrare lo stato di avanzamento del processo sperimentale. Il riepilogo dell'esperimento costituisce la definizione dell'obiettivo dell'esperimento. È l'input per la fase di pianificazione del processo di esperimento.

7.3 Esercizi

7.1. Perché è importante stabilire obiettivi chiari con un esperimento fin dall'inizio?

7.2. Scrivi un esempio di definizione di obiettivo per un esperimento che vorresti svolgere condotta.

7.3. Perché il contesto in un esperimento è importante?

7.4. Come si può caratterizzare il contesto?

7.5. Spiegare come una serie di studi possono essere utilizzati per il trasferimento tecnologico.

Capitolo 8

Pianificazione

Dopo la definizione dell'ambito dell'esperimento, avviene la pianificazione. Lo scoping determina le basi dell'esperimento – *il motivo per cui* l'esperimento viene condotto – mentre la pianificazione prepara il *modo in cui* l'esperimento verrà condotto.

Come in tutti i tipi di attività di ingegneria, l'esperimento deve essere pianificato e i piani devono essere seguiti per poter controllare l'esperimento. Il risultato dell'esperimento può essere disturbato o addirittura distrutto se non pianificato adeguatamente.

La fase di pianificazione di un esperimento può essere suddivisa in sette passaggi. L'input alla fase è la definizione dell'obiettivo dell'esperimento, vedere il Cap. 7. In base alla definizione dell'obiettivo, la *selezione del contesto* seleziona l'ambiente in cui verrà eseguito l'esperimento. Successivamente avviene la *formulazione delle ipotesi* e la *selezione delle variabili* indipendenti e dipendenti. Viene effettuata la *selezione dei soggetti*. Il *tipo di progettazione dell'esperimento* viene scelto in base all'ipotesi e alle variabili selezionate. Successivamente la *strumentazione* si prepara per l'implementazione pratica dell'esperimento. Infine la *valutazione di validità* mira a verificare la validità dell'esperimento. Il processo di pianificazione viene ripetuto fino a quando non è pronto un progetto completo dell'esperimento. Una panoramica della fase di pianificazione è fornita nella Fig. 8.1.

8.1 Selezione del contesto

Per ottenere i risultati più generali in un esperimento, questo dovrebbe essere eseguito in progetti software reali e di grandi dimensioni, con personale professionale. Tuttavia, condurre un esperimento comporta dei rischi, ad esempio che il nuovo metodo da esaminare non sia buono come previsto e causi ritardi. Un'alternativa è eseguire progetti offline parallelamente ai progetti reali. Ciò riduce i rischi ma causa costi aggiuntivi. Un'alternativa più economica è gestire progetti gestiti da studenti. Tali progetti sono più economici, più facili da controllare, ma più indirizzati a un determinato contesto rispetto ai progetti gestiti da professionisti con maggiore e diversa esperienza. Inoltre questi progetti lo fanno

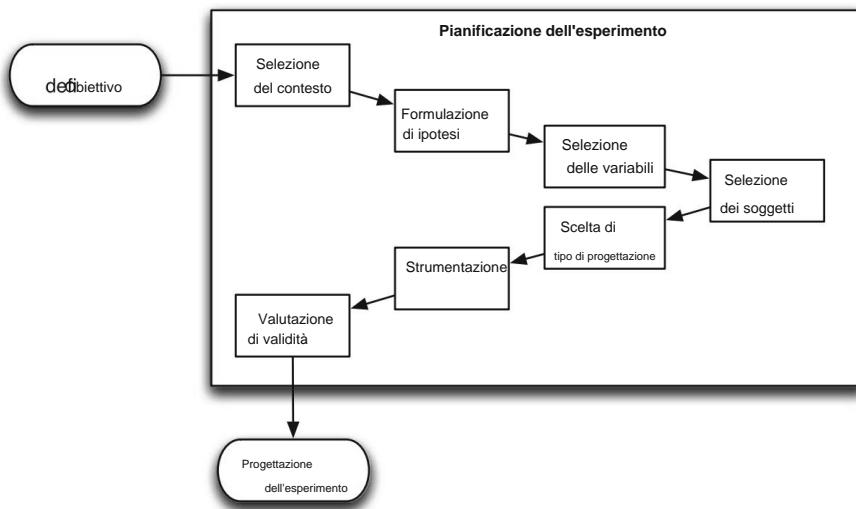


Fig. 8.1 Panoramica della fase di pianificazione

raramente affrontano problemi reali, ma problemi più legati alle dimensioni del giocattolo a causa dei vincoli costi e tempi. Questo compromesso implica un equilibrio tra il rendere gli studi validi per un contesto specifico o valido per il dominio generale dell'ingegneria del software, vedere oltre Setta. 8.7. Dato questo compromesso, vengono discussi gli esperimenti con gli studenti come soggetti in letteratura, ad esempio, da Host et al. [77].

Quindi, il contesto dell'esperimento può essere caratterizzato secondo quattro dimensioni:

- Off-line e on-line
- Studente vs. professionista
- Giocattolo contro problemi reali
- Specifico vs generale

Una situazione comune in un esperimento è che venga confrontato qualcosa di esistente a qualcosa di nuovo, ad esempio un metodo di ispezione esistente viene confrontato con a nuovo [18, 136, 139]. Ci sono due problemi legati a questo tipo di studi.

Innanzitutto, qual è il metodo esistente? È stato applicato per un certo periodo di tempo, ma raramente è ben documentato e non esiste un'applicazione coerente il metodo. In secondo luogo, apprendere un nuovo metodo può influenzare il comportamento di quello vecchio applicato.

Questa e altre questioni relative a ciò di cui ci occupiamo devono essere le persone presi in considerazione quando si pianifica un esperimento per ottenere i risultati valido.

8.2 Formulazione di ipotesi

La base per l'analisi statistica di un esperimento è la verifica delle ipotesi.

Un'ipotesi viene formulata formalmente e i dati raccolti nel corso dell'esperimento vengono utilizzati per, se possibile, respingere l'ipotesi. Se l'ipotesi può essere rifiutata, si possono trarre conclusioni, sulla base della verifica dell'ipotesi sotto determinati rischi.

Nella fase di pianificazione la definizione dell'esperimento viene formalizzata in ipotesi.

Bisogna formulare due ipotesi:

- | | |
|-------|---|
| Nullo | Un'ipotesi nulla, H_0 , afferma che non ci sono tendenze o modelli sottostanti reali nel contesto dell'esperimento; le uniche ragioni per le differenze nelle nostre osservazioni sono casuali. Questa è l'ipotesi che lo sperimentatore vuole respingere con la massima significatività possibile. |
| | Un'ipotesi di esempio è che un nuovo metodo di ispezione trovi in media lo stesso numero di difetti di quello vecchio, cioè $H_0: W_{\text{nuovo}} = W_{\text{vecchio}}$
dove indica la media e N è il numero di guasti
<i>nuovo</i> , trovato. |

Alternativa Un'ipotesi alternativa, H_a ; H_1 , ecc., è l'ipotesi a favore della quale si rifiuta l'ipotesi nulla. Un'ipotesi di esempio è che un nuovo metodo di ispezione trovi in media più difetti di quello vecchio, cioè $H_1: W_{\text{nuovo}} > W_{\text{vecchio}}$.

In letteratura sono descritti diversi test statisticci che possono essere utilizzati per valutare il risultato di un esperimento. Si basano tutti sul fatto che le ipotesi di cui sopra vengono formulate prima che i test statisticci vengano scelti ed eseguiti. I test statisticci sono elaborati nella Sez. 10.3.

Testare le ipotesi comporta diversi tipi di rischi. O il test rifiuta un'ipotesi vera oppure il test non rifiuta un'ipotesi falsa. Questi rischi sono definiti errore di tipo I ed errore di tipo II:

Errore di tipo I Un errore di tipo I si è verificato quando un test statisticco ha indicato uno schema o una relazione anche se in realtà non esiste uno schema reale. Cioè, la probabilità di commettere un errore di tipo I può essere espressa come: $P(\text{errore-tipo-I}) / DP(\text{reject } H_0 \mid H_0 \text{ true})$. Nell'ipotesi di esempio sopra, l'errore di tipo I è la probabilità di rifiutare H_0 anche se i due metodi in media trovano lo stesso numero di errori.

Errore di tipo II Un errore di tipo II si è verificato quando un test statisticco non ha indicato uno schema o una relazione anche se in realtà esiste uno schema reale. Cioè, la probabilità di commettere un errore di tipo II può essere espressa come: $P(\text{errore-tipo-II}) / DP(\text{non rifiutare } H_0 \mid H_0 \text{ falso})$. Nell'ipotesi di esempio sopra, l'errore di tipo II è la probabilità di non rifiutare H_0 anche se i due metodi in media hanno medie diverse.

La dimensione degli errori dipende da diversi fattori. Un esempio è la capacità del test statistico di rivelare un modello reale nei dati raccolti. Questo è indicato come il potere di un test:

Potenza La potenza di un test statistico è la probabilità che il test rivelì un modello vero se H_0 è falso.

Uno sperimentatore dovrebbe scegliere un test con la massima potenza possibile. La potenza può essere espressa come: Potenza DP .reject H_0 j H_0 false/ D 1 P .type-II-error)

Tutti questi fattori devono essere considerati quando si pianifica un esperimento.

8.3 Selezione delle variabili

Prima che qualsiasi progettazione possa iniziare dobbiamo scegliere le variabili dipendenti e indipendenti.

Le *variabili indipendenti* sono quelle variabili che possiamo controllare e modificare nell'esperimento. Scegliere le variabili giuste non è facile e di solito richiede la conoscenza del dominio. Le variabili dovrebbero avere qualche effetto sulla variabile dipendente e devono essere controllabili. Le scelte delle variabili indipendenti e dipendenti vengono spesso effettuate simultaneamente o in ordine inverso. La scelta delle variabili indipendenti comprende anche la scelta delle scale di misurazione, dell'intervallo delle variabili e dei livelli specifici ai quali verranno effettuati i test.

L'effetto dei trattamenti è misurato nelle *variabili dipendenti*. Spesso esiste una sola variabile dipendente e dovrebbe quindi essere derivata direttamente dall'ipotesi. La variabile nella maggior parte dei casi non è direttamente misurabile e dobbiamo invece misurarla tramite una misura indiretta. Questa misura indiretta deve essere validata attentamente, perché influenza il risultato dell'esperimento. L'ipotesi può essere perfezionata quando abbiamo scelto la variabile dipendente. La scelta della variabile dipendente significa anche che vengono determinati la scala di misurazione e l'intervallo delle variabili. Una ragione per avere una sola variabile dipendente è che se ce ne sono di più c'è il rischio che la minaccia "di pesca e del tasso di errore" alla validità della conclusione possa diventare troppo grande, come descritto nella Sez. [8.8.1](#).

8.4 Selezione dei soggetti

La selezione dei soggetti è importante quando si conduce un esperimento [\[144\]](#).

La selezione è strettamente connessa alla generalizzazione dei risultati dell'esperimento. Per generalizzare i risultati alla popolazione desiderata, la selezione deve essere rappresentativa per quella popolazione. La selezione dei soggetti è anche chiamata campione da una popolazione.

Il campionamento della popolazione può essere un campione probabilistico o non probabilistico. La differenza tra i due è che nel campionamento probabilistico, il

la probabilità di selezionare ciascun soggetto è nota mentre nel campionamento non probabilistico è sconosciuta. Esempi di *tecniche di campionamento probabilistico* sono:

- *Campionamento casuale semplice*: i soggetti vengono selezionati da un elenco della popolazione casuale.
- *Campionamento sistematico*: il primo soggetto viene selezionato dall'elenco della popolazione in modo casuale e poi ogni n-esima persona viene selezionata dall'elenco.
- *Campionamento casuale stratificato*: la popolazione è divisa in un numero di gruppi o strati con una distribuzione nota tra i gruppi. Viene quindi applicato il campionamento casuale all'interno degli strati.

Esempi di *tecniche di campionamento non probabilistico* sono:

- *Campionamento di convenienza*: vengono selezionate le persone più vicine e convenienti soggetti.
- *Campionamento per quote*: questo tipo di campionamento viene utilizzato per ottenere soggetti da vari elementi di una popolazione. Per ciascun elemento viene normalmente utilizzato il campionamento di convenienza.

Anche la dimensione del campione influisce sui risultati durante la generalizzazione. Più grande è il campione, minore diventa l'errore quando si generalizzano i risultati. La dimensione del campione è anche strettamente correlata alla potenza del test statistico, vedere Sez. 10.3.1. Esistono alcuni principi generali per la scelta della dimensione del campione:

- Se c'è una grande variabilità nella popolazione, è necessaria una dimensione del campione più ampia.
- L'analisi dei dati può influenzare la scelta della dimensione del campione. È quindi necessario considerare come analizzare i dati già in fase di progettazione dell'esperimento.

8.5 Progettazione dell'esperimento

Per trarre conclusioni significative da un esperimento, applichiamo metodi di analisi statistica sui dati raccolti per interpretare i risultati, come ulteriormente descritto nel Cap. 10. Per ottenere il massimo dall'esperimento, questo deve essere pianificato e progettato attentamente. Le analisi statistiche che possiamo applicare dipendono dal disegno scelto e dalle scale di misurazione utilizzate, vedere il Cap. 3. Pertanto progettazione e interpretazione sono strettamente correlate.

8.5.1 Scelta del disegno dell'esperimento

Un esperimento consiste in una serie di test dei trattamenti. Per ottenere il massimo dall'esperimento, la serie di test deve essere attentamente pianificata e progettata. La progettazione di un esperimento descrive come i test sono organizzati ed eseguiti. Più formalmente possiamo definire un esperimento come un insieme di test.

Come descritto sopra, la progettazione e l'analisi statistica sono strettamente correlate. La scelta del design influenza l'analisi e viceversa. Per progettare l'esperimento, dobbiamo esaminare l'ipotesi per vedere quale analisi statistica dobbiamo eseguire per rifiutare l'ipotesi nulla. Sulla base dei presupposti statistici, ad esempio delle scale di misurazione e su quali oggetti e soggetti siamo in grado di utilizzare, creiamo il disegno dell'esperimento. Durante la progettazione determiniamo quanti test dovrà effettuare l'esperimento per garantire che l'effetto del trattamento sia visibile. Una progettazione adeguata costituisce anche la base per consentire la replica. Nelle due sezioni seguenti vengono presentati i principi generali di progettazione e alcuni tipi di progettazione standard.

8.5.2 Principi generali di progettazione

Quando si progetta un esperimento è necessario considerare molti aspetti. I principi generali di progettazione sono *randomizzazione*, *blocco* e *bilanciamento* e la maggior parte dei progetti di esperimenti utilizza una combinazione di questi. Per illustrare i principi generali di progettazione, utilizziamo un esempio.

Esempio. Un'azienda condurrà un esperimento per studiare l'effetto sull'affidabilità di un programma quando si utilizza la progettazione orientata agli oggetti invece del principio di progettazione standard aziendale. L'esperimento utilizzerà il programma A come oggetto dell'esperimento. Il disegno dell'esperimento è del tipo "multi-test all'interno dello studio dell'oggetto", vedi cap. 7.

Randomizzazione. Uno dei principi di progettazione più importanti è la randomizzazione. Tutti i metodi statistici utilizzati per analizzare i dati richiedono che le osservazioni provengano da variabili casuali indipendenti. Per soddisfare questo requisito, viene utilizzata la randomizzazione. La randomizzazione si applica all'assegnazione degli oggetti, dei soggetti e all'ordine in cui vengono eseguiti i test. La randomizzazione viene utilizzata per mediare l'effetto di un fattore che potrebbe altrimenti essere presente. La randomizzazione viene utilizzata anche per selezionare soggetti rappresentativi della popolazione di interesse.

Esempio. La selezione delle persone (soggetti) sarà rappresentativa dei designer presenti in azienda, mediante selezione casuale dei designer disponibili. L'assegnazione a ciascun trattamento (progettazione orientata agli oggetti o principio di progettazione standard aziendale) viene selezionata in modo casuale.

Blocco. A volte abbiamo un fattore che probabilmente ha un effetto sulla risposta, ma quell'effetto non ci interessa. Se l'effetto del fattore è noto e controllabile, possiamo utilizzare una tecnica di progettazione chiamata blocking. Il blocco viene utilizzato per eliminare sistematicamente l'effetto indesiderato nel confronto tra i trattamenti. All'interno di un blocco, l'effetto indesiderato è lo stesso e possiamo studiare l'effetto dei trattamenti su quel blocco. Il blocco viene utilizzato per eliminare l'effetto indesiderato nello studio e pertanto gli effetti tra i blocchi non vengono studiati. Questa tecnica aumenta la precisione dell'esperimento.

Esempio. Le persone (soggetti) utilizzate, per questo esperimento, hanno esperienze diverse. Alcuni di loro hanno già utilizzato la progettazione orientata agli oggetti e altri no. Per ridurre al minimo l'effetto dell'esperienza, le persone sono raggruppate in due gruppi (blocchi), uno con esperienza di progettazione orientata agli oggetti e uno senza.

Bilanciamento. Se assegniamo i trattamenti in modo che ciascun trattamento abbia lo stesso numero di soggetti, abbiamo un disegno equilibrato. Il bilanciamento è auspicabile perché semplifica e rafforza l'analisi statistica dei dati, ma non è necessario.

Esempio. L'esperimento utilizza un disegno equilibrato, il che significa che in ciascun gruppo (blocco) è presente lo stesso numero di persone.

8.5.3 Tipi di progettazione standard

In questa sezione vengono presentati alcuni dei progetti di esperimenti utilizzati più frequentemente. I progetti spaziano da esperimenti semplici con un singolo fattore a esperimenti più complessi con molti fattori. La progettazione dell'esperimento è discussa in modo approfondito, ad esempio, da Montgomery [125] ed è elaborata più approfonditamente per l'ingegneria del software da Juristo e Moreno [88]. Per la maggior parte dei progetti viene formulata un'ipotesi di esempio e per ciascun progetto vengono suggeriti metodi di analisi statistica. I tipi di progettazione presentati in questa sezione sono adatti per esperimenti con:

- Un fattore con due trattamenti.
- Un fattore con più di due trattamenti.
- Due fattori con due trattamenti.
- Più di due fattori ciascuno con due trattamenti.

Un fattore con due trattamenti. Con questi esperimenti vogliamo confrontare i due trattamenti tra loro. Il più comune è confrontare le medie della variabile dipendente per ciascun trattamento. Vengono utilizzate le seguenti notazioni:

- La media della variabile dipendente per il trattamento i . y_{ij} La j -esima misura della variabile dipendente per il trattamento i .

Esempio di esperimento: l'obiettivo è verificare se un nuovo metodo di progettazione produce software con una qualità superiore rispetto al metodo di progettazione utilizzato in precedenza. Il fattore in questo esperimento è il metodo di progettazione e i trattamenti sono il nuovo e il vecchio metodo di progettazione. La variabile dipendente può essere il numero di difetti riscontrati nello sviluppo.

Design completamente randomizzato. Questo è un disegno sperimentale di base per confrontare due mezzi di trattamento. L'impostazione del progetto utilizza gli stessi oggetti per entrambi i trattamenti e assegna i soggetti in modo casuale a ciascun trattamento, vedere Tabella 8.1. Ogni soggetto utilizza un solo trattamento su un oggetto. Se abbiamo lo stesso numero di soggetti per trattamento il disegno è equilibrato.

Tabella 8.1 Esempio di assegnare i soggetti al trattamenti per un randomizzato progetto

Soggetti	Trattamento 1	Trattamento 2
1	X	
2		X
3		X
4	X	
5		X
6	X	

Tabella 8.2 Esempio di assegnare i trattamenti per a disegno abbinato

Soggetti	Trattamento 1	Trattamento 2
121		
212		
321		
421		
512		
612		

Esempio di ipotesi:

$$H_0 \quad 1^D \quad 2$$

$$L H_1 L 1 \neq 2; 1 < 2 \text{ o } 1 > 2$$

Esempi di analisi: t-test, Mann-Whitney, vedi sez. [10.3](#).

Progettazione di confronti accoppiati. A volte possiamo migliorare la precisione dell'esperimento effettuando confronti tra coppie di materiale sperimentale. In questo design, ogni soggetto utilizza entrambi i trattamenti sullo stesso oggetto. Questo a volte succede denominato design crossover. Questo tipo di design presenta alcune sfide, vale a dire ulteriormente discusso in relazione all'esempio nella Sez. [10.4](#). Per minimizzare l'effetto dell'ordine, in cui i soggetti applicano i trattamenti, l'ordine viene assegnato in modo casuale a ciascun soggetto, vedere la Tabella 8.2. Questo design non può essere applicato a tutti caso di confronto in quanto il soggetto può ricavare troppe informazioni dal primo trattamento per eseguire l'esperimento con il secondo trattamento. Il confronto per l'esperimento può consistere nel vedere se la differenza tra le misure accoppiate è zero. Se abbiamo lo stesso numero di soggetti che iniziano con il primo trattamento e con quello in secondo luogo, abbiamo un design equilibrato.

Esempio di ipotesi:

$$d \neq 0 \text{ e } d \text{ è la media della differenza.}$$

$$H_0 L d H_1 D_0$$

$$L p \neq 0; d < 0 \text{ o } d > 0$$

Esempi di analisi: t-test per dati appaiati, Sign test, Wilcoxon, vedi Sez. [10.3](#).

Un fattore con più di due trattamenti. Come con gli esperimenti con solo due trattamenti, vogliamo confrontare i trattamenti tra loro. Il confronto è spesso eseguito sui mezzi di trattamento.

Esempio di esperimento: l'esperimento esamina la qualità del software quando si utilizzano linguaggi di programmazione diversi. Il fattore nell'esperimento è il linguaggio di programmazione e i trattamenti possono essere C, CCC e Java.

Tabella 8.3 Esempio di affidare i trattamenti a i soggetti

	Soggetti	Trattamento 1	Trattamento 2	Trattamento 3
1			X	
2				X
3		X		
4		X		
5			X	
6				X

Tabella 8.4 Esempio di affidare i trattamenti a i soggetti

	Soggetti	Trattamento 1	Trattamento 2	Trattamento 3
113				2
231				2
323				1
421				3
532				1
612				3

Design completamente randomizzato. Un disegno completamente randomizzato richiede che l'esperimento viene eseguito in ordine casuale in modo che i trattamenti vengano utilizzati in un ambiente quanto più uniforme possibile. Il design utilizza un oggetto per tutti i trattamenti e i soggetti vengono assegnati in modo casuale ai trattamenti, vedere Tabella 8.3.

Esempio di ipotesi, dove a è il numero di soggetti:

$$H_0 \quad 1^D \quad 2^D \quad 3^D = D_{UN}$$

W H1 W i ≠ j per almeno una coppia .i; J /

Esempi di analisi: ANOVA (ANalysis Of VAriance) e Kruskal-Wallis, vedi Setta. 10.3.

Progettazione di blocchi completi randomizzati. Se la variabilità tra i soggetti è ampia, possiamo minimizzare questo effetto sul risultato utilizzando un blocco completo randomizzato progetto. Con questo disegno, ogni soggetto utilizza tutti i trattamenti e i soggetti formano a unità sperimentale più omogenea, cioè blocchiamo l'esperimento sui soggetti, vedere la Tabella 8.4. I blocchi rappresentano una restrizione alla randomizzazione. L'esperimento il design utilizza un oggetto per tutti i trattamenti e l'ordine in cui i soggetti lo utilizzano i trattamenti vengono assegnati in modo casuale. Il design di confronto accoppiato sopra è speciale caso di questo disegno con solo due trattamenti. Il disegno a blocchi completi randomizzati è uno dei progetti di esperimenti più utilizzati.

Esempio di ipotesi:

$$H_0 \quad 1^D \quad 2^D \quad 3^D = D_{UN}$$

H1 W i ≠ j per almeno una coppia .i; J /

W Esempi di analisi: ANOVA (ANalysis Of VAriance) e Kruskal-Wallis, vedere Setta. 10.3.

Due fattori. L'esperimento diventa più complesso quando aumentiamo da un fattore a due. La singola ipotesi per gli esperimenti con un fattore verrà divisa in tre ipotesi: un'ipotesi per l'effetto di uno dei fattori, una per l'altro e uno per l'interazione tra i due fattori. Utilizziamo le seguenti notazioni:

Tabella 8.5 Esempio di 2*2 disegno fattoriale

		Fattore A
		Trattamento A1 Trattamento A2
Fattore B	Trattamento B1	Soggetto 4, 6
	Trattamento B2	Soggetto 2, 3
		Oggetto 1, 7
		Argomento 5, 8

L'effetto del trattamento i sul fattore A.
 \bar{y}_j L'effetto del trattamento j sul fattore B.
 $\cdot\bar{y}_{ij}$ L'effetto dell'interazione tra \bar{y}_i e \bar{y}_j .

*Disegno fattoriale 2*2.* Questo disegno ha due fattori, ciascuno con due trattamenti. In questo disegno dell'esperimento, assegniamo casualmente i soggetti a ciascuna combinazione di trattamenti, vedere Tabella 8.5.

Esempio di esperimento: L'esperimento indaga la comprensibilità di il documento di progettazione quando si utilizza la progettazione strutturata o orientata agli oggetti basata su uno documenti con requisiti "buoni" e uno "cattivo". Il primo fattore, A, è il design metodo e il secondo fattore, B, è il documento dei requisiti. L'esperimento il disegno è un disegno fattoriale 2*2 poiché entrambi i fattori hanno due trattamenti e ciascuno è possibile la combinazione dei trattamenti

Esempio di ipotesi:

$$\begin{aligned} H_0: & DQ_i = 0 \quad \forall i \\ H_1: & W \text{ almeno uno } i \neq 0 \\ H_0: & L \bar{y}_1 P \bar{y}_2 P = 0 \\ H_1: & W \text{ almeno uno } \bar{y}_j \neq 0 \\ H_0: & W \cdot \bar{y}_{ij} D = 0 \text{ per tutti } i,j \\ H_1: & W \text{ almeno uno } \bar{y}_{ij} \neq 0 \end{aligned}$$

Esempio di analisi: ANOVA (ANalysis Of VAriance), vedi Sez. 10.3.

Design nidificato a due stadi. Se uno dei fattori, ad esempio B, nell'esperimento è simili ma non identici per trattamenti diversi dell'altro fattore, ad esempio A, abbiamo un design che si chiama design nidificato o gerarchico. Si dice che il fattore B sia annidato sotto il fattore A. Il disegno annidato a due stadi ha due fattori, ciascuno con due o più trattamenti. La progettazione e l'analisi dell'esperimento sono le stesse del 2*2 disegno fattoriale, vedere la Tabella 8.6.

Esempio di esperimento: L'esperimento esamina l'efficienza del test dell'unità test di un programma quando si utilizza la programmazione orientata alle funzioni o agli oggetti e se i programmi sono "incline a difetti" o "non inclini a difetti". Il primo fattore, A, è il linguaggio di programmazione e il secondo fattore, B, è la predisposizione ai difetti del programma. Il disegno dell'esperimento deve essere nidificato, poiché un programma funzionale "incline ai difetti/non incline ai difetti" non è la stessa cosa di un programma funzionale "incline ai difetti/non incline ai difetti" programma orientato agli oggetti.

Più di due fattori. In molti casi, l'esperimento deve considerare più di due fattori. L'effetto nella variabile dipendente può quindi essere non dipendente solo su ciascun fattore separatamente ma anche sulle interazioni tra i fattori.

Tabella 8.6 Esempio di progettazione nidificata a due stadi in cui B è nidificato sotto A

Fattore A			
Trattamento A1		Trattamento A2	
Fattore B		Fattore B	
Trattamento B10 Trattamento B20 Trattamento B100 Trattamento B200			
Oggetto 1, 3	Soggetto 6, 2	Argomento 7, 8	Argomento 5, 4

Tabella 8.7 Esempio di disegno fattoriale a 23

Fattore A	Fattore B	Fattore C	Soggetti
A1	B1	C1	2, 3
A2	B1	C1	1, 13
A1	B2	C1	5, 6
A2	B2	C1	10, 16
A1	B1	C2	7, 15
A2	B1	C2	8, 11
A1	B2	C2	4, 9
A2	B2	C2	12, 14

Queste interazioni possono avvenire tra due o più fattori. Questo tipo di disegni è chiamati disegni fattoriali. Questa sezione fornisce un'introduzione ai progetti in cui ciascuno factor ha solo due trattamenti ciascuno. Disegni in cui i fattori sono più di due i trattamenti sono presentati da Montgomery [125].

Disegno fattoriale 2k . Il disegno fattoriale 2^k è un caso speciale del fattoriale 2k disegno fattoriale, cioè quando $k = D$. Il disegno fattoriale 2k ha k fattori dove ciascun fattore ha due trattamenti. Ciò significa che ci sono 2^k diverse combinazioni di trattamenti. Per valutare gli effetti dei fattori k, è necessario che tutte le combinazioni siano valutate essere testato. I soggetti vengono assegnati in modo casuale alle diverse combinazioni. Un esempio di disegno fattoriale 23 è mostrato nella Tabella 8.7.

Le ipotesi e le analisi per questo tipo di progettazione sono dello stesso tipo di per il disegno fattoriale 2^k . Maggiori dettagli sulla cura della progettazione fattoriale 2k presentata di Montgomery [125].

Disegno fattoriale frazionario 2k . Quando il numero di fattori cresce in un fattoriale di 2k progettazione, il numero di combinazioni di fattori cresce rapidamente, ad esempio, ci sono 8 combinazioni per un disegno fattoriale a 23 e 16 per un disegno fattoriale a 24 . Spesso può si presuppone che gli effetti di certe interazioni di ordine superiore siano trascurabili e che gli effetti principali e gli effetti di interazione di ordine inferiore possono essere ottenuti eseguendo a frazione dell'esperimento fattoriale completo. Questo tipo di disegno viene quindi chiamato disegno fattoriale frazionario.

Il disegno fattoriale frazionario si basa su tre idee:

- **Il principio della scarsità degli effetti:** è probabile che il sistema sia guidato principalmente da alcuni degli effetti di interazione principali e di ordine inferiore.

Tabella 8.8 Esempio di una mezza frazione di 23 disegno fattoriale

Fattore A	Fattore B	Fattore C	Soggetti
A1	B1	C2	2, 3
A2	B1	C1	1, 8
A1	B2	C1	5, 6
A2	B2	C2	4, 7

Tabella 8.9 Esempio di una frazione di un quarto del disegno fattoriale a 25

Fattore A	Fattore B	Fattore C	Fattore D	Fattore E	Soggetti
A1	B1	C1	D2	E2	3, 16
A2	B1	C1	D1	E1	7, 9
A1	B2	C1	D1	E2	1, 4
A2	B2	C1	D2	E1	8, 10
A1	B1	C2	D2	E1	5, 12
A2	B1	C2	D1	E2	2, 6
A1	B2	C2	D1	E1	11, 15
A2	B2	C2	D2	E2	13, 14

- **La proprietà di proiezione:** un progetto più forte può essere ottenuto prendendo un sottoinsieme di fattori significativi dal disegno fattoriale frazionario.
- **Sperimentazione sequenziale:** combinando si può ottenere un design più forte esecuzioni sequenziali di due o più disegni fattoriali frazionari.

L'uso principale di questi disegni fattoriali frazionari è negli esperimenti di screening, dove lo scopo dell'esperimento è identificare i fattori che hanno grandi effetti sul sistema. Esempi di disegni fattoriali frazionari sono:

Metà del disegno fattoriale frazionario del disegno fattoriale 2k : metà del vengono scelte le combinazioni di un disegno fattoriale completo 2k . Le combinazioni sono selezionate in modo che se un fattore viene rimosso il disegno rimanente è un disegno fattoriale completo 2k1 , vedere la Tabella 8.8. I soggetti vengono assegnati in modo casuale alle combinazioni selezionate. Ci sono due frazioni alternative in questo progetto e se vengono utilizzate entrambe le frazioni sequenza, il disegno risultante è un disegno fattoriale completo di 2k .

Disegno fattoriale frazionario di un quarto del disegno fattoriale 2k : Un quarto viene scelta una delle combinazioni del disegno fattoriale completo 2k . Le combinazioni sono selezionati in modo tale che se due fattori vengono rimossi il disegno rimanente è un 2k2 completo disegno fattoriale, vedere la Tabella 8.9. Esistono tuttavia delle dipendenze tra i fattori nel disegno a un quarto poiché non è un disegno fattoriale completo.

Ad esempio, nella Tabella 8.9, il fattore D dipende da una combinazione del fattore A e B. Si può vedere, ad esempio, che per tutte le combinazioni di A1 e B1 abbiamo D2 e così via. In modo simile, il fattore E dipende da una combinazione di fattori A e C. Pertanto, se i fattori C ed E (o B e D) vengono rimossi, il disegno risultante diventano due repliche di un disegno fattoriale frazionario 231 e non di un disegno fattoriale 23 progetto. Quest'ultimo disegno si ottiene rimuovendo D ed E. Le due repliche

può essere identificato nella Tabella 8.9 notando che le prime quattro righe sono equivalenti alle quattro ultime righe della tabella, quando C ed E vengono rimosse, e quindi diventano due repliche di un disegno fattoriale a 2².

I soggetti vengono assegnati in modo casuale alle combinazioni selezionate. Ci sono quattro frazioni alternative in questo disegno e se tutte e quattro le frazioni vengono utilizzate in sequenza, il disegno risultante è un disegno fattoriale completo di 2^k. Se due delle frazioni vengono utilizzate in sequenza si ottiene un disegno a metà frazionario.

Maggiori dettagli sui disegni fattoriali frazionari sono presentati da Montgomery [125].

In sintesi, la scelta del corretto disegno sperimentale è cruciale, poiché un disegno inadeguato influenzerebbe senza dubbio la possibilità di poter trarre conclusioni corrette dopo lo studio. Inoltre, il progetto pone dei vincoli sui metodi statistici che possono essere applicati. Va infine sottolineato che è importante cercare di utilizzare se possibile un design semplice e cercare di sfruttare al meglio i soggetti a disposizione.

8.6 Strumentazione

Gli strumenti per un esperimento sono di tre tipi: oggetti, linee guida e strumenti di misurazione. Nella pianificazione di un esperimento si scelgono gli strumenti. Prima dell'esecuzione, gli strumenti vengono sviluppati per l'esperimento specifico.

Gli oggetti dell'esperimento possono essere, ad esempio, specifiche o documenti di codice. Quando si pianifica un esperimento, è importante scegliere oggetti appropriati. In un esperimento di ispezione, ad esempio, deve essere noto il numero di difetti negli oggetti da ispezionare. Ciò può essere ottenuto seminando errori o utilizzando un documento con un numero noto di errori. Utilizzando una vera versione iniziale di un documento in cui sono identificati i difetti si può ottenere quest'ultimo risultato.

Sono necessarie linee guida per guidare i partecipanti all'esperimento. Le linee guida includono, ad esempio, descrizioni dei processi e liste di controllo. Se nell'esperimento vengono confrontati metodi diversi, è necessario preparare le linee guida per i metodi per l'esperimento. Oltre alle linee guida, i partecipanti necessitano anche di formazione sui metodi da utilizzare.

Le misurazioni in un esperimento vengono condotte tramite la raccolta dati. Negli esperimenti ad alta intensità umana, i dati vengono generalmente raccolti tramite moduli manuali o interviste. Il compito di pianificazione da eseguire è quello di preparare moduli e domande per il colloquio e di convalidare i moduli e le domande con alcune persone con background e competenze simili a quelli dei partecipanti all'esperimento. Tra gli esercizi è mostrato un esempio di modulo utilizzato per raccogliere informazioni sull'esperienza dei soggetti, vedere Tabella A.1 nell'Appendice A.

L'obiettivo generale della strumentazione è fornire i mezzi per eseguire l'esperimento e monitorarlo, senza influire sul controllo dell'esperimento.

I risultati dell'esperimento saranno gli stessi indipendentemente da come si svolge l'esperimento

è strumentato. Se la strumentazione influisce sull'esito dell'esperimento, i risultati non sono validi.

La validità di un esperimento è elaborata nella Sez. 8.7 e altro su preparazione degli strumenti si trova nelle Sez. 9.1.2 e 9.2.2.

8.7 Valutazione di validità

Una domanda fondamentale riguardante i risultati di un esperimento è quanto siano validi i risultati. È importante considerare la questione della validità già nella fase di pianificazione per pianificare un'adeguata validità dei risultati dell'esperimento. La validità adeguata si riferisce al fatto che i risultati dovrebbero essere validi per la popolazione di interesse. Innanzitutto i risultati dovrebbero essere validi per la popolazione da cui viene estratto il campione.

In secondo luogo, potrebbe essere interessante generalizzare i risultati ad una popolazione più ampia. Si dice che i risultati abbiano validità adeguata se sono validi per la popolazione a cui si vuole generalizzare.

Una validità adeguata non implica necessariamente una validità più generale. Un esperimento condotto all'interno di un'organizzazione può essere progettato per rispondere ad alcune domande esclusivamente per quell'organizzazione, ed è sufficiente che i risultati siano validi all'interno di quella specifica organizzazione. D'altro canto, se si devono trarre conclusioni più generali, la validità deve coprire anche un ambito più generale.

Esistono diversi schemi di classificazione per diversi tipi di minacce alla validità di un esperimento. Campbell e Stanley definiscono due tipi, minacce alla validità interna ed esterna [32]. Cook e Campbell estendono l'elenco a quattro tipi di minacce alla validità dei risultati sperimentali. Le quattro minacce sono *conclusione, validità interna, costrutto e validità esterna* [37]. La prima categorizzazione viene talvolta citata in letteratura, ma la seconda è preferibile poiché è facilmente mappabile nelle diverse fasi coinvolte nella conduzione di un esperimento, vedere Fig. 8.2.

Ciascuna delle quattro categorie presentate da Cook e Campbell [37] è correlata a una questione metodologica nella sperimentazione. I principi di base di un esperimento sono presentati nella Fig. 8.2.

In alto abbiamo l'area teorica e in basso l'area di osservazione.

Vogliamo trarre conclusioni sulla teoria definita nelle ipotesi, sulla base delle nostre osservazioni. Nel trarre conclusioni abbiamo quattro passaggi, in ognuno dei quali esiste un tipo di minaccia alla validità dei risultati.

1. *Validità della conclusione.* Questa validità riguarda la relazione tra il trattamento e il risultato.
Vogliamo assicurarci che ci sia una statistica
relazione, cioè con un dato significato.
2. *Validità interna.* Se si osserva una relazione tra il trattamento e il risultato, dobbiamo assicurarci
che si tratti di una relazione causale e che non sia il risultato di un fattore su cui non abbiamo
alcun controllo o che non abbiamo misurato. Nell'altro
parole che il trattamento provoca il risultato (l'effetto).

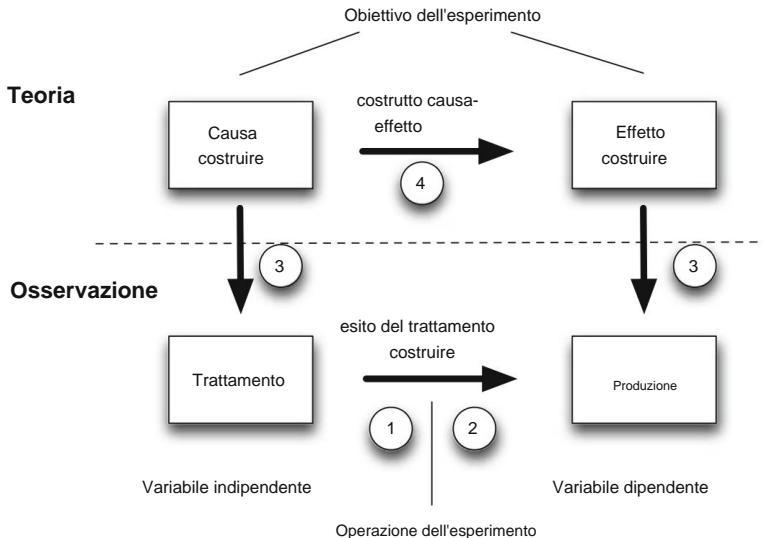


Fig. 8.2 Principi dell'esperimento (adattato da Trochim [171])

3. *Validità di costrutto.* Questa validità riguarda la relazione tra teoria e osservazione. Se la relazione tra causa ed effetto è causale, dobbiamo farlo garantire due cose: (1) che il trattamento riflette bene il costrutto della causa (vedi parte sinistra della Fig. 8.2) e (2) che il risultato riflette il costrutto del effetto bene (vedere la parte destra della Fig. 8.2).

4. *Validità esterna.* La validità esterna riguarda la generalizzazione. Se li è una relazione causale tra il costrutto della causa e l'effetto, can il risultato dello studio essere generalizzato al di fuori dell'ambito del nostro studio? C'è un rapporto tra trattamento e risultato?

La validità della conclusione è talvolta definita validità statistica della conclusione [37], e ha la sua controparte nell'affidabilità per l'analisi qualitativa, vedere la sez. 5.4.3.

Le minacce alla validità della conclusione riguardano questioni che influenzano la capacità di farlo trarre la conclusione corretta sulle relazioni tra il trattamento e il risultato di un esperimento. Tali questioni includono, ad esempio, la scelta dei test statistici, scelta delle dimensioni del campione, cura posta nell'implementazione e misurazione di un sperimentare.

Le minacce alla validità interna riguardano questioni che possono indicare una relazione causale, anche se non ce n'è. I fattori che influiscono sulla validità interna sono il modo in cui le materie vengono selezionate e divise in diverse classi, come sono le materie trattati e compensati durante l'esperimento, se si verificano eventi speciali durante l'esperimento ecc. Tutti questi fattori possono far sì che l'esperimento mostri un comportamento che non è dovuto al trattamento ma al fattore di disturbo.

Le minacce alla validità della costruzione si riferiscono alla misura in cui l'impostazione dell'esperimento riflette effettivamente il costrutto in studio. Ad esempio, il numero di corsi

sostenuto all'università in informatica può essere una misura inadeguata dell'esperienza del soggetto in un linguaggio di programmazione, ovvero ha una scarsa validità di costrutto. Il numero di anni di utilizzo pratico può essere una misura migliore, ovvero ha una validità di costrutto migliore.

Le minacce alla validità esterna riguardano la capacità di generalizzare i risultati dell'esperimento al di fuori del contesto sperimentale. La validità esterna è influenzata dal disegno dell'esperimento scelto, ma anche dagli oggetti dell'esperimento e dai soggetti scelti. I rischi principali sono tre: avere come soggetti partecipanti sbagliati, condurre l'esperimento in un ambiente sbagliato ed eseguirlo con tempistiche che influiscono sui risultati.

Un elenco dettagliato delle minacce alla validità è presentato nella Sez. 8.8. Questo elenco può essere utilizzato come lista di controllo per la progettazione di un esperimento. Nella valutazione della validità, ciascuno degli elementi viene controllato per vedere se ci sono minacce. Se ce ne sono, devono essere affrontati o accettati, poiché a volte è necessario accettare qualche minaccia alla validità. Potrebbe anche essere impossibile condurre un esperimento senza determinate minacce e quindi queste devono essere accettate e poi affrontate quando si interpretano i risultati.

La priorità tra i diversi tipi di minacce è ulteriormente discussa nella Sez. 8.9.

8.8 Descrizione dettagliata delle minacce alla validità

Di seguito viene discusso un elenco di minacce alla validità degli esperimenti basato su Cook e Campbell [37]. Non tutte le minacce sono applicabili a tutti gli esperimenti, ma questo elenco può essere visto come una lista di controllo. Le minacce sono riassunte nella Tabella 8.10 e lo schema di classificazione alternativo e limitato [32] è riassunto nella Tabella 8.11.

8.8.1 Conclusione Validità

Le minacce alla validità della conclusione riguardano questioni che influenzano la capacità di trarre la conclusione corretta sulle relazioni tra il trattamento e il risultato di un esperimento.

Basso potere statistico. La potenza di un test statistico è la capacità del test di rivelare un modello reale nei dati. Se la potenza è bassa c'è un alto rischio che si tragga una conclusione errata, vedere oltre par. 8.2 o più specificatamente non siamo in grado di respingere un'ipotesi errata.

Presupposti violati dei test statistici. Alcuni test prevedono ipotesi, ad esempio, su campioni normalmente distribuiti e indipendenti. La violazione delle ipotesi può portare a conclusioni errate. Alcuni test statistici sono più robusti rispetto ad altri rispetto ad ipotesi violate, vedere il Cap. 10.

La pesca e il tasso di errore. Questa minaccia contiene due parti separate. Cercare o "pescare" un risultato specifico è una minaccia, poiché le analisi non sono più indipendenti e i ricercatori possono influenzare il risultato cercando un risultato specifico.

Tabella 8.10 Minacce alla validità secondo Cook e Campbell [37]

Validità delle conclusioni	Validità interna
Basso potere statistico	Storia
Assunzione violata dei test statistici Pesca e tasso di errore Affidabilità delle misure Affidabilità dell'implementazione del trattamento Irrilevanze casuali nel contesto sperimentale Eterogeneità casuale dei soggetti	Maturazione Test Strumentazione Regressione statistica Selezione Mortalità Ambiguità sulla direzione causale influenza Interazioni con la selezione Diffusione dell'imitazione dei trattamenti Perequazione compensativa dei trattamenti Rivalità compensativa Demoralizzazione risentita
Validità di costrutto	Validità esterna
Spiegazione preoperatoria inadeguata dei costrutti. Interazione di selezione e trattamento Distorsione da monooperazione Distorsione da monometodo Costrutti e livelli di costrutti confondenti Interazione di diversi trattamenti Interazione tra test e trattamento Generalizzabilità limitata tra costrutti Ipotesi indovinata Apprensione valutativa Aspettativa dello sperimentatore	Interazione tra setting e trattamento Interazione tra anamnesi e trattamento

Tabella 8.11 Minacce alla validità secondo Campbell e Stanley [32]

Validità interna	Validità esterna
Storia	Interazione tra selezione e trattamento
Maturazione	Interazione tra anamnesi e trattamento
Test	Interazione tra setting e trattamento
Strumentazione	Interazione di diversi trattamenti
Regressione statistica	
Selezione	

Il tasso di errore riguarda il livello di significatività effettivo. Per esempio, condurre tre indagini con un livello di significatività di 0:05 significa che il livello di significatività totale è $1 - (1 - 0:05)^3$, che equivale a 0:14. Il tasso di errore (cioè il livello di significatività) dovrebbe quindi essere adeguato quando si conducono analisi multiple.

Affidabilità delle misure. La validità di un esperimento dipende fortemente da affidabilità delle misure. Questo a sua volta può dipendere da molti fattori diversi, come formulazione inadeguata delle domande, cattiva strumentazione o cattiva disposizione dello strumento. La base

Il principio è che quando si misura un fenomeno due volte, il risultato sarà lo stesso. Ad esempio, le righe di codice sono più affidabili dei punti funzione poiché non implicano il giudizio umano. In altre parole, le misure oggettive, che possono essere ripetute con lo stesso risultato, sono più affidabili delle misure soggettive, vedi anche Cap. 3.

Affidabilità dell'implementazione del trattamento. Per dare esecuzione al trattamento si intende l'applicazione dei trattamenti ai soggetti. Esiste il rischio che l'attuazione non sia simile tra le diverse persone che applicano il trattamento o tra diverse occasioni. L'implementazione dovrebbe quindi essere il più standard possibile per diversi argomenti e occasioni.

Irrilevanze casuali in ambito sperimentale. Elementi esterni all'ambiente sperimentale possono disturbare i risultati, come rumore fuori dalla stanza o un'improvvisa interruzione dell'esperimento.

Eterogeneità casuale dei soggetti. C'è sempre eterogeneità in un gruppo di studio.

Se il gruppo è molto eterogeneo, c'è il rischio che la variazione dovuta alle differenze individuali sia maggiore di quella dovuta al trattamento. La scelta di gruppi più omogenei influenzerà invece la validità esterna, vedi sotto. Ad esempio, un esperimento con studenti universitari riduce l'eterogeneità, poiché hanno conoscenze e background più simili, ma riduce anche la validità esterna dell'esperimento, poiché i soggetti non sono selezionati da una popolazione sufficientemente generale.

8.8.2 Validità interna

Le minacce alla validità interna sono influenze che possono influenzare la variabile indipendente rispetto alla causalità, all'insaputa del ricercatore. In questo modo minacciano la conclusione su una possibile relazione causale tra trattamento ed esito.

Le minacce alla validità interna sono talvolta classificate in tre categorie: *minacce a gruppo singolo*, *minacce a più gruppi* e *minacce sociali*.

Minacce di gruppo singolo. Queste minacce si applicano agli esperimenti con singoli gruppi. Non abbiamo un gruppo di controllo a cui non applichiamo il trattamento. Quindi, ci sono problemi nel determinare se il trattamento o un altro fattore abbia causato l'effetto osservato.

Storia. In un esperimento è possibile applicare trattamenti diversi allo stesso oggetto in momenti diversi. Allora c'è il rischio che la storia influenzi i risultati sperimentali, poiché le circostanze non sono le stesse in entrambe le occasioni. Ad esempio, se una delle occasioni dell'esperimento avviene il primo giorno dopo un giorno festivo o un giorno in cui si verifica un evento molto raro, e l'altra occasione è un giorno normale.

Maturazione. Questo è l'effetto che i soggetti reagiscono in modo diverso col passare del tempo. Esempi sono quando i soggetti vengono influenzati negativamente (stanchi o annoiati) durante l'esperimento, o positivamente (imparando) durante il corso dell'esperimento.

Test. Se il test viene ripetuto, i soggetti possono rispondere in modo diverso in momenti diversi poiché sanno come viene condotto il test. Se è necessario familiarizzare con i test, è importante che i risultati del test non vengano restituiti al soggetto, per non supportare un apprendimento involontario.

Strumentazione. Questo è l'effetto causato dagli artefatti utilizzati per l'esecuzione dell'esperimento, come moduli di raccolta dati, documenti da ispezionare in un esperimento di ispezione, ecc. Se questi sono progettati male, l'esperimento ne risente negativamente.

Regressione statistica. Ciò costituisce una minaccia quando i soggetti vengono classificati in gruppi sperimentali sulla base di un esperimento o di un caso di studio precedente, ad esempio i primi dieci o gli ultimi dieci. In questo caso potrebbe esserci un aumento o un miglioramento, anche se non viene applicato alcun trattamento. Ad esempio, se gli ultimi dieci in un esperimento vengono selezionati come soggetti in base a un esperimento precedente, probabilmente non saranno tutti tra gli ultimi dieci nel nuovo esperimento a causa della pura variazione casuale.

Gli ultimi dieci non possono essere peggio che rimanere tra gli ultimi dieci, e quindi l'unico cambiamento possibile è in meglio, relativamente alla popolazione più ampia da cui vengono selezionati.

Selezione. Questo è l'effetto della variazione naturale delle prestazioni umane. A seconda di come i soggetti vengono selezionati da un gruppo più ampio, gli effetti di selezione possono variare. Inoltre, l'effetto di consentire ai volontari di prendere parte a un esperimento può influenzare i risultati. I volontari sono generalmente più motivati e adatti a un nuovo compito rispetto all'intera popolazione. Pertanto il gruppo selezionato non è rappresentativo dell'intera popolazione.

Mortalità. Questo effetto è dovuto alle diverse tipologie di persone che abbandonano l'esperimento. È importante caratterizzare i dropout per verificare se sono rappresentativi del campione totale. Se i soggetti di una categoria specifica abbandonano, ad esempio, tutti i revisori senior in un esperimento di ispezione, la validità dell'esperimento ne risente notevolmente.

Ambiguità sulla direzione dell'influenza causale. La questione è se A causa B, B causa A o anche X causa A e B. Un esempio è se si osserva una correlazione tra la complessità del programma e il tasso di errore. La domanda è se l'elevata complessità del programma causa un alto tasso di errore, o viceversa, o se l'elevata complessità del problema da risolvere causa entrambi.

La maggior parte delle minacce alla validità interna possono essere affrontate attraverso la progettazione dell'esperimento. Ad esempio, introducendo un gruppo di controllo è possibile controllare molte delle minacce interne. D'altra parte, vengono invece introdotte più minacce di gruppo.

Minacce di più gruppi. In un esperimento su più gruppi, vengono studiati gruppi diversi. La minaccia per tali studi è che il gruppo di controllo e i gruppi sperimentali selezionati possono essere influenzati in modo diverso dalle minacce del singolo gruppo come sopra definite. Quindi ci sono interazioni con la selezione.

Interazioni con la selezione. Le interazioni con la selezione sono dovute a comportamenti diversi nei diversi gruppi. Ad esempio, l'interazione selezione-maturazione significa che gruppi diversi maturano a velocità diverse, ad esempio se due gruppi

applicare un nuovo metodo ciascuno. Se un gruppo apprende il nuovo metodo più velocemente dell'altro, grazie alla sua capacità di apprendimento, i gruppi selezionati maturano in modo diverso. La storia della selezione significa che gruppi diversi sono influenzati dalla storia in modo diverso, ecc.

Minacce sociali alla validità interna. Queste minacce sono applicabili agli esperimenti a gruppo singolo e a gruppi multipli. Di seguito vengono forniti esempi di un esperimento di ispezione in cui un nuovo metodo (lettura basata sulla prospettiva) viene confrontato con uno vecchio (lettura basata su lista di controllo).

Diffusione o imitazione di trattamenti. Questo effetto si verifica quando un gruppo di controllo apprende il trattamento dal gruppo nello studio sperimentale o cerca di imitare il comportamento del gruppo nello studio. Ad esempio, se un gruppo di controllo utilizza un metodo di ispezione basato su una lista di controllo e il gruppo sperimentale utilizza metodi basati sulla prospettiva, il primo gruppo potrebbe venire a conoscenza del metodo basato sulla prospettiva ed eseguire le proprie ispezioni influenzato dalla propria prospettiva. Quest'ultimo caso può verificarsi se il revisore è un esperto in una determinata area.

Perequazione compensativa dei trattamenti. Se a un gruppo di controllo viene concesso un compenso per essere un gruppo di controllo, in sostituzione di ciò non riceveranno trattamenti; ciò potrebbe influenzare l'esito dell'esperimento. Se al gruppo di controllo viene insegnato un altro nuovo metodo come compensazione per non aver appreso il metodo basato sulla prospettiva, la loro prestazione potrebbe essere influenzata da quel metodo.

Rivalità compensativa. Un soggetto che riceve trattamenti meno desiderabili può, in quanto perdente naturale, essere motivato a ridurre o invertire il risultato atteso dell'esperimento. Il gruppo che utilizza il metodo tradizionale può fare del suo meglio per dimostrare che il vecchio metodo è competitivo.

Demoralizzazione risentita. Questo è l'opposto della minaccia precedente. Un soggetto che riceve trattamenti meno desiderabili può arrendersi e non ottenere risultati così buoni come in genere. Il gruppo che utilizza il metodo tradizionale non è motivato a fare un buon lavoro, mentre imparare qualcosa di nuovo ispira il gruppo che utilizza il nuovo metodo.

8.8.3 Validità di costrutto

La validità di costrutto riguarda la generalizzazione del risultato dell'esperimento al concetto o alla teoria alla base dell'esperimento. Alcune minacce riguardano la progettazione dell'esperimento, altre a fattori sociali.

Minacce di progettazione. Le minacce progettuali alla validità del costrutto riguardano questioni legate alla progettazione dell'esperimento e alla sua capacità di riflettere il costrutto da studiare.

Spiegazione preoperatoria inadeguata dei costrutti. Questa minaccia, nonostante il suo titolo ampio, è piuttosto semplice. Vuol dire che i costrutti non sono sufficientemente definiti, prima di essere tradotti in misure o trattamenti. La teoria non è abbastanza chiara e quindi l'esperimento non può essere sufficientemente chiaro. Ad esempio, se si confrontano due metodi di ispezione e non è sufficientemente chiaro quale sia

'migliore' significa. Significa trovare il maggior numero di guasti, il maggior numero di guasti all'ora o i guasti più gravi?

Distorsione da monooperazione. Se l'esperimento include una singola variabile indipendente, caso, soggetto o trattamento, l'esperimento potrebbe sottorappresentare il costrutto e quindi non fornire il quadro completo della teoria. Ad esempio, se un esperimento di ispezione viene condotto con un singolo documento come oggetto, il costrutto causa è sottorappresentato.

Distorsione da monometodo. L'utilizzo di un unico tipo di misure o osservazioni comporta il rischio che, se tale misura o osservazione fornisce una distorsione nella misurazione, l'esperimento sarà fuorviante. Coinvolgendo diversi tipi di misure e osservazioni è possibile confrontarle tra loro. Ad esempio, se il numero di difetti riscontrati viene misurato in un esperimento di ispezione, dove la classificazione dei difetti si basa su un giudizio soggettivo, le relazioni non possono essere sufficientemente spiegate. Lo sperimentatore può influenzare le misure.

Costrutti e livelli di costrutti confondenti. In alcune relazioni non è principalmente la presenza o l'assenza di un costrutto, ma il livello del costrutto ad essere importante per il risultato. L'effetto della presenza del costrutto si confonde con l'effetto del livello del costrutto. Ad esempio, la presenza o l'assenza di conoscenze pregresse in un linguaggio di programmazione potrebbe non spiegare le cause di un esperimento, ma la differenza può dipendere dal fatto che i soggetti abbiano 1, 3 o 5 anni di esperienza con il linguaggio corrente.

Interazione di diversi trattamenti. Se il soggetto è coinvolto in più di uno studio, i trattamenti dei diversi studi potrebbero interagire. Quindi non è possibile concludere se l'effetto sia dovuto a uno dei trattamenti o ad una combinazione di essi trattamenti.

Interazione tra test e trattamento. Il test stesso, cioè l'applicazione dei trattamenti, può rendere i soggetti più sensibili o ricettivi al trattamento. Quindi il test fa parte del trattamento. Ad esempio, se il test prevede la misurazione del numero di errori commessi nella codifica, i soggetti saranno più consapevoli degli errori commessi e quindi cercheranno di ridurli.

Generalizzabilità limitata tra costrutti. Il trattamento può influenzare positivamente il costrutto studiato, ma involontariamente influenzare negativamente altri costrutti. Questa minaccia rende difficile generalizzare il risultato in altri risultati potenziali. Ad esempio, uno studio comparativo conclude che con un nuovo metodo si ottiene una migliore produttività. D'altro canto si può osservare che ciò riduce la manutenibilità, il che è un effetto collaterale non voluto. Se la manutenibilità non viene misurata o osservata, c'è il rischio che si traggano conclusioni basate sull'attributo produttività, ignorando la manutenibilità.

Minacce sociali alla costruzione della validità. Queste minacce riguardano questioni relative al comportamento dei soggetti e degli sperimentatori. Potrebbero, in base al fatto che fanno parte di un esperimento, agire diversamente da come fanno altrimenti, il che dà risultati falsi dall'esperimento.

Ipotesi indovinata. Quando le persone prendono parte a un esperimento potrebbero provare a capire quale sia lo scopo e il risultato previsto dell'esperimento. Quindi è probabile che basino il loro comportamento sulle loro ipotesi riguardo alle ipotesi, sia positivamente che negativamente, a seconda del loro atteggiamento nei confronti dell'ipotesi anticipata.

Apprensione valutativa. Alcune persone hanno paura di essere valutate. Una forma di tendenza umana è quella di cercare di apparire migliori quando vengono valutati, il che è confuso con l'esito dell'esperimento. Ad esempio, se si confrontano diversi modelli di stima, le persone potrebbero non riportare le loro vere deviazioni tra stima e risultato, ma alcuni valori falsi ma "migliori".

Aspettative dello sperimentatore. Gli sperimentatori possono influenzare i risultati di uno studio sia consciamente che inconsciamente in base a ciò che si aspettano dall'esperimento. La minaccia può essere ridotta coinvolgendo persone diverse che non hanno aspettative o che hanno aspettative diverse rispetto all'esperimento. Ad esempio, le domande possono essere poste in diversi modi per dare le risposte desiderate.

8.8.4 Validità esterna

Le minacce alla validità esterna sono condizioni che limitano la nostra capacità di generalizzare i risultati del nostro esperimento alla pratica industriale. Esistono tre tipi di interazioni con il trattamento: persone, luogo e tempo:

Interazione tra selezione e trattamento. Questo è l'effetto di avere una popolazione soggetta, non rappresentativa della popolazione a cui vogliamo generalizzare, cioè le persone sbagliate partecipano all'esperimento. Un esempio di questa minaccia è quello di selezionare solo i programmati in un esperimento di ispezione quando in genere alle ispezioni partecipano programmati, tester e ingegneri di sistema.

Interazione tra setting e trattamento. Questo è l'effetto di non avere il contesto sperimentale o il materiale rappresentativo, ad esempio, della pratica industriale. Un esempio è l'utilizzo di strumenti antiquati in un esperimento quando gli strumenti aggiornati sono comuni nell'industria. Un altro esempio è condurre esperimenti sui problemi dei giocattoli. Ciò significa "luogo" o ambiente sbagliato.

Interazione tra anamnesi e trattamento. Questo è l'effetto del fatto che l'esperimento viene condotto in un momento o giorno speciale che influenza i risultati. Se, ad esempio, viene condotto un questionario su sistemi critici per la sicurezza pochi giorni dopo un grave incidente correlato al software, le persone tendono a rispondere in modo diverso rispetto a pochi giorni prima o alcune settimane o mesi dopo.

Le minacce alla validità esterna vengono ridotte rendendo l'ambiente sperimentale il più realistico possibile. D'altro canto la realtà non è omogenea.

La cosa più importante è caratterizzare e riportare le caratteristiche dell'ambiente, come l'esperienza del personale, gli strumenti, i metodi al fine di valutare l'applicabilità in un contesto specifico.

8.9 Priorità tra i tipi di minacce alla validità

Esiste un conflitto tra alcuni tipi di minacce alla validità. I quattro tipi considerati sono validità interna, validità esterna, validità di conclusione e validità di costrutto. Quando si aumenta un tipo, un altro tipo potrebbe diminuire. Dare priorità tra i tipi di validità è quindi un problema di ottimizzazione, dato un certo scopo dell'esperimento.

Ad esempio, l'utilizzo di studenti universitari in un esperimento di ispezione consentirà probabilmente gruppi di studio più ampi, ridurrà l'eterogeneità all'interno del gruppo e fornirà un'implementazione affidabile del trattamento. Ciò si traduce in un'elevata validità delle conclusioni, mentre la validità esterna è ridotta, poiché la selezione non è rappresentativa se vogliamo generalizzare i risultati all'industria del software.

Un altro esempio è quello di far misurare ai soggetti diversi fattori compilando schemi per assicurarsi che i trattamenti e i risultati rappresentino realmente i costrutti oggetto di studio. Questa azione aumenterà la validità di costrutto, ma c'è il rischio che la validità della conclusione venga ridotta poiché misurazioni più noiose hanno la tendenza a ridurre l'affidabilità delle misure.

In diversi esperimenti, è possibile dare priorità a diversi tipi di validità in modo diverso, a seconda dello scopo dell'esperimento. Cook e Campbell [37] propongono le seguenti priorità per la verifica teorica e la ricerca applicata:

Test di teoria. Nella verifica della teoria, è molto importante dimostrare che esiste una relazione casuale (validità interna) e che le variabili nell'esperimento rappresentano i costrutti della teoria (validità di costrutto). L'aumento delle dimensioni dell'esperimento può generalmente risolvere i problemi di significatività statistica (validità della conclusione). Le teorie sono raramente correlate a contesti, popolazioni o periodi specifici a cui i risultati dovrebbero essere generalizzati. Quindi non c'è bisogno di problemi di validità esterna. Le priorità per gli esperimenti nella verifica teorica sono in ordine decrescente: interna, costrutto, conclusione ed esterna.

Ricerca applicata. Nella ricerca applicata, che è il campo di destinazione della maggior parte degli esperimenti di ingegneria del software, le priorità sono diverse. Ancora una volta, le relazioni studiate hanno la massima priorità (validità interna) poiché l'obiettivo chiave dell'esperimento è studiare le relazioni tra cause ed effetti. Nella ricerca applicata, la generalizzazione – dal contesto in cui viene condotto l'esperimento a un contesto più ampio – ha la massima priorità (validità esterna). Per un ricercatore non è tanto interessante mostrare un particolare risultato per l'azienda X, ma piuttosto che il risultato sia valido per aziende di una particolare dimensione o dominio di applicazione. In terzo luogo, il ricercatore applicato è relativamente meno interessato a quale dei componenti di un trattamento complesso causi realmente l'effetto (validità di costrutto). Ad esempio, in un esperimento di lettura, non è così interessante sapere se è la maggiore comprensione in generale da parte del revisore, o è la procedura di lettura specifica che aiuta i lettori a trovare più difetti. L'interesse principale è nell'effetto stesso. Infine, in contesti pratici è difficile ottenere set di dati di dimensioni sufficienti, quindi le conclusioni statistiche possono essere tratte con minore significatività (validità della conclusione).

Le priorità per gli esperimenti di ricerca applicata sono in ordine decrescente: interna, esterna, costrutto e conclusioni.

Si può concludere che è importante valutare e bilanciare le minacce alla validità dei risultati sperimentali durante la pianificazione di un esperimento. A seconda dello scopo dell'esperimento, ai diversi tipi di validità viene assegnata una priorità diversa.

I pericoli per un esperimento sono strettamente legati anche all'importanza pratica dei risultati. Potremmo, ad esempio, essere in grado di mostrare una significatività statistica, ma la differenza non ha alcuna importanza pratica. La questione è ulteriormente approfondita nella Sez. [10.3.14](#).

8.10 Esperimento di esempio

Questa descrizione è la continuazione dell'esempio introdotto nella Sez. [7.2](#). L'input alla fase di pianificazione è la definizione dell'obiettivo. Alcune delle questioni relative alla pianificazione sono state parzialmente affrontate nel modo in cui la definizione dell'obiettivo è formulata nell'esempio. Si afferma già che i soggetti saranno gli studenti e il testo indica anche che l'esperimento coinvolgerà più di un documento di requisiti.

La pianificazione è un'attività chiave quando si conduce un esperimento. Un errore nella fase di pianificazione può influenzare l'intero risultato dell'esperimento. La fase di pianificazione comprende sette attività, come mostrato in Fig. [8.1](#).

Selezione del contesto. Il tipo di contesto in molti casi è deciso, almeno in parte, dal modo in cui viene formulata la definizione dell'obiettivo. Si afferma implicitamente che l'esperimento verrà condotto off-line, anche se potenzialmente potrebbe far parte di un progetto studentesco, il che avrebbe significato on-line anche se non come parte di un progetto di sviluppo industriale.

L'esperimento verrà eseguito con una miscela di M.Sc. e dottorato di ricerca studenti.

Un esperimento offline con gli studenti implica che potrebbe essere difficile avere il tempo di esaminare un documento sui requisiti per un sistema reale a tutti gli effetti. In molti casi, esperimenti di questo tipo devono ricorrere a un documento di requisiti con caratteristiche limitate. In questo caso specifico, verranno utilizzati due documenti dei requisiti di un pacchetto di laboratorio (materiale disponibile on-line a scopo di replica). La scelta di utilizzare due documenti di requisiti ha alcune implicazioni per quanto riguarda la scelta del tipo di progetto, su cui torneremo. I documenti sui requisiti presentano alcune limitazioni per quanto riguarda le funzionalità e quindi in una certa misura devono essere considerati documenti sui requisiti "giocattolo".

L'esperimento può essere considerato generale nel senso che l'obiettivo è quello di confrontare due tecniche di lettura in generale (dal punto di vista della ricerca), e non si tratta di confrontare una tecnica di lettura esistente in un'azienda con una nuova tecnica di lettura alternativa. Quest'ultimo avrebbe reso l'esperimento specifico per la situazione dell'azienda. In entrambi i casi, ci sono alcune questioni da tenere in considerazione per garantire un confronto equo.

Nel caso della ricerca generale, è importante che il confronto sia equo, nel senso che il supporto per le due tecniche indagate sia comparabile. È di

ovviamente è facile trovare una lista di controllo molto scarsa e quindi fornire un buon supporto per PBR. Ciò favorirebbe il PBR e quindi l'esito dell'esperimento sarebbe definitivo essere sfidato. Questo è anche il motivo per cui non avere "supporto" non è un buon controllo. Un confronto/valutazione sperimentale deve basarsi sulla presenza di due comparabili metodi con supporto simile. Dovrebbe essere utilizzato "nessun supporto" come gruppo di controllo evitato. Sarebbe interessante solo se il gruppo che riceve il sostegno ottenessse risultati peggiori rispetto a coloro che non hanno alcun supporto, oppure è il "vecchio" modo di lavorare in azienda. Tuttavia, questa situazione è piuttosto rara e quindi raramente vale la pena eseguirla sperimentare in queste circostanze.

Nel caso specifico non vi è alcun problema di equità nella tipologia del sostegno a condizione che, finché una tecnica esistente viene confrontata con una nuova alternativa, allora va bene dal punto di vista del supporto. La sfida principale nel caso specifico è che i partecipanti conoscano molto bene la tecnica esistente, mentre una nuova tecnica bisogna insegnarglielo. Pertanto, la nuova tecnica potrebbe presentare uno svantaggio da allora non è così noto. D'altra parte, ha il vantaggio di potenzialmente essere più interessanti per le materie, poiché significa imparare una nuova tecnica. Pertanto, in questo caso la situazione non è così chiara, ma i potenziali pregiudizi sono presenti Il favore dell'una o dell'altra tecnica deve essere preso in considerazione dal ricercatore.

Formulazione di ipotesi. Nella definizione dell'obiettivo è espresso ciò che vorremmo per confrontare sia l'efficacia che l'efficienza quando si tratta di rilevare i guasti e quando utilizzando due diverse tecniche di lettura durante l'esecuzione dell'ispezione. Il primo il metodo è la lettura basata sulla prospettiva (PBR) e il secondo metodo è la lettura basata sulla lista di controllo (CBR). Il PBR si basa sul fatto che i revisori hanno prospettive diverse durante l'esecuzione dell'ispezione. La CBR si basa sull'avere una lista di controllo per diversi elementi che potrebbero essere correlati a difetti nei documenti relativi ai requisiti.

Il fatto che i documenti sui requisiti da utilizzare nell'esperimento siano stati utilizzati in esperimenti precedenti, significa che si presuppone che il numero di guasti sia noto, anche se non è da escludere che vengano riscontrati nuovi difetti. Va notato anche questo l'efficacia si riferisce al numero di guasti riscontrati sul numero totale di guasti, mentre l'efficienza comprende anche il tempo, cioè se vengono riscontrati più difetti per unità di tempo. Per poter formulare le ipotesi formali, lasciamo che N sia il numero di guasti e N_t il numero di guasti riscontrati per unità di tempo.

Se lasciamo:

- $E = N_{CBR}$ essere il numero di guasti rilevati utilizzando rispettivamente PBR e CBR, N_{PBR} e
- $N_{tPBR} = N_{tCBR}$ essere il numero di guasti rilevati per unità di tempo utilizzando PBR e rispettivamente CBR.

Quindi le ipotesi vengono formulate come segue:

Efficacia:

$$H_0: N_{NPBR} \leq N_{NCBR}$$

$$H_1: N_{NPBR} > N_{NCBR}$$

È opportuno notare che abbiamo scelto l'ipotesi alternativa in quanto qualsiasi differenza tra le due tecniche di lettura. In altre parole, l'ipotesi alternativa è formulata come un'ipotesi a due facce, senza alcuna assunzione che una tecnica sia migliore dell'altra.

Efficienza:

H0 W $NtPBR \leq NtCBR$

H1 W $NtPBR > NtCBR$

Le ipotesi fanno sì che si voglia dimostrare con significatività statistica che le due tecniche di lettura trovano un numero diverso di errori e che vengono trovati un numero diverso di errori per unità di tempo. Vorremmo confutare l'ipotesi nulla.

Va notato che non poter confutare l'ipotesi nulla non *implica* accettare l'ipotesi nulla. Questo tipo di risultato potrebbe essere dovuto al fatto che ci sono troppo pochi soggetti e non al fatto che le tecniche di lettura siano altrettanto efficaci nel rilevare i difetti.

Selezione delle variabili. La variabile indipendente è la tecnica di lettura e ha due livelli: rispettivamente PBR e CBR. Le variabili dipendenti sono il numero di guasti rilevati e il numero di guasti rilevati per unità di tempo. Ciò significa che dobbiamo garantire che i soggetti possano contrassegnare chiaramente i difetti riscontrati in modo che il ricercatore possa confrontare i difetti contrassegnati con l'insieme di difetti noti. Inoltre, dobbiamo garantire che i soggetti possano tenere traccia del tempo e compilare l'ora in cui è stato riscontrato un difetto specifico. Va notato che è importante tenere traccia del tempo per un guasto specifico, poiché un guasto può essere un falso positivo e quindi dobbiamo sapere anche quale tempo deve essere rimosso dal set di dati.

Selezione dei soggetti. Preferibilmente sarebbe possibile trovare i soggetti per l'esperimento in modo casuale. Tuttavia, nella maggior parte degli esperimenti il ricercatore tende ad essere costretto a utilizzare soggetti disponibili. Ciò significa che spesso gli studenti che partecipano ai corsi dell'università diventano soggetti di esperimenti condotti all'università, come nel caso di questo esperimento di esempio. In questo caso è importante che i soggetti abbiano comunque la libertà di negare la partecipazione, senza alcuna penalità per il singolo. Se la partecipazione all'esperimento dà crediti al corso, dovrebbero essere fornite opzioni alternative.

Se lo scopo dell'esperimento fosse quello di confrontare il rendimento dei due gruppi di studenti utilizzando i diversi metodi, allora il trattamento nell'esperimento sarà governato dalla selezione dei soggetti, cioè dalle caratteristiche dei gruppi di studenti. In effetti, questo lo renderebbe un quasi-esperimento. Indipendentemente, è importante caratterizzare i soggetti selezionati per aiutare a valutare la validità esterna dello studio.

Scelta del tipo di design. Una volta che sappiamo quali soggetti parteciperanno, è il momento di fare il passo successivo riguardo alla randomizzazione e decidere come dividere i soggetti in gruppi. Un buon approccio è spesso quello di utilizzare un pre-test per cercare di catturare l'esperienza dei soggetti e, in base al risultato del pre-test, dividere i soggetti in gruppi di esperienza dai quali selezioniamo casualmente i soggetti da inserire nei gruppi dell'esperimento. Questo viene fatto per cercare di garantire che i gruppi siano il più paritario possibile per quanto riguarda le esperienze precedenti

mantenendo la randomizzazione sui soggetti. Questo si chiama blocco, ovvero blocchiamo l'esperienza precedente per cercare di garantire che non influisca sull'esito dell'esperimento. Infine, l'obiettivo è nella maggior parte dei casi quello di avere gruppi ugualmente grandi, ovvero vogliamo una progettazione equilibrata. La scelta del tipo di disegno può essere influenzata dal numero di soggetti disponibili. Se si hanno molti soggetti, è possibile considerare più combinazioni sperimentalistiche o considerare di utilizzare ciascun soggetto per un solo trattamento. Con relativamente pochi soggetti, diventa più difficile progettare l'esperimento e utilizzare saggiamente i soggetti senza compromettere gli obiettivi dell'esperimento.

Il passo successivo è decidere il tipo di design. L'esperimento include un fattore di interesse primario (tecnica di lettura) con due trattamenti (rispettivamente PBR e CBR) e un secondo fattore che non è realmente di interesse per l'esperimento (documento dei requisiti). Sulla base delle decisioni prese in precedenza, il progetto naturale è un progetto completamente randomizzato in cui ciascun gruppo utilizza prima PBR o CBR su uno dei documenti dei requisiti e poi utilizza l'altra tecnica di lettura sull'altro documento dei requisiti. Tuttavia, anche le decisioni devono essere prese su ordine. Abbiamo due opzioni: (1) far sì che entrambi i gruppi utilizzino prima tecniche di lettura diverse su uno dei documenti dei requisiti e poi cambiare tecnica di lettura durante l'ispezione dell'altro documento dei requisiti, oppure (2) far sì che entrambi i gruppi utilizzino la stessa tecnica di lettura su requisiti diversi documenti. In entrambi i casi c'è un problema di ordine. Nel primo caso uno dei documenti dei requisiti verrà utilizzato prima dell'altro e nel secondo caso una tecnica di lettura verrà utilizzata prima dell'altra. Pertanto, dobbiamo considerare quale rappresenta la minima minaccia per l'esperimento.

Le minacce alla validità sono ulteriormente elaborate di seguito.

Un'altra opzione di progettazione sarebbe stata quella di consentire a un gruppo di utilizzare PBR su un documento dei requisiti e all'altro gruppo di utilizzare CBR sullo stesso documento. Il vantaggio sarebbe che nello stesso arco di tempo si potrebbe utilizzare un documento di requisiti più ampio. Lo svantaggio è che viene generata solo la metà dei punti dati. In un esperimento accade spesso che sia disponibile una certa quantità di tempo per l'esecuzione dell'esperimento. Pertanto, diventa una questione di come utilizzare il tempo nel modo più efficace, ovvero ottenere il miglior risultato possibile dall'esperimento per soddisfare le ipotesi formulate. La scelta del design è molto importante ed è sempre un compromesso. Diversi tipi di design presentano diversi vantaggi e svantaggi. Inoltre, la scelta costituisce anche la base per quale metodo statistico può essere applicato ai dati. Ciò è ulteriormente discusso nella Sez. [10.4](#).

In questo caso specifico viene scelto un disegno completamente randomizzato. A un gruppo viene innanzitutto assegnato l'utilizzo di PBR sul primo documento dei requisiti e all'altro gruppo viene assegnato l'utilizzo di CBR sullo stesso documento dei requisiti. Si sceglie questa alternativa poiché si ritiene che un ordine tra le tecniche di lettura sia peggiore di un ordine tra i documenti dei requisiti. Questo è particolarmente vero poiché l'interesse primario è nella differenza tra le tecniche di lettura e non nelle differenze tra i due documenti sui requisiti.

Strumentazione. Dato che l'esperimento si basa su un pacchetto di laboratorio, sono già disponibili i documenti dei requisiti e quindi anche l'elenco di quelli rilevati

difetti (almeno conosciuti finora). In caso contrario, dovrebbero essere individuati documenti con requisiti idonei, preferibilmente con un numero noto di difetti per poter determinare l'efficacia della tecnica di lettura.

Le linee guida per le due tecniche di lettura devono essere sviluppate o riutilizzate da altrove. In questo caso è importante garantire un confronto equo, come accennato in precedenza, fornendo un supporto comparabile per i due metodi.

I moduli per la compilazione dei difetti riscontrati devono essere sviluppati o riutilizzati da un altro esperimento. È fondamentale garantire la tracciabilità tra il documento dei requisiti e il modulo, ad esempio numerando i difetti nel documento dei requisiti e acquisendo le informazioni sul guasto nel modulo.

Valutazione di validità. Infine, devono essere valutate le minacce alla validità. È importante farlo in anticipo per garantire che le minacce siano ridotte al minimo. È quasi impossibile evitare tutte le minacce. Detto questo, ciò significa comunque che, se possibile, tutte le minacce dovrebbero essere identificate e, ove possibile, mitigate.

La valutazione delle minacce in questo specifico esempio è lasciata come esercizio; Vedere Esercizio 8.5 nella par. 8.11.

Passaggio successivo nel processo di esperimento. Sulla base dei passaggi descritti sopra per l'esempio, si spera che siamo pronti per eseguire l'esperimento. Tuttavia, prima di farlo, si consiglia ad alcuni colleghi di rivedere il disegno dell'esperimento. Inoltre, è positivo se sia possibile eseguire una prova dell'esperimento, anche se ciò significa utilizzare una o più persone che altrimenti avrebbero potuto essere soggetti all'esperimento.

Pertanto, è importante utilizzare saggiamente i potenziali soggetti.

8.11 Esercizi

8.1. Cosa sono l'ipotesi nulla e l'ipotesi alternativa?

8.2. Cos'è rispettivamente l'errore di tipo I e l'errore di tipo II, qual è il peggiore e perché?

8.3. In quali modi diversi possono essere campionati i soggetti?

8.4. Quali diversi tipi di progetti di esperimenti sono disponibili e in che modo il progetto si collega ai metodi statistici da applicare nell'analisi?

8.5. Quali sono le minacce (considerate tutti e quattro i tipi di minacce alla validità) che esistono nell'esempio della Sez. 8.10 e spiegare perché sono minacce, qual è il compromesso tra i diversi tipi di validità?

Capitolo 9

Operazione

Quando un esperimento è stato progettato e pianificato deve essere eseguito per raccogliere i dati che devono essere analizzati. Questo è ciò che intendiamo con il funzionamento di un esperimento. Nella fase operativa di un esperimento, i trattamenti vengono applicati ai soggetti. Ciò significa che questa parte dell'esperimento è la parte in cui lo sperimentatore incontra effettivamente i soggetti. Nella maggior parte degli esperimenti di ingegneria del software ci sono solo pochi altri momenti in cui i soggetti sono effettivamente coinvolti. Queste occasioni possono, ad esempio, essere un briefing prima che i soggetti si impegnino a partecipare all'esperimento e dopo l'esperimento quando i risultati dell'esperimento vengono presentati ai soggetti. Poiché gli esperimenti nell'ingegneria del software nella maggior parte dei casi riguardano gli esseri umani, sebbene sia possibile condurre esperimenti orientati alla tecnologia come discusso nella Sez. 2.4. Questo capitolo tratta in una certa misura di come motivare le persone a partecipare e prendere parte agli esperimenti.

Anche se un esperimento è stato progettato perfettamente e i dati raccolti vengono analizzati con metodi di analisi appropriati, il risultato non sarà valido se i soggetti non hanno partecipato seriamente all'esperimento. Poiché il campo della psicologia sperimentale si occupa anche di esperimenti che coinvolgono esseri umani, le linee guida per la conduzione di esperimenti in quel campo [4,29] sono in una certa misura applicabili anche all'ingegneria del software.

La fase operativa di un esperimento consiste di tre fasi: *preparazione* in cui vengono scelti i soggetti e preparati i moduli ecc., *esecuzione* in cui i soggetti svolgono i loro compiti in base ai diversi trattamenti e vengono raccolti i dati e *convalida dei dati* in cui i dati raccolti vengono convalidati. I tre passaggi sono visualizzati in Fig. 9.1 e sono ulteriormente descritti nel seguito di questo capitolo.

9.1 Preparazione

Prima che l'esperimento venga effettivamente eseguito è necessario effettuare alcuni preparativi. Migliori saranno questi preparativi, più facile sarà eseguire l'esperimento. Ci sono due aspetti importanti nella preparazione. Il primo è quello

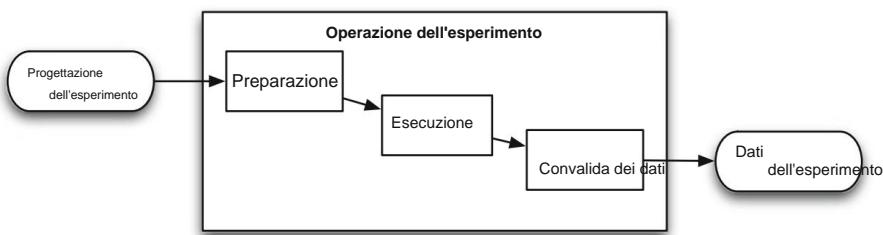


Fig. 9.1 Tre fasi del funzionamento dell'esperimento

selezionare e informare i partecipanti, e il secondo è preparare materiale come moduli e strumenti.

9.1.1 Impegnare i partecipanti

Prima di poter avviare un esperimento, è necessario trovare persone disposte a fungere da soggetti. È essenziale che le persone siano motivate e disposte a partecipare durante l'intero esperimento.

In molti casi è importante trovare persone che nell'esperimento lavorano con compiti simili ai loro compiti lavorativi ordinari. Ad esempio, se un esperimento prevede la scrittura di codice C con diversi tipi di strumenti, probabilmente avrebbe senso coinvolgere persone abituate a scrivere codice C e non coinvolgere programmatore Java. Se vengono scelte persone che non sono un insieme rappresentativo delle persone su cui vogliamo poter fare dichiarazioni, ciò costituirà una minaccia alla validità esterna dell'esperimento, vedi Cap. 8. La selezione dei soggetti, in termini di tecnica di campionamento, è discussa nella Sez. 8.4.

Quando si trovano le persone giuste ed è necessario convincere queste persone a partecipare all'esperimento. Diversi aspetti etici devono essere considerati quando le persone partecipano come soggetti.

Ottenere il consenso. I partecipanti devono concordare sugli obiettivi della ricerca. Se i partecipanti non conoscono l'intento del lavoro o il lavoro non è conforme a ciò che pensavano di dover fare quando hanno accettato di partecipare, c'è il rischio che non realizzino l'esperimento secondo gli obiettivi e le loro capacità personali. Ciò potrebbe comportare che i dati diventino non validi. È importante descrivere come verrà utilizzato e pubblicato il risultato dell'esperimento. Dovrebbe essere chiaro ai partecipanti che sono liberi di ritirarsi dall'esperimento.

A volte è necessario trovare un compromesso tra questo aspetto e il progetto rispetto alla validità. Se i partecipanti vengono influenzati dall'esperimento in quanto tali, ciò influirà sulla validità dell'esperimento.

Risultati sensibili. Se i risultati ottenuti nell'esperimento sono sensibili per i partecipanti, è importante assicurare ai partecipanti che i risultati della loro prestazione personale nell'esperimento saranno mantenuti riservati. A volte è difficile

per giudicare se il risultato è sensibile o meno, ma in generale si può dire che se il risultato avesse un significato per i partecipanti al di fuori dell'esperimento è in qualche modo sensibile. Ad esempio, se l'esperimento misura la produttività di un programmatore, il risultato indicherebbe quanto è abile il programmatore come programmatore e il risultato sarebbe sensibile. D'altra parte, se ai partecipanti venisse chiesto di utilizzare un metodo per i test di accettazione e normalmente non si occupano mai di questo tipo di test, il risultato dell'esperimento probabilmente non sarebbe così sensibile.

Incentivi. Un modo per attirare le persone verso un esperimento è offrire una sorta di incentivo. Il suo valore, tuttavia, non dovrebbe essere troppo elevato, poiché ciò potrebbe indurre le persone a partecipare solo per ricevere l'incentivo. Ciò non motiverebbe le persone a partecipare seriamente all'esperimento.

Divulgazione. Divulgazione significa rivelare tutti i dettagli dell'esperimento nel modo più aperto possibile ai soggetti dell'esperimento. Il contrario, ingannare o tradire i partecipanti, generalmente non è accettabile. Se sono disponibili metodi alternativi per condurre l'esperimento, è necessario utilizzare questi metodi. Se la non divulgazione è l'unica alternativa, dovrebbe essere applicata solo se riguarda aspetti che sono insignificanti per i partecipanti e non influenzano la loro volontà di partecipare all'esperimento. In caso di divulgazione parziale, la situazione dovrebbe essere spiegata e rivelata ai partecipanti il prima possibile.

Per una discussione più approfondita sugli aspetti etici nella sperimentazione, vedere la Sez. 2.11.

9.1.2 Problemi relativi alla strumentazione

Prima che l'esperimento possa essere eseguito, tutti gli strumenti sperimentali devono essere pronti, vedere par. 8.6. Ciò può includere gli oggetti dell'esperimento, le linee guida per l'esperimento e i moduli e gli strumenti di misurazione. Gli strumenti richiesti sono determinati dalla progettazione dell'esperimento e dal metodo che verrà utilizzato per la raccolta dei dati.

Se i soggetti stessi devono raccogliere i dati, ciò significa nella maggior parte dei casi che dei moduli devono essere distribuiti ai partecipanti. Una cosa da determinare quando vengono costruiti i moduli è se devono essere personali o se i partecipanti devono compilare in modo anonimo. Se non dovessero esserci studi aggiuntivi e quindi non vi fosse alcuna reale necessità per lo sperimentatore di distinguere tra i diversi partecipanti, potrebbe essere appropriato utilizzare forme anonime. Questo però significherà che non ci sarà la possibilità di contattare il partecipante se qualcosa viene compilato in modo poco chiaro.

In molti casi è opportuno preparare un set personale di strumenti per ogni partecipante. Questo perché molti progetti riguardano la randomizzazione e test ripetuti, in modo tale che partecipanti diversi dovrebbero essere soggetti a trattamenti diversi. Questo può essere fatto anche quando i partecipanti sono anonimi.

Se i dati devono essere raccolti durante le interviste, le domande dovrebbero essere preparate prima dell'esecuzione dell'esperimento. In questo caso potrebbe anche essere opportuno preparare domande diverse per i diversi partecipanti.

9.2 Esecuzione

L'esperimento può essere eseguito in diversi modi. Alcuni esperimenti, come semplici esperimenti di ispezione, possono essere condotti in un'occasione in cui tutti i partecipanti sono riuniti, ad esempio, in una riunione. Il vantaggio è che il risultato della raccolta dati può essere ottenuto direttamente durante l'incontro e non è necessario contattare i partecipanti e successivamente chiedere i rispettivi risultati. Un altro vantaggio è che lo sperimentatore è presente durante l'incontro e se sorgono dubbi questi possono essere risolti direttamente.

Alcuni esperimenti, tuttavia, vengono eseguiti in un arco di tempo molto più lungo ed è impossibile per lo sperimentatore partecipare a ogni dettaglio dell'esperimento e alla raccolta dei dati. Questo è, ad esempio, il caso quando l'esperimento viene eseguito in relazione a uno o più grandi progetti, in cui vengono valutati diversi metodi di sviluppo. Un esempio di tale esperimento è presentato da Ohlsson e Wohlin [128], dove è stato studiato un corso di sviluppo di software su larga scala per 2 anni. Ogni anno venivano condotti in parallelo sette progetti con un totale di circa 120 studenti. L'obiettivo dell'esperimento di Ohlsson e Wohlin [128] era quello di valutare diversi livelli di formalità nella raccolta dei dati sullo sforzo.

9.2.1 Raccolta dati

I dati possono essere raccolti manualmente dai partecipanti che compilano i moduli, manualmente supportati da strumenti, nelle interviste o automaticamente dagli strumenti.

Un vantaggio dell'utilizzo dei moduli è che non richiede molto sforzo da parte dello sperimentatore, poiché lo sperimentatore non deve prendere parte attiva alla raccolta. Uno svantaggio è che non c'è alcuna possibilità per lo sperimentatore di rivelare direttamente incoerenze, incertezze e difetti nei moduli, ecc. Questo tipo di difetti non può essere rivelato fino a dopo la raccolta dei dati o se i partecipanti sollevano l'attenzione sui difetti o hanno domande. Un vantaggio delle interviste è che lo sperimentatore ha la possibilità di comunicare meglio con i partecipanti durante la raccolta dei dati. Uno svantaggio è ovviamente che richiede uno sforzo maggiore da parte dello sperimentatore.

9.2.2 Ambiente sperimentale

Se un esperimento viene eseguito all'interno di un regolare progetto di sviluppo, l'esperimento non dovrebbe influenzare il progetto più del necessario. Questo perché lo scopo di eseguire l'esperimento all'interno del progetto è vedere gli effetti di diversi

trattamenti in un ambiente come quello oggetto del progetto. Se l'ambiente del progetto viene modificato troppo a causa dell'esperimento, l'effetto andrà perso.

Ci sono tuttavia alcuni casi in cui è opportuno con una certa interazione tra l'esperimento e il progetto. Se lo sperimentatore, ad esempio, rivela che alcune parti del progetto potrebbero essere eseguite meglio o che le stime non sono corrette, sarebbe opportuno che lo sperimentatore lo comunicasse al leader del progetto. Questo tipo di feedback diretto dall'esperimento al progetto può aiutare a motivare il personale del progetto a partecipare all'esperimento.

9.3 Convalida dei dati

Una volta raccolti i dati, lo sperimentatore deve verificare che i dati siano ragionevoli e che siano stati raccolti correttamente. Si tratta di aspetti come se i partecipanti abbiano compreso i moduli e quindi li abbiano compilati correttamente.

Un'altra fonte di errore è che alcuni partecipanti potrebbero non aver preso parte seriamente all'esperimento e pertanto alcuni dati dovrebbero essere rimossi prima dell'analisi. L'analisi dei valori anomali è ulteriormente discussa nella Sez. [10.2](#).

È importante verificare che l'esperimento sia stato effettivamente condotto nel modo previsto. È, ad esempio, importante che i soggetti abbiano applicato i trattamenti corretti nell'ordine corretto. Se si sono verificati malintesi di questo tipo, i dati ovviamente non sono validi.

Un modo per verificare che i partecipanti non abbiano frainteso le intenzioni dello sperimentatore, è quello di tenere un seminario, o in qualche altro modo presentare i risultati della raccolta dati. Ciò darà ai partecipanti la possibilità di riflettere sui risultati con cui non sono d'accordo. Aiuta anche a costruire una fiducia a lungo termine, come discusso nella Sez. [2.11](#).

9.4 Esempio di operazione

Il disegno dell'esperimento della Sez. [8.10](#) è l'input dell'operazione, che consiste in tre passaggi che devono essere affrontati.

Preparazione. Innanzitutto vanno individuati i soggetti. In questo esempio, il dottorato di ricerca. e M.Sc. gli studenti sono invitati come soggetti. Una volta ottenuto un potenziale gruppo di partecipanti, è importante convincerli a partecipare e ottenerne il loro impegno a partecipare all'esperimento. Dopo l'impegno iniziale, è necessario garantire il consenso da parte dei partecipanti. Si raccomanda di utilizzare i moduli di consenso anche se le norme formali potrebbero non richiederlo. Altre questioni da tenere in considerazione in relazione all'etica sono descritte nella Sez. [9.1.1](#). L'assegnazione dei soggetti al trattamento deve essere effettuata utilizzando una procedura di randomizzazione. Se il disegno include un fattore bloccante (tipo di studente), i soggetti dovrebbero essere divisi in base a quel fattore e poi assegnati in modo casuale a

trattamenti all'interno di ciascun gruppo di blocco. Se si sceglie un disegno equilibrato, la selezione deve terminare nello stesso numero di soggetti per ciascun gruppo.

Il prossimo passo è garantire la presenza delle infrastrutture necessarie. Ciò include, ad esempio, la prenotazione di una stanza adatta che garantisca una distanza sufficiente tra i soggetti. Copie di tutti i documenti e moduli devono essere disponibili per tutti i soggetti.

Dato che il tempo verrà raccolto, è necessario un orologio nella stanza. Non si può dare per scontato che ognuno abbia accesso al proprio orologio.

Esecuzione. Durante l'esecuzione è importante garantire che le persone siano adeguatamente distanziate nella stanza. Poiché si tratta di un esperimento di ispezione, dovrebbe essere possibile eseguire l'esperimento una volta con tutti i soggetti che eseguono l'ispezione contemporaneamente. Ciò significa anche che è facile fornire supporto per eventuali domande che potrebbero sorgere durante l'esperimento. A seconda che i dati debbano essere raccolti compilando manualmente moduli o utilizzando un computer, la preparazione deve essere effettuata di conseguenza.

Convalida dei dati. Infine, i dati devono essere convalidati. Può accadere che uno o più soggetti abbondonino l'esperimento molto presto e che i loro moduli dati debbano essere controllati attentamente per garantire che abbiano compilato i moduli in modo ragionevole.

Inoltre bisogna verificare che tutti abbiano capito come inserire i dati in modo corretto. In caso contrario, può accadere che i dati di uno o più soggetti debbano essere rimossi.

9.5 Esercizi

9.1. Quali fattori dovrebbero essere considerati nella scelta dei soggetti?

9.2. Perché le questioni etiche sono importanti nella sperimentazione?

9.3. Perché è necessario preparare attentamente la strumentazione prima di un esperimento?

9.4. Cos'è la validazione dei dati e perché dovrebbe essere effettuata prima dell'analisi statistica?

9.5. Come dovremmo gestire i soggetti che hanno un interesse personale nel risultato dell'esperimento?

Capitolo 10

Analisi e interpretazione

I dati sperimentali derivanti dall'operazione vengono immessi nell'analisi e nell'interpretazione.

Dopo aver raccolto i dati sperimentali nella fase operativa, vogliamo essere in grado di trarre conclusioni basate su questi dati. Per poter trarre conclusioni valide, dobbiamo interpretare i dati dell'esperimento.

L'interpretazione quantitativa può essere effettuata in tre fasi, come illustrato nella Fig. 10.1.

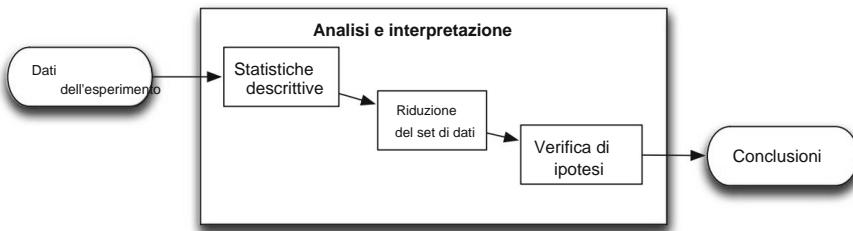
Nella prima fase, i dati vengono caratterizzati utilizzando *statistiche descrittive*, che visualizzano la tendenza centrale, la dispersione, ecc. Nella fase 2, i punti dati anomali o falsi vengono esclusi, riducendo così il set di dati a un insieme di punti dati validi. Nella terza fase, i dati vengono analizzati mediante *test di ipotesi*, in cui le ipotesi dell'esperimento vengono valutate statisticamente, ad un dato livello di significatività. Questi passaggi sono descritti più dettagliatamente nel seguito di questo capitolo.

10.1 Statistiche descrittive

La statistica descrittiva si occupa della presentazione e dell'elaborazione numerica di un set di dati.

Dopo aver raccolto i dati sperimentali, la statistica descrittiva può essere utilizzata per descrivere e presentare graficamente aspetti interessanti del set di dati. Tali aspetti includono misure che indicano, ad esempio, dove sono posizionati i dati su una certa scala e quanto è concentrato o diffuso il set di dati. L'obiettivo della statistica descrittiva è avere un'idea di come è distribuito il set di dati. La statistica descrittiva può essere utilizzata prima di effettuare test di ipotesi, al fine di comprendere meglio la natura dei dati e identificare punti di dati anomali o falsi (i cosiddetti *valori anomali*).

In questa sezione presentiamo una serie di statistiche descrittive e tecniche di grafico che possono aiutare a ottenere una visione generale di un set di dati. La scala di misurazione (vedi Cap. 3) restringe il tipo di statistiche che sono significative da calcolare. La tabella 10.1 mostra una sintesi di alcune di queste statistiche in relazione alle scale in cui sono ammissibili. Va tuttavia notato che le misure di un tipo di scala possono essere applicate alle scale più potenti, ad esempio la modalità può essere utilizzata per tutte e quattro le scale nella Tabella 10.1.

**Fig. 10.1** Tre fasi nell'interpretazione quantitativa**Tabella 10.1** Alcune statistiche rilevanti per ciascuna scala

Tipo di scala	Misura di centrale tendenza	Dispersione	Dipendenza
Nominale	Modalità	Frequenza	
Ordinale	Mediano, percentile	Intervallo di variazione	Spearman corr. coeff. Kendall corr. coeff.
Intervallo	Media, varianza, e portata	Deviazione standard	Pearson corr. coeff.
Rapporto	Media geometrica	Coefficiente di variazione	

10.1.1 Misure di Tendenza Centrale

Le misure della tendenza centrale, come media, mediana e moda, indicano una situazione "media" di un set di dati. Questo "punto medio" è spesso chiamato media e può essere interpretato come un valore medio stima dell'aspettativa della variabile stocastica da cui partono i dati nel set di dati vengono campionati.

Nel descrivere le misure di tendenza centrale, assumiamo di avere n punti dati $x_1 \dots x_n$, campionati da una variabile stocastica. La media (aritmetica), indicato \bar{x} , si calcola come:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

Il valore medio è significativo per le scale degli intervalli e dei rapporti. Noi, per esempio può calcolare la media per il set di dati .1; 1; 2; 4/ risultante in $\bar{x} = 2.0$.

La *mediana*, indicata con x_Q , rappresenta il valore medio di un set di dati, successivo a quello il numero di campioni superiori alla mediana è uguale al numero di campioni inferiori alla media. La mediana viene calcolata ordinando il campioni in ordine crescente (o discendente) e selezionando il campione centrale. Questo è ben definito se n è dispari. Se n è pari, la mediana può essere definita aritmetica media dei due valori medi. Quest'ultima operazione richiede che la scala sia almeno intervallo. Se la scala è ordinale, è possibile selezionare uno dei due valori medi scelta casuale, oppure la mediana può essere rappresentata come una coppia di valori.

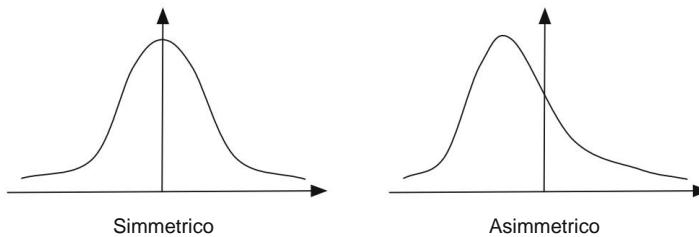


Fig. 10.2 Una distribuzione simmetrica ha gli stessi valori di media, mediana e moda, mentre possono differire se la distribuzione è asimmetrica

Il valore mediano è significativo per le scale ordinale, di intervallo e di rapporto. Ad esempio, potremmo calcolare la mediana per il set di dati .1; 1; 2; 4/ risultante in xQ D 1:5.

La mediana è un caso speciale del *percentile*, vale a dire il percentile del 50%, indicato con $x_{50\%}$, indicando che il 50% dei campioni si trova al di sotto di $x_{50\%}$. In generale x_p indica il percentile in cui $p\%$ dei campioni si trova al di sotto di questo valore. Il valore percentile è significativo per le scale ordinale, di intervallo e di rapporto.

La *modalità* rappresenta il campione più frequente. La modalità viene calcolata contando il numero di campioni per ciascun valore univoco e selezionando il valore con il conteggio più alto. La moda è ben definita se esiste un solo valore più comune di tutti gli altri. Se un numero dispari di campioni ha lo stesso conteggio di occorrenze, la modalità può essere selezionata come valore medio dei campioni più comuni. Quest'ultima operazione richiede che la scala sia almeno ordinale. Se la scala è nominale, la modalità può essere selezionata tra i campioni più comuni tramite scelta casuale o rappresentata come una coppia dei valori più comuni.

Il valore della moda è significativo per le scale nominale, ordinale, di intervallo e di rapporto. Ad esempio, possiamo calcolare la moda per il set di dati .1; 1; 2; 4/ dando una modalità di 1.

Una misura meno comune della tendenza centrale è la *media geometrica*, che è calcolato come radice n :esima del prodotto di tutti i campioni, come mostrato di seguito.

$$\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

La media geometrica è ben definita se tutti i campioni sono non negativi e significativi per la scala dei rapporti. La media (aritmetica) e la mediana sono uguali se la distribuzione dei campioni è simmetrica. Se la distribuzione è simmetrica e ha un unico massimo, tutte queste tre misure di tendenza centrale sono uguali. Tuttavia, se la distribuzione dei campioni è distorta, i valori della media, della mediana e della moda possono differire, vedere Fig. 10.2.

Se, ad esempio, la coda superiore della distribuzione è lunga, la media aumenta, mentre la mediana e la moda non vengono influenzate. Ciò indica che la media è una misura più sensibile. Tuttavia, richiede almeno una scala di intervalli e quindi potrebbe non essere sempre significativa.

10.1.2 Misure di dispersione

Le misure di tendenza centrale non forniscono informazioni sulla dispersione dei dati. Occorre quindi misurare il livello di variazione rispetto alla tendenza centrale, cioè vedere quanto sono sparsi o concentrati i dati. La *varianza* (campione), indicata con s^2 , è una misura comune di dispersione e viene calcolata come:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x})^2$$

Pertanto, la varianza è la media della distanza quadrata dalla media campionaria.

Può sembrare strano che il dividendo sia $n-1$ e non solo n , ma dividendo per $n-1$ la varianza ottiene alcune proprietà desiderabili. In particolare, la varianza campionaria è una stima imparziale e coerente della varianza della variabile stocastica. La varianza è significativa per le scale di intervallo e di rapporto.

La *deviazione standard*, indicata con s , è definita come la radice quadrata della varianza:

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

La deviazione standard è spesso preferita alla varianza poiché ha la stessa dimensione (unità di misura) dei valori dei dati stessi. La deviazione standard è significativa per le scale di intervallo e rapporto.

L' *intervallo* di un set di dati è la distanza tra il valore massimo e minimo dei dati:

$$\text{portata} = x_{\max} - x_{\min}$$

Il valore dell'intervallo è significativo per le scale di intervallo e rapporto.

L' *intervallo di variazione* è rappresentato dalla coppia $x_{\min}; x_{\max}$ inclusi il minimo e il massimo dei valori dei dati. Questa misura è significativa per le scale ordinali, di intervallo e di rapporto.

A volte la dispersione è espressa in percentuale della media. Questo valore è chiamato *coefficiente di variazione* e si calcola come:

$$\text{coefficiente di variazione} = \frac{s}{\bar{x}} \times 100$$

La misura del coefficiente di variazione non ha dimensione ed è significativa per la scala del rapporto.

Una visione generale della dispersione è data dalla *frequenza* di ciascun valore dei dati. Una tabella di frequenza viene costruita tabulando ciascun valore univoco e il conteggio delle occorrenze per ciascun valore. La *frequenza relativa* viene calcolata dividendo ciascuna frequenza per il numero totale di campioni. Per il set di dati $.1; 1; 1; 2; 2; 3; 4; 4; 4; 5; 6; 6/$ con 13 campioni possiamo costruire la tabella delle frequenze mostrata nella Tabella 10.2. La frequenza è significativa per tutte le scale.

Tabella 10.2 Una frequenza esempio di tabella

Valore	Frequenza	Frequenza relativa
1	3	23%
2	2	15%
3	1	8%
4	3	23%
5	1	8%
6	2	15%
7	1	8%

10.1.3 Misure di dipendenza

Quando il set di dati è costituito da campioni correlati in coppie (x_i, y_i) da due stocastici variabili, variabili X e Y, è spesso interessante esaminare la dipendenza tra questi X e Y.

Se X e Y sono legati tramite qualche funzione, y = f(x), vogliamo stimare questa funzione. Se sospettiamo che la funzione y = f(x) sia lineare e possa essere scritta sulla forma $y = C + bx$, potremmo applicare la regressione lineare. Regressione significa adattando i dati punta ad una curva, e nel nostro caso mostreremo come adattare una linea che riduce al minimo la somma delle distanze quadratiche da ciascun punto dati rendendolo lineare regressione. Prima di presentare le formule definiamo le seguenti abbreviazioni per alcune somme comunemente ricorrenti:

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{\sum x_i}{n}$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{\sum y_i}{n}$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum x_i}{n} \cdot \frac{\sum y_i}{n}$$

Le somme possono essere utilizzate per calcolare la retta di regressione $y = Ny - C - bx$ dove la pendenza della retta è:

$$b = \frac{S_{xy}}{S_{xx}}$$

e la linea interseca l'asse y in $C = \bar{y} - b\bar{x}$.

Se la dipendenza non è lineare, potrebbe essere possibile trovare una trasformazione di dati, in modo che la relazione diventi lineare e sia possibile utilizzare la regressione lineare. Se, per esempio, la relazione è esponenziale, $y = C e^{bx}$, questo implica che una trasformazione dei dati dà come risultato la relazione lineare $\log y = C + b \log x$.

Dopo la trasformazione logaritmica possiamo utilizzare la regressione lineare per calcolare i parametri della linea.

Per un singolo numero che quantifica quanto due set di dati, x_i e y_i , variano insieme, possiamo utilizzare la covarianza. Questa misura di dipendenza, denotata c_{xy} , è definita come:

$$c_{xy} = \frac{\sum xy}{n - 1}$$

La covarianza è significativa per le scale di intervallo e di rapporto. La covarianza dipende dalla varianza di ciascuna variabile e per poter confrontare le dipendenze tra diverse variabili correlate, la covarianza può essere normalizzata con le deviazioni standard di x_i e y_i .

Se lo facciamo otteniamo il coefficiente di correlazione r (chiamato anche coefficiente di correlazione di Pearson), che viene calcolato come:

$$r = \frac{c_{xy}}{\sqrt{s_x s_y}} = \frac{\sum xy}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Il valore r è compreso tra -1 e 1 e se non c'è correlazione r è uguale a zero. Non è però vero il contrario. I valori x_i e y_i possono essere fortemente correlati in modo non lineare anche se $r \neq 0$. Il coefficiente di correlazione (Pearson) misura solo la dipendenza lineare ed è significativo se le scale di x_i e y_i sono intervalli o rapporti e funziona bene per dati normalmente distribuiti.

Se la scala è ordinale o se i dati sono lontani dalla distribuzione normale, è possibile utilizzare il coefficiente di correlazione dell'ordine di rango di Spearman, indicato con r_s . La correlazione di Spearman viene calcolata nello stesso modo della correlazione di Pearson, tranne per il fatto che i ranghi (cioè i numeri d'ordine quando i campioni vengono ordinati) vengono utilizzati al posto dei valori dei campioni, vedere ad esempio Siegel e Castellan [157].

Un'altra misura di dipendenza è il coefficiente di correlazione dell'ordine di rango di Kendall, indicato con T . La correlazione di Kendall è adatta come misura per lo stesso tipo di dati della correlazione di Spearman, cioè almeno campioni ordinali in coppia. La correlazione di Kendall differisce, tuttavia, nella teoria sottostante poiché si concentra sul conteggio degli accordi e dei disaccordi nei ranghi tra i campioni, vedere ad esempio Siegel e Castellan [157].

Se abbiamo più di due variabili, possiamo applicare l'analisi multivariata, comprese tecniche come la regressione multipla, l'analisi delle componenti principali (PCA), l'analisi dei cluster e l'analisi discriminante. Queste tecniche sono descritte, ad esempio, da Manly [118] e Kachigan [90, 91].

10.1.4 Visualizzazione grafica

Quando si descrive un set di dati, le misure quantitative di tendenza centrale, dispersione e dipendenza dovrebbero essere combinate con tecniche di visualizzazione grafica.

I grafici sono molto illustrativi e forniscono una buona panoramica del set di dati.

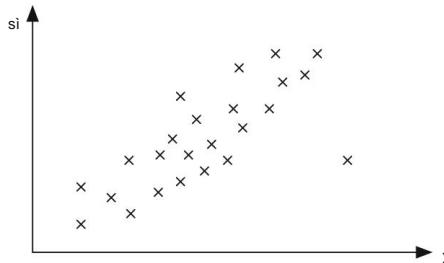


Fig. 10.3 Un grafico a dispersione

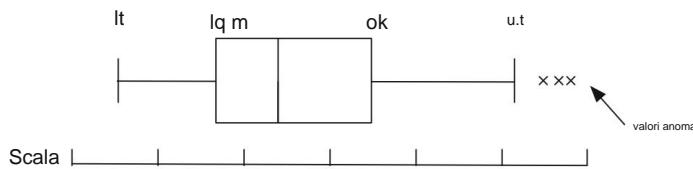


Fig. 10.4 Un box plot

Un grafico semplice ma efficace è il *grafico a dispersione*, in cui i campioni sono a coppie $(x_i ; y_i)$ sono tracciati in due dimensioni, come mostrato in Fig. 10.3.

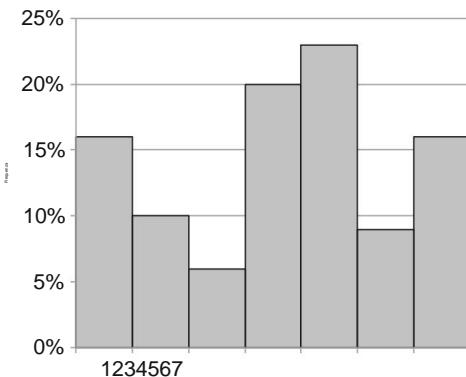
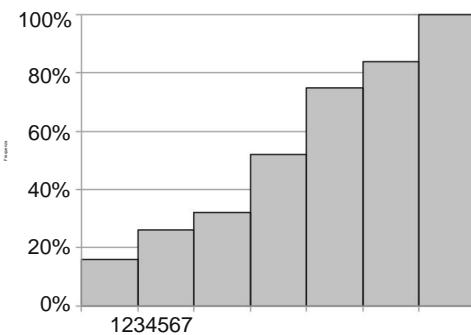
Il grafico a dispersione è utile per valutare le dipendenze tra le variabili. Esaminando il grafico a dispersione, è possibile vedere quanto sono diffusi o concentrati i punti dati e se esiste una tendenza alla relazione lineare. È possibile identificare valori atipici (outlier) e osservare la correlazione. Nella Figura 10.3, c'è una tendenza lineare con una correlazione positiva e possiamo osservare potenziali valori anomali. In questo caso particolare c'è un candidato anomalo.

Il *box plot* è utile per visualizzare la dispersione e l'asimmetria dei campioni.

Il box plot è costruito indicando graficamente diversi percentili, come mostrato in Fig. 10.4. I box plot possono essere realizzati in diversi modi. Abbiamo scelto un approccio sostanzioso, ad esempio, da Fenton e Pfleeger, e Frigge et al. [56, 60]. La differenza principale tra gli approcci è come gestire i baffi. Parte della letteratura propone che i baffi dovrebbero raggiungere rispettivamente i valori più basso e più alto, vedere ad esempio Montgomery [125]. Fenton e Pfleeger [56] propongono di utilizzare un valore, che è la lunghezza della scatola, moltiplicata per 1:5 e aggiunta o sottratta rispettivamente dai quartili superiore e inferiore.

La barra centrale nel riquadro m è la mediana. Il quartile inferiore lq è il percentile del 25% (la mediana dei valori inferiori a m) e il quartile superiore uq è il percentile del 75% (la mediana dei valori superiori a m). La lunghezza della scatola è $D = uq - lq$.

Le code del riquadro rappresentano il limite teorico entro il quale è probabile trovare tutti i punti dati se la distribuzione è normale. La coda superiore ut è $uq + 1.5D$ e la coda inferiore lt è $lq - 1.5D$ [60]. I valori di coda vengono troncati al punto dati effettivo più vicino, per evitare valori privi di significato (come righe di codice negativo).

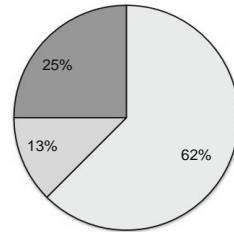
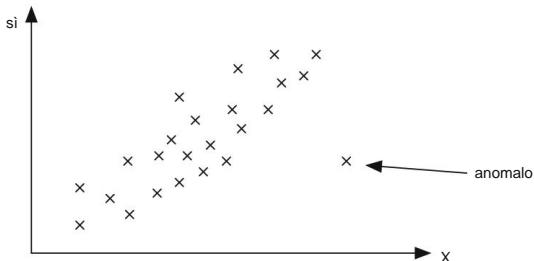
Fig. 10.5 Un istogramma**Fig. 10.6** Un istogramma cumulativo

I valori esterni alle code inferiore e superiore sono chiamati *valori anomali* e vengono visualizzati esplicitamente nel box plot. Nella Figura 10.4 ci sono tre valori anomali.

L' *istogramma* può essere utilizzato per fornire una panoramica della densità di distribuzione dei campioni di una variabile. Un istogramma è costituito da barre con altezze che rappresentano la frequenza (o la frequenza relativa) di un valore o di un intervallo di valori, come mostrato in Fig. 10.5. L'istogramma è quindi una rappresentazione grafica di una tabella di frequenza. Una distribuzione di particolare interesse è la distribuzione normale, poiché è un aspetto che dovrebbe essere preso in considerazione quando si analizzano i dati. Pertanto, un grafico potrebbe fornire una prima indicazione se i dati assomigliano o meno a una distribuzione normale. È anche possibile verificare la normalità dei dati. Ciò è ulteriormente discusso nella Sez. 10.3 quando si introduce il test Chi-2.

L' *istogramma cumulativo*, illustrato nella Figura 10.6, può essere utilizzato per fornire un quadro della funzione di distribuzione di probabilità dei campioni di una variabile. Ogni barra è la somma cumulativa delle frequenze fino alla classe di valori corrente.

Un *grafico a torta*, come illustrato in Fig. 10.7, mostra la frequenza relativa dei valori dei dati suddivisi in un numero specifico di classi distinte, costruendo segmenti in un cerchio con angoli proporzionali alla frequenza relativa.

Fig. 10.7 Un grafico a torta**Fig. 10.8** Un valore anomalo rilevato in un grafico a dispersione

10.2 Riduzione del set di dati

Nella sez. 10.3 vengono descritti numerosi metodi statistici. Tutti i metodi hanno in comune il fatto che il risultato del loro utilizzo dipende molto dalla qualità dei dati di input. Se i dati su cui vengono applicati i metodi statistici non rappresentano ciò che pensiamo rappresentino, allora le conclusioni che traiamo dai risultati dei metodi ovviamente non sono corrette.

Gli errori nel set di dati possono verificarsi come errori sistematici o come valori anomali, il che significa che il punto dati è molto più grande o molto più piccolo di quanto ci si potrebbe aspettare guardando gli altri punti dati, vedere Fig. 10.8.

Un modo efficace per identificare i valori anomali è disegnare grafici a dispersione, come mostrato nella Figura 10.8. Un altro modo è quello di disegnare box plot, come illustrato in Fig. 10.4. Sono disponibili alcuni metodi statistici per rilevare i valori anomali. Questi metodi possono, ad esempio, basarsi sul fatto che i dati provengono da una distribuzione normale e determinare la probabilità di trovare un valore come il valore più grande o più piccolo da questa distribuzione. Ciò può essere fatto, ad esempio, osservando la differenza tra possibili valori anomali e la media di tutti i valori o la differenza tra il valore anomalo e il suo valore più vicino, e quindi determinando la probabilità di trovare la differenza più grande di quella trovata. Questo studio è condotto per valutare se è possibile che l'outlier trovato possa provenire dalla distribuzione normale, anche se sembra un valore estremo.

Si noti che la riduzione dei dati come discusso qui è correlata alla convalida dei dati come discusso nel Cap. 9. La convalida dei dati si occupa dell'identificazione di dati falsi in base all'esecuzione dell'esperimento, ad esempio determinando se le persone hanno partecipato seriamente all'esperimento. Il tipo di riduzione dei dati discusso in questa sezione riguarda l'identificazione dei valori anomali non solo in base all'esecuzione dell'esperimento, ma

guardando invece i risultati dell'esecuzione sotto forma di dati raccolti e tenendo conto, ad esempio, delle statistiche descrittive.

Una volta identificati i valori anomali, è importante decidere cosa farne. Questo non dovrebbe basarsi solo sulle coordinate nel diagramma, qui è importante analizzare le ragioni degli outlier. Se il valore anomalo è dovuto ad un evento strano o raro che non si ripeterà mai più, il punto potrebbe essere escluso. Questo può accadere, ad esempio, se il punto è completamente sbagliato o franteso.

Se l'outlier è dovuto ad un evento raro che potrebbe verificarsi nuovamente, ad esempio se un modulo è stato implementato da personale inesperto, non è consigliabile escludere il valore dall'analisi, poiché nell'outlier sono presenti molte informazioni rilevanti.

Se il valore anomalo è dovuto a una variabile che non era stata considerata prima, come l'esperienza del personale, si può considerare di basare i calcoli e i modelli anche su questa variabile. È anche possibile derivare due modelli separati. Nel caso del personale con esperienza si intende un modello basato sul personale normale (con il valore anomalo rimosso) e un modello separato per il personale inesperto. Come farlo deve essere deciso caso per caso.

Non sono solo i dati non validi che possono essere rimossi dal set di dati. A volte è inefficace analizzare i dati ridondanti se la ridondanza è troppo grande. Un modo per identificare i dati ridondanti è attraverso l'analisi fattoriale e l'analisi delle componenti principali (PCA). Queste tecniche identificano fattori ortogonali che possono essere utilizzati al posto dei fattori originali. Sarebbe eccessivo descrivere questo tipo di tecniche in questo libro.

Fare riferimento invece, ad esempio, a Kachigan [90, 91] e Manly [118].

10.3 Verifica di ipotesi

10.3.1 Concetto di base

L'obiettivo del test delle ipotesi è vedere se è possibile rifiutare una certa ipotesi nulla, H_0 , sulla base di un campione proveniente da una distribuzione statistica. Cioè, l'ipotesi nulla descrive alcune proprietà della distribuzione da cui viene estratto il campione e lo sperimentatore vuole rifiutare che queste proprietà siano vere con un dato significato. L'ipotesi nulla è discussa anche nel Cap. 8. Un caso comune è che la distribuzione dipende da un singolo parametro. Impostare H_0 significa quindi formulare la distribuzione e assegnare un valore al parametro, che verrà testato.

Ad esempio, se uno sperimentatore osserva un veicolo e vuole dimostrare che il veicolo non è un'auto. Lo sperimentatore sa che tutte le automobili hanno quattro ruote, ma anche che esistono altri veicoli oltre alle automobili che hanno quattro ruote. Un esempio molto semplice di ipotesi nulla può essere formulato come " H_0 : il veicolo osservato è un'auto".

Per testare H_0 , viene definita un'unità di test, t , e viene data anche un'area critica, C , che è una parte dell'area su cui t varia. Ciò significa che il test di significatività può essere formulato come:

- Se $t \geq C$, rifiuta H_0 . • Se
... C , non rifiuta H_0

Nel nostro esempio, l'unità di prova t è il numero di ruote e l'area critica è 3 o ≥ 5 , rifiutare H_0 ,
 CD 1; 2; 3; 5; 6; ... Il test è: se t rifiuta altrimenti non
 H_0 .

Se si osserva che $t = 4$, significa che l'ipotesi non può essere scartata e non si può trarre alcuna conclusione. Questo perché potrebbero esserci altri veicoli oltre alle auto a quattro ruote.

L'ipotesi nulla dovrebbe quindi essere formulata negativamente, cioè l'intento del test è quello di rifiutare l'ipotesi. Se l'ipotesi nulla non viene rifiutata, non si può dire nulla sull'esito, mentre se l'ipotesi viene rifiutata, si può affermare che l'ipotesi è falsa con un dato significato ($\hat{\gamma}$), vedi sotto. Quando si effettua un test è in molti casi possibile calcolare la significatività più bassa possibile (spesso indicata con il valore p) con la quale è possibile rifiutare l'ipotesi nulla.

Questo valore viene spesso riportato dai pacchetti di analisi statistica.

L'area critica, C , può avere forme diverse, ma è molto comune che abbia b . Se C è costituito da uno a, b , dove $a < b$), l'intervallo è unilaterale. Se è composto da di questi intervalli, ad esempio a o due intervalli (t è bilaterale).

Tre importanti probabilità riguardanti la verifica delle ipotesi sono:

$\hat{\gamma} \text{ DP .errore-tipo-I/ DP .rifiuta } H_0 \text{ j } H_0 \text{ è vero/}$

$\hat{\gamma} \text{ DP .errore-tipo-II/ DP .non rifiutato } H_0 \text{ j } H_0 \text{ è falso/}$

Potenza D 1 $\hat{\gamma} \text{ DP .reject } H_0 \text{ j } H_0 \text{ è falso/}$

Queste probabilità sono discusse anche nel Cap. 8.

Cerchiamo qui di illustrare i concetti in un piccolo esempio che descrive un test semplice ma illustrativo chiamato *test binomiale*. Uno sperimentatore ha misurato un numero di guasti durante il funzionamento di un prodotto e li ha classificati come corruttivi (difetti che corrompono i dati del programma) e non corruttivi (difetti che non corrompono i dati del programma). La teoria dello sperimentatore è che i difetti non corruttivi sono più comuni di quelli corruttivi. Lo sperimentatore vuole quindi eseguire un test per vedere se la differenza nel numero di guasti dei diversi tipi è dovuta a coincidenze o se rivela una differenza sistematica.

L'ipotesi nulla è che non vi sia alcuna differenza nella probabilità di ricevere un guasto corruttivo e di ricevere un guasto non corruttivo. Cioè, l'ipotesi nulla può essere formulata come:

$H_0 \text{ WP .guasto corruttivo/ DP .guasto non corruttivo/ D } 1=2$

Si decide che $\hat{\gamma}$ dovrebbe essere inferiore a 0:10. Lo sperimentatore ha ricevuto i seguenti dati:

- Sono presenti 11 difetti non corrotti. • Ci sono quattro difetti che sono corruttivi.

Se l'ipotesi nulla è vera, la probabilità di ottenerne solo quattro (cioè quattro o meno) difetti corruttivi su 15

$$P .0 \text{-} 4 \text{ guasti corruttivi/ } DX_{\text{id0}} = \frac{4}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15} = \frac{4}{15} = 0.059$$

Cioè, se lo sperimentatore conclude che i dati ricevuti mostrano che gli errori non corruttivi sono più comuni di quelli corruttivi, la probabilità di commettere un errore di tipo I è 0:059. In questo caso lo sperimentatore può rifiutare H_0 perché $0:059 < 0:10$.

La probabilità di ricevere cinque o meno guasti corruttivi, se l'ipotesi nulla è vera, può essere calcolata pari a 0:1509. Questo è maggiore di 0:10, il che significa che 5 difetti corruttivi su 15 non sarebbero sufficienti per rifiutare H_0 . Lo sperimentatore può quindi decidere in modo più formale di interpretare i dati in un esperimento con 15 difetti ricevuti come:

- Se quattro o meno guasti sono dannosi, rifiutare H_0 .
- Se più di quattro guasti sono dannosi, non rifiutare H_0 .

Per riassumere, il numero di guasti corruttivi ricevuti (su 15 guasti) è pari a unità di test e l'area critica è 0, 1, 2, 3 e 4 (guasti corrutti).

Sulla base di ciò, è interessante determinare la potenza del test formulato. Poiché la potenza è la probabilità di rifiutare H_0 se H_0 non è vera, dobbiamo formulare cosa intendiamo con che H_0 non è vera. Nel nostro esempio questo può essere formulato come:

$$P.\text{guasto corruttivo/ } < P.\text{guasto non corruttivo/}$$

Poiché la somma delle due probabilità è uguale a 1, ciò può anche essere formulato come:

$$P.\text{guasto corruttivo/ } D \text{ a } < 1 = 2$$

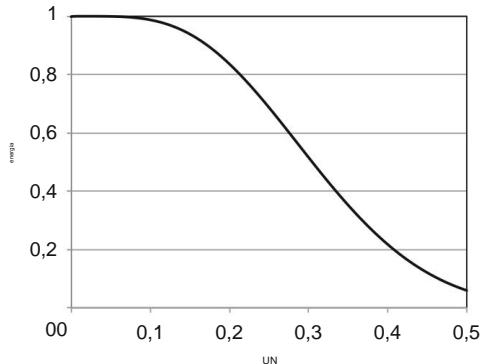
La probabilità di ricevere quattro o meno errori corruttivi su 15 errori (ovvero, la probabilità di rifiutare H_0 se H_0 è falso) è:

$$pDX_{\text{id0}} = \frac{4}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15} = \frac{4}{15} = 0.059$$

Questa probabilità è tracciata per diversi valori di a nella Figura 10.9.

Si può vedere che la potenza del test è elevata se la differenza tra le probabilità di ricevere un guasto corruttivo e un guasto non corruttivo è ampia. Se, ad esempio, un $D 0:05$ c'è una grande possibilità che ci siano quattro o meno errori corruttivi. D'altra parte, se la differenza è molto piccola, la potenza sarà minore. Se, ad esempio, un $D 0:45$ c'è una grande possibilità che ci siano più di quattro errori corruttivi.

Fig. 10.9 Potenza di un test binomiale unilaterale



Ci sono una serie di fattori che influenzano la potenza di un test. Innanzitutto, il test stesso può essere più o meno efficace. In secondo luogo, la dimensione del campione influisce sulla potenza. Una dimensione del campione maggiore significa una potenza maggiore. Un altro aspetto che incide sul potere è la scelta di un'ipotesi alternativa unilaterale o bilaterale. Un'ipotesi unilaterale dà un potere maggiore di un'ipotesi bilaterale.

Il potere nella sperimentazione dell'ingegneria del software è valutato e ulteriormente discusso da Dyba et al. [49].

10.3.2 Test parametrici e non parametrici

I test possono essere classificati in test parametrici e test non parametrici. I test parametrici si basano su un modello che coinvolge una distribuzione specifica. Nella maggior parte dei casi si presuppone che alcuni dei parametri coinvolti in un test parametrico siano distribuiti normalmente. Un test per la normalità è il test Chi-2, che verrà ulteriormente descritto di seguito quando si parleranno dei diversi tipi di test. I test parametrici richiedono inoltre che i parametri possano essere misurati almeno su una scala di intervalli. Se i parametri non possono essere misurati almeno su una scala di intervalli ciò significa generalmente che i test parametrici non possono essere utilizzati. In questi casi è disponibile un'ampia gamma di test non parametrici.

I test non parametrici non fanno lo stesso tipo di ipotesi riguardanti la distribuzione dei parametri dei test parametrici. Per derivare test non parametrici vengono fatte solo ipotesi molto generali. Il test binomiale, descritto nella sottosezione precedente è un esempio di test non parametrico. I test non parametrici sono più generali dei test parametrici. Ciò significa che i test non parametrici, se disponibili, possono essere utilizzati al posto dei test parametrici, ma i test parametrici generalmente non possono essere utilizzati quando è possibile utilizzare test non parametrici. Rispetto alla scelta del test parametrico o non parametrico ci sono due fattori da considerare:

Applicabilità Quali sono le ipotesi formulate dai diversi test? È importante che le ipotesi riguardanti la distribuzione dei parametri e le ipotesi relative alle scale sono realistiche.

Energia La potenza dei test parametrici è generalmente maggiore rispetto ai test non parametrici. Pertanto, i test parametrici richiedono meno dati punti, e quindi esperimenti più piccoli, rispetto ai test non parametrici se le ipotesi sono vere.

Anche la scelta tra metodi statistici parametrici e non parametrici lo è discusso da Briand et al. [27]. Lì viene descritto che anche se è un rischio usarlo metodi parametrici quando le condizioni richieste non sono soddisfatte, lo è in alcuni casi in cui vale la pena correre questo rischio. Le simulazioni hanno dimostrato che i metodi parametrici, come ad esempio poiché il test t, descritto di seguito, è abbastanza robusto rispetto alle deviazioni dalle precondizioni (scala dell'intervallo) purché le deviazioni non siano troppo grandi.

10.3.3 Panoramica dei test

Oltre al test binomiale sopra introdotto, vengono descritti i seguenti test in questa sezione:

prova t Uno dei test parametrici più utilizzati. Il test è abituato confrontare due medie campionarie. Cioè, il design è un fattore importante due trattamenti.

Mann-Whitney Questa è un'alternativa non parametrica al test t.

Test F Questo è un test parametrico che può essere utilizzato per confrontare due campioni distribuzioni.

Test t appaiato Un test t per un progetto di confronto appaiato.

Wilcoxon Questa è un'alternativa non parametrica al t-test appaiato.

Prova di segno Questa è un'alternativa non parametrica al t-test accoppiato. Il segno Il test è un'alternativa più semplice al test di Wilcoxon.

ANOVA (Analisi della varianza). Una famiglia di test parametrici che può essere utilizzato per progetti con più di due livelli di fattore. ANOVA i test possono, ad esempio, essere utilizzati nei seguenti progetti: One fattore con più di due livelli, un fattore e una variabile di blocco, disegno fattoriale e disegno annidato.

Kruskal-Wallis Questa è un'alternativa non parametrica all'ANOVA nel caso di uno fattore con più di due trattamenti.

Chi-2 Questa è una famiglia di test non parametrici che possono essere utilizzati quando i dati sono sotto forma di frequenze.

Le diverse prove possono essere ordinate rispetto al tipo di progetto e rispetto a se sono parametrici o non parametrici come nella Tabella 10.3.

Per tutti i test descritti, i seguenti elementi sono presentati separatamente tabelle per ogni prova:

Tabella 10.3 Panoramica dei test parametrici/non parametrici per diversi progetti

Design	Parametrico	Non parametrico
Un fattore, un trattamento Un		Chi-2, test binomiale
fattore, due trattamenti, completamente disegno randomizzato	test t, test F	Mann-Whitney, Chi-2
Un fattore, due trattamenti, confronto appaiato	Test t accoppiato	Wilcoxon, Prova dei segni
più di due trattamenti Più di un fattore	ANOVA ANOVAa	Kruskal-Wallis, Chi-2

^{UN} Questo test non è descritto in questo libro. Si veda invece, ad esempio, Marascuilo e Serlin [119] e Montgomery [125]

Ingresso Il tipo di misurazioni necessarie per rendere applicabile il test descrive l'input del test. Cioè, questo descrive cosa requisiti ci sono sulla progettazione dell'esperimento se il test dovesse essere applicabile.

Ipotesi nulla Viene fornita una formulazione dell'ipotesi nulla.

Calcoli Describe cosa calcolare in base ai dati misurati.

Criterio Il criterio per rifiutare l'ipotesi nulla. Ciò spesso comporta utilizzando tabelle statistiche e viene descritto da quale tabella utilizzare Appendice B. In questo libro le tabelle sono fornite solo per un livello di significato, ma per molti test si fa riferimento ad altri fonti in cui sono fornite tabelle più complete.

Non tutti i test sono descritti completamente qui. Per ulteriori informazioni riguardanti le prove fanno riferimento alla bibliografia riportata nel testo. Ad esempio, il Mann-Whitney vengono descritti il test di Wilcoxon, il test dei segni e il test di Kruskal-Wallis il caso più semplice con pochi campioni. Se ci sono molti campioni (per esempio, più di 35 circa per il test dei segni) in molti casi è difficile da eseguire i calcoli e le decisioni come descritto di seguito. In questi casi è possibile fai alcune approssimazioni perché ci sono così tanti campioni. Come farlo è descritto, ad esempio, da Siegel e Castellan [157]. Hanno anche descritto come farlo fare quando si verificano parità (due o più valori uguali) per quei test.

L'obiettivo delle descrizioni dei test è che siano possibili da utilizzare i test basati sulle descrizioni e sugli esempi. L'intenzione non è quella di fornire tutti i dettagli dietro le derivazioni delle formule.

Utilizzando il tipo di descrizione sopra delineato, il nostro semplice test di esempio, vedi Setta. 10.3.1, è riassunto nella Tabella 10.4.

Nella tabella sopra viene descritto il test binomiale per l'ipotesi nulla che il due eventi sono ugualmente probabili. È possibile formulare altre ipotesi nulle, come ad es P.evento 1/ D 0:3 e P.evento 2/ D 0:7. Per una descrizione di come eseguire il file test in questi casi, vedere ad esempio Siegel e Castellan [157].

Per la maggior parte dei test contenuti in questo capitolo vengono presentati esempi di utilizzo. IL gli esempi sono basati su dati fintizi. Inoltre, i test vengono principalmente presentati per un livello di significatività del 5% per il quale le tabelle sono fornite nell'Appendice B.

Tabella 10.4 Test binomiale

Articolo	Descrizione
Ingresso	Numero di eventi conteggiati per due diversi tipi di eventi (evento1 ed evento2)
H0	P.evento 1 / \bar{P} .evento 2 /
Calcoli	Calcola pD $\frac{1}{2N} \sum_{i=0}^N i! N!$ dove N è il numero totale di eventi e n è il numero dell'evento più raro
Criterio	Due lati (H1 $W P$.evento 1 / $\neq P$.evento 2/): rifiuta H0 se $p < \bar{y}=2$ Unilaterale (H1 $W P$.evento 1 / $< P$.evento 2/): rifiuta H0 se $p < \bar{y}$ e l'evento 1 è il evento più raro nel campione

Tabella 10.5 Prova t

Articolo	Descrizione
Ingresso	Due campioni indipendenti: $x_1; x_2; \dots; x_n$ e $y_1; y_2; \dots; y_m$
H0	$x = \bar{x}$ sì, cioè i valori medi attesi sono gli stessi
Calcoli	Calcola $t_0 D$ $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$ dove $s_p D$ s.n. 1/S2 $\sim E$, s_x^2 e s_y^2 si sono le varianze del singolo campione
Criterio	Due lati ($H1 W x \neq y$): rifiutare H0 se $ t_0 > \bar{t}_{\alpha/2} nCm2$. Qui, $\bar{t}_{\alpha/2}$ è il superiore \bar{y} punto percentuale della distribuzione t con f gradi di libertà, ovvero pari a $n + m - 2$. La distribuzione è tabellata, ad esempio, nella Tabella B.1 e di Montgomery [125] e Marascuilo e Serlin [119] Unilaterale ($H1 W x > y$): rifiutare H0 se $t_0 > \bar{t}_{\alpha} nCm2$

Tabelle più complete sono disponibili nei libri di statistica, ad esempio, di Marascuilo e Serlin [119] e Montgomery [125].

10.3.4 Prova t

Il t-test è un test parametrico utilizzato per confrontare due campioni indipendenti. Quello cioè, il design dovrebbe essere un fattore con due livelli. È possibile eseguire il test t basato su una serie di presupposti diversi, ma qui esiste un'alternativa spesso utilizzata descritto. Per maggiori informazioni si rimanda ad esempio a Montgomery [125], Siegel e Castellan [157], e Marascuilo e Serlin [119]. Il test viene eseguito come presentato nella Tabella 10.5.

Esempio di test t. Sono state confrontate le densità dei difetti in diversi programmi in due progetti. In uno dei progetti il risultato è

x_D 3:42; 2:71; 2:84; 1:85; 3:22; 3:48; 2:68; 4:30; 2:49; 1:54

Tabella 10.6 Mann-Whitney

Articolo	Descrizione
Ingresso	Due campioni indipendenti: $x_1; x_2; \dots; x_n$ e $y_1; y_2; \dots; y_m$ I due campioni provengono dalla stessa distribuzione NA.NAC1/ Calcoli Classifica tutti i campioni e calcola U_D NANB C 2
H_0	$T_e U_0 D$ NANB
Criterio	Le tabelle che forniscono i criteri per il rifiuto dell'ipotesi nulla sulla base dei calcoli sono fornite, ad esempio, nella Tabella B.3 e da Marascuilo e Serlin [119]

Rifiuta H_0 se $\min(U_e; U_0)$ è inferiore o uguale al valore nella Tabella B.3

e nell'altro progetto il risultato è

y_D 3:44; 4:97; 4:76; 4:96; 4:10; 3:05; 4:09; 3:69; 4:21; 4:40; 3:49

L'ipotesi nulla è che la densità dei difetti sia la stessa in entrambi i progetti, mentre l'ipotesi alternativa che non lo sia. Dai dati risulta che $n_D = 10$ e $m_D = 11$. I valori medi sono $x_N D = 2:853$ e $y_N D = 4:1055$.

Si può trovare che $S_2 D = 0:6506$; $S_2 D_0 = 4:112$; $Sp D = 0:7243$ e $t_0 D = 3:96$.

Il numero di gradi di libertà è $F D = n - 1$ $m - 1 = 10 - 1 = 9$. Nella Tabella B.1 si può vedere che $t_0:025; 19 D = 2:093$. Poiché $|t_0| > t_0:025; 19$ è possibile rifiutare l'ipotesi nulla con un test a due code al livello 0:05.

10.3.5 Mann-Whitney

Il test di Mann-Whitney è un'alternativa non parametrica al test t. È sempre possibile utilizzare questo test al posto del t-test se le ipotesi formulate dal t-test sembrano incerte. Il test, che si basa sui ranghi, non è descritto completamente qui. Maggiori dettagli sono presentati, ad esempio, da Siegel e Castellan [157] 1, e da Marascuilo e Serlin [119]. Il test è riassunto nella Tabella 10.6.

Esempio di Mann-Whitney. Quando vengono utilizzati gli stessi dati, come nell'esempio con il test t, si può vedere che $NA = \min(10, 11) = 10$ e $NB = \max(10, 11) = 11$. I ranghi del campione più piccolo (x) sono 9, 5, 6, 2, 8, 11, 4, 17, 3, 1 e i ranghi del campione più grande (y) sono 10, 21, 19, 20, 15, 7, 14, 13, 16, 18, 12. In base ai ranghi si può riscontrare che $TD = 66$; $UD = 99$ e $U_0 D = 11$. Poiché il più piccolo tra U_e e U_0 è inferiore a 26, vedere la Tabella B.3, è possibile rifiutare l'ipotesi nulla con un test a due code al livello 0:05.

¹Siegel e Castellan [157] descrivono il test di Wilcoxon-Mann-Whitney invece del test di Mann-Whitney. I due test sono, tuttavia, sostanzialmente uguali.

Tabella 10.7 Prova F

Articolo	Descrizione
Ingresso	Due campioni indipendenti: $x_1; x_2; \dots; x_n$ e $y_1; y_2; \dots; y_m$
H_0	$\frac{s_x^2}{s_y^2} = 1$, cioè le varianze sono uguali
Calcoli	Calcola F_0 $D = \frac{\max(S_{2x}; S_{2y}) / s_x^2}{\min(S_{2x}; S_{2y}) / s_y^2}$, dove S_{2x} e S_{2y} sono il campione individuale varianze
Criterio	Due lati (H_1 W \neq): rifiuta H_0 se $F_0 > F_{\bar{Y}=2; n_{\text{max}}; n_{\text{min}}}$, dove n_{max} è il numero di punteggi nel campione con la massima varianza campionaria e n_{min} è il numero di punteggi nel campione con la varianza campionaria minima. $F_{\bar{Y}=2; f_1; f_2}$ è il \bar{Y} punto percentuale superiore della distribuzione F con f_1 e f_2 gradi di libertà, che è tabulato, ad esempio, nella Tabella B.5 e da Montgomery [125], e Marascuilo e Serlin [119] Monofacciale (H_1 $L = \frac{s_x^2}{s_y^2} > \frac{s_y^2}{s_x^2}$): rifiutare H_0 se $F_0 > F_{\bar{Y}; n_{\text{max}}; n_{\text{min}}}$, e $S_{2x} > S_{2y}$

10.3.6 Prova F

Il test F è un test parametrico che può essere utilizzato per confrontare la varianza di due campioni indipendenti. Maggiori dettagli sul test sono presentati, ad esempio, Montgomery [125], Robson [144] e Marascuilo e Serlin [119]. La prova è eseguita come presentato nella Tabella 10.7.

Esempio di prova F. Quando vengono utilizzati gli stessi dati, come nell'esempio con il t-test, si trova che $S_x D 0:6506$ e $S_y D 0:4112$, il che significa che $F_0 D 1:58$. Si può anche vedere che $n_{\text{max}} D 10$ e $n_{\text{min}} D 11$.

Dalla Tabella B.5 si vede che $F_{0:025; 9; 10} D 3:78$. Poiché $F_0 < F_{0:025; 9; 10}$ lo è impossibile rifiutare l'ipotesi nulla con un test a due code al livello 0:05. Quello cioè, il test non esclude che i due campioni abbiano la stessa varianza.

10.3.7 Test t accoppiato

Il t-test per dati appaiati viene utilizzato quando due campioni risultanti da misure ripetute sono rispetto. Ciò significa che le misurazioni vengono effettuate rispetto, ad esempio, un argomento più di una volta. Un esempio di ciò è se vengono confrontati due strumenti. Se due gruppi utilizzassero indipendentemente i due diversi strumenti, il risultato sarebbe due potrebbero essere applicati campioni indipendenti e il test t ordinario. Se invece solo uno sarebbe stato utilizzato il gruppo e ogni persona avrebbe utilizzato entrambi gli strumenti, avremmo ripetuto misure. In questo caso il test esamina la differenza di prestazione per ogni persona con i diversi strumenti.

Il test, descritto più dettagliatamente, ad esempio, da Montgomery [125], e Marascuilo e Serlin [119], viene eseguito come presentato nella Tabella 10.8:

Esempio di test t accoppiato. Dieci programmati ne hanno sviluppati indipendentemente due programmi diversi. Hanno misurato lo sforzo richiesto e il risultato viene visualizzato nella Tabella 10.9.

Tabella 10.8 Test t accoppiato

Articolo	Descrizione
Ingresso	Campioni accoppiati: .x1; y1/; .x2; y2/ : : : .xn; si/
H0	D D 0, dove di D xi yi , ovvero la media attesa delle differenze è 0
Calcoli	Calcola t0 D Sd $= \frac{dN}{sPn}$, Due dove Sd D sPn iD1.di d MN 2 1
Criterio	lati .H1 W d ≠ 0/: rifiuta H0 se jt0j > t̄y=2;n1. Qui, t̄y;f è il y punto percentuale superiore della distribuzione t con f gradi di libertà. IL la distribuzione è tabulata, ad esempio, nella Tabella B.1 e da Montgomery [125], e Marascuilo e Serlin [119] Unilaterale .H1 W d > 0/: rifiuta H0 se jt0j > t̄y;n1

Tabella 10.9 Sforzo richiesto

Programmatore	1	2	3	4	5	6	7	8	9	10
Programma	105	137	124	111	151	150	168	159	104	102
1 Programma 2	86.1	115	175	94,9	174	120	153	17871.3	110	

L'ipotesi nulla è che lo sforzo richiesto per sviluppare il programma 1 sia lo stesso come sforzo richiesto per sviluppare il programma 2. L'ipotesi alternativa è che lo sia non. Per effettuare la prova vengono calcolati:

d RE f18:9; 22; 51; 16:1; 23; 30; 15; 19; 32:7; 9 g

SD D 27:358

t0D 00:39

Il numero di gradi di libertà è f D n 1 D 10 1 D 9. Nella Tabella B.1 ,
si può vedere che t0:025;9 D 2:262.

Poiché t0 < t0:025;9 è impossibile rifiutare l'ipotesi nulla con due code test al livello 0:05.

10.3.8 Wilcoxon

Il test di Wilcoxon è un'alternativa non parametrica al test t per dati appaiati. L'unica i requisiti sono che sia possibile determinare quale delle misure in una coppia è il più grande e che è possibile classificare le differenze. Il test, su cui si basa sui ranghi, non è descritto in dettaglio qui. Viene presentata una descrizione più dettagliata da, ad esempio, Siegel e Castellan [157], e Marascuilo e Serlin [119]. IL
Il test è riassunto nella Tabella 10.10.

Esempio di Wilcoxon. Quando vengono utilizzati gli stessi dati, come nell'esempio con il file t-test appaiato, si trova che i ranghi dei valori assoluti della differenza (d) sono 4, 6, 10, 3, 7, 8, 2, 5, 9, 1. Sulla base di questo T C e T possono essere calcolati come 32 e 23.

Poiché il più piccolo tra T C e T è maggiore di 8 (vedi Tabella B.4) ciò è impossibile rifiutare l'ipotesi nulla con un test a due code al livello 0,05.

Tabella 10.10 Wilcoxon

Articolo	Descrizione
Ingresso	Campioni accoppiati: .x1; y1/; .x2; y2/ : : : .xn; si/
H0	Se tutte le differenze (di D xi yi) sono classificate .1; 2; 3 : : / senza considerare il segno, allora la somma dei ranghi delle differenze positive è uguale alla somma di gradi delle differenze negative
Calcoli	Calcolare T C come la somma dei ranghi dei positivi di : s e T come la somma di i ranghi del negativo di : s
Criterio	Tabelle che possono essere utilizzate per determinare se H0 può essere rifiutato in base a TC , T e il numero di coppie, n, è disponibile. Vedi ad esempio la Tabella B.4 o Siegel e Castellan [157], e Marascuilo e Serlin [119]
	Rifiutare H0 se min(TC ; T) è inferiore o uguale al valore nella Tabella B.4

Tabella 10.11 Test dei segni

Articolo	Descrizione
Ingresso	Campioni accoppiati: .x1; y1/; .x2; y2/ : : : .xN ; siN /
H0	P./ C/ D P ./, dove C e rappresentano i due eventi che xi > yi e xi < yi
Calcoli	Rappresenta ogni differenza positiva .di D xi yi/ con una C e ogni differenza negativa differenza di a . Calcola pD $\frac{1}{2N} \sum_{i=0}^N X_n$ dove N è il numero totale di segni, e n è il numero di segni dei segni più rari
Criterio	Due lati .H1 W P .C/ \neq P .//: rifiuta H0 se p < $\ddot{\gamma}$ =2 Unilaterale .H1 W P .C/ < P .//: rifiuta H0 se p< $\ddot{\gamma}$ e l' evento C è il più evento raro nel campione

10.3.9 Prova dei segni

Il test dei segni è, come il test di Wilcoxon, un'alternativa non parametrica al test accoppiato prova t. Il test dei segni può essere utilizzato al posto del test di Wilcoxon quando non è possibile o necessario classificare le differenze, poiché si basa solo sul segno della differenza dei valori in ciascuna coppia. Ad esempio, non è necessario utilizzare Wilcoxon quando lo fa è possibile dimostrare la significatività con il test dei segni. Ciò è dovuto al test del segno ha una potenza inferiore. Inoltre, il test dei segni è più semplice da eseguire.

Il test è ulteriormente descritto, ad esempio, da Siegel e Castellan [157] e Robson [144], ed è riassunto nella Tabella 10.11.

Il lettore può riconoscere che questo test è un test binomiale in cui i due eventi sono C e rispettivamente.

Esempio di prova dei segni. Quando vengono utilizzati gli stessi dati, come nell'esempio con il file accoppiato t-test, si scopre che ci sono 6 differenze positive e 4 negative. Questo significa questo

$$pD = \frac{1}{210} \sum_{i=0}^{10} X_n = \frac{193}{512} = 0.3770$$

Poiché p > 0.025 è impossibile rifiutare l'ipotesi nulla con una doppia coda test al livello 0,05.

Tabella 10.12 ANOVA, un fattore, più di due trattamenti

Articolo	Descrizione
Ingresso	un campione: $x_{11}; x_{12}; \dots; x_{1n_1} x_{21}; x_{22}; \dots; x_{2n_2} \dots x_{a1}; x_{a2}; \dots; x_{an_a}$
H_0	$x_{11} = x_{12} = \dots = x_{1n_1} = x_{21} = x_{22} = \dots = x_{2n_2} = \dots = x_{a1} = x_{a2} = \dots = x_{an_a}$, cioè tutte le medie attese sono uguali
Calcoli	$\begin{aligned} SST &= \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{ij})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_{ij})^2}{N} \end{aligned}$ $\begin{aligned} STrattamento &= \sum_{i=1}^a \frac{\bar{x}_{i\cdot}^2 - \bar{x}_{\cdot\cdot}^2}{n_i} \\ &= \sum_{i=1}^a \frac{\bar{x}_{i\cdot}^2 - \bar{x}_{\cdot\cdot}^2}{N} \end{aligned}$ $SSError = SST - STrattamento$ $MSTreatment = SST / (a - 1)$ $MSError = SSError / (N - a)$ $F_0 = MSTreatment / MSError$ <p>dove N è il numero totale di misurazioni e un punto indice denota a sommatoria sull'indice tratteggiato, ad esempio \bar{x}_{ij}</p>
Criterio	Rifiutare H_0 se $F_0 > F_{\alpha/2; a-1, N-a}$. Qui $F_{\alpha/2; f_1, f_2}$ è il α punto percentuale superiore della distribuzione F con gradi di libertà f_1 e f_2 , che è tabulata, per esempio, nella Tabella B.5 e da Montgomery [125] e Marascuilo e Serlin [119]

Tabella 10.13 Tabella ANOVA per il test ANOVA descritto sopra

Fonte di variazione	Somma di piazze	Gradi di libertà	Quadrato medio	F0	
Tra i trattamenti	Trattamento	un 1	MSTreatment	F0	$\frac{MSTreatment}{MSError}$
Errore	SSError	N / a	MSError		
Totalle	SST	N1			

^{un} Questo è talvolta indicato all'interno dei trattamenti

10.3.10 ANOVA (ANalisi della varianza)

L'analisi della varianza può essere utilizzata per analizzare esperimenti da un numero diverso disegni. Il nome, analisi della varianza, viene utilizzato perché il metodo è basato su guardando la variabilità totale dei dati e la partizione della variabilità secondo componenti diversi. Nella sua forma più semplice il test confronta la variabilità dovuta a trattamento e la variabilità dovuta all'errore casuale.

In questa sezione viene descritto come utilizzare ANOVA nella sua forma più semplice. La prova può essere utilizzata per confrontare se un numero di campioni ha lo stesso valore medio. Questo è, il design è un fattore con più di due trattamenti. La prova è riassunta in Tabella 10.12.

I risultati di un test ANOVA sono spesso riepilogati in una tabella ANOVA. IL i risultati di un test per un fattore con più livelli possono, ad esempio, essere riepilogati come nella Tabella 10.13.

Si noti che il test ANOVA descritto è solo una variante dei test ANOVA. I test ANOVA possono essere utilizzati per numerosi progetti diversi, che ne coinvolgono molti

Tabella 10.14 Tabella ANOVA

Fonte di variazione	Somma dei quadrati	Gradi di libertà	Quadrato medio	F0
Tra i trattamenti	579.0515	2	289.5258	0,24
Errore	36.151	30	1.205	
Totale	36.730	32		

Tabella 10.15 Kruskal-Wallis

Articolo	Descrizione
Ingresso	un campione: $x_{11}, x_{12}, \dots, x_{1n_1} x_{21}, x_{22}, \dots, x_{2n_2} \dots x_{a1}, x_{a2}, \dots, x_{an_a}$
H0	Le mediane della popolazione dei campioni a sono uguali.
Calcoli	Tutte le misure sono classificate in una serie $.1; 2; \dots; n_1 C \ n_2 C \ \dots \ C \ n_a$, e il calcolo si basano su questi ranghi. Si veda ad esempio [119, 157].
Criterio	Si vedano, ad esempio, Siegel e Castellan [157] e Marascuilo e Serlin [119]

diversi fattori, variabili bloccanti, ecc. Sarebbe troppo lungo descrivere questi test in dettaglio qui. Fare riferimento invece, ad esempio, a Montgomery [125] e Marascuilo e Serling [119].

Esempio di ANOVA. Sono state le dimensioni dei moduli in tre diversi programmi misurato. Il risultato è:

Programma 1: 221, 159, 191, 194, 156, 238, 220, 197, 197, 194
 Programma 2: 173, 171, 168, 286, 206, 140, 226, 248, 189, 208, 213
 Programma 3: 234, 188, 181, 207, 266, 153, 190, 195, 181, 238, 191, 260

L'ipotesi nulla è che la dimensione media del modulo sia la stessa in tutti e tre programmi. L'ipotesi alternativa è che non lo sia. Sulla base dei dati sopra riportati È possibile calcolare la tabella ANOVA nella Tabella 10.14 .

Il numero di gradi di libertà sono $f_1 = 2$ a $f_2 = 30$. Nella Tabella B.5, si può vedere che $F_0 = 0,025; 2; 30$ è 4,18. Poiché $F_0 < F_0 = 0,025; 2; 30$ è impossibile rifiutare l'ipotesi nulla al livello 0,025.

10.3.11 Kruskal-Wallis

L'analisi unidirezionale della varianza per ranghi di Kruskal-Wallis non è parametrica alternativa all'analisi parametrica della varianza monofattoriale sopra descritta. Questo test può sempre essere utilizzato al posto dell'ANOVA parametrica se non è sicuro che il file le ipotesi di ANOVA sono soddisfatte. Il test, che si basa sui ranghi, non è descritto in dettaglio qui. Maggiori dettagli sono presentati, ad esempio, da Siegel e Castellan [157] e Marascuilo e Serlin [119].

Il test è riassunto nella Tabella 10.15.

Tabella 10.16 Tabella delle frequenze per dimensioni del modulo (variabili) di due sistemi (gruppi)

Dimensioni del modulo	Sistema 1	Sistema 2
piccolo	15	10
medio	20	19
grande	25	28

10.3.12 Chi-2

I test Chi-2 (a volte indicato con 2) possono essere eseguiti in diversi modi.

Tutti i test Chi-2, tuttavia, si basano su dati sotto forma di frequenze. UN esempio di frequenze per due sistemi con più moduli può essere quello per sistema 1 ci sono 15 moduli piccoli, 20 moduli medi e 25 moduli grandi, mentre per il sistema 2 sono previsti 10 moduli piccoli, 19 moduli medi e 28 grandi moduli. Ciò è riassunto nella Tabella 10.16.

In questo caso potrebbe essere eseguito un test Chi-2 per indagare se la distribuzione di i moduli piccoli, medi e grandi sono gli stessi per i due sistemi.

I test Chi-2 possono anche essere eseguiti con un gruppo di dati, per vedere se ne esiste uno la distribuzione della frequenza misurata è la stessa di una distribuzione teorica. Questa prova può, ad esempio, essere eseguito per verificare se i campioni possono essere visti normalmente distribuito.

Nella Tabella 10.17 è riassunto un test Chi-2, che può essere utilizzato per confrontare se le misurazioni di due o più gruppi provengono dalla stessa distribuzione.

Esempio di test Chi-2. Se viene eseguito un test Chi-2 sui dati nella Tabella 10.16 allora È possibile costruire la Tabella 10.18 .

L'ipotesi nulla è che la distribuzione dimensionale sia la stessa in entrambi i sistemi, e l'ipotesi alternativa è che le distribuzioni siano diverse. Sulla base dei dati la statistica del test può essere calcolata su X^2 D 1:12. Il numero di gradi di libertà è .r 1/k 1/ D 2 1 D 2. Nella Tabella B.2 si può vedere che D 5:99. Da $X^2 < \frac{2}{0:05;2}$ è impossibile rifiutare l'ipotesi nulla al livello 0,05.

Chi-2 Bontà del test di adattamento. Per verificare è possibile anche eseguire un test Chi-2 se le misurazioni vengono prese da una certa distribuzione, ad esempio, la distribuzione normale. In questo caso viene eseguito un test di bontà di adattamento secondo la Tabella 10.19

Se la bontà del fit test viene eseguita per una distribuzione continua, il possibile i valori che possono essere misurati devono essere divisi in intervalli in modo che ogni intervallo possa rappresentare un valore. Ciò deve essere fatto, ad esempio, per la distribuzione normale.

Se la distribuzione di H_0 è completamente specificata (ad esempio, $P .XD 1 / D 2=3; P .XD 2 / D 1=3$) allora nessun parametro deve essere stimato dalla misura dati (cioè è D 0). D'altro canto, ad esempio, se l'ipotesi nulla specifica soltanto che i valori rispettino una distribuzione normale, è necessario stimare due parametri. Sia il valore medio che la deviazione standard della distribuzione normale devono essere stimati, altrimenti non è possibile determinare i valori dei diversi valori attesi, E_i , per gli intervalli. Pertanto, in questo caso, è D 2.

Tabella 10.17 Campioni indipendenti Chi-2,k (gruppi)

Articolo	Descrizione																			
Ingresso	Dati come frequenze per k gruppi																			
H0	Le misurazioni dei gruppi k provengono dalla stessa distribuzione																			
Calcoli	Crea una tabella di contingenza. Un esempio di tabella di contingenza per due gruppi e tre variabili (vale a dire, le stesse dimensioni dei dati nella Tabella 10.16). costruito come:																			
	<table border="1"> <thead> <tr> <th>Variabile Gruppo1</th> <th>Gruppo2</th> <th>Combinata</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>n11</td> <td>n12</td> <td>R1</td> </tr> <tr> <td>2</td> <td>n21</td> <td>n22</td> <td>R2</td> </tr> <tr> <td>3</td> <td>n31</td> <td>n32</td> <td>R3</td> </tr> <tr> <td>Totale</td> <td>C1</td> <td>C2N</td> <td></td> </tr> </tbody> </table>	Variabile Gruppo1	Gruppo2	Combinata	1	n11	n12	R1	2	n21	n22	R2	3	n31	n32	R3	Totale	C1	C2N	
Variabile Gruppo1	Gruppo2	Combinata																		
1	n11	n12	R1																	
2	n21	n22	R2																	
3	n31	n32	R3																	
Totale	C1	C2N																		
	In questa tabella n_{ij} denota la frequenza per la variabile i e il gruppo j , Ci denota the la somma delle frequenze per il gruppo i e Ri denota la somma per la variabile i. N è la somma di tutte le frequenze																			
	Calcola $X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$ dove $E_{ij} = \frac{R_i C_j}{N}$ (la frequenza attesa se H0 è vero), r è il numero di variabili e k è il numero di gruppi																			
Criteria	Rifiutare H0 se $X^2 > \chi^2_f$, f è il numero di gradi di libertà determinati come $f = (r-1)(k-1)$. È il punto percentuale superiore del Chi-2 distribuzione con f gradi di libertà, che è tabulata, ad esempio, in Tabella B.2 e di Siegel e Castellan [157]																			

Tabella 10.18 Calcoli per il test Chi-2 (i valori attesi, E_{ij} , sono visualizzati tra parentesi)

Dimensione del modulo Combinato	Sistema	Sistema	
piccolo	1 15 (12.8205)	2 10 (12.1795)	R1D25
medio	20 (20)	19 (19)	R2D39
grande	25 (27.1795)	28 (25.8205)	R3D53
Totale	C1D60	C2D57	ND 117

Esempio: Chi-2 Bontà di adattamento test per la distribuzione normale. 60 studenti hanno sviluppato lo stesso programma e la dimensione misurata viene visualizzata nella Tabella 10.20.

L'ipotesi nulla è che i dati siano distribuiti normalmente e l'alternativa ipotesi che non lo sia. Sulla base dei dati, la media e la deviazione standard possono essere stimato: $\bar{x} = 794.9833$ e $s = 83.9751$.

L'intervallo può essere suddiviso in segmenti che hanno la stessa probabilità di includendo un valore se i dati sono effettivamente distribuiti normalmente con media \bar{x} e deviazione standard s . In questo esempio l'intervallo è diviso in dieci segmenti. Al fine per trovare il limite superiore (x) del primo segmento dovrebbe essere la seguente equazione risolto:

$P(X < x) = 0.10$ dove $X \sim N(\bar{x}, s)$, che in termini di standard normale distribuzione s corrisponde a

$P(X_s < x) = 0.10$ dove $X_s \sim N(0, 1)$, che è lo stesso di

$P(X_s < z) = 0.10$ dove $Z \sim N(0, 1)$ e $z = \frac{x - \bar{x}}{s}$.

Tabella 10.19 Chi-2, bontà di adattamento

Articolo	Descrizione
Ingresso	Dati come frequenze per un gruppo (ovvero, O1; O2;:::On, dove Oi rappresenta il numero di osservazioni nella categoria i). Confrontare con la Tabella 10.2
H0	Le misurazioni provengono da una certa distribuzione
Calcoli	Calcola $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, dove E_i è il numero atteso di osservatori
Criterio	Rifiutare H0 se $\chi^2 > \chi^2_{\alpha, f}$, dove f è il numero di gradi di libertà determinati poiché $f = n - 1$, ed n è il numero di parametri che devono essere stimati dai dati originali (vedi sotto). È il punto percentuale superiore del 1 - α della distribuzione Chi-2 con f gradi di libertà, che viene tabulata, ad esempio, nella Tabella B.2 e da Siegel e Castellan [157]. Questo è un test unilaterale

Tabella 10.20 Dimensioni misurate

757 758	892 734	800 979 938	866 690	877 773	778
679 888 799 811		657 750 891	724 775	810 940	854
784 843	867	743 816 813 618	715 706	906 679	845
708 855 777		660 870 843 790	741 766	677 801	850
821	877 713	680 667 752 875	811 999	808 771	832

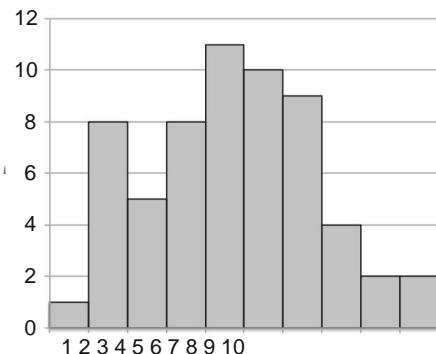
Tabella 10.21 Segmenti

Segmento	Limite inferiore	Limite superiore	Numero di valori
1		687,3	8
2	687,3	724,3	6
3	724,3	750,9	4
4	750,9	773,7	6
5	773,7	795	5
6	795	816,3	9
7	816,3	839	2
8	839	865,7	6
9	865,7	902,6	9
10	902,6		5

Queste equazioni possono essere risolte in diversi modi. Un modo è farlo ripetere e utilizzare un computer per trovare z o x . Un altro modo è utilizzare una tabella di distribuzione normale standard (che mostra $P(X < z)$ per diversi valori di z). Questo tipo di tabella è disponibile nella maggior parte dei libri di statistica. È anche possibile utilizzare una tabella specializzata che mostra direttamente i valori limite per i segmenti, cioè i valori di z . Questo tipo di tabella è presentata da Humphrey [82].

I limiti del segmento risultanti e il numero di valori che rientrano in ciascuno segmento sono mostrati nella Tabella 10.21.

Il numero atteso di valori (E_i) in ciascun segmento è $60=10 D 6$. Ciò significa che $X^2 D 7:3$. Il numero di gradi di libertà è $10 - 1 = 9$. Nella Tabella B.2, vede che è impossibile rifiutare H_0 al 14:07. Poiché $X^2 < si 0:05;2$, ipotesi nulla al livello 0,05. Se guardiamo un istogramma, vedi Fig. 10.10, dei dati, possiamo vedere che sembra essere distribuito abbastanza normalmente.

Fig. 10.10 Istogramma

Osservazioni finali del test Chi-2. Il test Chi-2 si basa su alcune ipotesi, che possono essere soddisfatte se i valori attesi, E_i , non sono troppo piccoli. Una regola pratica è che, se il numero di gradi di libertà (f) è uguale a 1, il test Chi-2 non dovrebbe essere utilizzato se una qualsiasi delle frequenze previste è inferiore a 5. Se $f > 1$ il test Chi-2 non dovrebbe essere utilizzato se più del 20% delle frequenze previste sono inferiori a 5 o qualcuna di esse è inferiore a 1. Va osservato che talvolta il test viene utilizzato anche se le frequenze previste non sono soddisfatte. In questi casi si tratta di un rischio calcolato.

Un modo per ottenere frequenze attese più ampie è combinare le categorie correlate in nuove categorie. Tuttavia, le nuove categorie devono essere significative. Per maggiori informazioni riguardanti il test Chi-2 fare riferimento, ad esempio, a Siegel e Castellan [157].

10.3.13 Verifica di adeguatezza del modello

Ogni modello statistico si basa su ipotesi specifiche riguardanti, ad esempio, distribuzione, indipendenza e scale. Se le ipotesi vengono invalidate dal set di dati, i risultati del test delle ipotesi non sono validi. Pertanto, è fondamentale verificare che tutte le ipotesi siano soddisfatte.

La verifica dell'adeguatezza del modello viene effettuata in base alle ipotesi. Di seguito descriviamo tre casi:

Normalità Se un test presuppone che i dati siano distribuiti normalmente, è possibile effettuare un test Chi-2 per valutare in che misura l'ipotesi è soddisfatta.
Il test Chi-2 è descritto sopra.

Indipendenza Se il test presuppone che i dati siano un campione di più variabili stocastiche indipendenti, è necessario verificare che non vi sia correlazione tra gli insiemi di campioni. Ciò può essere verificato con grafici a dispersione e calcolando i coefficienti di correlazione come discusso all'inizio di questa sezione.

Residui In molti modelli statistici esiste un termine che rappresenta i residui (errore statistico). Spesso si presume che i residui

sono normalmente distribuiti. Un modo comune per verificare questa proprietà è tracciare i residui in un grafico a dispersione e vedere che non ci sono tendenze specifiche nei dati (la distribuzione sembra casuale).

10.3.14 Trarre conclusioni

Una volta analizzati e interpretati i dati dell'esperimento, dobbiamo trarre conclusioni sull'esito dell'esperimento. Se le ipotesi vengono respinte possiamo trarre conclusioni sull'influenza delle variabili indipendenti sulle variabili dipendenti, dato che l'esperimento è valido, vedi Cap. 8.

Se, d'altra parte, l'esperimento non può rifiutare l'ipotesi nulla, non possiamo trarre alcuna conclusione sull'influenza delle variabili indipendenti sulla variabile dipendente. L'unica cosa che abbiamo dimostrato, in questo caso, è che non esiste alcuna differenza statisticamente significativa tra i trattamenti.

Se abbiamo trovato differenze statisticamente significative, vogliamo trarre conclusioni generali sulla relazione tra variabili indipendenti e dipendenti. Prima che ciò possa essere fatto dobbiamo considerare la validità esterna dell'esperimento, vedi Cap. 8. Possiamo generalizzare il risultato solo ad ambienti simili a quello sperimentale.

Sebbene il risultato dell'esperimento possa essere statisticamente significativo, non è necessariamente che abbia importanza pratica. Supponiamo, ad esempio, che il metodo X abbia dimostrato con un'elevata significatività statistica di essere il 2% più conveniente rispetto al metodo Y, sebbene vi sia un'elevata significatività statistica, il miglioramento derivante dal passaggio dal metodo Y al metodo X potrebbe non essere economicamente vantaggioso. Cioè, è necessario studiare la dimensione dell'effetto osservato dei diversi trattamenti e, sulla base di ciò, trarre conclusioni e presentare raccomandazioni. Kampenes et al. [92] forniscono una panoramica dei diversi concetti di dimensione dell'effetto e presentano una revisione sistematica su come questo viene gestito negli articoli pubblicati.

Potrebbe anche essere viceversa; anche se i risultati dell'esperimento potrebbero non essere statisticamente significativi o avere una bassa significatività statistica, le lezioni apprese dall'esperimento potrebbero comunque essere di importanza pratica. Il fatto che un'ipotesi nulla non possa essere rifiutata con un certo livello di significatività non significa che l'ipotesi nulla sia vera. Potrebbero esserci problemi con la progettazione dell'esperimento, come minacce reali alla validità o campioni di dati troppo pochi. Inoltre, a seconda della situazione e dell'obiettivo dello studio, potremmo accontentarci di una significatività statistica inferiore poiché i risultati hanno un'elevata importanza pratica. Questo problema è anche legato alla discussione riguardante le minacce alla validità, vedere Sez. 8.9.

Quando si trova una correlazione significativa tra una variabile A e una variabile B, non possiamo, in generale, trarre la conclusione che esiste una *relazione causale* tra A e B. Potrebbe esserci un terzo fattore C, che causa gli effetti misurabili su A e B.

Le conclusioni tratte in base al risultato dell'esperimento, sono input per una decisione, ad esempio, se un nuovo metodo sarà applicato nei progetti futuri, o se sono necessarie ulteriori sperimentazioni.

Va notato che ci sono anche degli svantaggi nell'utilizzare il test di ipotesi. Come sottolinea Miller [122], la maggior parte delle ipotesi nulle sono formulate in modo tale da essere sempre respinte, se vengono forniti dati sufficienti, e non è possibile ottenere effettivamente un campione rappresentativo dell'intera popolazione, ad esempio, di tutti gli ingegneri informatici del mondo. È necessario prestare sempre attenzione quando si intraprendono azioni basate sui risultati di un esperimento e il risultato dell'esperimento dovrebbe essere visto come un fattore nel processo decisionale.

10.4 Esempio di analisi

L'esempio è una continuazione dell'esempio nella Sez. 9.4. Sulla base dei dati sperimentali derivanti dall'esecuzione, il primo passo è applicare le statistiche descrittive, ovvero tracciare i dati. In relazione alle scale di misurazione dovranno essere utilizzati metodi statistici adeguati, come descritto nella Sez. 10.1. Un modo comunemente utilizzato per tracciare i dati è utilizzare i box plot. Forniscono un'eccellente opportunità per ottenere una panoramica dei dati e identificare i valori anomali. Se si identifica un valore anomalo, è importante capire se esiste una spiegazione sottostante. Ad esempio, può accadere che uno o più soggetti abbiano un background molto diverso rispetto agli altri e quindi è necessario garantire che i loro dati siano comparabili con quelli degli altri soggetti.

Potrebbe essere particolarmente critico se solo uno dei gruppi è interessato. In generale, dovremmo essere restrittivi nella rimozione dei dati, ovvero qualsiasi rimozione di dati dovrebbe essere ben motivata e documentata.

Una volta deciso quali dati includere nell'analisi dei dati, è tempo di dare un'occhiata all'analisi statistica. L'analisi statistica è sempre una sfida e ci sono molte opinioni diverse sull'uso dei diversi metodi statistici e su quando utilizzare metodi parametrici e non parametrici.

Il primo passo è verificare se i dati sono distribuiti normalmente, ad esempio tracciando un istogramma (Fig. 10.5) o utilizzando il test Chi-2 come descritto nella Sez. 10.3.12 oppure utilizzando altri test alternativi come ad esempio il cosiddetto test di Kolmogorov–Smirnov, il test W di Shapiro–Wilks, oppure il test di Anderson–Darling. Tuttavia, con una dimensione del campione ridotta, i dati potrebbero sembrare distribuiti normalmente senza esserlo effettivamente e i test di normalità potrebbero non rilevarli a causa della presenza di pochi punti dati. Alcuni test parametrici sono più robusti di altri rispetto alle deviazioni dalla normalità. Ad esempio, il test t è abbastanza robusto per la non normalità, il che non è il caso dell'ANOVA. Indipendentemente, potrebbe essere utile indagare se i dati sono distribuiti normalmente.

Considerato il design a due fattori (tecnica di lettura e documento dei requisiti) con due trattamenti ciascuno, vi è una grande necessità che i dati siano distribuiti normalmente.

Se i dati sono distribuiti normalmente è possibile utilizzare un'ANOVA. Se i dati non sono distribuiti normalmente allora c'è un problema, poiché non esiste una controparte non parametrica per questo tipo di progetto, come illustrato nella Tabella 10.3; se si ha un solo fattore con due trattamenti allora ci sono alternative non parametriche. Pertanto, anche se esistono design più semplici, il design scelto sembra abbastanza semplice ed è forte la tentazione di utilizzarlo. Tuttavia, potrebbe non essere una buona scelta utilizzare questo tipo di design.

Genera più punti dati, ma comporta alcune sfide quando si arriva all'analisi statistica. Pertanto, è importante essere consapevoli delle conseguenze in termini di analisi quando si seleziona il disegno dell'esperimento. Questo tipo di disegno viene talvolta definito disegno crossover, ovvero i primi soggetti utilizzano o sono esposti a un trattamento e poi sono esposti a un secondo trattamento. Alcune delle sfide sarebbero affrontate se fosse possibile avere un solo fattore, vale a dire la tecnica di lettura. Tuttavia, non è realistico utilizzare lo stesso documento sui requisiti per due ispezioni a meno che non passi molto tempo tra le due analisi. Kitchenham et al. [99] presentano alcune sfide statistiche legate alla progettazione del crossover. Detto questo, i progetti crossover non sono rari nell'ingegneria del software poiché si tratta di un compromesso tra le sfide statistiche e l'avere (troppo) pochi soggetti assegnati a ciascun trattamento, anche se altri sostengono che i progetti crossover non possono essere raccomandati nell'ingegneria del software [99].

Se si presuppone che i dati siano distribuiti normalmente, è possibile applicare un test ANOVA. Tuttavia, il problema è che se l'ANOVA mostra un risultato significativo, non è ancora noto quale differenza sia significativa. Per fare ciò è necessario utilizzare alcuni test aggiuntivi dopo l'ANOVA, ad esempio il test PLSD (Differenza Meno Significativa Protetta) di Fisher [125]. Il test richiede un'ANOVA significativa per essere utilizzato, ovvero è protetto da un'ANOVA significativa. Il PLSD di Fisher viene utilizzato per effettuare un confronto a coppie delle medie. Ancora una volta illustra alcune delle sfide statistiche che derivano come diretta conseguenza del disegno dell'esperimento. Pertanto, è necessario disporre di disegni sperimentali abbastanza semplici per poter effettuare un'analisi statistica corretta.

Se avessimo scelto di dividere i soggetti in due gruppi e poi avessimo assegnato loro l'utilizzo del PBR o del CBR sullo stesso documento dei requisiti, avremmo solo un fattore con due trattamenti. Ciò significa che si sarebbe potuto utilizzare un test t o un test di Mann-Whitney a seconda dell'esito del test di normalità. Non avremmo invece alcuna indicazione sull'interazione tra soggetto e trattamento. Se questo sia migliore o peggiore rispetto ad altri disegni alternativi deve essere deciso in ogni singolo caso a seconda del numero di soggetti e delle minacce alla validità identificate.

10.5 Esercizi

10.1. Cos'è la statistica descrittiva e a cosa serve?

10.2. Cos'è rispettivamente un test parametrico e non parametrico e quando possono essere applicati?

10.3. Qual è il potere di un test?

10.4. Cos'è un confronto accoppiato?

10.5. Spiegare brevemente il test ANOVA.

Capitolo 11

Presentazione e pacchetto

Una volta completato un esperimento, i risultati possono essere presentati a pubblici diversi, come definito nella Figura 11.1. Ciò potrebbe, ad esempio, essere fatto in un articolo per una conferenza o una rivista, in un rapporto per i decisori, in un pacchetto per la replica dell'esperimento o come materiale didattico. Il confezionamento potrebbe essere effettuato anche all'interno delle aziende per migliorare e comprendere diversi processi. In questo caso è opportuno archiviare le esperienze in una experience base secondo i concetti discussi da Basili et al. [16]. Tuttavia, qui ci concentriamo sulla rendicontazione accademica in riviste e conferenze. Se limitazioni di spazio impediscono una rendicontazione completa di tutti i dettagli, incoraggiamo la pubblicazione parallela di una relazione tecnica.

Jedlitschka e Pfahl propongono uno schema per la rendicontazione accademica degli esperimenti [86] che è stato successivamente valutato da Kitchenham et al. [101]. La proposta di Jedlitschka e Pfahl è riassunta nella Tabella 11.1 e brevemente elaborata nella Sez. 11.1.

11.1 Struttura del rapporto sull'esperimento

Estratto strutturato. L'abstract dovrebbe fornire al lettore un breve riassunto delle caratteristiche chiave dell'esperimento. È stato empiricamente dimostrato che gli abstract strutturati sono strumenti efficaci per facilitare l'estrazione dei dati [30] e per scrivere buoni abstract [31]. Gli elementi di un abstract strutturato sono:

- Contesto o contesto, • Obiettivi
o scopi, • Metodo, • Risultati
e • Conclusioni

Esempio. Per illustrare gli item viene presentato un esempio di abstract strutturato. In questo caso la lunghezza dell'abstract strutturato è limitata a 300 parole:

Contesto: all'interno di un'organizzazione, le persone hanno responsabilità e compiti lavorativi diversi, quindi è probabile che ruoli diversi abbiano priorità diverse quando

Tabella 11.1 Struttura di reporting proposta per i rapporti sugli esperimenti, di Jedlitschka e Pfahl [86]

Sezioni/sottosezioni	Contenuto
Titolo, paternità	
Estratto strutturato	Riassume il documento sotto titoli di background o contesto, obiettivi o scopi, metodo, risultati e conclusioni
Motivazione	Definisce lo scopo del lavoro e incoraggia i lettori a leggere il resto del libro carta
Dichiarazione del problema	Riporta qual è il problema; dove avviene e chi lo osserva
Obiettivi della ricerca	Definisce l'esperimento utilizzando lo stile formalizzato utilizzato in GQM
Contesto	Riporta fattori ambientali come impostazioni e posizioni
Lavoro correlato	Come lo studio attuale si collega ad altre ricerche
Progettazione sperimentale	Describe il risultato della fase di pianificazione sperimentale
Obiettivi, ipotesi e variabili	Presenta gli obiettivi di ricerca perfezionati
Progetto	Definire il tipo di disegno sperimentale
Soggetti	Definisce i metodi utilizzati per il campionamento dei soggetti e l'assegnazione dei gruppi
Oggetti	Definisce quali oggetti sperimentali sono stati utilizzati
Strumentazione	Definisce eventuali linee guida e strumenti di misurazione utilizzati
Raccolta dati procedura	Definisce il programma sperimentale, i tempi e le modalità di raccolta dei dati
Procedura di analisi	Specifica il modello di analisi matematica da utilizzare
Valutazione di validità	Describe la validità dei materiali, le procedure per garantire i partecipanti attenersi al metodo sperimentale e ai metodi per garantire la affidabilità e validità dei metodi e degli strumenti di raccolta dei dati
Esecuzione	Describe come è stato implementato il piano sperimentale
Campione	Descrizione delle caratteristiche del campione
Preparazione	Come sono stati formati e addestrati i gruppi sperimentali
Raccolta dati eseguito	Come è avvenuta la raccolta dei dati ed eventuali deviazioni dal piano
Procedura di validità	Come è stato seguito il processo di validità ed eventuali deviazioni dal piano
Analisi	Riassume i dati raccolti e descrive come sono stati analizzati
Statistiche descrittive	Presentazione dei dati mediante statistica descrittiva
Riduzione del set di dati	Describe qualsiasi riduzione del set di dati, ad esempio la rimozione di valori anomali
Verifica di ipotesi	Describe come sono stati valutati i dati e come era il modello di analisi convalidato
Interpretazione	Interpreta i risultati della sezione Analisi
Valutazione dei risultati e implicazioni	Spiega i risultati
Limitazioni dello studio	Discute le minacce alla validità
Inferenze	Come i risultati si generalizzano dati i risultati e le limitazioni
Lezione appresa	Descrizioni di cosa è andato bene e cosa no durante il corso l'esperimento
Conclusioni e lavoro futuro	Presenta una sintesi dello studio
Rapporto con l'esistente prova	Describe il contributo dello studio nel contesto precedente esperimenti
Impatto	Identifica i risultati più importanti
Limitazioni	Identifica i principali limiti dell'approccio, ovvero le circostanze in cui il i benefici attesi non verranno erogati
Lavoro futuro	Suggerimenti per altri esperimenti da approfondire
Ringraziamenti	Identifica eventuali contributori che non soddisfano i criteri di paternità
Riferimenti	Elenco tutta la letteratura citata
Appendici	Include dati grezzi e/o analisi dettagliate che potrebbero aiutare altri a farlo utilizzare i risultati



Fig. 11.1 Panoramica della presentazione e del pacchetto

si tratta di ciò che dovrebbe essere migliorato all'interno di un'azienda. Ciò è stato riscontrato in precedenti studi di marketing, ma è vero anche per il miglioramento del software?

OBIETTIVO: questo documento valuta il modo in cui i diversi ruoli in un'organizzazione di sviluppo software vedono i diversi problemi nel miglioramento dei processi software e se tali differenze potrebbero essere utilizzate per fornire miglioramenti dei processi più personalizzati all'interno di un'organizzazione e utilizza ciò come ipotesi di lavoro. **METODO:** è stato sviluppato un questionario quantitativo contenente cinque diverse domande ponderate relative al miglioramento del processo software. Sono stati quindi contattati ottantaquattro dipendenti di tutti i livelli di una società di telecomunicazioni svedese, di cui 63 hanno risposto. **RISULTATI:** i diversi ruoli erano in disaccordo in tre domande mentre erano d'accordo in due domande. Il disaccordo riguardava questioni relative all'importanza del miglioramento, all'urgenza dei problemi e alla minaccia per il successo della gestione dei processi, mentre le domande su cui i ruoli concordati si concentravano sulla comunicazione dei processi (documentazione e insegnamento). **CONCLUSIONE:** si conclude che è importante essere consapevoli e tenere conto delle diverse esigenze dei diversi ruoli. Ciò consentirà di fornire miglioramenti adattati a ruoli specifici che probabilmente aiuteranno a superare la resistenza ai miglioramenti dei processi. È anche importante esaminare altre aree e aziende (ad esempio il marketing) in cui potrebbe essere utile apportare miglioramenti ai processi.

Motivazione. La motivazione o introduzione stabilisce lo scopo e definisce l'obiettivo della ricerca, quindi riporta principalmente l'esito della fase di scoping (vedi Cap. 7). È possibile includere anche informazioni sull'intento del lavoro per chiarire e catturare l'interesse dei lettori. Ciò fornisce al lettore una comprensione del motivo per cui la ricerca è stata effettuata e perché ce n'è bisogno. Qui di seguito viene brevemente presentato il contesto in cui viene condotto l'esperimento.

Lavoro correlato. Il lavoro correlato è importante per fornire un quadro di come l'esperimento attuale è correlato al lavoro condotto in precedenza. Ogni rapporto sperimentale non necessita di una revisione sistematica completa della letteratura (vedi Cap. 4), sebbene essere sistematici nella ricerca della letteratura sia per lo più vantaggioso. In particolare, nel caso di studi di replica, dovranno essere riportati tutti gli studi precedenti.

Progettazione sperimentale. Qui si riporta l'esito della fase di pianificazione, vedi Cap. 8. Le ipotesi, che derivano dalla formulazione del problema, sono descritte in dettaglio. Viene presentato il disegno sperimentale, compreso il disegno

tipologia, variabili misurate, sia indipendenti che dipendenti, nonché la strumentazione.

Dovrebbe essere inclusa una descrizione di come i dati verranno raccolti e analizzati. Dovrebbe essere fornita una caratterizzazione dei soggetti. La discussione sulla conclusione dell'esperimento, sulla validità interna, costruttiva ed esterna dovrebbe essere fornita qui insieme alle possibili minacce contro i piani.

Lo scopo della descrizione di questi elementi è quello di consentire ad altre persone di comprendere il disegno in modo che sia visibile al lettore che i risultati sono affidabili e di consentire la replica dello studio. In breve, dovrebbe aiutare il lettore ad approfondire la comprensione di ciò che è stato fatto.

Esecuzione. La prima parte da descrivere è come viene preparata l'operazione, vedi Cap. 9. È importante includere descrizioni degli aspetti che faciliteranno la replica dell'esperimento e fornire informazioni su come sono state svolte le attività. La preparazione dei soggetti deve essere presentata. È importante fornire informazioni, ad esempio se hanno frequentato o meno alcune lezioni. Dovrebbe essere presentata anche l'esecuzione dell'esperimento e il modo in cui sono stati raccolti i dati durante l'esperimento.

Le procedure di validazione della raccolta dei dati sono un'altra questione che deve essere sottolineata e va segnalato se sono stati fatti passi avanti rispetto ai piani. Tutte le informazioni hanno lo scopo di dimostrare la validità dei dati e di evidenziare i problemi.

Analisi. Dovrebbe essere fornita una presentazione dell'analisi dei dati, in cui i calcoli sono descritti insieme alle ipotesi per l'utilizzo di alcuni modelli di analisi specifici. Devono essere incluse anche informazioni sulle dimensioni del campione, sui livelli di significatività e sull'applicazione dei test in modo che il lettore possa conoscere i prerequisiti per l'analisi.

Le ragioni delle azioni intraprese, ad esempio la rimozione dei valori anomali, dovrebbero essere descritte per evitare malintesi nell'interpretazione dei risultati. Per maggiori informazioni vedere il cap. 10.

Interpretazione. I risultati grezzi dell'analisi non sono sufficienti per fornire una comprensione dei risultati e delle conclusioni dell'esperimento. Occorre fornire anche un'interpretazione, cfr. cap. 10. Comprende il rifiuto dell'ipotesi o l'incapacità di rifiutare l'ipotesi nulla. L'interpretazione riassume come possono essere utilizzati i risultati dell'esperimento.

L'interpretazione dovrebbe essere fatta con riferimenti alla validità, vedi Cap. 8. Fattori che potrebbero aver avuto un impatto sui risultati dovrebbero essere descritti.

Conclusioni e ulteriore lavoro. Infine, le discussioni sui risultati e sulle conclusioni vengono presentate come un riassunto dell'intero esperimento insieme ai risultati, ai problemi, alle deviazioni dai piani e così via. I risultati dovrebbero anche essere correlati al lavoro riportato in precedenza. È importante affrontare le somiglianze e le differenze nei risultati.

In questa sezione potrebbero essere incluse anche idee per il lavoro futuro e informazioni su dove trovare ulteriori informazioni per ottenere una visione più approfondita dell'esperimento e per facilitarne la replica.

Appendici. Le informazioni non vitali per la presentazione potrebbero essere incluse negli allegati. Potrebbero trattarsi, ad esempio, dei dati raccolti e di ulteriori informazioni sui soggetti e sugli oggetti. Se l'intenzione è quella di produrre un pacchetto da laboratorio, il materiale utilizzato nell'esperimento potrebbe essere fornito qui.

11.2 Esercizi

11.1. Perché è importante documentare accuratamente un esperimento?

11.2. Cos'è un pacchetto laboratorio? Riesci a trovare qualche pacchetto di laboratorio su Internet?

11.3. Perché è importante segnalare il lavoro correlato?

11.4. Perché non è sufficiente fornire solo i risultati dell'analisi? In altre parole, perché è importante un'interpretazione speciale dei risultati?

11.5. Quali informazioni contenute nel rapporto sono più importanti quando si conduce una revisione sistematica della letteratura? Quando si replica un esperimento?

Parte III

Esperimenti di esempio

Capitolo 12

Illustrazione del processo dell'esperimento

L'obiettivo principale della presentazione di questo esperimento è illustrare la sperimentazione e le fasi del processo sperimentale introdotte nei capitoli precedenti. La presentazione dell'esperimento in questo capitolo si concentra sul processo dell'esperimento piuttosto che seguire la struttura del rapporto proposta nel Cap. 11.

L'obiettivo dell'esperimento presentato è quello di indagare le prestazioni nell'utilizzo del Personal Software Process (PSP) [82, 83] in base al background individuale delle persone che seguono il corso PSP. L'esperimento, come qui riportato, è parte di un'indagine più ampia sulle differenze individuali di prestazione all'interno della PSP. Dato che il "conto individuale" non può essere assegnato in modo casuale ai soggetti, l'esperimento è in realtà un quasi-esperimento.

Il PSP è un processo individuale per un approccio sistematico allo sviluppo del software. Il processo comprende, ad esempio, la misurazione, la stima, la pianificazione e il monitoraggio. Inoltre, il riutilizzo è una questione chiave, e in particolare il riutilizzo delle esperienze e dei dati individuali. Il corso PSP introduce il processo in sette passaggi incrementali aggiungendo nuove funzionalità al processo utilizzando modelli, moduli e script di processo.

Per semplicità vengono qui valutate solo due ipotesi. Il set di dati per lo studio più ampio può essere trovato nell'Appendice A.1.2. L'esperimento presentato in questo capitolo utilizza un sottoinsieme di dati.

12.1 Ambito

12.1.1 Definizione dell'obiettivo

Il primo passo è decidere se un esperimento è un modo adeguato per analizzare il problema in questione. In questo caso particolare, l'obiettivo dello studio empirico è quello di determinare le differenze nelle prestazioni individuali delle persone che utilizzano la PSP in base al loro background.

L'esperimento è motivato dalla necessità di comprendere le differenze nelle prestazioni individuali all'interno della PSP. È risaputo e accettato che gli ingegneri del software si comportano in modo diverso. Uno degli obiettivi dell'introduzione di un processo personale è fornire supporto agli individui per migliorare le proprie prestazioni. Per sostenere al meglio il miglioramento è importante capire quali differenze ancora ci si possono aspettare all'interno del PSP e se è possibile spiegare e quindi comprendere le differenze individuali.

Oggetto di studio. Oggetto di studio sono i partecipanti al corso Personal Software Process (PSP) e le loro capacità in termini di performance in base al loro background ed esperienza. Humphrey definisce il Personal Software Process nei suoi due libri sull'argomento [82, 83].

Scopo. Lo scopo dell'esperimento è valutare la performance individuale in base al background delle persone che seguono il corso PSP. L'esperimento fornisce informazioni su cosa ci si può aspettare in termini di prestazioni individuali quando si utilizza la PSP.

Prospettiva. La prospettiva è dal punto di vista dei ricercatori e degli insegnanti, cioè il ricercatore o l'insegnante vorrebbe sapere se ci sono differenze sistematiche nelle prestazioni del corso in base al background degli individui che entrano nel corso PSP. Ciò include anche le persone che potrebbero voler seguire il corso in futuro o introdurre la PSP nell'industria.

Focus sulla qualità. L'effetto principale studiato nell'esperimento è la prestazione individuale nel corso PSP. Qui si sottolineano due aspetti specifici. La scelta è quella di concentrarsi sulla Produttività (KLOC/tempo di sviluppo) e sulla Densità dei Difetti (guasti/KLOC), dove KLOC sta per migliaia di righe di codice.

Contesto. L'esperimento viene eseguito nel contesto della PSP. Inoltre, l'esperimento è condotto all'interno di un corso PSP tenuto presso il Dipartimento di Sistemi di Comunicazione dell'Università di Lund in Svezia. Questo studio proviene dal corso tenuto nel 1996-1997 e la differenza principale rispetto al PSP presentato da Humphrey [82] è che si è deciso di utilizzare uno standard di codifica e uno standard di conteggio delle linee. Inoltre, il corso si è svolto con C come linguaggio di programmazione obbligatorio, indipendentemente dal background degli studenti. La caratterizzazione del contesto sperimentale è "multi-test all'interno dello studio dell'oggetto", vedere Tabella 7.1. Lo studio si concentra sul PSP o più specificamente sui dieci programmi nel libro di Humphrey [82] indicati con 1A–10A.

Il corso PSP è seguito da un gran numero di persone (quest'anno in particolare, 65 studenti hanno terminato il corso). Pertanto, nello studio sono stati inclusi 65 soggetti, vedere sez. 7.1. Pertanto, da questa definizione lo studio può essere giudicato un esperimento controllato. La mancanza di randomizzazione degli studenti, cioè degli studenti iscritti al corso, fa però sì che allo studio manchi ancora un ingrediente importante per trasformarlo a pieno titolo in un esperimento controllato. Questo è quindi classificato come un quasi-esperimento, vedi Sez. 7.1.

12.1.2 Riepilogo dell'ambito

La sintesi è effettuata secondo la Sez. 7.2.

Analizzare l'esito del PSP ai fini della valutazione rispetto al background dei soggetti dal punto di vista dei ricercatori e degli insegnanti nel contesto del corso PSP.

12.2 Pianificazione

12.2.1 Selezione del contesto

Il contesto dell'esperimento è un corso PSP all'università, e quindi l'esperimento viene eseguito offline (non in uno sviluppo di software industriale), è condotto da studenti laureati (normalmente studenti al quarto anno di università). , e l'esperimento è specifico poiché focalizzato sulla PSP in un ambiente educativo.

La capacità di generalizzare da questo contesto specifico viene ulteriormente elaborata di seguito quando si discutono le minacce alla validità dell'esperimento. L'esperimento affronta un problema reale, ovvero le differenze nelle prestazioni individuali e la comprensione delle differenze.

L'uso del PSP come contesto sperimentale fornisce ad altri ricercatori ottime opportunità per replicare l'esperimento così come è ben definito. Inoltre, significa che non è necessario dedicare molti sforzi all'impostazione dell'esperimento in termini di definizione e creazione dell'ambiente in cui viene eseguito l'esperimento. Humphrey [82] definisce il contesto sperimentale, e quindi non c'è bisogno di preparare moduli per la raccolta dei dati e così via.

12.2.2 Formulazione di ipotesi

Un aspetto importante degli esperimenti è conoscere e dichiarare formalmente e chiaramente cosa verrà valutato nell'esperimento. Ciò porta alla formulazione di un'ipotesi (o più ipotesi). Qui si è scelto di concentrarsi su due ipotesi.

Informalmente sono:

1. Hanno frequentato il corso gli studenti sia del corso di Laurea Magistrale in Informatica e Ingegneria che del corso di Laurea in Ingegneria Elettrica. Gli studenti del programma di Informatica e Ingegneria normalmente hanno seguito più corsi di informatica e ingegneria del software, e quindi è previsto che ciò accada

hanno una produttività più elevata rispetto agli studenti del programma di Ingegneria Elettrica.

2. Nell'ambito della prima lezione, agli studenti è stato chiesto di compilare un sondaggio riguardante il loro background in termini di esperienze rispetto a questioni relative al corso, vedere Tabella A.1. Ciò può essere esemplificato, ad esempio, con la conoscenza in C.

Agli studenti veniva richiesto di utilizzare il C nel corso indipendentemente dalla loro precedente esperienza della lingua. Pertanto, non era richiesto che gli studenti avessero seguito un corso C prima di accedere al corso PSP, il che significava che alcuni studenti imparavano il C all'interno del corso PSP. Ciò non è secondo la raccomandazione di Humphrey [82]. L'ipotesi basata sull'esperienza in C è che gli studenti con più esperienza in C commettano meno errori per riga di codice.

Sulla base di questa formulazione informale delle ipotesi, è ora possibile enunciarle formalmente e anche definire quali misure sono necessarie per valutare le ipotesi.

1. Ipotesi nulla, H0: non c'è differenza nella produttività (misurata come righe di codice per tempo di sviluppo totale) tra gli studenti del programma di Informatica e Ingegneria (CSE) e del programma di Ingegneria Elettrica (EE).

H0: $\text{Prod}(\text{CSE}) = \text{Prod}(\text{EE})$

Ipotesi alternativa, H1: $\text{Prod}(\text{CSE}) \neq \text{Prod}(\text{EE})$

Misure necessarie: programma studentesco (CSE o EE) e produttività (LOC/ora).

2. Ipotesi nulla, H0: non c'è differenza tra gli studenti in termini di numero di errori per KLOC (1.000 righe di codice) in base alla conoscenza pregressa in C.

H0: il numero di errori per KLOC è indipendente dall'esperienza C.

Ipotesi alternativa, H1: il numero di guasti per KLOC cambia con l'esperienza C.

Misure necessarie: esperienza C e guasti/KLOC.

Le ipotesi implicano che sia necessario raccogliere i seguenti dati:

- Programma per studenti: misurato da CSE o EE (scala nominale) • La produttività è misurata come LOC/tempo di sviluppo. Pertanto, è necessario misurare la dimensione del programma (righe di codice secondo lo standard di codifica e lo standard di conteggio) e il tempo di sviluppo (minuti spesi per sviluppare il programma). Il tempo di sviluppo viene tradotto in ore quando viene calcolata la produttività.

Va notato che si è scelto di studiare la dimensione totale del programma (la somma dei dieci incarichi di programmazione) e il tempo di sviluppo per tutti e dieci i programmi. Pertanto, i compiti individuali non vengono studiati.

Le righe di codice vengono misurate contando le righe di codice utilizzando un programma contatore di righe (scala del rapporto). Le righe conteggiate sono righe di codice nuove e modificate. Il tempo di sviluppo è misurato in minuti (scala del rapporto).

La produttività viene quindi misurata su una scala proporzionale.

- L'esperienza C viene misurata introducendo una classificazione in quattro classi basate sulla precedente esperienza di C (scala ordinale). Le classi sono:

1. Nessuna esperienza precedente.
2. Leggi un libro o segui un corso.
3. Qualche esperienza industriale (meno di 6 mesi).
4. Esperienza industriale (più di 6 mesi).

L'esperienza in C è quindi misurata su una scala ordinale. • Guasti/KLOC viene misurato come il numero di guasti diviso per il numero di righe di codice.

Le ipotesi e le misure pongono vincoli sul tipo di test statistico da utilizzare, almeno formalmente. Le scale di misurazione determinano formalmente l'applicazione di metodi statistici specifici, ma potremmo voler allentare questi requisiti per altri motivi. Questo problema verrà discusso ulteriormente di seguito, quando si discuterà del tipo effettivo di progettazione nell'esperimento.

12.2.3 Selezione delle variabili

Le variabili indipendenti sono il programma dello studente e l'esperienza in C. Le variabili dipendenti sono produttività e difetti/KLOC.

12.2.4 Selezione dei soggetti

Le materie vengono scelte in base alla convenienza, ovvero i soggetti sono gli studenti che frequentano il corso PSP. Gli studenti rappresentano un campione di tutti gli studenti dei due programmi, ma non un campione casuale.

12.2.5 Progettazione dell'esperimento

Il problema è stato posto e sono state scelte le variabili indipendenti e dipendenti. Inoltre, sono state decise le scale di misurazione delle variabili. Pertanto, è ora possibile progettare l'esperimento. Il primo passo è affrontare i principi generali di progettazione:

Randomizzazione. L'oggetto non viene assegnato casualmente ai soggetti. Tutti gli studenti utilizzano la PSP e i suoi dieci compiti. L'obiettivo dello studio non è valutare la PSP rispetto a qualcosa'altro. I soggetti, come detto sopra, non sono selezionati in modo casuale; sono gli studenti che hanno scelto di frequentare il corso. Inoltre, le assegnazioni non vengono effettuate in ordine casuale. L'ordine, tuttavia, non è importante, poiché le misure utilizzate nella valutazione sono il risultato dello sviluppo dei dieci programmi.

Blocco. Non viene applicato alcun approccio sistematico al blocco. La decisione di misurare i dieci programmi e valutare in base ad essi, invece di considerare ogni singolo programma, può essere vista come un tentativo di bloccare le differenze tra i dieci programmi. Bloccando così l'impatto delle differenze tra i singoli programmi.

Bilanciamento. Sarebbe stato preferibile avere un set di dati bilanciato, ma lo studio sperimentale si basa su un corso in cui i partecipanti si sono iscritti al corso, e quindi non è possibile influenzare il background delle persone e di conseguenza non è possibile bilanciare il set di dati .

Tipi di design standard. Le informazioni disponibili vengono confrontate con la tipologia progettuale standard delineata nel Cap. 8. Entrambi i progetti possono essere trovati tra i tipi standard e i test statistici sono disponibili in questo libro.

1. La definizione, le ipotesi e le misure per la prima valutazione significano che il disegno è: un fattore con due trattamenti. Il fattore è il programma e i trattamenti sono CSE o EE. La variabile dipendente viene misurata su una scala di rapporti e quindi è adatto un test parametrico. In questo caso particolare verrà utilizzato il test t.
2. Il secondo disegno è del tipo "un fattore con più di due trattamenti". Il fattore è l'esperienza in C con quattro trattamenti, vedere la valutazione dell'esperienza sopra. La variabile dipendente viene misurata su una scala proporzionale ed è possibile utilizzare un test parametrico anche per questa ipotesi. Il test ANOVA è quindi adatto per la valutazione.

12.2.6 Strumentazione

Il background e l'esperienza degli individui vengono rilevati attraverso un sondaggio distribuito durante la prima lezione, vedere la Tabella A.1 nell'Appendice A. Questi dati forniscono l'input per la caratterizzazione degli studenti e quindi sono le variabili indipendenti nell'esperimento. Gli oggetti sono i programmi sviluppati all'interno del corso PSP. Le linee guida e le misurazioni sono fornite attraverso il PSP [82].

12.2.7 Valutazione di validità

Ci sono quattro livelli di minacce alla validità da considerare. *La validità interna* si concentra principalmente sulla validità dello studio reale. *La validità esterna* può essere suddivisa tra studenti PSP dell'Università di Lund nei prossimi anni, studenti dell'Università di Lund (o più realisticamente studenti dei programmi CSE ed EE), PSP in generale e sviluppo di software in generale. *La validità della conclusione* riguarda la relazione tra trattamento ed esito e la capacità di trarre conclusioni. *La validità di costrutto* riguarda la generalizzazione del risultato alla teoria alla base dell'esperimento.

La validità interna del corso probabilmente non è un problema. Il gran numero di prove (pari al numero degli studenti) garantisce una buona validità interna.

Per quanto riguarda le minacce esterne, è molto probabile che si ottengano risultati simili gestendo il corso in modo simile presso l'Università di Lund. È più difficile generalizzare i risultati ad altri studenti, cioè agli studenti che non frequentano il corso. Probabilmente non sono così interessati allo sviluppo di software e quindi provengono da una popolazione diversa. I risultati dell'analisi possono probabilmente essere generalizzati ad altri corsi PSP, dove è possibile confrontare i partecipanti in base al loro background in termini di informatica o ingegneria elettrica o esperienza di un particolare linguaggio di programmazione.

La principale minaccia per quanto riguarda la validità delle conclusioni è la qualità dei dati raccolti durante il corso PSP. Ci si aspetta che gli studenti forniscano molti dati come parte del loro lavoro con il corso. Esiste quindi il rischio che i dati siano falsificati o semplicemente non corretti a causa di errori. Tuttavia, non si ritiene che le incoerenze dei dati siano particolarmente legate a un background specifico, quindi il problema è probabilmente lo stesso indipendentemente dal background degli individui. La validità della conclusione non è quindi considerata critica.

La validità di costrutto include due minacce principali. La prima minaccia è che le misurazioni così come definite potrebbero non essere misure appropriate delle entità, ad esempio, il "LOC/Tempo di sviluppo" è una buona misura della produttività? La seconda grande minaccia alla validità di costrutto è che fa parte di un corso, in cui gli studenti vengono valutati. Ciò implica che gli studenti potrebbero influenzare i loro dati, poiché credono che ciò darà loro un voto migliore. All'inizio del corso è stato tuttavia sottolineato che il voto non dipendeva dai dati effettivi. Il voto si è basato sulla puntualità e correttezza dell'esposizione e sulla comprensione espressa nelle relazioni consegnate durante il corso.

I risultati sono stati trovati per PSP, ma è probabile che valgano per lo sviluppo di software in generale. Non vi è alcun motivo per cui persone provenienti da programmi di studio diversi o con esperienze di background diverse in un particolare linguaggio di programmazione si comportino diversamente tra PSP e lo sviluppo di software in generale.

Ciò è probabilmente valido quando si parla di differenze di background, sebbene la dimensione effettiva della differenza possa variare. La questione importante è che esiste una differenza e la dimensione effettiva della differenza è di minore importanza.

12.3 Funzionamento

12.3.1 Preparazione

I soggetti (studenti) non erano consapevoli di quali aspetti sarebbero stati studiati. Sono stati informati che i ricercatori volevano studiare i risultati del corso PSP confrontandoli con il background dei partecipanti. Erano, tuttavia,

non era a conoscenza delle reali ipotesi formulate. Gli studenti, dal loro punto di vista, non hanno partecipato principalmente ad un esperimento; stavano seguendo un corso. A tutti gli studenti è stato garantito l'anonymato.

Il materiale del sondaggio è stato preparato in anticipo. La maggior parte del resto del materiale, tuttavia, è stato fornito tramite il libro PSP [82].

12.3.2 Esecuzione

L'esperimento è durato 14 settimane, durante le quali i dieci compiti di programmazione sono stati consegnati regolarmente. I dati sono stati raccolti principalmente tramite moduli. Le interviste sono state utilizzate alla fine del corso, principalmente per valutare il corso e il PSP in quanto tale.

L'esperimento è stato, come detto in precedenza, condotto all'interno di un corso PSP e in un ambiente universitario. Non è stato consentito che l'esperimento influisca sugli obiettivi del corso. La differenza principale tra la gestione del PSP esclusivamente come corso è stata l'indagine iniziale del background degli studenti.

12.3.3 Convalida dei dati

I dati sono stati raccolti per 65 studenti. Dopo il corso, i risultati ottenuti dagli studenti sono stati discussi tra le persone coinvolte nel corso. I dati di sei studenti sono stati rimossi perché considerati non validi o quanto meno discutibili. Gli studenti non sono stati esclusi (in questa fase) dalla valutazione basata sui dati reali, ma a causa della nostra fiducia nei dati forniti e del fatto che i dati siano ritenuti rappresentativi o meno. I sei studenti sono stati rimossi a causa di:

- I dati di due studenti non sono stati compilati correttamente. •

Uno studente ha terminato il corso molto più tardi rispetto agli altri e ha trascorso un lungo periodo in cui non è stato svolto alcun lavoro con la PSP. Ciò potrebbe aver influito sui dati.

- I dati di due studenti sono stati rimossi perché consegnavano i compiti in ritardo e richiedevano un supporto notevolmente maggiore rispetto agli altri studenti, quindi si è ritenuto che i consigli aggiuntivi potessero aver influito sui loro dati. • Infine, uno studente è stato allontanato perché il suo background era completamente diverso da quello degli altri.

Ciò significa rimuovere sei studenti su 65, lasciando quindi 59 studenti per l'analisi statistica e l'interpretazione dei risultati.

12.4 Analisi e interpretazione

12.4.1 Statistiche descrittive

Come primo passo nell'analisi dei dati, vengono utilizzate statistiche descrittive per visualizzare i dati raccolti.

Programma di studio vs. produttività. La Figura 12.1 mostra la produttività per i due programmi di studio, dividendo la popolazione in classi in base alla produttività.

Alla prima classe appartengono quelli con una produttività compresa tra 5 e 10 righe di codice all'ora. Pertanto, l'ottava classe comprende quelli con una produttività compresa tra 40 e 45 righe di codice all'ora. Dalla Fig. 12.1 è possibile vedere che gli studenti del programma di Ingegneria Elettrica (EE) sembrano avere una produttività inferiore. Inoltre, è evidente che la variazione della distribuzione sembra essere maggiore tra gli studenti del corso di Informatica e Ingegneria (CSE). In totale, ci sono 32 studenti CSE e 27 studenti EE. Il valore medio per gli studenti CSE è 23,0 con una deviazione standard di 8,7, e per gli studenti EE il valore medio è 16,4 con una deviazione standard di 6,3. Per ottenere una comprensione ancora migliore dei dati, viene disegnato un box plot, vedere Fig. 12.2.

I baffi nel box plot sono costruiti come proposto da Frigge et al. [60] e discusso nel cap. 10. Per i baffi, un valore pari alla lunghezza della scatola moltiplicata per 1,5 e aggiunta o sottratta rispettivamente dai quartili superiore e inferiore. Ad esempio, per gli studenti CSE (vedi Fig. 12.2): mediana = 22,7, lunghezza della scatola = 29:4 17:6 D 11:8, la coda superiore diventa: 29:4 C 1:5 11:8 D 47: 1.

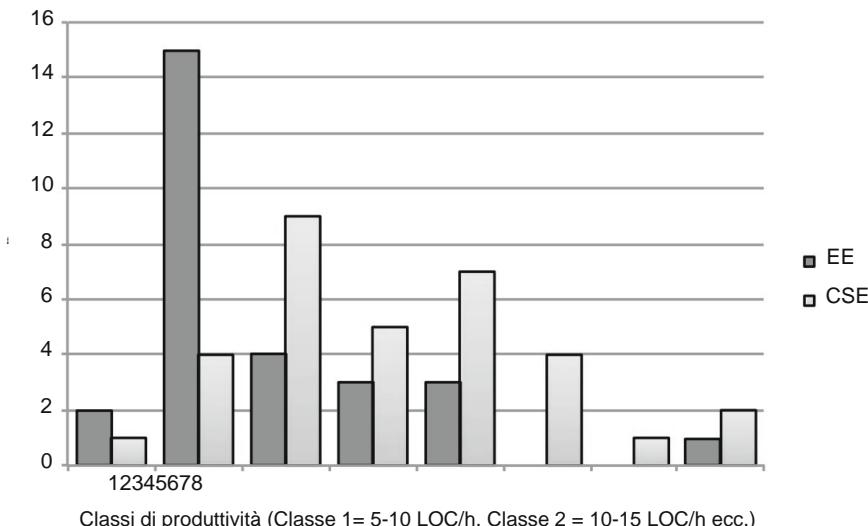


Fig. 12.1 Distribuzione di frequenza per la produttività (in classi)

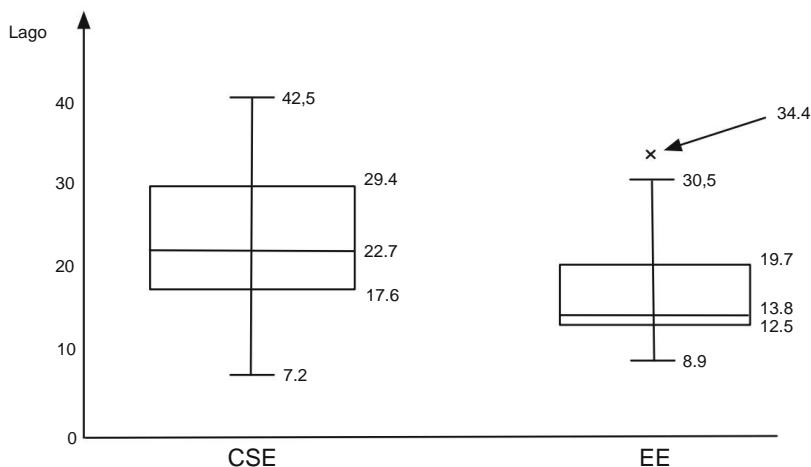


Fig. 12.2 Box plot della produttività per i due corsi di studio

Tabella 12.1 Errori/KLOC per le diverse classi di esperienza del C

Classe	Numero di studenti	Valore medio di guasti/KLOC	Valore medio di guasti/KLOC	Deviazione standard di guasti/KLOC
1	32	66.8	82.9	64.2
2	19	69.7	68.0	22.9
3	6	63.6	67.6	20.6
4	2	63	63.0	17.3

^{UN} Le diverse classi di esperienza sono spiegate nella Sez. 12.2.2

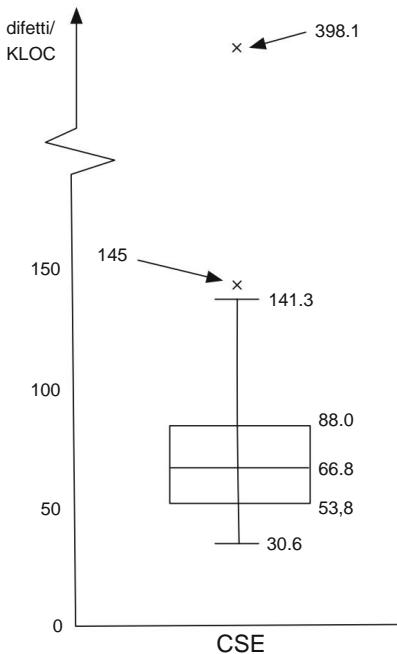
Esiste tuttavia un'eccezione a questa regola, vale a dire la coda superiore e quella inferiore non dovrebbe mai essere superiore o inferiore al valore massimo e minimo nel set di dati, quindi la coda superiore diventa 42,5, che è il valore più alto. Questa eccezione è introdotti per evitare valori negativi o altri tipi di valori non realistici. L'altro i valori in Fig. 12.2 si trovano in modo simile.

Dalla Fig. 12.2 si può vedere che esiste uno schema chiaro per gli studenti EE una produttività inferiore. Pertanto, potrebbe essere possibile identificare la differenza statisticamente in un test di ipotesi. Il test t viene utilizzato di seguito.

È anche importante considerare i valori anomali rispetto alle code superiore e inferiore. Per gli studenti CSE non esistono valori fuori coda. Per gli studenti EE, lì è un valore fuori dalla coda, cioè 34.4. Questo valore non è considerato un valore anomalo poiché lo è non è giudicato un valore estremo. È un valore insolito, ma si è determinati a mantenerlo il valore nell'analisi.

Esperienza C vs. difetti/KLOC. Il numero di studenti per ciascuna classe di C l'esperienza è mostrata nella Tabella 12.1, insieme ai valori medi e mediani, e deviazione standard per la rispettiva classe.

Fig. 12.3 Diagramma a riquadri per guasti/KLOC per la classe 1



Dalla Tabella 12.1 si può vedere che la distribuzione è sbilanciata verso nessuna o poca esperienza di C. Se si osservano i valori medi di errori/KLOC, sembra esserci una tendenza secondo cui gli studenti più esperti creano meno errori. La deviazione standard è, tuttavia, estremamente ampia e la mediana varia inaspettatamente rispetto al valore medio e all'ipotesi sottostante. La deviazione standard per la prima classe è molto elevata e si consiglia un'ulteriore indagine dei dati. Pertanto, il box plot può essere utilizzato anche per questo set di dati.

I box plot vengono costruiti per tutte e quattro le classi di esperienza. I grafici per le classi 2-4 non rivelano nulla, tutti i valori sono entro i confini dei baffi e quindi le code superiore e inferiore diventano rispettivamente uguali al valore più alto e più basso.

Il box plot per la prima classe è più interessante ed è mostrato in Fig. 12.3.

Dalla Fig. 12.3 si può vedere che la coda inferiore è uguale al valore più basso di guasti/KLOC. La coda superiore invece non è uguale al valore più alto e quindi sono presenti uno o più valori insoliti. In realtà ci sono due valori insoliti, vale a dire 145 e 398,1. Quest'ultimo valore è estremo; è più di dieci volte superiore al valore più basso. È anche quasi tre volte più alto del secondo valore più alto. Pertanto, è possibile concludere che l'elevata deviazione standard può essere spiegata con il valore estremo. Per la seconda ipotesi viene utilizzato il test ANOVA.

Le statistiche descrittive hanno fornito una migliore visione dei dati, sia in termini di cosa ci si può aspettare dalla verifica delle ipotesi sia di potenziali problemi causati da valori anomali.

Tabella 12.2 Difetti/KLOC per le diverse classi di esperienza C 1

Classe	Numero di studenti	Valore medio dei guasti/KLOC	Valore medio dei guasti/KLOC	Deviazione standard dei guasti/KLOC
1	31	66	72.7	29.0

12.4.2 Riduzione dei dati

Si può sempre discutere sulla riduzione dei dati, poiché non appena i punti dati vengono rimossi, le informazioni vanno perse. Si possono identificare due modi distinti per ridurre i dati:

- È possibile rimuovere singoli punti dati, ad esempio valori anomali, oppure • I dati possono essere analizzati e sulla base dell'analisi si può concludere che, a causa dell'elevata intercorrelazione tra alcune variabili, alcune misure dovrebbero essere combinate in misure più astratte .

Ciò significa che è possibile rimuovere i punti dati o ridurre il numero di variabili. Nel caso della rimozione dei punti dati, i principali candidati sono i valori anomali.

Non è affatto ovvio che tutti i valori anomali debbano essere rimossi, ma sono certamente candidati alla rimozione. È importante ricordare che i dati non dovrebbero essere rimossi semplicemente perché non si adattano alla convinzione o all'ipotesi. D'altra parte, è importante rimuovere i punti dati che potrebbero rendere non valida una relazione completamente valida, poiché, ad esempio, è incluso un valore anomalo estremo, cosa non prevista se si replica lo studio.

Per ridurre il numero di variabili sono necessari metodi statistici per la riduzione dei dati. Alcuni esempi sono l'analisi delle componenti principali e l'analisi fattoriale [90, 91, 118]. Questi tipi di metodi non vengono presi in considerazione in questa sede poiché l'obiettivo non è ridurre il numero di variabili.

Probabilmente è meglio essere restrittivi nel ridurre un set di dati, poiché c'è sempre il rischio di mirare a un determinato risultato. Pertanto, per i dati presentati sopra, si è scelto di rimuovere solo il valore anomalo estremo per il numero di guasti/KLOC. Dopo aver rimosso l'outlier estremo, i dati per la classe 1 sono riportati nella Tabella 12.2.

La rimozione del valore anomalo ha ridotto considerevolmente il valore medio e la deviazione standard. Il numero medio di guasti/KLOC è ancora più elevato per la classe 1. Tuttavia, le differenze tra le classi non sono così grandi. Dopo aver ridotto il secondo set di dati con un punto dati, non è possibile eseguire il test statistico. Qui si valutano le ipotesi.

12.4.3 Verifica di ipotesi

La prima ipotesi riguardante una maggiore produttività per gli studenti che seguono il corso di Informatica e Ingegneria viene valutata mediante un t-test. Viene applicato un test ANOVA per valutare l'ipotesi che più esperienza in C significhi meno errori/KLOC.

Tabella 12.3 Risultati del t-test

Fattore	Differenza media.	Gradi di libertà (DF)	valore t	valore p
CSE contro EE	6.1617	57	3.283	0,0018

Tabella 12.4 Risultati del test ANOVA

Fattore: C vs. guasti/KLOC	Gradi di libertà (DF)	Somma di piazze	Quadrato medio	Valore F	valore p
Tra i trattamenti	3	3483	1160.9	0,442	0,7236
Errore	55	144304	2623.7		

Programma di studio vs. produttività. I risultati del test t (spaiato, a due code) sono mostrati nella Tabella 12.3.

Dalla Tabella 12.3 si può concludere che H0 è rifiutato. C'è un significativo differenza di produttività per studenti provenienti da programmi di studio diversi. Il il valore p è molto basso quindi i risultati sono altamente significativi. Il vero motivo del la differenza deve essere ulteriormente valutata.

Esperienza C vs. difetti/KLOC. Questa ipotesi viene valutata con un test ANOVA (fattoriale). I risultati dell'analisi sono riportati nella Tabella 12.4.

I risultati dell'analisi non sono significativi, anche se alcune differenze lo sono osservato in termini di valore medio, vedi sopra, non è possibile dimostrare che esista a differenza significativa in termini di numero di guasti/KLOC basata sull'esperienza C.

Poiché il numero di studenti nelle classi 3 e 4 è molto limitato, le classi 2-4 lo sono raggruppati insieme per studiare la differenza tra la classe 1 e il resto. Un test t era eseguita per valutare se fosse possibile distinguere tra la classe 1 e la classe 1 raggruppamento delle classi 2-4 in un'unica classe. Non sono stati ottenuti risultati significativi.

12.5 Riepilogo

Abbiamo analizzato due ipotesi:

1. Programma di studi vs. produttività
2. Esperienza C vs. difetti/KLOC

Siamo in grado di dimostrare che gli studenti di Informatica e Ingegneria il programma è più produttivo. Questo però è in linea con le aspettative non formalmente indicato nell'ipotesi. L'aspettativa era basata sulla conoscenza che la maggior parte degli studenti del CSE ha seguito più studi di informatica e software ingegneria e corsi rispetto a quelli del programma di ingegneria elettrica.

Non è possibile dimostrare con alcuna significatività statistica che l'esperienza in C influenzi il numero di errori/KLOC. Questo è interessante in relazione a Humphrey [82] raccomanda di seguire il corso PSP con una persona conosciuta

linguaggio per concentrarsi sulla PSP e non sul linguaggio di programmazione in quanto tale.

I risultati ottenuti possono indicare uno o più dei seguenti risultati:

- La differenza diventerà significativa con più studenti. • Il numero di difetti introdotti non è influenzato in modo significativo dall'esperienza precedente. Potrebbe esserci la tendenza a commettere un certo numero di errori durante lo sviluppo del software. Il tipo di difetti può variare, ma il numero totale di difetti introdotti è più o meno lo stesso.
- Gli studenti inesperti possono scrivere programmi più grandi, che influiscono direttamente sul numero di errori/KLOC.

Probabilmente si possono trovare anche altre spiegazioni, ma tutte hanno una cosa in comune: la necessità di replicarsi. Pertanto, le repliche sono una questione importante che ci consente di comprendere, e quindi controllare e migliorare, il modo in cui viene sviluppato il software. Inoltre bisogna studiare anche altri fattori.

Dal punto di vista della validità, è ragionevole ritenere che gli studenti (in generale) di un corso di informatica abbiano una produttività maggiore rispetto agli studenti provenienti da altre discipline. Ciò è più o meno inerente al background educativo e non è una sorpresa.

Poiché non è stato possibile mostrare alcuna relazione statisticamente significativa tra l'esperienza in un linguaggio di programmazione e il numero di errori/KLOC; non ci sono conclusioni da generalizzare. Sono necessari ulteriori studi, attraverso la replica dell'esperimento PSP o studi simili in altri ambienti.

12.6 Conclusione

Lo studio presentato è un quasi-esperimento, poiché confronta fattori che non sono assegnati in modo casuale ai soggetti, ma piuttosto proprietà intrinseche dei soggetti (ad esempio il background educativo). È condotto con gli studenti come soggetti, il che fornisce una buona validità interna a scapito della validità esterna dei risultati. Essere condotto nel contesto PSP aiuterebbe la replicazione dello studio, poiché il contesto è molto ben definito.

Lo studio è stato condotto per diverse settimane, il che nella maggior parte dei casi avrebbe rappresentato una minaccia alla validità del costrutto. Tuttavia, poiché si tratta di un quasi-esperimento, la minaccia è minore. Non c'è possibilità di imbrogliare con il background educativo. Gli studenti sono stati informati sull'uso futuro dei dati raccolti, ma non è stato ottenuto il consenso esplicito, cosa che sarebbe stata preferibile.

Nell'analisi sono stati rimossi sei punti dati poiché non seguivano in modo coerente il processo di sperimentazione. Altri tre punti dati esterni alle code dei box-plot sono stati analizzati per essere considerati valori anomali. È stato rimosso solo un valore estremo, poiché avrebbe avuto un impatto significativo sulla deviazione standard e quindi avrebbe influenzato l'esito dell'analisi.

Capitolo

13 Le prospettive sono davvero diverse?: Ulteriori sperimentazioni sulla lettura dei requisiti basata su scenari¹

Background Questo capitolo presenta uno studio sperimentale così come è stato pubblicato, con l'obiettivo di mostrare un articolo di esempio tratto da una rivista internazionale. Inoltre, l'intenzione è che funzioni come uno studio adatto su cui esercitare le capacità di revisione. È importante notare che l'articolo è stato rivisto e rivisto in base al feedback prima di essere pubblicato sulla rivista *Empirical Software Engineering*.

Ciò significa che la qualità è superiore alla media degli articoli sperimentali presentati, sebbene anche gli standard degli articoli siano aumentati nel tempo rispetto alla pubblicazione originale. La revisione degli articoli scientifici è ulteriormente elaborata nell'Appendice A.2.

La lettura basata sulla prospettiva **astratta** (PBR) è una tecnica di ispezione basata su scenari in cui diversi revisori leggono un documento da diverse prospettive (ad esempio utente, progettista, tester). La lettura viene effettuata secondo uno scenario speciale, specifico per ciascuna prospettiva. Il presupposto di base alla base del PBR è che le prospettive trovino difetti diversi e una combinazione di più prospettive rilevi più difetti rispetto alla stessa quantità di lettura con un'unica prospettiva. Questo articolo presenta uno studio che analizza le differenze nelle prospettive. Lo studio è una replica parziale di studi precedenti. Si svolge in un ambiente accademico utilizzando studenti laureati come materie. Ciascuna prospettiva applica una tecnica di modellazione specifica: modellazione dei casi d'uso per la prospettiva dell'utente, partizionamento delle equivalenze per la prospettiva del tester e analisi strutturata per la prospettiva del design. Un totale di 30 soggetti sono stati divisi in 3 gruppi, fornendo 10 soggetti per prospettiva. I risultati dell'analisi mostrano che (1) non vi è alcuna differenza significativa tra le tre prospettive in termini di tasso di rilevamento dei difetti e numero di difetti rilevati all'ora, (2) non vi è alcuna differenza significativa nella copertura dei difetti delle tre prospettive e (3) uno studio di simulazione mostra che 30 soggetti sono sufficienti per rilevare differenze prospettive relativamente piccole con il test statistico scelto. I risultati suggeriscono che una combinazione di più prospettive potrebbe non fornire una maggiore copertura dei difetti

¹Questo capitolo è stato originariamente pubblicato in *Empirical Software Engineering: An International Journal*, vol. 5, n. 4, pp. 331–356 (2000).

rispetto alla lettura da una sola prospettiva, ma sono necessari ulteriori studi per aumentare la comprensione della differenza prospettica.

13.1 Introduzione

La convalida dei documenti dei requisiti viene spesso eseguita manualmente, poiché i documenti dei requisiti normalmente includono rappresentazioni informali di ciò che è richiesto da un sistema software previsto. Una tecnica comunemente utilizzata per la validazione manuale dei documenti software sono le ispezioni, proposte da Fagan [54]. Le ispezioni possono essere eseguite in diversi modi e utilizzate durante tutto il processo di sviluppo del software per (1) comprendere, (2) individuare difetti e (3) come base per prendere decisioni. Le ispezioni vengono utilizzate per individuare difetti nelle prime fasi del processo di sviluppo e hanno dimostrato di essere economicamente vantaggiose (ad esempio da Doolan [45]).

Una parte centrale del processo di ispezione è l' *individuazione dei difetti* effettuata da un singolo revisore che legge il documento e registra i difetti (una parte della preparazione, vedere Humphrey [81]). Tre tecniche comuni per il rilevamento dei difetti sono la lettura ad hoc, la lista di controllo e la lettura basata su scenari [137]. Il rilevamento ad hoc denota una tecnica non strutturata che non fornisce alcuna guida, il che implica che i revisori rilevano i difetti in base alla loro conoscenza ed esperienza personale. La tecnica di rilevamento della checklist fornisce un elenco di problemi e domande, acquisendo la conoscenza delle ispezioni precedenti, aiutando i revisori a focalizzare la loro lettura. Nell'approccio basato su scenari, diversi revisori hanno responsabilità diverse e sono guidati nella loro lettura da scenari specifici che mirano a costruire un modello, invece che da una semplice lettura passiva.

Uno scenario² qui denota uno script o una procedura che il revisore dovrebbe seguire. Sono state proposte due varianti di lettura basata su scenari: lettura basata su difetti [137] e lettura basata sulla prospettiva [18]. Il primo (in seguito denominato DBR) si concentra su specifiche classi di difetti, mentre il secondo (in seguito denominato PBR) si concentra sui punti di vista degli utenti di un documento.

Un'altra parte del processo di ispezione è la *compilazione dei difetti* in un elenco di difetti consolidato in cui vengono combinati tutti gli elenchi di difetti dei singoli revisori.

Questa fase può includere la rimozione di falsi positivi (difetti segnalati che non erano considerati difetti effettivi) nonché il rilevamento di nuovi difetti. Questo passaggio viene spesso eseguito in una *riunione di ispezione* strutturata a cui partecipa un *team* di revisori. L'efficacia dell'incontro di squadra è stata messa in discussione e studiata empiricamente da Votta [175] e Johnson e Tjahjono [87].

²Vi è qui un notevole rischio di confusione terminologica, poiché il termine *scenario* viene utilizzato anche nell'ingegneria dei requisiti per denotare una sequenza di eventi coinvolti in una situazione di utilizzo prevista del sistema in fase di sviluppo. Si dice spesso che un *caso d'uso* copra una serie di scenari correlati (di utilizzo del sistema). Nella lettura basata su scenari, tuttavia, il termine scenario è un concetto di meta-livello, che denota una procedura che il lettore di un documento dovrebbe seguire durante l'ispezione.

Questo articolo descrive la ricerca sulla lettura basata su scenari con un approccio PBR. Il metodo di ricerca è empirico e include un esperimento fattoriale formale in un ambiente accademico. L'esperimento presentato è una replica parziale di precedenti esperimenti nell'area e si concentra su ipotesi raffinate riguardanti le differenze tra le prospettive nel PBR. Il documento si concentra sul rilevamento dei difetti da parte dei singoli revisori, mentre gli aspetti relativi alle riunioni del team non sono inclusi.

La struttura del documento è la seguente. La sezione 13.2 fornisce una panoramica del lavoro correlato riassumendo i risultati degli esperimenti precedentemente condotti nelle ispezioni dei requisiti con un approccio basato su scenari. La sezione 13.3 include la dichiarazione del problema che motiva il lavoro presentato. Nella sez. 13.4 viene descritto il piano dell'esperimento, inclusa una discussione sulle minacce alla validità dello studio, e la sez. 13.5 riporta il funzionamento dell'esperimento. I risultati dell'analisi sono riportati nella Sez. 13.6 e sez. 13.7 include un'interpretazione dei risultati.

La sezione 13.8 fornisce una sintesi e le conclusioni.

13.2 Lavoro correlato

La letteratura esistente sull'ingegneria del software empirica include una serie di studi relativi alle ispezioni, dove la sperimentazione formale ha dimostrato di essere una strategia di ricerca rilevante [178]. L'esperimento presentato in questo documento si riferisce a precedenti esperimenti sulle ispezioni con un approccio basato su scenari. I risultati di una serie di esperimenti sull'ispezione basata su scenari dei documenti dei requisiti sono riepilogati di seguito.

1. Lo studio *del Maryland-95* [137] ha confrontato DBR con Ad Hoc e Checklist in un ambiente accademico. L'esperimento è stato eseguito due volte con 24 soggetti in ciascuna esecuzione. I documenti sui requisiti utilizzati erano un sistema di monitoraggio del livello dell'acqua (WLMS, 24 pagine) e un sistema di controllo della velocità per automobili (CRUISE, 31 pagine).

Risultato 1: i revisori DBR hanno tassi di rilevamento dei difetti significativamente più elevati rispetto ai revisori Ad Hoc o Checklist.

Risultato 2: i revisori DBR hanno tassi di rilevamento significativamente più elevati per i difetti che gli scenari sono stati progettati per scoprire, mentre tutti e tre i metodi hanno tassi di rilevamento simili per altri difetti.

Risultato 3: i revisori delle liste di controllo *non* hanno tassi di rilevamento significativamente più alti rispetto ai revisori ad hoc.

Risultato 4: gli incontri di riscossione *non* producono alcun miglioramento netto nel tasso di rilevamento: i guadagni degli incontri sono compensati dalle perdite degli incontri.

2. Lo studio *della NASA* [18] ha confrontato PBR con Ad Hoc in un ambiente industriale. L'esperimento consisteva in uno studio pilota con 12 soggetti e una seconda fase principale con 13 soggetti. Sono stati utilizzati due gruppi di documenti relativi ai requisiti; documenti requisiti generali: uno sportello automatico (bancomat, 17 pagine),

un sistema di controllo del parcheggio (PG, 16 pagine); e due documenti sui requisiti delle dinamiche di volo (27 pagine ciascuno).

Risultato 1: gli individui che applicano PBR a documenti generali hanno ottenuto risultati significativi tassi di rilevamento più elevati rispetto ad Ad Hoc.

Risultato 2: gli individui che applicano PBR a documenti specifici della NASA non hanno tassi di rilevamento significativamente più alti rispetto ad Ad Hoc.

Risultato 3: i team simulati che applicano PBR a documenti generali hanno risultati significativi Tassi di rilevamento leggermente più alti rispetto ad Ad Hoc.

Risultato 4: i team simulati che applicano PBR a documenti specifici della NASA hanno tassi di rilevamento significativamente più elevati rispetto ad Ad Hoc.

Risultato 5: i revisori con più esperienza non *hanno* tassi di rilevamento più elevati.

3. Lo studio di *Kaiserslautern* [34] ha confrontato PBR con Ad Hoc in un ambiente accademico utilizzando i documenti ATM e PG dello studio della NASA. L'esperimento consisteva in due sessioni rispettivamente con 25 e 26 soggetti.

Risultato 1: gli individui che applicano PBR a documenti generali hanno ottenuto risultati significativi tassi di rilevamento più elevati rispetto ad Ad Hoc.

Risultato 2: i team simulati che applicano PBR a documenti generali hanno risultati significativi Tassi di rilevamento leggermente più alti rispetto ad Ad Hoc.

Risultato 3: i tassi di rilevamento di cinque diverse classi di difetti *non* sono significativi non molto diverse tra le prospettive.

4. Lo studio di *Bari* [61] ha confrontato DBR con Ad Hoc e Checklist in un ambiente accademico utilizzando i documenti WLMS e CRUISE dello studio Maryland-95. L'esperimento è stato eseguito con 30 soggetti.

Risultato 1: DBR *non* ha registrato tassi di rilevamento dei difetti significativamente più elevati rispetto ad Ad Hoc o Checklist.

Risultato 2: i revisori DBR *non* hanno riscontrato tassi di rilevamento significativamente più elevati per i difetti che gli scenari erano stati progettati per scoprire, mentre tutti e tre i metodi avevano tassi di rilevamento simili per altri difetti.

Risultato 3: i revisori delle liste di controllo *non* hanno riscontrato tassi di rilevamento significativamente più elevati rispetto ai revisori ad hoc.

Risultato 4: le riunioni di raccolta *non* hanno prodotto alcun miglioramento netto nel rilevamento tasso: soddisfare i guadagni laddove compensati dalle perdite.

5. Lo studio di *Trondheim* [164] ha confrontato la versione del PBR dello studio della NASA con una versione modificata del PBR (di seguito denominata PBR2) in cui ai revisori venivano fornite maggiori istruzioni su come applicare la lettura basata sulla prospettiva. Lo studio è stato condotto in un ambiente accademico utilizzando i documenti ATM e PG dello studio della NASA. L'esperimento consisteva in una corsa con 48 soggetti.

Risultato 1: i revisori di PBR2 *non* hanno riscontrato tassi di rilevamento dei difetti significativamente più elevati rispetto a PBR.]

Risultato 2: gli individui che hanno applicato il PBR2 hanno esaminato un tempo significativamente più lungo rispetto a coloro che hanno applicato il PBR.

Risultato 3: gli individui che hanno applicato PBR2 hanno suggerito un potenziale significativamente inferiore difetti rispetto a chi ha applicato il PBR.

Risultato 4: gli individui che hanno applicato il PBR2 hanno avuto produttività ed efficienza significativamente inferiori rispetto a quelli che hanno applicato il PBR.

6. Lo studio di *Strathclyde* [124] ha confrontato la DBR con la Checklist in un ambiente accademico utilizzando i documenti WLMS e CRUISE dello studio del Maryland.

L'esperimento consisteva in una corsa con 50 soggetti.

Risultato 1: Nel documento WLMS, DBR *non* presentava tassi di rilevamento dei difetti significativamente più elevati rispetto alla Checklist.

Risultato 2: Nel documento CRUISE, DBR aveva tassi di rilevamento dei difetti significativamente più elevati rispetto alla Checklist.

Risultato 3: le riunioni di raccolta *non* hanno prodotto alcun miglioramento netto nel rilevamento tasso: i guadagni della riunione sono stati compensati dalle perdite della riunione.

7. Lo studio *Linköping* [147] ha confrontato DBR con Checklist in un ambiente accademico utilizzando i documenti WLMS e CRUISE dello studio del Maryland.

Altri difetti sono stati aggiunti all'elenco dei difetti totali. L'esperimento consisteva in una corsa con 24 soggetti.

Risultato 1: i revisori DBR *non* hanno avuto tassi di rilevamento dei difetti significativamente più alti rispetto ai revisori della checklist.

Risultato 2: i revisori DBR *non* hanno avuto tassi di rilevamento significativamente più alti rispetto a Revisori della lista di controllo.

8. Lo studio *del Maryland-98* [152] ha confrontato PBR con Ad Hoc in un ambiente accademico utilizzando i documenti ATM e PG dello studio del Maryland. L'esperimento consisteva in una corsa con 66 soggetti.

Risultato 1: i revisori PBR hanno avuto tassi di rilevamento dei difetti significativamente più elevati rispetto a Revisori ad hoc.

Risultato 2: gli individui con elevata esperienza nell'applicazione del PBR *non* hanno avuto tassi di rilevamento dei difetti significativamente più alti rispetto ad Ad Hoc.

Risultato 3: gli individui con un'esperienza media nell'applicazione del PBR hanno avuto risultati significativi tassi di rilevamento dei difetti più elevati rispetto ad Ad Hoc.

Risultato 4: gli individui con bassa esperienza nell'applicazione del PBR hanno avuto risultati significativi tassi di rilevamento dei difetti più elevati rispetto ad Ad Hoc.

Risultato 5: gli individui che hanno applicato PBR hanno avuto una produttività significativamente inferiore rispetto a coloro che hanno applicato Ad Hoc.

9. Lo studio *Lucent* [138] ha replicato lo studio Maryland-95 in un ambiente industriale utilizzando 18 sviluppatori professionisti della Lucent Technologies. IL

3I risultati 2–4 dello studio Maryland-98 applicano un livello di significatività di 0,10, mentre 0,05 è il livello di significatività scelto in tutti gli altri risultati.

Tabella 13.1 Riepilogo degli studi

Studio	Scopo	Materie ambientali	Significativo?
Maryland-95 DBR vs. AdHoc e checklist Academic	DBR vs. AdHoc e checklist Academic	24+24	sì
checklist Academic		30	NO
Strathclyde	DBR rispetto alla lista di controllo	Accademico	50
Collegamento	DBR rispetto alla lista di controllo	Accademico	24
Lucente	DBR vs. AdHoc e lista di controllo	Industriale	18
NASA	PBR contro AdHoc	Industriale	12+13
Kaiserslautern PBR contro AdHoc		Accademico	25+26
Trondheim	PBR contro PBR2	Accademico	48
Maryland-98	PBR contro AdHoc	Accademico	66

la replica ha avuto successo e ha completamente corroborato i risultati di Studio Maryland-95.

I risultati dei diversi studi variano sostanzialmente. Un tentativo di affrontare in modo sistematico la conoscenza combinata, acquisita da esperimenti e repliche è riportato da Hayes [74], dove la meta-analisi viene applicata ai risultati del Studi Maryland-95, Bari, Strathclyde, Linkoping e Lucent. Dalla meta-analisi si conclude che le dimensioni degli effetti per i metodi di ispezione non sono omogenee attraverso gli esperimenti. Gli studi Maryland-95 e Lucent mostrano risultati molto simili risultati, e un'interpretazione della meta-analisi identifica le caratteristiche che renderli diversi dagli altri tre studi: (1) sono condotti in un contesto dove i soggetti hanno più familiarità con la notazione utilizzata, (2) vengono condotti negli Stati Uniti, dove il controllo automatico della velocità è più comune nelle auto, che in Europa, dove il vengono eseguiti altri tre studi. Queste ipotesi però non sono realizzabili testare con i dati forniti, quindi sono necessarie ulteriori sperimentazioni.

La tabella 13.1 include un riepilogo degli studi presentati. Il Maryland-95, Gli studi della NASA, Kaiserslautern, Maryland-98 e Lucent indicano che un approccio basato su scenari fornisce un tasso di rilevamento più elevato. Gli studi di Bari, Strathclyde e Linkoping potrebbero, tuttavia, non corroborare questi risultati, il che motiva ulteriormente studi per aumentare la comprensione della lettura basata su scenari.

Molti studi hanno concluso che i veri incontri di squadra erano inefficaci in termini di rilevamento dei difetti. (Ci possono ovviamente essere altri buoni motivi per dirigere riunioni di gruppo oltre al rilevamento dei difetti, come la costruzione del consenso e la competenza condivisione e processo decisionale).

Lo studio qui presentato viene successivamente denominato studio *di Lund*. Il Lund lo studio è una replica parziale dello studio della NASA e si basa su un pacchetto di laboratorio [19] forniti dall'Università del Maryland per supportare le indagini empiriche della lettura basata su scenari. L'affermazione del problema che motiva lo studio di Lund è riportato nella sezione successiva.

13.3 Domande di ricerca

Gli studi precedenti, riassunti nella Sez. 13.2, si sono concentrati principalmente sul confronto della lettura basata su scenari con liste di controllo e tecniche ad hoc in termini di tassi di rilevamento dei difetti. L'obiettivo dello studio di Lund è, tuttavia, quello di indagare il presupposto di base alla base della lettura basata sugli scenari, secondo cui le diverse prospettive trovano difetti diversi. Un altro interesse è l'efficienza delle diverse prospettive in termini di difetti rilevati all'ora. Vengono affrontate le seguenti due domande:

1. Le prospettive rilevano difetti diversi?
2. Una prospettiva è superiore a un'altra?

Vengono affrontati due aspetti di superiorità: *l'efficacia*, ovvero quanta percentuale dei difetti esistenti viene rilevata (tasso di rilevamento), e *l'efficienza*, ovvero quanti difetti vengono rilevati per unità di tempo.

Le prospettive proposte da Basili et al. [18] sono progettista, tester e utente.

Gli utenti sono parti interessate importanti nel processo di sviluppo del software, e soprattutto quando i requisiti vengono individuati, analizzati e documentati. Il ruolo dell'utente in PBR è focalizzato sul rilevamento dei difetti ad alto livello di astrazione legati all'utilizzo del sistema, mentre il progettista si concentra sulle strutture interne e il tester si concentra sulla verifica.

Gli studi precedenti si sono concentrati principalmente sull'efficacia in termini di tasso di rilevamento. Da un punto di vista dell'ingegneria del software è importante anche valutare l'efficienza (ad esempio in termini di difetti rilevati per unità di tempo), poiché questo fattore è importante per la decisione di un professionista di introdurre una nuova tecnica di lettura.

I vincoli specifici del progetto e del dominio applicativo possono quindi, insieme alle stime dello sforzo necessario, costituire la base per un compromesso tra qualità e costo.

Uno degli scopi principali del PBR è che le prospettive rilevino diversi tipi di difetti al fine di ridurre al minimo la sovrapposizione tra i revisori. Quindi, una domanda naturale è se i revisori trovano o meno difetti diversi. Se rilevano gli stessi difetti, la sovrapposizione non viene ridotta al minimo e il PBR non funziona come previsto. Se tutte le prospettive riscontrano gli stessi tipi di difetti, potrebbe essere il risultato di (1) che l'approccio di lettura basato su scenari è inappropriate, (2) che le prospettive potrebbero non essere sufficientemente supportate dagli scenari di accompagnamento, o (3) che altre prospettive sono necessarie per ottenere una maggiore differenza di copertura. La soluzione ottimale è utilizzare prospettive senza sovrapposizioni e il più alto tasso di rilevamento dei difetti possibile, rendendo il PBR altamente affidabile ed efficace. Lo studio di Lund affronta la sovrapposizione indagando se le prospettive rilevano difetti diversi.

La domanda di ricerca 1 è interessante anche dal punto di vista della stima del contenuto di difetti. L'approccio di cattura-ricattura per la stima del contenuto dei difetti utilizza la sovrapposizione tra i difetti rilevati dai revisori per stimare il numero di difetti rimanenti in un artefatto software [51, 120]. La robustezza della cattura-ricattura utilizzando PBR è studiata da Thelin e Runeson [167], con l'obiettivo di indagare gli stimatori di cattura-ricattura applicati alle ispezioni PBR sotto l'ipotesi che PBR

funziona secondo il suo presupposto di base. Nello studio di Lund si esamina se le ipotesi del PBR siano reali. Pertanto, lo studio di Lund e lo studio di Thelin e Runeson [167] si completano a vicenda per rispondere alla domanda se le stime di cattura-ricattura possano essere utilizzate per le ispezioni PBR.

13.4 Pianificazione dell'esperimento

Questa sezione descrive la pianificazione dell'esperimento di lettura. La pianificazione include la definizione delle variabili dipendenti e indipendenti, le ipotesi da testare nell'esperimento, la progettazione dell'esperimento, la strumentazione e un'analisi delle minacce alla validità dell'esperimento [178].

L'esperimento di lettura è condotto in un ambiente accademico con stretti rapporti con l'industria. I soggetti sono studenti del quarto anno dei programmi di Master in Informatica e Ingegneria e Ingegneria Elettrica presso l'Università di Lund.

13.4.1 Variabili

Le variabili indipendenti determinano i casi per i quali vengono campionate le variabili dipendenti. Lo scopo è quello di indagare diverse prospettive e metodi di lettura, applicati a due oggetti (documenti dei requisiti). Gli oggetti di ispezione sono gli stessi del pacchetto di laboratorio dell'Università del Maryland [19], e anche la progettazione e la strumentazione si basano su questo pacchetto di laboratorio. Le variabili dello studio sono riassunte nella Tabella 13.2 insieme a brevi spiegazioni.

13.4.2 Ipotesi

Si presume che la lettura basata sulla prospettiva fornisca ispezioni più efficienti, poiché diversi revisori adottano prospettive diverse riducendo la sovrapposizione dei difetti [18]. L'obiettivo dello studio è verificare empiricamente se queste ipotesi sono vere. Di conseguenza, le ipotesi relative alla performance di diverse prospettive sono riportate di seguito. Le tre ipotesi riguardano l'efficienza, l'efficacia e la distribuzione sulle prospettive.

- H0;EFF . Si presuppone che le prospettive abbiano la stessa efficienza di accertamento, cioè il numero di difetti riscontrati per ora di ispezione non è diverso per le varie prospettive.

Tabella 13.2 Variabili

	Nome	Valori	Descrizione
Indipendente variabili	PERSP	fU,T,Dg	Viene applicata una delle tre prospettive ciascun soggetto: Utente, Tester e Designer.
	DOC	fATM,PGg	Gli oggetti di ispezione sono due requisiti documenti: uno per un bancomat (ATM) e uno per un sistema di controllo del garage (PG). Il documento ATM è di 17 pagine e contiene 29 difetti. Il P.G il documento è di 16 pagine e ne contiene 30 difetti.
Controllato variabile	ESPERIENZA Ordinale		L'esperienza con l'utente, il tester, il design le prospettive sono misurate su cinque livelli scala ordinale e utilizzata nell'allocazione dei soggetti alle prospettive. (Vedere Sette. 13.4.3 e 13.6.4)
Dipendente variabili	TEMPO	Intero	Il tempo trascorso da ciascun revisore in la preparazione individuale è registrata da tutti i soggetti. L'unità di tempo utilizzata è minuti.
	DEF	Intero	Il numero di difetti rilevati da ciascuno il revisore viene registrato, escluso falso positivi. I falsi positivi lo sono rimossi dagli sperimentatori, in ordine per garantire che tutti i candidati difettosi lo siano trattati allo stesso modo.
	EFF	60*DEF/TIME	L'efficienza di ricerca difetti, ovvero il numero di difetti rilevati all'ora, è calcolato come (DEF*60)/TEMPO.
VALUTARE	DEF/TOT		L'efficacia della ricerca dei difetti, ovvero il frazione dei difetti riscontrati rispetto al totale numero di difetti (chiamato anche rilevamento tasso) viene calcolato come DEF diviso per il numero totale di difetti noti contenuti nei documenti esaminati.
TROVATO	Intero		Il numero di revisori appartenenti a a certa prospettiva, che hanno trovato a certo difetto in un documento specifico è registrato. Questa variabile viene utilizzata per analizzare le distribuzioni di ricerca dei difetti per prospettive diverse.

- H0;TARIFFA . Si presuppone che le prospettive abbiano la stessa efficacia o tassi di rilevamento, cioè la frazione di difetti identificati non è diversa per i vari prospettive.
- H0;TROVATO. Si presuppone che le prospettive trovino gli stessi difetti, cioè il le distribuzioni sui difetti riscontrati sono le stesse per le diverse prospettive.

Tabella 13.3 Esperimento
progetto

	PERSP		
	Utente	Progettista	Tester
	DOC ATM5	5	5
PAG	5	5	5

13.4.3 Progettazione

Per verificare queste ipotesi viene utilizzato un esperimento con disegno fattoriale [125] a due fattori (PERSP e DOC). Il progetto è riassunto nella Tabella 13.3. L'esperimento varia le tre prospettive su due documenti.

L'assegnazione di un soggetto a una delle tre prospettive PBR (U, D, T), è stato condotto sulla base dell'esperienza riportata (vedere par. 13.6.4), simile allo studio della NASA [18]. L'obiettivo della prospettiva basata sull'esperienza Il compito è garantire che ogni prospettiva ottenga un'equa distribuzione dei soggetti sperimentati, in modo che il risultato dell'esperienza sia influenzato dalla prospettiva differenza piuttosto che esperienza della differenza. Il questionario sull'esperienza richiesto ai soggetti di valutare la loro esperienza con ciascuna prospettiva su un ordinale di cinque livelli scala. I soggetti sono stati poi ordinati tre volte, ottenendo un elenco ordinato di soggetti per ogni prospettiva con i più esperti per primi. All'interno della stessa esperienza livello, i soggetti sono stati disposti in ordine casuale. Successivamente sono stati assegnati i soggetti alle prospettive selezionando un argomento in cima all'elenco delle prospettive e rimuovendolo soggetto negli altri elenchi prima di continuare con la prospettiva successiva in un round Robin Fashion iniziando con una prospettiva selezionata casualmente, finché tutti i soggetti non lo furono assegnata una prospettiva.

Gli strumenti dell'esperimento di lettura consistono in due documenti relativi ai requisiti e modelli di segnalazione di tempi e difetti. Questi strumenti sono presi dal pacchetto di laboratorio dell'Università del Maryland [19] e vengono riutilizzati con una quantità minima cambiamenti.

Il disegno fattoriale sopra descritto viene analizzato con la statistica descrittiva (bar grafici e box plot) e analisi della varianza (ANOVA) [125] per le ipotesi H0;EFF e H0;TASSO . Per l'ipotesi H0;FOUND viene utilizzato il test del Chi-quadrato [157]. insieme ad un'analisi di correlazione [144].

13.4.4 Minacce alla validità

La validità dei risultati ottenuti negli esperimenti dipende da fattori nella impostazioni dell'esperimento. È possibile dare priorità a diversi tipi di validità a seconda del obiettivo dell'esperimento. In questo caso vengono analizzate le minacce a quattro tipi di validità [37, 178]: validità di conclusione, validità interna, validità di costrutto ed esterna validità.

La validità delle conclusioni riguarda l'analisi statistica dei risultati e la composizione dei soggetti. In questo esperimento vengono applicate tecniche statistiche ben note che sono resistenti alle violazioni dei loro presupposti. Una minaccia generale alla validità delle conclusioni è, tuttavia, il basso numero di campioni, che può ridurre la capacità di rivelare modelli nei dati. In particolare, esistono pochi campioni per il test del Chi-quadrato, che è ulteriormente elaborato nella Sez. 13.6.3.

La validità interna riguarda questioni che possono influenzare la variabile indipendente rispetto alla causalità, all'insaputa del ricercatore. Ci sono due minacce alla validità interna di questo esperimento: la selezione e la strumentazione. L'esperimento era una parte obbligatoria di un corso di ingegneria del software, quindi la selezione dei soggetti non è casuale, il che comporta una minaccia alla validità dell'esperimento.

Anche i documenti relativi ai requisiti utilizzati possono influenzare i risultati. I documenti sono piuttosto soggetti a difetti e ulteriori problemi nei documenti potrebbero essere considerati difetti. D'altra parte, è preferibile avere la stessa definizione di difetti degli studi precedenti per ragioni di confronto. Altre minacce alla validità interna sono considerate piccole. Ad ogni soggetto è stato assegnato un solo oggetto e un solo trattamento, quindi non c'è pericolo di maturazione nell'esperimento.

I soggetti hanno applicato prospettive diverse durante l'ispezione, ma la differenza tra le prospettive non è abbastanza grande da sospettare un'equalizzazione compensativa dei trattamenti o una rivalità compensativa. Ai soggetti è stato inoltre detto che il loro voto nel corso non dipendeva dalla loro prestazione nell'esperimento, ma solo dalla loro seria frequenza. Naturalmente c'è il rischio che i soggetti manchino di motivazione; potrebbero, ad esempio, considerare la loro partecipazione una perdita di tempo o potrebbero non essere motivati ad apprendere le tecniche. L'insegnante del corso in cui è stato effettuato l'esperimento ha, tuttavia, fatto un forte sforzo nel motivare gli studenti. Si affermava chiaramente che per superare il corso era necessaria una partecipazione seria. È opinione dell'insegnante che gli studenti abbiano fatto un tentativo molto serio durante l'ispezione.

La validità di costrutto riguarda la generalizzazione del risultato dell'esperimento al concetto o alla teoria alla base dell'esperimento. Una grave minaccia alla validità di costrutto è che le prospettive scelte o le tecniche di lettura per le prospettive potrebbero non essere rappresentative o adatte alla lettura basata su scenari. Ciò limita la portata delle conclusioni tratte da queste particolari prospettive e tecniche. Altre minacce alla validità di costrutto sono considerate piccole. I soggetti non sapevano quali ipotesi fossero state formulate e non erano coinvolti in alcuna discussione sui vantaggi e svantaggi del PBR, quindi non erano in grado di indovinare quali sarebbero stati i risultati attesi erano.

La validità esterna riguarda la generalizzazione del risultato dell'esperimento ad ambienti diversi da quello in cui viene condotto lo studio. La più grande minaccia alla validità esterna è l'uso degli studenti come soggetti. Tuttavia, questa minaccia viene ridotta utilizzando gli studenti del quarto anno che sono prossimi a completare la loro istruzione e iniziare a lavorare nell'industria. L'ambientazione vuole assomigliare a una situazione di ispezione reale, ma il processo a cui partecipano i soggetti non è parte di un vero progetto di sviluppo software. Anche gli incarichi devono essere realistici,

ma i documenti sono piuttosto brevi e i documenti sui requisiti software reali possono includere molte più pagine. Le minacce alla validità esterna relative alle impostazioni e agli incarichi sono, tuttavia, considerate limitate, poiché sia il processo di ispezione che i documenti somigliano in misura ragionevole a casi reali.

Si può concludere che ci sono minacce al costrutto, alla validità interna ed esterna. Tuttavia, questi sono quasi gli stessi degli studi originali. Pertanto, finché le conclusioni dell'esperimento non vengono tratte al di fuori dei limiti di queste minacce, i risultati sono validi.

13.5 Funzionamento dell'esperimento

L'esperimento è stato condotto nella primavera del 1998. A tutti gli studenti è stata data una lezione introduttiva di 2 ore in cui è stata fornita una panoramica dello studio insieme ad una descrizione della classificazione dei difetti. È stato somministrato un questionario sull'esperienza e ad ogni soggetto è stata assegnata una prospettiva, come descritto nella Sez. 13.4.3. Gli studenti sono stati informati che l'esperimento era una parte obbligatoria del corso, ma la valutazione si basava solo sulla seria partecipazione allo studio e non sulla prestazione individuale degli studenti. È stato garantito l'anonimato degli studenti.

Si è svolto un esercizio di 2 ore, in cui le tre prospettive PBR sono state descritte e illustrate utilizzando un documento sui requisiti per un sistema di noleggio video (VRS). Durante la seconda ora dell'esercizio, i soggetti hanno messo in pratica la propria tecnica di lettura prospettica del documento VRS e hanno avuto l'opportunità di porre domande.

Durante l'esercizio sono stati spiegati e utilizzati anche i moduli di raccolta dati. La lettura prospettica del documento VRS è stata completata dagli studenti in autonomia dopo le ore di aula.

Le dispense dell'esperimento, distribuite durante l'esercitazione, includevano i seguenti strumenti di strumentazione:

1. Classificazione dei difetti che descrive le classi di difetti da utilizzare nell'elenco dei difetti.
2. Registro di registrazione del tempo per registrare il tempo trascorso nella lettura.
3. Elenco difetti per la registrazione dei difetti riscontrati.
4. Istruzioni di lettura, specifiche per il punto di vista dell'utente, del progettista e del tester rispettivamente.
5. Moduli di modellazione, specifici per le prospettive dell'utente, del progettista e del tester rispettivamente.
6. Il documento dei requisiti (ATM o PG).

Agli studenti è stato chiesto di non discutere i documenti ATM o PG e i difetti che riscontrano. È stato loro consentito di discutere le prospettive del PBR in relazione al documento VRS prima di iniziare con la raccolta dei dati vera e propria.

13.6 Analisi dei dati

In questa sezione viene presentata l'analisi statistica dei dati raccolti. I dati sono stati raccolti dalle consegne dei soggetti. Ogni difetto nel registro dei difetti di ciascun soggetto è stato confrontato con l'elenco dei difetti originale "corretto" fornito dal pacchetto di laboratorio dell'Università del Maryland. In un incontro, gli autori hanno discusso ogni difetto e hanno deciso se corrispondeva a un difetto "corretto". Se non veniva trovato alcun difetto "corretto" corrispondente, il difetto segnalato veniva considerato un falso positivo.⁴ È stato inoltre raccolto il tempo trascorso segnalato e sono state calcolate le misure EFF, RATE e FOUND. I set di dati totali sono riportati nelle Tabelle 13.6–13.8.

13.6.1 Prestazioni individuali per prospettive diverse

I box-plot⁵ delle prestazioni individuali in termini di numero di difetti rilevati all'ora (EFF) e la frazione di difetti riscontrati rispetto al numero totale di difetti (RATE), sono mostrati nella Fig. 13.1. I box-plot sono divisi per documento e prospettiva.

Per EFF, la prospettiva Tester sul documento PG ha una media più alta rispetto alle prospettive Utente e Designer, mentre per il documento ATM, la prospettiva Designer ha una media più alta. Per RATE le medie del Designer sono più elevate rispetto alle prospettive Utente e Tester per entrambi i documenti. Ci sono, tuttavia, troppo pochi punti dati per gruppo per qualsiasi ulteriore interpretazione dei box-plot, rispetto ai valori anomali e all'asimmetria.

Quando vengono misurate più variabili dipendenti, l'analisi multivariata della varianza (MANOVA) può essere utilizzata per valutare se esiste qualche differenza statisticamente significativa nell'insieme totale delle medie. I risultati dei test MANOVA riguardanti l'effetto del PERSP non rivelano alcun significato e indicano l'assenza di effetti di interazione.

Inoltre, non ci sono differenze significative nelle medie di EFF, RATE per la variabile PERSP, come mostrato dall'analisi della varianza (ANOVA) nelle Tabelle 13.4 e 13.5. Da questa analisi si può concludere che le ipotesi nulle per EFF e RATE non possono essere rifiutate per nessuna delle tre prospettive.

⁴Alcuni dei difetti ritenuti falsi positivi potrebbero in realtà essere veri difetti se l'elenco dei difetti del pacchetto del laboratorio del Maryland è incompleto. Si è tuttavia deciso che, dal punto di vista della replicabilità, è importante utilizzare lo stesso elenco di difetti "corretti". Si ritiene che questa decisione non abbia avuto alcun impatto significativo sul risultato poiché c'erano solo pochi falsi positivi discutibili.

⁵I box-plot sono disegnati con l'altezza della scatola corrispondente al 25° e al 75° percentile, con il 50° percentile (la mediana) segnato nella scatola. I baffi corrispondono al 10° e al 90° percentile.

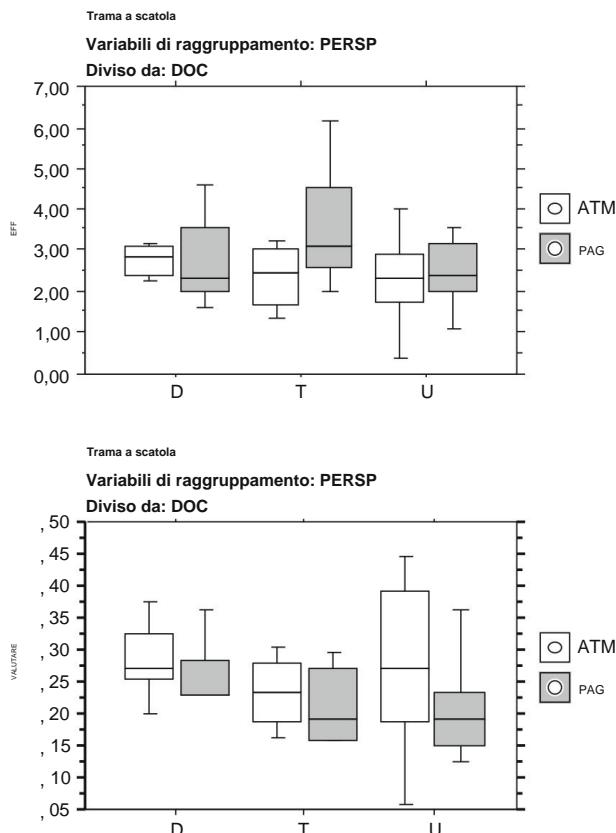


Fig. 13.1 Box plot per EFF e RATE divisi per DOC e PERSP

Tabella 13.4 Tabella ANOVA per EFF

	DF	Somma del valore F	quadrato medio	quadrato medio del quadrato	Valore p	Lambda	Potenza
PERSP	2	1.751	0,875	0,737	0,4893	1.473	0,156
DOC	1	1.640	1.640	1.380	0,2516	1.380	0,193
PERSP*DOC2		2.229	1.114	0,937	0,4055	1.875	0,187
Residuo	24	28.527	1.189				

Tabella 13.5 Tabella ANOVA per TASSO

	DF	Somma del valore F	quadrato medio	quadrato medio del quadrato	Valore p	Lambda	Potenza
PERSP	2	0,012	0,006	0,802	0,4602	1.604	0,166
DOC	1	0,011	0,011	1.488	0,2344	1.488	0,205
PERSP*DOC2		0,004	0,002	0,259	0,7739	0,518	0,085
Residuo	24	0,172	0,007				

13.6 Analisi dei dati

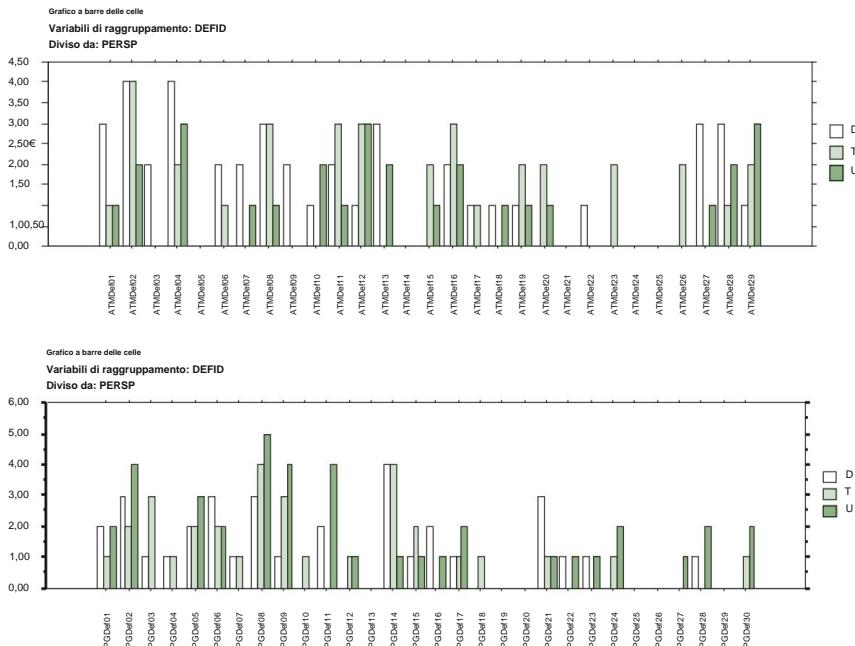


Fig. 13.2 Grafici a barre che illustrano la distribuzione del numero di revisori che hanno riscontrato ciascun difetto

13.6.2 Difetti riscontrati da diverse prospettive

In questa sezione viene studiata l'ipotesi H_0 : FOUND riguardante la sovrapposizione dei difetti riscontrati tra le prospettive. Le statistiche descrittive sotto forma di grafici a barre sono mostrate in Fig. 13.2. Per ciascun documento viene mostrata la distribuzione del numero di difetti riscontrati per prospettiva. Non sembrano esserci schemi particolari nelle diverse distribuzioni prospettiche; i risultati dei difetti di ciascuna prospettiva sembrano distribuiti in modo simile nello spazio dei difetti. Se ci fossero state grandi differenze nelle distribuzioni delle prospettive, il diagramma a barre avrebbe presumibilmente gruppi di difetti in cui una prospettiva avrebbe un numero elevato di risultati mentre le altre avrebbero un numero basso di risultati.

Per confrontare le distribuzioni dei difetti riscontrati per ciascuna prospettiva e indagare se esiste una differenza significativa tra i difetti rilevati dalle prospettive, viene creata una tabella di contingenza per la quale viene effettuato un test del Chi quadrato [157, pp. 191–194], come mostrato in Fig. 13.3. I difetti che nessuna prospettiva ha riscontrato sono esclusi dalle tabelle di contingenza (i “criteri di inclusione” nella Fig. 13.3), poiché questi casi non contribuiscono alla verifica delle differenze.

I valori p del Chi quadrato sono tutt'altro che significativi, il che indica che non è possibile con questo test e questo particolare set di dati mostrare una differenza nelle prospettive? distribuzioni per la ricerca dei difetti. Ci sono regole pratiche riguardo a quando il Chi Quadrato

Tabella riepilogativa per DEFID, PERSP
Criteri di inclusione: conteggi > 0 da PG.data

Num. Mancante	0
DF	46
Chi quadrato	33.951
Chi quadrato Valore P	,9058
G quadrato	*
G quadrato Valore P	*
Coef.	,494
V di Cramer	,402

Tabella riepilogativa per DEFID, PERSP
Criteri di inclusione: Conteggi > 0 da ATM.data

Num. Mancante	0
DF	46
Chi quadrato	41.676
Chi quadrato Valore P	,6538
G quadrato	*
G quadrato Valore P	*
Coef.	,535
V di Cramer	,448

Frequenze osservate per DEFID, PERSP Criteri di inclusione: conteggi > 0 da PG.data

	Totali DTU	
PGDef01	212	5
PGDef02	324	9
PGDef03	130	4
PGDef04	110	2
PGDef05	223	7
PGDef06	322	7
PGDef07	110	2
PGDef08	345	12
PGDef09	134	8
PGDef10	01	0
PGDef11	204	6
PGDef12	01	1
PGDef14	441	9
PGDef15	1 2	1
PGDef16	201	3
PGDef17	112	4
PGDef18	010	1
PGDef21	311	5
PGDef22	101	2
PGDef23	101	2
PGDef24	012	3
PGDef27	001	1
PGDef28	102	3
PGDef30	012	3
Totali	33 32 40	105

Frequenze osservate per DEFID, PERSP Criteri di inclusione: conteggi > 0 da ATM.data

	Totali DTU	
ATMDef01	311	5
ATMDef02	442	10
ATMDef03	200	2
ATMDef04	423	9
ATMDef06	210	3
ATMDef07	201	3
ATMDef08	331	7
ATMDef09	200	2
ATMDef10	102	3
ATMDef11	2 3	1
ATMDef12	133	7
ATMDef13	3 0	2
ATMDef15	02	1
ATMDef16	2 3	2
ATMDef17	110	2
ATMDef18	101	2
ATMDef19	121	4
ATMDef20	021	3
ATMDef22	100	1
ATMDef23	020	2
ATMDef26	020	2
ATMDef27	301	4
ATMDef28	312	6
ATMDef29	123	6
Totali	42 34 28	104

Fig. 13.3 Test Chi Quadrato e tabelle di contingenza per i difetti rilevati da U, T, D per DOC

si può usare un test [157, pp. 199–200], affermando che non più del 20% delle celle dovrebbe avere una frequenza prevista inferiore a 5, e nessuna cella dovrebbe avere una frequenza prevista inferiore a 1. Queste regole di pollici non sono soddisfatti dal set di dati in questo caso, ma si può sostenere che le regole sono troppo prudenti e poiché le frequenze attese nel nostro caso sono distribuite piuttosto uniformemente, il test del Chi quadrato può ancora essere valido (vedere più avanti la sezione 13.6 .3).

Documento ATM

Analisi di correlazione

	Correlazione valore p .480	95% inferiore	95% superiore
Utente, Tester	,0076 ,499 ,0052	,138	,720
Utente, progettista	,258 ,1789	,162	,732
Tester, progettista		-,120	,570

In questo calcolo sono state utilizzate 29 osservazioni.

Analisi di correlazione

Criteri di inclusione: Utente > 0 OR Tester > 0 OR Designer > 0 da ATM-ctable.data

	Valore p di correlazione	95% inferiore	95% superiore
Utente, Tester	,357	,0867	-,054
Utente, progettista	,352	,0915	-,059
Tester, progettista	,043	,8449	-,367

In questo calcolo sono state utilizzate 24 osservazioni.

Documento PG

Analisi di correlazione

	Valore p di correlazione	95% inferiore	95% superiore
Utente, Tester	,463	,0092	,123
Utente, progettista	,543	,0016	,228
Tester, progettista	,601	,0003	,307

In questo calcolo sono state utilizzate 30 osservazioni.

Analisi di correlazione

Criteri di inclusione: Utente > 0 OR Tester > 0 OR Designer > 0 da PG-ctable.data

	Correlazione valore p ,319	95% inferiore	95% superiore
Utente, Tester	,1300 ,414 ,0438		,640
Utente, progettista	,493 ,0134	-,097,012	,700
Tester, progettista		,112	,748

In questo calcolo sono state utilizzate 24 osservazioni.

Fig. 13.4 Analisi di correlazione delle prospettive per ciascun documento

Il test del Chi Quadrato non fornisce una misura del grado di differenza. Al fine per analizzare quanto diverse (o simili) sono le prospettive, è necessaria un'analisi di correlazione presentato nella Fig. 13.4, utilizzando il coefficiente di correlazione di Pearson [143, pp. 338–340].

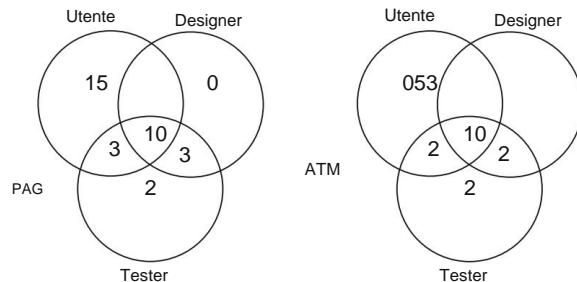
Per ciascun documento sono previste due diverse analisi di correlazione, una per tutti difetti "corretti" inclusi e uno in cui sono inclusi solo i difetti che lo erano trovato da almeno un recensore. Quest'ultima opzione potrebbe essere auspicata, come ci interessa le differenze nell'insieme dei difetti riscontrati da ciascuna prospettiva; i difetti che nessuna ricerca prospettica non contribuisce alle differenze tra le prospettive.

Il valore p indica se il coefficiente di correlazione è significativo e gli intervalli di confidenza presentati indicano l'intervallo in cui è il coefficiente di correlazione

probabilmente lo sarà.

L'analisi delle correlazioni indica che esistono correlazioni significativamente positive tra le prospettive, il che significa che quando una prospettiva trova un difetto è probabile

Fig. 13.5 Copertura dei difetti per i documenti PG e ATM



che anche gli altri lo trovino. L'unico coefficiente di correlazione tutt'altro che significativo è la correlazione Designer-Tester per il documento ATM.

Un altro modo per analizzare qualitativamente la sovrapposizione tra le prospettive sono i diagrammi di Venn, come utilizzati nello studio della NASA [18, p. 151].

A scopo di confronto includiamo tali diagrammi per i dati dello studio di Lund, come mostrato in Fig. 13.5. Ogni difetto è classificato in una delle sette classi a seconda delle combinazioni di prospettive che hanno una misura TROVATA maggiore di zero. I numeri nei diagrammi di Venn indicano quanti difetti appartengono a ciascuna classe. Ad esempio, per il documento PG, ci sono dieci difetti riscontrati da tutti i punti di vista, mentre cinque difetti sono stati rilevati sia dal punto di vista dell'utente che da quello del progettista e solo un difetto è stato riscontrato esclusivamente dal punto di vista dell'utente.

Questo tipo di analisi è molto sensibile al numero di soggetti. È sufficiente che un solo revisore trovi un difetto perché la classificazione cambi. La probabilità che venga trovato un difetto aumenta con il numero di revisori e, se disponiamo di un numero elevato di revisori, sarà più probabile che i difetti vengano inclusi nella classe in cui tutti i punti di vista lo hanno trovato. Ciò significa che questo tipo di analisi non è molto robusta e non fornisce interpretazioni significative nel caso generale. Nel nostro caso, possiamo almeno dire che l'analisi della copertura dei difetti nella Figura 13.5 non contraddice i nostri risultati precedenti e che non possiamo rifiutare l'ipotesi che le prospettive siano simili rispetto agli insiemi di difetti che trovano. I difetti riscontrati da tutte le prospettive rappresentano di gran lunga la classe più ampia.

13.6.3 La dimensione del campione è sufficientemente grande?

Il risultato dello studio di Lund è che non è possibile rilevare alcuna differenza significativa tra le prospettive. Sorge la domanda se ciò sia dovuto alla mancanza di differenze nei dati o al fatto che i test statistici non siano in grado di rivelare le differenze, ad esempio a causa della quantità limitata di dati. Per valutare il test Chi-quadrato, i set di dati di rilevamento dei difetti prospettici vengono simulati con variazioni stocastiche tra le prospettive e il test Chi-quadrato viene applicato ai dati simulati.

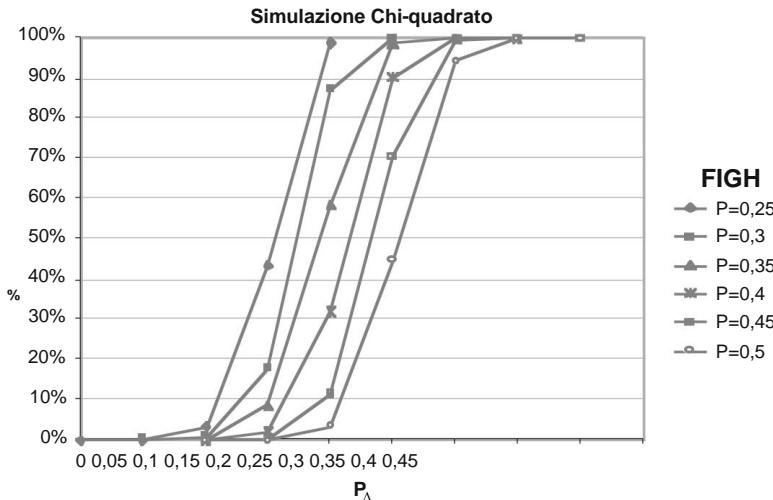


Fig. 13.6 Frazione dei risultati significativi dei test riguardanti H0;FOUND

La simulazione è progettata per assomigliare all'esperimento presentato nella sezione precedente. La differenza è che nel caso della simulazione, la probabilità di rilevamento di un difetto specifico da parte di una prospettiva è una variabile indipendente. Inoltre, viene applicata solo la variabile dipendente FOUND, poiché l'aspetto temporale non è modellato. Il modello di simulazione è progettato come segue:

- Il numero di difetti in ciascun documento simulato è 30. • Per ogni ispezione simulata, vengono utilizzate tre prospettive con dieci revisori per prospettiva. Si presuppone che un documento contenga tre diversi tipi di difetti, che hanno diverse probabilità di essere rilevati. Una prospettiva ha un'alta probabilità (*PHIGH*) di rilevare un terzo dei difetti e una bassa probabilità (*PLOW*) di rilevare gli altri due terzi dei difetti. La differenza tra *PHIGH* e *PLOW* è indicata con *P*. I livelli di probabilità sono impostati su valori compresi tra 0,05 e 0,5 in incrementi di 0,05, che sono valori attorno alla media misurata nello studio di Lund. • Vengono simulate 1.000 esecuzioni per ciascuna ispezione.

L'ipotesi *H0;FOUND* viene testata con il test del Chi-Quadro e i risultati sono presentati in Fig. 13.6. Ogni esperimento simulato viene testato separatamente. La figura mostra la frazione di test rifiutati per ciascun caso. Per tutti i casi di simulazione con *P* maggiore di 0,3, il test può mostrare una differenza significativa tra le prospettive simulate. Per casi simulati con *PHIGH* inferiore a 0,25, le differenze possono essere mostrate se *P* è maggiore di 0,2. I test sono condotti con un livello di significatività pari a 0,05. Lo studio di simulazione mostra che è possibile rilevare differenze nel FOUND con il test Chi-Square, anche se le differenze prospettiche sono piccole e la dimensione del campione è piccola.

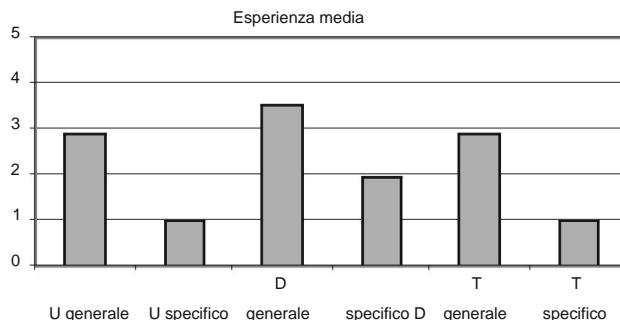


Fig. 13.7 Esperienza media dei soggetti riguardo alla loro esperienza generale della loro prospettiva ed esperienza specifica con la loro tecnica di modellazione

13.6.4 Esperienza dei soggetti

L'esperienza è stata misurata attraverso un questionario che copre ciascuna prospettiva in generale, nonché l'esperienza con le tecniche di modellazione specifiche delle tre prospettive (modellazione dei casi d'uso, partizionamento delle equivalenze e analisi strutturata). L'esperienza viene misurata per ogni prospettiva generale e ogni tecnica di modellazione specifica su una scala ordinale a cinque livelli: 1 D nessuno, 2 D studiato in classe o da un libro, 3 D praticato in un progetto in classe, 4 D utilizzato in un progetto nell'industria, 5 D utilizzato su più progetti nell'industria.

La Figura 13.7 mostra l'esperienza media di ciascun soggetto rispetto alla prospettiva a cui è stato assegnato, sia per la prospettiva in generale che per la specifica tecnica di modellazione.

Si può vedere che l'assegnazione dei soggetti (secondo l'algoritmo spiegato nella sezione 13.4.3) ha, come previsto, prodotto un profilo di esperienza relativamente equilibrato rispetto alle prospettive. Si può anche notare che gli studenti avevano pochissima esperienza industriale.

13.7 Interpretazioni dei risultati

In questa sezione l'analisi dei dati viene interpretata rispetto alle ipotesi formulate nella Sez. 13.4.2. Le prime due ipotesi vengono testate utilizzando ANOVA e la terza ipotesi viene testata utilizzando un test Chi-quadrato. Le seguenti tre ipotesi nulle non possono essere rifiutate:

- H0;EFF Si presuppone che le prospettive trovino lo stesso numero di difetti per ora. Questa ipotesi *non* può essere respinta.
- H0;RATE Si presuppone che le prospettive trovino lo stesso numero di difetti. Questo ipotesi *non* può essere respinta.

- H0; FOUND Si presuppone che le prospettive trovino gli stessi difetti. Questa ipotesi non può essere respinta.

Si può quindi concludere che non vi è alcuna differenza significativa tra le tre prospettive, utente, progettazione e test. Ciò vale per tutte e tre le ipotesi, ovvero non vi è alcuna differenza significativa in termini di efficacia o efficienza. Inoltre, non vi è alcuna differenza significativa nel tempo trascorso utilizzando le diverse prospettive, quindi il tempo trascorso non è favorevole a nessuna delle tecniche. La mancanza di differenza tra le tre prospettive, se il risultato è replicabile e generalizzabile, pregiudica gravemente i capisaldi della PBR. Si presuppone che i vantaggi del PBR risiedano nel fatto che le diverse prospettive si concentrano su diversi tipi di difetti e quindi rilevano diversi insiemi di difetti. Questo studio non mostra alcuna differenza statisticamente significativa tra gli insiemi di difetti riscontrati dalle tre prospettive, e quindi i vantaggi del PBR possono essere messi in discussione.

Le minacce alla validità delle conclusioni dei risultati sono rappresentate dal basso numero di campioni, in particolare per il test Chi-quadrato. Tuttavia, uno studio di simulazione rivela che il test Chi-quadrato può rilevare differenze tra le prospettive su 30 soggetti per differenze relativamente piccole nella probabilità di rilevamento. Inoltre, i grafici a barre sui difetti riscontrati da diverse prospettive (vedi Fig. 13.2) non indicano alcun modello chiaro, che supporti i risultati non significativi. Le statistiche ANOVA vengono applicate entro limiti accettabili e non mostrano alcuna differenza tra le prospettive. Le prospettive specifiche e le tecniche di lettura delle prospettive potrebbero anche rappresentare una minaccia alla validità dei risultati, quando si tenta di applicare i risultati alla lettura basata su scenari in generale.

La minaccia alla validità relativa alla motivazione dei soggetti può essere valutata confrontando i tassi di rilevamento dello studio di Lund con altri studi. Il tasso di rilevamento PBR individuale per lo studio della NASA [18] era in media di 0,249 per lo studio pilota e di 0,321 per quello principale, mentre lo studio di Lund mostra un tasso di rilevamento PBR individuale medio di 0,252. I tassi sono comparabili, supportando l'ipotesi che i soggetti di questo studio fossero motivati quanto nello studio della NASA.

Altre minacce alla validità della Sez. 13.4.4 non vengono considerati diversamente alla luce del risultato.

13.8 Riepilogo e Conclusioni

Lo studio riportato in questo documento è focalizzato sulla valutazione della lettura basata sulla prospettiva (PBR) dei documenti dei requisiti. Lo studio è una replica parziale di precedenti esperimenti in un ambiente accademico basato sul pacchetto di laboratorio dell'Università del Maryland [19].

L'obiettivo dello studio presentato è duplice:

1. Esaminare le differenze nelle prestazioni delle prospettive in termini di efficacia (tasso di rilevamento dei difetti) ed efficienza (numero di difetti rilevati all'ora).

2. Investigare le differenze nella copertura dei difetti delle diverse prospettive, e quindi valutare i presupposti di base alla base del PBR supponendo che prospettive diverse trovino difetti diversi.

L'impostazione dell'esperimento include due documenti di requisiti e scenari per tre prospettive (*l'utente* che applica la modellazione dei casi d'uso, *il progettista* che applica l'analisi strutturata e *il tester* che applica il partizionamento di equivalenza). Un totale di 30 studenti di Master sono stati divisi in 3 gruppi, assegnando 10 materie per prospettiva.

In sintesi i risultati dell'analisi dei dati mostrano che:

1. Non vi è alcuna differenza significativa tra il punto di vista dell'utente, del progettista e del tester in termini di tasso di rilevamento dei difetti e numero di difetti rilevati all'ora.
2. Non vi è alcuna differenza significativa nella copertura dei difetti delle tre prospettive.

L'interpretazione di questi risultati suggerisce che una combinazione di più prospettive potrebbe non fornire una maggiore copertura dei difetti rispetto alla lettura con una sola prospettiva.

I risultati contraddicono le principali ipotesi alla base del PBR. Alcuni degli studi precedenti, riassunti nella Sez. 13.2, hanno mostrato vantaggi significativi con la lettura basata su scenari rispetto all'ispezione ad hoc, ma in nessuno degli studi riportati nella sez. 13.2.

Inoltre, gli studi precedenti nella Sez. 13.2 non hanno preso in considerazione l'efficienza (numero di difetti rilevati all'ora), ma si concentrano sul tasso di rilevamento come principale variabile dipendente. Dal punto di vista dell'ingegneria del software, dove il costo e l'efficienza di un metodo sono di interesse centrale, è molto interessante studiare non solo il tasso di rilevamento, ma anche se un metodo può funzionare bene con uno sforzo limitato.

Esistono numerose minacce alla validità dei risultati, tra cui:

1. L'ambientazione potrebbe non essere realistica.
2. Le prospettive potrebbero non essere ottimali.
3. I soggetti potrebbero non essere motivati o formati a sufficienza.
4. Il numero di soggetti potrebbe essere troppo piccolo.

Si può sostenere che le minacce alla validità sono sotto controllo, sulla base delle seguenti considerazioni: (1) gli oggetti di ispezione sono simili ai documenti sui requisiti industriali; (2) Le prospettive sono motivate da una visione del processo di ingegneria del software; (3) I soggetti erano studenti del quarto anno con un interesse speciale per l'ingegneria del software che frequentavano un corso opzionale che avevano scelto per proprio interesse e, inoltre, molte aziende hanno una grande percentuale di dipendenti con nuovi esami; (4) Lo studio di simulazione presentato mostra che è possibile rilevare differenze relativamente piccole tra le prospettive con l'analisi scelta per un dato numero di punti dati.

Un singolo studio, come questo, non è una base sufficiente per cambiare l'atteggiamento nei confronti del PBR. Condurre le stesse analisi sui dati degli esperimenti esistenti e su nuove repliche con lo scopo di valutare le differenze tra le prospettive porterà maggiore chiarezza sui vantaggi e sugli svantaggi delle tecniche PBR e darà anche un migliore controllo sulle minacce alla validità.

13.9 Dati sulla prestazione individuale

Tabella 13.6 Dati per ciascun soggetto

Id	Prospettiva	Documento	Tempo	Difetti	Efficienza	Valutare
1	U	ATM	187	8	2.567	0,276
2	D	PAG	150	8	3.200	0,267
3	T	ATM	165	9	3.273	0,310
4	U	PAG	185	11	3.568	0,367
5	D	ATM	155	8	3.097	0,276
6	T	PAG	121	8	3.967	0,267
7	U	ATM	190	7	2.211	0,241
8	D	PAG	260	7	1.615	0,233
9	T	ATM	123	6	2.927	0,207
10	U	PAG	155	6	2.323	0,200
11	D	ATM	210	11	3.143	0,379
12	T	PAG	88	9	6.136	0,300
13	U	ATM	280	11	2.357	0,379
14	D	PAG	145	11	4.552	0,367
15	T	ATM	170	5	1.765	0,172
16	U	PAG	120	6	3.000	0,200
17	D	ATM	190	9	2.842	0,310
18	T	PAG	97	5	3.093	0,167
19	U	ATM	295	2	0.407	0,069
20	D	PAG	180	7	2.333	0,233
21	T	ATM	306	7	1.373	0,241
22	U	PAG	223	4	1.076	0,133
23	D	ATM	157	6	2.293	0,207
24	T	PAG	130	6	2.769	0,200
25	U	ATM	195	13	4.000	0,448
26	D	PAG	200	7	2.100	0,233
27	T	ATM	195	8	2.462	0,276
28	U	PAG	125	5	2.400	0,167
29	D	ATM	200	8	2.400	0,276
30	T	PAG	150	5	2.000	0,167

13.10 Dati sui difetti rilevati da Perspectives

13.10.1 Documento PG

Tabella 13.7 Difetti id D# riscontrati (1) o non riscontrati (0) dai soggetti che leggono il documento PG

Individui prospettiva dell'utente	Tester prospettiva	Designer prospettiva	
RE#2 8 14 20 26	S 4 10 16 22 28	S 6 12 18 24 30	S
1001012100001100102			
2111014101002011103			
3000000110103000101			
4 000000000101100001			
5001113100012011002			
6110002011002100113			
7 000000100001010001			
8111115111104110103			
9111104111003010001			
10 000000001001000000			
11110114000000011002			
12100001100001000000			
13 00000000000000000000			
14 000011110114110114			
15001001010012100001			
16 100001000000011002			
17 000112100001010001			
18 000000100001000000			
19 00000000000000000000			
20 00000000000000000000			
21 00100100001101013			
22 0001010000000000011			
23 0010010000000000011			
24 001102000011000000			
25 00000000000000000000			
26 00000000000000000000			
27 01000100000000000000			
28 011002000000100001			
29 00000000000000000000			
30101002001001000000			
S 8 7 11 7 7 40 11	6 6 4 5 32	8 9 5 6 5 33	

13.10.2 Documento ATM

Tabella 13.8 Difetti se D# riscontrato (1) o non riscontrato (0) da parte dei soggetti che leggono il documento ATM

Individui prospettiva dell'utente	Tester prospettiva						Designer prospettiva											
	D#1	7	13	19	25	S	3	9	15	21	27	S	5	11	17	23	29	S
1000101000101110013																		
2101002101114111014																		
3 000000000000100102																		
4011103110002111014																		
5 00000000000000000000																		
6 000000000011001102																		
701000100000001012																		
8001001010113110013																		
9 000000000000011002																		
10011002000000010001																		
11100001010113101002																		
12111003001113010001																		
13101002000000011103																		
14 00000000000000000000																		
15010001100012000000																		
16011002111003011002																		
17 000000100001000101																		
18 001001000000010001																		
19100001110002000101																		
20 001001001102000000																		
21 00000000000000000000																		
22 00000000000000000011																		
23 000000001012000000																		
24 00000000000000000000																		
25 00000000000000000000																		
26 000000010102000000																		
27100001000000110013																		
28 101002100001101013																		
29111003100012000101																		
S 8 7 11 2 028							8	6	5	7	8 34		8	11	9	6	8 42	

Ringraziamenti Gli autori desiderano innanzitutto ringraziare gli studenti che hanno partecipato come soggetti nell'esperimento. Vorremmo anche dare un riconoscimento speciale a Forrest Shull presso l'Università del Maryland che ha fornito supporto al pacchetto laboratorio UMD e ha dato molti benefici commenti su una bozza di questo documento. Siamo inoltre grati per tutti i commenti costruttivi espressi

dai revisori anonimi. Grazie anche a Claes Wohlin, Martin Host e Hakan Petersson del Dipartimento di Sistemi di Comunicazione dell'Università di Lund, che hanno esaminato attentamente questo articolo.

Un ringraziamento speciale ad Anders Holtsberg del Centro di Scienze Matematiche, Università di Lund, per il suo aiuto di esperti in analisi statistiche. Questo lavoro è in parte finanziato dal Consiglio Nazionale dell'Industria e Sviluppo tecnico (NUTEK), Svezia, concessione 1K1P-97-09690.

Appendici

Appendice A

Esercizi

Le spiegazioni dei diversi tipi di esercizi si trovano nella Prefazione. In sintesi, l'obiettivo è quello di fornire quattro tipologie di esercizi:

Comprensione	Questi esercizi mirano a evidenziare le questioni più importanti di ciascun capitolo. Gli esercizi sono disponibili nei capp. 1–11 .
Formazione	L'obiettivo di questi esercizi è quello di incoraggiare la pratica della sperimentazione. Ciò include la formulazione di ipotesi e l'esecuzione dell'analisi statistica.
Revisione	I capitoli 12 e 13 includono esempi di esperimenti. Lo scopo di questa parte è quello di fornire aiuto nella revisione e nella lettura degli esperimenti pubblicati.
Compiti	Questi esercizi sono formulati per promuovere la comprensione di come gli esperimenti possono essere utilizzati nell'ingegneria del software per valutare metodi e tecniche.

Le domande di tipo comprensione si trovano alla fine di ogni capitolo, mentre le altre tre tipologie di esercizi si trovano in questa appendice.

A.1 Formazione

Gli esercizi vengono preferibilmente risolti utilizzando un pacchetto di programmi statistici o tabelle tratte da libri di statistica. È possibile utilizzare le tabelle dell'Appendice [B](#), ma le tabelle fornite riguardano solo il livello di significatività del 5%, quindi se vengono utilizzati altri livelli di significatività è necessario utilizzare altre fonti. Va ricordato che l'Appendice [B](#) è stata fornita principalmente per spiegare gli esempi riportati nel Cap. [10](#).

A.1.1 Dati normalmente distribuiti

L'esempio probabilmente più complicato dei metodi statistici nel Cap. 10 è la bontà del fit test per la distribuzione normale, vedere Sez. 10.2.12. È pertanto opportuno garantire una buona comprensione di tale test.

1. Eseguire il test di bontà dell'adattamento, sugli stessi dati, vedere la Tabella 10.20, utilizzando invece 12 segmenti.

A.1.2 Esperienza

Nel cap. 12, il risultato del corso Personal Software Process viene confrontato con il background degli studenti che seguono il corso. L'analisi condotta nel cap. 12 è solo parziale. Il set completo di dati è fornito nelle tabelle A.2 e A.3. Nella Tabella A.1 è presentato il materiale di indagine distribuito nella prima lezione. L'esito dell'indagine è presentato nella tabella A.2. L'esito del corso PSP è presentato nella Tabella A.3, dove sono state utilizzate le seguenti sette misure per misurare l'esito del corso:

Taglia	Il numero di righe di codice nuove e modificate per i dieci programmi.
Tempo	Il tempo di sviluppo totale per i dieci programmi.
Prod.	La produttività misurata come numero di righe di codice per ora di sviluppo.
Difetti	Il numero di errori registrati per i dieci programmi. Ciò include tutti gli errori rilevati, inclusi ad esempio gli errori di compilazione.
Errori/KLOC	Il numero di errori per ogni 1.000 righe di codice.
Pred. Dimensioni	L'errore relativo assoluto nella previsione delle dimensioni del programma. Le cifre mostrano l'errore in percentuale assoluta, ad esempio, sia la sovrastima che la sottostima con il 20% vengono mostrate come 20% senza alcun segno che indichi la direzione dell'errore di stima.
Pred. Tempo	L'errore relativo assoluto nella previsione del tempo di sviluppo.

Sulla base della presentazione nel cap. 12 e i dati delle Tabelle A.2 e A.3 rispondono alle seguenti domande.

1. Come si può migliorare l'indagine? Pensa a ciò che costituisce una buona misura di background, esperienza e capacità.
2. Definire ipotesi aggiuntive rispetto a quelle del Cap. 12, sulla base dei dati disponibili.
Motivare perché queste ipotesi sono interessanti.
3. Che tipo di campionamento è stato utilizzato?
4. Analizza le ipotesi che hai affermato. Quali sono i risultati?
5. Discuti la validità esterna dei tuoi risultati. I risultati possono essere generalizzati al di fuori del PSP?
I risultati possono essere generalizzati agli ingegneri del software industriale?

Tabella A.1 Caratterizzazione degli studenti

Zona	Descrizione	Risposta
Programma di studio (denominato Linea)	Risposta: Informatica e Ingegneria o Ingegneria Elettrica	
Conoscenze generali in informatica e ingegneria del software (denominate SE)	1. Poco, ma curioso del nuovo corso 2. Non è la mia specialità (concentrarmi su altre materie) 3. Abbastanza buono, ma non il mio obiettivo principale (una di un paio di aree) 4. Obiettivo principale dei miei studi	
Conoscenze generali di programmazione (denominate Prog.)	1. Solo 1-2 corsi 2. 3 o più corsi, nessuna esperienza industriale 3. Alcuni corsi e qualche esperienza industriale 4. Più di tre corsi e più di 1 anno di esperienza industriale	
Conoscenza su la PSP (indicata PSP)	1. Di cosa si tratta? 2. Ne ho sentito parlare 3. Una comprensione generale di cosa si tratta 4. Ho letto del materiale	
Conoscenza in C (indicato con C)	1. Nessuna conoscenza preliminare 2. Leggere un libro o seguire un corso 3. Qualche esperienza industriale (meno di 6 mesi) 4. Esperienza industriale	
Conoscenza di C++ (denotato C++)	1. Nessuna conoscenza preliminare 2. Leggere un libro o seguire un corso 3. Qualche esperienza industriale (meno di 6 mesi) 4. Esperienza industriale	
Numero di corsi (indicato come Corsi)	È stato fornito un elenco dei corsi e agli studenti è stato chiesto di scrivere un sì o un no se avevano frequentato o meno il corso. Inoltre, è stato chiesto loro di integrare l'elenco dei corsi se avessero letto qualcosa altro che ritenevano fosse particolarmente rilevante	

A.1.3 Programmazione

In un esperimento, 20 programmati hanno sviluppato lo stesso programma, 10 di loro hanno utilizzato il linguaggio di programmazione A e 10 il linguaggio B. La lingua A è più recente e l'azienda prevede di passare alla lingua A se è migliore della lingua B. Durante lo sviluppo, la dimensione del programma, il tempo di sviluppo, il numero totale di difetti rimossi e il numero di difetti rimossi nei test sono cambiati.

stato misurato.

Tabella A.2 Informazioni provenienti dall'indagine di base

Soggetto	Linea	SE	Progr.	PSP	C	C++	Corsi
11212112							
21321214							
32322227							
41323213							
51323215							
62432117							
72322127							
81322114							
92432119							
10	2	4	2	1	1	1	7
11	1	2	2	1	2	1	3
12	2	4	3	2	1	1	9
13	2	4	3	2	3	3	8
14	2	3	2	2	1	1	6
15	1	3	2	2	1	1	5
16	2	4	2	1	1	1	10
17	1	3	3	1	1	1	5
18	2	4	3	2	1	3	6
19	2	4	3	3	3	3	8
20	1	1	1	1	1	1	2
21	2	3	3	2	2	2	10
22	2	3	2	3	1	1	5
23	1	3	2	2	1	1	4
24	1	2	1	1	1	1	3
25	2	4	3	1	2	2	7
26	1	3	2	2	1	1	5
27	2	4	3	2	3	2	7
28	1	3	2	3	1	1	2
29	2	4	2	3	1	1	7
30	2	3	3	1	2	3	6
31	1	3	2	2	2	2	5
32	2	3	3	1	2	2	10
33	2	4	3	1	1	1	5
34	1	2	2	1	2	2	3
35	1	2	1	1	1	1	2
36	1	2	1	2	1	1	2
37	1	2	2	2	2	2	2
38	2	4	2	2	2	1	6
39	1	2	1	2	1	1	2
40	2	4	3	1	4	4	7
41	2	3	3	2	2	2	8
41	2	4	3	2	2	2	9

(continua)

Tabella A.2 (continua)

Soggetto	Linea	SE	Progr.	PSP	C	C++	Corsi
43	1	3	2	1	1	1	3
44	1	4	3	2	3	2	7
45	2	4	2	2	2	1	6
46	2	2	4	2	4	4	7
47	2	4	3	2	3	2	7
48	1	2	2	2	1	1	2
49	1	3	3	1	1	1	3
50	2	3	2	3	1	1	8
51	2	4	2	4	2	2	8
52	2	4	3	3	3	2	8
53	2	4	3	3	2	2	10
54	1	2	1	2	1	1	2
55	1	2	2	2	1	1	4
56	2	3	2	1	1	1	8
57	1	2	3	1	1	1	4
58	2	4	3	3	1	1	6
59	1	2	2	2	2	1	4

Ai programmatore è stato assegnato in modo casuale un linguaggio di programmazione e l'obiettivo dell'esperimento è valutare se il linguaggio ha qualche effetto sul quattro variabili misurate. I dati raccolti sono riportati nella Tabella A.4. I dati sono fittizio.

1. Quale disegno è stato utilizzato nell'esperimento?
2. Definire le ipotesi per la valutazione.
3. Utilizzare i box plot per indagare le differenze tra le lingue in termini di tendenza centrale e dispersione rispetto a tutti e quattro i fattori. Ce n'è? valore anomalo e, in tal caso, dovrebbe essere rimosso?
4. Si supponga che sia possibile utilizzare test parametrici. Valutare l'effetto della programmazione linguaggio sulle quattro variabili misurate. Da quali conclusioni si possono trarre i risultati?
5. Valutare l'effetto del linguaggio di programmazione sulle quattro variabili misurate utilizzando un test non parametrico. Quali conclusioni si possono trarre dai risultati? Confrontare i risultati con quelli ottenuti utilizzando test parametrici.
6. Discutere la validità dei risultati e se è opportuno utilizzare un test parametrico.
7. Supponiamo che i programmatore partecipanti abbiano scelto la programmazione lingua stessa. Quali conseguenze ha ciò sulla validità dell'art risultati? Le conclusioni sono ancora valide?

Tabella A.3 Esito del corso PSP

Soggetto	Misurare	Prod. ora	Difetti	Guasti/KLOC	Pred. misurare	Pred. tempo
1	839	3.657	13:8	53	63:2	39,7
2	1; 249	3.799	19:7	56	44:8	44,1
3	968	1.680	34:6	71	73:3	29,1
4	996	4.357	13:7	35	35:1	24,3
5	794	2.011	23:7	32	40:3	26,0
6	849	2.505 20:3 4.017		26	30:6	61,1
7	1;	21:7 118 2.673 26:4 61			81:1	36,5
8	4551;	1.552 28:9 2.479 26:8			51:8	34,6
9	177 747			41	54:9	51,0
10	1; 107			59	53:3	22,6
11	729	3.449	12:7	27	37:0	26,9
12	999	3.105	19:3	63	63:1	26,0
13	881	2.224 23:8		44	49:9	47,9
14	730	2.395	18:3	94	128:8	63,0
15	1; 145	3.632	18:9	70	61:1	33,3
16	1; 803	3.193 33:9 2.702		98	54:4	52,9
17	800	17:8 2.089 29:9		60	75:0	34,3
18	1; 042	3.648 15:1 6.807		64	61:4	49,3
19	918			43	46:8	49,7
20	1; 115		9:8	26	23:3	34,1
21	890	4.096	13:0	108	121:3	19,3
22	1; 038	3.609	17:3	98	94:4	21,4
23	1; 251	6.925	10:8 498		398:1	21,8
24	623	4.216	8:9	53	85:1	40,5
25	1; 319	1.864 42:5		92	69:7	43,7
26	800	4.088	11:7	74	92:5	42,6
27	1; 267	2.553 29:8 1.648		88	69:5	53,0
28	945	34:4 4.144 2.869		42	44:4	33,3
29	724	23:7	10:5	49	67:7	32,8
30	1; 131	102			90:2	29,2
31	1; 021	2.235 27:4 3.215		49	48:0	18,0
32	840	15:7 5.643 2.678		69	82:1	85,6
33	985	4.321	10:5	133	135:0	27,3
34	590		13:2	33	55:9	83,0
35	727		10:1	48	66:0	17,0
36	955	3.836	14:9	76	79:6	33,3
37	803	4.470	10:8	56	69:7	18,2
38	684	1.592 25:8 4.188		28	40:9	35,0
39	913	13:1 1.827 39:4		45	49:3	25,3
40	1; 200			61	50:8	31,6
41	894	2.777	19:3	64	71:6	21,3
42	1; 545	3.281	28:3 136		88:0	35,0

(continua)

Tabella A.3 (continua)

Soggetto	Misurare	Prod. ora	Difetti	Guasti/KLOC	Pred. misurare	Pred. tempo
43	995	2.806 21:3 2.464	71	71:4	15.6	38.3
44	807	19:7 2.462 26:3	65	80:5	43.3	26.4
45	1; 078		55	51:0	49.1	51.6
46	944	3.154	18:0	71	75:2	59.0
47	868	1.564	33:3	50	57:6	50.4
48	701	3.188	13:2	31	44:2	21.2
49	1;	4.823	13:8	86	77:7	19.3
50	1071; 535	2.938	31:3	71	46:3	29.6
51	858	7.163	7:2	97	113:1	58.4
52	832	2.033 24:6 3.160	84	101:0	48.4	25.6
53	975	18:5 3.337 4.583	115	117:9	29,5	31.5
54	715		12:9	40	55:9	41.7
55	947		12:4	99	104:5	41.0
56	926	2.924	19:0	77	83:2	32,5
57	711	3.053	14:0	78	109:7	22.8
58	1; 283	7.063	10:9	186	145:0	46,5
59	1; 261	3.092 24:5		54	42:8	27.4
						45.3

A.1.4 Progettazione

Questo esercizio si basa sui dati ottenuti da un esperimento condotto da Briand,

Bunse e Daly. L'esperimento è ulteriormente descritto da Briand et al. [28].

Viene progettato un esperimento per valutare l'impatto dei principi di progettazione orientata agli oggetti di qualità quando si intende modificare un determinato progetto. La qualità i principi di progettazione valutati sono i principi forniti da Coad e Yourdon [35].

Nell'esperimento vengono utilizzati due sistemi con un progetto per ciascun sistema. Uno dei design è un progetto "buono" realizzato utilizzando i principi di progettazione e l'altro è un progetto "cattivo" progettare non utilizzando i principi. I due progetti sono documentati allo stesso modo in termini di layout e contenuto e sono della stessa dimensione, cioè sono sviluppati essere il più simile possibile, tranne che per il fatto di seguire o meno i principi di progettazione. L'obiettivo dell'esperimento è valutare se i principi della progettazione di qualità si allentano analisi dell'impatto durante l'identificazione delle modifiche alla progettazione.

Il compito di ciascun partecipante è quello di intraprendere due analisi di impatto separate, una per ogni progetto di sistema. Contrassegnare tutti i punti del disegno che devono essere modificati ma non modificarli effettivamente rende le analisi di impatto. La prima analisi di impatto è per una mutata esigenza del cliente e il secondo è per un miglioramento del funzionalità dei sistemi. Durante l'attività vengono raccolte quattro misure:

Mod Time: tempo impiegato per identificare i luoghi per la modifica.

Mod Comp: Rappresenta la completezza dell'analisi di impatto ed è definita

Tabella A.4 Dati per l'esercizio di programmazione

Programmazione lingua	Dimensioni del programma (LOC)	Sviluppo tempo (minuti)	Numero totale di difetti	Numero di prove difetti
UN	1; 408	3.949	89	23
UN	1; 529	2.061	69	16
UN	946	3.869	170	41
UN	1; 141	5.562	271	55
UN	696	5.028	103	39
UN	775	2.296	75	29
UN	1;	2.980	79	11
UN	2051;	2.991	194	28
UN	159 862	2.701	67	27
UN	1; 206	2.592	77	15
B	1; 316	3.986	68	20
B	1;	4.477	54	10
B	7871;	3.789	130	23
B	1051;	4.371	48	13
B	583 1; 381	3.325	133	29
B	944	5.234	80	25
B	1;	4.901	64	21
B	4921;	3.897	89	29
B	217 936	3.825	57	20
B	1; 441	4.015	79	18

$$\text{Mod Comp D} = \frac{\text{Numero di posti corretti trovati}}{\text{Numero totale di posti da trovare}}$$

Mod Corr: Rappresenta la correttezza dell'analisi di impatto ed è definita come:

$$\text{Mod Corr D} = \frac{\text{Numero di posti corretti trovati}}{\text{Numero totale di luoghi indicati come trovati}}$$

Mod Rate: Il numero di posti corretti trovati per unità di tempo, ovvero:

$$\text{Tasso di modifica D} = \frac{\text{Numero di posti corretti trovati}}{\text{Tempo di identificazione}}$$

L'esperimento viene condotto in due occasioni, in modo da lasciare che ciascun partecipante lavorare sia con il buon design che con il cattivo design. I soggetti erano casuali assegnato a uno dei due gruppi, A o B. Il gruppo A ha lavorato con il buon design al prima occasione e pessima progettazione nella seconda. Il gruppo B ha studiato per primo il cattivo design e poi il buon design. I dati raccolti sono riportati nella Tabella A.5.

Tabella A.5 Dati per l'esercizio di progettazione

Partecipante	Gruppo	Buon design orientato agli oggetti				Cattiva progettazione orientata agli oggetti			
		Tempo	ModComp	Nord	Scarsità	Tempo	I scarsità	I tempo	Scarsità
P01	B	–	0,388	0,75	–	–	0,238	0,714	–
P02	B	–	0,018	1	–	–	0,095	1	–
P03	UN	20	0,028	1	0,45	25	0,19	1	0,16
P04	B	22	0,018	1	0,818	25	0,238	1	0,2
P05	B	30	0,008	1	0,667	35	0,476	0,909	0,286
P07	UN	–	0	–	–	38	0,476	1	0,263
P09	UN	–	0,455	1	–	–	0,476	1	–
P10	B	–	0,409	0,9	–	–	0,381	1	–
P11	UN	45	0,545	0,923	0,267	50	0,714	1	0,3
P12	B	–	0,773	1	–	–	0,714	1	–
P13	UN	40	0,773	1	0,425	40	0,762	1	0,4
P14	B	30	0,909	1	0,667	30	0,333	0,875	0,233
P15	B	–	0,864	1	–	40	0,238	1	0,125
P16	B	30	0,773	1	0,567	–	–	–	–
P17	B	–	0,955	1	–	–	0,286	0,75	–
P18	B	–	0	–	–	–	0,19	1	–
P19	UN	29	0,818	1	0,621	27	0,667	1	0,519
P20	UN	9	0,591	1	1,444	15	0,19	0,8	0,267
P21	B	20	0,591	1	0,65	35	0,19	1	0,114
P22	B	30	0,682	1	0,5	20	0,714	1	0,75
P23	B	–	0,818	1	–	–	0,476	1	–
P24	UN	30	0,773	1	0,567	40	0,762	1	0,4
P25	UN	–	0,955	1	–	–	0,667	0,875	–
P26	B	25	0	0	0	25	0,095	0,5	0,08
P27	UN	27	0,773	0,944	0,63	36	0,389	0,7	0,194
P28	UN	25	0,773	1	0,68	30	0,667	1	0,467
P29	B	44	0,773	1	0,386	23	0,762	1	0,696
P31	UN	–	0,409	1	–	–	0,286	0,75	–
P32	UN	30	0,909	1	0,667	–	0,5	1	–
P33	UN	65	0,818	1	0,277	–	0,619	1	–
P34	UN	50	0,636	0,933	0,28	30	0,4	0,889	0,267
P35	UN	10	0,591	1	1,3	10	0,667	1	1,4
P36	UN	13	1	1	1,692	–	0,619	1	–

1. Quale disegno è stato utilizzato nell'esperimento?
2. Definire le ipotesi per la valutazione.
3. Come dovrebbero essere trattati i valori mancanti nella Tabella A.5 ?
4. Si supponga che sia possibile utilizzare test parametrici. Valutare l'effetto dei principi di progettazione della qualità sulle quattro variabili misurate. Quali conclusioni si possono trarre dai risultati?

5. Valutare l'effetto dei principi di progettazione della qualità sulle quattro variabili misurate utilizzando test non parametrici. Quali conclusioni si possono trarre dai risultati?
Confrontare i risultati con quelli ottenuti utilizzando test parametrici.
6. Discutere la validità dei risultati e se è opportuno utilizzare test parametrici.
7. I partecipanti all'esperimento sono studenti che seguono un corso di ingegneria del software che si sono offerti volontari come soggetti. Da quale popolazione viene prelevato il campione?
Discutere in che modo questo tipo di campionamento influenzera la validità esterna dell'esperimento? Come è possibile effettuare diversamente il campionamento?

A.1.5 Ispezioni

Questo esercizio fa riferimento all'esperimento di esempio nel Cap. 13.

1. Riscrivi l'abstract nel cap. 13 deve essere un abstract strutturato, come definito in Cap. 11.
2. Condurre le fasi di definizione dell'ambito e di pianificazione per una replica *esatta* dell'esperimento. In particolare, definire quanti soggetti dovrebbero essere arruolati per raggiungere un determinato livello di fiducia nell'analisi.
3. Condurre la fase di definizione dell'ambito per una replica *differenziata* dell'esperimento. Definire tre diversi modelli di obiettivo per tre repliche alternative. Discutere i pro e i contro di ciascuna alternativa rispetto a costi, rischi e guadagni (vedere anche Fig. 2.1).

A.2 Revisione

Di seguito è riportato un elenco di domande importanti da considerare quando si legge o si rivede un articolo che presenta un esperimento. Utilizzare l'elenco e rivedere gli esempi presentati nei capp. 12 e 13, ed anche alcuni esperimenti presentati in letteratura.

L'elenco seguente dovrebbe essere visto come una lista di controllo in aggiunta alle normali domande durante la lettura di un articolo. Un esempio di una domanda normale potrebbe essere; l'abstract è una buona descrizione del contenuto dell'articolo? Alcuni aspetti specifici da considerare quando si legge un articolo sperimentale sono:

- L'esperimento è comprensibile e interessante in generale? • L'esperimento ha qualche valore pratico? • Sono stati riassunti e referenziati altri esperimenti che affrontano il problema?

- Qual è la popolazione nell'esperimento? • Il campione utilizzato è rappresentativo della popolazione? • Le variabili dipendenti e indipendenti sono chiaramente definite? • Le ipotesi sono formulate chiaramente? • Il tipo di progettazione è chiaramente indicato? • Il progetto è corretto? • La strumentazione è descritta correttamente? • La validità dell'esperimento è considerata attentamente e convincente? • I diversi tipi di minacce alla validità vengono affrontati adeguatamente? • I dati sono stati validati? • La potenza statistica è sufficiente? Ci sono abbastanza soggetti nell'esperimento? • Vengono applicati i test statistici appropriati? Vengono utilizzati test parametrici o non parametrici e vengono utilizzati correttamente? • Il livello di significatività utilizzato è appropriato? • I dati sono interpretati correttamente? • Le conclusioni sono corrette? • I risultati non sono sopravvalutati? • È possibile replicare lo studio? • Vengono forniti i dati? • È possibile utilizzare i risultati per eseguire una meta-analisi? • Sono delineati ulteriori lavori e sperimentazioni nel settore?

A.3 Incarichi

Queste assegnazioni si basano sul seguente scenario generale. Un'azienda vorrebbe migliorare il proprio modo di lavorare modificando il processo software. Verrai consultato come esperto nella valutazione di nuove tecniche e metodi in relazione al processo esistente. L'azienda vorrebbe sapere se modificare o meno il proprio processo software.

Dovresti cercare la letteratura appropriata, rivedere la letteratura esistente sull'argomento, applicare il processo di sperimentazione e scrivere un rapporto contenente una raccomandazione per l'azienda. La raccomandazione dovrebbe discutere sia i risultati dell'esperimento che altre questioni rilevanti per prendere la decisione se modificare o meno il processo. Altre questioni rilevanti includono i costi e i benefici per apportare il cambiamento. Se non riesci a trovare i costi corretti, sei tenuto a fare delle stime. Quest'ultimo può essere in termini di costi relativi.

I compiti sono intenzionalmente abbastanza aperti per consentire l'interpretazione e la discussione. Ogni incarico viene descritto in termini di prerequisiti necessari per eseguire l'incarico e quindi viene brevemente descritto il compito effettivo. Va notato che i compiti seguenti sono esempi di possibili esperimenti che possono essere condotti. La questione importante da tenere a mente è che l'obiettivo principale è che i compiti forniscano pratica nell'uso degli esperimenti come parte di una procedura di valutazione.

Infine, va notato che alcune organizzazioni forniscono i cosiddetti pacchetti di laboratorio che possono essere utilizzati per replicare gli esperimenti. I pacchetti di laboratorio sono importanti in quanto ci consentono di basarci sul lavoro di altri e quindi, si spera, di arrivare a risultati più generalmente validi mediante la replica. Alcuni pacchetti di laboratorio possono essere trovati effettuando una ricerca su Internet. Potrebbe anche essere utile contattare lo sperimentatore originale per ottenere supporto e magari anche un pacchetto di laboratorio non pubblicato.

A.3.1 **Test unitario e revisioni del codice**

L'azienda vuole valutare se sia conveniente introdurre revisioni del codice. I test unitari vengono eseguiti oggi, anche se su codice non rivisto. È questo il modo migliore per farlo?

Prerequisiti

- Programmi idonei con difetti rilevabili durante le revisioni o i test. • Un metodo di revisione, che può essere ad hoc, ma è preferibile che sia qualcosa di più realistico, ad esempio un approccio basato su una lista di controllo. In questo caso è necessaria una lista di controllo.
- Un metodo di prova, che può anche essere ad hoc, ma preferibilmente è basato su, for esempio, partizionamento per utilizzo o equivalenza.

Compito

- Valutare se è conveniente introdurre revisioni del codice.

A.3.2 **Metodi di ispezione**

Sono disponibili diversi modi per condurre le revisioni. L'azienda intende introdurre il miglior metodo di ispezione tra due possibili scelte. Quale dei due metodi è meglio introdurre in azienda?

Prerequisiti

- Dovrebbero essere disponibili artefatti software idonei da revisionare. •

Due metodi di revisione con supporto adeguato in termini, ad esempio, di liste di controllo o di descrizione di diverse prospettive di lettura, vedere anche Appendice A.1.5.

Compito

- Supponendo che l'azienda intenda introdurre revisioni degli artefatti software scelti, quale metodo dovrebbero introdurre? Determinare quale dei metodi di ispezione è il migliore per individuare i difetti. Il metodo migliore è anche conveniente?

A.3.3 Notazione dei requisiti

È importante scrivere le specifiche dei requisiti in modo che tutti i lettori li interpretino facilmente e nello stesso modo. L'azienda ha diverse notazioni tra cui scegliere. Qual è il modo migliore per rappresentare i requisiti?

Prerequisiti

- Una specifica dei requisiti scritta in diverse notazioni, ad esempio:
linguaggio naturale e diverse rappresentazioni grafiche.

Compito

- Valutare se è vantaggioso modificare la notazione aziendale per le specifiche dei requisiti.
Supponiamo che oggi l'azienda utilizzi il linguaggio naturale.

Appendice B

Tabelle statistiche

Questa appendice contiene tabelle statistiche per un livello di significatività del 5%. Tabelle più elaborate si possono trovare nella maggior parte dei libri di statistica, ad esempio [119], e sono disponibili anche su Internet. L'obiettivo principale qui è quello di fornire alcune informazioni, in modo che i test spiegati nel Cap. 10 diventino comprensibili e in modo che gli esempi forniti possano essere seguiti. Ciò è importante anche se per i calcoli vengono utilizzati pacchetti statistici, poiché è importante comprendere i calcoli sottostanti prima di applicare semplicemente i diversi test statistici. Vale anche la pena notare che le tabelle sono una scorciatoia, ad esempio i valori per t-test, F-test e Chi-2 possono essere calcolati dalle rispettive distribuzioni.

Sono incluse le seguenti tabelle statistiche:

- t-test (vedi Sez. 10.3.4, 10.3.7 e Tabella B.1) •
- Chi-2 (vedi Sez. 10.3.12 e Tabella B.2) •
- Mann-Whitney (vedi Sez. 10.3.5 e Tabella B.3) •
- Wilcoxon (vedi Sez. 10.3.8 e Tabella B.4) • F-test (vedi Sez. 10.3.6, 10.3.10, Tabella B.5)

Tabella B.1 Valori critici

t-test a due code (5%), vedere
Sette. 10.3.4 e 10.3.7

Gradi di libertà	valore t
1	12:706
2	4:303
3	3:182
4	2:776
5	2:571
6	2:447
7	2:365
8	2:306
9	2:262
10	2:228
11	2:201
12	2:179
13	2:160
14	2:145
15	2:131
16	2:120
17	2:110
18	2:101
19	2:093
20	2:086
21	2:080
22	2:074
23	2:069
24	2:064
25	2:060
26	2:056
27	2:052
28	2:048
29	2:045
30	2:042
40	2:021
60	2:000
120	1:980
1	1:960

Tabella B.2 Valori critici

Test Chi2 a una coda (5%), vedere
Setta. [10.3.12](#)

Gradi di libertà	2
1	3:84
2	5:99
3	7:81
4	9:49
5	11:07
6	12:59
7	14:07
8	15:51
9	16:92
10	18:31
11	19:68
12	21:03
13	22:36
14	23:68
15	25:00
16	26:30
17	27:59
18	28:87
19	30:14
20	31:41
21	32:67
22	33:92
23	35:17
24	36:42
25	37:65
26	38:88
27	40:11
28	41:34
29	42:56
30	43:77
40	55:76
60	79:08
80	101:88
100	124:34

Tabella B.3 Valori critici Mann-Whitney a due code (5%), vedere sez. [10.3.5](#)

NB 5678	9	10	11	12
N / A				
3 011 22334				
4 123 44567				
5 235 6	7	8	9	11
6	10	11	13	14
7	12	14	16	18
8	13	15	17	19
9	17	20	23	26
10		23	26	29
11			30	33
12				37

Tabella B.4 Valori critici

coppia abbinata a due code

Test di Wilcoxon (5%), vedi

Setta. [10.3.8](#)

N	T
6	0
7	2
8	3
9	5
10	8
11	10
12	13
13	17
14	21
15	25
16	29
17	34
18	40
19	46
20	52
22	66
25	89

Si prega di notare che nella Tabella B.3, NA è per il campione più piccolo e NB per quello più grande campione.

Si prega di notare che la Tabella B.5 fornisce il punto superiore dello 0:025% della distribuzione F essendo f1 e f2 i gradi di libertà. Ciò equivale a F0:0025;f1;f2 .

Riferimenti

1. Anastas, JW, MacDonald, ML: Progettazione della ricerca per il servizio sociale e l'essere umano Servizi, 2a ed. Columbia University Press, New York (2000)
2. Andersson, C., Runeson, P.: Un modello di processo a spirale per casi di studio sul monitoraggio della qualità del software: metodo e metriche. Softw. Processo: improvvisazione. Pratica. 12(2), 125–140 (2007). doi: 10.1002/spip.3113
3. Andrews, AA, Pradhan, AS: Questioni etiche nell'ingegneria del software empirico: i limiti di politica. Impero. Softw. L'Ing. 6(2), 105–110 (2001)
4. American Psychological Association: Principi etici degli psicologi e codice di condotta. Sono. Psicologo. 47, 1597–1611 (1992)
5. Avison, D., Baskerville, R., Myers, M.: Controllo dei progetti di ricerca-azione. Inf. Tecnologia. Persone 14(1), 28–45 (2001). doi: 10.1108/09593840110384762
6. Babbie, ER: Metodi di ricerca basata sull'indagine. Wadsworth, Belmont (1990)
7. Basili, VR: Valutazione quantitativa della metodologia dell'ingegneria del software. In: Atti della prima conferenza informatica panpacifica, vol. 1, pp. 379–398. Società informatica australiana, Melbourne (1985)
8. Basili, VR: Sviluppo software: un paradigma per il futuro. In: Atti della 13a conferenza annuale internazionale sui software e sulle applicazioni per computer, COMPSAC'89, Orlando, pp. 471–485. IEEE Computer Society Press, Washington (1989)
9. Basili, VR: Il paradigma sperimentale nell'ingegneria del software. In: HD Rombach, VR Basili, RW Selby (a cura di) Problemi di ingegneria del software sperimentale: valutazione critica e direttive future. Appunti delle lezioni di informatica, vol. 706. Springer, Berlino Heidelberg (1993)
10. Basili, VR: Evoluzione e confezionamento delle tecnologie di lettura. J.Sist. Softw. 38(1), 3–12 (1997)
11. Basili, VR, Weiss, DM: Una metodologia per la raccolta di dati validi di ingegneria del software. IEEE Trans. Softw. L'Ing. 10(6), 728–737 (1984)
12. Basili, VR, Selby, RW: Confronto dell'efficacia delle strategie di test del software. IEEE Trans. Softw. L'Ing. 13(12), 1278–1298 (1987)
13. Basili, VR, Rombach, HD: Il progetto TAME: verso un software orientato al miglioramento ambienti. IEEE Trans. Softw. L'Ing. 14(6), 758–773 (1988)
14. Basili, VR, Green, S.: Valutazione del processo software al SEL. Software IEEE. 11(4), pp. 58–66 (1994)
15. Basili, VR, Selby, RW, Hutchens, DH: Sperimentazione nell'ingegneria del software. IEEE Trans. Softw. L'Ing. 12(7), 733–743 (1986)
16. Basili, VR, Caldiera, G., Rombach, HD: Fabbrica di esperienze. In: JJ Marciniak (a cura di) Enciclopedia dell'ingegneria del software, pp. 469–476. Wiley, New York (1994)

17. Basili, VR, Caldiera, G., Rombach, HD: paradigma Goal Question Metrics. In: JJ Marciniak (a cura di) Encyclopedia of Software Engineering, pp. 528–532. Wiley (1994)
18. Basili, VR, Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgard, S., Zelkowitz, MV: L'indagine empirica della lettura basata sulla prospettiva. Impero. Morbido. L'Ing. 1(2), 133–164 (1996)
19. Basili, VR, Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgard, S., Zelkowitz, MV: Pacchetto laboratorio per l'indagine empirica della lettura basata sulla prospettiva. Rapporto tecnico, Università del Maryland (1998). URL http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/pacchetto_pbr/manual.html
20. Basili, VR, Shull, F., Lanubile, F.: Costruire conoscenza attraverso famiglie di esperimenti. IEEE Trans. Softw. L'Ing. 25(4), 456–473 (1999)
21. Baskerville, RL, Wood-Harper, AT: Una prospettiva critica sulla ricerca-azione come metodo per la ricerca sui sistemi informativi. J.Inf. Tecnologia. 11(3), 235–246 (1996). doi: 10.1080/026839696345289
22. Benbasat, I., Goldstein, DK, Mead, M.: La strategia di ricerca dei casi negli studi sui sistemi informativi. MIS Q. 11(3), 369 (1987). doi: 10.2307/24868423. Bergman, B., Klefsjö, B.: Qualità dalle esigenze del cliente alla soddisfazione del cliente. Letteratura studentesca, Lund (2010)
24. Brereton, P., Kitchenham, BA, Budgen, D., Turner, M., Khalil, M.: Lezioni dall'applicazione del processo di revisione sistematica della letteratura all'interno del dominio dell'ingegneria del software. J.Sist. Softw. 80(4), 571–583 (2007). doi: 10.1016/j.jss.2006.07.009 25. Brereton, P., Kitchenham, BA, Budgen, D.: Utilizzo di un modello di protocollo per la pianificazione del caso di studio. In: Atti della 12a Conferenza internazionale sulla valutazione e l'accertamento nell'ingegneria del software. Università di Bari, Italia (2008)
26. Briand, LC, Differding, CM, Rombach, HD: linee guida pratiche per il miglioramento dei processi basati sulla misurazione. Softw. Processo: improvvisazione. Pratica. 2(4), 253–280 (1996)
27. Briand, LC, El Emam, K., Morasca, S.: Sull'applicazione della teoria della misurazione nell'ingegneria del software. Impero. Softw. L'Ing. 1(1), 61–88 (1996)
28. Briand, LC, Bunse, C., Daly, JW: un esperimento controllato per valutare le linee guida di qualità sulla manutenibilità dei progetti orientati agli oggetti. IEEE Trans. Softw. L'Ing. 27(6), 513–530 (2001)
29. British Psychological Society: principi etici per condurre ricerche con partecipanti umani. Psicologo 6(1), 33–35 (1993)
30. Budgen, D., Kitchenham, BA, Charters, S., Turner, M., Brereton, P., Linkman, S.: Presentazione dei risultati dell'ingegneria del software utilizzando abstract strutturati: un esperimento randomizzato. Impero. Softw. L'Ing. 13, 435–468 (2008). doi: 10.1007/s10664-008-9075-7 31. Budgen, D., Burn, AJ, Kitchenham, BA: Reporting di progetti informatici attraverso abstract strutturati: un quasi-esperimento. Impero. Softw. L'Ing. 16(2), 244–277 (2011). doi: 10.1007/s10664-010-9139-3 32. Campbell, DT, Stanley, JC: Disegni sperimentali e quasi sperimentali per la ricerca. Azienda Houghton Mifflin, Boston (1963)
33. Chrissis, MB, Konrad, M., Shrum, S.: CMMI(R): Linee guida per l'integrazione dei processi e miglioramento del prodotto. Relazione tecnica, SEI (2003)
34. Ciolkowski, M., Differding, CM, Laitenberger, O., Münch, J.: Indagine empirica sulla lettura basata sulla prospettiva: un esperimento replicato. Relazione tecnica, 97-13. ISERN (1997)
35. Coad, P., Yourdon, E.: Design orientato agli oggetti, 1a ed. Prentice-Hall, Englewood (1991)
36. Cohen, J.: Kappa ponderato: accordo su scala nominale con previsione di disaccordo su scala o credito parziale. Psicologo. Toro. 70, 213–220 (1968)
37. Cook, TD, Campbell, DT: Quasi-sperimentazione – Problemi di progettazione e analisi sul campo Impostazioni. Azienda Houghton Mifflin, Boston (1979)
38. Corbin, J., Strauss, A.: Fondamenti di ricerca qualitativa, 3a ed. SAGE, Los Angeles (2008)
39. Cruzes, DS, Dyba, T.: Sintesi della ricerca nell'ingegneria del software: uno studio terzario. Inf. Softw. Tecnologia. 53(5), 440–455 (2011). doi: 10.1016/j.infsof.2011.01.004

40. Dalkey, N., Helmer, O.: Un'applicazione sperimentale del metodo Delphi all'uso di esperti. *Gest. Sci.* 9(3), 458–467 (1963)
41. DeMarco, T.: Controllo dei progetti software. Yourdon Press, New York (1982)
42. Demming, WE: Fuori dalla crisi. Centro del MIT per gli studi di ingegneria avanzata, MIT Press, Cambridge, Massachusetts (1986)
43. Dieste, O., Griman, A., Juristo, N.: Sviluppare strategie di ricerca per individuare esperimenti rilevanti. *Impero. Softw. L'Ing.* 14, 513–539 (2009). URL <http://dx.doi.org/10.1007/s10664-008-9091-7>
44. Dietrich, Y., Ronkkö, K., Eriksson, J., Hansson, C., Lindeberg, O.: Sviluppo del metodo cooperativo. *Impero. Softw. L'Ing.* 13(3), 231–260 (2007). doi: 10.1007/s10664-007-9057-1 45. Doolan, EP: Esperienze con il metodo di ispezione di Fagan. *Softw. Pratica. Esp.* 22(2), 173–182 (1992)
45. Dyba, T., Dingsøyr, T.: Studi empirici sullo sviluppo agile del software: una revisione sistematica. *Inf. Softw. Tecnologia.* 50(9-10), 833–859 (2008). doi: DOI:10.1016/j.infsof.2008.01.006 47. Dyba, T., Dingsøyr, T.: Forza dell'evidenza nelle revisioni sistematiche nell'ingegneria del software. In: Atti del 2° Simposio internazionale ACM-IEEE sull'ingegneria e la misurazione del software empirico, ESEM '08, Kaiserslautern, pp. 178–187. ACM, New York (2008). doi: <http://doi.acm.org/10.1145/1414004.1414034> 48. Dyba, T., Kitchenham, BA, Jørgensen, M.: Ingegneria del software basata sull'evidenza per professionisti. *Software IEEE.* 22, 58–65 (2005). doi: <http://doi.ieeecomputersociety.org/10.1109/MS.2005.6>
49. Dyba, T., Kampenes, VB, Sjøberg, DIK: Una revisione sistematica del potere statistico negli esperimenti di ingegneria del software. *Inf. Softw. Tecnologia.* 48(8), 745–755 (2006). doi: 10.1016/j.infsof.2005.08.009 50. Easterbrook, S., Singer, J., Storey, M.-A., Damian, D.: Selezione di metodi empirici per la ricerca sull'ingegneria del software. In: F. Shull, J. Singer, DI Sjøberg (a cura di) Guida all'ingegneria del software empirico avanzato. Springer, Londra (2008)
51. Eick, SG, Loader, CR, Long, MD, Votta, LG, Vander Wiel, SA: Stima del contenuto dell'errore del software prima della codifica. In: Atti della 14a conferenza internazionale sull'ingegneria del software, Melbourne, pp. 59–65. ACM Press, New York (1992)
52. Eisenhardt, KM: Costruire teorie dalla ricerca di casi di studio. *Accade. Gest. Rev.* 14(4), 532 (1989). doi: 10.2307/258557 53. Endres, A., Rombach, HD: Un manuale di ingegneria del software e dei sistemi – Osservazioni empiriche, leggi e teorie. Pearson Addison-Wesley, Harlow/New York (2003)
54. Fagan, ME: Ispezioni di progettazione e codice per ridurre gli errori nello sviluppo del programma. *IBM Sist. J.* 15 (3), 182–211 (1976)
55. Fenton, N.: Misurazione del software: una base scientifica necessaria. *IEEE Trans. Softw. L'Ing.* 3(20), 199–206 (1994)
56. Fenton, N., Pfleeger, SL: Software Metrics: A Rigorous and Practical Approach, 2a ed. International Thomson Computer Press, Londra (1996)
57. Fenton, N., Pfleeger, SL, Glass, R.: Scienza e sostanza: una sfida per gli ingegneri del software. *Software IEEE.* 11, 86–95 (1994)
58. Fink, A.: The Survey Handbook, 2a ed. SAGE, Thousand Oaks/Londra (2003)
59. Flyvbjerg, B.: Cinque malintesi sulla ricerca di casi di studio. In: Qualitative Research Practice, edizione tascabile concisa, pp. 390–404. SAGE, Londra (2007)
60. Frigge, M., Hoaglin, DC, Iglewicz, B.: Alcune implementazioni del boxplot. *Sono. Statistica.* 43(1), 50-54 (1989)
61. Fusaro, P., Lanubile, F., Visaggio, G.: Un esperimento replicato per valutare le tecniche di ispezione dei requisiti. *Impero. Softw. L'Ing.* 2(1), 39–57 (1997)
62. Glass, RL: La crisi della ricerca sul software. *Software IEEE.* 11, 42–47 (1994)
63. Glass, RL, Vessey, I., Ramesh, V.: Ricerca nell'ingegneria del software: un'analisi della letteratura. *Inf. Softw. Tecnologia.* 44(8), 491–506 (2002). doi: 10.1016/S0950-5849(02)00049-6

64. Gomez, OS, Juristo, N., Vegas, S.: Tipi di replica nelle discipline sperimentali. In: Atti del 4th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, Bolzano (2010)
65. Gorschek, T., Wohlin, C.: Modello di astrazione dei requisiti. Richiedi. L'Ing. 11, 79–101 (2006). doi: 10.1007/s00766-005-0020-7
66. Gorschek, T., Garre, P., Larsson, S., Wohlin, C.: Un modello per il trasferimento tecnologico nella pratica. Software IEEE. 23(6), 88–95 (2006)
67. Gorschek, T., Garre, P., Larsson, S., Wohlin, C.: Valutazione industriale del modello di astrazione dei requisiti. Richiedi. L'Ing. 12, 163–190 (2007). doi: 10.1007/s00766-007-0047-z 68. Grady, RB, Caswell, DL: Metriche del software: definizione di un programma a livello aziendale. Prentice-Hall, Englewood (1994)
69. Grant, EE, Sackman, H.: Un'indagine esplorativa sulle prestazioni del programmatore in condizioni online e offline. IEEE Trans. Elettrone del fattore umano. HFE-8(1), 33–48 (1967)
70. Gregor, S.: La natura della teoria nei sistemi informativi. MIS D. 30(3), 491–506 (2006)
71. Hall, T., Flynn, V.: Questioni etiche nella ricerca sull'ingegneria del software: un'indagine sulle attuali pratica. Impero. Softw. L'Ing. 6, 305–317 (2001)
72. Hannay, JE, Sjøberg, DIK, Dyba, T.: Una revisione sistematica dell'uso della teoria negli esperimenti di ingegneria del software. IEEE Trans. Softw. L'Ing. 33(2), 87–107 (2007). doi: 10.1109/TSE.2007.12
73. Hannay, JE, Dyba, T., Arisholm, E., Sjøberg, DIK: L'efficacia della programmazione in coppia: una meta-analisi. Inf. Softw. Tecnologia. 51(7), 1110–1122 (2009). doi: 10.1016/j.infsof.2009.02.001
74. Hayes, W.: Sintesi della ricerca nell'ingegneria del software: un caso per la meta-analisi. In: Atti del 6° Simposio internazionale sulla metrica del software, Boca Raton, pp. 143–151 (1999)
75. Hetzel, B.: Far funzionare la misurazione del software: costruire una misurazione efficace Programma. Wiley, New York (1993)
76. Hevner, AR, March, ST, Park, J., Ram, S.: Design science in information Systems Research. MIS D. 28(1), 75–105 (2004)
77. Ospite, M., Regnell, B., Wohlin, C.: Usare gli studenti come soggetti – uno studio comparativo di studenti e professionisti nella valutazione dell'impatto del lead-time. Impero. Softw. L'Ing. 5(3), 201–214 (2000)
78. Host, M., Wohlin, C., Thelin, T.: Classificazione del contesto sperimentale: incentivi ed esperienza dei soggetti. In: Atti della 27a conferenza internazionale sull'ingegneria del software, St. Louis, pp. 470–478 (2005)
79. Host, M., Runeson, P.: Liste di controllo per la ricerca di casi di studio sull'ingegneria del software. In: Atti del primo simposio internazionale sull'ingegneria e la misurazione del software empirico, Madrid, pp. 479–481 (2007)
80. Hove, SE, Anda, B.: Esperienze derivanti dalla conduzione di interviste semi-strutturate nella ricerca empirica sull'ingegneria del software. In: Atti dell'11° Simposio internazionale sui parametri del software dell'IEEE, pp. 1–10. IEEE Computer Society Press, Los Alamitos (2005)
81. Humphrey, WS: Gestione del processo software. Addison-Wesley, Lettura (1989)
82. Humphrey, WS: Una disciplina per l'ingegneria del software. Addison Wesley, Lettura (1995)
83. Humphrey, WS: Introduzione al processo del software personale. Addison Wesley, Lettura (1997)
84. IEEE: glossario standard IEEE della terminologia dell'ingegneria del software. Rapporto tecnico, IEEE Norma 610.12-1990, IEEE (1990)
85. Iversen, JH, Mathiassen, L., Nielsen, PA: Gestione del rischio nel miglioramento dei processi software: un approccio di ricerca-azione. MIS D. 28(3), 395–433 (2004)
86. Jedlitschka, A., Pfahl, D.: Linee guida per la rendicontazione di esperimenti controllati nell'ingegneria del software. In: Atti del 4° simposio internazionale sull'ingegneria del software empirico, Noosa Heads, pp. 95–104 (2005)
87. Johnson, PM, Tjahjono, D.: Ogni ispezione ha davvero bisogno di un incontro? Impero. Softw. L'Ing. 3(1), 9–35 (1998)

88. Juristo, N., Moreno, AM: *Nozioni di base sulla sperimentazione dell'ingegneria del software*. Springer, Kluwer Academic Publishers, Boston (2001)
89. Juristo, N., Vegas, S.: Il ruolo delle repliche non esatte negli esperimenti di ingegneria del software. *Impero. Softw. L'Ing.* **16**, 295–324 (2011). doi: 10.1007/s10664-010-9141-9
90. Kachigan, SK: *Analisi statistica: un'introduzione interdisciplinare all'univariato e Metodi multivariati*. Radius Press, New York (1986)
91. Kachigan, SK: *Analisi statistica multivariata: un'introduzione concettuale*, 2a ed. Radius Press, New York (1991)
92. Kampenes, VB, Dyba, T., Hannay, JE, Sjø berg, DIK: Una revisione sistematica della dimensione dell'effetto negli esperimenti di ingegneria del software. *Inf. Softw. Tecnologia.* **49**(11–12), 1073–1086 (2007). doi: 10.1016/j.infsof.2007.02.015
93. Karahasanovic, A., Anda, B., Arisholm, E., Hove, SE, Jørgensen, M., Sjøberg, D., Welland, R.: Raccolta di feedback durante gli esperimenti di ingegneria del software. *Impero. Softw. L'Ing.* **10**(2), 113–147 (2005). doi: 10.1007/s10664-004-6189-4. URL <http://www.springerlink.com/index/10.1007/s10664-004-6189-4>
94. Karlstrom, D., Runeson, P., Wohlin, C.: Punti di vista aggregati per il miglioramento strategico dei processi software. *Procedimento IEE Softw.* **149**(5), 143–152 (2002). doi: 10.1049/ip-sen:20020696 95. Kitchenham, BA: Il ruolo delle repliche nell'ingegneria del software empirica: un avvertimento. *Impero. Softw. L'Ing.* **13**, 219–221 (2008). URL <10.1007/s10664-008-9061-0> 96. Kitchenham, BA, Charters, S.: Linee guida per l'esecuzione di revisioni sistematiche della letteratura nell'ingegneria del software (versione 2.3). Rapporto tecnico, Rapporto tecnico EBSE EBSE-2007-01, Keele University e Durham University (2007)
97. Kitchenham, BA, Pickard, LM, Pfleeger, SL: Casi di studio per la valutazione di metodi e strumenti. *Software IEEE.* **12**(4), 52–62 (1995)
98. Kitchenham, BA, Pfleeger, SL, Pickard, LM, Jones, PW, Hoaglin, DC, El Emam, K., Rosenberg, J.: Linee guida preliminari per la ricerca empirica nell'ingegneria del software. *IEEE Trans. Softw. L'Ing.* **28**(8), 721–734 (2002). doi: 10.1109/TSE.2002.1027796. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=10277969>. Kitchenham, B., Fry, J., Linkman, SG: Il caso contro i progetti incrociati nell'ingegneria del software. In: Atti dell'11° workshop internazionale sulla tecnologia del software e sulla pratica ingegneristica, Amsterdam, pp. 65–67. IEEE Computer Society, Los Alamitos (2003)
100. Kitchenham, BA, Dyba, T., Jørgensen, M.: Ingegneria del software basata sull'evidenza. In: Atti della 26a conferenza internazionale sull'ingegneria del software, Edimburgo, pp. 273–281 (2004)
101. Kitchenham, BA, Al-Khilidari, H., Babar, MA, Berry, M., Cox, K., Keung, J., Kurniawati, F., Staples, M., Zhang, H., Zhu, L.: Valutazione delle linee guida per la rendicontazione di studi empirici di ingegneria del software. *Impero. Softw. L'Ing.* **13**(1), 97–121 (2007). doi: 10.1007/s10664-007-9053-5. URL <http://www.springerlink.com/index/10.1007/s10664-007-9053-5> 102. Kitchenham, BA, Jeffery, DR, Connaughton, C.: Metriche fuorvianti e analisi infondate. *Software IEEE.* **24**, 73–78 (2007). doi: 10.1109/MS.2007.49 103. Kitchenham, BA, Brereton, P., Budgen, D., Turner, M., Bailey, J., Linkman, SG: Revisioni sistematiche della letteratura nell'ingegneria del software – una revisione sistematica della letteratura. *Inf. Softw. Tecnologia.* **51**(1), 7–15 (2009). doi: 10.1016/j.infsof.2008.09.009. URL <http://www.dx.doi.org/10.1016/j.infsof.2008.09.009104>. Kitchenham, BA, Pretorius, R., Budgen, D., Brereton, P., Turner, M., Niazi, M., Linkman, S.: Revisioni sistematiche della letteratura nell'ingegneria del software – uno studio terziario. *Inf. Softw. Tecnologia.* **52**(8), 792–805 (2010). doi: 10.1016/j.infsof.2010.03.006 105. Kitchenham, BA, Sjøberg, DIK, Brereton, P., Budgen, D., Dyba, T., Host, M., Pfahl, D., Runeson, P.: Possiamo valutare la qualità degli esperimenti di ingegneria del software? In: Atti del 4° Simposio internazionale ACM-IEEE sull'ingegneria e la misurazione del software empirico. ACM, Bolzano (2010)
106. Kitchenham, BA, Budgen, D., Brereton, P.: Utilizzo degli studi di mappatura come base per ulteriori ricerche: un caso di studio partecipante-osservatore. *Inf. Softw. Tecnologia.* **53**(6), 638–651 (2011). doi: 10.1016/j.infsof.2010.12.011

107. Laitenberger, O., Atkinson, C., Schlich, M., El Emam, K.: An sperimentale confronto di tecniche di lettura per il rilevamento dei difetti nei documenti di progettazione UML. *J.Sist. Softw.* 53(2), 183–204 (2000)
108. Larsson, R.: Metodologia dell'indagine sui casi: analisi quantitativa dei modelli attraverso i casi di studio. *Accade. Gest. J.* 36 (6), 1515–1546 (1993)
109. Lee, AS: Una metodologia scientifica per casi di studio MIS. *MIS Q.* 13(1), 33 (1989). doi: 10.2307/248698. URL <http://www.jstor.org/stable/248698?origin=crossref>
110. Lehman, MM: Programma, cicli di vita e leggi dell'evoluzione del software. *Proc. IEEE* 68(9), 1060–1076 (1980)
111. Lethbridge, TC, Sim, SE, Singer, J.: Studiare gli ingegneri del software: tecniche di raccolta dati per studi sul campo del software. *Impero. Softw. L'Ing.* 10, 311–341 (2005)
112. Linger, R.: Modelli di processo della camera bianca. *Software IEEE.* pagine 50–58 (1994)
113. Linkman, S., Rombach, HD: La sperimentazione come veicolo per il trasferimento di tecnologia software – una famiglia di tecniche di lettura del software. *Inf. Softw. Tecnologia.* 39(11), 777–780 (1997)
114. Lucas, WA: Il metodo dell'indagine del caso: aggregazione dell'esperienza del caso. Rapporto tecnico, R-1515-RC, The RAND Corporation, Santa Monica (1974)
115. Lucas, HC, Kaplan, RB: un esperimento di programmazione strutturata. *Calcola. J.* 19(2), 136–138 (1976)
116. Lyu, MR (a cura di): Manuale di ingegneria dell'affidabilità del software. McGraw-Hill, New York (1996)
117. Maldonado, JC, Carver, J., Shull, F., Fabbri, S., Doria, E., Martimiano, L., Mendonc,a, M., Basili, V.: Lettura prospettica: una replica esperimento focalizzato sull'efficacia del singolo revisore. *Impero. Softw. L'Ing.* 11, 119–142 (2006). doi: 10.1007/s10664-006-5967-6
118. Manly, BFJ: Metodi statistici multivariati: A Primer, 2a ed. Chapman e Hall, Londra (1994)
119. Marascuilo, LA, Serlin, RC: Metodi statistici per le scienze sociali e comportamentali. WH Freeman and Company, New York (1988)
120. Miller, J.: Stima del numero di difetti rimanenti dopo l'ispezione. *Softw. Test. Verif. Affidabile.* 9(4), 167–189 (1999)
121. Miller, J.: Applicazione di procedure meta-analitiche a esperimenti di ingegneria del software. *J.Sist. Softw.* 54(1), 29–39 (2000)
122. Miller, J.: Test di significatività statistica: una panacea per gli esperimenti sulla tecnologia software? *J. Sist. Softw.* 73, 183–192 (2004). doi: <http://dx.doi.org/10.1016/j.jss.2003.12.019>
123. Miller, J.: Replicare esperimenti di ingegneria del software: un calice avvelenato o il Santo Graal. *Inf. Softw. Tecnologia.* 47(4), 233–244 (2005)
124. Miller, J., Wood, M., Roper, M.: Ulteriori esperienze con scenari e liste di controllo. *Impero. Softw. L'Ing.* 3(1), 37–64 (1998)
125. Montgomery, DC: Progettazione e analisi degli esperimenti, 5a edizione. Wiley, New York (2000)
126. Myers, GJ: Un esperimento controllato nel test del programma e nelle procedure dettagliate/ispezioni del codice. *Comune. ACM* 21, 760–768 (1978). doi: <http://doi.acm.org/10.1145/359588.359602>
127. Noblit, GW, Hare, RD: Meta-Ethnography: Synthesizing Qualitative Studies. Pubblicazioni Sage, Newbury Park (1988)
128. Ohlsson, MC, Wohlin, C.: Uno studio sulla stima dell'impegno progettuale. *Inf. Softw. Tecnologia.* 40(14), 831–839 (1998)
129. Owen, S., Brereton, P., Budgen, D.: Analisi del protocollo: una pratica trascurata. *Comune. ACM* 49(2), 117–122 (2006). doi: 10.1145/1113034.1113039
130. Paultk, MC, Curtis, B., Chrissis, MB, Weber, CV: Modello di maturità delle capacità per il software. Rapporto tecnico, CMU/SEI-93-TR-24, Software Engineering Institute, Pittsburgh (1993)
131. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Studi di mappatura sistematica nell'ingegneria del software. In: Atti della 12a Conferenza internazionale sulla valutazione e l'accertamento nell'ingegneria del software, Electronic Workshops in Computing (eWIC). BCS, Università di Bari, Italia (2008)
132. Petersen, K., Wohlin, C.: Contesto nella ricerca sull'ingegneria del software industriale. In: Atti del 3° simposio internazionale ACM-IEEE sull'ingegneria e la misurazione del software empirico, Lake Buena Vista, pp. 401–404 (2009)

133. Pfleeger, SL: Progettazione sperimentale e analisi nell'ingegneria del software, parte 1–5. ACM Sigsoft, Softw. L'Ing. Note, 19(4), 16–20; 20(1), 22–26; 20(2), 14–16; 20(3), 13–15; **20**, (1994)
134. Pfleeger, SL, Atlee, JM: Ingegneria del software: teoria e pratica, 4a ed. Pearson Prentice-Hall, Upper Saddle River (2009)
135. Pickard, LM, Kitchenham, BA, Jones, PW: Combinazione di risultati empirici nell'ingegneria del software. Inf. Softw. Tecnologia. 40(14), 811–821 (1998). doi: 10.1016/S0950-5849(98)00101-3
136. Porter, AA, Votta, LG: Un esperimento per valutare diversi metodi di rilevamento dei difetti per le ispezioni dei requisiti software. In: Atti della 16a Conferenza internazionale sull'ingegneria del software, Sorrento, pp. 103–112 (1994)
137. Porter, AA, Votta, LG: Confronto dei metodi di rilevamento per l'ispezione dei requisiti software: un esperimento replicato. IEEE Trans. Softw. L'Ing. 21(6), 563–575 (1995)
138. Porter, AA, Votta, LG: Confronto dei metodi di rilevamento per l'ispezione dei requisiti software: una sperimentazione replicata: una replica utilizzando soggetti professionali. Impero. Softw. L'Ing. 3(4), 355–380 (1998)
139. Porter, AA, Siy, HP, Toman, CA, Votta, LG: Un esperimento per valutare i costi-benefici delle ispezioni del codice nello sviluppo di software su larga scala. IEEE Trans. Softw. L'Ing. 23(6), 329–346 (1997)
140. Potts, C.: La ricerca sull'ingegneria del software rivisitata. Software IEEE. pagine 19–28 (1993)
141. Rainer, AW: La strategia di ricerca di casi di studio longitudinale e cronologica: una definizione e un esempio da IBM Hursley Park. Inf. Softw. Tecnologia. 53(7), 730–746 (2011)
142. Robinson, H., Segal, J., Sharp, H.: Studi empirici etnograficamente informati sulla pratica del software. Inf. Softw. Tecnologia. 49(6), 540–551 (2007). doi: 10.1016/j.infsof.2007.02.007 143. Robson, C.: Real World Research: A Resource for Social Scientists and Practitioners-Researchers, 1a ed. Blackwell, Oxford/Cambridge (1993)
144. Robson, C.: Real World Research: una risorsa per scienziati sociali e professionisti-ricercatori, 2a ed. Blackwell, Oxford/Madden (2002)
145. Runeson, P., Skoglund, M.: Strategie di ricerca basate sui riferimenti nelle revisioni sistematiche. In: Atti della 13a Conferenza internazionale sulla valutazione empirica e sulla valutazione nell'ingegneria del software. Laboratori elettronici di informatica (eWIC). BCS, Università di Durham, Regno Unito (2009)
146. Runeson, P., Host, M., Rainer, AW, Regnell, B.: Case Study Research nel software Ingegneria. Linee guida ed esempi. Wiley, Hoboken (2012)
147. Sandahl, K., Blomkvist, O., Karlsson, J., Krysander, C., Lindvall, M., Ohlsson, N.: An Extended Replication of an Experiment for Assessment Methods for Software Requirements. Impero. Softw. L'Ing. 3(4), 381–406 (1998)
148. Seaman, CB: Metodi qualitativi negli studi empirici dell'ingegneria del software. IEEE Trans. Softw. L'Ing. 25(4), 557–572 (1999)
149. Selby, RW, Basili, VR, Baker, FT: Sviluppo di software per camere bianche: un'analisi empirica valutazione. IEEE Trans. Softw. L'Ing. 13(9), 1027–1037 (1987)
150. Shepperd, M.: Fondamenti della misurazione del software. Prentice-Hall, Londra/New York (1995)
151. Shneiderman, B., Mayer, R., McKay, D., Heller, P.: Indagini sperimentali sull'utilità di diagrammi di flusso dettagliati nella programmazione. Comune. ACM **20**, 373–381 (1977). doi: 10.1145/ 359605.359610
152. Shull, F.: Sviluppo di tecniche per l'utilizzo di documenti software: una serie di studi empirici. Dottorato di ricerca tesi, Dipartimento di Informatica, Università del Maryland, USA (1998)
153. Shull, F., Basili, VR, Carver, J., Maldonado, JC, Travassos, GH, Mendonc,a, MG, Fabbri, S.: Replicating software engineering sperimentali: affrontare il problema della conoscenza tacita. In: Atti del primo simposio internazionale sull'ingegneria del software empirico, Nara, pp. 7–16 (2002)
154. Shull, F., Mendoncc,a, MG, Basili, VR, Carver, J., Maldonado, JC, Fabbri, S., Travassos, GH, Ferreira, MC: Knowledge-sharing Issues in Experimental Software Engineering. Impero. Softw. L'Ing. 9, 111–137 (2004). doi: 10.1023/B:EMSE.0000013516.80487.33

155. Shull, F., Carver, J., Vegas, S., Juristo, N.: Il ruolo delle repliche nell'ingegneria del software empirica. Impero. Softw. L'Ing. 13, 211–218 (2008). doi: 10.1007/s10664-008-9060-1 156. Sieber, JE: Protezione dei soggetti di ricerca, dei dipendenti e dei ricercatori: implicazioni per l'ingegneria del software. Impero. Softw. L'Ing. 6(4), 329–341 (2001)
157. Siegel, S., Castellan, J.: Statistiche non parametriche per le scienze comportamentali, 2a ed. Edizioni internazionali McGraw-Hill, New York (1988)
158. Singer, J., Vinson, NG: Perché e come l'etica della ricerca è importante per te. Sì, tu! Impero. Softw. L'Ing. 6, 287–290 (2001). doi: 10.1023/A:1011998412776 159. Singer, J., Vinson, NG: Questioni etiche negli studi empirici dell'ingegneria del software. IEEE Trans. Softw. L'Ing. 28(12), 1171–1180 (2002). doi: 10.1109/TSE.2002.1158289. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1158289>
160. Simon S.: L'ultimo teorema di Fermat. Quarto Stato, Londra (1997)
161. Sjøberg, DIK, Hannay, JE, Hansen, O., Kampenes, VB, Karasanovic, A., Liborg, N.- K., Rekdal, AC: A Survey of Controlled Experiments in Software Engineering. IEEE Trans. Softw. L'Ing. 31(9), 733–753 (2005). doi: 10.1109/TSE.2005.97. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1514443>
162. Sjøberg, DIK, Dyba, T., Anda, B., Hannay, JE: Building teorie in ingegneria del software. In: Shull, F., Singer, J., Sjøberg D. (a cura di) Guida all'ingegneria del software empirica avanzata. Springer, Londra (2008)
163. Sommerville, I.: Ingegneria del software, 9a ed. Addison-Wesley, Wokingham, Inghilterra/Lettura (2010)
164. Sørungard, S.: Verifica della conformità del processo negli studi empirici sullo sviluppo del software. Dottorato di ricerca tesi, Università norvegese di scienza e tecnologia, Dipartimento di informatica e scienza dell'informazione, Norvegia (1997)
165. Stake, RE: L'arte della ricerca sui casi di studio. Pubblicazioni SAGE, Thousand Oaks (1995)
166. Staples, M., Niazi, M.: Esperienze che utilizzano linee guida di revisione sistematica. J. Sist. Softw. 80(9), 1425–1437 (2007). doi: 10.1016/j.jss.2006.09.046
167. Thelin, T., Runeson, P.: Stime di cattura-ricattura per la lettura basata sulla prospettiva: un esperimento simulato. In: Atti della prima conferenza internazionale sul miglioramento dei processi software incentrati sul prodotto (PROFES), Oulu, pp. 182–200 (1999)
168. Thelin, T., Runeson, P., Wohlin, C.: Un confronto sperimentale tra lettura basata sull'uso e basata su checklist. IEEE Trans. Softw. L'Ing. 29(8), 687–704 (2003). doi: 10.1109/TSE.2003.1223644
169. Tichy, WF: Gli informatici dovrebbero sperimentare di più? Calcolo IEEE. 31(5), 32–39 (1998)
170. Tichy, WF, Lukowicz, P., Prechelt, L., Heinz, EA: Valutazione sperimentale nel computer scienza: uno studio quantitativo. J. Sist. Softw. 28(1), 9–18 (1995)
171. Trochim, WMK: Base di conoscenza dei metodi di ricerca, 2a ed. Cornell Custom Publishing, Cornell University, Ithaca (1999) 172. van Solingen, R., Berghout, E.: The Goal/Question/Metric Method: A Practical Guide for Quality Improvement and Software Development. McGraw-Hill Internazionale, Londra/Chicago (1999)
173. Verner, JM, Sampson, J., Tosic, V., Abu Bakar, NA, Kitchenham, BA: Linee guida per studi di casi multipli su base industriale nell'ingegneria del software. In: Terza conferenza internazionale sulle sfide della ricerca nella scienza dell'informazione, Fez, pp. 313–324 (2009)
174. Vinson, NG, Singer, J.: Una guida pratica alla ricerca etica che coinvolge gli esseri umani. In: Shull, F., Singer, J., Sjøberg, D. (a cura di) Guida all'ingegneria del software empirica avanzata. Springer, Londra (2008)
175. Votta, LG: Ogni ispezione necessita di un incontro? In: Atti del simposio ACM SIGSOFT sui fondamenti dell'ingegneria del software, ACM Software Engineering Notes, vol. 18, pp. 107–114. ACM Press, New York (1993)
176. Wallace, C., Cook, C., Summet, J., Burnett, M.: Linguaggi e ambienti informatici incentrati sull'uomo. In: Atti dei simposi sui linguaggi e ambienti di calcolo umano centrico, Arlington, pp. 63–65 (2002)

177. Wohlin, C., Gustavsson, A., Host, M., Mattsson, C.: Un quadro per l'introduzione della tecnologia nelle organizzazioni del software. In: Atti della conferenza sul miglioramento dei processi software, Brighton, pp. 167–176 (1996)
178. Wohlin, C., Runeson, P., Host, M., Ohlsson, MC, Regnell, B., Wessl en, A.: Experimentation in Software Engineering: An Introduction. Kluwer, Boston (2000)
179. Wohlin, C., Aurum, A., Angelis, L., Phillips, L., Dittrich, Y., Gorschek, T., Grahn, H., Henningsson, K., Kagstr om, S., Low, G., Roveg ard, P., Tomaszewski, P., van Toorn, C., Winter, J.: Fattori di successo che alimentano la collaborazione tra industria e mondo accademico nella ricerca sul software. Software IEEE. (Prestampa) (2011). doi: 10.1109/MS.2011.92
180. Yin, RK: Progettazione e metodi di ricerca di casi di studio, 4a ed. Pubblicazioni Sage, Beverly Colline (2009)
181. Zelkowitz, MV, Wallace, DR: Modelli sperimentali per la validazione della tecnologia. IEEE Calcola. 31(5), 23–31 (1998)
182. Zendler, A.: Una teoria preliminare dell'ingegneria del software investigata da esperimenti pubblicati. Impero. Softw. L'Ing. 6, 161–180 (2001). doi: <http://dx.doi.org/10.1023/A:1011489321999>

Indice

- Scala assoluta, [39](#)
- Trasformazione ammissibile, [38](#)
- Ipotesi alternativa, [91](#)
- Analisi e interpretazione, [80, 123](#)
- Analisi della varianza (ANOVA), [97, 98, 172](#)
 - Un fattore, più di due trattamenti, [143](#)
- Metodo analitico, [6](#)
- Dati
 - sull'anonimato, [35](#) partecipazione, [35](#)
 - Ricerca applicata, [111](#)
 - Presupposti dei test statistici, [104, 135](#)
 - Media, [124](#)
- Equilibrio, [95](#)
- Test binomiale, [133, 138](#)
- Blocco, [94](#)
- Box plot, [129, 131, 169](#)
- Ciclo di capitalizzazione, [27](#)
- Caso di studio, [10, 14, 55](#)
 - analisi dei dati, [65](#)
 - raccolta dati, [61](#)
 - pianificazione, [58](#) processo, [58](#) protocollo, [60](#) reporting, [69](#)
- Relazione casuale, [149](#)
- Tendenza centrale, [124](#)
- Chi- χ^2 , [130, 135](#)
 - bontà di adattamento, [147.000](#) campioni indipendenti, [146](#)
- Analisi dei cluster, [128](#)
- Coefficiente di variazione, [126](#)
- Riferimento aziendale, [15](#)
- Disegno completamente randomizzato, [95, 97](#)
- Validità della conclusione, [104](#)
- Riservatezza, [34](#)
- Effetti confondenti, [15](#)
- Fattori confondenti, [15, 149](#)
- Consenso, [118](#)
- Validità di costrutto, [108](#)
- Contesto, [11, 86, 89](#)
 - caso di studio, [56](#) in vitro, [25](#) in vivo, [25](#)
- Ciclo di controllo, [27](#)
- Campionamento di convenienza, [93](#)
- Coefficiente di correlazione, [128](#)
- Covarianza, [128](#)
- Design incrociato, [96, 151](#)
- Istogramma cumulativo, [130](#)
- Analisi
 - dei dati, [65](#)
 - raccolta, [61, 120](#)
 - riduzione, [131](#)
 - convalida, [121, 131](#)
- Dipendenza, [127](#)
- Variabile dipendente, [92](#)
- Statistica descrittiva, [123](#)
- Sintesi descrittiva, [50](#)
- Disegno
 - completamente randomizzato, [95, 97](#)
 - crossover, [96, 151](#)
 - fattoriale 2*2, [98](#)
 - fattoriale frazionario 2k , [99](#)
 - gerarchico, [98](#)
 - nidificato, [98](#) confronto appaiato, [96](#)

- blocco completo randomizzato, 97 test per, 136 annidati a due fasi, 98
 Principi di progettazione, 94
 Minacce progettuali, 108
 Divulgazione, 119
 Analisi discriminante, 128
 Dispersione, 126
 Dimensione dell'effetto, 38
 Metodi empirici, 6, 18
 Metodo ingegneristico, 6
 Comitato di revisione etica, 34
 Etica, 33, 118
 Controllo dell'esecuzione, 18
 Esercizi, xiv, 203
 Sperimentatore delle aspettative, 35, 110 variabile stocastica, 124
 Base dell'esperienza, 27
 Fabbrica dell'esperienza, 24, 27
 Modelli di esperienza, 27
 Esperimento, 11, 16, 73 progettazione, 75,
 93 orientato all'uomo, 16, 76 offline, 16, 90 on-line, 16, 90 processo, 76 reporting, 153 orientato alla tecnologia, 16, 76
 Attributo esterno, 41
 Validità esterna, 110
 Disegno fattoriale 2*2, 98 disegno fattoriale 2k , 99 disegno fattoriale frazionario 2k , 99 Prova F, 140
 Fattore, 75, 95
 Analisi fattoriale, 132
 Disegno fattoriale, 98
 Disegno fisso, 9, 76
 Progettazione flessibile, 9, 76
 Apezzamento di bosco, 50
 Disegno fattoriale frazionario, 99
 Frequenza, 126
 Metodo Obiettivo/Domanda/Metrico (GQM), 24, 85 Bontà di adattamento, 145 GQM.
Vedere Metodo obiettivo/domanda/metrica Visualizzazione grafica, 128
 Progettazione gerarchica, 98
 Istogramma, 130
 Ipotesi, 73, 91
 Verifica di ipotesi, 132
 Variabile indipendente, 92
 Incentivi, 119
 Consenso informato, 33, 34
 Strumentazione, 101, 107, 119
 Attributo interno, 41
 Validità interna, 106
 Scala degli intervalli, 39
 Intervistatore, 14
 Interviste, 13
 Costo dell'indagine, 18
 Coefficiente di correlazione dell'ordine di rango di Kendall, 128
 Kruskal-Wallis, 97, 144
 Regressione lineare, 127
 Studio longitudinale, 19
 Mann-Whitney, 96, 139
 Studi di mappatura, 23, 52
 Media aritmetica, 124
 geometrica, 125
 Affermazione significativa, 38
 Affermazione priva di significato, 38
 Misura, 37
 diretta, 41
 indiretta, 41
 oggettiva, 40
 soggettiva, 40
 valida, 38
 Misurazione, 37
 Controllo della misurazione, 18
 Mediana, 124
 Meta-analisi, 23, 48
 Metriche, 37
 Modo, 125
 Mortalità, 107
 Minacce di più gruppi, 107
 Analisi statistica multivariata, 73, 128
 Sintesi narrativa. *Vedi* Sintesi descrittiva Progettazione annidata, 98 Scala nominale, 39 Test non parametrici, 135

- Campione non probabilistico, 93
 Distribuzione normale, 130
 Normalità, 130, 135, 146
 Ipotesi nulla, 91, 132
- Oggetto, 75
 Oggetto di studio, 85
 Operazione, 80
 Scala ordinale, 39
 Valori anomali, 123, 129–131, 170
- Disegno a confronto appaiato, 96 Test parametrici, 135 Coefficiente di correlazione di Pearson, 128 Percentile, 125 Processo software personale (PSP), 161 Prospettiva, 86 Grafico a torta, 130 Pianificazione, 58, 78, 89 Popolazione, 92 Potenza, 92, 104, 134, 136 Presentazione e pacchetto, 80 Analisi delle componenti principali (PCA), 128, 132 Campione probabilistico, 93 Processo, 41 Prodotto, 41 PSP. *Vedi* Processo software personale (PSP) Distorsione nella pubblicazione, 47 Scopo, 85
- Ricerca qualitativa, 9
 Ricerca quantitativa, 9 Attenzione alla qualità, 85 Paradigma del miglioramento della qualità, 24, 26 Quasi-esperimento, 8, 11, 73, 86, 174 Questionari, 13 Campionamento per quote, 93
- Campionamento casuale, 93 Randomizzazione, 94 Disegno a blocchi completo randomizzato, 97 Gamma, 126 Scala di rapporto, 40 Frequenza relativa, 126 Affidabilità, 105 Replica, 18 ravvicinati, 20 differenziati, 20 Ridimensionamento, 38 Risorse, 41
- Convenienza di campionamento, 93 non probabilità, 93 probabilità, 93 quota, 93 casuale, 93 casuale stratificato, 93 sistematico, 93 Scala, 38 Tipo di scala, 39 Grafico a dispersione, 128 Metodo scientifico, 6 Scoping, 78, 85 Studi di scoping. Vedere Studi di mappatura Selezione di soggetti, 92, 107 interazioni, 107 Risultati sensibili, 118 Test dei segni, 96, 142 Simulazione, 5 Minacce a gruppo singolo, 106 Snowballing, 23, 47 Minacce sociali, 108, 109 Coefficiente di correlazione dell'ordine di classificazione di Spearman, 128 Deviazione standard, 126 Regressione statistica, 107 Test statistico unilaterale, 133 bilaterale, 133 Campionamento casuale stratificato, 93 Soggetti, 75, 92 Incentivo, 35 Studenti, 35, 90, 174 Sondaggio, 10, 12 descrittivo, 13 esplicativo, 13 esplorativo, 13 di sintesi, 22 descrittivo, 50 Revisioni sistematiche della letteratura, 22, 45 Campionamento sistematico, 93
- test t, 96, 138 accoppiati, 96, 141 Trasferimento tecnologico, 30 Prova, 76, 94 Teoria dell'ingegneria del software, 21 test, 111 Minacce alla validità, 102 progettazione, 108 gruppo multiplo, 107 priorità, 111

- unico gruppo, 106
- sociale, 108, 109
- Trasformazione, 38 , 127
- Trattamento, 75, 95
- Design nidificato a due stadi, 98
- Errore di tipo I, 91
- Errore di tipo II, 91
- Validità, 68 , 102, 104, 111
 - conclusione, 104
 - costruire, 108
- Variabile, 74 , 92
 - dipendente, 74 , 92
 - indipendente, 74 , 92
 - risposta, 74
- Varianza, 126
- Intervallo di variazione, 126
- Baffi, 129 , 169
- Wilcoxon, 96 , 142