UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior de Ingenieros Informáticos

Departamento de Lenguajes y Sistemas Informáticos e
Ingeniería de Software

Evidencias de la Existencia de Errores Estadísticos en Experimentos de Ingeniería del Software: Cuatro Estudios Empíricos

TESIS DOCTORAL

Tesis doctoral que presenta para optar al título de

Doctor en Software, Sistemas y Computación

Autor: Rolando Patricio Reyes Chicango

Madrid, 2018

Esta tesis doctoral se presenta al ETSI Informáticos de la Universidad Politécnica de Madrid en cumplimiento de los requisitos para el grado de Doctor en Software, Sistemas y Computación

Tesis Doctoral

Doctorado en Software, Sistemas y Computación

Título de la tesis doctoral: **Evidencias de la Existencia de Errores Estadísticos en Experimentos de Ingeniería del Software: Cuatro Estudios Empíricos**

Autor: Rolando Patricio Reyes Chicango

Universidad del título presentado: Universidad Politécnica de Madrid

Marzo 2018

Directores:

Natalia Juristo

Doctora en Informática

Dpto. Lenguajes y Sistemas Informáticos e Ingeniería de Software

Universidad Politécnica de Madrid

Óscar Dieste Tubío

Doctor en Informática

Dpto. Lenguajes y Sistemas Informáticos e Ingeniería de Software

Universidad Politécnica de Madrid

ETSI Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n
28660 Boadilla del Monte (Madrid)
Spain

*El presente trabajo de Investigación Doctoral
y
todo el esfuerzo que representa,
esta dedicado a mi madre Inés y
a mi futura esposa Johana
y
todas aquellas personas
que me apoyaron en todo momento.*

Rolando.

Agradecimientos

Primeramente quiero agradecer a Dios y a la persona más importante, mi madre, de quién siempre he sentido su apoyo, y de quién sin sus consejos y fuerza no hubiera podido iniciar y concluir con este trabajo doctoral.

A Dña. Natalia Juristo, D. Oscar Dieste y D. Efraín Fonseca, quienes desde el primer momento, han mostrado una apertura incondicional, dándome la oportunidad de trabajar a su lado, y de quienes han sembrado en mí sus conocimientos y sabiduría. Gracias por su inmensa paciencia, sus orientaciones, sus consejos, su tiempo y por la confianza depositada para la realización de este trabajo de tesis doctoral.

Agradezco al Grupo de Investigación de Ingeniería de Software Empírica (GRISE) de la Universidad Politécnica de Madrid, por concederme el privilegio de participar en sus investigaciones.

También quiero agradecer a la Armada del Ecuador, noble institución a la cual orgullosamente pertenezco, así como a la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), que me brindaron la oportunidad de iniciar mi camino en la investigación científica en Ingeniería de Software Experimental.

Finalmente, a todas y cada una de las personas que de una u otra manera han hecho posible este logro,

Muchas gracias.

Resumen

Contexto: En 2005, J.P.A. Ioannidis advertía a los investigadores que muchos resultados experimentales, independientemente del campo científico, son probablemente falsos. Las razones son múltiples; algunas de ellas están relacionadas con errores estadísticos. Las disciplinas experimentalmente maduras (p. ej., Medicina y Psicología) monitorizan con frecuencia la calidad y rigor de sus investigaciones.

Problema: En Ingeniería de Software (IS), con contadas excepciones, apenas se ha prestado atención a la existencia de errores estadísticos en sus trabajos experimentales. Una de las pocas excepciones es la reciente investigación de Jørgensen et al. Este trabajo demuestra la existencia de *publication bias* y *research bias* en IS. Ambos *biases* tienen una naturaleza esencialmente estadística. Aunque los autores relacionan dichos *biases* con la ausencia de poder estadístico de los experimentos, cabe preguntarse si también podrían deberse a la existencia de malas prácticas en IS.

Objetivos: Explorar, identificar, analizar y evaluar los errores estadísticos en IS, con el fin de establecer sus causas y, a partir de las mismas, proponer mecanismos para evitarlos o, al menos, minimizar su impacto.

Metodología: Hemos realizado una investigación basada en métodos mixtos: (1) revisiones sistemáticas para construir una tipología de errores y sus respectivas causas, (2) revisiones y encuestas para identificar la existencia de errores estadísticos experimentos de IS, y (3) entrevistas con experimentadores para establecer claramente las causas de los errores y alternativas de solución.

Resultados: Nuestra investigación pone de manifiesto que las malas prácticas no son las causas principales de los errores estadísticos en IS. En su lugar, las causas más relevantes parecen ser: a) El insuficiente conocimiento, por parte de los investigadores, de la metodología experimental y técnicas estadísticas asociadas, así como b) el insuficiente conocimiento, también por parte de los investigadores, de los fenómenos que ocurren en nuestra disciplina, lo que lleva a la frecuente realización de investigaciones de carácter exploratorio.

Conclusiones: Al menos en el actual estadio de evolución de la experimentación en IS, parece que los errores estadísticos se producen por un problema de *inmadurez* de la propia disciplina. En consecuencia, las alternativas de solución van en la dirección de atacar dicha inmadurez, proporcionando a los investigadores: a) Formación adecuada y b) guías claras de como realizar y analizar experimentos.

Abstract

Context: In 2005, J.P.A. Ioannidis warned researchers that many experimental results are probably false, regardless of the scientific field. The reasons are multiple, but some of them are related to statistical errors. The quality and rigor of research in mature experimental disciplines (e.g.: Medicine, Psychology, etc.) are frequently monitored.

Problem: With few exceptions, little attention has been paid to the presence of statistical errors in Experimental Software Engineering. One of these few exceptions is the recent research made by Jørgensen et al., which demonstrates the existence of bias in publications and research on Software Engineering (SE). These biases are essentially of statistical nature. Although Jørgensen et al. relate these biases with the absence of statistical power, it is worth asking whether biases in SE could be due to the existence of bad practices.

Aim: The aim of the research is to explore, identify, analyze and evaluate statistical errors in experimental SE, in order to identify their causes, and propose mechanisms to avoid them or, at least, minimize their impact.

Research Method: We conducted a research based on mixed methods: (1) systematic reviews to construct a typology of errors and their underlying causes, (2) surveys to identify the existence of statistical errors in SE experiments, and (3) interviews with experimenters to clearly establish the causes of the errors and alternatives for solution.

Results: Our research shows that bad practices are not the main cause of statistical errors in experimental SE. Instead, the most relevant causes seem to be: a) researchers' lack of knowledge about the experimental methodology and the associated statistical techniques, and b) researchers' lack of knowledge about the phenomena that occur in experimental SE, which in turn lead to the conduction of exploratory research.

Conclusions: It seems that statistical errors are produced by the immaturity of the experimental SE discipline itself, at least in the current stage of its evolution. Consequently, the alternatives for solution point in the direction to address such immaturity, providing researchers with: a) adequate training, and b) clear guidelines on how to perform and analyze experiments.

Lista de Publicaciones

- I *Statistical Errors in Experimental Software Engineering: A Preliminary Literature Review*, Oscar Dieste, Rolando P. Reyes Ch., Efraín R. Fonseca C., Natalia Juristo, **aceptado en ICSE 2018**.
- II *Statistical Errors in Experimental Disciplines: A Systematic Literature Review*, Rolando P. Reyes Ch., Oscar Dieste, Efraín R. Fonseca C., Natalia Juristo, *en desarrollo, para enviar a la revista IST en Septiembre 2018*.
- III *How Hypothesis Testing is Used at ESEM and What Researchers Say About It: Do We Practice What We Preach?*, Rolando P. Reyes Ch., Oscar Dieste, Efraín R. Fonseca C., Natalia Juristo, *en revisión por directores, para enviar a la revista EMSE en Marzo 2018*.
- IV *Pruebas de Hipótesis en Experimentos de Ingeniería de Software: Desviaciones Respecto a la Teoría*, Rolando P. Reyes Ch., Oscar Dieste y Efraín R. Fonseca C., **publicado en CIBSE - XIX Ibero-American Conference on Software Engineering, Abril 2016, pages 157-170**.
- V *Statistical Errors with Hypothesis Testing in Experimental Practice. Which way to go?: An Interview with the Expert*, Rolando P. Reyes Ch., Oscar Dieste y Natalia Juristo, *en desarrollo, previsto su envío al congreso ESEM en Mayo 2018*.

Otras Publicaciones

- VI *Defectos Metodológicos de la Práctica Experimental en Ingeniería de Software: Causas, Consecuencias y Algunas Direcciones para su Solución*, Rolando P. Reyes Ch., **Publicado en Simposio Doctoral, CIBSE - XX Ibero-American Conference on Software Engineering, Mayo 2017, pages 42-49**.

Contribuciones de Autor

En los artículos I-V, el doctorando contribuyó con la investigación y elaboración de los artículos, así como en la realización del estudio de campo, recopilación de datos y análisis. A continuación se detallan las contribuciones del doctorando por cada artículo:

- En los artículos I y II, el doctorando llevó acabo una gran parte de la Revisión Sistemática de Literatura (SLR), en concreto la obtención y piloteo de las cadenas de búsqueda, selección de estudios primarios, extracción, recopilación de datos, análisis y síntesis.

Asimismo, el doctorando contribuyó a la tipología preliminar de errores estadísticos y a la obtención de los catálogos de errores estadísticos y sus causas más probables.

- Para los artículos III, IV y V, el doctorando realizó el trabajo de campo. Para ello, creó, piloteo y refinó los cuestionarios y las guías de las entrevistas. También realizó las sesiones de encuestas y entrevistas a varios investigadores que han publicado artículos en de ICSE, ESEM, EMSE y TSE. El análisis, síntesis y obtención de resultados fue realizado por el doctorando, elaborando esquemas, clasificaciones, árboles y tipologías.
- El reporte de todos artículos fue llevado por el doctorando bajo la supervisión de sus directores.

Índice

1. INTRODUCCIÓN	1
1.1. Área de investigación	2
1.2. Problema de investigación	3
1.3. Importancia del problema	4
1.4. Objetivo de la investigación	4
1.5. Preguntas de investigación	5
1.6. Ejecución	6
1.7. Resultados	6
1.8. Estructura de la tesis	8
2. ESTADO DE LA CUESTIÓN	9
2.1. Situación actual en IS	10
2.2. Recopilaciones de errores estadísticos	11
2.3. Estudios pertenecientes a las ciencias experimentales	12
2.4. Estudios pertenecientes a la Ingeniería de Software Experimental	15
2.5. Características del estado de la cuestión	17
2.5.1. Evidencias de la existencia de errores estadísticos	18
2.5.2. Tipos de errores estadísticos	18
2.5.3. Alternativas de solución	20
2.5.4. Características principales identificadas	20
3. PLANTEAMIENTO DEL PROBLEMA Y OBJETIVOS	23
3.1. Planteamiento del problema	24
3.2. Objetivos de investigación	25
3.2.1. Objetivo general	25
3.2.2. Objetivos específicos	26
4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO	29
4.1. Diseños de investigación usados comúnmente en la investigación IS	30
4.1.1. Diseño de investigación cuantitativa	30
4.1.2. Diseño de investigación cualitativa	31
4.1.3. Diseño de investigación de métodos mixtos	33
4.2. Proceso de selección del método de investigación	34

4.2.1. Identificación de los desafíos planteados por los objetivos de investigación	34
4.2.2. Evaluación de los métodos de investigación aplicables . . .	36
4.3. Plan de trabajo	40
4.3.1. Etapa 1 - Exploración	41
4.3.2. Etapa 2 - Análisis	43
4.3.3. Etapa 3 - Diagnóstico	45
5. EJECUCIÓN Y RESULTADOS	48
5.1. Objetivo 1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	49
5.1.1. RQ1.1: ¿Qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales? . .	49
5.1.2. RQ1.2: ¿Hasta qué punto los errores estadísticos ocurren también en la IS?	52
5.2. Objetivo 2: Determinar la naturaleza de los errores estadísticos cometidos en IS.	55
5.2.1. RQ2.1: ¿Qué problemas se observan en la aplicación de las técnicas estadísticas en IS?	55
5.2.2. RQ2.2: ¿Es la comunidad de IS consciente de la existencia del problema?	58
5.3. Objetivo 3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	61
5.3.1. RQ3.1: ¿Cuáles son las causas de los errores estadísticos en las disciplinas experimentales?	61
5.3.2. RQ3.2: ¿Cuales son las causas más probables de los errores estadísticos en IS?	63
5.3.3. RQ3.3: ¿Qué medidas se pueden tomar Para evitar la comisión de errores?	66
6. Discusión	68
6.1. Discusión	69
6.1.1. Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	69
6.1.2. Determinar la naturaleza de los errores estadísticos cometidos en IS.	70
6.1.3. Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia . . .	72
7. AMENAZAS A LA VALIDEZ	75
7.1. Validez interna	76
7.2. Validez externa	78
7.3. Validez del constructo	80
7.4. Aspectos de fiabilidad	82

8. Conclusiones y futuras líneas de investigación	87
8.1. Conclusiones	88
8.2. Líneas futuras de investigación	89
Bibliografía	91

Índice de Figuras

2.1. Características del Estado de la Cuestión	22
3.1. Relación entre las Características del Estado de la Cuestión y la Problemática de Investigación.	25
3.2. Contribución de los Objetivos y Preguntas de Investigación con la Resolución de la Problemática Identificada.	28
4.1. Diseño secuencial de métodos mixtos	41
4.2. Plan de Trabajo Utilizado en el Tesis Doctoral	42
5.1. Clasificación de Errores Estadísticos en Artículos de Investigación Experimental	50
5.2. Relación de Problemas, Objetivos, Hipótesis de Investigación e Hipótesis Estadísticas y Colas	56
5.3. (a) Evolución Temporal de Estudios Primarios (b) Causas que son Nombradas por las Disciplinas	63

Índice de Tablas

1.1. Referencias de los artículos obtenidos durante la tesis doctoral . .	7
1.2. Mapeo entre los objetivos de investigación y los resultados de la investigación.	8
2.1. Tópicos considerados en el libro de [Good and Hardin, 2012] . . .	11
2.2. Tópicos considerados en el libro de [Huck, 2009]	12
2.3. Tópicos considerados en el libro de [Vickers, 2010]	12
2.4. Defectos estadísticos considerados en los trabajos de [Bakker and Wicherts, 2011, Haque and George, 2007]	14
2.5. Defectos estadísticos considerados en los trabajos de [Brewer, 1985]	15
2.6. Defectos estadísticos considerados en los trabajos de [Kilkenny et al., 2009]	15
4.1. Diseños de Investigación Usados Comúnmente en la Investigación IS	30
4.2. Desafíos para O1 y sus preguntas de investigación	35
4.3. Desafíos para O2 y sus preguntas de investigación	36
4.4. Desafíos para O3 y sus preguntas de investigación	37
4.5. Evaluación de métodos y herramienta aplicables para RQ 1.1 y RQ 1.2	38
4.6. Evaluación de métodos y herramienta aplicables para RQ2.1 y RQ 2.2	39
4.7. Evaluación de métodos y herramienta aplicables para RQ3.1 . . .	40
4.8. Evaluación de Métodos y Herramientas de Apoyo para RQ3.2 y RQ3.3	40
5.1. Lista de verificación de errores estadísticos. La pregunta Q7 no aparece en la Tabla por motivos metodológicos discutidos en el Artículo I	52
5.2. Porcentaje de errores estadísticos en publicaciones ICSE	53
5.3. Tabla cruzada de RQs vs. hipótesis de investigación	57
5.4. Tabla cruzada entre RQs vs. hipótesis estadística	57
5.5. Catálogo de errores estadístico de nivel de orden superior (High-theme)	62

5.6. Causas principales de la aparición de defectos metodológicos de pruebas de hipótesis	64
5.7. Medidas se pueden tomar para evitar la comisión de errores con pruebas de hipótesis	66
7.1. Amenazas de validez interna	79
7.2. Aspectos de validez externa	80
7.3. Aspectos de validez constructo	83
7.4. Aspectos de fiabilidad	86

1 | INTRODUCCIÓN

Resumen: El proceso experimental es considerado como el método más fiable para obtener conocimiento científico (p. ej.: [Box et al., 2008, Montgomery, 2010]); por esa razón se encuentra situado en el nivel superior de las jerarquías de evidencias [Pilcher and Bedford, 2011, Evans, 2003]. En Ingeniería de Software (IS), el proceso experimental es una realidad bien asentada; prueba de ello, es la amplia literatura al respecto, publicada en revistas y congresos de alto impacto (sección 1.1).

En las ciencias experimentales se presta especial atención al modo en que los investigadores usan las técnicas estadísticas durante el proceso experimental. Esta introspección ocurre tanto en disciplinas jóvenes (Educación) como asentadas (Medicina), y ha dado lugar al descubrimiento de numerosos errores estadísticos (sección 1.2). Por el contrario, apenas hemos encontrado literatura al respecto en IS, lo que supone un aparente desconocimiento o desinterés acerca del problema de la existencia de errores estadísticos y su impacto en los resultados de la investigación experimental. Uno de los pocos estudios existentes en IS es el de [Jørgensen et al., 2016], quien demostró la existencia de resultados cuestionables o probablemente falsos en IS (sección 1.3).

La presente tesis doctoral tiene como objetivo estudiar en qué medida los errores estadísticos afectan a la investigación experimental en IS (sección 1.4); para lo cual proponemos una aproximación exploratoria que se resume en explorar, identificar, analizar y evaluar los errores estadísticos que ocurren en la experimentación en IS (sección 1.5), y con ello poder establecer sus causas y proponer mecanismos que permitan evitar su comisión o, al menos, minimizar su impacto (secciones 1.7 y 1.6).

1.1. Área de investigación

Un experimento es una investigación formal, rigurosa y controlada donde un conjunto de variables bajo estudio (variables independientes o factores) toman distintos valores o niveles, con el fin de investigar los efectos (variables dependientes o respuesta) que se producen en cada nivel [Juristo and Moreno, 2001]. Para que los resultados de un experimento sean fiables, es necesario que se hayan realizado adecuadamente las etapas de planificación, ejecución y análisis [Wohlin et al., 2012]. De estas tres etapas, la planificación y análisis son las más importantes, ya que durante las mismas se asegura que todas las variables que puede afectar a los resultados del experimento se encuentran perfectamente controladas (factores), bloqueadas (variables de bloques), o aleatorizadas (el resto de potenciales variables). Adicionalmente, durante la etapa de planificación del experimento se asegura haber tomado en cuenta que todas las amenazas a la validez hayan sido bien comprendidas y adecuadamente afrontadas.

En IS, aparentemente, la experimentación es una metodología bien comprendida. Prueba de ello es la existencia de varios libros o guidelines que detallan claramente el proceso experimental en IS. Por ejemplo, los libros de [Juristo and Moreno, 2001] y [Wohlin et al., 2012] han establecido de forma sólida los principios de la experimentación en IS, junto con otros libros clásicos como: [Box et al., 2005], [Montgomery et al., 2006] y [Cook et al., 1979] que discuten en detalle aspectos de diseño, análisis y amenazas a la validez, respectivamente.

Asimismo, existen aspectos que apuntan a la madurez del proceso experimental, como son la existencia de una revista especializada (Empirical Software Engineering Journal - EMSE), conferencias temáticas en experimentación (International Symposium on Empirical Software Engineering and Measurement - ESEM), varios cursos reglados respecto a Ingeniería de Software Experimental, en distintas universidades, a nivel de doctorado e incluso de master (, Master EMSE-UPM).

Desconocemos el número exacto de experimentos que se han publicado hasta el momento en IS, pero puede cifrarse en cientos [Sjøberg et al., 2005, Bezerra et al., 2015]. Por ello, creemos poder afirmar que la aplicación de la experimentación ha sido un éxito en IS, situación similar a la que han experimentado otras disciplinas científicas clásicas (Física, Química, Medicina, etc.). Los experimentos, en sus distintas variantes (experimentos controlados, evaluaciones comparativas, cuasi-experimentos, etc.), se han utilizado en virtualmente todas las áreas de la IS, como por ejemplo: requerimientos [Aranda et al., 2016], diseño [Salman et al., 2015], implementación [Mäntylä et al., 2014], testing [Darwish et al., 2017],

mantenimiento [Fernández-Sáez et al., 2013], gestión de la configuración [Wang et al., 2017], herramientas-métodos [Madeyski and Kawalerowicz, 2017], calidad [Maier and Berry, 2017], etc. En pocas palabras, la investigación en IS se ha visto enormemente beneficiada con el proceso experimental obteniendo evidencias científicas de mayor fiabilidad [Evans, 2003, Basili et al., 1986].

1.2. Problema de investigación

Ahora bien, realizar adecuadamente la etapa de análisis de los datos experimentales, no es fácil. Se requiere de considerables conocimientos de matemáticas avanzadas (p. ej.: Álgebra Lineal, Estadística, etc) que, en líneas generales, no siempre están al alcance de los investigadores (independientemente de la disciplina, no específicamente en IS) [Garfield, 1995, Garfield and Ben-Zvi, 2007]. No alcanzar la suficiente experiencia, habilidad y conocimiento estadístico para enfrentar la etapa de análisis e interpretación de los datos supone una brecha entre la formalización del proceso experimental y la estadística [Smith III et al., 1994]. Esta brecha hace posible la comisión de errores estadísticos, que en muchos de los casos no son detectados en los procesos de revisión por pares, y que pueden poner en cuestión los resultados científicos obtenidos. Algo parecido ocurre respecto a la etapa de planificación. Las decisiones tomadas durante esta etapa, p. ej.: selección de variables, sujetos experimentales, métricas escala, entre otras; y sobre todo la formalización de las hipótesis del experimento (H_0 y H_1), tienen un gran impacto en los resultados experimentales [Wohlin et al., 2012]. Sin embargo, muchos investigadores ignoran la importancia de estos conceptos [Smith III et al., 1994].

Se han publicado, en las distintas ciencias, una considerable cantidad de trabajos donde se reconoce la existencia de errores estadísticos. La mayoría de estos artículos críticos se derivan de algún tipo de Revisión Sistemática de Literatura (SLR), donde evalúan artículos experimentales publicados en revistas tanto de alto como de bajo impacto, como por ejemplo: [Welch and Gabbe, 1996, Olsen, 2003, Ercan et al., 2008] en la Medicina, [Bakker and Wicherts, 2011, Nickerson, 2000] en la Psicología, entre otros. Las comisiones de errores estadísticos que comúnmente se mencionan en la literatura se refieren a conceptos sencillos como: tamaño de la muestra [Altman, 1998], p-valor [Altman, 1998], niveles de significación estadística [Ercan et al., 2008] y pruebas de hipótesis [Welch and Gabbe, 1996], entre otros.

Al contrario de lo que ocurre en otras disciplinas experimentales, hemos logrado encontrar muy poca información relacionada con errores estadísticos en IS, y los pocos trabajos encontrados son bastante especializados. Por ejemplo, [Dybå et al., 2006] con la potencia estadística, [Miller, 1999] con meta-análisis o [Vegas et al., 2016, Kitchenham et al., 2003] con análisis de diseños cross-over.

1.3. Importancia del problema

La diferencia entre IS y otras ciencias experimentales no puede ser más acusada. Mientras en Medicina, una ciencia experimental establecida y paradigmática, se han identificado múltiples comisiones de errores estadísticos en sus publicaciones, en IS este aspecto puede estar siendo pasado por alto, razón por la cual la discusión al respecto en la comunidad IS es casi inexistente. Bien es cierto que la falta de referencias a errores estadísticos en IS podría considerarse “algo normal y justificable”, dado que la IS Experimental es aún una disciplina joven y, al igual que ha sucedido con otras disciplinas experimentales en sus inicios, la atención que se presta a la comisión de errores estadísticos es muy limitada.

A este respecto, no deja de ser significativo el reciente artículo de [Jørgensen et al., 2016], donde se estudia la existencia de sesgos en estudios experimentales en IS. Los resultados indican claramente la existencia de “research bias” y “publication bias” en IS, los cuales a su vez tienen como origen una amplia diversidad de errores estadísticos. Creemos que lo mencionado por [Jørgensen et al., 2016] representa la punta del iceberg de un problema mucho más amplio, lo que podría suponer que la comunidad de IS se enfrenta a errores estadísticos comparables, o incluso peores, a lo reportado en otras ciencias experimentales.

Los errores estadísticos pueden manifestarse de distintos modos. No obstante, no estudiar este fenómeno, al menos de forma exploratoria, sería perjudicial a largo plazo para la comunidad de IS, debido a la acumulación de resultados cuestionables que podrían poner en peligro la misma práctica experimental en IS. Al respecto, no dejan de ser ilustrativas las advertencias de J.P.A. Ioannidis [Ioannidis, 2005], investigador en la Medicina, quien alertó que muchos resultados experimentales, independientemente del campo científico, son probablemente falsos debido a la cuestionable aplicación de la estadísticas así como al crecimiento de un entorno de investigación muy competitivo asociado a la cultura de “*publish or perish*”. Este es un grave problema que conviene afrontar desde este mismo momento en IS.

1.4. Objetivo de la investigación

El objetivo general de esta tesis se resume en *explorar, identificar, analizar y evaluar los errores estadísticos que puede ocurrir en la experimentación en IS, con el fin de establecer las causas más realistas y proponer mecanismos para evitarlos o, al menos, minimizar su impacto*. Más concretamente, hemos establecido tres objetivos concretos de investigación:

- **O1:** Identificar si existen errores estadísticos en IS, y establecer su preva-

lencia.

- **O2:** Determinar la naturaleza de los errores estadísticos cometidos en IS.
- **O3:** Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.

1.5. Preguntas de investigación

Para llevar a cabo una mejor gestión de la tesis doctoral, hemos desagregado los objetivos propuesto en preguntas de investigación (RQ) con el fin de alcanzar resultados tangibles. Las preguntas de investigación se detallan a continuación:

Para O1:

- **RQ1.1:** ¿Qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales?
- **RQ1.2:** ¿Hasta qué punto estos errores estadísticos ocurren también en la IS?

Para O2:

- **RQ2.1:** ¿Qué problemas se observan en la aplicación de las técnicas estadísticas en IS?
- **RQ2.2:** ¿Es la comunidad de IS consciente de la existencia del problema?

Para O3:

- **RQ3.1:** ¿Cuáles son las causas de los errores estadísticos en las disciplinas experimentales?
- **RQ3.2:** ¿Cuales son las causas más probables de los errores estadísticos en IS?
- **RQ3.3:** ¿Qué medidas se pueden tomar para evitar la comisión de errores?

Responder a las preguntas de investigación planteadas es complejo, debido a que la información sobre el fenómeno de estudio es casi inexistente en la comunidad de IS. En consecuencia creemos que la alternativa más razonable es realizar una investigación exploratoria [Runeson et al., 2012] de tal manera que observemos lo que está ocurriendo en la IS, abordando el problema planteado desde la perspectiva de las bases literarias de otras disciplinas, progresando posteriormente hacia la literatura en IS e incluso estudiando, la actuación de los propios investigadores en la práctica experimental.

Dada la poca definición y diversidad del fenómeno bajo investigación, hemos decidido aplicar una estrategia basada en métodos mixtos (cualitativos y cuan-

titativos) para lograr enfrentar los desafíos de nuestra aproximación exploratoria [Creswell, 2002]. Al respecto podemos indicar que la combinación de métodos (cualitativos y cuantitativos) permite proporcionar una comprensión y resultados más amplios del problema de investigación [Creswell and Clark, 2007].

1.6. Ejecución

Se elaboró un plan de trabajo que fue organizado en 3 etapas (*exploración, análisis y diagnóstico*), mediante las cuales fue posible obtener respuestas a las preguntas y objetivos de investigación.

La primera *etapa de exploración* tuvo la finalidad de explorar *errores comunes en otras disciplinas* considerando las bases literarias de una *Revisión de Literatura*. Establecidos los errores comunes, realizamos una *revisión de artículos* en IS. Se logró establecer preliminarmente *la existencia de los errores comunes de las disciplinas, también en la IS*. Esta etapa fue realizada en aproximadamente 8 meses.

La segunda *etapa de análisis* consistió en determinar formalmente los diferentes *problemas en la aplicación de técnicas estadísticas en IS*, para lo cual se utilizó una *revisión de artículos*. Con los resultados obtenidos de los problemas encontrados, se realizó una *encuesta a los experimentadores de IS* con el fin de establecer la aplicación de las técnicas estadísticas durante su práctica experimental y la *conciencia del problema en IS*. La etapa en cuestión fue desarrollada en aproximadamente 10 meses.

Finalmente, con la tercera *etapa de diagnóstico*, se establecieron las *causas a errores estadísticos en las disciplinas experimentales* a partir de las bases literarias de una *Revisión Sistemática de Literatura (SLR)*. Con los resultados obtenidos, sumados a los resultados de la *etapa de análisis*, realizamos una *entrevista a los experimentadores* donde logramos establecer las *causas a errores estadísticos en IS*. Utilizamos la misma entrevista para obtener las *medidas más realistas para evitar dichos errores estadísticos*. Esta etapa comprendió una duración de 12 meses.

El estudio de la temática de la tesis y la planificación de la investigación requirieron 6 meses, antes de comentar la investigación propiamente dicha.

1.7. Resultados

La presente tesis doctoral ha dado lugar a varios resultados que se encuentran plasmados en publicaciones indicados en la Tabla 1.1 los cuales pueden ser

revisados en un apéndice web en el siguiente enlace: <https://goo.gl/T4gSva>. Los principales resultados se detallan a continuación:

1. Una **tipología preliminar de errores estadísticos** obtenida a partir de la literatura estadística general, descrita en el **Artículo I**.
2. La **obtención de evidencias preliminares acerca de la existencia de errores estadísticos** en publicaciones de IS, descrita en el **Artículo I**.
3. Un **catálogo formal de las causas más probables de dichos errores estadísticos**, obtenido a partir de la literatura de las distintas ciencias experimentales, descrito en el **Artículo II**.
4. La **demonstración, más allá de una duda razonable, de la existencia de errores estadísticos (asociados a un error en particular: las pruebas de hipótesis)** en publicaciones de IS, descritas en los **Artículos III y IV**.
5. La **identificación de causas probables de aparición de errores estadísticos en la IS**, así como **recomendaciones específicas para evitar su comisión o, al menos, minimizar el impacto**, descritas en el **Artículo V**.

Tabla 1.1: Referencias de los artículos obtenidos durante la tesis doctoral

No.	Referencias
I	Statistical Errors in Experimental Software Engineering: A Preliminary Literature Review , Oscar Dieste, Rolando P. Reyes Ch., Efraín R. Fonseca C., Natalia Juristo, <i>aceptado en ICSE 2018</i>
II	Statistical Errors in Experimental Disciplines: A Systematic Literature Review , Rolando P. Reyes Ch., Oscar Dieste, Efraín R. Fonseca C., Natalia Juristo, <i>en desarrollo, para enviar a la revista IST en Septiembre 2018</i> .
III	How Hypothesis Testing is Used at ESEM and What Researchers Say About It: Do We Practice What We Preach? , Rolando P. Reyes Ch., Oscar Dieste, Efraín R. Fonseca C., Natalia Juristo, <i>para enviar a la revista EMSE en Marzo 2018</i> .
IV	Pruebas de Hipótesis en Experimentos de Ingeniería de Software: Desviaciones Respecto a la Teoría , Rolando P. Reyes Ch., Oscar Dieste y Efraín R. Fonseca C., <i>publicado en CIBSE - XIX Ibero-American Conference on Software Engineering, Abril 2016, pages 157-170</i> .
V	Statistical Errors with Hypothesis Testing in Experimental Practice. Which way to go?: An Interview with the Expert , Rolando P. Reyes Ch., Oscar Dieste y Natalia Juristo, <i>en desarrollo, previsto su envío al congreso ESEM en Mayo 2018</i> .

Los resultados obtenidos nos han permitido satisfacer todos los objetivos de investigación, tal y como se muestra en la Tabla 1.2. Una “r” indica que el resultado tiene relación con el objetivo de investigación. Una “P” indica que el resultado aporta en un alto porcentaje al objetivo de investigación.

Tabla 1.2: Mapeo entre los objetivos de investigación y los resultados de la investigación.

Objetivos	Resultado	1	2	3	4	5
	RQ's					
O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia	RQ1.1: ¿Qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales?.	P				
	RQ1.2: ¿Hasta qué punto estos errores estadísticos ocurren también en la IS?.	r				
O2: Determinar la naturaleza de los errores estadísticos cometidos en IS	RQ2.1: ¿Qué problemas se observan en la aplicación de las técnicas estadísticas en IS?.	r		P	P	
	RQ2.2: ¿Es la comunidad de IS consciente de la existencia del problema?.			P	P	
O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	RQ3.1: ¿Cuáles son las causas de los errores estadísticos en las disciplinas experimentales?.		P			
	RQ3.2: ¿Cuales son las causas más probables de los errores estadísticos en IS?.			P		P
	RQ3.3: ¿Qué medidas se pueden tomar para evitar la comisión de errores?.					P

1.8. Estructura de la tesis

La estructura de la tesis doctoral es la siguiente:

- En el capítulo 2 se analiza el estado de la cuestión referente a la problemática planteada.
- El planteamiento del problema y objetivos de la investigación de la tesis se describen en el Capítulo 3.
- En el Capítulo 4 se indica el método de investigación y el plan de trabajo.
- Un resumen de las ejecución y resultados se encuentra en el Capítulo 5.
- La discusión de los hallazgos obtenidos se muestra en el Capítulo 6.
- En el Capítulo 7, se describen las amenazas a la validez de la tesis.
- Las conclusiones y trabajos futuros se describen en el Capítulo 8.

2 | ESTADO DE LA CUESTIÓN

Resumen: En las disciplinas científicas con tradición experimental, es común que aparezcan estudios específicos acerca de la existencia de errores estadísticos en sus trabajos experimentales. Ejemplos de estos estudios son: Educación [Brewer, 1985], Medicina o Ciencias de la Salud [Welch and Gabbe, 1996, Altman, 1998, Olsen, 2003, Ercan et al., 2008], Psicología [Bakker and Wicherts, 2011, Nickerson, 2000], entre otros. La IS no es ajena a esta tendencia, aunque la reciente adopción de la experimentación como metodología de investigación produce que los trabajos acerca de la existencia e impacto de los errores estadísticos sean solo incipientes.

En este Capítulo, descubriremos recopilaciones de errores estadísticos (sección 2.2) y trabajos que reportan errores estadísticos en las distintas ciencias (sección 2.3), así como en IS (sección 2.4). La descripción será necesariamente parcial, limitándonos algunos trabajos bien conocidos de las disciplinas, ya que no existen trabajos de síntesis a este respecto. Finalizaremos el capítulo con una síntesis de características del estado de la cuestión (sección 2.5) para explicar la problemática en el Capítulo 3.

2.1. Situación actual en IS

La incorporación de la experimentación a la IS se ha realizado mediante libros como el de [Juristo and Moreno, 2001] y [Wohlin et al., 2012], complementados con otros libros más clásicos como: [Box et al., 2005], [Montgomery, 2010] y [Cook et al., 1979]. No obstante, la existencia de estos libros no parece suficiente para que los investigadores puedan dominar el proceso experimental. Esto es particularmente grave en el caso de la estadística. La experimentación exige utilizar métodos y técnicas matemáticas complejas, que no están al alcance de los investigadores (en general, no específicamente en IS). La reciente adopción de la experimentación en IS agrava este problema. En muchos casos, los actuales investigadores de IS son autodidactas. Todo ello podría significar que existe una brecha entre la formalización de la experimentación en IS y la aplicación correcta de las técnicas estadísticas.

Esta brecha ha sido observada en muchas disciplinas experimentales, p. ej., Educación [Brewer, 1985], Medicina o Ciencias de la Salud [Welch and Gabbe, 1996, Altman, 1998, Olsen, 2003, Ercan et al., 2008], Psicología [Bakker and Wicherts, 2011, Nickerson, 2000], independientemente de su grado de madurez. Es por ello la disponibilidad de un conjunto relativamente grande de publicaciones que brindan información sobre la existencia de problemas estadísticos en prácticamente todas las disciplinas. No todas las publicaciones son recientes; han sido publicadas desde la adopción generalizada de la investigación experimental en sus respectivas áreas. Los problemas informados tienen un amplio alcance [Midi et al., 2012], que incluyen: la definición de hipótesis estadísticas [Cho and Abe, 2013, Nickerson, 2000], la interpretación de los valores p [Nuzzo et al., 2014], el cálculo del tamaño de la muestra [Altman, 1998, Ercan et al., 2008], la importancia de los niveles [Nickerson, 2000], e intervalos de confianza [Cho and Abe, 2013], entre otros.

En la literatura de IS, apenas existe información que aborde la problemática de los errores estadísticos. Los trabajos existentes se describen en la sección 2.4. En contrapartida a lo mencionado anteriormente, existe un gran número de trabajos en otras ciencias. Hemos seleccionado algunos de ellos, los cuales se describen a continuación en las secciones 2.2 y 2.3. Para la selección de los trabajos en esta sección, hemos utilizado un proceso de búsqueda no sistemático basado en criterios de inclusión y exclusión¹ con el fin de establecer un grupo de artículos de las disciplinas experimentales (denominado “Quasi-Gold Estandar”), que hemos posteriormente utilizado para la realización de una SLR. La creación del “Quasi-Gold Estandar” fue realizado por tres investigadores del Grupo de Investigación en Ingeniería de Software Empírica (GRISE), quienes mediante técnicas de screening (p. ej., título, abstract, palabras claves) y alcanzando un alto valor

¹Los criterios de inclusión y exclusión se encuentran basados en nuestra pregunta de investigación RQ3.1, detallada en la sección 3.2.2

de acuerdo entre investigadores, validaron la idoneidad de los artículos.

2.2. Recopilaciones de errores estadísticos

La mayor parte de los libros de texto de temática estadístico-experimental (p. ej., [Basili, 1993], [Box et al., 2005], [Montgomery, 2010], entre otros), tienen como objetivo enseñar al lector el uso de la metodología experimental y las técnicas estadísticas asociadas. Sin embargo, existen algunos libros como los de [Good and Hardin, 2012], [Huck, 2009] y [Vickers, 2010], que han tratado de elaborar, además de la enseñanza de los conceptos metodológicos, una serie de advertencias acerca de errores estadísticos que ocurren comúnmente en la práctica, así como la forma de poder reconocerlos y evitarlos.

Dichos libros [Good and Hardin, 2012, Vickers, 2010, Huck, 2009] suponen un buen punto de partida para la tesis doctoral, debido a que: (1) no están relacionados con ninguna disciplina específica (aunque existe cierto sesgo hacia la ciencias de la salud) y (2) se centran en errores estadísticos graves, en muchos casos inspirados en la investigación experimental real. A continuación, describiremos brevemente libros mencionados:

[Good and Hardin, 2012]

Se enfoca principalmente en errores estadísticos que suelen aparecer por mitos populares o usos inapropiados de las técnicas estadísticas durante la planificación (diseño) y análisis experimental. Es interesante señalar que los autores no especifican, ni catalogan, ni clasifican a detalle los errores estadísticos que estudian. Sin embargo, se deduce que los autores utilizan una clasificación basada en la ocurrencia del error estadístico dentro de las fases experimentales (diseño, análisis, interpretación, etc). En total, en este libro se discuten aproximadamente 15 tópicos que engloban diferentes tipos de errores estadísticos, tal y como se muestra en la Tabla 2.1.

Tabla 2.1: Tópicos considerados en el libro de [Good and Hardin, 2012]

Fuentes de error	Hipótesis	Recopilación de datos
Evaluación de la calidad de los datos	Estimación	Pruebas de Hipótesis
Fortalezas y Limitaciones de procedimientos estadísticos	Presentación de resultados	Interpretación de informes
Gráficos	Regresión univariada	Métodos alternativos de regresión
Regresión multivariable	Modelado de cuentas y datos correlacionados	Validación

[Huck, 2009]

Este autor discute aproximadamente 52 tipos de errores estadísticos. La mayoría de ellos se refieren a errores en conceptos y técnicas bastantes sencillas, tales como: pruebas de hipótesis, tamaño de la muestra, t-test, entre otros. En todos los casos, el autor aporta referencias bibliográficas donde se cometen explícitamente los errores estadísticos bajo estudio. Asimismo, se plantea el modo correcto de utilización y como evitar el error estadístico. La organización de los errores estadísticos en este libro se encuentra dividido en 12 tópicos, tal y como se muestran en la Tabla 2.2.

Tabla 2.2: Tópicos considerados en el libro de [Huck, 2009]

Estadística descriptiva	Probabilidad	T test
Distribución estadística	Muestreo	ANOVA Y ANCOVA
Correlación biviariada	Estimación	Poder estadístico
Confiabilidad y validez	Prueba de hipótesis	Regresión

[Vickers, 2010]

El autor de este libro trata de enseñar y explicar los errores estadísticos basándose en ejemplos que suelen aparecer en la vida diaria, sin profundizar en la conceptualización estadística y matemática de cada error estadístico. Su enfoque literario es bastante dinámico, resaltando la ejemplaridad y practicidad para quienes no tienen grandes conocimientos estadísticos. Debido a ello, el autor resume en el último capítulo los errores estadísticos abordados, los cuales son clasificados 5 tipos de errores detallados en la Tabla 2.3.

Tabla 2.3: Tópicos considerados en el libro de [Vickers, 2010]

Describing Data	Data Distributions
Variation of Study Results: Confidence Intervals	Hypothesis Testing
Some Common Statistical Errors, and What They Teach Us	

2.3. Estudios pertenecientes a las ciencias experimentales

En la comunidad científica abundan los estudios que alertan acerca de la existencia de errores estadísticos en sus respectivas comunidades, como son: Educación [Brewer, 1985], Medicina o Ciencias de la Salud [Welch and Gabbe, 1996, Altman, 1998, Olsen, 2003, Ercan et al., 2008], Psicología [Bakker and Wicherts,

2011, Nickerson, 2000], entre otras comunidades. La mayoría de estos artículos críticos han derivado sus resultados a partir de estudios primarios obtenidos mediante Revisiones Sistemática de Literatura (SLR). Las revisiones evalúan errores estadísticos en revistas y publicaciones de alto y bajo impacto en sus respectivas comunidades. Los errores estadísticos reportados son muy diversos y nada complejos; por mencionar algunos ejemplos: pruebas estadísticas inadecuadas, reporte de resultados erróneos, no reportar valores atípicos, introducir sesgos en los resultados, etc.

[Welch and Gabbe, 1996, Olsen, 2003, Ercan et al., 2008]

Son artículos críticos que pertenecen a la Medicina. En estos trabajos se analizaron **145, 141 y 164 artículos experimentales** respectivamente, pertenecientes a revistas de alto y bajo impacto en su comunidad como: *American Journal of Obstetrics and Gynecology*² y *Journal of Advanced Nursing*³, entre otros. Los resultados demostraron que más de la mitad de artículos revisados (52.6 %, 54 % y 50 % respectivamente) contenían **errores estadísticos en el análisis** (p. ej., muestreo, tipo de muestreo, otros.), **descripciones estadísticas inadecuadas o incompletas, falencias en el reporte** (p. ej., medias y desviación estándar) y uso de **métodos estadísticos inadecuados** para probar la hipótesis estadística (p. ej., t-test, repeated-measures ANOVA, Wilcoxon signed-rank test). Los autores recomiendan que los investigadores deben **elegir correctamente en el diseño las pruebas estadísticas antes del análisis de los datos** o bien **consultar a los estadísticos** en cada etapa experimental; mientras que para la comunidad, los autores recomiendan aumentar la **conciencia estadística, reconocimiento y comprensión** de los peligros del uso inadecuado de las técnicas estadísticas, así como la necesidad de crear **directrices** acerca del uso de procedimientos estadísticos.

[Bakker and Wicherts, 2011, Haque and George, 2007]

Los autores pertenecen a la comunidad de la Psicología. En sus trabajos **evaluaron 281 y 50 artículos experimentales** respectivamente, pertenecientes a revistas de alto y bajo impacto en su comunidad como: *British Journal of Psychiatry*⁴, *Journal of Child Psychology*⁵, *Journal of Personality and Social Psychology*⁶, *Journal of Black Psychology*⁷, *Journal of Applied Developmental Psycho-*

²<http://www.sciencedirect.com/science/journal/00029378?sdc=1>

³[http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1365-2648](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1365-2648)

⁴<http://bjp.rcpsych.org/>

⁵[http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1469-7610](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1469-7610)

⁶<http://www.apa.org/pubs/journals/psp/>

⁷<http://journals.sagepub.com/home/jbp>

logy⁸ y *Psychiatric Bulletin*⁹. Los resultados demuestran que las revistas de bajo impacto contienen mayor número de artículos experimentales con problemas estadísticos. No obstante, en el mismo artículo [Bakker and Wicherts, 2011], evidencian la existencia de problemas y errores estadísticos en ambos tipos de revistas (alto y bajo impacto). Los problemas estadísticos son numerosos, tal y como se encuentra detallado en la Tabla 2.4. Los autores recomiendan **consultar a estadísticos, describir los procedimientos estadísticos con suficiente detalle** en los reportes, realizar continuamente **revisiones estadísticas** y, sobre todo, aumentar la **conciencia estadística** en la comunidad.

Tabla 2.4: Defectos estadísticos considerados en los trabajos de [Bakker and Wicherts, 2011, Haque and George, 2007]

Haque y George, 2007	Bakker y Wicherts, 2011
Resultados estadísticos inexactos o incompletos	Uso de métodos estadísticos inadecuados
Valores informados de forma inexacta o incompleta (ej. p-valor, muestras aleatorias)	Problemas en la gestión de sujetos experimentales (ej. aleatorización, tamaño de la muestra, entre otros)
Incoherencias estadísticas (Bonferroni)	Errores en el uso, presentación e interpretaciones estadísticas (ej. Mean, Desviación Estándar, t-test, p-valor, ANOVA)
Aplicación de pruebas estadísticas imposibles	Problemas estadísticos en Pruebas de hipótesis (ej. definición y planeación)

[Brewer, 1985]

El artículo pertenece a la comunidad de la Educación, en el cual se detalla la **evaluación de 18 manuales para la enseñanza estadística** que fueron publicados por editoriales como: *Academic Press*, *Addison-Wesley*, *McGraw-Hill*, *Prentice-Hall*, *John Wiley*, entre otros. En 6 manuales de estadística se encontraron varios errores estadísticos como: **imprecisiones en la teoría de inferencia estadística, interpretaciones erróneas y desviaciones durante la explicación del uso de distintas técnicas estadísticas**, especialmente en técnicas sencillas como: pruebas de hipótesis, nivel de confianza y distribución muestral. El autor propuso una clasificación a los errores estadísticos encontrados en 5 categorías, tal y como se muestra en la Tabla 2.5. El autor recomienda **aumentar las discusiones sobre los problemas estadísticos, mejorar los incentivos para una colaboración estadística entre investigadores** y elaborar directrices estadísticas, entre otras.

⁸<https://www.journals.elsevier.com/journal-of-applied-developmental-psychology/>

⁹<http://pb.rcpsych.org/>

CAPÍTULO 2. ESTADO DE LA CUESTIÓN

Tabla 2.5: Defectos estadísticos considerados en los trabajos de [Brewer, 1985]

"Half-truth" (creencias para extrapolar la verdad algo que falla pasarlo como una condición correcta)	"Cart-before-the-horse" (suposiciones se confunden con cálculos o resultados)
"Definitional errors" (creencias de un incorrecto entendimiento o definición)	"Unitary inference" (Diferentes formas de hacer interferencia o intentar un término de otro)
"Constant-cum-variable" (alterar una variable a una constante, generalmente para hacer una declaración aparentemente más fuerte o más significativa)	

[Kilkenny et al., 2009]

El artículo pertenece a la comunidad de ciencias biológicas. En esta investigación se señala la existencia de problemas o errores estadísticos en los diseños experimentales de **271 artículos publicados entre los años 2003 - 2005**. Los artículos experimentales evaluados fueron seleccionados al azar a partir de las bases de datos indexadas **Medline**¹⁰ y **EMBASE (Elsevier)**¹¹. Los resultados demuestran que más del 60 % de artículos evaluados poseen problemas en la **transparencia de la presentación de informes, solidez del análisis estadístico** y otros problemas que son señalados en la Tabla 2.6. El autor recomienda a los investigadores describir los métodos estadísticos y resultados experimentales de forma exhaustiva, precisa y transparente; en especial los resultados no significativos. Asimismo, el autor da recomendaciones a los editores con el fin de asegurar que los estudios a ser publicados cumplan con criterios rigurosos, así como fomentar la conciencia de la existencia de problemas en el proceso experimental.

Tabla 2.6: Defectos estadísticos considerados en los trabajos de [Kilkenny et al., 2009]

Omitir información esencial (hipótesis, objetivo del estudio, test estadístico, entre otros)	Sesgos en la Selección de sujetos experimentales	Resultados con errores o con variabilidad.
Uso incorrecto de técnicas estadísticas	Falta de información estadística en el reporte.	Diseños experimentales erróneos.

2.4. Estudios pertenecientes a la Ingeniería de Software Experimental

El estudio acerca de errores estadísticos en IS es algo que parece estar en sus inicios. Hemos logrado encontrar muy poca información al respecto, y los po-

¹⁰<https://www.nlm.nih.gov/bsd/pmresources.html>

¹¹<https://www.elsevier.com/promo/rd-solutions/embase>

cos trabajos encontrados son bastantes especializados; por ejemplo, [Dybå et al., 2006] con la potencia estadística, [Miller, 1999] con meta-análisis, etc. Únicamente hemos encontrado un reciente trabajo de [Jørgensen et al., 2016] en el que se estudia en profundidad la presencia e impacto de los errores estadísticos en la investigación experimental. A ello podemos sumar el comentario emitido hace varios años atrás en [Kitchenham et al., 2002], donde uno de sus coautores (David C. Hoaglin, importante estadístico en Medicina) menciona que varios artículos experimentales de IS que fueron revisados y publicados en el *IEEE Transactions on Software Engineering* realizaban un cuestionable uso de la estadística.

La escasa literatura encontrada se detalla a continuación:

[Miller, 1999]

Este trabajo se orienta a evaluar la correcta utilización de la técnica del meta-análisis. El autor utiliza varios casos de estudio como estrategia para diagnosticar defectos metodológicos. Se discute el papel protagónico del meta-análisis como parte del análisis estadístico en la investigación en IS, y el extraño procedimiento usado por parte de algunos investigadores durante la aplicación del meta-análisis en sus experimentos. Los resultados obtenidos, según el autor, son poco alentadores a pesar de su sencillez. Se discute convincentemente que con el meta-análisis se obtienen resultados muy dispares que no conducen a una única conclusión confiable. La razón de esta deficiencia a creencias en varios componentes de los experimentos usados como estudios primarios.

A pesar de estos resultados, el autor finaliza con una serie de recomendaciones para que la **comunidad mejore el control e informes de los trabajos empíricos como una forma de embarcarse en un período de reflexión y mejora** en el uso del meta-análisis con experimentos replicados que utilizan la misma hipótesis.

[Dybå et al., 2006]

Los autores realizan una revisión acerca del poder estadístico en IS. En el estudio se considera una muestra de 5.453 artículos publicados en nueve revistas alto impacto y tres congresos de IS, en la década 1993-2002. Mediante una Revisión Sistemática de Literatura (SLR) se obtienen 103 experimentos controlados que conforman el conjunto de estudios primarios. En total se evaluaron 459 pruebas estadísticas en las que se evidenció que el poder estadístico está sustancialmente por debajo de los estándares comúnmente aceptados. Las evidencias obtenidas invitan a la comunidad a poner más atención en la gestión de las muestras y en la creación de diseños experimentales adecuados para garantizar niveles aceptables de potencia estadística. **Los autores creen en la necesidad de que los investigadores incrementen y mejoren sus conocimientos acerca del poder estadístico.**

[Kitchenham et al., 2003] y [Vegas et al., 2016]

Los autores de ambos artículos evalúan los diseños cross-over utilizados comúnmente en publicaciones experimentales de la comunidad IS. Se evidencia que varios tipos de diseño cross-over no siempre son correctos, por lo que no es posible obtener las ventajas que brindan este tipo de diseños como por ejemplo el aumento del poder estadístico cuando existen abandonos, errores en la asignación de sujetos a tratamientos y carry-over, entre otros.

Los resultados demostraron que algunos diseños cross-over en artículos experimentales de IS, fueron guiados por suposiciones erróneas o cuestionables. Estos problemas fueron en general pasados por alto al momento de realizar replicaciones experimentales. **El uso inadecuado de los diseños cross-over en la comunidad de IS podría haber limitado o cuestionado la validez de ciertos resultados.**

[Jørgensen et al., 2016]

Es el más reciente trabajo referente a problemas estadísticos en IS. El autor considera las advertencias de [Ioannidis, 2005], el cual afirma la existencia de resultados posiblemente falsos en las disciplinas experimentales. La principal contribución del trabajo es la evidencia de **la presencia de sesgos en publicaciones experimentales de IS, así como la existencia de resultados falsos.**

Los autores recalcan que los sesgos pasan casi siempre desapercibidos por los actuales mecanismos de publicación en IS, que no siempre premian a las buenas prácticas estadísticas. Finalizan recomendando que para aumentar la confiabilidad de los experimentos en IS es necesaria dos situaciones: (1) mejorar las prácticas estadísticas y de investigación y (2) fortalecer el poder estadístico en los estudios experimentales.

2.5. Características del estado de la cuestión

Las literaturas anteriormente detalladas: literatura general relacionada al tema de estudio (sección 2.2), literatura de las disciplinas experimentales (sección 2.3), y literatura de la IS (sección 2.4), representan una introducción al fenómeno bajo estudio.

Los trabajos y literatura acerca de los errores y problemas estadísticos en la IS son muy limitados. En cierto sentido, no es extraño el estado literario en la IS, ya que probablemente se encuentre atravesando un estudio similar al que siguieron las disciplinas experimentales asentadas cuando iniciaron este tipo de estudios. Un claro ejemplo es la Medicina, donde su primer experimento formal

se llevó a cabo en la década de los 40 [Bhatt, 2010]. Sin embargo, las primeras publicaciones acerca de errores estadísticos aparecieron aproximadamente en la década de los 70 [Gore et al., 1977]. Parece claro que los primeros trabajos de investigación relacionados con errores estadísticos, aparecen tardíamente en cada ciencia experimental. Dado que la IS es una disciplina relativamente joven, la ausencia de tales publicaciones podría no ser una sorpresa.

Ahora bien, la ausencia de información produce incertidumbre acerca de lo que puede estar sucediendo en comunidad de IS acerca del fenómeno bajo estudio, y asimismo, nos genera dudas acerca de si en la comunidad de IS *existe la suficiente conciencia* en lo referente a la presencia y el impacto de los errores estadísticos en su proceso experimental.

Es evidente que la cantidad de errores estadísticos puede ser muy elevada, y que apenas en nuestro estado de la cuestión hemos visto la punta del iceberg. El análisis de las características del estado de la cuestión (ver Figura 2.1) que se detalla a continuación, nos permitirá sumarizar las características del fenómeno bajo estudio.

2.5.1. Evidencias de la existencia de errores estadísticos

La existencia de errores estadísticos en trabajos experimentales es un hecho bien asentado en las ciencias experimentales. La cantidad de evaluaciones que realizan a sus publicaciones experimentales en revistas y congresos en alto y bajo impacto es notable. Las evaluaciones no son recientes en algunos casos; se extienden a lo largo de períodos de una o dos décadas (p. ej., Medicina y Psiquiatría). Todo ello implica que las disciplinas experimentales no dudan ni escatiman esfuerzos en realizar evaluaciones continuas, con el fin de llegar a obtener una mayor rigurosidad y veracidad en sus resultados experimentales.

2.5.2. Tipos de errores estadísticos

En la literatura general

En la literatura general, se ha logrado evidenciar dos situaciones: por un lado, la gran cantidad de errores estadísticos que se cometen en los experimentos de cada comunidad, por citar algunos ejemplos:

- **Interpretaciones erróneas o desviaciones en la explicación de técnicas estadísticas** [Bakker and Wicherts, 2011, Brewer, 1985],
- **Aplicación de pruebas estadísticas imposibles o métodos y técnicas estadísticas inadecuados** [Olsen, 2003, Ercan et al., 2008, Welch and

Gabbe, 1996],

- **Resultados estadísticos inexactos o incompletos** [Haque and George, 2007, Welch and Gabbe, 1996, Olsen, 2003, Ercan et al., 2008],
- **Problemas y sesgos en la gestión de sujetos experimentales** [Bakker and Wicherts, 2011, Kilkenny et al., 2009],
- **Presentación e interpretaciones estadísticas** [Bakker and Wicherts, 2011, Kilkenny et al., 2009],
- **Falta u omisión de información estadística esencial en el reporte, descripciones estadísticas inadecuadas e incompletas** [Olsen, 2003, Ercan et al., 2008, Welch and Gabbe, 1996, Kilkenny et al., 2009],
- **Suposiciones incorrectas con técnicas estadísticas** [Brewer, 1985, Kilkenny et al., 2009],
- **Diseños experimentales erróneos** [Brewer, 1985],
- **Alteraciones de variables** [Brewer, 1985],
- **Problemas estadísticos en pruebas de hipótesis** [Bakker and Wicherts, 2011], entre otros;

Y por otro lado, los errores estadísticos aparecen en todas las fases del proceso experimental, por citar algunos ejemplos [Huck, 2009, Good and Hardin, 2012, Vickers, 2010]:

- **Diseño** (p. ej., pruebas de hipótesis, otros),
- **Ejecución** (p. ej., Tamaño de la muestra, selección de sujetos,
- **Análisis** (p. ej., multi-way ANOVA, 2-way ANOVA, regresión, t-test, otros) y
- **Reporteo** (p. ej., estadística descriptiva, gráficos y tablas, otros).

En la literatura específica de IS

Para el caso de la IS, la literatura es muy limitada. Los pocos estudios encontrados, por lo general, son bastante especializados. Recordemos que la IS es aún un campo científico muy joven, y por ello quizás requiera algún tiempo y esfuerzo para tratar y madurar este tipo de temática. Sin embargo, las evidencias encontradas en la disciplinas experimentales frente a las evidencias de IS no pueden ser más reveladoras; por ejemplo, mientras en la Medicina, una ciencia experimental establecida y paradigmática, se identifican múltiples errores estadísticos [Welch and Gabbe, 1996, Olsen, 2003, Ercan et al., 2008], en IS aparentemente este aspecto parece ser ignorado o tal vez es aún desconocido.

2.5.3. Alternativas de solución

En la mayoría o casi todos los trabajos revisados, los autores (a modo de discusión) suelen describir recomendaciones o alternativas de solución para sus comunidades, algunas tan sencillas como:

- **Elegir las pruebas estadísticas antes del análisis de los datos** [Olsen, 2003, Welch and Gabbe, 1996],
- **Consultar a los estadísticos** en cada etapa experimental [Olsen, 2003, Welch and Gabbe, 1996, Ercan et al., 2008],
- **Describir los procedimientos estadísticos con suficiente detalle** [Haque and George, 2007],
- **Generar colaboración estadística** para la elaboración de directrices [Brewer, 1985], entre otros
- Recomendaciones mucho más tajantes, enérgicas y algunas preocupantes como: aumentar la **conciencia estadística** en la comunidad [Olsen, 2003, Welch and Gabbe, 1996, Ercan et al., 2008, Kilkenny et al., 2009, Bakker and Wicherts, 2011],
- **Reconocer y comprender** los peligros del uso inadecuado de la estadística.
- Necesidad de crear **directrices acerca de procedimientos estadísticos** [Olsen, 2003, Welch and Gabbe, 1996, Ercan et al., 2008, Bakker and Wicherts, 2011],
- Realizar **revisiones estadísticas sólidas** [Bakker and Wicherts, 2011],
- **Aumentar las discusiones sobre las advertencias estadísticas** [Brewer, 1985],
- **Embarcarse en período de reflexión y mejora** [Miller, 1999], entre otros.

2.5.4. Características principales identificadas

La revisión de literatura, nos ha permitido identificar cuatro características fundamentales acerca de la situación actual, tal y como se indica en la Figura 2.1:

A breves rasgos, una característica que es evidente y se refiere a *estudios formales acerca de la existencia de errores estadísticos* en las distintas ciencias, por ejemplo: Medicina, Psicología, Educación, Biología, etc. Una segunda característica se refiere a *los errores estadísticos que abarcan todas las fases del proceso experimental*, desde la fase de Diseño hasta la fase de Reporte. La tercera característica se refiere a que los *errores estadísticos se encuentran relacionados en su*

mayoría a técnicas estadísticas sencillas como por ejemplo: p-valor, pruebas de hipótesis, tamaño de la muestra, etc. y, finalmente, la cuarta característica se refiere a *las alternativas de solución*, que son planteadas en cada comunidad. Apparentemente, estas alternativas de solución son muy genéricas, como por ejemplo: crear directrices estadísticas, investigación colaborativa, fomentar una conciencia estadística en los investigadores, etc.

Figura 2.1: Características del Estado de la Cuestión

3 | PLANTEAMIENTO DEL PROBLEMA Y OBJETIVOS

Resumen: El estado de la cuestión nos ha permitido desvelar que existen errores estadísticos en todas las ciencias, los cuales, afectan a la totalidad de las fases del proceso experimental. En Ingeniería de Software (IS), la situación está considerablemente menos clara, ya que la discusión acerca de errores estadísticos hace referencia a varios aspectos bastante específicos. En general, la situación dista de ser satisfactoria, ya que la causas de los errores no están bien diagnosticadas, lo que limita el planteamiento de alternativas de solución. A dicha situación se une el aparente desinterés de la comunidad. En este Capítulo, planteamos cuatro problemas principales respecto al fenómeno de estudio a partir de las características del estado de la cuestión del capítulo anterior (ver sección 3.1). Determinados los problemas, finalizaremos el capítulo estableciendo y explicando los objetivos y preguntas de investigación de la presente tesis doctoral (ver sección 3.2)

3.1. Planteamiento del problema

Las características del estado de la cuestión, identificadas en el Capítulo 2, apuntan a la existencia de varios problemas o incertidumbres, que afectan a las ciencias en general y a la IS en particular. Nos planteamos el problema de si la experimentación en IS puede estar libre de errores estadísticos, tal y como la escasa literatura crítica lo sugiere, o si esta literatura sólo ha vislumbrado la punta del iceberg. Para lograr resolver este problema, lo hemos desagregado en problemas más específicos con la finalidad de poderlos afrontar más fácilmente, tal y como se indica en la Figura 3.1.

Hemos podido comprobar que en las ciencias experimentales se han realizado estudios acerca de la presencia de errores estadísticos (C1), los cuales abarcan la totalidad de las fases experimentales (C2). Sin embargo, todos los estudios realizados son parciales, típicamente restringidos a una o pocas revistas dentro de una disciplina específica. Se aprecia:

P1: La ausencia de un catálogo bien definido de errores estadísticos, ya sea de tipo general o multidisciplinar o ya sea de tipo particular

Sin dicho catálogo, es complejo plantear acción alguna de diagnóstico o mejora ya que:

- Es más difícil identificar qué errores afectan a un área o conjunto de artículos partiendo de cero (esto es construyendo una clasificación ad-hoc), que basándose en guías previas.
- Por la misma razón, es necesario determinar la prevalencia de los errores estadísticos que producen resultados más inciertos.

La existencia de errores estadísticos en las distintas ciencias (C1) es un hecho bien establecido. De forma bastante coherente, se han propuesto diversas alternativas de solución (C4). Sin embargo, estas alternativas son muy genéricas. Hacen referencias a aspectos tales como: aumentar la conciencia estadística [Bakker and Wicherts, 2011, Haque and George, 2007, Welch and Gabbe, 1996, Olsen, 2003, Ercan et al., 2008], o a generar colaboración estadística entre investigadores [Brewer, 1985], entre otras.

Sin embargo,

P2: Existe una falta de Información específica acerca de las causas de los errores estadísticos

Por lo cual, a su vez:

P3: Es difícil de plantear soluciones o recomendaciones de mejora realmente efectivas

En IS, se ha publicado un número bastante reducido de trabajos relacionados con los errores estadísticos. Todos estos trabajos abordan errores estadísticos relacionados a técnicas estadísticas sencillas (C3).

Parece que:

P4: Existe una aparente ignorancia o desinterés en IS acerca de los errores estadísticos

En resumen y en nuestra opinión, consideramos que existe una reducida probabilidad de que la IS esté libre de errores estadísticos. De hecho, los errores estadísticos en IS podrían ser incluso peores que en otras ciencias.

Figura 3.1: Relación entre las Características del Estado de la Cuestión y la Problemática de Investigación.

3.2. Objetivos de investigación

3.2.1. Objetivo general

De acuerdo a los problemas identificados (P1-P4), nos hemos planteado como objetivo de esta investigación el **explorar, identificar, analizar y evaluar los errores estadísticos en IS, con el fin de establecer sus causas y proponer mecanismos para evitarlos o, al menos, minimizar su impacto**. De esta forma pretendemos que la comunidad de investigadores de IS dispongan, como mínimo, de información referente al rigor estadístico en estudios de IS, para que en su caso se inicien discusiones acerca de esta problemática dentro de nuestra comunidad.

Ahora bien, el objetivo general es muy amplio y complicado de afrontar metodológicamente. Por ello, para poder facilitar una mejor selección de la metodología de investigación (ver Capítulo 4), fue necesario descomponer nuestro objetivo general en objetivos más específicos y, a su vez, en preguntas de investigación concretas.

3.2.2. Objetivos específicos

Es clara la falta de información acerca de errores estadísticos y sus causas más probables (P1 y P2). Por ello, se ve razonable realizar una identificación de los errores estadísticos en las disciplinas experimentales. Ello da lugar a nuestro primer objetivo específico:

O1: *Identificar si existen errores estadísticos en IS, y establecer su prevalencia.*

El objetivo específico planteado nos empuja a recurrir como fuente primaria de información a las distintas disciplinas experimentales, ya la literatura existente en IS es muy escasa. De las diversas ciencias obtendremos información acerca de frecuencia de ocurrencia de una amplia y variada tipología de errores estadísticos. Sin embargo, la identificación de los errores estadísticos no sería suficiente si no determinamos (al menos de forma preliminar) también hasta qué punto dichos errores estadísticos podrían también ocurrir en la IS. Por tal motivo, las preguntas de investigación (RQs) que se plantean para este objetivo son:

- **RQ1.1:** ¿Qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales?
- **RQ1.2:** ¿Hasta qué punto estos errores estadísticos ocurren también en la IS?

La respuesta a las preguntas de investigación RQ1.1 y RQ1.2 suponen únicamente un punto de partida. En caso de que se confirmara la existencia de errores en IS, necesitamos identificarlos con mayor precisión, ya que de lo contrario nuestro trabajo tendría poco impacto en la práctica; nótese que el IS los investigadores o bien ignoran la existencia de errores, o bien no los consideran importantes (P3). Por lo tanto, nuestro segundo objetivo específico es:

O2: *Determinar la naturaleza de los errores estadísticos cometidos en IS.*

Para determinar la naturaleza de los errores estadísticos cometidos en IS, es necesario lograr dos hitos: por un lado, estudiar en detalle uno o varios errores estadísticos comunes, para poder identificar cuándo y cómo ocurren. En segundo lugar, deberemos enfrentar a los investigadores de IS con dichos errores, para averiguar: a) si reconocen los errores y b) cuál puede ser la razón que justifique la comisión de dichos errores. Por ello, las RQs que se plantean para este objetivo son:

- **RQ2.1:** ¿Qué problemas se observan en la aplicación de las técnicas estadísticas en IS?
- **RQ2.2:** ¿Es la comunidad de IS consciente de la existencia del problema?

Una vez hallamos clarificado la naturaleza de los errores estadísticos en IS, tendremos la oportunidad de conocer a ciencia cierta las causas más realistas de estos problemas y, con ello, sugerir recomendaciones o medidas paliativas a la comunidad IS. Por lo tanto, nuestro tercer objetivo específico es:

O3: *Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.*

Para establecer las causas más realistas y por ende obtener recomendaciones adecuadas y acordes a la IS, es necesario 2 realizar dos actividades: Por un lado, lograr identificar las causas de la aparición de errores estadísticos en las distintas disciplinas experimentales, lo que nos proporcionará un punto de partida para entender la situación en IS. Por otro lado, deberemos recurrir de nuevo a los propios investigadores de IS con el fin de establecer y priorizar las causas que pudieran dar lugar a la aparición de errores estadísticos. Nótese que identificar únicamente las causas no es suficiente para que la IS pueda evitar la comisión de errores estadísticos; por ello, es necesario obtener recomendaciones o medidas paliativas para evitar la comisión de errores estadísticos en la comunidad de la IS. Por todo ello, las RQs que se plantean para este objetivo son:

- **RQ3.1:** ¿Cuáles son las causas de los errores estadísticos en las disciplinas experimentales?
- **RQ3.2:** ¿Cuales son las causas más probables de los errores estadísticos en IS?
- **RQ3.3:** ¿Qué medidas se pueden tomar para evitar la comisión de errores?

Con la consecución de los objetivos específicos planteados anteriormente, lograremos conseguir al menos que la comunidad de IS pueda adentrarse en tres problemas metodológicos importantes: (1) reconocer la existencia de errores estadísticos dentro de su proceso de experimentación, (2) lograr conciencia acerca de este fenómeno entre los investigadores de la IS, e (3) iniciar discusiones o estudios similares a esta tesis doctoral como una forma de afrontar y minimizar este problema, que no es ajeno a ninguna de las ciencias experimentales [Ioannidis, 2005].

Figura 3.2: Contribución de los Objetivos y Preguntas de Investigación con la Resolución de la Problemática Identificada.

4 | MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Resumen: La definición de los objetivos y preguntas de investigación, nos permite tener una mayor claridad acerca de las metas o hitos que se pretenden alcanzar con la presente tesis doctoral. Sin embargo, alcanzarlos suponen varias dificultades y desafíos que deben ser enfrentados. Para ello, se definió como metodología de investigación a un conjunto de métodos que encajan dentro de los métodos mixtos. Al respecto, en este capítulo revisaremos brevemente los diferentes diseños y métodos de investigación que son comúnmente usados en IS (sección 4.1). También se explica el procedimiento de la selección del método de investigación indicado, realizando con anterioridad una identificación, análisis y evaluación de cada uno de los desafíos que podríamos encontrar con nuestros objetivos y preguntas de investigación frente al método seleccionado (sección 4.2). Finalizaremos el capítulo proponiendo varias etapas enmarcadas dentro de nuestro plan de trabajo que se compaginan plenamente con el método de investigación seleccionado para la tesis doctoral (sección 4.3)

4.1. Diseños de investigación usados comúnmente en la investigación IS

Según [Creswell, 2002], los diseños de investigación son planes y procedimientos de investigación que abarcan varios métodos de recolección (p. ej., experimentos, encuestas, otros.) y análisis de datos, tal y como se muestra en la Tabla 4.1. Para un mejor detalle y descripción explicamos los tres tipos de diseños de investigación que propone el autor en las siguientes secciones.

Tabla 4.1: Diseños de Investigación Usados Comúnmente en la Investigación IS

Diseños de Investigación			
	Cuantitativa	Cualitativa	Métodos Mixtos
Métodos	Experimentos Encuestas otros.	Entrevistas Etnografía Casos de estudio Investigación Narrativa	Secuencial Concurrente

4.1.1. Diseño de investigación cuantitativa

Es un diseño comúnmente utilizado para probar teorías (presuntivas, planteamientos o inducciones) que examinan la comparativa entre variables cuantitativas con el fin de lograr identificar (o cuantificar) las relaciones existentes por medio de procedimientos estadísticos [Creswell, 2002]. Tradicionalmente, los métodos que suelen utilizarse dentro de este diseño de investigación son: experimentos y no-experimentos (p. ej., encuestas, otros) [Easterbrook et al., 2008].

Experimentos

Un experimento (en IS) se caracteriza por tener al menos una hipótesis cuantificable, en donde una o más variables independientes (o tratamientos) se utilizan para medir el efecto sobre una o más variables dependientes (o variables respuesta) a partir de la muestra de una población aleatoria [Wohlin et al., 2012, Juristo and Moreno, 2001]. Las características más representativas de un experimento IS son: (1) se utiliza de forma controlada a ciertas variables de la hipótesis para determinar su relación con otras variables, (2) se requiere directa o indirectamente sujetos experimentales (o a veces no se los requiere), (3) se utiliza una muestra

representativa a partir de una población (bien definida) que es seleccionada aleatoriamente, (4) los resultados son reproducibles o replicables y (5) es necesario mucho esfuerzo y recursos para su ejecución [Easterbrook et al., 2008].

Encuestas

Una encuesta provee una descripción numérica de las tendencias, actitudes u opiniones de una muestra representativa de una población [Creswell, 2002]. Es muy utilizado en estudios longitudinales (especialmente en ciencias sociales) y secciones cruzadas. Las características más representativas de una encuesta en IS son: (1) permite una descripción numérica de tendencias, actitudes u opiniones, (2) los resultados permiten generalizar desde una muestra representativa hacia su población en particular y (3) utiliza cuestionarios con diferentes tipos de preguntas (abiertas, cerradas, mixtas) que deben ser validadas para permitir rigurosidad en los resultados.

4.1.2. Diseño de investigación cualitativa

Es un diseño utilizado en investigaciones de tipo exploratorio [Wohlin et al., 2012], mediante el cual se analizan a diferentes objetos, sujetos o grupos en su entorno natural, social o humano para obtener resultados en base a la observación [Wohlin et al., 2012, Creswell, 2002]. No se suele probar teorías o hipótesis; más bien, sus métodos permiten generar teorías e hipótesis. Los métodos utilizados de este tipo de diseño han llegado a ser muy utilizados en las últimas décadas (especialmente en ciencias sociales) [Creswell, 2002] debido a que permiten realizar una exploración y descubrimiento con mayor profundidad respecto a la información que provienen de pensamientos y experiencias de la gente. Es posible que la información obtenida con este diseño proporcione una interpretación y comprensión del porqué de las cosas y cómo estas funcionan. En IS, los métodos cualitativos más susceptibles a ser utilizados son: entrevistas, etnografías y casos de estudio, entre otros [Easterbrook et al., 2008].

Entrevistas

La recopilación de datos a través de entrevistas es una de las fuentes de recolección de datos más utilizadas y más importantes en los estudios de IS [Runeson et al., 2012]. Su importancia radica en dos aspectos: por un lado, que es un método que permite la recolección de datos primarios o permite la validación de datos previos; y por otro lado, que este método permite obtener conocimiento de valioso interés cuando no está disponible la información en literaturas o no existe más que en la mente de la población que se investiga. Las características más

importantes de estos métodos son [Runeson et al., 2012]: (1) las preguntas se encuentran íntimamente relacionadas a los objetivos de investigación por lo que es posible profundizar durante una sesión, (2) se puede o no utilizar una guía para la entrevista (no estructurada, semi-estructurada, estructurada), (3) la profundidad en las sesiones depende del conocimiento que haya obtenido previamente el entrevistador, (4) permite al investigador estar en el campo para observar de cerca el fenómeno, (5) se aplica en una población representativa (bien definida) y (6) permite obtener conocimiento valioso a partir de pensamientos y experiencias de una población, entre otros.

Etnografía

La etnografía es un método de investigación cualitativa, en el cual el investigador estudia a un grupo cultural en su entorno natural (sin modificarlo) con el fin de recopilar datos observacionales durante un prolongado periodo tiempo [Creswell, 2002]. Las características más representativas de una etnografía en IS son: (1) permite conocer cómo se crea una cultura técnica de prácticas y estrategias de comunicación de trabajo colaborativo en una comunidad, (2) se enfoca en cómo los miembros de la comunidad (por sí mismos) le dan sentido a su entorno determinado por sus propias prácticas, (3) se aplica a cualquier tamaño de comunidad (grandes o simples), (4) permite observaciones detalladas, evitando prejuicios y (5) permite observar, aprender y crear pre-concepciones propias [Creswell, 2002].

Caso de estudio

Los casos de estudio permiten explorar en profundidad un fenómeno real dentro de su contexto, especialmente cuando los límites entre el fenómeno y el contexto no son claramente evidentes [Yin, 2013]. Este tipo de método está limitado por el tiempo y la actividad, razón por la cual se recopila información utilizando una variedad de procedimientos de recopilación de datos como: entrevistas (estructuradas, semi-estructuradas y no-estructuradas), observaciones, datos de documentos, entre otros [Runeson et al., 2012]. Las características más representativas de un caso de estudio en IS son: (1) permite entender un fenómeno generalmente a nivel de una unidad de análisis bien definida, (2) suele complementar con otros métodos de investigación, (3) ofrece entendimiento de cómo y porqué sucede un fenómeno, aunque se limita a relacionar causas, (4) ayuda a probar una teoría bien formulada, (5) permite generalizar una teoría a través de la replicación en más casos, y (6) requiere esfuerzo medio para su realización [Creswell, 2002].

4.1.3. Diseño de investigación de métodos mixtos

Para [Creswell, 2002], una investigación basada en métodos mixtos es una aproximación que combina (o asocia) a varios métodos de investigación con el fin de analizar complementariamente a un fenómeno de estudio. Este tipo de diseño de investigación ha sido relevante en las ciencias sociales y humanas [Creswell, 2002]. Su empleo (en la mayoría de casos) se centran en la incorporación de diseños de investigación (y sus métodos) tanto cualitativos como cuantitativos, los mismos que unidos permiten ampliar o comprender mejor los resultados entre uno u otro enfoque. La principal consideración de los métodos mixtos es que depende en gran medida de varios aspectos, por ejemplo, de la sincronización de los métodos, la ponderación del diseño de la investigación, la manera de mezclar los resultados entre métodos y la teorización que se haya realizado previamente. Este tipo de diseño tiene dos tipos de métodos mixtos que se detallan a continuación:

Método mixto secuencial

El método mixto secuencial se utiliza cuando existe la intención de explorar un tema en específico como parte de la investigación y posteriormente ampliarlo obteniendo conocimiento mediante la recopilación de datos a partir de otras fuentes de información. En consecuencia, los resultados cuantitativos y cualitativos son recopilados y enlazados uno detrás de otro durante el análisis de la investigación [Creswell, 2002]. Un ejemplo es la realización de una entrevista (cualitativo) con fines exploratorios. Posteriormente, los resultados obtenidos son validados o ampliados con una encuesta (cuantitativo), aplicado a la misma u otra población, dependiendo de lo que el investigador desee generalizar; o alternatively, iniciar la investigación con una encuesta (cuantitativo) acerca de una teoría o concepto probado y posteriormente aplicar sobre la misma u otra población, una entrevista (cualitativo) para profundizar la exploración.

Método mixto concurrente (simultáneo)

El diseño de método mixto concurrente permite recopilar e integrar simultáneamente tanto datos cuantitativos como datos cualitativos. En este diseño, el investigador recopila ambas formas de datos, al mismo tiempo, y luego integra la información en un análisis e interpretación global de los resultados [Creswell, 2002]. Un ejemplo común es realizar una encuesta (cuantitativo) y entrevista (cualitativo) simultáneamente para posteriormente integrar (o fusionar) sus resultados en números (cuantitativos).

4.2. Proceso de selección del método de investigación

La decisión de seleccionar uno u otro diseño de investigación (incluyendo sus métodos internos específicos), depende casi siempre de tener una clara visión de la naturaleza del problema y de lo que se desea alcanzar con la investigación. La selección del método es guiada por la afinidad que tiene el método de investigación con los objetivos y preguntas de investigación, así como también las dificultades que podrían ser afrontadas. Un aspecto adicional que necesita ser considerado para la selección del método. Son los recursos necesarios para su ejecución (p. ej., tiempo, logística, etc.).

El proceso de selección del método de investigación de la presente tesis doctoral, esta basado en dos actividades: (1) identificación de los desafíos planteados por los objetivos de investigación y (2) la evaluación de los métodos de investigación aplicables. A continuación se detalla cada una de estas actividades:

4.2.1. Identificación de los desafíos planteados por los objetivos de investigación

O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia

Este objetivo supone alcanzar dos hitos referente a tipologías de errores estadísticos y su presencia en IS (RQ1.1 - RQ1.2), tal y como se indica en la Tabla 4.2. Alcanzar los hitos propuestos supone una serie de dificultades, de las cuales identificaremos las que comúnmente son encontradas durante la teorización del fenómeno de estudio (D1.1 - D1.4). Estas son: la existencia de una *amplia literatura general* (D1.1) acerca de errores estadísticos en las disciplinas experimentales, a excepción de la IS que es bastante *limitada* (D1.2). Asimismo, la revisión de una abundante literatura, conlleva a cierta complejidad y tiempo en el *manejo y clasificación* de la información acerca de errores estadísticos (D1.3). Finalmente, no podemos olvidar la dificultad de lograr establecer adecuadamente la existencia de errores estadísticos en la IS (D1.4).

O2: Determinar la naturaleza de los errores estadísticos cometidos en IS.

Este objetivo supone alcanzar hitos referentes a *la existencia de conciencia de errores estadísticos en IS* y el grado de conciencia de los investigadores acerca de la existencia de los mismos. Los hitos están asociados a RQ2.1 y RQ2.2, tal y

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.2: Desafíos para O1 y sus preguntas de investigación

		Preguntas de Investigación	
		RQ1.1	RQ1.2
Desafíos	D1.1: Amplia literatura general, incluyendo libros y artículos.	X	X
	D1.2: Limitada cantidad de estudios con este tipo de información en IS.	X	X
	D1.3: Revisión, manejo y clasificación complicada por la gran cantidad de información proveniente de la amplia literatura general.	X	X
	D1.4: Es difícil determinar la existencia preliminar de diferentes tipos de errores estadísticos en IS.		X
		Tipología preliminar de errores estadísticos en las disciplinas experimentales.	Comisión de evidencias preliminares de la existencia de errores estadísticos en IS
Hitos			

como se indica en la Tabla 4.3. Ambos hitos presentan una serie de dificultades (D2.1 - D2.8) que se orientan principalmente a la *limitada cantidad de estudios referente a errores estadísticos en IS* (D2.2), lo que hace prever que no será fácil establecer la *información referente a evidencias y percepción* acerca del fenómeno de estudio (D2.8). Por esta razón, se dificulta *determinar con seguridad la existencia de varios tipos de errores estadísticos*, al menos inicialmente (D2.1) como para poder establecer *estimaciones* o al menos indagaciones preliminares (D2.5).

Para el caso particular del hito de *obtener el grado de conciencia de los investigadores en IS*, existen diferentes dificultades, inclusive más complicadas que el hito anterior. Las dificultades van desde la limitada información acerca de *opiniones, creencias o actitudes sobre errores estadísticos que en IS* (D2.3) hasta una *escasa conciencia o desinterés acerca de la existencia de errores estadísticos* (D2.4). Todo ello se podría complicar aún más si no se *garantiza la privacidad* del experimentador como para que pueda ser sincero durante la contestación de preguntas acerca del tema de estudio (D2.7).

Al respecto, es posible que para cada hito, sea necesario preguntar a una considerable población de experimentadores, que no siempre se encuentran cerca geográficamente, por lo desde supone la *dificultad de concentrar a una gran cantidad de experimentadores en un solo lugar* (D2.6).

O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia

Este objetivo supone alcanzar tres hitos que se orientan a la evaluación de las causas y obtención de recomendaciones necesarias para evitar o al menos minimizar el impacto de los errores estadísticos en IS. Los hitos se encuentran

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.3: Desafíos para O2 y sus preguntas de investigación

		Preguntas de Investigación	
		RQ2.1	RQ2.2
Desafíos	D2.1: Es difícil determinar con seguridad la existencia de muchos tipos de errores estadísticos.	X	X
	D2.2: Limitada cantidad de estudios con información relacionada (descriptiva y analítica) o la información de errores estadísticos no se encuentra estandarizada en IS	X	X
	D2.3: Son limitadas las opiniones, creencias o actitudes formales en la comunidad acerca de errores estadísticos.		X
	D2.4: Escasa conciencia o desinterés acerca de la existencia de errores estadísticos.		X
	D2.5: Dificultad de realizar estimaciones iniciales acerca del fenómeno de estudio.	X	X
	D2.6: Dificultad en concentrar una gran cantidad de experimentadores en un instante.		X
	D2.7: Garantizar la privacidad para un mejor resultado y una mayor colaboración.		X
	D2.8: Obtener información acerca de evidencias y percepciones.	X	X
		Evidencia formal de la existencia de errores estadísticos con alguna técnica estadística seleccionada.	Percepción de los investigadores acerca de la problemática de estudio en la IS
Hitos			

asociados a RQ3.1, RQ3.2 y RQ3.3, tal y como se indica en la Tabla 4.4. Los desafíos aparentemente son similares a los hitos de los anteriores objetivos, que se resumen en la *limitada cantidad de estudios con información de las causas en IS* (D3.1 y D3.2), *limitada información puntual del fenómeno, lo que complica la clara identificación de causas realistas y sus recomendaciones* (D3.3); lo que supone como única forma de obtener información razonablemente válida sea mediante *la obtención de comportamientos y experiencias que se encuentran en la cabeza del investigador de IS* (D3.4). A ello, se suma la *limitación de encontrar investigadores que conozcan o al menos manejen adecuadamente el fenómeno de estudio o que posean experiencia en este tipo de fenómenos* (D3.1). Aquellos investigadores que logremos encontrar, muchos de ellos pueden encontrarse distantes geográficamente, haciendo *difícil la interacción* con ellos (D3.6).

4.2.2. Evaluación de los métodos de investigación aplicables

Las particularidades y características de los métodos de investigación indicados en la sección 4.1, sumado a la identificación de desafíos y dificultades de nuestras preguntas de investigación en la sección 4.2.1, nos permitió evaluar los métodos de investigación y herramientas de apoyo que podrían ser adecuados para nuestra investigación en la presente tesis doctoral [Genero et al., 2014].

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.4: Desafíos para O3 y sus preguntas de investigación

		Preguntas de Investigación		
		RQ3.1	RQ3.2	RQ3.3
Desafíos	D3.1: Pocos investigadores en IS con experiencia en este tipo de fenómenos.	X	X	X
	D3.2: Limitada cantidad de estudios con este tipo de información en IS (concepto vs la práctica).	X	X	X
	D3.3: Se desconocen las causas principales y recomendaciones realistas para la mitigación de errores estadísticos en IS.	X	X	X
	D3.4: La única información razonablemente válida se encuentra en la cabeza del investigador de IS.		X	X
	D3.5: Información puntual del fenómeno no existe, por tal no hay una identificación clara de las causas y sus recomendaciones del fenómeno.		X	X
	D3.6: Es difícil establecer la interacción con investigadores expertos en errores estadísticos		X	X
		Causas en disciplinas	Causas en IS	Recomendaciones en IS
		Hitos		

Vemos como una opción metodológica el método “*Quick and dirty*” [Wheeler, 2014] con el apoyo del *análisis temático* [Cruzes and Dybå, 2011] como estrategia de investigación para alcanzar los hitos de RQ1.1, tal y como se muestra en la Tabla 4.5.

Consideramos que este método se justifica dado que es necesario obtener *resultados preliminares* para dar una mejor orientación a la tesis doctoral, en otras palabras, con la información obtenida previamente se podrá brindar una *introducción acerca del fenómeno de estudio* basado en diferentes *bibliografías específicas* (p. ej., libros) acerca de errores estadísticos, dada la extensa literatura en las disciplinas (D1.1) y la limitada información en IS (D1.2 y D1.4). Asimismo, el manejo y clasificación de la información obtenida (D1.3), supone tiempo y esfuerzo de varios investigadores. Por ello que el método seleccionado más el apoyo de un análisis temático [Cruzes and Dybå, 2011] supone una mejor *organización, clasificación y codificación* de la información obtenida (D1.1, D1.2, D1.3), a la vez que permite obtener la suficiente rigurosidad en la validez de los resultados. También considero utilizar cuestionarios *como instrumento de revisión y recolección de información* que reflejen la existencia de tipos de errores estadísticos en publicaciones de la IS (D1.4).

Los hitos de RQ2.1 y RQ2.1 no parecen fáciles de afrontar dado que requieren de la interacción con investigadores. Por ello que consideramos como una buena opción la aplicación de una *encuesta* (cuantitativo) con el apoyo un análisis temático, tal y como se detalla en la Tabla 4.6. Suponemos que en la encuesta se podrían formular preguntas de diferente tipo (p. ej., abiertas, cerradas) acerca

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.5: Evaluación de métodos y herramienta aplicables para RQ 1.1 y RQ 1.2

		RQ1.1 y RQ 1.2			
		Hito 1: Tipología preliminar de errores estadísticos en las disciplinas experimentales.			
		Hito 2: Comisión de evidencias preliminares de la existencia de errores estadísticos en IS			
		Dificultades			
		D1.1	D1.2	D1.3	D1.4
Método/Estrategia	Quick and dirty	Permite obtener Información con rapidez (al menos preliminar).	x	x	x
		Sus resultados proporcionan una introducción a un tema específico.	x	x	x
		No implica realizar un extenso plan de búsqueda.	x	x	x
		Utiliza un número limitado de puntos de entrada de búsqueda (ej. Libros, revistas).	x	x	x
		Los resultados obtenidos permite dar una visión de lo que se busca.	x	x	x
Herramientas de Apoyo	Análisis temático	Permite análisis en profundidad de un tema específico.	x	x	x
		Permite la recopilación de cualquier fuente de información.	x	x	x
		Permite ordenar, codificar y clasificar la información obtenida.	x	x	x
		Permite la formulación de temas a cualquier nivel de abstracción.	x	x	x
		Con los resultados se puede realizar gráficas descriptivas de la información obtenida.	x	x	x
	Cuestionario de revisión	Permite obtener de manera sistemática y ordenada información necesaria y relevante relacionada al tema de estudio.	x	x	x
		Permite formular preguntas para poder indagar un tema en particular o variable de evaluación.	x	x	x
		Su uso maximiza la validez de la información obtenida en la evaluación.	x	x	x
		Permite categorizar las preguntas (ej. dicotómicas) como para evaluar la existencia o no de alguna evidencia.	x	x	x

del fenómeno de estudio, Una encuesta es útil cuando la información es aparentemente inexistente en la comunidad (D2.2), y que el único lugar donde podemos encontrar conocimiento acerca del fenómeno se encuentra en la cabeza de cada investigador dadas por sus creencias, opiniones o actitudes (D2.3 y D2.4). La información obtenida mediante la encuesta se podrá *organizar, clasificar y codificar* con un análisis temático [Cruzes and Dybå, 2011] para minimizar las dificultades de su manejo (D2.1, D2.2, D2.5, D2.8).

Finalmente, para alcanzar los hitos RQ3.1, RQ3.2 y RQ3.3 se ha considerado adecuado utilizar 2 métodos de investigación. Por un lado, para RQ3.1 se ve factible realizar una Revisión Sistemática de Literatura (SLR) [Kitchenham et al., 2010, Brereton et al., 2007] con el fin de obtener información referida a las causas de los problemas estadísticos que operan en las disciplinas experimentales. Se ve necesario el apoyo de un análisis temático [Cruzes and Dybå, 2011], para la organización y clasificación de la información, tal y como se muestra en la Tabla 4.7. El SLR se justifica dado a que sus procedimientos de búsqueda, extracción y síntesis son adecuados al tema de estudio, dada la existencia de una amplia literatura (actual y pasada) en las disciplinas experimentales (D1.1 y D1.3), mientras el análisis temático [Cruzes and Dybå, 2011] se justifica por las ventajas que proporciona para la *organización, clasificación y codificación* de la información referente a las causas a errores estadísticos en los estudios primarios (D1.1, D1.2, D1.3).

Para el caso de RQ3.2 y RQ3.3, se ve apropiado el uso de una *entrevista* a experimentadores de IS. El uso de la entrevista se justifica como un medio adecuado que permite lograr establecer causas realistas y recomendaciones más

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.6: Evaluación de métodos y herramienta aplicables para RQ2.1 y RQ 2.2

		RQ2.1 y RQ2.2							
		Hito 1: Evidencia formal de la existencia de errores estadísticos con alguna técnica estadística seleccionada.							
		Hito 2: Percepción de los investigadores acerca de la problemática de estudio en la IS							
		Dificultades							
		D2.1	D2.2	D2.3	D2.4	D2.5	D2.6	D2.7	D2.8
Método/Estrategia	Encuesta	Generaliza a partir de una muestra hacia una población.	x	x	x	x	x	x	x
		Aplica la definición de una o varias preguntas de investigación, acerca de la naturaleza de una población particular.	x	x	x	x	x	x	x
		Utiliza a una población representativa (bien definida).	x	x	x	x	x	x	x
		Aplican cuestionarios por lo que su formulación debe ser validada previa su aplicación.	x	x	x	x	x	x	x
		Se requiere un esfuerzo y tiempo mediano-alto para ser llevada a cabo.	x	x	x	x	x	x	x
Herramientas de Apoyo	Análisis temático	Permite análisis en profundidad de un tema específico		x	x		x	x	
		Permite la recopilación de cualquier fuente de información.		x	x		x	x	
		Permite ordenar, codificar y clasificar la información obtenida.		x	x		x	x	
		Permite la formulación de temas a cualquier nivel de abstracción.		x	x		x	x	
		Con los resultados se puede realizar gráficas descriptivas de la información obtenida.		x	x		x	x	
	Cuestionario de revisión	Permite obtener de manera sistemática y ordenada información necesaria y relevante relacionada al tema de estudio.	x	x		x			x
		Permite formular preguntas para poder indagar un tema en particular o evaluación.	x	x		x			x
		Su uso maximiza la validez de la información obtenida en la evaluación.	x	x		x			x
		Permite categorizar las preguntas (ej. Dicotómicas) como para evaluar la existencia o no de alguna evidencia.	x	x		x			x

adecuadas, a partir de las experiencias de experimentadores (D3.1, D3.4) cuando el background literario del tema en IS(D3.3) es bastante limitado (D3.2). El uso del análisis temático [Cruzes and Dybå, 2011] y análisis narrativo se justifican por el manejo adecuado de la información obtenida, tal y como se muestra en la Tabla 4.8.

La evaluación de los métodos de investigación, nos permitió concluir que es prácticamente imposible llegar a obtener resultados para todas nuestras preguntas de investigación con un único método de investigación. En su lugar, la aplicación y combinación de los varios métodos de investigación (p. ej., encuestas y entrevistas) permitirá alcanzar los objetivos planteados para la presente tesis doctoral. La combinación de los métodos seleccionados sugieren también la combinación de varios diseños de investigación (p. ej., cualitativo o cuantitativo), por lo que la metodología de investigación de la presente tesis doctoral puede definirse como un método mixto de investigación [Creswell, 2002] con una organización secuencial, tal y como se muestra en la Figura 4.1

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Tabla 4.7: Evaluación de métodos y herramienta aplicables para RQ3.1

		RQ3.1			
		Hito 1: Catálogo formal de las causas más probables de dichos errores estadísticos en las disciplinas experimentales .			
		Dificultades			
		D3.1	D3.2	D3.3	
Metodo/Estrategia	Revisión Sistemática SLR	Permite obtener estudios primarios relacionados a un tema de estudio.	x	x	x
		Permite realizar la búsqueda de estudios de forma sistemática a partir de bases de datos bibliográficas.	x	x	x
		Permite recolectar información relevante (sin sesgos) o evidencias sobre un problema en específico.	x	x	x
		Utiliza métodos explícitos y sistemáticos en su conducción que limita el sesgo (error sistemático).	x	x	x
		Proporciona resultados válidos y confiables sobre los cuales sacar conclusiones y tomar decisiones.	x	x	x
Herramientas de Apoyo	Análisis temático	Permite análisis en profundidad de un tema específico.	x		x
		Permite la recopilación de cualquier fuente de información.	x		x
		Permite ordenar, codificar y clasificar la información obtenida.	x		x
		Permite la formulación de temas a cualquier nivel de abstracción.	x		x
		Con los resultados se puede realizar gráficas descriptivas de la información obtenida.	x		x

Tabla 4.8: Evaluación de Métodos y Herramientas de Apoyo para RQ3.2 y RQ3.3

		RQ3.2 y RQ3.3					
		Hito: Causas y recomendaciones para evitar o al menos minimizar el impacto del fenómeno de estudio en IS					
		Dificultades					
		D3.1	D3.2	D3.3	D3.4	D3.5	D3.6
Metodo/Estrategia	Entrevista	Las preguntas se encuentran íntimamente relacionadas a los objetivos de la investigación (no estructurada, semi-estructurada, estructurada)	x	x	x	x	x
		Se puede o no utilizar una guía para la entrevista para profundizar en un tema específico.	x	x	x	x	x
		Permite obtener profundidad en las sesiones dependiendo de lo que busque el investigador.	x	x	x	x	x
		Aplican cuestionarios por lo que su formulación debe ser validada previa su aplicación	x	x	x	x	
		Permite al investigador conocer mucho más sobre el campo de estudio.	x	x	x	x	x
		Permite soporte de grabación	x	x	x	x	x
		Permite análisis en profundidad de un tema específico	x	x	x	x	
Herramientas de Apoyo	Análisis temático	Permite la recopilación de cualquier fuente de información	x	x		x	
		Permite ordenar, codificar y clasificar la información obtenida		x		x	
		Permite la formulación de temas a cualquier nivel de abstracción		x		x	
		Con los resultados se puede realizar gráficas descriptivas de la información obtenida		x	x	x	x

4.3. Plan de trabajo

El plan de trabajo de la tesis doctoral responde al *diseño secuencial de métodos mixtos* que fue descrito en la sección anterior. Para tal efecto, creamos varias etapas que contienen a las actividades de los métodos de investigación. Se crearon en total 3 etapas de investigación que están fuertemente unidas al objetivo de la tesis doctoral. Las etapas son las siguientes: (1) etapa de *exploración* de los errores estadísticos en las distintas ciencias experimentales, incluida la IS; (2) etapa de *análisis* de las evidencias y percepción de los errores estadísticos en la IS y finalmente (3) etapa de *diagnóstico* de las causas de errores estadísticos y establecimiento de recomendaciones. Tales etapas se muestran en la Figura 4.2

Figura 4.1: Diseño secuencial de métodos mixtos

y se describen a continuación:

4.3.1. Etapa 1 - Exploración

La primera etapa tiene como finalidad determinar, de forma exploratoria, qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales (RQ1.1), así como determinar hasta qué punto los errores estadísticos encontrados durante la exploración, pueden también ocurrir en IS (RQ1.2). La etapa posee dos tareas principales: (1) tarea de *Exploración preliminar de errores estadísticos* para identificar los errores estadísticos comúnmente discutidos en monografías especializadas o literatura general, y (2) tarea de *Exploración y evaluación inicial de errores estadísticos en IS*, que tiene como finalidad determinar si dichos errores estadísticos ocurren o aparecen en artículos experimentales de alta calidad publicados en IS, para lo cual se aplica un cuestionario de evaluación. Los resultados de esta etapa nos brindan una introducción al fenómeno de estudio cuando en la IS, aparentemente, este tipo de información es casi inexistente. Los resultados de esta etapa son reportados en el Artículo I (ver apéndice web en el enlace: <https://goo.gl/T4gSva>).

Tarea 1: Exploración preliminar de errores estadísticos

Existe una buena cantidad de literatura, como por ejemplo [Good and Hardin, 2012, Huck, 2009, Vickers, 2010], que describen errores estadísticos que ocurren

CAPÍTULO 4. MÉTODO DE INVESTIGACIÓN Y PLAN DE TRABAJO

Figura 4.2: Plan de Trabajo Utilizado en el Tesis Doctoral

en las ciencias. En IS, son pocos los trabajos existentes, y los que hay son bastante especializados como: poder estadístico [Dybå et al., 2006], heterogeneidad en el meta-análisis [Miller, 1999] y la aplicación de diseños cross-over [Kitchenham et al., 2003, Vegas et al., 2016].

En cualquier caso, no existe una tipología preliminar de errores estadísticos comunes (Hito de RQ1.1), ni en las ciencias experimentales, ni en IS. Esto dificulta seriamente a nuestra investigación, ya que la ausencia de una tipología preliminar impide realizar una evaluación consistente para obtener posteriormente evidencias preliminares.

Por este motivo, en la presente tarea realizamos una revisión “quick and dirty” [Wheeler, 2014] y un análisis temático [Cruzes and Dybå, 2010], que nos permita, al menos de forma exploratoria, seleccionar literatura inicial (solo libros) que facilite, recopilar información acerca de los varios errores estadísticos que comúnmente aparece en esta literatura especializada. Una vez identificada la literatura, se obtiene información de los errores estadísticos, la cual es: organizada, clasificada y codificada. De esta manera, es posible crear varios temas con diferentes niveles de abstracción obteniendo como posible resultado una tipología preliminar de los errores estadísticos comunes.

Tarea 2: Exploración y evaluación inicial de errores estadísticos en IS

Como se explicó en la tarea anterior, la IS posee una muy limitada literatura referente a los errores estadísticos, y dicha literatura es muy especializada. Por ello que encontrar evidencias de errores estadísticos específicos de IS, es casi imposible. La tarea anterior nos permite obtener una tipología preliminar de errores estadísticos comunes que podrían estar sucediendo en las disciplinas experimentales. No obstante, el resultado solo es un insumo para conocer hasta qué punto los errores estadísticos también ocurren en la IS (RQ1.2).

Por ello, con los errores obtenidos en la tipología preliminar, revisaremos varios artículos experimentales publicados en 10 ediciones (2006-2015) de *Int'l Conference on Software Engineering (ICSE)*. Conocemos que en el ICSE, los métodos estadísticos, no juegan un papel prominente, sin embargo, también es considerado el lugar donde se pueden encontrar una gran cantidad de experimentos pertenecientes a las diferentes áreas de IS (p. ej., requerimientos, testing, etc.). Por ello que es un buen comienzo de nuestra investigación. Utilizaremos cuestionarios de revisión que nos ayudarán durante la recogida y evaluación de información relacionado con los errores estadísticos. Las principales actividades de esta tarea consisten en la elaboración y validación del cuestionario de revisión, la selección de los errores estadísticos comunes a partir de la tipología obtenida en la Tarea 1, y la revisión-evaluación de artículos experimentales.

4.3.2. Etapa 2 - Análisis

La segunda etapa tiene como finalidad determinar la prevalencia de los errores estadísticos más frecuentes encontrados en la primera etapa dentro de los

estudios publicados en IS (RQ2.1) y a partir de estas evidencias, establecer si la comunidad de IS es realmente consciente de este problema (RQ2.2). La etapa se encuentra definida principalmente en dos tareas: (1) tarea de *Identificación de errores estadísticos en publicaciones de IS*, para identificar errores estadísticos en publicaciones de IS, y la (2) tarea de *Análisis de la percepción de los investigadores en IS* que tiene como finalidad determinar el grado de consciencia en los experimentadores acerca de los errores estadísticos encontrados en la tarea anterior. Los resultados de esta etapa nos brindan el grado situacional del fenómeno de estudio en IS, como se ha indicado con anterioridad los estudios referentes al tema, son casi inexistentes. Los resultados de esta etapa son reportados en los Artículos III y IV (ver apéndice web en el enlace: <https://goo.gl/T4gSva>).

Tarea 3: Identificación de errores estadísticos en publicaciones de IS

En la IS es muy difícil determinar con seguridad la existencia de distintos tipos de errores estadísticos dada su limitada o casi inexistente cantidad de trabajos que identifican o evalúan errores estadísticos en artículos experimentales. Al no existir en la comunidad una clara evidencia formal de la existencia de errores estadísticos (Hito RQ2.1), se dificulta seriamente nuestra investigación, ya que la ausencia de evidencias no permite determinar la percepción de los investigadores acerca del fenómeno en IS.

Por tal motivo, en la presente tarea realizaremos una revisión y evaluación de artículos experimentales publicados en 7 ediciones (2007-2013) del *Empirical Software Engineering and Measurement (ESEM)*, que mediante cuestionarios de revisión permitirán guiar al investigador en la evaluación y recogida de información valiosa para nuestro estudio. El ESEM es un congreso donde la metodología experimental es el tema principal y es también considerado uno de los foros más representativos de la experimentación en IS, con lo que es posible obtener evidencias más claras de la existencia de errores estadísticos en IS.

Las principales actividades consisten en la selección de una técnica estadística con base a los resultados de las Tareas 1 y 2, la mismas que fueron utilizadas para establecer errores estadísticos en IS. Luego, se realizará la elaboración y validación del cuestionario que será utilizado para la revisión. Finalmente, se revisan los artículos experimentales para obtener evidencias formales que avalen la existencia de errores estadísticos en IS.

Tarea 4: Análisis de la percepción de los investigadores en IS

La falta de literatura acerca de los errores estadísticos en IS ha generado una falta de discusión del fenómeno entre los investigadores de IS. Por esta razón, los

investigadores son probablemente apenas conscientes de la existencia de errores estadísticos en IS. Esto dificulta seriamente a nuestra investigación, ya que la ausencia de literatura impide averiguar las causas más realistas del apareamiento de este fenómeno. Por ello, es necesario hacer conscientes a los propios experimentadores acerca de la existencia de este fenómeno.

En la presente tarea, realizamos una encuesta [Runeson et al., 2012] a experimentadores que han publicado sus experimentos durante los años 2006-2015 en el *Empirical Software Engineering and Measurement* (ESEM), y en otros congresos y revistas de alto impacto como: *Empirical Software Engineering Journal* (EMSEJ), *IEEE Transactions on Software Engineering Journal* (TSEJ) e *International Conference on Software Engineering* (ICSE).

Las principales actividades consisten en: la elaboración, validación y pilotaje de la encuesta con base a los resultados de la tarea anterior. Se seleccionan a los experimentadores a quienes se les aplicará la encuesta. Finalmente, se analizará y cuantificará la información obtenida mediante un análisis temático [Cruzes and Dybå, 2010]

4.3.3. Etapa 3 - Diagnóstico

La tercera y última etapa tiene como finalidad establecer las causas del apareamiento de errores estadísticos en las disciplinas experimentales (RQ3.1), así como establecer las principales causas de la aparición de errores estadísticos en estudios experimentales en IS (RQ3.2). También se intentará plantear qué medidas se pueden tomar para evitar la comisión de errores (RQ3.3). Para ello, en esta etapa se definieron tres tareas principales: (1) *Identificación de posibles causas de los errores estadísticos en disciplinas experimentales*, (2) *Análisis de las causas de los errores estadísticos en IS*, y (3) *Plantear recomendaciones para mitigar el impacto de los errores estadísticos en IS*. Los resultados finales de esta etapa son las causas de los errores y recomendaciones de solución orientadas a la realidad de la práctica experimental de IS, las cuales son reportados en nuestro Artículo V (ver apéndice web en el enlace: <https://goo.gl/T4gSva>).

Tarea 5: Identificación de posibles causas de los errores estadísticos en disciplinas experimentales

La literatura referente a las causas de errores estadísticos en las distintas ciencias es bastante amplia. Podemos hablar de cientos de errores estadísticos, y por lo tanto de un número elevado de causas. En la presente tarea realizamos una revisión sistemática de literatura (SLR) [Kitchenham et al., 2010, Brereton et al., 2007] que permitirá la búsqueda y selección de estudios primarios que evidencian diversos tipos de errores estadísticos en la práctica experimental, y sus posibles

causas de apareamiento (Hito de RQ3.1). La búsqueda de estudios referente al tema se realizará en bases de datos indexadas como: SCOPUS, Web of science, IEEE Explorer y ACM. La información de errores estadísticos y las causas contenidas en los estudios primarios será organizada, clasificada y codificada mediante análisis temático [Cruzes and Dybå, 2011] para obtener mayor validez en los resultados.

Tarea 6: Análisis de las causas de errores estadísticos en IS

Ya hemos indicado anteriormente la falta de literatura y trabajos acerca del apareamiento de errores estadísticos en IS. Este hecho ha motivado que en la comunidad IS se haya discutido muy poco acerca de este tema entre sus investigadores. A eso se suma lo complejo de determinar las causas en la literatura general, dado que la mayoría de las causas son bastantes genéricas como para aplicar directamente a la IS. Por esta razón, consideramos que una opción para obtener esta información es acudir al propio conocimiento de varios investigadores de IS que manejen muy bien este tema (Hito de RQ3.2).

Para superar este desafío, realizamos una entrevista [Runeson et al., 2012] a investigadores con experiencia en la práctica experimental de IS. Los investigadores serán entrevistados durante el congreso: Ibero-American Conference on Software Engineering (CIBSE), así como en otros foros.

Las principales actividades de esta tarea consistirán en la elaboración, validación y piloteo de la entrevista. Una vez realizadas las entrevistas se procederá a la recolección, limpieza y análisis de la información. Para esto último, se pretende utilizar un análisis temático [Cruzes and Dybå, 2011] para la clasificación y codificación en miras de obtener una etiquetación de causas. A modo de triangulación, el etiquetado tendrá en consideración los resultados de la Tarea 5.

Tarea 7: Recomendaciones para mitigar el impacto de los errores estadísticos en IS

La literatura u opiniones con recomendaciones realistas y adecuadas para minimizar el impacto de errores estadísticos en IS son prácticamente inexistentes, dada la limitada difusión del tema entre los investigadores de IS.

Sabemos de la existencia de esta carencia; por lo que podemos considerar, al menos para esta tesis doctoral, obtener a partir de investigadores con experiencia en el manejo de la estadística en la práctica experimental, recomendaciones que éstos pueden aportar a nuestra investigación (Hito de RQ3.3).

Realizamos una entrevista [Runeson et al., 2012] a investigadores que manejan adecuadamente la experimentación y de forma aceptable la estadística en IS.

Realizar una entrevista a investigadores con este perfil no es fácil, debido a la relativa escasez de investigadores de este tipo, así como a motivos logísticos que cada sesión puede presentar. Por ello, esta tarea se realiza de forma simultánea con la Tarea 6, y de esta forma aprovechar para obtener información de recomendaciones en paralelo a la identificación de las causas. De igual forma, se utilizará un análisis temático para el manejo de la información resultante.

5 | EJECUCIÓN Y RESULTADOS

Resumen: En este capítulo, resumimos la ejecución de la investigación y los principales resultados alcanzados. Los detalles de la investigación de la presente tesis doctoral se encuentran disponibles en los artículos I a V, las cuales están incluidas en el apéndice web en el enlace: <https://goo.gl/T4gSva>.

Los resultados se presentan separados por preguntas de investigación. La visión integrada de los mismos se expondrá en el Capítulo 6 (Discusión).

5.1. Objetivo 1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.

5.1.1. RQ1.1: ¿Qué problemas de diseño y errores estadísticos ocurren comúnmente en las disciplinas experimentales?

Hemos revisado 3 libros: [Good and Hardin, 2012], [Vickers, 2010] y [Huck, 2009] dedicados a errores y problemas estadísticos con especial énfasis en la investigación real. De ellos, hemos extraído 94 fragmentos de texto que contienen claras indicaciones de la existencia de errores y problemas estadísticos. A pesar que 94 fragmentos de texto parecen un gran número, podría haber sido mayor si hubiéramos seleccionado secciones del texto que aludían a los problemas desde una perspectiva general. El análisis de los fragmentos se realizó aplicando síntesis temática. Los fragmentos fueron codificados para, posteriormente, crear temas (códigos de nivel de abstracción más alto) en los cuales se agrupan los códigos de menor nivel. Se considera criterios de inclusión y exclusión y validaciones cruzadas entre investigadores.

El compendio de fragmentos de texto, códigos y temas está disponible en <https://goo.gl/5dZbYd>. El desarrollo extenso de la respuesta a esta pregunta de investigación se refleja en el Artículo I (ver apéndice web en el enlace: <https://goo.gl/T4gSva>), de la presente tesis doctoral.

Resultados

El resultado del análisis se muestra en la Figura 5.1, la cual posee una estructura similar a un árbol, que representa a los códigos de cada error estadístico como hojas, mientras que los temas constituyen las raíces del árbol. Nótese que en la gráfica, los nodos se vuelven progresivamente más abstractos cuando atravesamos el gráfico de derecha a izquierda. Los nodos se encuentran conectados mediante arcos de grosor variable. El grosor del arco es proporcional al número de veces que los códigos/temas aparecen en los fragmentos de texto. Ello permite identificar de forma sencilla los errores más frecuentes. La figura completa se encuentra disponible en <https://goo.gl/qovXQw>.

Los errores estadísticos comunes pueden clasificarse en tres grupos: a) Experimentación, b) meta-análisis, y c) predicción. Cabe recalcar que la mayoría de

Figura 5.1: Clasificación de Errores Estadísticos en Artículos de Investigación Experimental

errores estadísticos se encuentran relacionados con la experimentación (90 veces). Sólo se encontraron tres referencias al meta-análisis y una a predicción. Esto se debe, en gran medida, a que los libros utilizados [Good and Hardin, 2012, Vickers, 2010, Huck, 2009] están principalmente relacionados con la investigación empírica con énfasis en la experimentación. Otra razón importante es que el meta-análisis y la predicción, en términos históricos, han representado un papel secundario en la investigación empírica, frente a la experimentación.

Para la **Experimentación**, que es el área de interés para esta investigación, se identificaron errores estadísticos en las fases de definición de objetivos, análisis, diseño y reporte. La mayor cantidad de errores estadísticos tiene lugar en la fase de **análisis**, alcanzando 63 de 94 (67 %) de los errores estadísticos identificados. Hay dos fuentes principales de errores estadísticos en esta fase:

- **Aplicación de técnicas inferenciales:** Los investigadores a menudo hacen suposiciones erróneas sobre las pruebas estadísticas clásicas (p. ej., robustez del test-t), conceptos relacionados (p. ej., p-valores, colas), selección de pruebas en circunstancias en las que no se deben aplicar (p. ej., orden de las hipótesis alternativas) o simplemente aplican técnicas de manera subóptima (p. ej., pruebas de baja potencia).
- La **interpretación de los resultados** se basa especialmente en el modelado lineal, que es frecuentemente mencionado. La regresión lineal múltiple es el ejemplo más conocido de modelado lineal. El problema más frecuentemente mencionado es el débil razonamiento subyacente a la definición del modelo lineal. También se informan otros problemas, como la violación de suposiciones y su uso más allá de los límites permitidos (por ejemplo, fuera de la fase lineal). Los problemas con la interpretación de resultados se extiende a muchos otros conceptos supuestamente básicos, como los intervalos de confianza, la significación estadística o los p-valores.

La fase de **diseño** se menciona 9 veces (9.5 %), especialmente en problemas estadísticos que se refieren a:

- **Gestión de unidades experimentales** (p. ej., muestreo y asignación).
- **Asignación al azar** inadecuada.
- **Ausencia del cálculo del tamaño de la muestra.**

Se han identificado 10 errores estadísticos (10.6 %) para el **reporte**. Los problemas más relevantes se refieren a:

- **Ausencia de notificación en los artículos acerca de la realización de ensayos múltiples.**
- **Ausencia de reporte de incidentes o ausencia de estadísticas descriptivas** (p. ej., medidas de tendencia central)

La **definición de objetivos** es la fase que tiene menos errores estadísticos (6 veces, un 6.38 %). Aunque el número de errores estadísticos es menor comparada a las otras fases, los errores son de gran importancia, especialmente en lo referido al **no planteamiento de las hipótesis nulas / alternativas**.

Aspectos claves:

Las principales conclusiones alcanzadas en esta fase de la investigación son las siguientes:

- Los errores en las disciplinas experimentales afectan a la totalidad de las fases del proceso experimental.
- Los errores no son de naturaleza compleja. Afectan a conceptos básicos, tales como la definición de hipótesis, test sencillos o el reporte de información esencial (p. ej., estadísticos descriptivos) de un experimento.

5.1.2. RQ1.2: ¿Hasta qué punto los errores estadísticos ocurren también en la IS?

La tipología preliminar de errores estadísticos resultado de RQ1.1 nos ha permitido desarrollar una lista de verificación para ser aplicada en la búsqueda de errores en publicaciones experimentales de IS.

Tabla 5.1: Lista de verificación de errores estadísticos. La pregunta Q7 no aparece en la Tabla por motivos metodológicos discutidos en el Artículo I

#	Pregunta	Comentario
Q1.1	Are hypotheses null explicitly defined?	Podría parecer anticuado por las críticas a la "Null Hypothesis y Significance Testing" (NHST), sin embargo, hoy en día el NHST sigue siendo el principal enfoque estadístico utilizado en IS.
Q1.2	Are hypotheses alternate explicitly defined?	
Q2	Has the required sample size been calculated?	Podrían no ser aplicables en algunos tipos de experimentos.
Q3	Have subjects been randomly selected?	
Q4	Have subjects been randomly assigned to treatments?	
Q5	Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed?	Pondremos énfasis sólo en las condiciones más habituales (p. ej.: normalidad) antes de aplicar virtualmente cualquier prueba paramétrica.
Q6	In cases where linear models are applied, has the model definition been discussed?	Es una de las técnicas estadísticas más frecuentemente mencionadas
Q8	Do researchers avoid calculating and discussing power post hoc?	Aparece como una fuente de problemas durante el análisis y la presentación de informes.
Q9	Is multiple testing reported and accounted for, e.g., Bonferroni?	
Q10	Are descriptive statistics, such as means and counts, reported?	

Hemos aplicado la lista de verificación mostrada en la Tabla 5.1 para evaluar los experimentos publicados en 10 ediciones (2006-2015) del ICSE. Hemos exa-

minado los artículos del ICSE debido al prestigio que tiene como la conferencia más importante en IS. Adicionalmente, ICSE es un congreso generalista (esto es, las publicaciones no primordialmente empíricas). Por ello, esperamos que nuestra evaluación arroje la prevalencia de errores estadísticos comunes en la mejor investigación en IS; la investigación de menor calidad en IS probablemente esté experimentando una peor situación.

Hemos encontrado 21 experimentos controlados publicados entre 2006 y 2015 (a los cuales hemos denominado como experimentos “standalone”). También hemos podido localizar un gran número de publicaciones que no son experimentos en su totalidad, sino que son publicaciones que contienen experimentos como una sección de validación. A estas publicaciones las hemos denominado experimentos “as evaluation”. Para no sobre-representar a este grupo de trabajos, hemos revisado únicamente 30 artículos, a razón de 3 artículos por año. Se considera criterios de inclusión y exclusión, y validaciones cruzadas entre investigadores.

La selección de publicaciones se encuentra disponible en: <https://goo.gl/qKWEHK>, y la evaluación de errores estadísticos en ambos tipos de experimento se encuentran disponible en: <https://goo.gl/YcXAYk> y <https://goo.gl/vUm9V4>. El desarrollo extenso de esta sección se refleja en el Artículo I del apéndice web en el enlace: <https://goo.gl/T4gSva>

Tabla 5.2: Porcentaje de errores estadísticos en publicaciones ICSE

Fase		Standalone Experiments			Experiments as <i>Evaluation Sections</i>		
		Si	No	N/A	Si	No	N/A
Planificación	Q1.1	66.7 %	33.3 %	0.0 %	13.3 %	83.3 %	3.3 %
	Q1.2	57.1 %	42.9 %	0.0 %	6.7 %	90.0 %	3.3 %
Diseño	Q2	0.0 %	100.0 %	0.0 %	3.3 %	96.7 %	0.0 %
	Q3	28.6 %	71.4 %	0.0 %	13.3 %	86.7 %	0.0 %
	Q4	66.7 %	28.6 %	4.76 %	20.0 %	0.0 %	80.0 %
Análisis	Q5	61.9 %	33.3 %	4.76 %	13.3 %	20.0 %	66.7 %
	Q6	4.8 %	0.0 %	95.24 %	3.3 %	0.0 %	96.7 %
	Q8	85.7 %	9.5 %	4.76 %	36.7 %	0.0 %	63.3 %
Reporteo	Q9	9.5 %	71.4 %	19.07 %	3.3 %	26.7 %	70.0 %
	Q10	95.2 %	0.0 %	4.76 %	76.7 %	13.3 %	10.0 %

Resultados

La Tabla 5.2 muestra el porcentaje de artículos que se ven afectados por los errores estadísticos indicados por cada pregunta de la lista de verificación (Q1-Q10). Todas las preguntas, a excepción de Q6 y Q9, se responden negativamente para algunos artículos. En el caso de Q6, ninguno de los 21 experimentos utilizan modelos lineales, lo que resultó en que la totalidad de las respuestas sean “N/A” (No Aplica). Los aspectos más relevantes que hemos encontrado son:

Experimentos “standalone” y “as evaluation”

- La prevalencia de errores estadísticos para ambos experimentos es sustancial, especialmente en Q1, Q2, Q3 y Q5 (definición hipótesis, cálculo del

tamaño de la muestra, selección aleatoria y verificación de las suposiciones de los tests, respectivamente).

- La selección aleatoria (Q3) presenta altas prevalencias. Sin embargo, este problema no es fácil de resolver en experimentos humanos debido a la molestia de reunir cohortes. A su vez, la selección aleatoria podría aplicarse eficazmente en investigaciones no humanas, por ejemplo, cuando se extraen datos de los repositorios de código.

Experimentos “standalone”

- Los experimentos “standalone” definen las hipótesis nulas (Q1.1) cinco veces más frecuentemente que los experimentos “as evaluation” (66.7 % vs. 13.3 %) e hipótesis alternativas (Q1.2) ocho veces más (57.1 % vs. 6.7 %).
- La prueba múltiple (Q9) es un problema omnipresente en los experimentos de IS. La mayoría de los estudios no informan o corrigen las múltiples pruebas usando métodos adecuados, por ejemplo, Bonferroni.

Experimentos “as evaluation”

- El número de respuestas “N/A” es mucho mayor para este tipo de artículos con respecto a experimentos “standalone”. Esto se debe a que la asignación aleatoria Q4 no suele ser aplicable, ya que uno de los diseños más habituales en experimentos “as evaluation” es el de “matched pairs”.
- La estadística inferencial no se utiliza en muchas ocasiones, por lo que Q5-Q9 no son aplicables. En su lugar, el análisis se realiza mediante estadística descriptiva.

Aspectos claves:

La investigación acerca de los errores estadísticos en artículos experimentales “standalone” y “as evaluation” (concretamente, en el marco de la conferencia ICSE 2006-2015) arroja los siguientes aspectos clave:

- La IS enfrenta los mismos errores estadísticos que fueron identificados en otras disciplinas.
- Los errores estadísticos se encuentran tanto en experimentos “as evaluation” que en experimentos “standalone”.
- Las mayores prevalencias de errores estadísticos (>25-30 %) se encuentran en:
 - (a) La no existencia de la hipótesis estadísticas,
 - (b) La no existencia del cálculo del tamaño de la muestra,
 - (c) La falta de evaluación de los supuestos de las pruebas estadísticas, y
 - (d) La realización de múltiples pruebas que no son corregidas.

5.2. Objetivo 2: Determinar la naturaleza de los errores estadísticos cometidos en IS.

La consecución del Objetivo 1 nos ha permitido concluir que en IS se producen los mismos errores que en otras disciplinas experimentales. En este Objetivo 2, profundizaremos en nuestra comprensión de la naturaleza y causas de dichos errores.

Las pruebas de hipótesis aparecen en el catálogo de errores resultado de RQ1.1. La prueba de hipótesis es un concepto que tiene relación con múltiples aspectos de la investigación experimental, es sencillo y además ha sido partícipe de confusiones y malos usos a lo largo de los años. Por esta razón, consideramos la prueba de hipótesis como un buen punto de arranque para nuestra investigación en profundidad.

Adicionalmente, las pruebas de hipótesis se encuentran ligadas al actual sistema de incentivos a la producción científica. Por ejemplo, la realización de múltiple testing podría estar relacionada con el planteamiento de hipótesis post-hoc (HAR-King, por sus siglas en inglés). Por consiguiente, la investigación en profundidad de las pruebas de hipótesis nos permitirá comprender mejor tanto la vertiente técnica como social de los errores estadísticos en IS.

5.2.1. RQ2.1: ¿Qué problemas se observan en la aplicación de las técnicas estadísticas en IS?

Para dar respuesta a esta pregunta de investigación, utilizaremos la misma estrategia que en el *objetivo 1*: una revisión de artículos experimentales. Previo a la realización de la revisión, se realizó un análisis en profundidad de los problemas específicos en los que puede descomponerse el problema general referido a la prueba de hipótesis. Logramos establecer 3 sub-problemas, que se refieren a:

1. *“Ausencia o falta de relación entre Hipótesis de Investigación e Hipótesis Estadística”*, que hace referencia al grado en que la definición y planteamiento de las pruebas de hipótesis siguen los lineamientos generalmente recomendados en la práctica experimental;
2. *“Uso arbitrario de las pruebas de hipótesis 1-cola y 2-colas”* que suelen ocurrir durante la práctica experimental y
3. *“Asociación incorrecta de técnicas estadísticas y pruebas de hipótesis”* que se aplica durante la fase de análisis de los datos experimentales.

Los sub-problemas representan el fundamento de nuestra profundización en el conocimiento de los errores estadísticos en IS. Para ello, hemos revisado publi-

caciones experimentales del ESEM entre el 2007 y 2013. Elegimos ESEM debido a que es la conferencia principal sobre Ingeniería de Software Experimental. Los criterios de exclusión pueden resumirse en el hecho de que las publicaciones reportasen experimentos “standalone”. De una muestra de 311 trabajos del ESEM en los años señalados, 37 trabajos experimentales fueron seleccionados para revisión, incluyendo validación cruzada entre investigadores. La evaluación de los experimentos se encuentra disponible en <https://goo.gl/9Y9sT1>. El desarrollo extenso de esta sección se refleja en el Artículo III, del apéndice web en el enlace: <https://goo.gl/T4gSva>.

Resultados

“Ausencia o falta de relación entre Hipótesis de Investigación e Hipótesis Estadística”

En la Figura 5.2 se muestran cuantitativamente los problemas encontrados, de los cuales se desprende los siguientes resultados:

- El 100 % de experimentos definen el objetivo general de la investigación.
- El 59 % de experimentos contienen hipótesis de investigación.
- El 30 % de experimentos declaran la hipótesis estadísticas.
- Para el caso de las colas, la situación es preocupante; sólo el 22 % de publicaciones definen las colas.

Figura 5.2: Relación de Problemas, Objetivos, Hipótesis de Investigación e Hipótesis Estadísticas y Colas

“Uso arbitrario de las pruebas de hipótesis 1-cola y 2-colas”

El estudio del “uso arbitrario de las pruebas de hipótesis 1-cola y 2-colas”, también se muestra en la Figura 5.2, de la cual se desprende los siguientes resultados:

- La mayoría de las publicaciones experimentales (78 %), no especifican las colas.
- No existe una relación clara entre el planteamiento de hipótesis estadísticas y la definición de colas: Ocurre con frecuencia que, en experimentos sin hipótesis estadísticas, si aparezcan definidos los tipos de cola.
- Los investigadores en IS, tienden a utilizar frecuentemente preguntas de investigación.
- Las preguntas de investigación aparecen independientemente de si existen hipótesis de investigación (ver Tabla 5.3). Curiosamente, las preguntas de investigación aparecen más frecuentemente en la ausencia de hipótesis estadísticas (ver Tabla 5.4).

Tabla 5.3: Tabla cruzada de RQs vs. hipótesis de investigación

	Tiene hipótesis de investigación	No tiene hipótesis de investigación	Total
Tiene RQs	13 (59 %)	13 (87 %)	26 (100 %)
No tiene RQs	9 (41 %)	2 (13 %)	11 (100 %)
Total	22 (100 %)	15 (100 %)	

Tabla 5.4: Tabla cruzada entre RQs vs. hipótesis estadística

	Tiene hipótesis estadística	No tiene hipótesis estadística	Total
Tiene RQs	19 (63 %)	49 (75 %)	68 (100 %)
No tiene RQs	11 (37 %)	16 (25 %)	27 (100 %)
Total	30 (100 %)	65 (100 %)	

“Asociación incorrecta de técnicas estadísticas y pruebas de hipótesis”.

Los resultados relacionados con la “asociación incorrecta de técnicas estadísticas y pruebas de hipótesis” son muy escasos. Los más relevantes se detallan a continuación:

- Existen pocos errores estadísticos desde el punto de vista técnico. Sólo un trabajo define su hipótesis con 1-cola y usa erróneamente una prueba de 2-colas.
- En técnicas estadísticas “complejas” (p. ej., ANOVA y Kruskal-Wallis) no se suelen especificar las colas.
- En técnicas estadísticas “simples” (p. ej., t-test y Mann-Whitney) comúnmente se definen las colas.

Aspectos claves:

La naturaleza de los errores estadísticos en lo tocante a las prueba de hipótesis en artículos experimentales se resume en:

- Los objetivos de investigación son descritos mayoritariamente.
- La definición de las hipótesis de investigación suelen encontrarse mayoritariamente.

- Casi no se logra identificar la descripción de las hipótesis estadísticas. Ello es compatible con un fenómeno de planteamiento de hipótesis post-hoc (HAR-King).
- Predomina la ausencia de la descripción del tipo de cola (1-2 colas) en las hipótesis estadísticas, por lo que no es posible determinar la direccionalidad a priori. Esto es podría ser compatible con un fenómeno de p-hacking.
- Se omite la especificación de colas en los test estadísticos durante el análisis, por lo que no es posible descartar p-hacking, aunque en nuestra opinión ello podría deberse a un fenómeno de automatismo en el uso de las prueba de hipótesis .
- Sólo se observan algunas discordancias entre el tipo de cola especificada en las hipótesis y el tipo de cola usada en los tests, pero esto puede deberse al hecho de que el tipo de cola es frecuentemente no especificado.

En resumen: podemos decir que cuando los investigadores no definen el tipo de cola, desconocemos los motivos subyacentes; pero cuando definen las colas, por lo general lo hacen correctamente.

5.2.2. RQ2.2: ¿Es la comunidad de IS consciente de la existencia del problema?

Los resultados de RQ2.1 son interesantes, pero no nos permiten conocer las motivaciones que guían a los investigadores en la realización de las pruebas de hipótesis durante el diseño y análisis de experimentos en IS.

Por esa razón, realizamos una encuesta a varios investigadores que han publicado experimentos en el ESEM, EMSE, TSE y ICSE entre los años 2012 y 2015. Previo a la encuesta, se realizó la selección de experimentos en estos congresos y revistas, siguiendo los mismos criterios de inclusión-exclusión utilizadas para RQ2.1. De las publicaciones se obtuvieron 403 correos electrónicos, de los cuales:

- 95 correos electrónicos son de experimentadores del ESEM,
- 93 correos electrónicos del EMSE,
- 124 correos electrónicos de TSE y
- 91 correos electrónicos del ICSE.

Se realizó una encuesta on-line con 21 preguntas basadas en los 3 sub-problemas anteriormente mencionados. El ratio de respuesta fue del $45/403 = 11.20\%$, comparable con los ratios de respuesta de otros surveys en la comunidad IS [de la Vara et al., 2014].

El instrumento de recogida de información se encuentra disponible en <https://www.researchgate.net/publication/312222222>

[//goo.gl/cuUeLx](https://goo.gl/cuUeLx). Las respuestas, y demás documentación están disponibles en <https://goo.gl/QE1kSr>. El desarrollo extenso de esta sección se refleja en el Artículo III, del apéndice web en el enlace: <https://goo.gl/T4gSva>.

Resultados

“Ausencia o falta de relación entre Hipótesis de Investigación e Hipótesis Estadística”

En general, los investigadores indican un mayor grado de adhesión a la práctica experimental estándar que la revisión de literatura reportada en RQ2.1 nos haría suponer:

- El 86 % y 59 % de experimentadores afirman definir la hipótesis de investigación y estadística, respectivamente.
- El 56 % de experimentadores dicen definir las colas con la hipótesis alternativa.
- Las razones para no definir las hipótesis estadísticas se resumen en: “*se asume que están por defecto o son explícitas*”, producen “*sobrecarga cognitiva al lector*”, debido a “*olvidos*” o “*dificultad de interpretación*”, entre otros.

En resumen, los investigadores dicen que hacen cosas que realmente no hacen. En otras palabras, los investigadores parece que realizan **prácticas cuestionables pero no las quieren reconocer**.

“Uso arbitrario de la prueba de 1-cola y 2-colas”

En lo que respecta a la práctica de los experimentadores respecto al “uso arbitrario de la prueba de 1-cola y 2-colas”, se obtienen las siguientes conclusiones:

- La prueba de 1-cola se lo utiliza fundamentalmente para estudios confirmatorios (69 %), y 2-colas para estudios exploratorios (66 %).
- Los 2 tipos de cola son de uso en general, aunque existe un cierto sesgo (66 %) por el uso por las 2-colas.
- La mayor parte de los investigadores (14 %) rechaza por algún motivo la 1-cola, mientras que muy pocos investigadores rechazan la prueba de 2-colas (3 %).
- No se encontró evidencia del cambio de direccionalidad de 1-cola a 2-colas.
- Un 21 % de los experimentadores mencionaron que efectivamente realizan cambio de direccionalidad de 2-colas a 1-cola, a pesar que este procedimiento es incorrecto. Lo preocupante es que de este porcentaje, el 10.5 % logró obtener resultados significativos.

En resumen, los investigadores muestran una preferencia por el uso de 2-colas, lo cual puede estar relacionado con la realización de investigación exploratoria. Se observa, además, una cierta evidencia acerca de la realización de prácticas

cuestionables en IS, tales como: el cambio de 2-colas a 1-cola con la intención de obtener resultados estadísticamente significativos.

“Asociación incorrecta de técnicas estadísticas y pruebas de hipótesis”

Aunque se aprecian algunos sesgos en el uso de la 1-cola y 2-colas, las desviaciones de la práctica estándar son muy evidentes cuando estudiamos la asociación entre las pruebas de hipótesis y la estadística inferencial:

- Con la excepción de la regresión lineal, el 79 % de investigadores dicen aplicar 1-cola con los test ANOVA y Kruskal-Wallis, que son inherentemente test de 2-colas.
- El 59 % de los experimentadores afirman aplicar 2-colas con el Chi-squared test, que es inherentemente 1-cola.
- Un 7 % de los experimentadores afirman haber ensayado varias técnicas estadísticas durante el análisis, al menos en algunas ocasiones.

En resumen, existen problemas probablemente conceptuales, a la hora de seleccionar al tipo de cola asociado con los tests estadísticos.

Aspectos claves:

El modo en que los experimentadores aplican las pruebas de hipótesis está distanciado de lo que observamos en sus publicaciones, así como de lo recomendado por la vertiente más ortodoxa del proceso experimental. Al reflexionar sobre los resultados de la revisión (RQ2.1) y de la encuesta (RQ2.2), podemos concluir lo siguiente:

- Los investigadores desconocemos la naturaleza de los fenómenos de interés en IS. Por ello, muchos experimentos realizados en IS son de carácter exploratorio.
- Caemos en la tentación de realizar HARKing y p-hacking, lo cual (posiblemente) justificamos debido al carácter exploratorio de los estudios (p. ej., planteamiento de preguntas de investigación en lugar de hipótesis estadísticas, no definición del tipo de cola, cambios de 2-colas a 1-cola, ocultación de información en las encuestas por parte de los investigadores).
- Generalmente hablando, los conocimientos estadísticos de los investigadores no son suficientes para afrontar la complejidad de fenómenos y diseños experimentales que estudiamos en IS.

5.3. Objetivo 3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.

Las respuestas a RQ2.1 y RQ2.2 sugieren un conjunto de causas (p. ej., desconocimiento de los fenómenos de interés en IS, HARKing, p-hacking y falta de conocimientos estadísticos por parte de los investigadores) para los errores estadísticos en experimentos en IS. Estas causas se han inferido a partir de una revisión de literatura y de los conocimientos que un conjunto de investigadores nos ha transmitido mediante una encuesta.

Nuestro tercer objetivo es profundizar en el conocimiento de las causas de los problemas estadísticos. Para ello, realizaremos entrevistas en las que vamos a preguntar explícitamente a los investigadores acerca de los problemas estudiados en RQ2.1 y RQ2.2. Trataremos de identificar las causas y posibles acciones paliativas de dichos problemas mediante análisis narrativo.

5.3.1. RQ3.1: ¿Cuáles son las causas de los errores estadísticos en las disciplinas experimentales?

Quisimos formalizar la búsqueda de las causas de los errores estadísticos que se producen. Para ello, utilizaremos una Revisión Sistemática de Literatura (SLR). La SLR comenzó con un Quasi-Gold Estándar (QGS) de 19 artículos basados en los criterios de inclusión-exclusión. El QGS nos brindó los términos apropiados para construir varias cadenas de búsqueda tentativas. Escogimos finalmente una cadena de búsqueda con el mejor recall (71 %) y una precisión aceptable (4 %), con la cual logramos obtener 661 estudios candidatos. Luego de llevar a cabo los procedimientos de identificación, screening y elegibilidad con altos niveles de coincidencias y acuerdos entre investigadores, se lograron obtener 69 estudios primarios. El porcentaje de coincidencias entre investigadores en el proceso de obtención de artículos primarios fue del 80.95 %. En términos de inter-rater agreement obtuvimos el valor de $\kappa = 0,99$ que representa a un “Almost perfect agreement” [Fleiss et al., 2013].

Con los estudios primarios se realizó un proceso de extracción de información, identificando 111 fragmentos de texto que mencionan causas de apareamiento de errores estadísticos. Mediante análisis temático, logramos organizar a los fragmentos de texto en 9 temas de orden superior (codificado como HTC), 48 temas de orden medio (codificado como MTC) y 86 temas de orden bajo (codificado como LTC). La información detallada se encuentra disponible en el siguiente enlace: <https://goo.gl/kgd1ng>. El desarrollo extenso de esta sección se refleja en el Ar-

título II, del apéndice web en el enlace: <https://goo.gl/T4gSva>. El resumen del análisis realizado se muestra en la Tabla 5.5

Tabla 5.5: Catálogo de errores estadístico de nivel de orden superior (High-theme)

High-Themes	No. Prim. Stud	%	No. Med.- Th.	No. Low- Th.	No. Cau- ses
HTC01: [Statistical Misunderstood]	13	18,84 %	3	2	3
HTC02: [statistical concepts]	9	13,04 %	2	1	3
HTC03: [Deliberate use]	8	11,59 %	6	2	6
HTC04: [Fraud]	8	11,59 %	5	1	5
HTC05: [Lack of ...]	37	53,62 %	11	12	62
HTC06: [Assumptions]	17	24,64 %	6	6	12
HTC07: [Professional Ethics]	15	21,74 %	5	17	17
HTC08: [Education]	48	69,57 %	5	19	20
HTC09: [Experimental Design]	28	40,58 %	4	25	26
Total			47	86	111

Resultados

Las causas de apareamiento de errores estadísticos en las diferentes ciencias se concentran en 9 causas principales. Las causas más importantes están relacionadas con:

- HTC08: Educación (69.57 %),
- HTC05: Carencias del investigador (experiencia, asesoría) (53.62 %),
- HTC09: Diseño experimental (40.58 %),
- HTC06: Suposiciones (24.64 %)
- HTC07: Ética profesional (21.74 %), y
- HTC01: Incomprensión de la estadísticos (18.84 %),

y las causas menos frecuentes, aunque no por ello menos importantes, son:

- HTC02: Aplicación de conceptos estadísticos (13.04 %),
- HTC03: Uso deliberado de la estadística (11.59 %), y
- HTC04: Fraude (11.59 %).

La caracterización de los estudios primarios se muestra en la Figura 5.3a. Hay dos aspectos particularmente relevantes. Por una parte, se puede observar que el número de estudios publicados relacionados con problemas estadísticos es puramente testimonial en las décadas de los 70's, 80's y 90's, donde aparecen entre 0-1 estudios por año. Sin embargo, a partir del año 98, el número de estudios aumenta y la tendencia parece creciente. Por otra parte, los primeros estudios surgen en el área de Medicina, pero en fechas recientes el interés en el tema relacionado a problemas estadísticos ha surgido en otras áreas, incluso en TIC.

Figura 5.3: (a) Evolución Temporal de Estudios Primarios (b) Causas que son Nombradas por las Disciplinas

Aspectos claves:

Los aspectos claves encontrados en la SLR acerca de las causas de errores estadísticos en las disciplinas experimentales son:

- No aparecen estudios primarios pertenecientes a la IS. La Medicina y Psicología poseen el mayor número de publicaciones que aportan a causas de errores estadísticos en su práctica experimental.
- Los estudios primarios sugieren causas que son bastante generales para todas las ciencias.
- Las causas más frecuentes se encuentran relacionadas con la educación y carencias del investigador y al diseño experimental, entre otros, y en mucha menor frecuencia con el fraude.

5.3.2. RQ3.2: ¿Cuales son las causas más probables de los errores estadísticos en IS?

Realizamos 5 entrevistas a investigadores que poseen un perfil orientado a la experimentación, y que asumen responsabilidades en los aspectos estadísticos de sus actividades de investigación. Los investigadores pertenecen a diferentes grupos de investigación de Ingeniería de Software Empírica en Iberoamérica.

Luego de un análisis narrativo a partir de la información extraída en las entrevistas, obtuvimos 123 fragmentos de texto que enuncian posibles causas de problemas estadísticos, los cuales catalogamos con 36 etiquetas únicas. Para

poder clasificar las etiquetas, se crearon en total 6 temas orientados hacia aspectos relacionados con: el Investigador, la Investigación Experimental en IS, la Metodología, la Educación, la Estadística en la Experimentación y Control de los sesgos. La información de la síntesis de las causas se encuentra en el siguiente enlace: <https://goo.gl/3fApsW>. El desarrollo extenso de esta sección se refleja en el Artículo V, del apéndice web en el enlace: <https://goo.gl/T4gSva>.

Tabla 5.6: Causas principales de la aparición de defectos metodológicos de pruebas de hipótesis

Contexto	Alta importancia	Alta especificidad	Alta importancia y alta especificidad
Conocimiento acerca de defectos estadísticas	<ul style="list-style-type: none"> - Desinterés - Limitada literatura acerca de experimentación en IS 	<ul style="list-style-type: none"> - Literatura acerca de errores estadísticos basada en caso particulares 	<ul style="list-style-type: none"> - Desconocimiento del fenómeno investigado
Pruebas de hipótesis sistemáticas	<ul style="list-style-type: none"> - Desconocimiento del investigador 	<ul style="list-style-type: none"> - Incertidumbre debido a poco poder estadístico - Experimentación es demasiado rígida - Planteamiento de suposiciones - Atomización de la investigación - Falta de continuidad en la investigación - Carencia de una práctica estándar - Es difícil contestar a una pregunta de investigación cuando está formada por hipótesis - Pregunta de investigación similar a la hipótesis - Se usan preguntas de investigación cuando no se desea profundizar - Pregunta de investigación relacionada con estudios exploratorios - Pregunta de investigación ligada a una hipótesis 	<ul style="list-style-type: none"> - Exploración iterativa - Poco espacio para el reporte - Investigación Exploratoria no requiere hipótesis
Sesgos de la prueba de 1-2 colas	<ul style="list-style-type: none"> - Desconocimiento del investigador 	<ul style="list-style-type: none"> - Contradicción en literaturas estadísticas - Preconcepción acerca del uso de la estadística 	<ul style="list-style-type: none"> - Fishing
Coherencia de las pruebas de hipótesis con la estadística inferencial	<ul style="list-style-type: none"> - Desconocimiento del investigador 	<ul style="list-style-type: none"> - La asesoría estadística es costosa 	<ul style="list-style-type: none"> - Inclusión de investigadores no expertos en ISE en el área de experimentación

Resultados:

Las causas pueden clasificarse de acuerdo a dos dimensiones: importancia y especificidad. Los detalles de esta clasificación pueden encontrarse en el Artículo V (ver apéndice web en el enlace: <https://goo.gl/T4gSva>.) de esta tesis doctoral. Nótese que aunque las causas se han obtenido tomando como base a los errores estadísticos relacionados con las pruebas de hipótesis, creemos que su validez es más general.

En la Tabla 5.6 resumimos las causas más relevantes que ocasionan errores estadísticos. Podemos destacar los siguientes aspectos:

- En todos los sub-problemas estudiados (“Ausencia o falta de relación entre Hipótesis de Investigación e Hipótesis Estadística”, “Uso arbitrario de la prueba de 1-cola y 2-colas”, “Asociación incorrecta de técnicas estadísticas y pruebas de hipótesis”) el desconocimiento de investigador aparece como la causa más frecuentemente mencionada. Desde nuestra perspectiva, ello probablemente significa que el desconocimiento del investigador es la causa de mayor importancia en la comisión de errores estadísticos.
- Existen varias causas que comparten una alta importancia y alta especificidad. Todas ellas, con dos excepciones, coinciden con causas ya hipotetizadas durante la respuesta a las preguntas de investigación RQ2.1 y RQ2.2: desconocimiento del fenómeno investigado, realización de investigación exploratoria, y fishing (que en el contexto informal de las entrevistas puede entenderse indistintamente como HARKing o p-hacking). Las excepciones son:
 - Disponer de poco espacio en los manuscritos para el reporte de experimentos.
 - El apareamiento de investigadores no adecuadamente formados en estadística, dentro del área de investigación de Ingeniería del Software Experimental (ISE).
- Existen otro buen número de causas específicas. Estas causas específicas podrían subsumirse en gran medida en las causas de alta importancia mencionadas anteriormente.

Aspectos clave:

Creemos poder afirmar que los sesgos y errores que subyacen a los errores estadísticos en IS, coinciden con lo mencionado a las ciencias, además de algunas causas específicas de la IS. Parece ser que el **desconocimiento del investigador** es la principal causa de los errores estadísticos en pruebas de hipótesis en experimentos de la IS. Los sesgos (HARKing, p-hacking) apenas se comentan como causa de los errores estadísticos; concretamente, sólo se menciona una vez, en relación al uso inapropiado de 1-2 colas.

Aparte de estos dos aspectos, se menciona un buen número de diversas causas. Algunas de ellas ya habían sido detectadas durante el estudio de RQ2.1 y RQ2.2 (desconocimiento de los fenómenos investigados en IS y realización de investigación exploratoria). Sin embargo, hay uno nuevo que nos resulta especialmente llamativo: *la aparición de investigadores no expertos en experimentación, pero que realizan y publican experimentos*.

5.3.3. RQ3.3: ¿Qué medidas se pueden tomar Para evitar la comisión de errores?

Con el mismo enfoque de análisis narrativo de las entrevistas realizado para la RQ3.2, se obtuvieron 55 fragmentos de texto que enuncian varias recomendaciones, las cuales resultaron en 20 etiquetas únicas. La información de la síntesis de las recomendaciones se encuentra en el enlace: <https://goo.gl/6NDXCQ>. El desarrollo extenso de esta sección se refleja en el Artículo V, del apéndice web en el enlace: <https://goo.gl/T4gSva>.

Tabla 5.7: Medidas se pueden tomar para evitar la comisión de errores con pruebas de hipótesis

Contexto	Alta importancia	Alta especificidad	Alta importancia y alta especificidad
Conocimiento acerca de defectos estadísticos	- Guías claras de cómo aplicar los métodos estadísticos	-	- Estudiar problemas relevantes
Pruebas de hipótesis sistemáticas	- Mejorar la educación estadística	- Aumentar la libertad del investigador - Regresar a la bases filosóficas	- Definir protocolo de investigación
Sesgos de la prueba de 1-2 colas	- Mejorar la educación estadística	- Tomar suposiciones conservadoras	- Incentivar la educación ética experimental - Mecanismos para minimizar el sesgo
Coherencia de las pruebas de hipótesis con la estadística inferencial	- Guías claras de como aplicar los métodos estadísticos	- No tener prisa durante la investigación	- Disponer de asesoría estadística

Resultados:

Las recomendaciones más relevantes para evitar o al menos minimizar el impacto de los errores estadísticos de las pruebas de hipótesis en la práctica experimental de IS se resumen en la Tabla 5.7. Nótese que aunque las recomendaciones se han obtenido como base a las causas de los errores estadísticos relacionados con las pruebas de hipótesis, creemos que su validez es más general.

Dado que la investigación relacionada con RQ3.2 y RQ3.3 se realizó mediante entrevista abierta, las causas de los problemas estadísticos, y las recomendaciones de solución, surgían espontáneamente en la conversación. Por ello, no es posible relacionar unívocamente causa-recomendación. No obstante, es razonable suponer que la superposición de las Tablas 5.6 y 5.7 corresponde aproximadamente con la relación causa-recomendación. Los resultados que hemos obtenido son los siguientes:

- En todos los sub-problemas estudiados (“Ausencia o falta de relación entre Hipótesis de Investigación e Hipótesis Estadística”, “Uso arbitrario de la prueba de 1-cola y 2-colas”, “Asociación incorrecta de técnicas estadísticas y pruebas

de hipótesis”), el *mejorar la educación estadística* y crear *guías claras de cómo aplicar los métodos estadísticos* aparecen como las recomendaciones de alta importancia como para evitar el **desconocimiento del investigador**

- Para las restantes causas, las recomendaciones de solución son particulares. De nuevo, en nuestra opinión destacan las recomendaciones de alta importancia y especificidad, tanto debido a sus concreción como posibilidad práctica de implementación:
 - Estudiar problemas relevantes,
 - Incentivar la educación ética experimental,
 - Establecer mecanismos para minimizar el sesgo (fishing), entre los cuales se encuentra la definición explícita de protocolos de investigación, y
 - Disponer de asesoría estadística, en particular para ayudar a los investigadores no expertos en estadística que se inician en la Ingeniería del Software Experimental.

Aspectos clave:

En resumen, vemos que **mejorar la Educación Estadística** y disponer de **guías claras de cómo aplicar los métodos estadísticos** es el camino principal para minimizar o al menos paliar la mayoría de errores estadísticos que aparecen en la IS.

6 | Discusión

Resumen: En este capítulo resumimos las principales contribuciones de esta investigación, reflexionando acerca de cómo estas contribuciones han permitido el logro de los objetivos planteados.

6.1. Discusión

El principal objetivo de esta investigación fue explorar, identificar, analizar y evaluar los errores estadísticos en IS, con el fin de establecer sus causas y proponer mecanismos para evitarlos o, al menos, minimizar su impacto. Este objetivo general se descompuso en tres objetivos específicos:

- Identificar los errores estadísticos comunes y sus causas probables en las distintas ciencias, incluida la IS.
- Determinar la naturaleza de los errores estadísticos cometidos en IS.
- Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.

En lo que sigue, discutiremos en qué medida los resultados de la presente tesis doctoral han permitido el logro de los objetivos planteados.

6.1.1. Identificar si existen errores estadísticos en IS, y establecer su prevalencia.

El primer sub-objetivo tenía un carácter fundamentalmente académico, de fundamentación de la investigación reflejada en la presente Tesis Doctoral. En efecto, si ciencias asentadas como la medicina o psicología cometen errores en sus publicaciones experimentales, resultaría de lo más sorprendente que ello no ocurriera en IS. No obstante, es un aspecto que necesitábamos comprobar explícitamente. Para ello, hemos realizado dos actividades, correspondientes a las preguntas de investigación RQ1.1 - RQ1.2:

- En la primera de estas actividades (RQ1.1), hemos realizado una revisión “Quick and Dirty” de varios libros especializado en errores estadísticos en las ciencias experimentales. Esta revisión nos permitió definir una **tipología preliminar de errores estadísticos**¹. Esta tipología no contiene errores sofisticados, o relacionados con técnicas estadísticas complejas. Al contrario, la mayoría de estos errores son sorprendentemente simples: ausencia de hipótesis estadísticas, aleatorización incorrecta, o multiple testing.
- La existencia de errores estadísticos en las ciencias experimentales (recogidos en la tipología de errores indicada anteriormente) no implica que dichos

¹ Esta tipología no pretende ser un catálogo completo de errores estadísticos; para ello estamos realizando una SLR, que no ha podido ser completada a tiempo de la defensa de la presente tesis.

errores ocurran en IS. Podría ser perfectamente posible (aunque poco probable) que la IS estuviera libre de errores. Para comprobar este extremo, la segunda actividad (RQ2.2) consistió en **revisar la existencia de errores en artículos experimentales de IS**. Para ello, revisamos 10 ediciones del Int'l Conference on Software Engineering (ICSE), el cual representa, probablemente, el foro no especializado en investigación experimental donde se publican los trabajos experimentales de mayor calidad. Los resultados no han dejado lugar a dudas: **los mismos errores que afectan a las ciencias experimentales (p. ej., aleatorización incorrecta, múltiple testing, etc.) también ocurren en IS**. Esta constatación justifica la realización del resto de actividades de investigación incluidas en la presente Tesis Doctoral.

6.1.2. Determinar la naturaleza de los errores estadísticos cometidos en IS.

Una vez firmemente establecida la existencia de errores en IS, nuestro interés se derivó a conocer por qué se producen dichos errores. De forma creemos que natural, intentamos estudiar una serie de errores distintos, con la finalidad de aumentar la generalizabilidad de nuestras conclusiones. Sin embargo, pronto nos dimos cuenta de que ello no resultaba una tarea sencilla. El análisis de publicaciones sólo permite determinar la existencia o no de un error (ej: multiple testing), pero en general **los motivos que motivan la existencia del error quedan ocultos** (ej: los autores no explican el porqué de su decisión).

Esta dificultad fue superada cuando prestamos atención a los errores relacionados con las pruebas de hipótesis. A diferencia de otros tipos de errores, **los errores relacionados con las pruebas de hipótesis dejan un rastro explícito** a lo largo del paper (ej: la relación entre el tipo de cola de las hipótesis estadísticas y de los tests estadísticos) que puede utilizarse para inferir por qué los investigadores cometían dichos errores. Por este motivo, nos decidimos a estudiar este error en profundidad. Para ello, hemos realizado dos actividades, correspondientes a las preguntas de investigación RQ2.1 - RQ2.2:

- La primera actividad realizada (RQ2.1) consistió en estudiar los errores relacionados con las pruebas de hipótesis cometidos en artículos experimentales publicados en el Int'l Symposium on Empirical Software Engineering and Metrics (ESEM). En este caso, hemos escogido una conferencia especializada en investigación experimental (al contrario que el ICSE), para "eliminar ruido" en la investigación, como por ejemplo el análisis de errores simples cometidos por investigadores novatos. Esto podría desvirtuar (ej: simplificar) los resultados de la investigación.

La revisión de artículos puso de manifiesto que la prueba de hipótesis era un aspecto propenso a errores en los trabajos experimentales. Esto no supuso

una sorpresa, ya que la revisión de artículos del ICSE ya apuntaba hacia esta conclusión. Resultaron más interesantes otros aspectos surgidos durante la investigación:

- **Casi nunca se describen las hipótesis estadísticas.**
 - **Apenas se describe el tipo de cola** (1-2 colas) en las hipótesis estadísticas.
 - **Se omiten la especificación de las colas** de los tests estadísticos.
- Los resultados de la primera actividad eran perfectamente **compatibles con los fenómenos de p-hacking o HARKing**, esto es, seleccionar las hipótesis y tests estadísticos que arrojen resultados significativos. Este fenómeno ya ha sido reportado con anterioridad en IS [Jørgensen et al., 2016]. Sin embargo, nos parecía aventurado afirmar la existencia de mala práctica en IS, sin más. Por ello, como segunda actividad (RQ2.2) realizamos una encuesta a investigadores experimentales en IS.

En esta encuesta, preguntamos (de forma anónima) a los investigadores por qué realizaban ciertas prácticas (ej: no especificar las colas de los tests). La encuesta nos permitió profundizar en las motivaciones que subyacen a las acciones de los investigadores (en lo que respecta a la prueba de hipótesis), mientras que el anonimato nos sirvió para asegurar la sinceridad de las respuestas. Las respuestas obtenidas fueron bastante diversas; no obstante, pudimos identificar algunos patrones claros:

1. **Caemos en la tentación de realizar HARKing y p-hacking**, lo cual justificamos debido al carácter exploratorio de los estudios (ej: modificar la cola de los tests de 2-colas a 1-cola).
2. Los investigadores desconocemos la naturaleza de los fenómenos de interés en IS. Por ello **muchos experimentos realizados en IS son de carácter exploratorio**.
3. Generalmente hablando, **los conocimientos estadísticos de los investigadores no son siempre suficientes** para afrontar la complejidad de la investigación experimental.

Los patrones anteriores están ordenados de menor a mayor nivel de especificidad. En primer lugar, surge de manera evidente que los investigadores experimentales en IS realizamos p-hacking. No obstante, los problemas de las pruebas de hipótesis (ej: no especificar el tipo de cola) no parecen tener una relación directa con el p-hacking. En su lugar, **parece que cometemos un p-hacking técnico** (ej: cambiar el tipo de cola de los tests) justificado por la naturaleza exploratoria de la investigación.

Cuando hablamos de naturaleza exploratoria, nos referimos a que muchos experimentos no se basan en una teoría subyacente que desean probar o

refutar, sino que pretenden examinar un fenómeno esencialmente desconocido utilizando para ello la metodología experimental. **Dado que no hay teorías subyacente, tampoco pueden definirse hipótesis estadísticas** o los tipos de cola.

Cabría preguntarse a qué se debe que los investigadores en IS utilicemos una metodología experimental para realizar estudios exploratorios. El tercer patrón encontrado es la respuesta a esta pregunta: aunque hay grandes diferencias entre investigadores, probablemente una cantidad sustancial de nosotros no disponemos de los conocimientos suficientes para aplicar correctamente la metodología experimental. Esto tampoco es extraño. De todos es conocido que la experimentación es una metodología reciente en IS, y que **muchos investigadores hemos aprendido a experimentar "en las trincheras"**. Ello explica a la perfección las carencias de conocimientos antedichas.

6.1.3. Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia

La investigación realizada hasta este punto señalaba que las carencias inherentes de nuestra disciplina (conocimiento de los fenómenos, formación en estadística), además de algún pecadillo (p-hacking) explicaban la comisión de errores en las pruebas de hipótesis. Quedaba pendiente determinar si: a) las mismas causas operaban en otros errores, y b) estudiar modos paliar dichas carencias.

La mejor forma alcanzar dichos sub-objetivos sería realizar repetidamente el estudio llevado a cabo en RQ2.1-RQ2.2, y sintetizar un conjunto de causas mediante saturación teórica, tal y como se aplica en *grounded theory*. Sin embargo, la repetición prácticamente *sine die* del mismo estudio no es factible en la mayoría de los proyectos de investigación, incluidas las tesis doctorales. Por esta razón, hemos optado por una estrategia más incierta, pero más acotada en tiempo y esfuerzo: **realizar entrevistas a experimentadores** para conocer sus opiniones acerca de los defectos estadísticos en IS.

Más concretamente, hemos confeccionado una entrevista, basándonos en las respuestas a las RQ2.1-RQ2.2, en la que inicialmente hemos inquirido acerca de los errores relacionados con las pruebas de hipótesis, para ir posteriormente elevando el nivel de abstracción hacia errores más generales. Antes de realizar estas entrevistas, hemos realizado una SLR acerca de las causas de los errores estadísticos en las ciencias, para poder afrontar las entrevistas con más garantías. En total, hemos realizado tres actividades, correspondientes a las preguntas de investigación RQ3.1 - RQ3.3:

- La primera actividad realizada (RQ3.1) fue la SLR. Después de revisar 69

estudios primarios, pudimos establecer que:

- Los estudios primarios sugieren causas que son bastante generales para todas las ciencias. No obstante, ninguno de los estudios primarios pertenecía a la IS.
- Las causas predominantes se encuentran relacionadas a la educación del investigador y errores en el planteamiento del diseño experimental. El fraude (ej: p-hacking) se enuncia como causa en mucha menor frecuencia.

Los resultados de la SLR estaban sorprendentemente alineados con nuestra respuesta a RQ2.2, esto es, la causa principal de los errores estadísticos en las ciencias era **la deficiente educación estadística del investigador, y no el fraude**. Naturalmente, existían otras causas (ej: el deficiente diseño experimental, antes citado), pero esta Tesis Doctoral no podía penetrar en ellas. El estudio de dichas causas debe ser relegado a las futuras líneas de investigación.

- La segunda y tercera actividades realizadas (RQ3.2 - RQ3.3) consistieron en una entrevista, como ya hemos indicado anteriormente, referida a las causas de los errores y alternativas de solución, respectivamente. Para la defensa de la presente Tesis Doctoral hemos realizado únicamente 5 entrevistas, aunque realizaremos algunas más en el futuro para asegurar la saturación teórica que nos requerirán los revisores de la correspondiente publicación.

En lo que respecta a las causas de los errores (RQ3.2), es difícil resumir cabalmente los resultados. Sin embargo, nos decantamos (y reafirmamos) en que la causa principal de los errores estadísticos en IS parece ser **el desconocimiento, por parte de los investigadores, de las técnicas estadísticas aplicadas a la investigación experimental**.

Además de esta causa principal, los entrevistados han mencionado otras múltiples posibilidades, algunas más importantes y/o específicas que otras. Varias de ellas ya habían sido detectadas durante el estudio de RQ2.1 y RQ2.2 (p. ej., desconocimiento de los fenómenos investigados en IS). Sin embargo, hay uno que nos resulta especialmente llamativo y que no había surgido hasta el momento: **la aparición de investigadores no expertos en experimentación**, pero que realizan y publican experimentos, cometiendo errores con relativa frecuencia. Esto se detecta fácilmente, por ejemplo, durante los procesos de revisión por pares.

Finalmente, en lo que se refiere a los mecanismos para evitar o paliar errores (RQ3.3), los experimentadores entrevistados propusieron múltiples opciones. Sin embargo, dos de ellas destacan por su reiteración en múltiples entrevistas:

- Mejorar la educación estadística.
- Disponer de guías claras de cómo aplicar los métodos estadísticos.

Estas alternativas de solución no hacen otra cosa que dar mayor verosimilitud a nuestra opinión de que los deficientes (en general) conocimientos estadísticos de los experimentadores son la causa principal de los problemas estadísticos en IS.

7 | AMENAZAS A LA VALIDEZ

Resumen: Para que los resultados de la tesis doctoral posean el suficiente rigor, se tomaron en consideración las diferentes amenazas que [Yin, 2013] y [Creswell, 2002] mencionan en sus libros. Más concretamente, se consideró mitigar a las *amenazas a la validez interna* (sección 7.1) que se refieren aquellos aspectos que amenazan la relación causa-efecto entre las variables independientes y dependientes. Asimismo, se examinaron las *amenazas a la validez externa* (sección 7.2) para evitar los aspectos que pueden limitar la generalización de los resultados de la investigación a otros contextos. También se mitigaron las *amenazas a la validez de constructo* (sección 7.3) que influyen negativamente en la relación que hay entre los aspectos teóricos y las observaciones concretas realizadas en la tesis. Finalmente, se consideraron algunos aspectos de fiabilidad (sección 7.4) recomendados por [Creswell, 2002].

Según [Creswell, 2002], utilizar una estrategia de métodos mixtos tiene la ventaja sobre otras aproximaciones metodológicas que las bondades de un método (p. ej., cuantitativo) puede mitigar las amenazas a la validez del otro método (p. ej., cualitativo), y viceversa. Sin embargo, este chequeo cruzado no es suficiente. Para asegurar el rigor metodológico y la validez de los resultados de la presente tesis doctoral, es necesario considerar explícitamente las amenazas a la validez que operan sobre los métodos cuantitativos y cualitativos, lo cual realizaremos a continuación.

7.1. Validez interna

Las amenazas a la validez interna son aspectos que ponen en peligro la capacidad del investigador al momento que desea extraer inferencias correctas sobre los datos de una población [Creswell, 2002]. Las medidas paliativas permiten rechazar interpretaciones alternativas que pueden provocar resultados no fiables. Las amenazas a la validez interna aparecen especialmente cuando se involucran participantes (p. ej., historia, maduración, regresión, selección y mortalidad) o procedimientos relacionados al uso de tratamientos donde el investigador puede manipular (p. ej., desmoralización compensatoria-resentida y la rivalidad compensatoria)[Campbell and Stanley, 1963].

En la Tabla 7.1 se encuentra un resumen de las amenazas de validez interna que han sido afrontadas en las tareas de cada etapa de investigación de esta tesis. Esta información se desarrolla a continuación:

La amenaza de **historia** produce que el tiempo sea el factor de amenaza, cuando ocurren eventos o acontecimientos que pueden influir en el resultado obtenido más allá del tratamiento aplicado. En nuestra tesis doctoral no opera este tipo de amenaza, debido a que las sesiones con los métodos utilizados (p. ej., encuestas, entrevistas) se realizan en tiempos distintos y por una sola vez. Por otra parte, las sesiones son de corto tiempo (p. ej., 20 a 30 min) de tal forma que sus resultados no se alteran en el tiempo (p. ej., cansancio por parte de los encuestados o entrevistados).

La amenaza de **maduración** aparece cuando los participantes en una sesión logran una madurez en el transcurso del tiempo, de tal forma que les permita alterar o influir en los resultados. En la presente tesis doctoral podría operar este tipo de amenaza, aunque muy levemente o de modo poco probable, debido a que el mismo sujeto puede haber participado de forma anónima en las encuestas y, además, haber sido seleccionado para las entrevistas (**Tareas 4 y 6**). No obstante, los temas tratados son un tanto diferentes entre métodos, por lo que no existe preocupación por este tipo de amenaza.

La amenaza de **regresión** aparece cuando los participantes son clasificados por grupos basados en casos de estudios previos; influyendo cuando participantes que poseen puntuaciones más altas son seleccionados para ser parte de algún procedimiento empírico. En nuestra tesis doctoral no opera esta amenaza, debido a que no hemos seleccionado o clasificado participantes por ningún modo. Sólo hemos seleccionado congresos y revistas de alto impacto, por el rigor de los resultados de sus publicaciones. Nuestro procedimiento no parece implicar este tipo de amenaza.

La amenaza de **selección** de participantes aparece cuando los seleccionados poseen características especiales que permiten obtener ciertos resultados (p. ej., son más brillantes). En nuestra tesis doctoral no opera esta amenaza, debido a que los participantes en las tareas de encuesta y entrevistas fueron voluntarios.

La amenaza de **mortalidad** aparece cuando los participantes abandonan o desertan durante un estudio empírico debido a diferentes razones propias del participante. Esta amenaza podría haber aparecido en el cuestionario (**Tarea 4**). Sin embargo, para minimizarla nos apoyamos en una muestra amplia de participantes, lo que nos permitió soportar posibles abandonos (que en el cuestionario son muy comunes). La mitigación resultó exitosa debido a que nuestro ratio de participación se encuentra en valores aceptables.

La amenaza de **difusión e imitación de tratamientos** aparece cuando los participantes de los grupos de control y experimental se comunican entre sí, influyendo en el aprendizaje acerca del tratamiento en los participantes del grupo de control. En nuestra tesis doctoral no opera esta amenaza, debido a que en el contexto de nuestro estudio no existen grupos de control y experimental.

La amenaza de **igualdad compensatoria/resentimiento/desmoralización** aparece cuando las compensaciones dadas a los participantes no son las mismas entre ellos, produciendo un resentimiento que puede afectar a los resultados. En nuestra tesis doctoral, esta amenaza no opera debido a que en el contexto del estudio todos los participantes son voluntarios, siendo el único ofrecimiento por nuestra parte la entrega del reporte de los resultados al final del estudio.

La amenaza de **rivalidad compensatoria** ocurre cuando los participantes en el grupo de control sienten que están siendo devaluados debido a que no participan en el estudio empírico. En consecuencia, los participantes en el grupo de control estarían tentados a reducir sus esfuerzos obteniendo de este modo un resultado no deseable. En nuestra tesis doctoral no opera esta amenaza, dado a que todos los participantes son considerados por igual en los procedimientos empíricos utilizados.

La amenaza de **testing** aparece cuando se aplica la misma prueba varias veces a los participantes, a tal punto que ellos pueden responder de forma distinta debido a que van conociendo cómo se conduce la prueba y las respuestas a la misma. En nuestra investigación, esta amenaza no opera ya que hemos aplicado

nuestros procedimientos empíricos una sola vez.

La amenaza de **instrumentación** aparece cuando existen cambios en el proceso de medición de la variable dependiente a lo largo del tiempo, debido a cambios en los instrumentos de recogida de datos (p. ej., formularios, documentos utilizados, etc.). Aparentemente, esta amenaza podría aparecer en las **tareas 3, 4, 6 y 7** debido a que son procedimientos empíricos donde se utilizan instrumentos de recogida de datos. Para mitigar esta amenaza, se utilizó un procedimiento previo de evaluación y piloteo con investigadores externos a fin de mejorar su diseño y evitar cambios o modificaciones posteriores.

7.2. Validez externa

Las amenazas de validez externa surgen cuando el investigador generaliza más allá del o de los grupos utilizados en el estudio empírico, hacia otros grupos que no se encuentran bajo estudio, o hacia contextos no estudiados, o a situaciones pasadas o futuras. Estas amenazas ocurren porque los investigadores realizan inferencias incorrectas a partir de las características de los individuos seleccionados para la muestra y la singularidad del entorno, entre otras. Para que un estudio tenga validez externa, es fundamental que la muestra de la población estudiada sea representativa y que se haya seleccionado a los participantes aleatoriamente.

En la Tabla 7.2, se encuentra un resumen de las amenazas de validez externa aplicadas en las tareas de cada etapa de investigación. La forma como fueron mitigadas se detalla a continuación:

La amenaza de **interacción entre selección y tratamiento** ocurre cuando los participantes de un estudio empírico poseen características muy particulares, por lo que el investigador no puede generalizar a características más generales. Esta amenaza también ocurre cuando el participante está involucrado en más de un tratamiento, a tal punto que su presencia en varios tratamientos no hace posible concluir si el efecto se debe a un tratamiento o a la interacción entre tratamientos. Esta amenaza podría aparecer en las **Tareas 4, 6 y 7** debido a las características propias de los experimentadores participantes en nuestra investigación. Sin embargo, nuestros participantes poseen una dilatada experiencia en el campo de la experimentación en IS, por lo que creemos que es posible generalizar los resultados a la población general de experimentadores en IS. Una consideración frente a la interacción de participantes y tratamientos, fue evitar los posibles efectos de confusión, especialmente en las encuestas, brindando la suficiente información en cada pregunta. En el caso de las entrevistas, se realizó en el idioma materno del entrevistado (español y portugués), para asegurar una mayor efectividad en las respuestas.

Tabla 7.1: Amenazas de validez interna

Validación Interna											
Objetivos	Tareas	Historia	Madurez	Regresión	Selección	Mortalidad	Difusión e Imitación	Igualdad Compensatoria	Rivalidad Compensatoria	Testing	Instrumen.
O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	Tarea 1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Tarea 2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
O2: Determinar la naturaleza de los errores estadísticos cometidos en IS.	Tarea 3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	x
	Tarea 4	N/A	x	N/A	N/A	x	N/A	N/A	N/A	N/A	x
O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	Tarea 5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Tarea 6	N/A	x	N/A	N/A	N/A	N/A	N/A	N/A	N/A	x
	Tarea 7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	x

La amenaza de **interacción del setting y el tratamiento** ocurre por la artificialidad de las situaciones empíricas debido a la configuración del tratamiento, por lo que el investigador no puede generalizar a los individuos en otros entornos, imposibilitando la generalización a situaciones reales. Esta amenaza podría operar en las **tareas 4, 6 y 7**. La amenaza es minimizada gracias a que el contexto se encuentra basada en opiniones, conocimientos o creencias de lo que realizan los participantes durante la práctica experimental; por lo que las respuestas obtenidas son más naturales y reales. De la misma forma, las preguntas de los instrumentos se encuentran basados en la práctica experimental real. Asimismo, el tiempo de las sesiones de encuestas y entrevistas no fue un obstáculo para obtener información, pues se dejó este factor a libertad de los participantes.

La amenaza de **interacción historia y tratamiento** aparece cuando el investigador necesita realizar estudios empíricos adicionales en nuevos escenarios con el fin de observar si ocurren los mismos resultados. En nuestra investigación, este tipo de amenazas no parecen operar. Las entrevistas y encuestas no dependen del momento específico del tiempo.

Tabla 7.2: Aspectos de validez externa

Objetivos	Etapa	Validación Externa		
		Interacción entre selección y tratamiento	Interacción del setting y el tratamiento	Interacción historia y tratamiento
O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	Tarea 1	N/A	N/A	N/A
	Tarea 2	N/A	N/A	N/A
O2: Determinar la naturaleza de los errores estadísticos cometidos en IS.	Tarea 3	N/A	N/A	N/A
	Tarea 4	x	x	x
O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	Tarea 5	N/A	N/A	N/A
	Tarea 6	x	x	x
	Tarea 7	x	x	x

7.3. Validez del constructo

Los problemas de validez del constructo ocurren cuando los investigadores utilizan definiciones y medidas inadecuadas de las variables, propiciando que las variables no representen correctamente los constructos teóricos. En la Tabla 7.3 se encuentra un resumen de las amenazas de validez del constructo relacionadas con las tareas cada etapa de la investigación. La forma en cómo fueron mitigadas se detalla a continuación:

La amenaza de **explicación pre-operacional inadecuada del constructo** aparece cuando el constructo (teoría) no está definido con la suficiente claridad, antes de convertirse en métrica o tratamiento. Esta amenaza podría aparecer en **tareas 4, 6 y 7**. Esta amenaza se mitigó gracias a la rigurosidad de la revisión

de literatura preliminar y de la revisión sistemática de literatura posterior, a fin de establecer la teoría correcta del dominio del problema. Asimismo, con la ayuda de aspectos de fiabilidad (ver sección 7.4) entre investigadores, se determinó la información que deberían proporcionar los sujetos participantes en la investigación durante las sesiones de encuesta y entrevista.

La amenaza debido al **empleo de operacionalizaciones únicas** aparece cuando el número de variables dependientes es muy bajo, evitando que los estudios empíricos identifiquen correctamente los fenómenos subyacentes. En esta investigación no estudiamos un único tipo de error estadístico, sino que consideramos cualquier tipo: omisiones, equivocaciones, conceptos erróneos, etc. Además de los errores, también estudiamos sus causas y consecuencias. Creemos que esta diversidad de aspectos suponen una operacionalización múltiple de la teoría. Sin embargo, no podemos asegurar que esta amenaza no este afectando a nuestros resultados.

La amenaza de **sesgo debido al empleo de un método operacionalizaciones único** aparece cuando se usa una sola métrica de medición, por lo que puede aumentar el riesgo de que los resultados no sean fiables y por lo tanto el estudio empírico sea engañoso. Esta amenaza fue mitigada mediante la utilización de diversos mecanismos de adquisición de datos, como son: surveys, encuestas y entrevistas.

La amenaza de **confusión entre constructo y niveles de constructo** aparece cuando en algunas relaciones empíricas no es relevante la presencia o ausencia del constructo, sino el nivel del constructo ensayado. Creemos que esta amenaza no opera, ya que no estamos usando un diseño experimental tratamiento vs placebo, sino que estamos estudiando un fenómeno concreto: *los errores estadísticos*. Adicionalmente, tal y como hemos indicado anteriormente, no estudiamos un error concreto, sino distintas modalidades (equivalentes a niveles) en que los errores pueden aparecer (omisiones, confusiones, etc).

La amenaza de **adivinación de la hipótesis** ocurre cuando un participante puede adivinar la hipótesis del experimento, mismo que puede alterar los resultados de modo que la hipótesis se vea favorecida o perjudicada. En nuestra investigación este tipo de amenazas no pueden operar, ya que no partimos de una hipótesis (o, si se quiere, la falsedad de la hipótesis nula es evidente). En este sentido, esta investigación corresponde a un estudio observacional, donde la hipótesis de trabajo tienen una importancia relativa.

La amenaza de **recelo de evaluación** aparece cuando la persona sabe que está siendo evaluada, y en consecuencia, puede cambiar su comportamiento. Esta amenaza podría aparecer en las **tareas 4, 6 y 7**. Esta amenaza se mitigó indicando a los participantes que las sesiones de encuesta y entrevista son confidenciales y/o anónimas. Al inicio de las sesiones, a los participantes también se les mencionó que las sesiones no pretenden evaluar al participante en modo

alguno, sino que pretenden obtener una percepción acerca de la práctica de la experimentación en IS. De este modo, esperamos que los participantes aporten sus respuestas con naturalidad y sin sesgos.

La amenaza de **expectativas del experimentador** ocurre cuando los resultados no son conformes a lo que el experimentador esperaba y, en consecuencia, éste podría (de forma involuntaria o no) sesgar los resultados del estudio empírico hacia el resultado esperado. Esta amenaza podría aparecer en la **tarea 6 y 7** durante la entrevista. Por ello, se mitigó esta amenaza con la participación de investigadores adicionales (directores e invitado) durante el análisis de los datos obtenidos. Asimismo, las entrevistas fueron piloteadas previamente para evitar que las preguntas a redactar pudieran ser sesgadas por el investigador.

7.4. Aspectos de fiabilidad

Como se mencionó en el capítulo 3, la tesis doctoral utiliza métodos cuantitativos y cualitativos. [Yin, 2013] y [Creswell, 2002] mencionan que al aplicar métodos cualitativos, es necesario establecer medidas que aseguren la fiabilidad para obtener resultados consistentes y confiables. En la Tabla 7.4 se encuentra un resumen de las medidas de fiabilidad aplicadas en las tareas de investigación de la presente tesis, así como la forma de cómo estas fueron aplicadas:

Una estrategia sugerida es la **documentación de procedimientos**. Según [Yin, 2013], los investigadores cualitativos necesitan documentar los procedimientos utilizados en sus estudios empíricos, tanto como sea posible. En nuestra investigación utilizamos esta estrategia para **todas las tareas (1-7)**. Prueba de ello son los artículos obtenidos de esta tesis donde se detallan los pasos y procedimientos utilizados en cada una de las tareas realizadas. Asimismo, existe una gran cantidad de información disponible en diferentes enlaces y páginas web referida a la recolección de datos, selección de artículos, etc. Los artículos de la tesis se encuentran disponibles en el apéndice web en el enlace: <https://goo.gl/T4gSva>.

La estrategia de **intercoder agreement (or cross-checking)** fue utilizada dentro de nuestra investigación en prácticamente **las tareas (2,3,5-7)**. La validación cruzada se aplicó específicamente en los análisis temáticos [Cruzes and Dybå, 2011] con el apoyo de varios investigadores (tesista, directores e investigador invitado), así como en las encuestas y revisiones de literatura. Como evidencia del agreement se realizó el cálculo de Fleiss' κ [Fleiss et al., 2013] para medir el nivel de acuerdo entre investigadores. En la mayoría de los casos hemos obtenido un "moderate agreement", lo que significa un nivel aceptable de fiabilidad.

La estrategia de **triangulación** permite que diferentes fuentes de información sean utilizadas para construir una justificación coherente del tema bajo estudio;

Tabla 7.3: Aspectos de validez constructo

Objetivos	Tareas	Validez del constructo						
		Explicación pre-operacional inadecuada del constructo	Empleo de operacionalizaciones únicas	Sesgo debido al empleo de un método de operacionalización único	Confusión entre constructo y niveles de constructo	Adición de la hipótesis	Recelo de evaluación	Expectativas del experimentador
O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	Tarea 1	N/A	-	x	N/A	N/A	N/A	N/A
	Tarea 2	N/A	-	x	N/A	N/A	N/A	N/A
O2: Determinar la naturaleza de los errores estadísticos cometidos en IS.	Tarea 3	N/A	-	x	N/A	N/A	N/A	N/A
	Tarea 4	x	-	x	N/A	N/A	x	N/A
O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	Tarea 5	N/A	-	x	N/A	N/A	N/A	N/A
	Tarea 6	x	-	x	N/A	N/A	x	x
	Tarea 7	x	-	x	N/A	N/A	x	x

permitiendo que los temas puedan ser sustentados a partir de la convergencia de varias fuentes de datos, perspectivas, teorías, etc. En esta tesis doctoral, existe una pequeña triangulación entre la información obtenida en las **tareas 1, 2, 3 y 4**, donde se realizan observaciones preliminares y posteriormente una observación formal del fenómeno bajo estudio. También podemos decir que podría existir una triangulación entre las **tareas 5, 6 y 7**, donde se establecen las causas de los errores estadísticos en las disciplinas, y se contrastan con las causas encontradas en la IS.

La estrategia de **comprobación de información entre investigadores** tiene como finalidad determinar la exactitud de los hallazgos cualitativos basados en descripciones o fragmentos específicos. La idea de la comprobación es determinar si los hallazgos son iguales entre investigadores. Si bien es cierto que la tesis doctoral fue realizada por el doctorando, durante toda la investigación existió comprobación de sus observaciones por parte de sus tutores en prácticamente *las tareas (2,3,5-7)*. El caso más resaltante fue en los análisis temáticos, donde se extrajeron fragmentos de texto por parte de varios investigadores, y posteriormente se evaluó la bondad de la extracción. En este procedimiento se contó con el soporte del **intercoder agreement (or cross-checking)** indicado anteriormente.

La estrategia de **descripción sustanciosa y gruesa** se refiere a la cantidad de información que se necesita incluir en el reporte para transmitir de forma correcta los hallazgos de la investigación. En todas las **tareas 1-7** se realizaron descripciones densas de los hallazgos encontrados. Prueba de ello son las discusiones en cada uno de los artículos resultado de la presente tesis doctoral.

La estrategia de **aclaración del sesgo que posiblemente el investigador aporta al estudio**, se refiere a indicar explícitamente en qué aspectos de la investigación esta el investigador especialmente interesado. Esta auto-reflexión crea una narrativa abierta y honesta que resuena bien con los lectores. Los artículos obtenidos de esta tesis doctoral contienen comentarios en la discusión donde se mencionan la interpretación de los hallazgos partiendo de trabajos relacionados o estudios propios anteriormente realizados, así como de opiniones personales del investigador.

La estrategia de presentar información **negativa o discrepante** que va en contra de las conclusiones de la propia investigación, se refiere a que en la investigación cualitativa pueden existir resultados desde diferentes perspectivas que no siempre son coincidentes, por lo que es necesario discutir la información contraria como una forma de aumentar la credibilidad de la investigación. En nuestra tesis doctoral, los resultados de las **tareas 1-7** son discutidas en sus respectivos reportes. Las discusiones son realizadas respecto a las evidencias encontradas, a pesar de que no sean del agrado de la comunidad o que pueden contradecir la opinión mayoritaria. Debido a ello, consideramos que nuestra investigación se vuelve más realista y, por tanto, válida.

La estrategia de **permanecer tiempo prolongado en el campo**, se refiere a que el doctorando debe permanecer cerca de las fuentes de información como una estrategia para aumentar la validez de los hallazgos. Esta estrategia fue acogida satisfactoriamente por el doctorando, quien realizó todos los estudios de campo, especialmente en las **tareas 3, 4, 5, 6 y 7**. De esta manera, el doctorando logró desarrollar una comprensión profunda del fenómeno estudiado, logrando transmitir de mejor manera el fenómeno de estudio. Por ello, creemos que la experiencia del doctorando, con los participantes en su entorno real, ayudó a que los hallazgos sean más precisos y válidos.

La estrategia de **interrogatorio por pares** consiste en localizar a investigadores (debriefers) que revisan y hacen preguntas acerca del estudio. Para nuestra tesis doctoral se utilizó como *debriefers* a los directores de la tesis (N. Juristo y O. Dieste) quienes revisaron y aportaron su interpretación a lo obtenido por el doctorando, añadiendo mayor validez a los hallazgos. Los debriefers aparecen en ciertas ocasiones como co-autor en los artículos de la tesis doctoral.

La estrategia de un **auditor externo**, para revisar toda o parte de la investigación es una muy buena opción cuando se requiere que los hallazgos sean especialmente fiables. A diferencia de un “peer reviewer”, el auditor no está familiarizado completamente con la investigación por lo que puede proporcionar una evaluación objetiva durante todo el proceso de investigación y al final del estudio. Esta estrategia fue muy bien acogida en la presente tesis doctoral. La figura de auditor externo fue utilizada de dos formas. En primer lugar, se presentaron los resultados de la investigación en ponencias de congresos de Ingeniería de Software Experimental, así como en un simposio doctoral con el fin de que investigadores externos proporcionen una evaluación o sugerencias acerca de la investigación. La segunda forma fue con el involucramiento un investigador externo (E. Fonseca), el cual examinó independientemente todos los aspectos de la investigación, en especial la revisión de los artículos de la tesis doctoral.

Finalmente, la estrategia de **uso de paquetes de software informático cualitativo** se utilizó frecuentemente esta tesis doctoral, especialmente en las **tareas 1, 4, 5, 6 y 7**. Hemos utilizado NVivo 11 con el fin de mantener y mejorar la consistencia de la codificación en los análisis temáticos.

Tabla 7.4: Aspectos de fiabilidad

Objetivos	Tareas	Aspectos de Fiabilidad										
		Documen- tación de procedi- mientos	Intercoder agree- ment	Triangulación	Compro- bación de infor- mación	Descripción sustan- ciosa y gruesa	Aclaración del inves- tigador al estudio	Presentar informa- ción negativa o discre- pante	Tiempo prolonga- do en el campo	Interrogatorio por pares	Auditor externo	Software informáti- co cualitati- vo
O1: Identificar si existen errores estadísticos en IS, y establecer su prevalencia.	Tarea 1	x	N/A	x	N/A	x	x	x	N/A	x	x	x
	Tarea 2	x	x	x	x	x	x	x	N/A	x	x	N/A
O2: Determinar la naturaleza de los errores estadísticos cometidos en IS.	Tarea 3	x	x	x	x	x	x	x	x	x	x	N/A
	Tarea 4	x	N/A	x	N/A	x	x	x	x	x	x	x
O3: Identificar las causas de los errores estadísticos en IS, y proponer recomendaciones para evitar su ocurrencia.	Tarea 5	x	x	x	x	x	x	x	x	x	x	x
	Tarea 6	x	x	x	x	x	x	x	x	x	x	x
	Tarea 7	x	x	x	x	x	x	x	x	x	x	x

8 | Conclusiones y futuras líneas de investigación

Resumen: Finalmente, enunciamos las conclusiones de la investigación, y las posibles futuras líneas de trabajo.

8.1. Conclusiones

En la presente Tesis Doctoral, hemos estudiado qué errores estadísticos ocurren en los trabajos experimentales en IS, y hemos determinado (por supuesto, de forma únicamente tentativa) razones para su aparición y alternativas de solución a los mismos.

Este tipo de investigación no es habitual en IS, aunque se realiza con frecuencia en otras ciencias, como por ejemplo la Educación [Brewer, 1985], Ciencias de la Salud [Welch and Gabbe, 1996, Altman, 1998, Olsen, 2003, Ercan et al., 2008] o Psicología [Bakker and Wicherts, 2011, Nickerson, 2000], por citar algunos ejemplos ya presentados anteriormente. En IS, la única investigación perfectamente encuadrada en esta línea de investigación es Jørgensen et al. [Jørgensen et al., 2016].

La falta de introspección de la IS, en lo referente al modo en que lleva a cabo sus investigaciones experimentales, es perfectamente entendible. La investigación experimental en IS es reciente, y los investigadores hemos estado más ocupados en conseguir resultados que en reflexionar cómo dichos resultados han sido conseguidos. No obstante, algunas reflexiones, aunque localizadas y parciales, ya empiezan a estar disponibles [Kitchenham et al., 2003, Dybå et al., 2006, Vegas et al., 2016]. Esperamos que la presente investigación contribuya a promocionar esta línea de trabajo.

Nuestra investigación ha demostrado, dentro de los límites de la misma, la existencia de errores estadísticos en IS. En sí mismo, este hallazgo no es sorprendente, ya que sería realmente inusual que la IS estuviera libre de la plaga de errores que afectan a otras disciplinas. Reflexionar sobre las causas de los errores es algo mucho más interesante.

Tomemos como partida el artículo de Jørgensen et al., antes citado. Este trabajo determina la existencia (que no dudamos en absoluto) de *publication bias* y *research bias* en IS. Ambos sesgos apuntan a la realización de prácticas cuestionables en IS como el p-hacking y HARKing¹. Sin embargo, nuestra investigación pone de manifiesto que las malas prácticas no son (no al menos principalmente) las causas de los errores estadísticos en IS. En su lugar, las causas más relevantes parecen ser:

- El insuficiente conocimiento, por parte de los investigadores, de la metodología experimental y técnicas estadísticas asociadas.

¹Aunque Jørgensen et al. [Jørgensen et al., 2016] no mencionan tal cosa, y se centran principalmente en el problema del bajo poder estadístico de los experimentos en IS.

- El insuficiente conocimiento, también por parte de los investigadores, de los fenómenos que ocurren en nuestra disciplina, lo que lleva a la frecuente realización de investigaciones de carácter exploratorio.

En otras palabras; al menos en el actual estadio de evolución de la experimentación en IS, parece que los errores estadísticos se producen por un problema de *inmadurez* de la propia disciplina. En consecuencia, las alternativas de solución van en la dirección de atacar dicha inmadurez, proporcionando a los investigadores:

- Formación adecuada.
- Guías claras de como realizar y analizar experimentos.

Ambas alternativas de solución parecen factibles, aunque no por ello su aplicación resulte más sencilla.

8.2. Líneas futuras de investigación

Los resultados de esta investigación han sido prometedores. Hemos logrado alcanzar todos los objetivos planteados. Sin embargo, como ya hemos indicado a lo largo de la Discusión (ver Capítulo 6) y Conclusiones (sección 8.1, inmediatamente anterior), ello no implica que todas las facetas del área de investigación hayan sido exploradas; al contrario, queda todavía mucho por hacer. Algunas futuras líneas de trabajo son las siguientes:

- Investigar otras técnicas o aspectos estadísticos incluidos en el catálogo de errores que hemos obtenido en esta Tesis doctoral. Un aspecto prometedor, por poner un ejemplo, sería determinar si los experimentadores conocen y comprueban adecuadamente las condiciones de aplicación de distintos test estadísticos.
- Utilizar otras metodologías de investigación. En la presente Tesis Doctoral, hemos basado nuestros hallazgos en revisiones de literatura e información obtenida de los experimentadores (usando encuestas y entrevistas). Sin embargo, otras aproximaciones son posibles. Por ejemplo, podría utilizarse etnografía, action research, o incluso experimentos controlados.
- Establecer mecanismos para implementar las recomendaciones de mejora planteadas por los experimentadores; por ejemplo, una línea interesante de trabajo consistiría en desarrollar guías claras para realizar ciertos estudios experimentales (paralelos, cross-over, etc.).
- Un aspecto que también creemos muy interesante es el establecimiento de mecanismos que permitan distinguir la investigación exploratoria y confir-

CAPÍTULO 8. CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

matoria. Un ejemplo paradigmático sería estudiar en qué medida establecer sistemas de pre-registro de experimentos en IS.

Finalmente, algunos de los experimentadores con los que hemos colaborado han insistido en que el investigador debe tener libertad, y no apresurarse en la realización de sus investigaciones. Estamos de acuerdo. Una futura línea de esta tesis doctoral, tan importante como la anterior, es explorar el impacto que las actuales políticas de publicación ejercen en los investigadores, y estudiar qué mejoras puede inducir una investigación más pausada y responsable.

Bibliografía

- [Altman, 1998] Altman, D. G. (1998). Statistical reviewing for medical journals. *Statistics in medicine*, 17(23):2661–2674.
- [Aranda et al., 2016] Aranda, A. M., Dieste, O., and Juristo, N. (2016). Effect of domain knowledge on elicitation effectiveness: an internally replicated controlled experiment. *IEEE Transactions on Software Engineering*, 42(5):427–451.
- [Bakker and Wicherts, 2011] Bakker, M. and Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3):666–678.
- [Basili, 1993] Basili, V. R. (1993). The experimental paradigm in software engineering. In *Experimental Software Engineering Issues: Critical Assessment and Future Directions*, pages 1–12. Springer.
- [Basili et al., 1986] Basili, V. R., Selby, R. W., and Hutchens, D. H. (1986). Experimentation in software engineering. *IEEE Transactions on Software Engineering*, SE-12(7):733–743.
- [Bezerra et al., 2015] Bezerra, R. M., da Silva, F. Q., Santana, A. M., Magalhaes, C. V., and Santos, R. E. (2015). Replication of empirical studies in software engineering: An update of a systematic mapping study. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–4. IEEE.
- [Bhatt, 2010] Bhatt, A. (2010). Evolution of clinical research: A history before and beyond james lind. *Perspectives in Clinical Research*, 1(1):6–10.
- [Box et al., 2005] Box, G., Hunter, J., and Hunter, W. (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley series in probability and statistics. Wiley-Interscience.
- [Box et al., 2008] Box, G. E., Hunter, J. S., and Hunter, W. G. (2008). *Estadística para Investigadores Diseño, innovación y descubrimiento*. Editorial Reverté, S. A., 2nd edition.
- [Brereton et al., 2007] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review

- process within the software engineering domain. *Journal of systems and software*, 80(4):571–583.
- [Brewer, 1985] Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational and Behavioral Statistics*, 10(3):252–268.
- [Campbell and Stanley, 1963] Campbell, D. T. and Stanley, J. C. (1963). Experimental designs for research on teaching. *Handbook of research on teaching*, pages 171–246.
- [Cho and Abe, 2013] Cho, H.-C. and Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, 66(9):1261–1266.
- [Cook et al., 1979] Cook, T. D., Campbell, D. T., and Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*, volume 351. Houghton Mifflin Boston.
- [Creswell, 2002] Creswell, J. (2002). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.
- [Creswell and Clark, 2007] Creswell, J. W. and Clark, V. L. P. (2007). *Designing and conducting mixed methods research*. Sage Publications.
- [Cruzes and Dybå, 2010] Cruzes, D. S. and Dybå, T. (2010). Synthesizing evidence in software engineering research. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, number 1, pages 1–10, New York, NY, USA. ACM.
- [Cruzes and Dybå, 2011] Cruzes, D. S. and Dybå, T. (2011). Recommended steps for thematic synthesis in software engineering. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 275–284. IEEE.
- [Darwish et al., 2017] Darwish, R., Gwosuta, L. N., and Torkar, R. (2017). A controlled experiment on coverage maximization of automated model-based software test cases in the automotive industry. In *IEEE International Conference on Software Testing, Verification and Validation (ICST)*, pages 546–547. IEEE.
- [de la Vara et al., 2014] de la Vara, J. L., Borg, M., Wnuk, K., and Moonen, L. (2014). Survey on safety evidence change impact analysis in practice: Detailed description and analysis.
- [Dybå et al., 2006] Dybå, T., Kampenes, V. B., and Sjøberg, D. I. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8):745–755.

- [Easterbrook et al., 2008] Easterbrook, S., Singer, J., Storey, M.-A., and Damian, D. (2008). Selecting empirical methods for software engineering research. In *Guide to advanced empirical software engineering*, pages 285–311. Springer.
- [Ercan et al., 2008] Ercan, I., Yang, Y., Özkaya, G., Cangur, S., Ediz, B., Kan, I., et al. (2008). Misusage of statistics in medical research.
- [Evans, 2003] Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of clinical nursing*, 12(1):77–84.
- [Fernández-Sáez et al., 2013] Fernández-Sáez, A. M., Chaudron, M. R., Genero, M., and Ramos, I. (2013). Are forward designed or reverse-engineered uml diagrams more helpful for code maintenance?: a controlled experiment. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 60–71. ACM.
- [Fleiss et al., 2013] Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.
- [Garfield, 1995] Garfield, J. (1995). How students learn statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 25–34.
- [Garfield and Ben-Zvi, 2007] Garfield, J. and Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3):372–396.
- [Genero et al., 2014] Genero, M., Cruz-Lemus, J. A., and Piattini, M. (2014). *Métodos de Investigación en Ingeniería del Software*. RA-MA Editorial.
- [Good and Hardin, 2012] Good, P. I. and Hardin, J. W. (2012). *Common errors in statistics (and how to avoid them)*. John Wiley & Sons.
- [Gore et al., 1977] Gore, S. M., Jones, I. G., and Rytter, E. C. (1977). Misuse of statistical methods: critical assessment of articles in *bmj* from january to march 1976. *BMJ*, 1(6053):85–87.
- [Haque and George, 2007] Haque, M. S. and George, S. (2007). Use of statistics in the psychiatric bulletin: author guidelines. *The Psychiatrist*, 31(7):265–267.
- [Huck, 2009] Huck, S. W. (2009). *Statistical misconceptions*. Routledge.
- [Ioannidis, 2005] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.
- [Jørgensen et al., 2016] Jørgensen, M., Dybå, T., Liestøl, K., and Sjøberg, D. I. (2016). Incorrect results in software engineering experiments: How to improve research practices. *Journal of Systems and Software*, 116:133–145.
- [Juristo and Moreno, 2001] Juristo, N. and Moreno, A. M. (2001). *Basics of Software Engineering Experimentation*. Kluwer Academic Publishers.

- [Kilkenny et al., 2009] Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F., Cuthill, I. C., Fry, D., Hutton, J., and Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one*, 4(11):e7824.
- [Kitchenham et al., 2003] Kitchenham, B., Fry, J., and Linkman, S. (2003). The case against cross-over designs in software engineering. In *International Workshop on Software Technology and Engineering Practice*, pages 65–67. IEEE.
- [Kitchenham et al., 2002] Kitchenham, B., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., Rosenberg, J., et al. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734.
- [Kitchenham et al., 2010] Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., and Linkman, S. (2010). Systematic literature reviews in software engineering—a tertiary study. *Information and Software Technology*, 52(8):792–805.
- [Madeyski and Kawalerowicz, 2017] Madeyski, L. and Kawalerowicz, M. (2017). Software engineering needs agile experimentation: A new practice and supporting tool. In *Software Engineering: Challenges and Solutions*, pages 149–162. Springer.
- [Maier and Berry, 2017] Maier, A. and Berry, D. M. (2017). Improving the identification of hedonic quality in user requirements - a controlled experiment. In *International Requirements Engineering Conference (RE)*, pages 213–222. IEEE.
- [Mäntylä et al., 2014] Mäntylä, M. V., Petersen, K., Lehtinen, T. O., and Lassenius, C. (2014). Time pressure: a controlled experiment of test case development and requirements review. In *International Conference on Software Engineering (ICSE)*, pages 83–94. ACM.
- [Midi et al., 2012] Midi, H., Imon, A., and Jaafar, A. (2012). The misconceptions of some statistical techniques in research. *Jurnal Teknologi*, 47(1):21–36.
- [Miller, 1999] Miller, J. (1999). Can results from software engineering experiments be safely combined? In *Software Metrics Symposium (METRICS)*, pages 152–158. IEEE.
- [Montgomery, 2010] Montgomery, D. (2010). *Applied Statistics and Probability for Engineers 5th Edition IS Version with WileyPLUS Set*. Wiley Plus Products Series. John Wiley & Sons, Incorporated.
- [Montgomery et al., 2006] Montgomery, D. C., Bermúdez Rojas, T., Walker, H., Maisel, L., Arley, N. B., Maddala, G., Collazos, A., Forsythe, E., Boonpragob, K. R., Kay, L., et al. (2006). *Probabilidad y estadística aplicadas a la ingenie-*

Applied statistics and probability for engineers. Number 519.5 M787. CATIE, Turrialba (Costa Rica).

- [Nickerson, 2000] Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241.
- [Nuzzo et al., 2014] Nuzzo, R. et al. (2014). Statistical errors. *Nature*, 506(7487):150–152.
- [Olsen, 2003] Olsen, C. H. (2003). Review of the use of statistics in infection and immunity. *Infection and immunity*, 71(12):6689–6692.
- [Pilcher and Bedford, 2011] Pilcher, J. and Bedford, L. A. (2011). Hierarchies of evidence in education. *The Journal of Continuing Education in Nursing*, 42(8):371–377.
- [Runeson et al., 2012] Runeson, P., Host, M., Rainer, A., and Regnell, B. (2012). *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons.
- [Salman et al., 2015] Salman, I., Misirli, A. T., and Juristo, N. (2015). Are students representatives of professionals in software engineering experiments? In *International Conference on Software Engineering (ICSE)*, Volume 1, pages 666–676. IEEE Press.
- [Sjøberg et al., 2005] Sjøberg, D. I., Han Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N.-K., and Rekdal, A. C. (2005). A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9):733 – 753.
- [Smith III et al., 1994] Smith III, J. P., Disessa, A. A., and Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The journal of the learning sciences*, 3(2):115–163.
- [Vegas et al., 2016] Vegas, S., Apa, C., and Juristo, N. (2016). Crossover designs in software engineering experiments: Benefits and perils. *IEEE Transactions on Software Engineering*, 42(2):120–135.
- [Vickers, 2010] Vickers, A. (2010). *What is a P-value anyway?: 34 stories to help you actually understand statistics*. Addison-Wesley Longman.
- [Wang et al., 2017] Wang, S., Ali, S., Gotlieb, A., and Liaaen, M. (2017). Automated product line test case selection: industrial case study and controlled experiment. *Software & Systems Modeling*, 16(2):417–441.
- [Welch and Gabbe, 1996] Welch, G. E. and Gabbe, S. G. (1996). Review of statistics usage in the american journal of obstetrics and gynecology. *American journal of obstetrics and gynecology*, 175(5):1138–1141.

- [Wheeler, 2014] Wheeler, J. (2014). Quick and dirty research: Opportunities for people who are too busy to do research to do research. *International Journal of Solution-Focused Practices*, 2(1).
- [Wohlin et al., 2012] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.
- [Yin, 2013] Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.