TESIS DE MAGISTER
EN INGENIERIA DEL SOFTWARE

REVISIÓN DE RESULTADOS EXPERIMENTALES EN TÉCNICAS DE PRUEBA Y DE EDUCCIÓN DE CONOCIMIENTOS.

Autor: Ing. Mario Luis Guerini

Directores: M.Ing. Enrique Fernández

Dr. Ramón García-Martínez

Buenos Aires, 2007

TABLA DE CONTENIDOS

CAPÍTULO 1

1. INTRODUCCIÓN

Este capítulo incluye consideraciones preliminares sobre la problemática a tratar (sección 1.1) y una descripción del trabajo de tesis (sección 1.2)

1.1 CONSIDERACIONES PRELIMINARES

En Informática, de acuerdo a la norma 610.12 de IEEE, se debe aplicar conocimiento científico para el desarrollo, operación y mantenimiento de sistemas software. Para ello cuenta con métodos, técnicas y herramientas para ser utilizadas en cada actividad de acuerdo a las condiciones que se disponga. Sin embargo, en la actualidad generalmente no se cuenta con técnicas ni métodos que cuenten con una justificación científica ni un "estudio objetivo de su efectividad" [Juristo y Moreno, 2001]. Por lo tanto, es necesario un marco que permita a los ingenieros poder conocer cuales son los mejores métodos y herramientas que se deben aplicar a través de un método científico y por lo tanto objetivo. Este marco es la investigación experimental, utilizada también en otros ámbitos para brindar información objetiva sobre hipótesis que se desean probar. De esta forma, como afirma [Pfleeger, 1999], se "permitirá ganar más entendimiento de que hace un software bueno y como hacerlo mejor".

El presente trabajo tiene como objetivo evaluar la calidad de los estudios experimentales vigentes desarrollados dentro del ámbito de la Ingeniería del Software.

Como el proceso de identificación y validación de estudios es largo y requiere de un gran esfuerzo en cuanto a la recolección y validación de estudios, el presenta trabajo

se ha desarrollado sobre la base de dos revisiones sistemáticas desarrolladas previamente:

- La primera [Davis, A., 2006] desarrollada sobre estudios experimentales vinculados a la educción de conocimientos, aporta información vinculada a como son los estudios experimentales hechos en contextos en los cuales el factor humano **tiene** una alta incidencia sobre los resultados.
- La segunda [Juristo, Moreno y Vegas, 2003] fue desarrollada sobre experimentos vinculados a técnicas que permiten la generación automática de casos de prueba, aporta información vinculada a como son los estudios experimentales hechos en contextos en los cuales el factor humano **no tiene** una alta incidencia sobre los resultados.

De esta forma se logró acceder a un conjunto de estudios experimentales de buena calidad dentro de un marco de tiempo acotado.

Como resultado de esta evaluación se pretende identificar:

- Puntos débiles y fuertes del actual contexto experimental de la IS (calidad de los experimentos y de los reportes)
- Los valores típicos de los distintos parámetros estadísticos (medias, varianzas y sujetos experimentales).

En base a este estudio, se podrá determinar, por un lado, cuan fiables son los resultados experimentales en IS, y por otro, que aspectos deberían tenerse en cuenta a la hora de combinar los resultados dentro de un proceso de agregación.

1.2 DESCRIPCIÓN DE LA COMPOSICIÓN DEL TRABAJO DE TESIS

En el capítulo 1, el presente, se introducen los temas tratados en esta tesis describiéndose brevemente su objetivo y composición

En el capítulo 2, se describen las distintas técnicas de educción de conocimientos y prueba analizadas en los estudios en cuestión. Como así también se hace un repaso de las técnicas estadísticas aplicadas en el presente informe y en los estudios analizados.

En el capítulo 3, se define el problema tratado en el informe.

En el capítulo 4, se presenta los materiales que se utilizan en el experimento, como así también una descripción del método que se utiliza para analizar los estudios.

El capítulo 5, contiene la experimentación realizada. El mismo consta de dos partes. En la primera se analizan siete informes sobre técnicas de prueba y luego se vuelcan los resultados en una tabla resumen. En la segunda parte, se resumen los resultados de todos los experimentos, destacando la cantidad de información que cada uno de ellos contiene.

El capítulo 6, contiene las conclusiones alcanzadas con el desarrollo del trabajo.

El capítulo 7, contiene referencias a la bibliografía analizada en el informe como también la bibliografía consultada.

El anexo, contiene una descripción de las técnicas estadísticas usadas en el informe.

CAPÍTULO 2

2. ESTADO DE LA CUESTIÓN

En ingeniería del software, se dispone de una amplia variedad de técnicas para la educción de conocimientos [Juristo, INCO] como así también de técnicas de prueba [Myers, 1979; Gao et al, 2003]. Pero a pesar de que existan numerosos estudios experimentales, al momento de seleccionar una técnica para llevar a cabo un proyecto, no se cuenta con información empírica acerca de cual es la técnica más eficaz o eficiente para ser aplicada en cada caso en particular. Esto tiene que ver con la escasa tradición de experimentar y consultar estudios experimentales dentro del contexto de la IS [Juristo y Moreno, 2001], pero antes de utilizar los conocimientos experimentales disponibles es importante conocer cual es la calidad de los mismos. En tal sentido el presente trabajo buscará determinar cual es el nivel de calidad de un conjunto de estudios empíricos considerados, a priori, bien desarrollados, ya que los mismos fueron seleccionados dentro del marco de una Revisión Sistemática.

2.1 TÉCNICAS DE EDUCCIÓN Y PRUEBA

A continuación se describen las técnicas de educción de conocimientos y técnicas de prueba mas difundidas y sobre la cuales se han identificado estudios experimentales a evaluar.

2.1.1 Técnicas de educción de conocimientos

Hay dos clases de técnicas para la educción de conocimientos [Burton, et al]:

- Las denominadas técnicas directas, en estas técnicas se le pregunta directamente al experto los conocimientos que se desea adquirir, de ahí el

nombre, es decir, en ellos el experto reporta los conocimientos que él puede articular directamente.

- Las denominadas técnicas indirectas, estas técnicas se usan porque no siempre los expertos pueden acceder a los detalles de sus conocimientos o procesos mentales. Es posible que los expertos perciban relaciones complejas o alcancen conclusiones perfectas, sin saber exactamente cómo lo hicieron. En estos casos, es necesario usar métodos indirectos de educción de conocimientos, en los cuales no se le pregunta directamente al experto por lo que sabe. Al contrario, se le proporcionan otras tareas, por ejemplo, el grado de similitud de dos objetos, o la contemplación de unos objetos varias veces desde distintos puntos de vista, etc., y, a partir de estos resultados, el ingeniero de conocimiento (IC) infiere los conocimientos subyacentes a la resolución del problema planteado.

2.1.1.1 Métodos de Educción DIRECTOS

2.1.1.1.1 Entrevista

La entrevista con el experto es el método más común y familiar para educir conocimientos. La entrevista consiste en una interacción sistemática de un IC con un experto para extraer los conocimientos de experiencia de éste. Al conversar con el experto, se revelan sus objetivos al resolver problemas, cómo están relacionados u organizados sus pensamientos, y los procesos a través de los cuales hace un juicio, resuelve un problema o diseña una solución.

Las entrevistas pueden ser abiertas o estructuradas. En una entrevista no estructurada, o abierta, el IC pregunta, más o menos espontáneamente, cuestiones al experto. En las entrevistas estructuradas, el IC, una vez marcado el tema y la profundidad con que se desea tratarlo, planifica con anterioridad todas las preguntas que debe plantear al experto durante la sesión [Juristo, INCO].

2.1.1.1.2 Cuestionarios

La técnica de cuestionarios consiste en realizar una entrevista estructurada al experto de forma indirecta, a través de cuestionarios. [Juristo, INCO]

2.1.1.1.3 Educción directa de atributos

El proceso de educción es muy simple y consiste en preguntar directamente a los expertos cuáles son los atributos que deben ser tenidos en cuenta en la tarea en cuestión. La educción directa se usa como punto de partida para la educción de atributos. [Juristo, INCO]

2.1.1.2 Métodos de Educción INDIRECTOS

2.1.1.2.1 Observación de tareas habituales

Esta técnica consiste en observar a un experto trabajar en un problema real habitual. En la observación de tareas habituales, el IC no interfiere la actuación del experto en la solución de sus tareas reales cotidianas y solamente registra las mismas para luego analizarlas [Juristo, INCO].

2.1.1.2.2 Incidentes críticos

En la técnica de incidentes críticos se le pide al experto que describa casos especialmente interesantes o difíciles que se le hayan presentado. El experto, además, deberá contar cómo los resolvió. Es decir, el experto expondrá lo que recuerda sobre cómo solucionó tales casos críticos. Esta forma presenta la ventaja de que el experto puede olvidar detalles esenciales a la hora de resolver casos normales, pero los casos especialmente complejos le estimulan de modo que le hacen comentar detalles que, en otro momento, hubiera pasado por alto [Juristo, INCO].

2.1.1.2.3 Clasificación de conceptos

Esta técnica consiste en obtener, a partir de un simple glosario o texto, un conjunto de conceptos que cubran ampliamente el dominio. A continuación,

se transfiere cada concepto a una ficha y se le pide al experto que las clasifique en una serie de grupos, describiendo lo que cada grupo tiene en común. A continuación, los grupos pueden compararse para formar jerarquías [Juristo, INCO].

2.1.1.2.4 Análisis de Protocolos

Un tipo de técnica de educción similar a la observación de tareas habituales es el análisis del protocolo (AP). La diferencia de este método con las tareas habituales es que en el AP no hay un intervalo entre el acto de pensar del experto, y el acto de reportarlo. En ambas técnicas existe alguien, en este caso un experto, o usuario, que está desarrollando su comportamiento para efectuar una tarea normal con un problema específico. En el AP, además de registrar las sesiones y anotar el comportamiento después de realizado, se le pide al experto que piense en voz alta mientras efectúa la tarea. Con el AP, se va más lejos que con la simple observación del comportamiento del experto en su trabajo habitual: toma de notas, lectura de medidas, búsqueda de informaciones, etc. Se busca capturar, y después estudiar, todo lo que dice el experto en el momento en que trata un problema [Juristo, INCO].

2.1.1.2.5 Emparrillado

Esta técnica se basa en la idea de que cada persona tiene su propio modelo o visión del mundo que le rodea. El emparrillado incluye un diálogo inicial con el experto, una sesión de valoración, y un análisis de los grupos, de los conceptos y de las dimensiones sobre las cuales fueron valorados los elementos. Esencialmente, es una sesión de valoración y recuerdo de forma libre, en la cual el IC efectúa inferencias acerca de las relaciones entre los conceptos y la calidad de las relaciones en las dimensiones a las que el experto presta atención [Juristo, INCO].

2.1.1.2.6 Educción de atributos por rangos

En la educción por rangos, los expertos responden acerca de la prioridad de los atributos en cuestión y las razones por el orden elegido. La elección de los atributos utilizando alguna otra técnica cualitativa de selección [Bech-Larsen et al, 1999]

2.1.1.2.7 Descripción ideal de atributos

En esta técnica a los expertos se les pide que den la descripción ideal del objeto de análisis para cada categoría posible. Esta aproximación es consistente con el concepto de ideales de categorías [Barsalou, 1985]. En esta perspectiva de clasificación, el mejor ejemplo de una categoría de un determinado producto puede ser el ideal, en lugar del producto promedio [Breivik y Supphellen, 2003].

2.1.2. Técnicas de prueba de software.

Las técnicas de pruebas de software se pueden agrupar en las siguientes categorías:

- Técnicas Aleatorias: Los casos de prueba se generan aleatoriamente.
- Técnicas Funcionales: Se utilizan las especificaciones del problema para generar los casos de prueba. El programa se ve como una caja negra.
- Técnicas de Flujo de Control: Se requiere conocimiento del código fuente. Se seleccionan caminos dentro del programa que deben ser ejecutados al ingresar los casos de prueba.
- Técnicas de Flujo de Datos: También requiere conocimiento del código fuente. En este caso, los caminos se eligen de forma de explorar secuencias de eventos relacionadas con el estado de las variables.
- Técnicas de Mutación: En la mutación se introducen fallas al programa creando varios mutantes, cada uno con una falla. Los casos de prueba se hacen pasar por los mutantes con la intención de hacer fallar al programa. Cuando ocurre la falla de un mutante, se dice que éste ha sido matado y no se prueba más ese mutante muerto. Asimismo, el caso de prueba que hizo fallar al mutante, se marca como un caso de prueba útil para la detección de fallas. Al

cabo de ejecutar todos los casos de prueba, los mutantes pueden sobrevivir debido a dos razones: una es que sean equivalentes al programa inicial, y otra es que no haya habido un caso de prueba lo suficientemente bueno como para detectar el fallo. En este caso, se deben generar más casos de prueba para poder matar todos los mutantes no equivalentes. El score de mutación es el porcentaje de mutantes no equivalentes que mata un conjunto de prueba, cuyo valor ideal es el 100%

Para una descripción más detallada de las técnicas se puede consultar [Beizer, 1990; Myers, 1979]. Los estudios sobre estas técnicas se pueden realizar de dos formas: una es intra familias y la otra es entre familias. En el primer caso se comparan técnicas pertenecientes a la misma categoría, mientras que en el segundo, la comparación se hace entre categorías distintas. En las siguientes secciones se describen las técnicas de prueba sobre las que se estudian los experimentos.

2.1.2.1 Técnicas de Flujo de Control

De esta familia se analizará solamente el criterio de Branch Testing. Este criterio establece que los casos de prueba deben hacer que se ejecuten todas las sentencias del programa al menos una vez, o dicho de otro modo, que todas las decisiones se lleguen a evaluar para el caso en que den verdaderas al menos una vez y lo mismo para el caso en que den falsas.

2.1.2.2 Técnicas de Flujo de Datos

Sean s_i y s_j; $1 \leq i$; $j \leq n$ dos sentencias en un programa (P) donde la variable x es definida y usada respectivamente. Nos referimos a esta definición como $d_i(x)$, y al uso como $u_j(x)$. El par $(d_i(x); u_j(x))$ es llamado *par du*. Un *par du* puede ser *p-use* o *c-use* dependiendo de si s_j es un predicado o no, respectivamente. Este par es *factible* si existe un caso de prueba t en el dominio (D) de todos los casos de pruebas posibles, tal que la ejecución de P en t cause que el control del programa vaya de s_i a s_j, pasando por una o más sentencias que no definan a x. Un camino de este tipo es llamado un *camino libre de definición* con

respecto a x. Un caso de prueba t perteneciente a un conjunto de pruebas T cubre un par c-use ($d_i(x)$; $u_j(x)$) si la ejecución de P en t causa la ejecución de un camino libre de definición con respecto a x de s_i a s_j. Un par p-use ($d_i(x)$; $u_j(x)$) es considerado cubierto si un camino libre de definición con respecto a x desde s_i a s_j a s_k es ejecutado por cada sucesor s_k de s_j. Nos referimos al conjunto de todos los pares c-use y p-use como *all-uses*. Un conjunto de prueba T puede ser evaluado contra el criterio *all-uses* computando la razón entre el número total de *all-uses* cubierto contra el número factible de *all-uses*. Una razón de 1 implica que T es adecuado con respecto al criterio *all-uses*.

2.1.2.3 Técnicas de Mutación

Las técnicas de mutación se basan en la modelización de las faltas típicas que se comente al hacer un programa, mediante lo que se conocen como operadores de mutación (dependientes del lenguaje de programación). Cada operador de mutación se aplica sobre el programa, dando lugar a una serie de mutantes (programas exactamente igual al base, pero con una sentencia modificada, precisamente la originada por el operador de mutación). Una vez que se tiene generado el conjunto de mutantes, se generan casos de prueba que ejerciten la parte mutada del mismo. Tras generar casos de prueba para cubrir todos los mutantes, teóricamente se tienen cubiertas todas las posibles faltas cometidas (en la práctica, sólo las faltas modelizadas por los operadores de mutación).

En este artículo se examinarán los cinco criterios de generación de casos de pruebas descriptos en la siguiente tabla:

TÉCNICA	CRITERIO DE GENERACIÓN DE MUTANTES
Mutación (Standard o fuerte)	Se seleccionaron todos los operadores usados por Monthra [MONTHRA, 1987] excepto el operador GOTO debido a que no era utilizado en los programas de prueba escritos en el lenguaje C.

Tabla 2.1. Técnicas de mutación

TÉCNICA	CRITERIO DE GENERACIÓN DE MUTANTES
Mutación abs/ror	Solamente utiliza los operadores abs y ror para generar los mutantes. El operador abs reemplaza el valor de cada variable *x* por abs(*x*), -abs(*x*) y zpush(*x*). El operador zpush hace que el mutante muera inmediatamente si su argumento es cero, lo que requiere que los datos de prueba fuercen a que toda expresión adquiera el valor cero. Cuando se aplica a un programa que contiene una asignación z := x+1, el operador abs genera seis mutantes, obtenidos de reemplazar la asignación por z := abs(x)+ 1, z := -abs(x)+ 1, z := zpush(x)+ 1, z := abs(x + 1), z := -abs(x + 1), y z := zpush(x + 1). El operador ror genera mutantes reemplazando cada operador relacional por otros operadores relacionales. Por ejemplo, cuando se aplica a un programa que contiene un predicado "if (x = 0) then", el operador genera las siguientes siete condiciones: "if (x < 0) then", "if (x ≤ 0) then", "if (x ≠ 0) then", "if (x > 0) then", "if (x ≥ 0) then", "if (true) then", y "if (false) then"
Mutación 10%	En este caso, se seleccionan aleatoriamente el diez por ciento de los mutantes generados para cada tipo de mutación.
Mutación selectiva	En la mutación selectiva se descartan los mutantes que fueron generados con los operadores de mutación que generan más mutantes. La mutación selectiva N-selective descarta los mutantes generados con los N operadores más populosos, siendo N un número natural.
Mutación débil	En la mutación débil, los mutantes son evaluados antes de finalizar la ejecución del programa mutante. Es decir que la comparación entre el programa original y el mutante se realiza en un estado intermedio del mismo, lo que permite que se reduzcan los tiempos de prueba.

Tabla 2.1. Técnicas de mutación (Cont.)

2.2 TÉCNICAS ESTADÍSTICAS

Luego de obtener datos experimentales es deseable obtener conclusiones a partir de los mismos, para lo que es necesario interpretar estos datos. La interpretación cuantitativa de los datos se puede llevar a cabo en tres pasos [Wohlin et al., 2000]:

1. Estadística descriptiva
2. Reducción de conjunto de datos
3. Tests de hipótesis

En el primer paso, los datos son interpretados usando estadística descriptiva, que permite obtener la tendencia, dispersión y dependencia. En el segundo paso, se eliminan las mediciones anormales para obtener un conjunto de mediciones válidas. En el tercer paso, los datos se analizan usando tests de hipótesis, donde las hipótesis del experimento son evaluadas estadísticamente con un cierto nivel de confianza.

2.2.1 Estadística descriptiva

La estadística descriptiva trata con la presentación y el procesamiento de los datos numéricos. Luego de juntar los datos experimentales, la estadística descriptiva se usa para describir y presentar gráficamente ciertos aspectos del conjunto de mediciones obtenidas. El objetivo es obtener una sensación de cómo los datos están distribuidos para comprender la naturaleza de los datos e identificar mediciones anormales.
Los indicadores se pueden agrupar en:

- Indicadores de tendencia, como la media, la mediana, la moda y la media geométrica.
- Indicadores de dispersión, como la varianza, la desviación estándar, el intervalo de variación y el coeficiente de variación.
- Indicadores de dependencia, como la regresión lineal o de otro tipo, la covarianza y el coeficiente de correlación

2.2.2 Reducción de conjunto de datos

En la próxima sección se enumeran varios métodos estadísticos. Lo que todos los métodos tienen en común, es que los resultados que se obtienen por usarlos dependen

fuertemente de la calidad de los datos de entrada. Es por eso, que si los datos ingresados no se corresponden con lo que deben representar, las conclusiones que se obtendrán no serán válidas.

En el caso de mediciones donde el factor humano no tiene una alta incidencia, estos ocurren principalmente como errores sistemáticos o como mediciones atípicas. Cuando el factor humano tiene incidencia, los datos anormales se pueden deber a que los participantes no tomaron en serio el experimento.

Estos valores a excluir del conjunto de prueba deben ser analizados para encontrar la causa de su desviación de forma de asegurarse que no deberían repetirse y que no son parte del fenómeno bajo estudio. En algunos casos estos valores suelen agregar una nueva variable para considerar en el modelo.

2.2.3 Tests de hipótesis

El objetivo del test de hipótesis es verificar si es posible rechazar una determinada hipótesis, basándose en una muestra de alguna distribución estadística.

Las pruebas pueden ser clasificadas en paramétricas y no paramétricas. Las pruebas paramétricas están basadas en un modelo que involucra una distribución específica. Además los parámetros deben ser mediciones de intervalo o razón; por lo que para mediciones nominales u ordinales, se deben usar pruebas no paramétricas.

Las pruebas no paramétricas tienen supuestos menos estrictos por lo que son más generales; es decir, que siempre se pueden usar en lugar de los paramétricos, pero la inversa no siempre es cierta.

Con respecto a la elección de qué tipo de prueba realizar, existen dos factores a considerar:

1. Aplicabilidad: Es importante que los supuestos sean válidos: tipo de medición adecuado y restricciones del modelo cumplidas.

2. Poder: El poder de las pruebas paramétricas es generalmente mayor que las no paramétricas. Es por eso que los tests paramétricos requieren menor cantidad de datos y, por lo tanto, experimentos menores.

Entre los tests paramétricos se encuentran el t-test, F-test y ANOVA (análisis de varianza). Mann-Whitney, Wilcoxon, Kruskal-Wallis y Chi-2 son ejemplos de tests no paramétricos.

Una descripción de las técnicas estadísticas utilizadas se encuentra en el Anexo.

CAPÍTULO 3

3. DEFINICIÓN DEL PROBLEMA

La experimentación de software trata de buscar reglas empíricas que proporcionen evidencias sobre las ventajas o desventajas de los distintos métodos, técnicas o herramientas empleadas en la construcción de sistemas software [Juristo y Moreno, 2001.]. Si no se realiza una revisión formal sobre los resultados de los experimentos se corre el riesgo de sostener conclusiones erróneas. Es posible también obtener resultados incorrectos si se estiman parámetros sobre distribuciones estadísticas que fueron asumidas erróneamente.

Se han realizado estudios empíricos sobre técnicas de educción de conocimientos y de prueba de software con el objeto de crear una base de conocimientos que permita seleccionar la técnica más adecuada para cada problema. Sin embargo, para que los estudios sean más que una mera descripción de un caso en particular, es necesario que el análisis de los datos de los estudios cumpla con ciertas condiciones que permitan proyectar la validez de sus conclusiones, como así también aplicar métodos de triangulación y meta-análisis.

En la actualidad muchos de los intentos por aplicar meta-análisis en Ingeniería del Software (IS) han fracasado, en general por problemas en las publicaciones en el armado de la presentación del artículo experimental. Por tal motivo se considera necesario evaluar en detalle la calidad de los estudios experimentales que hoy día se publican dentro del ámbito de la IS. De forma tal de poder establecer puntos débiles y, de existir, fuertes de los estudios realizados con vistas a su posterior validación y generalización.

CAPÍTULO 4

4. PLANTEO DE LA SOLUCIÓN

Con el fin asegurar la calidad de los conocimientos obtenidos en el presente trabajo, se va a seguir una estrategia de investigación clásica [Kumar, 1996; Creswell, 2003; Marczyk et al., 2005] identificando los materiales y métodos necesarios para desarrollar el proceso de investigación.

4.1 MATERIALES

Para el desarrollo del método de agregación utilizaremos dos tipos de materiales:

4.1.1 Estudios experimentales

Se cuenta con un conjunto de 37 estudios experimentales procedentes de sendas revisiones sistemáticas desarrolladas dentro del área de educción de requisitos y pruebas de software.

4.1.2 Técnicas estadísticas

Se cuenta con un grupo de técnicas estadísticas –paramétricas y no paramétricas- que permiten evaluar la validez de las conclusiones halladas. Estas técnicas se describen en el Capítulo 2 y en el Anexo de este informe.

4.2 MÉTODO

Para validar la calidad de los estudios y determinar cuales son los parámetros típicos de los estudios en IS, se realizará un proceso de verificación que se describe a continuación:

4.2.1 Revisión de experimentos

Se analizan los datos publicados por el estudio, con el objeto de corroborar si los mismos son factibles de ser analizados y evaluados. A continuación se describen los datos a extraer de cada experimento:

Estimador	(Sí/No) ¿Publica valor cuantitativo de alguna métrica?
Error en Estimador	(Sí/No) ¿Publica valor del error observado en el estimador?
Cantidad de sujetos	(Sí/No) ¿Publica la cantidad de sujetos sobre los que se realizó el estudio para obtener el estimador y su error?
Fuentes publicadas	(Sí/No) ¿Publica el código fuente sobre el que se realizó el estudio?
Datos sin procesar publicados	(Sí/No) ¿Publica los valores medidos, antes de realizar los cálculos estadísticos?
Técnicas estadísticas aplicadas	(Sí/No) ¿Aplica y expone técnicas estadísticas para hallar el estimador y su error?
Aporta Conclusiones	(Sí/No) ¿Llega a conclusiones que pueden ser generalizadas o simplemente publican los valores obtenidos?
Fundamenta conclusiones estadísticamente	(Sí/No) ¿Respalda las conclusiones obtenidas aplicando técnicas estadísticas?
Cantidad de Sujetos	Cantidad de sujetos utilizados en el estudio.
Variable estimada	Variable estimada en el estudio.
Media	Valor de la media del parámetro estimado
Desviación estándar	Desviación estándar hallada

4.2.2 Verificación de conclusiones

Se verifica si el informe llega a alguna conclusión, ya que en algunos casos los estudios son análisis de casos que no permiten una generalización de los resultados obtenidos. En el caso en que se publiquen las conclusiones, se verifica si las mismas están soportadas por técnicas estadísticas. Para los estudios sobre técnicas de prueba de software, en el caso en que no hayan sido aplicadas técnicas estadísticas, se aplican para verificar la validez de las conclusiones.

Una forma de revisión de los resultados puede formularse a partir de la utilización de métodos de análisis no paramétricos sobre las estadísticas que han permitido generar dichos resultados. Los métodos de análisis no paramétricos no requieren suposición alguna sobre el tipo de distribución al que responden las variables estudiadas [Ledesma 1980; Montgomery, 2002; García, 2004].

En algunos casos se validan las hipótesis de trabajo y los procedimientos usados para alcanzar las conclusiones halladas.

4.2.3 Síntesis de resultados en tabla resumen

Los resultados obtenidos luego del análisis de los estudios sobre técnicas de prueba se resumen en una tabla. En la misma se vuelca la información más destacada de cada estudio y el tipo de análisis estadístico realizado sobre el mismo.

4.2.4 Tabla resumen de parámetros

Con los datos recogidos durante el análisis de los experimentos, se determinarán los valores típicos para los parámetros estadísticos publicados.

4.2.5 Conclusiones

Finalmente en el Capítulo 6 se exponen las conclusiones halladas en el informe.

CAPÍTULO 5

5. EXPERIMENTACIÓN

5.1. REVISIÓN DE EXPERIMENTOS

Se analizaron 37 trabajos, de los cuales 9 se encuentran en el ámbito de las pruebas de software y los 28 restantes en el ámbito de la educción de conocimientos.

A continuación se describen los resultados obtenidos de la evaluación de los estudios:

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
1	Si	No	Si	No	No	Si	Si	No	10	Efectividad (mutación)	-	-	Ver Sección 5.2.1
2	Si	Si	Si	No	No	Si	Si	No	10	Efectividad (mutación)	-	-	Ver Sección 5.2.2
3	Si	No	Si	No	No	No	Si	No	29	Cant. casos de prueba	-	-	Ver Sección 5.2.3
4	Si	Si	Si	No	No	Si	Si	No	8	Efectividad	-	-	Ver Sección 5.2.4
5	No	No	Si	No	Si	No	No	No	143	Cant. casos de prueba	-	-	Ver Sección 5.2.5
6	Si	No	Si	No	No	No	Si	No	10	Efectividad (mutación)	-	-	Ver Sección 5.2.6
7	Si	No	Si	No	No	No	Si	No	11	Score mutación	-	-	Ver Sección 5.2.7
8	Si	Si	Si	No	No	Si	No	No	9	Efectividad	-	-	Halla medias y varianza para los datos intermedios pero no para las variables finales.
9	Si	No	Si	No	No	No	Si	No	10	Score mutación	-	-	
10	No	No	Si	N/A	No	No	Si	No	17	Cantidad atributos	-	-	Publica solamente cantidad de atributos educidos

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
11	Si	Si	Si	N/A	No	Si	Si	No	30	Cantidad atributos	-	-	Análisis ANOVA para ver diferencias entre grupos.
12	No	No	Si	No	No	No	Si	No	6	-	-	-	Sólo análisis cualitativo
13	Si	Si	Si	No	No	Si	No	No	8	Resp. a cuestionarios	-	-	Análisis ANOVA para ver diferencias entre grupos.
14	Si	Si	Si	No	No	Si	Si	Si	21	Eficiencia (CI)	10,81	4,65	Compara Cognitive Interview (CI) vs. Standard Interview (SI)
									21	Eficiencia (SI)	5,619	2,97	
									32	Efectividad (CI)	5,93	1,43	
									32	Efectividad (SI)	4,02	1,87	
15	No	No	Si	No	No	No	Si	No	10	-	-	-	Sólo análisis cualitativo
16	No	No	No	No	No	No	Si	No	-	-	-	-	Sólo análisis cualitativo
17	Si	Si	Si	No	No	Si	Si	Si	8	Cantidad conceptos	12,75	5,6	Análisis de mapa de conceptos y ANOVA para ver diferencias entre grupos
									8	Cantidad de enlaces	18,5	8,25	
									8	Complejidad	6,75	3,99	
18	Si	Si	Si	No	No	Si	Si	No	32	Tiempo y cláusulas educidas	-	-	Análisis ANOVA para ver diferencias entre grupos.
19	No	No	Si	No	No	No	No	No	1	Reglas obtenidas	-	-	Solamente caso de estudio
20	Si	No	Si	No	No	Si	No	No	4	Tiempo y cláusulas educidas	-	-	
21	Si	Si	Si	No	No	Si	Si	Si	25	Cant. de ordenamientos	-	-	Análisis ANOVA para ver diferencias entre grupos.
22	Si	Si	Si	No	No	Si	Si	Si	15	Cant. Req. (1) Total	30,8	16,3	Estudia tres técnicas de educción de requerimientos: Sintáctica (1), semántica (2) y
									15	Cant. Req. (2) Total	40,93	12,1	
									15	Cant. Req. (3) Total	66,27	20,7	
									15	Cant. Req. (1) Metas	8,47	5,15	

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
									15	Cant. Req. (2) Metas	8,4	3,52	característica s de la tarea (3). Analiza la cantidad de requisitos educidos para cuatro clases distintas de requerimiento s: Metas, Procesos, Tareas e Información.
									15	Cant. Req. (3) Metas	6,67	5,71	
									15	Cant. Req. (1) Proc.	19,87	8,41	
									15	Cant. Req. (2) Proc.	26,27	9,31	
									15	Cant. Req. (3) Proc.	40,4	10.51	
									15	Cant. Req. (1) Tareas	1,73	2,14	
									15	Cant. Req. (2) Tareas	3,07	2,52	
									15	Cant. Req. (3) Tareas	4,6	3,81	
									15	Cant. Req. (1) Inform.	1,4	2,52	
									15	Cant. Req. (2) Inform.	3,13	2,45	
									15	Cant. Req. (3) Inform.	14,6	6,56	
23	Si	Si	Si	No	No	Si	Si	Si	10	Cant. Reglas (N/P)	5,2	-	Estudia la adquisición de conocimiento s para técnicas estructuradas (E) y no estructuradas (N) de entrevistas. Utiliza personas expertas (X) y principiantes (P) como fuente de conocimiento s
									10	Cant. Reglas (E/P)	9,9	-	
									10	Cant. Reglas (N/X)	6,1	-	
									10	Criterio/con tenido (N/P)	2,98	-	
									10	Criterio/con tenido (E/P)	3,84	-	
									10	Criterio/con tenido (N/X)	3,16	-	
									10	Criterio cualitativo (N/P)	0,299	-	
									10	Criterio cualitativo (E/P)	0,419	-	
									10	Criterio cualitativo (N/X)	0,328	-	
24	Si	Si	Si	No	No	Si	Si	Si	30	Cant. Atributos (1)	5,93	-	Analiza diversas variables para distintas técnicas de educción: (1) Ordenamient o “Triadic”
									30	Cant. Atributos (2)	5,57	-	
									30	Cant. Atributos (3)	4,1	-	

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
									30	Cant. Atributos (4)	4,3	-	(2) Ordenamiento Libre (3) Ordenamiento Directo (4) Ordenamiento por Rangos (5) Eligiendo de Lista de Atributos. Realiza análisis ANOVA para encontrar diferencias entre los grupos.
									30	Cant. Atributos (5)	2,6	-	
									30	Importancia (1)	4,264	-	
									30	Importancia (2)	4,659	-	
									30	Importancia (3)	4,722	-	
									30	Importancia (4)	4,885	-	
									30	Importancia (5)	5,054	-	
									30	Atrib. Concretos (1)	9,83	-	
									30	Atrib. Concretos (2)	9,53	-	
									30	Atrib. Concretos (3)	7,7	-	
									30	Atrib. Concretos (4)	8,6	-	
									30	Atrib. Concretos (5)	6,47	-	
25	Si	Si	Si	No	No	Si	Si	Si	43	Cant. Atributos (1)	4,49	1,4	Compara 4 técnicas de educción de requisitos: (1) Educción libre (2) Descripción de un ideal (3) Ordenamiento por rangos (4) Ord. p/ rangos para conjuntos dispares. Estas comparaciones las realiza con 2 fuentes distintas de conocimientos: autos y restaurantes
									39	Cant. Atributos (2)	4,33	2	
									37	Cant. Atributos (3)	4,32	1,67	
									38	Cant. Atributos (4)	3,61	1,41	
									43	Importancia (1)	5,82	0,83	
									39	Importancia (2)	5,77	0,81	
									37	Importancia (3)	5,38	1,09	
									35	Importancia (4)	5,25	1,03	
									41	Variabilidad (1)	1,12	0,52	
									39	Variabilidad (2)	1,13	0,53	

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
									37	Variabilidad (3)	1,3	0,49	(no tabulado).
									38	Variabilidad (4)	1,4	0,67	
26	Si	Si	Si	No	No	Si	Si	No	20	Tipo de conocimiento educido	-	-	Análisis ANOVA para ver diferencias entre grupos.
27	No	No	Si	No	Si	Si	Si	Si	7	-	-	-	Análisis ANOVA para ver diferencias entre grupos.
28	Si	Si	Si	No	No	Si	Si	Si	4	1)Cant cláusulas (c/re)	113,3	20,2	Analiza el producto de la técnica de emparrillado denominado Laddering. El objetivo es estudiar el aprendizaje y la realimentación utilizando esta técnica en seis sesiones de educción. En algunos casos se daba realimentación (c/re) y en otros no había realimentación (s/re). Luego se midió el esfuerzo y la ganancia de 4 técnicas de educción. Finalmente se comparan 3 herramientas distintas de Laddering.
									4	1)Cant cláusulas (s/re)	138,5	60,5	
									4	2)Cant cláusulas (c/re)	131,5	57,9	
									4	2)Cant cláusulas (s/re)	138,3	55,4	
									4	3)Cant cláusulas (c/re)	153,8	53,7	
									4	3)Cant cláusulas (s/re)	145,5	71,6	
									4	4)Cant cláusulas (c/re)	144,5	62,7	
									4	4)Cant cláusulas (s/re)	141,3	59	
									4	5)Cant cláusulas (c/re)	149	46,5	
									4	5)Cant cláusulas (s/re)	149,3	49,1	
									4	6)Cant cláusulas (c/re)	74,5	38,6	
									4	6)Cant cláusulas (s/re)	104,3	45,3	
									8	Entrevista – Esfuerzo	39,5	14,3	
									8	Auto-reporte – Esf.	26,75	5,14	

Publicación	Estimador	Error en Estimador	Cantidad de sujetos	Fuentes publicadas	Datos sin procesar publicados	Técnicas estadísticas aplicadas	Aporta Conclusiones	Fundamenta concl. Estadísticamente	Cantidad de Sujetos	Variable medida	Media	Desviación estándar	Comentarios
									8	Laddered – Esfuerzo	40,75	16	
									8	Orden. de cartas- Esf.	29,75	13,6	
									8	Entrevista – Ganancia	274	102	
									8	Auto-reporte – Gan.	145	74	
									8	Laddered – Gan.	521,4	420	
									8	Orden. de cartas- Gan.	144	52	
									16	Laddering Tool - ALTO	50,94	13,5	
									16	Laddering Graphical	87,69	14,2	
									16	Laddering Textual	102,6	23,8	
29	Si	Si	Si	No	No	Si	Si	Si	12	-	-	-	Análisis ANOVA para ver diferencias entre grupos.
30	Si	Si	Si	No	No	Si	Si	Si	20	-	-	-	Análisis ANOVA para ver diferencias entre grupos.
31	Si	Si	Si	No	No	Si	Si	Si	19	-	-	-	Análisis ANOVA para ver diferencias entre grupos.
32	Si	Si	Si	No	No	Si	Si	Si	25	-	-	-	Análisis ANOVA para ver diferencias entre grupos.
33	No estudia técnicas de prueba o educción de conocimientos												
34	No estudia técnicas de prueba o educción de conocimientos												
35	No estudia técnicas de prueba o educción de conocimientos												
36	No estudia técnicas de prueba o educción de conocimientos												
37	No estudia técnicas de prueba o educción de conocimientos												

Tabla 5.1. Evolución de los estudios empíricos

5.2. VERIFICACIÓN DE CONCLUSIONES

En este apartado se revisan los resultados presentados en los experimentos que publicaron los datos necesarios para poder realizar una constatación de resultados (los datos primarios del estudio).

5.2.1 Aplicación de métodos no paramétricos a los resultados del estudio 1

Objetivo:

> En este estudio se compara la efectividad de diversas técnicas de prueba de software. Las conclusiones a las que se llega, no fueron sometidas a técnicas estadísticas para validarlas, por lo que en este caso se aplican las mismas a los valores obtenidos en el informe. Esto se hará comparando todas las técnicas de forma conjunta (tomadas de a cuatro) y de a pares.

5.2.1.1 Comparación de las Técnicas todas juntas

Para comparar las cuatro técnicas, se usará el H-Test, cuyo primer paso es la asignación de rangos a las mediciones tomadas (Ver Anexo para información sobre el método aplicado).
Los porcentajes de efectividad fueron tomados del estudio y a cada uno de ellos se le asignó un rango tal como lo indica la técnica.

En la Tabla 5.2 se observa cómo se asigna un rango a cada valor

mutación		**mutación abs/ror**		**mutación 10%**		**All-uses**	
Efectividad (%)	Rango	Efectividad (%)	Rango	Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	31.5	100.00	31.5	86.21	13	56.67	3
100.00	31.5	100.00	31.5	96.67	20.5	96.67	20.5
100.00	31.5	76.67	11.5	60.00	5.5	100.00	31.5
100.00	31.5	86.67	14	96.67	20.5	90.00	15.5
93.33	17.5	60.00	5.5	66.67	8	60.00	5.5

mutación		mutación abs/ror		mutación 10%		All-uses	
70.00	9	53.33	2	60.00	5.5	40.00	1
100.00	31.5	100.00	31.5	93.33	17.5	100.00	31.5
100.00	31.5	100.00	31.5	100.00	31.5	100.00	31.5
100.00	31.5	100.00	31.5	76.67	11.5	96.67	20.5
100.00	31.5	100.00	31.5	73.33	10	90.00	15.5

Tabla 5.2. Asignación de rangos a las observaciones

Aplicando las fórmulas (5) y (6), se llega a que $H = 8.32$.

Los valores de χ para distintos valores de significación α son:

$$\chi^2_{0.05,3} = 7.81$$

$$\chi^2_{0.025,3} = 9.35$$

$$\chi^2_{0.01,3} = 11.3$$

Por lo tanto, se puede rechazar la hipótesis para un valor de significación de 0.05%, pero no para valores menores. A continuación, se debe efectuar una prueba comparando las técnicas de a pares, para hallar de cuáles se obtienen resultados diferentes.

5.2.1.2 Comparación de las Técnicas agrupadas de a dos

A continuación se comparan las técnicas mediante el U-Test. Si se reemplazan las variables por sus respectivos valores, la fórmula (4) del Anexo II, se transforma en:

$$z = \frac{105 - R_1}{13.23}$$

La hipótesis de que no hay diferencia en los resultados con ambas técnicas se rechazará si el módulo de z es mayor que 1.96, lo que equivale a decir que el nivel de significación es del 5%.

1) Mutación fuerte vs. mutación abs/ror.

mutación		mutación abs/ror	
Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	13.50	100.00	13.50
100.00	13.50	100.00	13.50
100.00	13.50	76.67	4.00
100.00	13.50	86.67	5.00
93.33	6.00	60.00	2.00
70.00	3.00	53.33	1.00
100.00	13.50	100.00	13.50
100.00	13.50	100.00	13.50
100.00	13.50	100.00	13.50
100.00	13.50	100.00	13.50

Tabla 5.3. Asignación de rangos a las observaciones

$R_1 = 117$

$R_2 = 93$

Por lo tanto; $z = -0.907$ y no se puede rechazar la hipótesis.

2) Mutación fuerte vs. mutación 10%.

mutación		mutación 10%	
Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	16.00	86.21	7.00
100.00	16.00	96.67	10.50
100.00	16.00	60.00	1.50
100.00	16.00	96.67	10.50
93.33	8.50	66.67	3.00
70.00	4.00	60.00	1.50
100.00	16.00	93.33	8.50
100.00	16.00	100.00	16.00
100.00	16.00	76.67	6.00
100.00	16.00	73.33	5.00

Tabla 5.4. Asignación de rangos a las observaciones

En este caso los valores hallados para R_1 y R_2 son:

$R_1 = 140.5$

$R_2 = 69.5$

Por lo tanto; z = -2.68 y se puede rechazar la hipótesis. Es decir, que en este caso existe evidencia de que la mutación fuerte es más efectiva que la mutación 10%, tal como suponía el estudio.

3) Mutación fuerte vs. all-uses.

Mutación		All-uses	
Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	15.00	56.67	2.00
100.00	15.00	96.67	8.50
100.00	15.00	100.00	15.00
100.00	15.00	90.00	5.50
93.33	7.00	60.00	3.00
70.00	4.00	40.00	1.00
100.00	15.00	100.00	15.00
100.00	15.00	100.00	15.00
100.00	15.00	96.67	8.50
100.00	15.00	90.00	5.50

Tabla 5.5. Asignación de rangos a las observaciones

Hallando los siguientes valores:

$R_1 = 131$

$R_2 = 79$

Por lo tanto; z = -1.965. En este caso el valor de z está apenas (0.005) por encima del umbral del 5% de significación. La hipótesis, de todas formas, puede ser rechazada y puede afirmarse que la mutación fuerte es más efectiva que el criterio de all-uses. Nuevamente en este caso los resultados respaldan las conclusiones del estudio.

4) Mutación abs/ror vs. mutación 10%.

mutación abs/ror		Mutación 10%	
Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	17.00	86.21	9.00
100.00	17.00	96.67	12.50
76.67	7.50	60.00	3.00
86.67	10.00	96.67	12.50
60.00	3.00	66.67	5.00
53.33	1.00	60.00	3.00
100.00	17.00	93.33	11.00
100.00	17.00	100.00	17.00
100.00	17.00	76.67	7.50
100.00	17.00	73.33	6.00

Tabla 5.6. Asignación de rangos a las observaciones

$R_1 = 123.5$

$R_2 = 83.5$

Por lo tanto; $z = -1.39$ y no se puede rechazar la hipótesis.

5) Mutación abs/ror vs. all-uses.

mutación abs/ror		All-uses	
Efectividad (%)	Rango	Efectividad (%)	Rango
100.00	16.00	56.67	3.00
100.00	16.00	96.67	10.50
76.67	6.00	100.00	16.00
86.67	7.00	90.00	8.50
60.00	4.50	60.00	4.50
53.33	2.00	40.00	1.00
100.00	16.00	100.00	16.00
100.00	16.00	100.00	16.00
100.00	16.00	96.67	10.50
100.00	16.00	90.00	8.50

Tabla 5.7. Asignación de rangos a las observaciones

De donde se obtienen los siguientes valores de R_1 y R_2:

$R_1 = 115.5$

$R_2 = 94.5$

Por lo tanto; $z = -0.79$ y no se puede rechazar la hipótesis.

6) Mutación 10% vs. all-uses.

mutación 10%		All-uses	
Efectividad (%)	Rango	Efectividad (%)	Rango
86.21	9.00	56.67	2.00
96.67	14.50	96.67	14.50
60.00	4.00	100.00	18.50
96.67	14.50	90.00	10.50
66.67	6.00	60.00	4.00
60.00	4.00	40.00	1.00
93.33	12.00	100.00	18.50
100.00	18.50	100.00	18.50
76.67	8.00	96.67	14.50
73.33	7.00	90.00	10.50

Tabla 5.8. Asignación de rangos a las observaciones

$R_1 = 97.5$

$R_2 = 112.5$

Por lo tanto; $z = 0.567$ y no se puede rechazar la hipótesis.

Como se ha visto hasta aquí, solamente en dos de las seis comparaciones efectuadas, se puede rechazar la hipótesis y afirmar que existe diferencias en los resultados obtenidos por las técnicas comparadas.

5.2.2 Aplicación de Métodos No Paramétricos a los resultados del estudio 2

Objetivo:

> Este análisis es similar al caso anterior, donde las conclusiones halladas no fueron sometidas a técnicas estadísticas que las respalden, sino que se obtuvieron analizando “a ojo” las mediciones tomadas.

5.2.2.1 Comparación de las Técnicas agrupadas de a dos

Para poder obtener resultados generales se promediaron los valores de efectividad para cada programa, de esta forma cada programa aporta con el mismo peso a los resultados finales.

A continuación se comparan las técnicas mediante el U-Test. Si se reemplazan las variables por sus respectivos valores, la fórmula (4) del Anexo, se transforma en:

$$z = \frac{105 - R_1}{13.23}$$

La hipótesis de que no hay diferencia en los resultados con ambas técnicas se rechaza si el módulo de *z* es mayor que 1.96, lo que equivale a decir que el nivel de significación es del 5%.

1) Mutación fuerte vs. All-uses.

	Mutación fuerte		All-uses	
	Efectividad	Rango	Efectividad	Rango
determinant	4%	1	17%	3
find1	99%	15	82%	14
find2	35%	6	30%	4
matinv1	100%	18	79%	12
matinv2	48%	8	14%	2
strmatch1	100%	18	100%	18
strmatch2	75%	11	44%	7
textformat.o	68%	9	100%	18
textformat.r	70%	10	100%	18
transpose	80%	13	34%	5

Tabla 5.9. Asignación de rangos a las observaciones

$R_1 = 109$

$R_2 = 101$

Por lo tanto; *z* = -0.302 y no se puede rechazar la hipótesis y no se puede afirmar que un criterio sea más efectivo que el otro. Notar que lo que se rechaza es que exista un criterio que sea, para todos los programas, más efectivo que el otro.

5.2.3. Estudio de las técnicas aplicadas por el estudio 3

Objetivo:

Verificar el método de estimación de casos de prueba.

En el informe de Weyuker se intenta encontrar la forma de estimar empíricamente la cantidad de casos de prueba que son necesarios para cubrir los criterios de prueba all-c-uses, all-p-uses, all-uses y all-du-paths.

Con el objeto de hallar una fórmula para estimar la cantidad necesaria de casos de prueba para cubrir cada criterio, primero se debe encontrar la o las variables dependientes de las que dependerá la cantidad de casos de prueba. Según [Weyuker, 1984], la cota superior de casos de prueba depende únicamente de la cantidad de sentencias de decisión, por lo que Weyuker supone que esta variable será la misma para estimar el número de casos de prueba. Luego se debe encontrar la forma en que se relacionan estas variables. La forma escogida fue la lineal, pero sin argumento teórico que sustente esta decisión (la cantidad máxima de casos de prueba encontrada en [Weyuker, 1984] es cuadrática para los tres primeros criterios y exponencial para el cuarto), por lo que los datos medidos empíricamente deberían sustentarla. Finalmente se deben estimar los parámetros de la ecuación, para lo que el estimador elegido fue el de mínimos cuadrados.

Weyuker analiza los programas del conjunto de programas Software Tools in Pascal por [Kernighan y Plauger, 1981]. De los más de 100 programas del conjunto, se eligen los 29 que tienen 5 o más sentencias de decisión (***d***). A cada uno de estos programas se le calcularon los casos de de prueba (***t***) para cada uno de los criterios a estudiar. De esta forma se obtuvieron 29 pares ordenados de la forma (***d***,***t***). Mediante el estimador de mínimos cuadrados, se ajustaron estos puntos encontrados a una recta.

En el informe, están publicadas las ecuaciones de la recta a las que se llega para cada criterio:

all-c-uses:

$t = 0.52\ d + 1,87$

all-p-uses:

$t = 0.76\ d + 1{,}01$

all-uses:

$t = 0.81\ d + 1{,}12$

all-du-paths:

$t = 0.93\ d + 1{,}10$

Finalmente se llega a la conclusión de que la relación es lineal basándose en que el coeficiente de d es menor que la unidad. Este argumento no tiene ningún respaldo estadístico.

Como se explica en el Anexo, los errores en las mediciones deben cumplir ciertos criterios, que no fueron verificados en el informe. También se describe cómo, el coeficiente de correlación, puede ayudar a verificar si las fórmulas halladas son correctas o no. Este coeficiente tampoco fue publicado, ni tampoco los valores de los pares ordenados (d_i,t_i) para corroborar las ecuaciones obtenidas. Por lo tanto, no hay forma de verificar la validez de las fórmulas halladas. Ni siquiera fue publicado el valor máximo de d_i, por lo que es imposible conocer cuál sería el rango de validez de la fórmula hallada en el caso en que ésta hubiera sido correcta.

El capítulo IV del informe de Weyuker contiene información obtenida por [Shimeall y Leveson, 1988] para el criterio de all-p-uses. En este caso, los pares ordenados (d_i,t_i) fueron publicados y se encuentran en la Tabla 5.9.

Decisiones	Test cases
173	104
196	115
209	103
246	103
301	113
325	110
334	117
434	107

Tabla 5.10. Mediciones efectuadas por Shimeall y Leveson

En este caso, las mediciones se tomaron sobre ocho programas y haciendo un análisis sobre estos datos, se puede determinar si el estimador de mínimos cuadrados para hallar una relación lineal es adecuado o no.

El estimador de mínimos cuadrados obtiene ecuación $t = 0.02\ d + 104{,}28$ cuya gráfica se encuentra en la Figura 5.11.

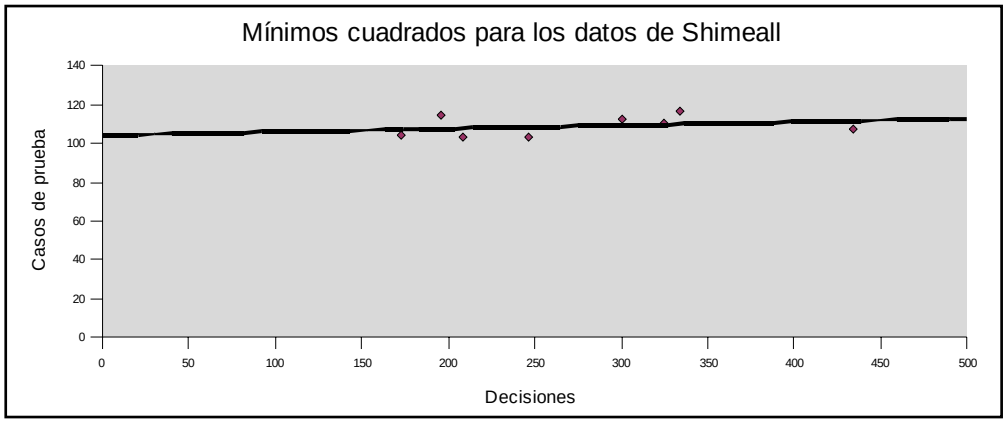

Figura 5.11. Cálculo de regresión para los datos de Shimeall y Leveson

Si se comparan las fórmulas obtenidas por Weyuker y Shimeall, éstas difieren considerablemente:

t = 0.76 d + 1,01

t = 0.02 d + 104,28

En la Figura 5.12 se observan ambas rectas:

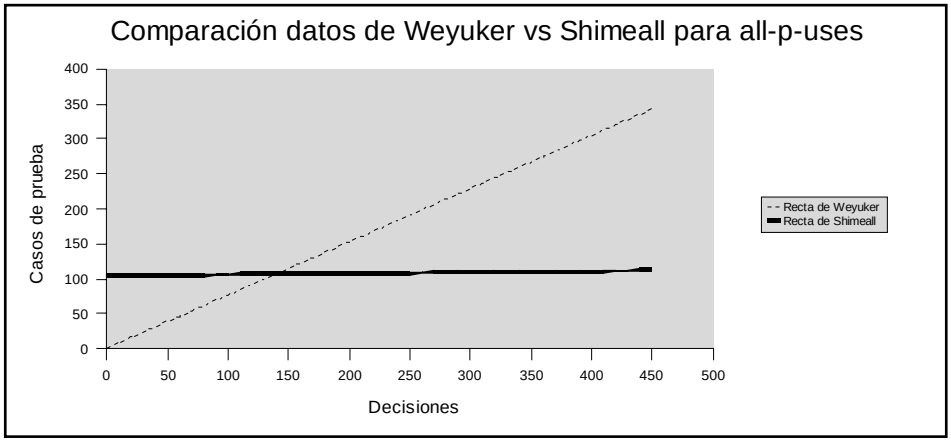

Figura 5.12. Comparación de las rectas de Weyuker vs. Shimeall.

El gráfico muestra que ambos resultados no son compatibles, salvo que se especifique el rango de número de decisiones en el que cada uno tiene validez.

Por otro lado el cálculo del coeficiente de correlación para la recta de Shimeall da 0,27. Este valor es bajo, lo que indica que la relación de estas variables no es lineal y que por lo tanto la ecuación encontrada no se corresponde con los datos medidos.

5.2.4. Análisis de los resultados obtenidos por el estudio 4

Objetivo:

Verificar el método de estimación de casos de prueba.

Este documento analiza dos técnicas de prueba de caja blanca: los criterios de branch testing y de all-uses. El objetivo del mismo es observar cómo varía la efectividad de las técnicas en función de la cobertura del criterio. Por ejemplo, para el caso de branch testing, verifica qué tan bueno es un conjunto de casos de prueba, en función de qué tanto cubre todas las sentencias de un programa.

Con el objeto de medir la efectividad y la cobertura, se toma un programa y se le introducen 33 errores reales; es decir, que ya fueron corregidos durante la construcción del programa. Como estos errores se agregan de a uno por vez, se obtienen 33 programas y cada uno con un error.

Se generan 10000 casos de prueba aleatoriamente. Luego, se ejecutan 5000 de los casos de prueba sobre los 33 programas y se seleccionan solamente aquellos donde el índice de fallos es menor al 1,5 %; es decir, aquellos programas que fallaron con no más de 75 de los 5000 casos de prueba usados. A continuación, los casos de prueba se agrupan de forma aleatoria para generar conjuntos de casos de prueba. A estos conjuntos, se les mide el porcentaje de cobertura sobre cada una de las dos técnicas y la efectividad, que puede ser 1 o 0, dependiendo de si hacen fallar al programa o no.

Los resultados que se obtienen son gráficas que muestran que a medida que la cobertura aumenta, así también lo hace la efectividad, lo que constituye una de las conclusiones del informe.
Otra conclusión a la que se llega es que estos criterios son mejores que la generación de casos de prueba de forma aleatoria. En este caso la conclusión no está justificada debido a dos razones:

1. No se publica la curva de distribución de cobertura de los conjuntos de casos de prueba generados aleatoriamente.
2. Los programas analizados fueron seleccionados de forma que se induce a llegar a esta conclusión.

La curva mencionada en el punto 1. permite observar qué tan probable es que un conjunto de casos de prueba aleatorio tenga una determinada cobertura. Sin esto, es imposible afirmar que los conjuntos de prueba aleatorios tienen baja cobertura y que por lo tanto, no son efectivos. Aunque parezca trivial, es necesaria la publicación de este dato, para asegurarse que es difícil alcanzar una alta cobertura con conjuntos generados aleatoriamente.

El punto 2. se refiere a que el análisis es tendencioso ya que se eligieron los 11 programas que contenían fallas difíciles de encontrar con casos de prueba generados aleatoriamente. Por lo tanto, no se podría llegar a la conclusión de que estas técnicas son mejores que las técnicas aleatorias, si no que lo único que se puede afirmar es que son mejores, cuando las técnicas aleatorias tienen una baja efectividad.

5.2.5. Análisis de los resultados del estudio 5

En este informe se medió el número de casos de prueba necesarios para satisfacer el criterio all-du-paths con el objetivo de verificar si este criterio era utilizable en la práctica. En el informe se publican todas las mediciones hechas, pero no se hace ningún análisis estadístico de los resultados, como tampoco se llega a ninguna fórmula que permita, en base a ciertos parámetros del programa, estimar la cantidad de casos de prueba.

5.2.6. Análisis de los resultados del estudio 6

Objetivo:

Verificar el método de estimación de casos de prueba.

Este estudio compara las técnicas de mutación selectiva contra la mutación fuerte, llegando a la conclusión cualitativa de que la mutación selectiva es una alternativa más eficiente que la mutación fuerte. Sin embargo, durante el desarrollo de la experimentación se obtienen fórmulas para lograr estimar la cantidad de mutantes en función de ciertas variables del programa. Las variables independientes son:

Vars: Cantidad de variables que tiene el programa
Varrefs: Cantidad de referencias a variables que tiene el programa
Lines: Cantidad de líneas de código que tiene el programa
Units: Cantidad de *Program Units* que tiene el programa

Debido a que se estima que las fórmulas no son lineales, se crean variables independientes formadas por la multiplicación de las variables independientes mencionadas. De esa forma, se llega a las siguientes ecuaciones:

$$\text{Mutantes} = a_0 + a_1 \text{Lines} + a_2 \text{Varrefs} + a_3 \text{Vars Varrefs}$$

$$\text{Mutantes} = a_0 + a_1 \text{Lines} + a_2 \text{Lines Lines}$$

$$\text{Mutantes} = a_0 + a_1 \text{V}ars + a_2 \text{Vars Vars}$$

$$\text{Mutantes} = a_0 + a_1 \text{V}ars + a_2 \text{Varrefs} + a_3 \text{Units} + a_4 \text{Vars Varrefs}$$

A estas fórmulas se le calcula el coeficiente de determinación y dan valores que varían entre 0,90 y 0,97 por lo que se supone que las fórmulas son formas correctas de estimación de la cantidad de mutantes.

En el informe no se publican los coeficientes a_i hallados para cada fórmula, por lo que no se puede comparar los resultados contra nuevos experimentos.

5.2.7. Aplicación de Métodos No Paramétricos a los resultados del estudio 7

En este artículo, se comparan 4 técnicas de mutación débil. Se generan conjuntos de casos de prueba que tienen una puntuación del 100% para cada una de las técnicas y luego, se aplican estos mismos conjuntos de prueba a la técnica de mutación fuerte y

se mide la puntuación. Los resultados figuran en la Tabla 5.13. Cabe aclarar que los mismos fueron extrapolados de un gráfico ya que la tabla no fue publicada.

5.2.7.1 Comparación de las Técnicas todas juntas

Para comparar las cuatro técnicas, se usará el H-Test, cuyo primer paso es la asignación de rangos a las mediciones tomadas:

EX-Weak/1		ST-Weak/1		BB-Weak/1		BB-Weak/N	
Efectividad (%)	Rango	Efectividad (%)	Rango	Efectividad (%)	Rango	Efectividad (%)	Rango
98.7	22	99.4	29	99.3	27.5	99.0	25
97.7	15	99.3	27.5	99.1	26	98.3	20.5
99.9	34.5	99.9	34.5	99.6	30.5	98.9	23.5
97.2	12	98.1	18	98.3	20.5	97.4	14
100.0	40.5	100.0	40.5	100.0	40.5	100.0	40.5
96.7	10	100.0	40.5	99.9	34.5	99.9	34.5
64.0	1	85.5	2	86.0	3	92.5	4
98.9	23.5	100.0	40.5	100.0	40.5	100.0	40.5
94.8	6.5	97.2	12	97.8	16.5	97.8	16.5
94.2	5	94.8	6.5	95.5	8.5	95.5	8.5
98.2	19	99.6	30.5	99.8	32	97.2	12

Tabla 5.13. Asignación de rangos a las observaciones

Aplicando las fórmulas (5) y (6) del Anexo, se llega a que $H = 3.16$.

Los valores de χ para distintos valores de significación α son:

$$\chi^2_{0.5,3} = 2.37$$

$$\chi^2_{0.25,3} = 4.11$$

Por lo tanto, se puede rechazar la hipótesis para un valor de significación que se encuentra entre el 25% y el 50%. Estos valores son demasiado altos como para rechazar la hipótesis y afirmar que las técnicas obtienen resultados distintos.

5.2.7.2 Comparación de las Técnicas agrupadas de a dos

Al momento de comparar las técnicas de a dos, se eligieron aquellas dos con los valores de R_i más distintos. En este caso el valor de R_1 (EX-Weak/1) fue de 189 y el valor de R_2 (ST-Weak/1) fue de 281,50.

A continuación se comparan las técnicas mediante el U-Test. Si se reemplazan las variables por sus respectivos valores, la fórmula (4) se transforma en:

$$z = \frac{126.5 - R_1}{15.23}$$

La hipótesis de que no hay diferencia en los resultados con ambas técnicas se rechazará si el módulo de *z* es mayor que 1.96, lo que equivale a decir que el nivel de significación es del 5%.

EX-Weak/1		ST-Weak/1	
Efectividad (%)	Rango	Efectividad (%)	Rango
98.70	12.00	99.40	15.00
97.70	9.00	99.30	14.00
99.90	17.50	99.90	17.50
97.20	7.50	98.10	10.00
100.00	20.50	100.00	20.50
96.70	6.00	100.00	20.50
64.00	1.00	85.50	2.00
98.90	13.00	100.00	20.50
94.80	4.50	97.20	7.50
94.20	3.00	94.80	4.50
98.20	11.00	99.60	16.00

Tabla 5.14. Asignación de rangos a las observaciones

$R_1 = 105$

$R_2 = 148$

Por lo tanto; $z = 1.41$ y no se puede rechazar la hipótesis por lo que no se puede apreciar una diferencia entre los resultados usando un método o el otro.

5.3. SÍNTESIS DE RESULTADOS

Los resultados obtenidos en el análisis de las técnicas de prueba se encuentran en la Tabla 5.15, donde se resumen los campos más destacados.

<table>
<tr><th>DOMINIO</th><th>APLICADO A</th><th>TÉCNICAS UTILIZADAS</th><th>MEDICIONES DOCUMENTADAS</th><th>RESULTADOS OBTENIDOS</th><th>RESUMEN DEL ANÁLISIS</th></tr>
<tr>
<td>Comparación de la efectividad en la detección de fallas entre las técnicas de prueba de mutación y de flujo de datos [1]</td>
<td>Cinco programas distintos:
FIND (1*)
SORT (2*)
STRMATCH (3*)
POSITION (2*)
STAT (2*)
* Cantidad de veces que se ejecutó el programa.</td>
<td>1- Mutación fuerte
2- Mutación abs/ror
3- Mutación 10%
4- All-uses</td>
<td>1- Efectividad en la detección de fallas
$\frac{N}{T} \cdot 100\ \%$
N = Número de conjunto de pruebas que exponen al menos una falla.
T = Número total de conjuntos de prueba generados.</td>
<td>
<table>
<tr><th colspan="3">Mutación</th><th>Datos</th></tr>
<tr><th>Fuerte</th><th>abs/ror</th><th>10%</th><th>alluses</th></tr>
<tr><td>100.00</td><td>100.00</td><td>86.21</td><td>56.67</td></tr>
<tr><td>100.00</td><td>100.00</td><td>96.67</td><td>96.67</td></tr>
<tr><td>100.00</td><td>76.67</td><td>60.00</td><td>100.00</td></tr>
<tr><td>100.00</td><td>86.67</td><td>96.67</td><td>90.00</td></tr>
<tr><td>93.33</td><td>60.00</td><td>66.67</td><td>60.00</td></tr>
<tr><td>70.00</td><td>53.33</td><td>60.00</td><td>40.00</td></tr>
<tr><td>100.00</td><td>100.00</td><td>93.33</td><td>100.00</td></tr>
<tr><td>100.00</td><td>100.00</td><td>100.00</td><td>100.00</td></tr>
<tr><td>100.00</td><td>100.00</td><td>76.67</td><td>96.67</td></tr>
<tr><td>100.00</td><td>100.00</td><td>73.33</td><td>90.00</td></tr>
</table>
</td>
<td>Aunque los resultados parecen confirmar la hipótesis del autor, la aplicación de técnicas estadísticas, muestran que la evidencia no es suficiente para afirmar que las variantes de la técnica de mutación sean más efectivas que las técnicas de all-uses.
Ver sección 5.2.1</td>
</tr>
<tr>
<td>Comparación de la efectividad en la detección de fallas entre la técnica de mutación fuerte y la técnica all-uses [2].</td>
<td>DETERMINANT (6*)
FIND1 (7*)
FIND2 (10*)
MATINV1 (11*)
MATINV2 (5*)
STRMATCH1 (4*)
STRMATCH2 (3*)
TEXTFORMAT.O (10*)
TEXTFORMAT.R (5*)
TRANSPOSE (16*)</td>
<td>1- Mutación fuerte
2- All-uses</td>
<td>Se define la efectividad de un criterio C sobre un programa P como la probabilidad de que un conjunto de pruebas que satisface al criterio C y generado mediante la estrategia G exponga una falla en P.
* La estrategia de generación G utilizada fue la aleatoria.</td>
<td>
<table>
<tr><th>Programa</th><th>Mutación</th><th>alluses</th></tr>
<tr><td>determinant</td><td>0.04</td><td>0.17</td></tr>
<tr><td>find1</td><td>0.99</td><td>0.82</td></tr>
<tr><td>find2</td><td>0.35</td><td>0.30</td></tr>
<tr><td>matinv1</td><td>1.00</td><td>0.79</td></tr>
<tr><td>matinv2</td><td>0.48</td><td>0.14</td></tr>
<tr><td>strmatch1</td><td>1.00</td><td>1.00</td></tr>
<tr><td>strmatch2</td><td>0.75</td><td>0.44</td></tr>
<tr><td>textformat.r</td><td>0.68</td><td>1.00</td></tr>
<tr><td>textformat.o</td><td>0.70</td><td>1.00</td></tr>
<tr><td>transpose</td><td>0.80</td><td>0.34</td></tr>
</table>
</td>
<td>En este caso, la hipótesis de que la técnica de mutación utilizada es más efectiva que la de all-uses, no está soportada por las técnicas estadísticas.
Ver sección 5.2.2</td>
</tr>
</table>

Tabla 5.15. Resumen de estudios empíricos relevados

DOMINIO	APLICADO A	TÉCNICAS UTILIZADAS	MEDICIONES DOCUMENTADAS	RESULTADOS OBTENIDOS	RESUMEN DEL ANÁLISIS
Estudio empírico para estimar el costo de usar los criterios de prueba de flujo de datos [3].	Se eligen los 29 programas de la plataforma Software Tools de Kernighan y Plauger que tienen 5 o más sentencias de decisión.	1- all-c-uses 2- all-p-uses 3- all-uses 4- all-du-paths	Se midió la cantidad de sentencias de decisión (***d***) y la cantidad de casos de prueba necesarios (***t***) para satisfacer cada criterio	**all-c-uses**: $t = 0.52\ d + 1{,}87$ **all-p-uses**: $t = 0.76\ d + 1{,}01$ **all-uses**: $t = 0.81\ d + 1{,}12$ **all-du-paths**: $t = 0.93\ d + 1{,}10$	Faltan publicar datos para poder corroborar la validez de las ecuaciones encontradas: rango de validez de las ecuaciones, coeficiente de correlación, pares ordenados de mediciones. De todas formas todo pareciera indicar que la linearización hallada no es correcta. Ver sección 5.2.3
Evaluación empírica de la capacidad de detectar errores de dos técnicas de prueba de caja blanca [4].	Se utilizaron 33 versiones de un programa real, cada una de ellas con un error real, que fue insertado	1- Branch testing 2- all-uses	Se midió la efectividad de la técnica de prueba calculando el porcentaje de conjuntos de prueba que encontraron el error en cada uno de los programas analizados.	Se graficó la efectividad en función del porcentaje de cobertura de cada uno de los criterios. Se observó que a medida que el porcentaje de cobertura es mayor, mayor es la efectividad medida.	Falta publicar datos para corroborar que ambos criterios son más efectivos que seleccionar los casos de prueba aleatoriamente: se debería publicar cómo se distribuyen los conjuntos de prueba aleatorios en función de la cobertura que cada uno de ellos alcanza. Los resultados son tendenciosos ya que el autor selecciona solamente 11 de los 33 programas de forma que se induce a alcanzar los resultados obtenidos. Ver sección 5.2.4

Tabla 5.15. Resumen de estudios empíricos relevados (cont.)

DOMINIO	APLICADO A	TÉCNICAS UTILIZADAS	MEDICIONES DOCUMENTADAS	RESULTADOS OBTENIDOS	RESUMEN DEL ANÁLISIS
Estimación del número de casos de prueba necesarios para satisfacer el criterio all-du-paths [5].	Se analizaron 143 subrutinas (procedimientos y funciones) de un programa para analizar textos en lenguaje natural	1- all-du-paths	Se midieron las siguientes variables: LOCs, cantidad de nodos, cantidad de bordes, cantidad de du-paths, cantidad de du-paths no redundantes y número mínimo de caminos completos para satisfacer el criterio de all-du-paths.	Se observó que el número de caminos completos que satisfacen el criterio de all-du-paths fue: 0 ≤ Cantidad ≤ 10 80,4% 11 ≤ Cantidad ≤ 25 10,5% 26 ≤ Cantidad ≤ 50 4,2% 51 ≤ Cantidad ≤ 100 2,8% 101 ≤ Cantidad ≤ 400 1,4% 401 ≤ Cantidad 1,4%	En este caso, el autor publica todos los datos encontrados lo que permite hacer todo tipo de análisis sobre los mismos. Sin embargo, al no llegar a ninguna conclusión en particular, no se puede realizar el análisis. Ver sección 5.2.5
Comparación de la técnica de mutación selectiva contra la mutación fuerte [6]	Se analizaron 28 programas en Fortran-77, cuyo tamaño variaba entre 8 y 164 líneas de código.	1- 2-selective 2- 4-selective 3- 6-selective	Se midió la cantidad de mutantes generados por cada una de las técnicas. Se calculó el score de los test sets adecuados con estas tres técnicas frente a la mutación fuerte.	Las conclusiones cualitativas sobre los porcentajes de adecuación hallados son correctas. El estudio incluye también 4 fórmulas no lineales para estimar la cantidad de mutantes en función de distintos parámetros: líneas de código, variables y referencias a variables	No se publicaron los coeficientes de las fórmulas halladas, sólo la forma de las mismas. Ver sección 5.2.6

Tabla 5.15. Resumen de estudios empíricos relevados (cont.)

<table>
<tr><th>DOMINIO</th><th>APLICADO A</th><th>TÉCNICAS UTILIZADAS</th><th>MEDICIONES DOCUMENTADAS</th><th>RESULTADOS OBTENIDOS</th><th>RESUMEN DEL ANÁLISIS</th></tr>
<tr>
<td>Comparación de diversas técnicas de mutación débil entre ellas y contra la mutación fuerte [7].</td>
<td>Se analizaron 11 programas de entre 10 y 29 sentencias.</td>
<td>1- EX-weak/1
2- ST-weak/1
3- BB-weak/1
4- BB-weak/N</td>
<td>Se midió la cantidad de mutantes generados por cada una de las técnicas. Se calculó el score de los test sets adecuados con estas cuatro técnicas frente a la mutación fuerte.</td>
<td>
<table>
<tr><th>Prog</th><th>EX/1</th><th>ST/1</th><th>BB/1</th><th>BB/N</th></tr>
<tr><td>bub</td><td>98.7</td><td>99.4</td><td>99.3</td><td>99.0</td></tr>
<tr><td>cal</td><td>97.7</td><td>99.3</td><td>99.1</td><td>98.3</td></tr>
<tr><td>euclid</td><td>99.9</td><td>99.9</td><td>99.6</td><td>98.9</td></tr>
<tr><td>find</td><td>97.2</td><td>98.1</td><td>98.3</td><td>97.4</td></tr>
<tr><td>insert</td><td>100.0</td><td>100.0</td><td>100.0</td><td>100.0</td></tr>
<tr><td>mid</td><td>96.7</td><td>100.0</td><td>99.9</td><td>-</td></tr>
<tr><td>pat</td><td>64.0</td><td>85.5</td><td>86.0</td><td>92.5</td></tr>
<tr><td>quad</td><td>98.9</td><td>100.0</td><td>100.0</td><td>-</td></tr>
<tr><td>trismall</td><td>94.8</td><td>97.2</td><td>97.8</td><td>-</td></tr>
<tr><td>trityp</td><td>94.2</td><td>94.8</td><td>95.5</td><td>-</td></tr>
<tr><td>warshall</td><td>98.2</td><td>99.6</td><td>99.8</td><td>97.2</td></tr>
</table>
</td>
<td>Los valores obtenidos no permiten afirmar que alguna de las técnicas de mutación débil sea preferible a las restantes.
Sin embargo, la conclusión cualitativa de que las técnicas de mutación débil son casi tan efectivas como la mutación fuerte es correcta.
Ver sección 5.2.7</td>
</tr>
</table>

Tabla 5.15. Resumen de estudios empíricos relevados (cont.)

5.4. TABLA RESUMEN DE PARÁMETROS

A continuación se describen los valores típicos obtenidos del análisis de los datos reportados en los experimentos de educción de requisitos:

	Promedio	*Máximo*	*Mínimo*	*Moda*
Medias	77.29	521.4	3.61	
Desvió estándar	17.40	144	1.17	
Sujetos	16	43	2	4
Relación Media - Desvió estándar	0.38	0.72	0.03	
Porcentaje de estudios que no publican medias	7.69			
Porcentaje de estudios que no publican desvío estándar	53.85			
Porcentaje de estudios que no publican Cantidad de sujetos	15.38			

Tabla 5.16. Detalle de los valores típicos de los experimentos

CAPITULO 6

6. CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

6.1 CONCLUSIONES

A continuación se describen las conclusiones obtenidas de la evaluación de los parámetros estadísticos publicados en los estudios empíricos analizados:

- El 50 % de los estudios no publican todas las variables estadísticas (medias, varianzas y cantidad de sujetos). Por lo tanto, no podrían formar parte de un proceso de agregación estándar mediante Meta-Análisis.
- La Varianza es la variable estadística menos reportada, el 53.85 % de los estudios no la reporta, esto imposibilitaría poder aplicar Meta-Análisis en más de la mitad de los experimentos.
- Si bien el promedio de sujetos experimentales por experimento es de 16, existen varios estudios que trabajan con solo 2 sujetos experimentales y la moda obtenida es igual a 4. Estos valores son muy bajos y limitan la potencia de los test estadísticos paramétricos.
- La relación que existe entre el promedio de las medias y Desvío estándar casi del 40 %, lo cual implica que para una media de 100 el desvío estándar esperado es de 40. Este nivel de variación es bastante elevado, lo cual podría indicar una baja homogeneidad entre los sujetos experimentales.

A continuación se describen las conclusiones obtenidas de la evaluación de los resultados publicados en los estudios donde se publicaban los datos originales:

- Comparando la técnica de mutación frente a la técnica all-uses en términos de efectividad; se infiere que no existe evidencia concreta que permita aseverar que una técnica es mejor que otra. Las pruebas realizadas solamente permiten

afirmar que la técnica de mutación fuerte es más efectiva que una de sus variantes (mutación 10%) y que all-uses para el caso en que se probaron 5 programas. Sin embargo, en el experimento en el que se utilizan 10 programas, no se llega a la misma conclusión; es decir, que no se puede afirmar que una técnica sea más efectiva que la otra. Los resultados obtenidos demuestran que el programa que se mide es una variable que no puede ser obviada al momento de sacar conclusiones, lo que dificulta el análisis de los resultados ya que en principio los resultados a obtener serían distintos para cada programa. Una posible forma de sortear este obstáculo podría ser hallar la técnica más efectiva en función de ciertas métricas del programa (caminos posibles, líneas de código, cantidad de bifurcaciones, etc.), para, de acuerdo a las mismas, elegir la técnica apropiada.

- En el caso de las técnicas de flujo de datos, se observa que los resultados fundamentan las conclusiones cualitativas, pero que no respaldan las fórmulas cuantitativas a las que se llega. Algo similar ocurre con las técnicas de flujos de control, donde los resultados hallados, aunque parecen ser correctos a primera vista, no están totalmente fundados, y en algunos casos son inducidos por la forma en que se seleccionan los objetos experimentales.

En ninguno de los estudios donde se aplican técnicas estadísticas, se verifica que las variables independientes cumplan con las condiciones que impone el uso de la técnica (confirmación de la distribución de los datos, homogeneidad de varianzas, etc.). Por lo tanto, las relaciones halladas entre las variables es meramente descriptiva, aunque sugiera una cierta correlación entre las misma, pero lejos se está de hallar una relación causal.

Para los experimentos sobre técnicas de prueba se observan los siguientes puntos:

- La media de la cantidad de sujetos es de 12,1 con una desviación estándar de 6,8 y una moda de 10 (Solamente se contabilizaron 8 de los 9 estudios, ya que uno de ellos presenta un valor atípico. Este estudio tiene 143 sujetos, debido a que analiza las funciones de 3 programas y por lo tanto no se puede comparar con los otros estudios que tratan al programa como una unidad.

- Los estudios de técnicas de prueba cumplen a lo sumo con el primero de los pasos de análisis repasado en el capítulo 2.2: estadística descriptiva.

Los estudios realizados sobre educción de conocimientos presentan un mayor formalismo estadístico que los hechos en pruebas de software. En general las diferencias entre grupos se fundamentan con la técnica ANOVA, aunque en ninguno de los casos se verifican las hipótesis necesarias para aplicar la misma. En un 74% de los estudios se realizó un test de hipótesis sobre los datos, pero en ninguno de los casos se hizo una reducción del conjunto de pruebas.

Los experimentos sobre técnicas de educción de conocimientos presentan las siguientes características:

- La media de sujetos es de 18,3 con una desviación estándar de 11,9 y una moda de 30

- Falta de homogeneidad en los estudios.

- Solamente el 30% publica la cantidad de sujetos, la media y la varianza -o desviación estándar.

- El porcentaje de la desviación estándar sobre la media tiene un valor de 45,9%, con una desviación estándar de 27,0%.

Todos estos puntos dificultan la posibilidad de realizar un análisis de agregación estándar mediante Meta-análisis sobre los estudios.

6.2 FUTURAS LÍNEAS DE INVESTIGACIÓN

- Ampliar la cantidad de estudios analizados, tomando nuevas revisiones sistemáticas o realizando una.
- En base a los valores de los parámetros estadísticos obtenidos identificar cuales son las técnicas estadísticas que mejor se adaptan a este contexto.

- Hacer una evaluación más detallada de los estudios analizando no sólo los parámetros estadísticos, sino también otros aspectos que hacen a la calidad de los mismos.

CAPÍTULO 7

7. REFERENCIAS

7.1 BIBLIOGRAFÍA ANALIZADA

[1] Mathur A. y Wong W., 1993. *Comparing the fault detection effectiveness of mutation and data flow testing: An empirical study*. Tech. Report SERC-TR-146-P, Software Engineering Research Center.

[2] Frankl P., Weiss S., Hu C., 1997. *All-uses versus mutation testing: An experimental comparison of effectiveness*. J. Systems and Software, Sept. 1997, 38(3): 235-253.

[3] E. J. Weyuker. *The Cost of Data Flow Testing: An Empirical Study*, IEEE Transactions on Software Engineering. Vol 16. No 2. February 1990.

[4] Frankl, P. y Iakounenko, O., 1998. *Further Empirical Studies of Test Effectiveness*. In Proceedings of the ACM SIGSOFT International Symposium on Foundations on Software Engineering, pages 153-162, Lake Buena Vista, Florida, USA.

[5] Bieman, J. y Schultz, J., 1989. *Estimating the Number of Test Cases Required to Satisfy the All-du-paths Testing Criterion*. ACM. Pages 179-186.

[6] Offutt, A.J., Rothermel, G. and Zapf, C., 1993. *An Experimental Evaluation of Selective Mutation*. Proceedings of the 15th International Conference on Software Engineering. Pages 100-107. Baltimore, USA. IEEE

[7] Offutt, A.J. and Lee, S.D., 1994. *An Empirical Evaluation of Weak Mutation*. IEEE Transactions on Software Engineering. Vol. 20(5). Pages 337-344.

[8] Frankl P., Weiss S., Hu C. *An Experimental Comparison of the Effectiveness of Branch Testing and Data Flow Testing*

[9] Offutt, A.J. and Lee, S.D. *An Experimental Determination of Sufficient Mutant Operators*

[10] Eva Hudlicka, et al. *Requirements Elicitation with Indirect Knowledge Elicitation Techniques: Comparison of Three Methods*

[11] Leonard Adelman. *Measurement Issues in Knowledge Engineering*

[12] Malcom Eva, et al. *Requirements Acquisition for rapid applications development*

[13] Victoria Goodrich, Lome Olfman. *An Experimental Evaluation of Task and Methodology Variables For Requirements Definition Phase Success*

[14] Janete W. Moody, et al. *Enhancing Knowledge Elicitation using the Cognitive Interview*

[15] J. Michael Moore. *A Comparison of Questionnaire-Based and GUI-Based Requirements Gathering*

[16] Beth W. Crandall. *A Comparative Study Of Think-Aloud And Critical Decision Knowledge Elicitation Methods*

[17] Lee A. Freeman. *The effects of concept maps on requirements elicitation and system models during information systems development*

[18] A. M. Burton, et al. *A Formal Evaluation of Knowledge Elicitation Techniques For Expert Systems*

[19] R. Schweickert, et al. *Comparing Knowledge Elicitation Techniques: A case Study*

[20] A. M. Burton, et al. *The Efficacy of Knowledge Elicitation Techniques: a comparison across domains and levels of expertise*

[21] G. Rugg, et al. *A Comparison of sorting techniques in knowledge acquisition.*

[22] Glenn Browne, et al. *An Empirical Investigation of User Requirement Elicitation: Comparing the Effectiveness of Prompting Techniques*

[23] Ritu Agarwal, et al. *Knowledge Aquisition Using Structured Interviewing: An Empirical Investigation.*

[24] Tino Bech-Larsen, Niels Asger Nielsen, 1999. *A comparison of five elicitation techniques for elicitation of attributes of low involvement products*. Journal of Economic Psychology 20.

[25] Einar Breivik, Magne Supphellen, 2003. *Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques*. Journal of Economic Psychology 24.

[26] Fowles, Salas, et al. *The Utility of Event-Based Knowledge Elicitation*

[27] Jones, Miles et. al. *The use of a prototype system for evaluating knowledge elicitation techniques*

[28] Corbridge, C., et al. *Laddering: Technique and tool use in knowledge acquisition*

[29] Tor J. Larser, et al. *An experimental comparison of abstract and concrete representations in system analysis*

[30] George M. Marakas, et al. *Semantic Structuring in Analyst Acquisition and Representation of Facts in Requirements Analysis*

[31] Anna L. Rowe, Nancy J. Cooke ,et al. *Toward an On-Line Knowledge Assessment Methodology: Building on the Relationship Between Knowing and Doing*

[32] Robert W. Zmud. *The Use of mental Imagery to Facilitate Information Identification in Requirement Analysis*

[33] Dag I.K. Sjoberg, et al. *A Survey of Controlled Experiments in Software Engineering*

[34] Tore Dyba, et al. *A systematic review of statistical power in software engineering experiments*

[35] Abbie Griffin. *The Voice of the Customer*

[36] Jonathan Alan Silver, et al. *Understanding Customer's Needs: A Systematic Approach to the "Voice of the Customer"*.

[37] Mitzi G. Pitts, Glenn J. Browne. *Stopping Behavior of Systems Analysts During Information Requirements Elicitation*

7.2 BIBLIOGRAFÍA CONSULTADA

Abdi, H, 2007. *"Coefficients of correlation, alienation and determination."*, in N.J. Salkind (ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage.

Barsalou, L. W., 1985. *Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories*. Journal of Experimental Psychology Learning, Memory, and Cognition, 11, 629–654.

Beizer, B. 1990. *Software Testing Techniques*. International Thomson Computer Press.

Bieman, J. y Schultz, J., 1992. *An Empirical Evaluation (and specification) of the All du-paths Testing Criterion*. Software Engineering Journal. Pages 43-51, January.

Davis, A.; Dieste o.; Hickey, A.; Juristo, N.; Moreno, A.; 2006; *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*; 14th IEEE International Requirements Engineering Conference (RE'06) pp. 179-188

García, R., 2004. *Inferencia Estadísitica y Diseño de Experimentos*. Eudeba. Buenos Aires.

Jerry Zeyu Gao, H.-S. Jacob Tsao y Ye Wu, 2003. *Testing and Quality Assurance for Component-Based Software*. Artech House.

Juristo N., INCO. *Módulo V: TÉCNICAS de INGENIERÍA DEL CONOCIMIENTO*. Unidad 23 del material para el Magíster en Ingeniería del Software.

Juristo N. y Moreno A., 2001. *Basics of Software Engineering Experimentation*. Kluwer Academic Publisher. Dordrecht.

Juristo N., Moreno A. y Vegas S. 2003. *Limitations of Empirical Testing Technique Knowledge*. Lecture notes on empirical software engineering archive, pages 1-38. World Scientific Publishing Co., Inc. River Edge, NJ, USA.

Kernighan, B. W. and Plauger, P. J., Software Tools in Pascal. Reading, MA: Addison-Wesley, 1981

Ledesma, D. 1980. *Estadística Médica*. Eudeba. Buenos Aires.

Monthra, 1987. *The Mothra Software Testing Environment User's Manual*. Technical Report SERC-TR-4-P, Software Engineering Research Center, Purdue University.

Montgomery, D. y Runger, G. 2002. *Probabilidad y Estadística*. Limusa Wiley. Mexico DF.

Myers, G. 1979. *The Art of Software Testing*. Wiley-interscience.

Pfleeger S.L.. *Albert Einstein and Empirical Software Engineering*. Computer. 1999: Octubre: 32-37.

Shimeall, T. J. and Leveson, N. G., "*An empirical comparison of software fault tolerance and fault elimination*," in Proc. Second Workshop Software Testing, Verification, and Analysis, Banff, Alta., Canada, July 1988, pp. 180-187.

Weyuker, E. J., *The complexity of data flow criteria for test data selection*, Inform. Processing Left., vol. 19, no. 2, pp. 103-109, Aug. 1984.

Wohlin C., Runeson P., Höst M., et al., 2000. *Experimentation in Software Engineering. An Introduction*. Kluwer Academic Publishers.

ANEXO

I. REVISIÓN DE TÉCNICAS DE ANÁLISIS ESTADÍSTICO

I.1. ESTIMADOR DE MÍNIMOS CUADRADOS

El estimador de mínimos cuadrados es un método de determinar la curva que mejor describe la relación entre una variable dependiente y otra independiente, minimizando la suma de los cuadrados de la diferencia entre el valor esperado y el observado. Este método es óptimo cuando se cumplen las condiciones del teorema de Gauss-Markov Abdi, 2007. El teorema de Gauss–Markov dice que en un modelo lineal en el que los errores tienen esperanza igual a cero, son incorrelacionados y de igual varianza; el mejor estimador lineal insesgado de los coeficientes es el de los mínimos cuadrados. De forma más general, el mejor estimador insesgado de cualquier combinación lineal de coeficientes es el estimador de mínimos cuadrados. Los errores no deben estar normalmente distribuidos ni necesitan ser independientes, solamente deben ser incorrelacionados. Tampoco deben estar idénticamente distribuidos, sino que deben ser homocedásticos. Se dice que existe homocedasticidad cuando la varianza de los errores estocásticos de la regresión son los mismos para cada observación.

El *coeficiente de correlación* Wikipedia, comúnmente llamado correlación, permite estimar qué tan bien se ajustan los datos medidos a la recta hallada. La correlación puede tomar valores entre -1 y 1 y cuanto más grande sea el valor absoluto de este parámetro, mejor se ajustan los datos a una recta. Sin embargo, la correlación por sí sola, no es indicador suficiente ya que, como se muestra en la Figura 1, valores iguales de correlación pueden corresponder a puntos que claramente no se encuentran sobre una recta y que tampoco sean lineales. En esta figura se puede observar cómo el método de mínimos cuadrados halla la misma recta y el mismo coeficiente de correlación para los puntos dados.

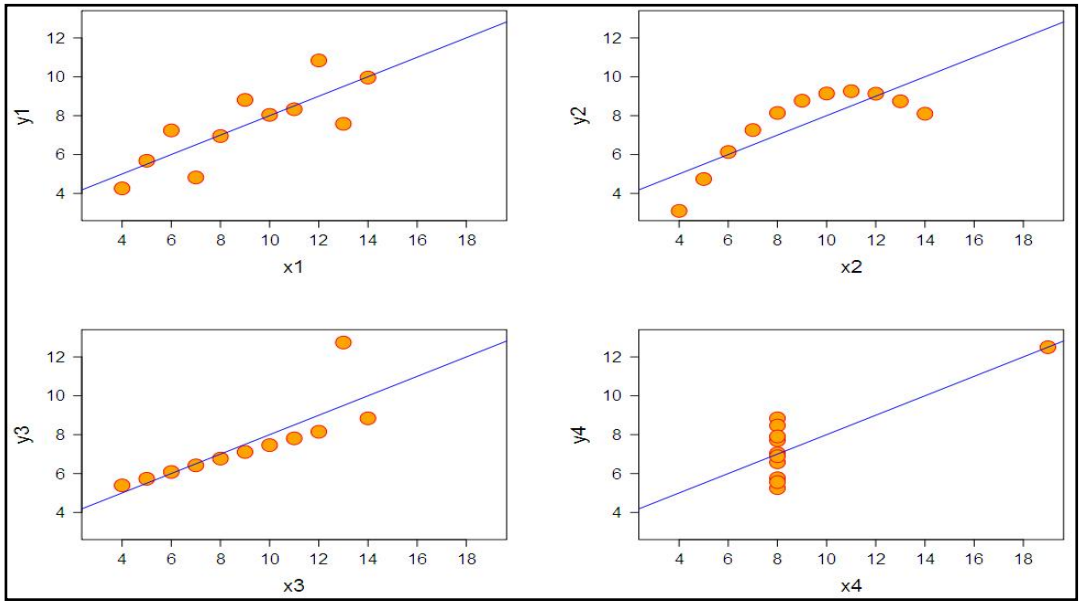

Figura 1. Cuatro conjuntos de datos con el mismo coeficiente de correlación de 0,81

Debido a esto, no es posible basarse solamente en este índice sino que también es necesaria la inspección del gráfico que tiene los puntos medidos y la recta hallada.

I.2. APLICACIÓN DE MÉTODOS PARAMÉTRICOS

I.2.1. Descripción de la prueba ANOVA

Esta técnica se denomina análisis de varianza debido a que el método se basa en observar la variabilidad total de los datos y la variabilidad de acuerdo a diversos componentes. En su forma más sencilla, compara la variabilidad debida al tratamiento dado con respecto a la variabilidad debida a errores aleatorios.

A continuación se describe el método para comparar si un número de muestras tiene el mismo valor de media.

ANOVA, un factor, más de dos tratamientos	
Entrada	***a*** muestras: $x_{11}, x_{12}, ..., x_{1n_1}; x_{21}, x_{22}, ..., x_{2n_2}; ...; x_{a1}, x_{a2}, ..., x_{an_a}$
H_0	$\mu_{x_1} = \mu_{x_2} = ... = \mu_{x_n}$;es decir, todas las medias son iguales
Cálculos	Calcular

ANOVA, un factor, más de dos tratamientos	
	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij}^2 - \frac{x_{\bullet\bullet}}{N})^2$ $SS_{Tratamiento} = \sum_{i=1}^{a}(\frac{x_{i\bullet}^2}{N} - \frac{x_{\bullet\bullet}^2}{N})^2$ $SS_{Error} = SS_T - SS_{Treatment}$ $MS_{Treatment} = SS_{Treatment}/(a-1)$ $MS_{Error} = SS_{Error}/(N-a)$ $F_0 = MS_{Treatment} / MS_{Error}$ donde N es el número total de mediciones y el punto como subíndice indica sumatoria sobre ese índice: $x_{i\bullet} = \sum_j x_{ij}$
Criterios	Se rechaza H_0 si $F_0 > F_{\alpha,a-1,N-a}$ donde F_{α,f_1,f_2} es la distribución F con f_1 y f_2 grados de libertad.

I.3. APLICACIÓN DE MÉTODOS NO PARAMÉTRICOS

I.3.1. Descripción del Test de Mann-Whitney U o U-Test

El U-Test se aplica cuando hay solamente dos alternativas para la variable independiente Para efectuarlo, se ordenan las observaciones y_{ij} en orden ascendente y se reemplazan por sus rangos R_{ij}, donde el valor de rango 1 se le asigna a la observación más pequeña. En el caso en que haya un empate, se promedia el valor de los rangos para las observaciones empatadas.

Sean R_1 y R_2 la suma de los rangos para cada alternativa; y sean N_1 y N_2, las replicaciones de cada alternativa; entonces la estadística del test estará dada por la siguiente fórmula:

(1) $$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1$$

donde la distribución de U es simétrica, y su media y varianza están dadas por:

(2) $$\mu_U = \frac{N_1N_2}{2}$$

(3) $$\sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12}$$

Si N_1 y N_2 son ambas mayores que 7, entonces la distribución de U es aproximadamente normal, de forma que:

(4) $$z = \frac{U - \mu_U}{\sigma_U}$$

está normalmente distribuida con media 0 y varianza 1.

I.3.2. Descripción del Test de Kruskal-Wallis o H-Test

En el caso en que se tengan más de dos alternativas se debe aplicar el H-Test. Primero se deben ordenar las muestras y asignarles un determinado rango al igual que se hace en el U-Test. Sea R_i la suma de los rangos de la observación de la i-ésima técnica, la estadística del H-Test es:

(5) $$H = \frac{1}{S^2}\left(\sum_{i=1}^{a}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

donde a es la cantidad de técnicas a comparar, N es la cantidad total de mediciones, n_i es la cantidad de mediciones por cada método y la varianza de los rangos S^2 está dada por la siguiente fórmula:

(6) $$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{a}\sum_{j=1}^{n_i}R_{ij}^2 - \frac{N(N+1)^2}{4}\right)$$

donde R_{ij} es el rango perteneciente al método i y a la medición j.

Cuando, para todo i, n_i es mayor que 5, H tiene una distribución que se aproxima a $\chi^2_{\alpha,a-1}$, si la hipótesis es cierta. Entonces si $H > \chi^2_{\alpha,a-1}$, la hipótesis debe ser rechazada.