La Importancia del Uso de la Evidencia Empírica en Ingeniería de Software

Enrique Fernández[1], Oscar Dieste[2], Patricia Pesado[3], Ramón García-Martínez[4]

1. Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.
2. Grupo de Ingeniería de Software Experimental. Facultad de Informática. Universidad Politécnica de Madrid.
3. Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP - CIC
4. Grupo de Investigación en Sistemas de Información. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.

odieste@fi.upm.es, enriquefernandez@educ.ar, ppesado@unlp.edu.ar, rgarcia@.unla.edu.ar

1. Introducción

Un caso de estudio [1] propuesto por la literatura en Ingeniería de Software Empírica describe a un Dr. Smith, que investiga sobre técnicas de testing en la universidad. Recientemente a desarrollado una nueva técnica de inspección de código que, a priori, parece que mejora el desempeño de otras técnicas como por ejemplo la técnica basada en perspectivas. Por tal motivo, decide realizar un estudio empírico que le permita validar esta hipótesis. Para ello hace un llamado a estudiantes de los últimos años de la carrera de Ingeniería en Sistemas para que participen en el proyecto. Como resultado de la convocatoria consigue reclutar a 16 estudiantes, los cuales son entrenados 8 en la nueva técnica y 8 en la técnica basada en perspectivas. Con posterioridad, cada grupo aplica la técnica de inspección correspondiente al mismo programa, registrando el número de defectos encontrados. Los resultados obtenidos (ya agregados por grupo) son los que se indican en la tabla 1:

Nueva técnica	Técnica basada en perspectivas
Media (M_e) = 12.000 defectos Desvío Estándar (S_e) = 2.673	Media (M_c) = 11.125 defectos Desvío Estándar (S_c) = 2.800

Tabla 1. Resultados del estudio experimental del Dr. Smith

En base a estos valores, el Dr Smith, realiza un contraste de hipótesis (un t-test suponiendo varianzas iguales, con $\alpha = 0.05$. Dicho test arroja un p-value de 0.53, por tanto no puede asegurarse que el nuevo método mejore el desempeño del método preexistente. El resultado del experimento desilusiona profundamente al Dr. Smith, pero como necesita imperiosamente publicar para la renovación de su contrato de investigación, escribe un artículo que envía a la International Conference on

Empirical Software Engineering (ICESE). Al finalizar el proceso de revisión, el Dr. Smith, recibió la siguiente evaluación:

Originalidad:	*Accept*
Importancia:	*Strong Reject*
Valoración Geenral:	*Reject*
Comentarios:	*Su trabajo es interesante pero tiene dos grandes falencias, en primer lugar ha sido desarrollado con muy pocos sujetos experimentales (que, además, no son profesionales). En segundo lugar, los resultados de los estudios son no significativos, por lo cual no aporta información relevante para los profesionales del área.*

Figura 1. Resultados de la evaluación del trabajo

El ejemplo anterior, aunque ficticio, es representativo de muchas investigaciones reales en Ingeniería del Software (IS) empírica. Por una parte, muchos investigadores interpretan de un modo excesivamente restrictivo (e incorrecto en muchos casos, como se comprobará más adelante) los contrastes de hipótesis, focalizándose únicamente en si son o no significativas a nivel $\alpha = 0.05$. Del mismo modo, desarrolladores de la industria renuncian a tomar como evidencia estudios experimentales que fueron construidos con estudiantes, al considerar que estos trabajos no pueden extrapolarse a entornos reales. Sin embargo el problema que se presenta es que existe escasez de profesionales o estudiantes avanzados dispuestos a participar de trabajos experimentales a un costo accesible por los investigadores de las universidades o empresas (ver los trabajos de revisión [2][3][4][5]). Además, dependiendo del caso bajo estudio, muchas veces es necesario contar con un nivel de infraestructura de altos costos, lo cual restringe adicionalmente las posibilidades de experimentación. Estos factores implican una alta limitante para que los investigadores de IS puedan generar conocimiento validado empíricamente.

Afortunadamente existen algunas alternativas que permiten aprovechar los resultados de estudios con pocos sujetos que arrojan diferencias no significativas. En este trabajo nos centraremos en uno de ellos: el meta-análisis. En esencia, el meta-análisis es una técnica estadística que permite acumular (agregar) varios estudios, aumentando, de esta forma, el número de sujetos experimentales que intervienen en el contraste de hipótesis, mejorando de este modo su potencia. Procederemos del siguiente modo: En la sección 2 describiremos como afecta el tamaño de la muestra a los test de hipotesis; en la sección 3 se presenta una recopilación de cómo aprovechar los experimentos hechos con pocos sujetos experimentales; en la sección 4 se presenta un conjunto de problemas vinculados al meta-análisis; y en la sección 5 se presentan las conclusiones.

2. Estado de la Cuestión

Es bien conocido que cualquier test estadístico está sometido a dos tipos de errores: α, o error de tipo I, y β, o error de tipo II. Dichos errores se producen por la incertidumbre asociada a estimar parámetros (medias y desvío típico) de una población (por ejemplo la técnica de inspección identificada por el Dr. Smith) a partir de una muestra de la misma (los sujetos que han ensayado las técnicas). Tal y como indica la tabla 2, α es el error asociado a aceptar la hipótesis alternativa (H_1) cuando en la población se verifica la hipótesis nula (H_0); y β es la probabilidad asociada al evento justamente inverso.

	H_0	H_1
H_0	Correct decision $(1-\alpha)$	β (Tipe II error)
H_1	α (Tipe I error)	Correct decision $(1-\beta)$

Tabla 2. Tipos de error de un test estadístico

Para un investigador, el error más importante es α. La razón es muy sencilla: todos nosotros intentamos desarrollar nuevos métodos y técnicas que hagan mas eficiente el desarrollo del software. Pero necesitamos demostrar que dichos métodos y técnicas son efectivamente mejores, razón por la cual acudimos a la realización de experimentos. Esperando que H_1 sea cierta. Esta convicción la habremos obtenido de muchos modos: students' assignments, casos de estudio o simplemente cabezonería (los tenure track tienen bastante que ver con esto). Por este motivo, tras realizar el experimento, y presentar nuestros datos, queremos que todo el mundo acepte nuestras conclusiones. Si el test arroja que H_1 es cierto (o sea, acontece la segunda fila de la tabla 2 que es lo que realmente queremos) deseamos que el error del test sea el menor posible. Por ello, el valor de α se fija en valores muy pequeños, tales como: 0.1, 0.05 o incluso 0.01 (10%, 5% y 1% respectivamente).

Ahora bien que ocurre si el test arroja que H_0 es cierto, tal y como le ha ocurrido al Dr. Smith (y a nosotros, muchas –demasiadas- veces)? Por mucho que nos pese (ya que el resultado es contrario a nuestras expectativas) deberemos asegurarnos de que el error (β) sea el menor posible o, lo que es equivalente, que es muy probable que H_0 sea cierto en la población (o sea, la técnica del Dr. Smith no mejora la del Dr. Basili, cosa por otra parte esperable).

Lamentablemente α y β no son independientes: Según la teoría estadística, un test de hipótesis, se caracteriza por 5 factores [6]: α, β, la diferencia entre las medias (d), el nivel de variación de la variable respuesta (s) (medido a través de la varianza o la desviación típica) y el número de sujetos experimentales (n) o, dicho con mayor precisión, el tamaño de la muestra. La relación entre estos factores se muestra en la función 1 (z representa comúnmente la distribución normal tipificada):

$$z_{1-\beta} = \sqrt{\frac{n}{2}}\frac{d}{S} - z_{1-\alpha} \qquad (1)$$

Estos 5 factores forman un sistema cerrado, haciendo que una disminución o incremento en cualquiera de los factores provoque incrementos o disminuciones en los demás factores (con la excepción de d, que permanece siempre fijo, ya que representa la diferencia real entre los métodos o técnicas comparados en un experimento). Por ello, los libros estadísticos recomiendan fijar, en primer lugar, los errores de tipo I (α) y II (β), y en base a estos factores y la diferencia de medias y varianza (factores propios del contexto experimental, y que no pueden ser manipulados por el investigador), establecer el tamaño de la muestra. Este es uno de los motivos (no el único, pero si uno de los más importantes) por el cual se exige que los experimentos posean un alto número de sujetos experimentales. En caso contrario, y suponiendo que α sea fijado en 0.05, β tiene habitualmente valores muy altos. Volviendo al ejemplo de la sección 1, aplicando la función 1 obtenemos un $\beta = 0.83$, es decir, el test detectará diferencias significativas en el 17% de las veces y por el contrario no será capaz de identificarlas el 83% de las veces a pesar de que estas puedan existir. La influencia del número de sujetos en el error β es más clara todavía cuando se muestra como disminuye el error de tipo II al test aplicado por el Dr. Smith a medida que se incorporan sujetos experimentales y conservando fijos los otros factores (diferencia de medias, varianza y error de tipo I), como se ve en la figura 2.

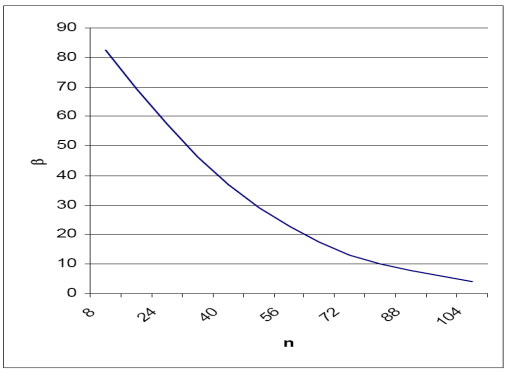

Figura 2. Reducción del error Tipo II

En la literatura es bastante habitual utilizar el término "potencia estadística" en lugar de β para referirse al error de tipo II. Sencillamente, la potencia se estima como $1 - \beta$. Para que un experimento sea considerado fiable, es habitual fijar los valores $\alpha = 0.05$ (esto es, una fiabilidad de 0.95 o del 95%) y $\beta = 0.2$ (esto es, una potencia de 0.8 o del 80%).

Afortunadamente, en los últimos años se han desarrollado una serie de estrategias que permiten paliar los problemas de baja potencia de los test de hipótesis producida por la baja cantidad de sujetos experimentales utilizados. Particularmente en este trabajo presentaremos al meta-análisis. El meta-análisis es una estrategia que permite

combinar los resultados de varios experimentos previamente desarrollados aumentando el número de sujetos totales y el poder estadístico del contraste. Esto será discutido en las siguientes secciones.

3. Como Aprovechar los Experimentos con Pocos Sujetos

Aunque habitualmente el término meta-análisis se asocia a la medicina, su desarrollo tal y como ahora lo conocemos tuvo lugar en la psicología, por razones bastante similares a los problemas experimentados por el Dr. Albert [7]. Las manifestaciones externas eran no obstante distintas. Es bien conocido que la psicología tiene una antigua tradición experimental. Lamentablemente, en muchos casos los tratamientos estudiados (por ejemplo psicoterapia) tienen efectos muy pequeños en los pacientes, por lo que, en concordancia con lo indicado en la sección 2, los experimentos necesitan un número de sujetos experimentales muy elevado (las recomendaciones usuales están en torno de los 150 [8]). A medida que se acumulan estudios de bajo poder estadístico, predominan aquellos que reportan efectos no significativos (como sucede en el caso del Dr. Smith) frente a los que si detectan efectos significativos. Observados en conjunto, parecería obvio que un elevado número de experimentos no significativos sugeriría que los tratamientos no poseen efecto alguno. Pero, sabemos que esta impresión puede ser falsa [8], por ello no debería analizarse los resultados de los estudios individualmente, sino, que es necesario contar con una estrategia que permita mirarlos como parte de un experimento mayor.

El meta-análisis, término acuñado por primera vez por Gene V. Glass [9], consiste en la agregación de los resultados de un conjunto de experimentos que analizan el desempeño de un par de tratamientos con el fin de dar una estimación cuantitativa sintética de todos los estudios disponibles [10]. Esencialmente la estrategia consiste en promediar los resultados de los experimentos individuales (para conocer los detalles técnicos, consúltese el Apéndice de este artículo). El valor así obtenido representa el efecto que teóricamente se habría obtenido en un experimento singular con un tamaño de muestra mayor que el de los experimentos individuales y que, por lo tanto, cuando se acumulan varios experimentos posee un menor error tipo II que cualquiera de ellos (los valores exactos los indicaremos inmediatamente a continuación).

Cabe aclarar que si todos los estudios fueran igualmente fiables, bastaría con un simple promedio para obtener una conclusión final. El problema es que no todos los estudios son igualmente fiables. Cuanto más grande sea el tamaño de la muestra, y por ende menor el nivel de variación, menos influencia del azar habrá en el resultado del estudio. Para lograr esto se debe, en primer, lugar compatibilizar los resultados de los distintos experimentos, pera ello se utiliza la métrica “tamaño de efecto”, el cual es una medida no escalar que se obtiene como la diferencia de la medias de los tratamientos dividido por el desvío estándar conjunto. Una vez obtenido el tamaño de efecto para cada experimento, el cual nos indica que nivel de diferencia existe entre las medias de ambos tratamientos (esto en general se interpreta como: 0.2 = baja, 0.5 = media y 0.8 = alta). En segundo lugar, se deben combinar los efectos de los distintos experimentos mediante un promedio ponderado (siendo el método más

conocido el denominado "diferencia de medias ponderadas" (WMD)), donde cada uno de los experimentos es ponderado por la inversa de su varianza haciendo de esta forma que el resultado de los estudios más fiables afecte mas a la conclusión final que el resultado de los estudios menos fiables. De nuevo, para mayores detalles de acerca de formulas concretas que aplicar remitirse al Apéndice.

Siguiendo con el ejemplo, supongamos que el Dr. Smith publicó su trabajo en la página Web del laboratorio al cual pertenece, porque, a pesar de que no fue aprobado, él consideraba que la información contenida era en realidad valiosa para la comunidad. Este trabajo fue encontrado por el Dr. Thomas quién lo ve interesante y decide desarrollar una replicación del mismo, porque tiene la impresión de que la nueva técnica es realmente eficiente. En este caso el Dr. Thomas pudo reclutar 8 estudiantes avanzados de la carrera de ingeniería informática (asignando cuatro a cada una de las técnicas). Una vez realizado el experimento, se obtuvieron los siguientes resultados:

Nueva técnica	Técnica basada en perspectivas
Media (M_e) = 13.000 defectos	Media (M_c) = 12.000 defectos
Desvío Estándar (S_e) = 1.800	Desvío Estándar (S_c) = 1.700

Tabla 3. Resultados del estudio experimental del Dr. Thomas

Con estos resultados, el Dr. Thomas, realiza un t-test (suponiendo varianzas iguales con $\alpha = 0.05$) y también obtiene diferencias no significativas (p-value 0.57). Dicho resultado estuvo fuertemente influenciado por la baja potencia del test, que en este caso es de solo el 0.11.

Que pasaría si estos dos estudios se combinaran mediante meta-análisis para obtener un nuevo resultado: Podrían obtenerse diferencias significativas? Mejoraría la potencia del test?. La respuesta a la primer pregunta es no; el tamaño muestral es todavía demasiado reducido para arrojar resultados significativos (solo se cuenta con 12 sujetos por técnica). Tal y como puede verse en la figura 3 (que muestra el poder estadístico del meta-análisis para una población con un tamaño de efecto de 0.5 y α=0.05) se necesitan cerca de 70 sujetos experimentales para que un meta-análisis alcance el poder estadístico considerado habitualmente como discriminante ($1-\beta = 0.8$). Sin embargo, el poder estadístico ha mejorado en parte, miestras los test de los Dr. Smith y Thomas tenían una potencia del 0.17 y 0.11 respectivamente, el meta-análisis alcanza una potencia del 0.13. Obviamente, de haber podido utilizar 8 + 4 = 12 sujetos por grupo en un único experimento, habría sido posible conseguir un poder de 0.22 (ver figura 2). Ahora bien, el meta-análisis permite ir incrementando paulatinamente la potencia estadística a medida que se van incorporando experimentos, permitiendo que los trabajos hechos con pocos sujetos experimentales y, en general, con bajos costos pueden complementarse, permitiendo de esta forma una suma de esfuerzos, por ello su importancia.

Una conclusión obvia de la situación anterior es que, a medida que haya más experimentos (aunque le número de sujetos sea reducido), su agregación mediante meta-análisis producirá mayores aumentos de potencia y, por lo tanto, la posibilidad de detectar falsos negativos.

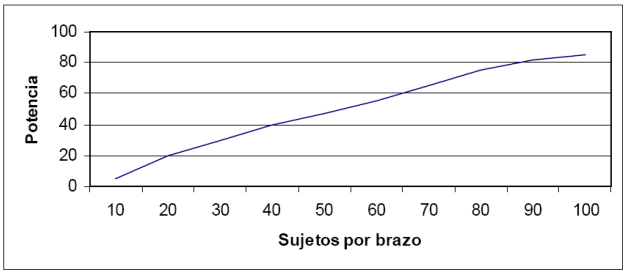

Figura 3. Incremento de la potencia estadística en un meta-análisis

A modo de ejemplo, supóngase la existencia de tres replicaciones mas del trabajo del Dr. Smith, cuyos resultados se muestran en la tabla 4, indicando con: N‘s el número de sujetos utilizados, M‘s las medias y S‘s los desvío estándar; y con (e) a los resultados vinculado a la nueva técnica y con (c) a los resultados vinculados a la técnica de Basili (nótese que todos ellos aportan resultados no significativos):

Estudio	Ne	Me	Se	Nc	Mc	Sc	p-value	Potencia
3	9.0	11.0	1.8	9.0	10.1	1.7	0.30	17.58
4	10.0	10.0	1.4	10.0	9.1	1.5	0.20	20.34
5	12.0	9.0	1.6	12.0	8.1	1.8	0.20	24.63

Tabla 4. Resultados de los estudios identificados

La figura 4 muestra gráficamente como va incrementándose la potencia del meta-análisis con la incorporación de estos estudios.

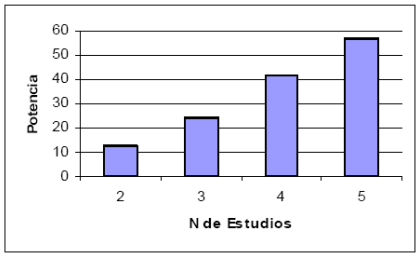

Figura 4. Incremento de la potencia estadística del MA

Es importante destacar que a pesar de que no se ha alcanzado el nivel de potencia deseado del 80%, igualmente con una potencia de casi 57% (que es muy superior a la que ofrece el mejor experimento individual, que alcanza el 24%) se ha llegado a lograr que el test permita que las diferencias se conviertan en significativas. Y este es un hecho notable, ya que el ejemplo se ha diseñado partiendo de la base de que la eficiencia de las técnicas de testing son distintas (Me ≅ 9, Mc ≅ 8 y S ≅ 2), aunque no muy diferentes.

Hemos mostrado hasta ahora cómo el meta-análisis puede utilizarse para aprovechar estudios con pocos sujetos experimentales, que arrojan aisladamente resultados no significativos pero que, conjuntamente, pueden proporcionar valiosas evidencias.

4. Limitaciones del Modelo Teórico

Si bien las funciones de estimación de la potencia de un meta-análisis permite estimar con precisión la potencia estadística de un meta-análisis cuando el conjunto de experimentos que hacen parte del proceso de agregación son homogéneos [11] (las diferencias entre los resultados de los distintos experimentos son mínimas), según nuestra óptica, esta función, posee falencias para poder ser aplicada en el actual contexto de la IS. Dicho problema radica en que como los estudios son pequeños, las variaciones en los resultados, en general, son grandes por influencias del error experimental. Cuando esto sucede la potencia del meta-análisis tiende a decaer como lo indican soslayadamente Hedges y Olkin [11]. Este hecho lo hemos analizado mediante una prueba de MonteCarlo ("Submited to International Journal of Empirical Software Engineering"), en la cual simulamos resultados de estudios pertenecientes a una misma población, pero sin forzar homogeneidad entre los grupos a agregar, variando el número de sujetos por experimentos entre 4 y 20 y combinando entre 2 y 10 experimentos por meta-análisis. Este estudio nos permitió determinar, cuando no se posee homogeneidad, que para tamaños de efectos de bajo (0.2) es casi imposible conseguir que el test arroje diferencia significativas trabajando con 200 sujetos por brazo (tamaño máximo de muestra simulada) debido a la baja potencia estadística; que para efectos medios y altos (0.5 y 0.8) la heterogeneidad no están condicionante ya que se ha podido alcanzar una buena potencia estadística con cantidades de sujetos no demasiado elevadas (aproximadamente 80 y 30 sujetos por brazo respectivamente). En la figura 5 se presenta una comparativa en la potencia estimada para cada uno de los valores de efecto típicos.

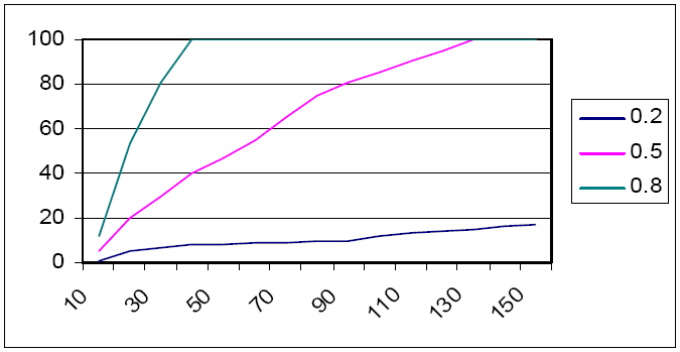

Figura 5. Potencia simulada para un meta-análisis

5. Conclusiones

En el presente trabajo se ha mostrado que existen alternativas para que los investigadores puedan generar piezas de conocimiento empírico de una forma más eficiente de lo que se hace hoy en día. Se ha mostrado la capacidad que tiene el meta-análisis para aumentar la potencia de los experimentos, permitiendo que un conjunto de estudios pequeños que individualmente no arrojan diferencias significativas tomados en conjunto si permiten su identificación. De esta forma pueden solucionarse, en parte, los problemas vinculados a la dificultad de reunir un número apreciable de sujetos experimentales por un único investigador, ya que mediante el meta-análisis de pequeñas replicaciones de estudios, se puede lograr conformar un experimento de gran envergadura.

En resumen, podemos afirmar que: (1) Vale la pena hacer experimentos a pesar de que los mismos no tengan muchos sujetos experimentales, ya que pueden combinarse y conformar un estudio de mayor nivel; (2) Vale la pena publicar estudios a pesar de que los mismos no den resultados significativos, ya que esto muchas veces puede deberse a la falta de potencia del método estadísticos; y (3) Si esta estrategia fuera aplicada a tecnologías realmente importantes en IS (UML o partición de equivalencia por citar algún ejemplo), el esfuerzo combinado de los investigadores permitiría decidir si dichas tecnologías son realmente útiles o no, proporcionando el fundamento necesario para que la construcción de software sea realmente un proceso ingenieril.

6. Bibliografía

[1] Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2; pp. 133–164.

[2] Miller, J; 1999: Can Results from Software Engineering Experiments be Safely Combined? IEEE METRICS, 152-158

[3] N. Juristo, A. Moreno, Basics of Software Engineering Experimentation, Kluwer Academic Publishers, 2001.

[4] Tonella P., Torchiano M., Du Bois B., Systä T.; 2007; *Empirical studies in reverse engineering: state of the art and future trends*; Empir Software Eng 12:551–571.

[5] Dyba, T., Aricholm, E.; Sjoberg, D.; Hannay J.; Shull, F.; 2007; Are two heads better than one? On the effectiveness of pair programming. IEEE Software;12-15.

[6] Cohen, J.; *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.) 1988. ISBN 0-8058-0283-5.

[7] Gurevitch, J. and Hedges, L.; 2001; *Meta-analysis: Combining results of independent experiments.* Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), pp. 347–369. Oxford University Press, Oxford.

[8] Fisher RA (1925). *Statistical Methods for Research Workers* (first ed.). Edinburgh: Oliver & Boyd.

[9] Glass, G; 1976; Primary, secondary, and meta-analysis of research. Educational Researcher 5: 3-8

[10] Cochrane; 2008; Curso Avanzado de Revisiones Sistemáticas; www.cochrane.es/?q=es/node/198

[11] Hedges, L.; Olkin, I.; 1985; Statistical methods for meta-analysis. Academic Press