JOURNAL OF move AND THERAPEUTIC SCIENCE

2023; 5(1):m41

El Cálculo del Tamaño Muestral en Ciencias de la Salud: Recomendaciones y Guía Práctica

Calculating the Sample Size in Health Sciences: Recommendations and Practical Guide

Rubén Fernández Matías[1]

1. Unidad de Investigación, Servicio de Fisioterapia y Rehabilitación, Hospital Universitario Fundación Alcorcón, Madrid, España

Correspondencia:

Rubén Fernández Matías, FT, MSc.
Hospital Universitario Fundación Alcorcón
Servicio de Fisioterapia, Calle Budapest 1,
28922, Alcorcón, Madrid, España
Teléfono: 916219727
E-mail: ruben.fernanmat@gmail.com

Conflicto de Intereses:

Los autores declaran no tener ningún conflicto de intereses. Este proyecto no ha sido presentado en ningún evento científico

Financiación:

Los autores declaran no haber recibido financiación/compensación para el desarrollo de esta investigación.

DOI: 10.37382/jomts.v5i1.915

Recepción del Manuscrito:

28-Mayo-2023

Aceptación del Manuscrito:

24-Julio-2023

RESUMEN

El cálculo de tamaño muestral es uno de los aspectos más importantes en la planificación de la mayoría de las investigaciones, pudiendo derivar una muestra insuficiente a una inutilidad de la investigación en sí misma. Tradicionalmente se han utilizado los cálculos de tamaño muestral basados en potencia, pero actualmente se han empezado implementar los cálculos basados en precisión. En el presente escrito se presentan una serie de recomendaciones para cálculos para ensayos clínicos aleatorizados, modelos de regresión lineal y logística múltiples, análisis de reproducibilidad y de modelos predictivos multivariables, junto con algunos ejemplos prácticos de su implementación, así como algunas consideraciones con respecto a realización y utilización de datos de estudios piloto a la hora de planificar un cálculo de tamaño muestral.

Palabras clave: Tamaño de la Muestra; Estadística; Metodología.

ABSTRACT

Sample size calculation is one of the most important aspects in the planning of most research, and an insufficient sample can lead to the uselessness of the research itself. Traditionally, power-based sample size calculations have been used, but now precision-based calculations have begun to be implemented. This paper presents recommendations for calculations for randomised clinical trials, multiple linear and logistic regression models, reproducibility analysis, and multivariable predictive models, along with some practical examples of their implementation, as well as some considerations regarding the development and use of pilot study data when planning a sample size calculation.

Keywords: Sample Size; Statistics; Methodology

INTRODUCCIÓN

El cálculo de tamaño muestral es uno de los aspectos más importantes en la planificación de la mayoría de las investigaciones. Resulta inadecuado, tanto desde un punto de vista metodológico como ético, realizar investigaciones con una muestra insuficiente, así como con una muestra excesiva para los objetivos planteados.("World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects," 2013) Una muestra insuficiente puede hacer que una investigación no tenga utilidad alguna, y una muestra excesiva suponer un gasto no justificado de recursos materiales y humanos, exponiendo a un mayor conjunto personas a posibles efectos adversos innecesarios, sin embargo, este segundo problema no es tan habitual.

Las muestras pequeñas presentan una situación paradójica. En ellas es donde más falta hacen ciertas correcciones en múltiples análisis estadísticos, pero donde menos útiles resultan tales correcciones. Por ejemplo, las muestras pequeñas se ven más influenciadas por la presencia de valores perdidos, sin embargo, los métodos de manejo de valores perdidos como la imputación múltiple por ecuaciones encadenadas, requieren de muestras grandes para poder implementarse adecuadamente, siendo poco eficientes en muestras pequeñas.(Barnes et al., 2006) Del mismo modo, las muestras pequeñas tienden a producir más sobreajuste de modelos multivariables, sobreestimando la variabilidad explicada por el modelo. Este sobreajuste puede corregirse en parte con métodos de penalización (*shrinkage)*, sin embargo, dichos métodos no funcionan bien cuando hay mucho sobreajuste (más habitual en muestras pequeñas) y cuando el tamaño muestral es pequeño.(Riley et al., 2020)

Distintas guías de reporte, tales como la CONSORT (Schulz et al., 2010) (ensayos clínicos), STROBE (Vandenbroucke et al., 2007) (estudios observacionales), STARD (J. F. Cohen et al., 2016) (estudios de agudeza diagnóstica), GRRAS (Kottner et al., 2011) (estudios de fiabilidad) y TRIPOD (Collins et al., 2015) (estudios de modelos predictivos) incluyen un apartado sobre el cálculo del tamaño muestral. Sin embargo, a pesar de su importancia, el reporte del cálculo del tamaño muestral sigue siendo pobre dentro de múltiples campos de ciencias de la salud, entre ellos el de rehabilitación, con un elevado porcentaje de estudios con reporte inadecuado que imposibilita su replicación, o sin un cálculo previo realizado.(Arienti et al., 2021; Copsey et al., 2018; Gonzalez et al., 2018) Además, los estudios realizados en el campo de la rehabilitación suelen presentar tamaños muestrales que podrían considerarse pequeños, por ejemplo, con medianas de 73 (primer y tercer cuartiles, 50-120) (Copsey et al., 2018) y 60 (primer y tercer cuartiles, 34-109) (Gonzalez et al., 2018) en dos revisiones publicadas sobre ensayos aleatorizados.

El objetivo de este escrito es presentar múltiples recomendaciones de buenas prácticas para el cálculo del tamaño muestral, así como facilitar una guía para su realización con distintos análisis estadísticos. En el Material Suplementario 1 se recogen ejemplos prácticos de todos los cálculos presentados, y en el Material Suplementario 2 se presenta un recopilatorio de herramientas de cálculo de tamaño muestral con sus respectivas guías.

MASTERCLASS

Tipos de Cálculo de Tamaño Muestral: Potencia y Precisión

Principalmente pueden distinguirse dos métodos de cálculo de tamaño muestral en la literatura, los basados en potencia y los basados en precisión.(J. M. Bland, 2009)

Los cálculos basados en potencia se basan en el establecimiento a priori, de una probabilidad esperada de obtener un resultado estadísticamente significativo, partiendo de unas determinadas asunciones sobre la distribución poblacional de los datos.(J. M. Bland, 2009) Por ejemplo, si estamos haciendo una comparación entre dos grupos con una t-Student, y sabemos que la diferencia real entre las dos poblaciones es de 10 puntos, con una desviación estándar en ambas de 15 puntos y una distribución normal. Si asumimos como estadísticamente significativo un valor de $p < 0,05$, necesitaríamos una muestra de 37 sujetos por grupo, para que, si realizamos infinitos estudios con muestreos

probabilísticos, en un 80% de ellos obtengamos un valor de $p < 0,05$.

Por tanto, los cálculos basados en potencia tienen una relación directa con la interpretación dicotómica de los resultados de una investigación basados exclusivamente en el valor-p. Este método de cálculo de tamaño muestral tiene como objetivo elaborar investigaciones que respondan a la pregunta de si parece existir o no un "efecto", usando como método de decisión un punto de corte del valor-p (ej. $p < 0,05$).

Por su parte, los cálculos basados en precisión se basan en el establecimiento a priori de la amplitud deseada del intervalo de confianza, partiendo también de unas determinadas asunciones sobre la distribución poblacional de los datos,(J. M. Bland, 2009) es decir, calcular estimaciones precisas del parámetro poblacional deseado. Por tanto, este método tiene como objetivo elaborar investigaciones que respondan a la pregunta de cuál es el efecto deseado, no solo si "existe o no". Por ejemplo, asumiendo los datos del anterior ejemplo, si deseásemos una amplitud esperada del intervalo de confianza al 95% de 5 puntos, necesitaríamos 276 sujetos por grupo. Dicha amplitud implicaría que, si repetimos el estudio infinitas veces con muestreos probabilísticos, entonces en un 95% de dichos estudios, la media observada estaría dentro del intervalo [7,5; 12,5]. Es decir, las estimaciones de la diferencia poblacional a partir de un estudio se alejarían poco del valor real.

La selección de uno u otro método de cálculo de tamaño muestral depende esencialmente del objetivo de la investigación. Si deseamos responder a la pregunta dicotómica de si existe o no un "efecto", basándonos en un punto de corte de valor-p, entonces el cálculo de tamaño muestral basado en potencia puede ser considerado. Si, por el contrario, el objetivo es estimar cual es el valor de un "efecto" o parámetro, entonces es necesario un cálculo basado en precisión.

Desde hace décadas, existe una crítica contundente hacia la interpretación dicotómica de los resultados de investigación basada exclusivamente en un punto de corte del valor-p.(Wasserstein & Lazar, 2016) Ya en 1986, Gardner y Altman recomendaban la preferencia del reporte e interpretación del intervalo de confianza en contraposición al valor-p,(Gardner & Altman, 1986). También cabe destacar la crítica de Jacob Cohen sobre esta temática.(J. Cohen, 1994) Debido a la relación entre los cálculos basados en potencia y la interpretación dicotómica del valor-p, este tipo de cálculos también han sido duramente criticados, recomendando algunos autores en su lugar los cálculos basados en precisión.(J. M. Bland, 2009; Kelley & Maxwell, 2003; Rothman & Greenland, 2018)

Estudios Piloto o de Factibilidad: Qué, para qué y cómo

Un estudio piloto o de factibilidad tiene como objetivo evaluar la viabilidad de realización del futuro estudio de investigación. Normalmente, este tipo de diseños se plantean en relación a los ensayos clínicos, para analizar si la metodología propuesta para la realización de estos es viable, en por ejemplo aspectos como participación/reclutamiento de sujetos, adherencia al tratamiento, tolerancia de los sujetos al mismo, tasa de abandonos, etc.(Eldridge et al., 2016) Sin embargo, este tipo de diseños son más conocidos en el contexto de estimación de parámetros como la diferencia media o la desviación estándar, para poder usarlos para calcular la muestra para el ensayo clínico definitivo.(Bell et al., 2018)

Debido a esta utilidad de estimación de parámetros necesarios para realizar el cálculo de tamaño muestral del ensayo definitivo, existe tendencia en la literatura a olvidar que también es necesario establecer a priori, e incluso calcular, la muestra necesaria para el propio piloto en sí mismo. Este es un punto que se recoge incluso en la extensión de la declaración CONSORT para estudios piloto o de factibilidad.(Eldridge et al., 2016) La razón de ello es poder estimar con cierta precisión dichos parámetros que se usarán para el cálculo del futuro ensayo,(Bell et al., 2018) ya que una inadecuada estimación puede derivar en cálculos erróneos con una pérdida de potencia y/o precisión deseadas para el ensayo definitivo.(Bell et al., 2018; Teare et al., 2014)

Existen varias propuestas de recomendaciones y cálculos de tamaño muestral para estudios piloto.(Cocks & Torgerson, 2013; Teare et al., 2014; Whitehead et al., 2016) Teare y cols.(Teare et al., 2014) recomiendan una muestra mínima de 70 sujetos (35 por grupo) para la estimación de la desviación estándar de una variable continua, y de al menos 120

sujetos (60 por grupo), en el caso de que la variable resultado sea binaria.

¿Qué, para qué y cómo?

Con un estudio piloto se pueden estimar distintos parámetros necesarios para realizar un cálculo de tamaño muestral. En el ámbito de rehabilitación lo más habitual es disponer de variables de resultado continuas en ensayos clínicos, de modo que esté apartado se centrará en las mismas. Existen dos parámetros que pueden ser necesarios para realizar un cálculo de tamaño muestral de una variable continua: la media y la deviación estándar.

La diferencia media tiene una interpretación clínica más sencilla que la desviación estándar, es decir, podemos razonar que consideramos una diferencia grande o pequeña, o una diferencia plausible, para la comparación entre dos intervenciones dadas. Sin embargo, la desviación estándar no tiene esa interpretación clínica directa, no tenemos capacidad de predecir cual puede ser una desviación estándar plausible, es un parámetro que necesita ser estimado.(Bell et al., 2018; Whitehead et al., 2016)

En el año 2018 se publicó la guía DELTA^2, que tiene como objetivo proporcionar una serie de recomendaciones para seleccionar la diferencia esperada u objetivo de un ensayo clínico, a la hora de realizar el cálculo del tamaño muestral.(Cook et al., 2018) Podemos distinguir los siguientes métodos (aunque no los únicos):

1. Mínima diferencia relevante de interés.
2. Magnitud del efecto estandarizado, usando puntos de corte de la *d* de Cohen de 0,2 (tamaño pequeño), 0,5 (tamaño medio) y 0,8 (tamaño grande).
3. Estimación de la diferencia media o el efecto estandarizado esperados (realistas o plausibles) a través de un estudio piloto, o de literatura previa publicada.

La diferencia esperada utilizada en el cálculo debe facilitar una muestra suficiente para encontrar un efecto que sea tanto clínicamente relevante como plausible.(Cook et al., 2018) Por ejemplo, si el conjunto de literatura publicada estudiando un tipo de intervención para una patología, por ejemplo ejercicio terapéutico, hubiera encontrado diferencias medias en una variable resultado de 3 a 7 puntos, aunque una diferencia de 20 puntos sea relevante, no es plausible y no debería utilizarse para el cálculo del tamaño muestral.

Con respecto a la relevancia clínica, hay ciertos aspectos que deben ser tenidos en consideración. Por un lado, es desaconsejable la utilización de la mínima diferencia detectable, extraída de estudios de fiabilidad, como diferencia esperada para los cálculos de tamaño muestral. El motivo es que, una variable resultado con poca fiabilidad, deriva en una mayor variabilidad, disminuyendo por tanto la potencia de un estudio, de modo que se requeriría más muestra. Sin embargo, dicha disminución de fiabilidad producirá también un valor más alto de la mínima diferencia detectable,(Weir, 2005) y a mayor diferencia esperada, menor tamaño muestral calculado. Por tanto, la utilización de la mínima diferencia detectable solo conduce a una infraestimación de la muestra necesaria para realizar una investigación.

Por otro lado, debe tenerse en cuenta que la mínima diferencia clínicamente relevante es un parámetro orientado a mediciones individuales, es decir, un parámetro que intenta estimar que cambio es relevante para un sujeto individual.(Cook et al., 2018) Este parámetro a nivel del individuo no tiene por qué corresponderse con el parámetro de tendencia central poblacional, de modo que no debe ser tenido en cuenta de manera única para definir la diferencia esperada para realizar cálculos de tamaño muestral.

Algunos autores recomiendan establecer siempre la diferencia esperada en base a literatura previa publicada de ensayos clínicos ya realizados (no pilotos) y el conocimiento en la materia del equipo investigador de lo que sería la mínima diferencia relevante y plausible,(Dechartres et al., 2013) más que en la realización de un piloto para su estimación.(Sim, 2019) Esta recomendación se debe en parte, a que los tamaños muestrales pequeños tienden a sobreestimar el tamaño del efecto, pudiéndose infraestimar la muestra necesaria para el ensayo clínico definitivo. Además, aunque hay correcciones propuestas para el sesgo presente en la *d* de Cohen con muestras pequeñas,(Lakens, 2013) dichas correcciones no

corrigen lo suficiente para prevenir tales sobrestimaciones.

Por su parte, la desviación estándar debe ser estimada siempre, sin opciones a establecer la misma por criterio clínico del grupo investigador. Dicha estimación puede venir dada de una revisión de la literatura previa publicada de ensayos clínicos similares al que se pretende realizar, así como de la posible realización de un estudio piloto previo, con una muestra mínima según las recomendaciones de Teare y cols.(Teare et al., 2014)

La desviación estándar sufre también de sesgos en su estimación con muestras pequeñas, que tienden a infraestimar la misma, pudiendo derivar también en una infraestimación del tamaño muestral necesario para el ensayo clínico definitivo.(Vickers, 2003) Existen dos métodos propuestos para tratar de corregir el cálculo de tamaño muestral para este sesgo, el método del límite superior del intervalo de confianza (UCL) propuesto por Browne,(Browne, 1995) y el método de la distribución-t no central (NCT) de Julious y Owen.(Julious & Owen, 2006) El primer método se basa en utilizar el valor del límite superior del intervalo de confianza al X% a una cola de la varianza, recomendando Browne un intervalo al 80% como suficiente.(Browne, 1995) El método NCT se basa en usar un factor de inflación teniendo en cuenta que estamos usando una estimación de la varianza poblacional a partir de una muestra, basándose en una función de distribución acumulada y un parámetro de no centralidad.(Julious & Owen, 2006) En el artículo publicado por Whitehead y cols.(Whitehead et al., 2016) se recogen las fórmulas exactas para realizar ambas correcciones, así como tablas con los factores de corrección ya calculados según ambos métodos para varios tamaños muestrales piloto.

Cálculos Basados en Precisión: Primeros pasos

Mientras que en los cálculos basados en potencia el foco de relevancia clínica se pone en la diferencia media estimada (estimación puntual), en los cálculos basados en precisión ese aspecto de relevancia clínica se pone en la estimación por intervalo, es decir, en el intervalo de confianza al X%.

Supongamos que comparamos dos grupos, observándose una diferencia media entre ambos de 10 puntos con un intervalo de confianza al 95% de [7,5; 12,5]. Podemos distinguir los siguientes conceptos:(Kelley & Maxwell, 2003; Riley et al., 2020)

- Estimación puntual**:** 10 puntos.
- Estimación por intervalo: [7,5; 12,5] puntos.
- Margen de error (MoE): 2,5 puntos (mitad de la amplitud del intervalo de confianza).
- Margen de error multiplicativo (MMOE): 1,25 (1 + margen de error/estimación puntual).

En los cálculos basados en precisión, el foco se pone en el MoE (o el MMOE), ya que esa mitad de la amplitud del intervalo de confianza es la que presenta una interpretación clínica directa.(Kelley & Maxwell, 2003) La pregunta que debemos realizarnos sería, ¿cambiaría la interpretación de los resultados y la implicación clínica de los mismos si la estimación puntual real se alejase ± 2,5 puntos del valor observado? Si la respuesta es no, entonces ese MoE asumido sería adecuado para la investigación.

El MoE debe ser principalmente establecido en función del conocimiento técnico del equipo investigador, en materia de plausibilidad biológica, investigaciones previas realizadas en sujetos con la misma patología y/o intervenciones similares, etc.(Lai & Kelley, 2012) Sin embargo, esto no siempre es tarea fácil, debido a la dificultad de establecer que es un MoE clínicamente relevante para algunos parámetros a estimar, como los coeficientes de regresión estandarizados. Por ello, en algunos casos dicho MoE vendrá definido por recomendaciones previas de expertos sobre puntos de corte prestablecidos.(Kelley & Maxwell, 2003)

Un intervalo de confianza es un intervalo de extremos aleatorios, es decir, para una misma población dada y un mismo tamaño muestral reclutado, los valores de los extremos de dicho intervalo variarán de un muestreo a otro, haciéndolo también como consiguiente la amplitud del mismo y por tanto el MoE.(Kelley & Maxwell, 2003) Beal (S. Beal, 1991; S. L. Beal, 1989) y Grieve (Grieve, 1989, 1991) introducen en la década de los 90 el concepto de potencia asociada a los intervalos de confianza para cálculos de tamaño muestral.

Este procedimiento consiste en tener en cuenta esa variabilidad de la amplitud del intervalo de confianza, y calcular el tamaño muestral no para un MoE esperado, sino para tener una probabilidad de un X% de obtener un MoE igual o inferior al valor esperado en el estudio que se va a realizar.(S. Beal, 1991; S. L. Beal, 1989; Grieve, 1989, 1991) Se recomienda un valor del 80% de probabilidad.(S. Beal, 1991; S. L. Beal, 1989; Grieve, 1989, 1991; Kelley & Maxwell, 2003) Por ejemplo, en el caso anteriormente mencionado para una diferencia media entre dos grupos, con valor esperado (valor al que tiende un parámetro cuando se replica un experimento un número de veces que tiende al infinito y se haya la media de todas las replicaciones) del intervalo de confianza de 5 puntos (276 sujetos por grupo), solamente la mitad de los intervalos obtenidos simulando un experimento 10.000 veces son iguales o inferiores a 5, es decir, hay una potencia del 50% (Figura 1.A). Si quisiéramos una potencia del 80%, tendríamos que elevar la muestra a 292 sujetos por grupo (Figura 1.B).

Por tanto, en los cálculos de tamaño muestral basados en precisión, debemos especificar como mínimo los siguientes parámetros estimados:(Kelley & Maxwell, 2003)

- Nivel de confianza (normalmente 95%).
- Error estándar del parámetro a estimar.
- MoE esperado.
- Probabilidad esperada de obtener un MoE igual o inferior al valor esperado (normalmente 80%).

Un punto importante que debe tenerse en consideración es que, a mayor muestra, mayor

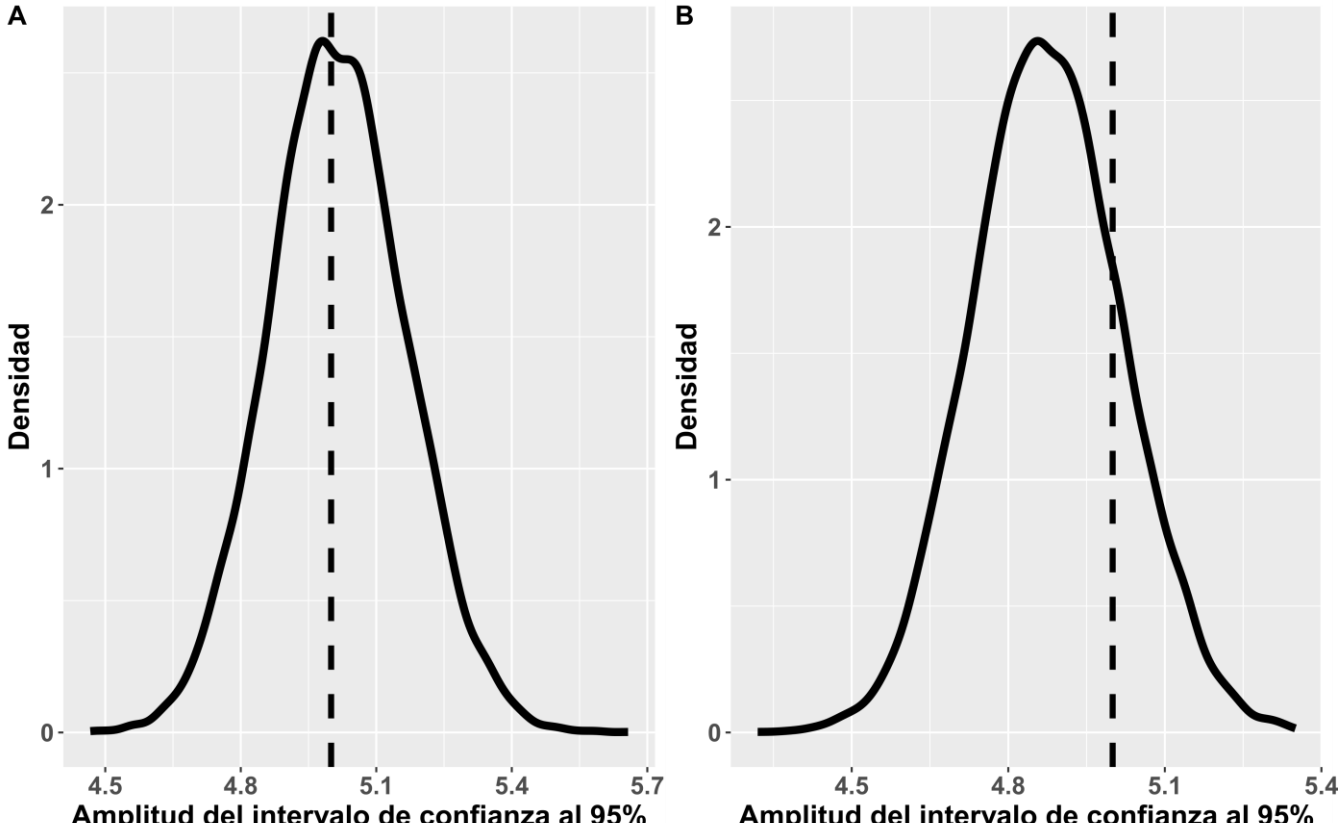


Figura 1. Gráficos de densidad ejemplificando el concepto de potencia estadística para una precisión deseada. En la Figura 1.A se muestra un gráfico de densidad para un intervalo al 95% de confianza esperado de 5 puntos (desviación estándar asumida de 15 puntos), para un tamaño muestral de 276 sujetos por grupo, sin tener en cuenta el concepto de potencia (10.000 simulaciones). En la Figura 1.B se muestra otro gráfico de densidad, con las mismas asunciones, pero para una potencia deseada del 80%, usando un tamaño muestral de 292 sujetos por grupo (10.000 simulaciones). Las líneas verticales discontinuas marcan el valor esperado de 5 puntos.

precisión en la estimación de los parámetros de interés.(Kelley & Maxwell, 2003) Existe cierta preocupación con respecto a la utilización de muestras grandes por un incremento en la tasa de "falsos positivos" o "errores tipo I", sin embargo, esta creencia es errónea. La distribución de los valores-p y, por tanto, el número de veces que se observaría un valor-p considerado como "significativo" (ej. $p < 0,05$), solo depende del tamaño muestral cuando la hipótesis alternativa es cierta, pero no cuando la hipótesis nula lo es, si no fuera así, los test de contraste de hipótesis no tendrían sentido. Por ejemplo, realizando 10.000 simulaciones de Monte Carlo, para una prueba t-Student entre dos muestras independientes siendo la hipótesis nula cierta, el porcentaje de valores-p menores de 0,05 para 30 sujetos por grupo es del 4.95%, y para 3.000 sujetos por grupo del 4.84%. En la Figura 2 se muestra ausencia de relación entre el valor-p y el tamaño muestral para una prueba t-Student para dos muestras independientes cuando la hipótesis nula es cierta, para muestras que oscilan de 30 a 3000 sujetos, con 10 simulaciones de Monte Carlo para cada tamaño muestral (un total de 2.980 simulaciones).

Incrementar el tamaño muestral solo aumentaría la precisión de la estimación de una desviación de la hipótesis nula, ya sea por ejemplo porque un tratamiento tenga un efecto, o porque se hayan

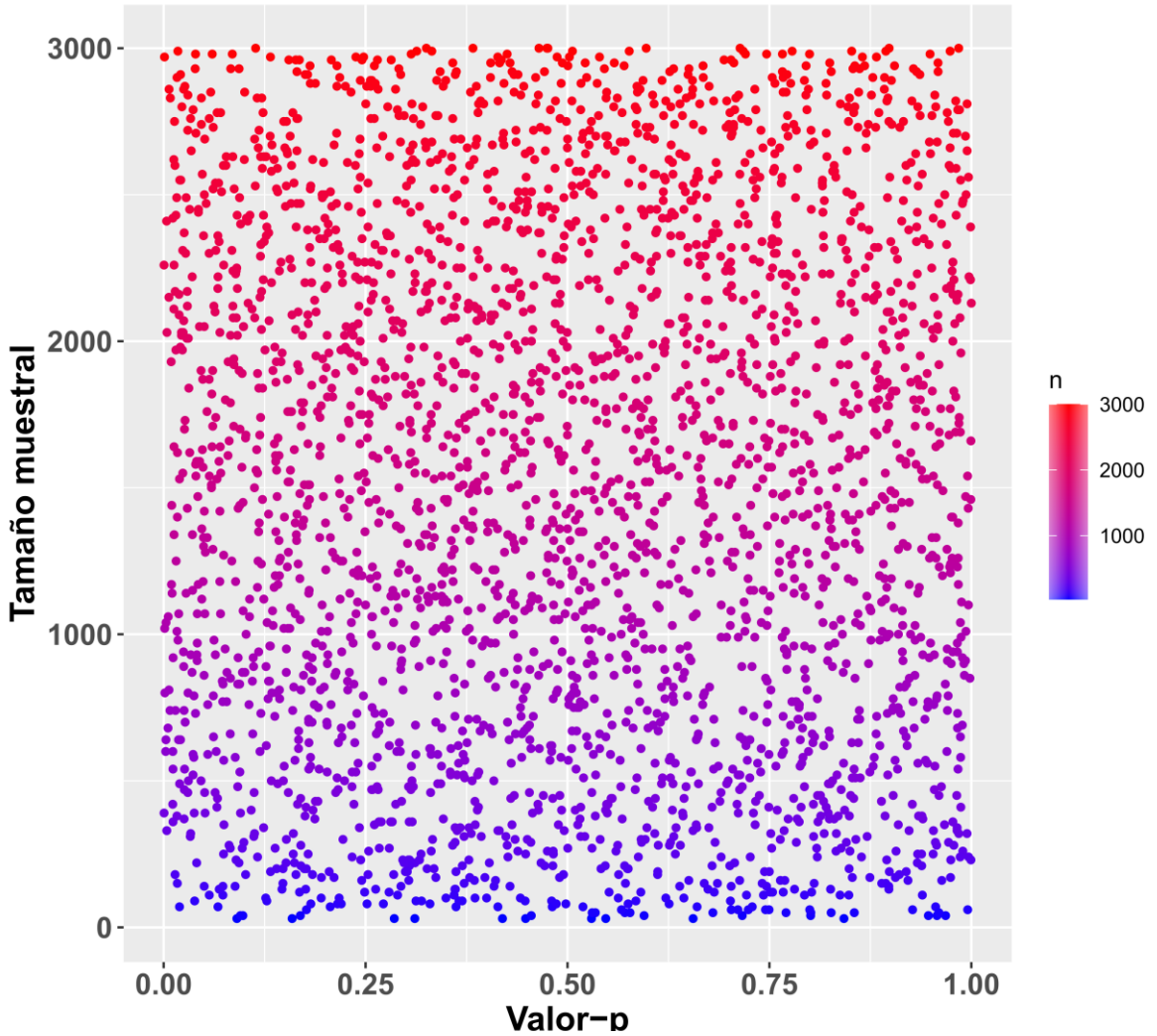


Figura 2. Gráfico de dispersión para la relación entre el valor-p y el tamaño muestral (por grupo) para una prueba t-Student para dos muestras independientes, cuando la hipótesis nula es cierta (2.980 simulaciones). Los colores reflejan el tamaño muestral empleado para la simulación de cada punto del gráfico (azul menor tamaño muestral, rojo mayor tamaño muestral).

introducido errores sistemáticos en el estudio. Por tanto, si la preocupación es obtener un resultado significativo con muestras grandes, que pueda deberse a errores metodológicos en la obtención de los datos y no a un "efecto real", deben tomarse medidas sobre dichos aspectos metodológicos, en lugar de utilizarse muestras pequeñas, que también estarían expuestas a los mismos.

Actualmente existen dos paquetes principales en R para el cálculo de tamaño muestral basado en precisión. El paquete *'presize'*,(Haynes et al., 2021) que permite implementar cálculos basados en una precisión estimada para múltiples análisis estadísticos, y el paquete *'MBESS'*,(Kelley, 2007) que permite implementar dichos cálculos teniendo en cuenta el concepto de potencia de Beal y Grieve.

Recomendaciones para Ensayos Aleatorizados

Habitualmente en el campo de la rehabilitación, se utilizan variables resultado continuas (o que se tratan como continuas) en los ensayos clínicos aleatorizados, de modo que las recomendaciones aquí recogidas se centrarán en este caso concreto.

En un ensayo clínico aleatorizado con una variable continua resultado, dos o más grupos y dos o más momentos de medición, el análisis estadístico recomendado es el análisis de la covarianza (ANCOVA), introduciendo la medición basal como covariable en el modelo. Este tipo de análisis constituyen la especificación matemática correcta del modelo lineal, que optimiza la potencia estadística, en comparación al ANOVA o la comparación bruta de las mediciones post-tratamiento.(Vickers, 2001; Vickers & Altman, 2001) La superioridad del ANCOVA con respecto a la diferencia (post menos pre, es decir, un ANOVA), o la utilización solo de la medición post-tratamiento depende de la correlación entre la medición basal y la post-tratamiento.(Vickers, 2001; Vickers & Altman, 2001) En la Figura 3 se muestra el tamaño muestral necesario para una comparación de dos grupos con dos mediciones (basal y post-tratamiento), en función de la correlación entre la medición basal y la post-tratamiento, para una potencia del 90%, pudiéndose apreciar la superioridad del ANCOVA.

Como se puede apreciar, a mayor correlación entre la medición basal y la post-tratamiento, menor tamaño muestral necesario. Es importante no sobreestimar la correlación estimada al realizar el cálculo de tamaño muestral, ya que esto podría derivar en una disminución de la potencia y/o precisión deseadas en el ensayo clínico. De acuerdo con datos calculados en base a literatura previa publicada, una estimación sensata de correlación sería de 0.5, con un rango de valores plausibles entre 0.4 y 0.6.(Walters et al., 2019)

Por otro lado, al igual que lo comentado en la mejora de la potencia, la utilización de un ANCOVA, para ajustar las comparaciones por tratamiento para las diferencias basales encontradas por la asignación aleatoria, también incrementa la precisión de la diferencia media estimada en una magnitud de $\sqrt{1-\rho^2}$, siendo ρ la correlación entre la medición basal y la post-tratamiento.(Borm et al., 2007)

Cuando se realiza un ensayo clínico aleatorizado con dos o más grupos y múltiples mediciones, se suele realizar primero un ANOVA o ANCOVA para analizar la interacción tiempo-por-grupo, y posteriormente llevar a cabo las comparaciones por pares *post hoc* de diferencias medias entre los grupos. La recomendación actual es realizar el cálculo de tamaño muestral para dichas comparaciones *post hoc*, en lugar de para la interacción tiempo-por-grupo. Esto se debe, en primer lugar, a que, dado un tamaño muestral fijo, la potencia de una t-Student de una comparación *post hoc* es menor que la potencia para la interacción tiempo-por-grupo del ANOVA mixto. Por tanto, los cálculos basados en dicha interacción pueden derivar en una potencia insuficiente para los contrastes posteriores. Por otro lado, el tamaño del efecto de eta cuadrado parcial de la interacción tiempo-por-grupo no tiene una interpretación clínica directa, pero las diferencias medias de las comparaciones *post hoc* sí, por encontrarse en las mismas unidades de medida que la variable resultado utilizada. Por ello, primero se explicarán consideraciones para el cálculo basado en la precisión de una diferencia de medias ajustada de un ANCOVA, y posteriormente se explicarán algunos matices para los cálculos basados en potencia para el eta cuadrado parcial de un ANOVA mixto.

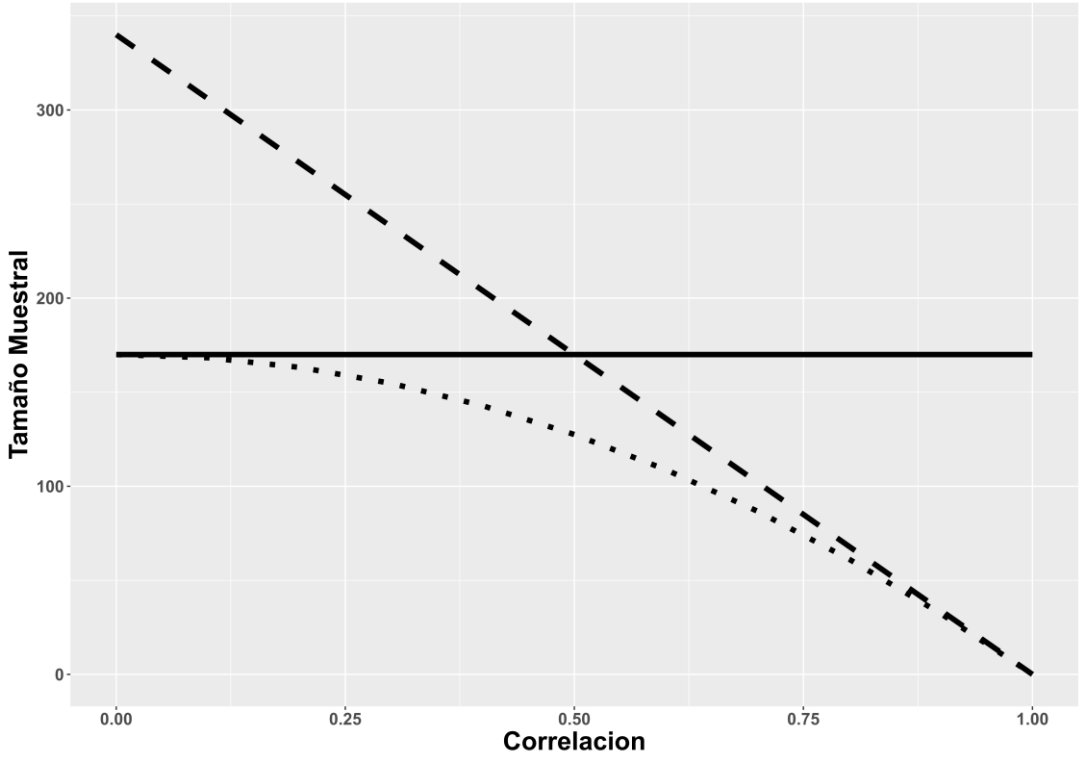


Figura 3. Gráfico líneas mostrando la relación entre el tamaño muestral necesario para una potencia estadística del 90% y la correlación entre la medición basal y la post-tratamiento, para un ensayo aleatorizado con dos grupos y dos mediciones por grupo (diferencia media = 0,5, desviación estándar = 1). La línea continua hace referencia a un diseño de comparación post-tratamiento, la línea discontinua a un diseño de comparación de la diferencia post-pre, y la línea con puntos a un diseño de análisis de la covarianza.

Calculo basado en precisión para diferencia de medias ajustada

El cálculo de tamaño muestral basado en precisión para una diferencia de medias ajustada entre dos grupos se puede realizar con la función *'ss.aipe.c.ancova()'* del paquete *'MBESS'* (Kelley, 2007), cuyo procedimiento se explica con detalle en el artículo de Lai y Kelley de 2012.(Lai & Kelley, 2012) Se requiere especificar los siguientes parámetros:

- Desviación estándar de la medición post-tratamiento en ambos grupos (o varianza residual del ANOVA).
- Correlación entre la medición basal y la medición post-tratamiento (valor recomendado de 0,5).
- Valor de MoE esperado.
- Probabilidad de obtener un valor igual o inferior al MoE esperado (valor recomendado del 80%).
- Nivel X% del intervalo de confianza (normalmente 95%).
- Número de comparaciones múltiples (en caso de haber más de dos grupos).
- Porcentaje esperado de pérdidas (recomendado un valor entre el 10% y el 20%).
- Diferencia media esperada (solo sería necesario si se va a utilizar el MMOE, en caso contrario este valor no sería necesario para el cálculo).

El primer paso consiste en seleccionar cual es el momento de seguimiento principal de interés del estudio, por ejemplo, la comparación entre los tratamientos al año de seguimiento. Es importante ya que esto influirá en las estimaciones que se realizarán de la desviación estándar post-tratamiento y

correlación entre la medición basal y la post-tratamiento.

La desviación estándar post-tratamiento es crucial en el cálculo basado en precisión, ya que a menor desviación estándar, mayor precisión y por tanto menor muestra necesaria. Por ello, se debe tener cuidado de no infraestimar su valor al realizar los cálculos.(Vickers, 2003) Se recomienda revisar la literatura previa publicada, con intervenciones y momentos de medición similares, para analizar los valores plausibles de la misma para la variable resultado de interés. La recomendación conservadora sería seleccionar de dichos valores, aquel que sea más grande, así como utilizar alguna de las dos propuestas de corrección para la desviación estándar, incluso aunque esta se haya extraído de otro ensayo clínico y no de un estudio piloto.(Whitehead et al., 2016) En la función de R *'ss.aipe.c.ancova()'* se requiere especificar la varianza residual del ANOVA, que para este caso concreto de cálculo de una diferencia de medias ajustada y con los dos grupos con igual tamaño muestral, es igual a la media de las varianzas individuales de cada grupo.

Con respecto a la correlación entre la medición basal y la post-tratamiento, a mayor correlación, mayor precisión y por tanto menor tamaño muestral necesario, de modo que también debe tenerse cuidado en no sobreestimar dicha correlación. Un valor conservador y plausible ya comentado es de 0,5.(Walters et al., 2019) La correlación entre dos mediciones repetidas disminuye según aumenta la distancia temporal entre las mismas, de forma que puede ser adecuado especificar valores de 0,4 si la diferencia de interés es a dos años o más de seguimiento.(Walters et al., 2019)

En caso de existir más de dos grupos de tratamiento, es aconsejable ajustar el X% del intervalo de confianza, siendo una opción conservadora la corrección de Bonferroni.(j. M. Bland & Altman, 1995) Esta corrección se haría dividiendo el nivel de significación deseado (ej. 5%) entre el número de comparaciones múltiples y restando a 100 el valor obtenido (ej. para tres comparaciones, 5/3 = 1,67, y por tanto el X% del intervalo de confianza a especificar en el cálculo sería 100 – 1,67 = 98,33%).

*Cálculo basado en potencia para un ANOVA con G*Power*

En caso de desear realizar un cálculo basado en potencia para la interacción tiempo-por-grupo de un ANOVA mixto en G*Power,(Faul et al., 2007) hay ciertos puntos que deben tenerse en consideración. Los parámetros a especificar para poder realizar el cálculo son:

Tamaño del efecto (debe especificarse en G*Power si es según SPSS o según Cohen). Puede especificarse directamente el valor de *f* o el de coeficiente eta cuadrado parcial (η_p^2).

- Umbral crítico de significación (normalmente 0.05).
- Potencia estadística deseada (recomendada del 90%).
- Número de grupos.
- Número de mediciones repetidas.
- Corrección para el incumplimiento de la asunción de esfericidad.
- Porcentaje esperado de pérdidas (recomendado un valor entre el 10% y el 20%).

El programa G*Power permite especificar cuatro tipos de tamaño del efecto en los ANOVA de medidas repetidas, mediante una pestaña localizada en la parte inferior denominada "Opciones":(Lakens, 2013) según G*Power, según G*Power con correlación implícita, según SPSS y según Cohen (recomendada). La opción que viene por seleccionada por defecto es la de según G*Power. Este punto es el más relevante, dado que una mal especificación de dicho tamaño del efecto puede derivar en una gran infraestimación de la muestra necesaria.(Lakens, 2013) Debe seleccionarse la opción de según SPSS, en caso de basarse de datos de eta cuadrado parcial extraídos de literatura previa, o según Cohen, en caso de basarse en sus recomendaciones de puntos de corte de un tamaño pequeño (f(V) = 0,10; η_p^2 = 0.01), mediano (f(V) = 0,25; η_p^2 = 0,06) y grande (f(V) = 0,40; η_p^2 = 0,14). De no hacerlo, la muestra siempre se infraestimará, ya que los tamaños del efecto según SPSS y Cohen ya utilizan la correlación entre medidas repetidas en su cálculo, mientras que el tamaño del efecto de GPower no, de modo que al especificar uno de los dos anteriores

como si fuese el de GPower, se utiliza dos veces la correlación entre medidas repetidas en el cálculo del tamaño muestral, derivando en una infraestimación de la muestra necesaria.(Lakens, 2013)

Por ejemplo, asumiendo un tamaño del efecto medio (f(V) = 0,25) según Cohen, 2 grupos y 5 momentos de medición, con una correlación de 0,5, una potencia del 90% y una corrección del incumplimiento de esfericidad de 1, el tamaño muestral sin cambiar la pestaña de "Opciones" resulta en 30 sujetos (15 por grupo), mientras que, especificándolo correctamente en dicha pestaña según Cohen, el tamaño muestral resultante es de 230 sujetos (115 por grupo).

La asunción de esfericidad es el otro parámetro relevante a tener en cuenta en este tipo de cálculos. Esta asunción puede definirse como la asunción de igualdad de varianzas de las diferencias entre los niveles de tratamiento en un ANOVA con medidas repetidas.(Field et al., 2012) Una violación de dicha asunción conllevaría una pérdida de potencia en un ANOVA, requiriéndose por tanto mayor muestra para un tamaño del efecto fijado. Por defecto, en G*Power el valor de corrección es igual 1, que implica la asunción de cumplimiento perfecto de esfericidad. Sin embargo, esta situación no es habitual en la práctica real,(Faul et al., 2007) de modo que es recomendable utilizar un valor más conservador como 0,75, asumiendo una posible violación de la asunción en el estudio a realizar.

Recomendaciones para Análisis de Regresión Lineal Múltiple

Existe mucha literatura publicada con respecto al cálculo de tamaño muestral en el caso de análisis de regresión lineal múltiple, tanto sobre cálculos basados en potencia como en precisión, para el coeficiente de determinación (R^2),(Algina & Olejnik, 2000) y para los coeficientes de regresión individuales del modelo (β).(Hsieh et al., 1998; Kelley & Maxwell, 2003) No se profundizará en este apartado en el cálculo basado en dichos análisis con el objetivo de elaboración de un modelo predictivo multivariable, ya que se abordará de manera independiente por requerir de algunas consideraciones especiales.(Riley et al., 2019a)

En el campo de los análisis de regresión múltiple, tanto lineal como logística u otros, se ha publicado mucha literatura acerca del concepto de "sujetos por variable" (SPV) y de "sujetos por parámetro" (SPP) (o eventos por variable [EPV] y por parámetro [EPP] en el caso de regresiones logísticas, siendo los eventos el número de sujetos de la categoría con menor frecuencia). El método de cálculo de tamaño muestral basado en SPV o SPP consiste en reclutar un número predefinido de sujetos por cada variable o parámetro predictor a incluir en el modelo.(van Smeden et al., 2019) Existen multitud de propuestas en la literatura, de 10 sujetos,(Harrell, 2001) 15-20 sujetos,(Schmidt, 1971) y e incluso valores llamativamente pequeños como 2 sujetos por variable.(Austin & Steyerberg, 2015) Estos cálculos ha sido duramente criticados por algunos autores en la literatura,(Van Smeden et al., 2016; van Smeden et al., 2019) siendo actualmente desaconsejados, debido a que se han visto ineficientes por ignorar múltiples consideraciones a tener en cuenta en el cálculo del tamaño muestral.

De entre las dos opciones, la basada en el coeficiente de determinación R^2 presenta una menor utilidad por si sola. Las propuestas de cálculos para una estimación precisa de dicho coeficiente son útiles dentro de un marco de elaboración de un modelo predictivo multivariable,(Riley et al., 2019a) pero no en una investigación de factores pronóstico, donde el objetivo del estudio son los coeficientes de regresión.(Riley et al., 2013)

La elección de un cálculo basado en potencia (Hsieh et al., 1998) o en precisión (Kelley & Maxwell, 2003) de los coeficientes de regresión individuales dependerá del objetivo del estudio. Como ya se ha comentado, los cálculos basados en potencia tienen como objetivo dar respuestas dicotómicas de si existe o no una asociación o efecto. En el contexto de una regresión múltiple, el cálculo basado en potencia (Hsieh et al., 1998) se realizará si el objetivo del estudio es realizar un análisis preliminar de que posibles variables pueden presentar una asociación con la variable dependiente de interés, para utilizar las mismas en futuras investigaciones basadas en precisión o en la posible elaboración de modelos predictivos. Por otro lado, si el objetivo es estimar de manera precisa uno o varios coeficientes de regresión,

para discutir sobre la fuerza de la asociación de una o más variables independientes con la variable dependiente, entonces el cálculo de elección es el basado en precisión.(Kelley & Maxwell, 2003)

Una propuesta adecuada para el cálculo basado en potencia es la de Hsieh y cols.(Hsieh et al., 1998) del año 1998, basada en la corrección del cálculo para un coeficiente de correlación de Pearson con la transformación de Fisher. Con respecto al cálculo basado en precisión con una potencia deseada para el MoE, una propuesta óptima es la de Kelley y Maxwell del año 2003,(Kelley & Maxwell, 2003) que puede implementarse con la función *'ss.aipe.reg.coef'* del paquete *'MBESSS*' en R.

Factores que influyen en el tamaño muestral en análisis de regresión lineal múltiple

Son tres los parámetros principales que influyen en la muestra necesaria para un análisis de regresión lineal múltiple:(Hsieh et al., 1998; Kelley & Maxwell, 2003)

- Coeficiente de determinación del modelo multivariable prediciendo la variable dependiente de interés (R^2).
- Coeficiente de determinación del modelo multivariable prediciendo una de las variables independientes por el resto de las variables independientes (R^2_{XXj}).
- Número de parámetros predictores potenciales a introducir en el modelo.

Cuanto mayor sea el coeficiente de determinación R^2 menor será la muestra necesaria, ya que se mejora la potencia y precisión en la estimación de los coeficientes de regresión individuales.(Hsieh et al., 1998; Kelley & Maxwell, 2003) Por el contrario, el coeficiente R^2_{XXj} tiene el efecto contrario, cuanto mayor es la asociación entre el resto de predictores y ese predictor de interés, más disminuye la potencia y precisión de su coeficiente de regresión.(Hsieh et al., 1998; Kelley & Maxwell, 2003) Este punto es importante, ya que en algunas situaciones, será necesario realizar el cálculo de tamaño muestral para las asunciones especificadas para cada parámetro predictor individual (ya que variará el valor de R^2_{XXj}), y quedarnos con el cálculo de mayor muestra, es decir, usar el valor más grande estimado de R^2_{XXj}.(Hsieh et al., 1998; Kelley & Maxwell, 2003)

Para la estimación de R^2 y R^2_{XXj} existen dos posibilidades, calcular sus valores a través de una estimación del vector de correlaciones entre los predictores y la variable dependiente, así como la matriz de correlaciones entre los distintos predictores, o estimar su valor de manera directa, ya sea en base a criterios de conocimiento técnico de los investigadores de mínimo tamaño esperado, o en literatura previa, en cuyo caso deben usarse siempre valores de R^2 ajustados.(Hsieh et al., 1998; Kelley & Maxwell, 2003) Para la primera opción, dichas matrices pueden elaborarse basándose en literatura previa publicada, o en puntos de corte de tamaños del efecto pequeños, medianos y grandes,(Kelley & Maxwell, 2003) procurando no sobreestimar el valor de R^2 ni infraestimar el de R^2_{XXj}.(Kelley & Maxwell, 2003) En el caso de utilizar puntos de corte, mi recomendación propia es asumir una correlación máxima de 0,30 entre el predictor de interés y la variable dependiente, y una correlación mínima de 0,4 entre los distintos predictores, para los cálculos basados en potencia, así como una potencia del 90%.

Por otro lado, a mayor número de parámetros predictores potenciales a incluir en el modelo, mayor muestra necesaria.(Hsieh et al., 1998; Kelley & Maxwell, 2003; Riley et al., 2020) Se habla de parámetros potenciales porque deben tenerse en cuenta todos los posibles parámetros que se valoran incluir en el modelo, aunque finalmente no se incluyan todos por realizarse métodos de selección automática de variables, como en una regresión por pasos hacia detrás. Es decir, si en una investigación se planifica medir 20 potenciales parámetros, aunque se estime que finalmente la regresión lineal múltiple quedará constituida solo por 15 parámetros, el cálculo de tamaño muestral debe realizarse para 20 y no para 15.(Riley et al., 2020) También debe tenerse en cuenta la posible inclusión de asociación no lineales y/o interacciones, que incrementan el número de parámetros, así como la presencia de variables multicategóricas, que tienen asociado más de un parámetro predictor para su adecuada codificación.(Riley et al., 2020)

Por último, y quizás el punto más controvertido, es el establecimiento de un MoE aceptable en el caso de coeficientes de regresión para un cálculo basado en precisión. Kelley y Maxwell recomiendan establecer el valor de MoE para el coeficiente de regresión estandarizado parcial, en lugar de para el coeficiente sin estandarizar, por la dificultad en establecer un MoE "óptimo" sin estandarizar.(Kelley & Maxwell, 2003) Un valor de MoE de 0,10 o 0,15 para los coeficientes parcialmente estandarizados puede ser considerado aceptable para este tipo de cálculos.

Recomendaciones para Análisis de Regresión Logística Binaria Múltiple

La situación de la investigación con respecto al cálculo de tamaño muestral para una regresión logística binaria múltiple es similar a lo ya comentado para una regresión lineal múltiple. Existen distintas propuestas de valores de EPV o EPP que han sido duramente criticados en la literatura, desaconsejándose su utilización actualmente.(Van Smeden et al., 2016; van Smeden et al., 2019)

Dada la finalidad habitual de este tipo de análisis en investigación, solamente se explicará el cálculo basado en potencia por Hsieh y cols.,(Hsieh et al., 1998) ya que se asume que si se desea precisión en la estimación de los coeficientes, es porque el estudio presentará un diseño de modelo predictivo multivariable,(Riley et al., 2019c, 2020) que se abordará en la sección correspondiente.

Para la realización del cálculo, se requiere de la estimación de los siguientes parámetros:(Hsieh et al., 1998)

- Probabilidad de ocurrencia del evento de la variable dependiente (ej. prevalencia de dolor lumbar).
- Diferencia media entre las dos categorías de la variable independiente y desviación estándar en cada categoría, o especificación directa de la *d* de Cohen (en el caso de que la variable independiente de interés sea continua).
- Probabilidad de ocurrencia del evento de la variable dependiente en cada una de las dos categorías de la variable independiente de interés, así como prevalencia de ocurrencia del evento de la variable independiente de interés (en caso de que la variable independiente de interés sea dicotómica).
- Coeficiente de determinación del modelo multivariable de regresión lineal prediciendo la variable independiente de interés por el resto de las variables independientes (R^2_{XXj}).

La probabilidad de ocurrencia del evento de la variable dependiente hace referencia a la prevalencia de la variable a predecir con el modelo de regresión logística, como por ejemplo la prevalencia de dolor lumbar. Es importante no modificar de manera experimental esta frecuencia (por ejemplo, reclutando 300 sujetos con y 300 sin dolor lumbar), sino que debe estimarse su valor en función de la prevalencia estimada en la población bajo estudio, para que los resultados puedan ser extrapolables y no se cometan errores en la estimación de los coeficientes.

En el caso de que la variable independiente de interés del modelo sea cuantitativa, el cálculo propuesto por Hsieh y cols.(Hsieh et al., 1998) consiste en calcular la muestra necesaria para una prueba t-Student para dos muestras independientes, ajustando el ratio de sujetos entre los grupos de acuerdo a la prevalencia de ocurrencia del evento de la variable dependiente, y finalmente corrigiendo la muestra resultante para el valor esperado de R^2_{XXj} resultante de predecir dicha variable independiente de interés por el resto de variables independientes introducidas en el modelo. Para el establecimiento de dichos valores se aconseja seguir las recomendaciones previas realizadas en secciones anteriores.

En el caso de que la variable independiente de interés sea dicotómica, deben especificarse las probabilidades indicadas para la variable independiente, así como para la variable dependiente dentro de cada nivel de la independiente. Estos valores deben ser estimados en base a las prevalencias esperadas en la población diana, es decir, si la prevalencia esperada de mujeres es del 30%, no es aconsejable modificar experimentalmente dicho valor dentro del estudio para que haya una prevalencia del 50% para cada sexo.

Es recomendable realizar el cálculo para los distintos predictores de interés y utilizar el de mayor muestra resultante, para una potencia esperada del 80% (recomendación propia).

Recomendaciones para Análisis de Reproducibilidad

Actualmente existen numerosas propuestas de distintos estadísticos de reproducibilidad dentro de la teoría clásica del test, tanto para variables cuantitativas(Gwet, 2021b) como categóricas,(Gwet, 2021a) en función de la escala de medición de las mismas y otras asunciones. En el caso de variables categóricas, algunas propuestas son el coeficiente Kappa de Cohen y Fleiss,(Sim & Wright, 2005) el coeficiente AC1 de Gwet,(Gwet, 2008, 2021a) el coeficiente pi de Scott,(Gwet, 2008, 2021a) el coeficiente Kappa generalizado de Conger,(Gwet, 2008, 2021a) o el coeficiente Alpha de Krippendorff.(Gwet, 2008, 2021a) Por su parte, en el caso de variables cuantitativas, existen distintas variantes del coeficiente de correlación intraclase, y otras propuestas como el coeficiente de Finn, o los coeficientes de concordancia ponderados.(Gwet, 2021b) Del mismo modo, existen numerosas propuestas en la literatura de cálculos de tamaño muestral para dichos coeficientes.(Bonett, 2002; Cantor, 1996; Gwet, 2008, 2021a, 2021b; S. Liu & Luo, 2010; Saito et al., 2006; Walter & Donner A, 1998; Zou, 2012)

Kilem Li Gwet es uno de los investigadores actuales que más ha profundizado en el campo de los análisis de reproducibilidad, con dos libros publicados, uno destinado a las variables categóricas,(Gwet, 2021a) y otro a las variables cuantitativas.(Gwet, 2021b) En ellos presenta propuestas de cálculo de tamaño muestral basados en precisión para múltiples coeficientes de reproducibilidad, junto con tablas de cálculos ya realizados para ellos en función de múltiples asunciones.(Gwet, 2021a, 2021b) Actualmente se encuentra en proceso de elaboración de un paquete de R para poder implementar dichos cálculos (comunicación personal), pero de momento la única y mejor opción es la implementación directa de las fórmulas y/o tablas reportadas en sus libros, que recomiendo encarecidamente a todos aquellos interesados en calcular el tamaño muestral necesario para este tipo de investigaciones, ya que no existe una solución única cerrada y los cálculos varían mucho de un coeficiente a otro y por tanto de un diseño a otro.(Gwet, 2021a, 2021b) No obstante, a continuación se resumen algunos de los conceptos a tener en cuenta para planificar los cálculos con análisis de reproducibilidad basados en precisión.

En primer lugar, un aspecto fundamental a tener en cuenta es que la precisión de la estimación de los coeficientes de reproducibilidad se incrementa a medida que lo hace el valor real del coeficiente. Es decir, dado un tamaño muestral fijo, el MoE disminuye según aumenta el valor real del coeficiente de reproducibilidad bajo estudio.(Gwet, 2021a, 2021b) Este punto es importante ya que una sobreestimación de la reproducibilidad esperada también conllevará una sobreestimación de la precisión de tal estimación, pudiendo suponer un mayor detrimento para el estudio por una excesiva infraestimación de la muestra necesaria.(Gwet, 2021a, 2021b)

En relación a los cálculos para variables categóricas, influyen en el mismo el número de evaluadores utilizados, el número de categorías, el valor del coeficiente de reproducibilidad esperado, y la distribución de los sujetos en las distintas categorías. Es complicado, casi imposible en algunos casos, estimar la distribución de sujetos entre las distintas categorías. Por ello, Gwet hace una propuesta basada en la máxima varianza posible del coeficiente de reproducibilidad, mediante el cálculo del estadístico "*C*", que facilita tabulado para distintos escenarios (número de categorías y evaluadores) y varios coeficientes de reproducibilidad, permitiendo el cálculo del tamaño muestral según dos fórmulas propuestas en función de si el tamaño poblacional es finito conocido o infinito, para una precisión deseada.(Gwet, 2021a) Cabe destacar, dada la metodología propuesta de máxima varianza de Gwet, que el concepto de probabilidad para una precisión deseada no presentaría aplicación en estos cálculos.

Por otro lado, en el cálculo de tamaño muestral para variables cuantitativas influyen el valor esperado del coeficiente de reproducibilidad, el número de evaluadores y el número de mediciones realizadas a cada sujeto (además de otros factores como la distribución de la variable, aunque este aspecto no se

puede tener en cuenta en los cálculos de momento, asumiéndose una distribución aproximada a la normal). Gwet facilita fórmulas basadas en precisión para distintos tipos del coeficiente de correlación intraclase, así como tablas con cálculos ya realizados en función de varias combinaciones de los tres parámetros mencionados.(Gwet, 2021b) A pesar de que Gwet no implementa en sus cálculos el concepto de potencia para una precisión deseada, este puede implementarse modificando el valor del estadístico de las distribuciones propuestas por Gwet en sus cálculos.

Recomendaciones para Modelos Predictivos Multivariables

Los análisis de regresión multivariables pueden utilizarse, además de para evaluar asociaciones controlando para otras covariables, con el objetivo de predecir un evento,(Moons, Altman, et al., 2009; Moons, Royston, et al., 2009) ya sea la mejoría de discapacidad al año de seguimiento de un paciente (ej. mediante una regresión lineal),(Kent et al., 2020; Riley et al., 2019b) la presencia una determinada patología (mediante una regresión logística), o el tiempo de supervivencia con una enfermedad (mediate una regresión de hazards proporcionales).(Riley et al., 2019c) Este tipo de modelos pueden tener la finalidad de predecir el pronóstico de una enfermedad,(Steyerberg et al., 2013) la mejoría con uno tratamiento,(Hingorani et al., 2013) o el diagnóstico del paciente.(Steyerberg et al., 2013) Debido a esta finalidad, este tipo de metodología requiere de algunas consideraciones adicionales a lo ya comentado con respecto al cálculo de tamaño muestral para estos análisis.(Riley et al., 2020; Royston et al., 2009)

Los cálculos se pueden implementar con el paquete de R *'pmsampsize'* del grupo PROGRESS.(Riley et al., 2020) A continuación se explican algunos de los conceptos clave de este tipo de cálculos, sin embargo, se recomienda la lectura de la serie de artículos del grupo PROGRESS a fin de poder entender en profundidad los mismos y poder implementarlos correctamente.(Pate et al., 2023; Riley et al., 2019a, 2019c, 2020) De forma resumida, los cálculos se basan en intentar asegurar que se cumplen una serie de criterios:(Riley et al., 2020)

- Poco sobreajuste en el modelo, con un *shrinkage* esperado del 10% o menos.
- Pequeña diferencia absoluta (0,05) entre el valor observado del coeficiente R^2 (pseudo-R^2 de Nagelkerke en el caso de análisis de regresión logística y de hazards proporcionales) y su valor ajustado.
- Estimación precisa de la desviación estándar residual (regresión lineal).
- Estimación precisa del valor medio de la variable dependiente (regresión lineal), o el riesgo general en la población diana para un punto de seguimiento de interés (regresión logística y de hazards proporcionales).

El concepto de sobreajuste hace referencia a un exceso de capacidad aparente de predicción del modelo multivariable en los datos utilizados para elaborar el mismo, derivado un escaso tamaño muestral para el número de parámetros predictores incluidos en el modelo. El resultado es un modelo que aparenta ser muy bueno prediciendo, pero que presentará una escasa validez externa al intentar predecir la variable de interés en otros sujetos diferentes a los utilizados para elaborar el modelo.(Riley et al., 2020) Aunque existen métodos propuestos (penalización o *shrinkage*) para corregir el modelo para dicho sobreajuste, estos no funcionan bien con muestras pequeñas, motivo por el cual es necesario utilizar una muestra suficiente que minimice el potencial sobreajuste del modelo.(Riley et al., 2020) De manera simplificada, los métodos de penalización se basan en "arrastrar" los valores de los coeficientes de regresión hacia el cero. Dado que se asume un posible sobreajuste, esos coeficientes se habrán sobreestimado, de modo que los métodos de penalización tienen como objetivo disminuir el valor de los mismos, multiplicándolos por un (o varios) factor de corrección, denominado factor de *shrinkage*.(Riley et al., 2019b) El objetivo de este criterio es que dicho factor de *shrinkage* sea de 0.90 o superior, es decir, que se requiera poca corrección para el posible sobreajuste del modelo, de forma que esa penalización funcione adecuadamente y se mejore la validez externa del modelo.(Riley et al., 2019b)

El segundo criterio presenta cierta relación con el primero, acerca de la no sobreestimación del ajuste del

modelo y su capacidad de predicción. Por ello, se necesita utilizar un tamaño muestral que minimice la diferencia entre los coeficientes R^2 (que reflejan el ajuste general del modelo) y sus valores ajustados. Para ello, los miembros del grupo PROGRESS recomiendan una diferencia máxima esperada de 0,05 entre ambos valores.(Riley et al., 2020)

En relación a los modelos con variables continuas, también es necesario estimar de manera precisa la desviación estándar residual del modelo de regresión, ya que este valor se utiliza para la estimación del coeficiente R^2, los errores estándar de los coeficientes de regresión y de la constante, y los intervalos de confianza de las predicciones realizadas con el modelo.(Riley et al., 2019b)

El cuarto criterio hace referencia a la estimación precisa de la media poblacional de la variable resultado de interés, en caso una regresión lineal (ej. media de discapacidad de los sujetos al año de seguimiento de 32 puntos), o de la prevalencia del evento en cuestión en la población de interés, para un momento de seguimiento dado, en el caso de análisis de regresión logística o de hazards proporcionales (ej. riesgo de cáncer de pulmón a los 5 años de seguimiento de 0.23). Esto se debe a que las predicciones a realizar se "anclan" a dicha medida de tendencia central. Para este criterio, el grupo PROGRESS establece como aceptable un valor absoluto de MMOE de 1.1 o menor en el caso de variables continuas,(Riley et al., 2019b, 2020) y un margen de error de ± 0,05 de la prevalencia en el caso de variables dicotómica(Riley et al., 2019c, 2020)

Finalmente, un último aspecto a tener en cuenta, tanto para estudios de modelos predictivos como para otros diseños, es que en general, cuanta más muestra mejor, siempre que se tengan en cuenta algunos aspectos metodológicos para asegurar la ética de la investigación.("World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects," 2013) Por ejemplo, si se dispone de una base de datos ya medidos de 10,000 sujetos, que puede utilizarse, carece de sentido utilizar solo 2,000 por el cálculo de tamaño muestral realizado, lo aconsejable es utilizar la muestra íntegra de 10,000 sujetos que permitirá obtener estimaciones más precisas de los parámetros, así como incluir más posibles variables, relaciones no lineales y/o interacciones (en cuyo caso la muestra debería ser incrementada con respecto a los cálculos aquí recogidos, ya que se requiere de una mayor muestra para detectar posibles interacciones entre variables).

Otros Cálculos de Tamaño Muestral

Aunque no es posible profundizar en todos, cabe destacar que también existen propuestas de cálculo de tamaño muestral para otros análisis estadísticos, como los análisis de mediación,(Pan et al., 2018; Schoemann et al., 2017) diferencias en las pendientes de dos líneas de regresión (interacción entre una variable dicotómica y una continua),(Shieh, 2009, 2018) interacción entre dos variables continuas,(Shieh, 2010) diferencias de medias ajustadas para covariables en estudios observacionales,(X. S. Liu, 2010) o el intervalo de concordancia de Bland-Altman,(Jan & Shieh, 2018) entre otros. Por ello, recomiendo a cualquier investigador revisar la literatura pertinente para el caso concreto de investigación que se pretenda realizar.

CONCLUSIÓN

El cálculo de tamaño muestral es uno de los apartados fundamentales a tener en cuenta en la planificación de la mayoría de los diseños de investigación. Los cálculos basados en precisión pueden ser preferibles en algunas situaciones a los basados en potencia. Es recomendable ser conservadores en los cálculos para disminuir la posibilidad de infraestimar el tamaño muestral necesario.

AGRADECIMIENTOS

El autor agradece a todas las personas cercanas que le apoyan durante sus labores de investigación, con especial consideración a sus compañeros/as del XXX.

FRASES DESTACADAS

- El cálculo de tamaño muestral es uno de los puntos más importantes del diseño de la mayoría de las investigaciones.
- Los cálculos basados en precisión pueden ser preferibles en algunas situaciones en lugar de los basados en potencia.
- Los cálculos deben ser conservadores para evitar una infraestimación de la muestra necesaria.

HIGHLIGHTS

- Sample size calculation is one of the most important considerations in the design of most research studies.
- Sample size calculations based on precision may be preferable in some situations than power-based ones.
- Calculations must be conservative in aim to avoid underestimation of the needed sample size.

REFERENCIAS

Algina, J., & Olejnik, S. (2000). Determining Sample Size for Accurate Estimation of the Squared Multiple Correlation Coefficient. *Multivariate Behavioral Research*, *35*(1), 119–137. https://doi.org/10.1207/S15327906MBR3501_5

Arienti, C., Armijo-Olivo, S., Minozzi, S., Tjosvold, L., Lazzarini, S. G., Patrini, M., & Negrini, S. (2021). Methodological Issues in Rehabilitation Research: A Scoping Review. *Archives of Physical Medicine and Rehabilitation*, *102*(8), 1614-1622.e14. https://doi.org/10.1016/J.APMR.2021.04.006

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/J.JCLINEPI.2014.12.014

Barnes, S. A., Lindborg, S. R., & Seaman, J. W. (2006). Multiple imputation techniques in small sample clinical trials. *Statistics in Medicine*, *25*(2), 233–245. https://doi.org/10.1002/SIM.2231

Beal, S. (1991). Response to "Confidence intervals and sample sizes." *Biometrics*, *47*(4), 1602–1603.

Beal, S. L. (1989). Sample Size Determination for Confidence Intervals on the Population Mean and on the Difference Between Two Population Means. *Biometrics*, *45*(3), 969. https://doi.org/10.2307/2531696

Bell, M. L., Whitehead, A. L., & Julious, S. A. (2018). Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. *Clinical Epidemiology*, *10*, 153–157. https://doi.org/10.2147/CLEP.S146397

Bland, j. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, *310*(6973), 170. https://doi.org/10.1136/BMJ.310.6973.170

Bland, J. M. (2009). The tyranny of power: is there a better way to calculate sample size? *BMJ (Clinical Research Ed.)*, *339*(7730), 1133–1135. https://doi.org/10.1136/BMJ.B3985

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*(9), 1331–1335. https://doi.org/10.1002/sim.1108

Borm, G. F., Fransen, J., & Lemmens, W. A. J. G. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology*, *60*(12), 1234–1238. https://doi.org/10.1016/J.JCLINEPI.2007.02.006

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine*, *14*(17), 1933–1940. https://doi.org/10.1002/SIM.4780141709

Cantor, A. B. (1996). Sample-Size Calculations for Cohen's Kappa. *Psychological Methods*, *1*(2), 150–153.

Cocks, K., & Torgerson, D. J. (2013). Sample size calculations for pilot randomized trials: a confidence interval approach. *Journal of Clinical Epidemiology*, *66*(2), 197–201. https://doi.org/10.1016/J.JCLINEPI.2012.09.002

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., Irwig, L., Levine, D., Reitsma, J. B., De Vet, H. C. W., & Bossuyt, P. M. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*, *6*(11), e012799. https://doi.org/10.1136/BMJOPEN-2016-012799

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, *13*(1), 1–10. https://doi.org/10.1186/S12916-014-0241-Z/TABLES/1

Cook, J. A., Julious, S. A., Sones, W., Hampson, L. V., Hewitt, C., Berlin, J. A., Ashby, D., Emsley, R., Fergusson, D. A., Walters, S. J., Wilson, E. C. F., Maclennan, G., Stallard, N., Rothwell, J. C., Bland, M., Brown, L., Ramsay, C. R., Cook, A., Armstrong, D., … Vale, L. D. (2018). DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Trials*, *19*(1). https://doi.org/10.1186/S13063-018-2884-0

Copsey, B., Thompson, J. Y., Vadher, K., Ali, U., Dutton, S. J., Fitzpatrick, R., Lamb, S. E., & Cook, J. A. (2018). Sample size calculations are poorly conducted and reported in many randomized trials of hip and knee osteoarthritis: results of a systematic review. *Journal of Clinical Epidemiology*, *104*, 52–61. https://doi.org/10.1016/J.JCLINEPI.2018.08.013

Dechartres, A., Trinquart, L., Boutron, I., & Ravaud, P. (2013). Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *BMJ (Clinical Research Ed.)*, *346*(7908). https://doi.org/10.1136/BMJ.F2304

Eldridge, S. M., Chan, C. L., Campbell, M. J., Bond, C. M., Hopewell, S., Thabane, L., Lancaster, G. A., Altman, D., Bretz, F., Campbell, M., Cobo, E., Craig, P., Davidson, P., Groves, T., Gumedze, F., Hewison, J., Hirst, A., Hoddinott, P., Lamb, S. E., … Tugwell, P. (2016). CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*, *355*. https://doi.org/10.1136/BMJ.I5239

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Ed.)*, *292*(6522), 746. https://doi.org/10.1136/BMJ.292.6522.746

Gonzalez, G. Z., Moseley, A. M., Maher, C. G., Nascimento, D. P., Costa, L. da C. M., & Costa, L. O. (2018). Methodologic Quality and Statistical Reporting of Physical Therapy Randomized Controlled Trials Relevant to Musculoskeletal Conditions. *Archives of Physical Medicine and Rehabilitation*, *99*(1), 129–136. https://doi.org/10.1016/J.APMR.2017.08.485

Grieve, A. (1989). Confidence intervals and trial sizes (Letter). *Lancet*, *i*, 337.

Grieve, A. (1991). Confidence intervals and sample sizes. *Biometrics*, *47*(4), 1597–1603. https://doi.org/https://doi.org/10.2307/2532411

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2021a). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Volume 1: Analysis of Categorical Ratings* (5th ed.). AgreeStat Analytics.

Gwet, K. L. (2021b). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Volume 2: Analysis of Quantitative Ratings* (5th ed.). AgreeStat Analytics.

Harrell, F. E. (2001). *Regression modeling strategies*. Springer-Verlag.

Haynes, A. G., Lenz, A., Stalder, O., & Limacher, A. (2021). `presize`: An R-package for precision-based sample size calculation in clinical research. *Journal of Open Source Software*, *6*(60), 3118. https://doi.org/10.21105/JOSS.03118

Hingorani, A. D., Van Der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., Hemingway, H., Briggs, A., Brunner, N., Croft, P., Hayden, J., Kyzas, P., Malats, N., Peat, G., Perel, P., Roberts, I., & Timmis, A. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ*, *346*. https://doi.org/10.1136/BMJ.E5793

Hsieh, F., Bloch, D., & Larsen, M. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, *17*(14), 1623–1634.

Jan, S. L., & Shieh, G. (2018). The Bland-Altman range of agreement: Exact interval procedure and sample size determination. *Computers in Biology and Medicine*, *100*, 247–252. https://doi.org/10.1016/J.COMPBIOMED.2018.06.020

Julious, S. A., & Owen, R. J. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics*, *5*(1), 29–37. https://doi.org/10.1002/PST.197

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: an R package. *Behavior Research Methods*, *39*(4), 979–984. https://doi.org/10.3758/BF03192993

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply

significant. *Psychological Methods*, *8*(3), 305–321. https://doi.org/10.1037/1082-989X.8.3.305

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P. A., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., & Steyerberg, E. W. (2020). The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine*, *172*(1), 35–45. https://doi.org/10.7326/M18-3667

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: sample size planning via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology*, *65*(2), 350–370. https://doi.org/10.1111/J.2044-8317.2011.02029.X

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV). https://doi.org/10.3389/fpsyg.2013.00863

Liu, S., & Luo, J. (2010). A Study on the Current Development of Body Shape during Infancy in Shanghai. In Jiang, Y and Zou, YL and Zhang, JG and Chen, JQ (Ed.), *PROCEEDINGS OF THE 2010 INTERNATIONAL SYMPOSIUM ON CHILDREN AND YOUTH FITNESS AND HEALTH, VOL 1* (pp. 256–259).

Liu, X. S. (2010). Sample Size for Confidence Interval of Covariate-Adjusted Mean Difference. *Http://Dx.Doi.Org/10.3102/1076998610381401*, *35*(6), 714–725. https://doi.org/10.3102/1076998610381401

Moons, K. G. M., Altman, D. G., Vergouwe, Y., & Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*, *338*(7709), 1487–1490. https://doi.org/10.1136/BMJ.B606

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: what, why, and how? *BMJ*, *338*(7706), 1317–1320. https://doi.org/10.1136/BMJ.B375

Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology*, *18*(1), 1–11. https://doi.org/10.1186/S12874-018-0473-2/FIGURES/3

Pate, A., Riley, R. D., Collins, G. S., van Smeden, M., Van Calster, B., Ensor, J., & Martin, G. P. (2023). Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Statistical Methods in Medical Research*, *32*(3). https://doi.org/10.1177/09622802231151220

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & Van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical Research Ed.)*, *368*. https://doi.org/10.1136/BMJ.M441

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G. M., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., & Hemingway, H. (2013). Prognosis

Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine*, *10*(2). https://doi.org/10.1371/JOURNAL.PMED.1001380

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., & Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/SIM.7993

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., & Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/SIM.7993

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., & Collins, G. S. (2019c). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*, *38*(7), 1276–1296. https://doi.org/10.1002/SIM.7992

Rothman, K. J., & Greenland, S. (2018). Planning Study Size Based on Precision Rather Than Power. *Epidemiology (Cambridge, Mass.)*, *29*(5), 599–603. https://doi.org/10.1097/EDE.0000000000000876

Royston, P., Moons, K. G. M., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ*, *338*(7707), 1373–1377. https://doi.org/10.1136/BMJ.B604

Saito, Y., Sozu, T., Hamada, C., & Yoshimura, I. (2006). Effective number of subjects and number of raters for inter-rater reliability studies. *Statistics in Medicine*, *25*(9), 1547–1560. https://doi.org/10.1002/SIM.2294

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, *31*(3), 699–714. https://doi.org/10.1177/001316447103100310/ASSET/001316447103100310.FP.PNG_V03

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science*, *8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ (Online)*, *340*(7748), 698–702. https://doi.org/10.1136/bmj.c332

Shieh, G. (2009). Detection of interactions between a dichotomous moderator and a continuous predictor in moderated multiple regression with heterogeneous error variance. *Behavior Research Methods*, *41*(1), 61–74. https://doi.org/10.3758/BRM.41.1.61

Shieh, G. (2010). Sample size determination for confidence intervals of interaction effects in moderated multiple regression with continuous predictor and moderator variables. *Behavior Research Methods*, *42*(3), 824–835. https://doi.org/10.3758/BRM.42.3.824

Shieh, G. (2018). Power and sample size calculations for comparison of two regression lines with heterogeneous variances. *PLoS ONE*, *13*(12). https://doi.org/10.1371/JOURNAL.PONE.0207745

Sim, J. (2019). Should treatment effects be estimated in pilot and feasibility studies?

Pilot and Feasibility Studies, *5*(1). https://doi.org/10.1186/S40814-019-0493-7

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., & Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine*, *10*(2). https://doi.org/10.1371/JOURNAL.PMED.1001381

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: A simulation study. *Trials*, *15*(1), 1–13. https://doi.org/10.1186/1745-6215-15-264/FIGURES/8

Van Smeden, M., De Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, *16*(1), 1–12. https://doi.org/10.1186/S12874-016-0267-3/TABLES/4

van Smeden, M., Moons, K. G. M., de Groot, J. A. H., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, *28*(8), 2455–2474. https://doi.org/10.1177/0962280218784726/ASSET/IMAGES/LARGE/10.1177_0962280218784726-FIG4.JPEG

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., & Egger, M. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). *Epidemiology*, *18*(6), 805–835. https://doi.org/10.1097/EDE.0b013e3181577511

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology*, *1*(1), 1–4. https://doi.org/10.1186/1471-2288-1-6/TABLES/1

Vickers, A. J. (2003). Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, *56*(8), 717–720. https://doi.org/10.1016/S0895-4356(03)00141-0

Vickers, A. J., & Altman, D. G. (2001). Statistics Notes: Analysing controlled trials with baseline and follow up measurements. *BMJ : British Medical Journal*, *323*(7321), 1123. https://doi.org/10.1136/BMJ.323.7321.1123

Walter, S., & Donner A, M. E. (1998). Sample size and optimal designs for reliability studies. *Stat Med*, *17*(1), 101–110.

Walters, S. J., Jacques, R. M., Henriques-Cadby, I. B. D. A., Candlish, J., Totton, N., & Shu Xian, M. T. (2019). Sample size estimation for randomised controlled trials with repeated assessment of patient-reported outcomes: what correlation between baseline and follow-up outcomes should we assume? *Trials*, *20*(1), 566. https://doi.org/10.1186/S13063-019-3671-2

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. In *American Statistician* (Vol. 70, Issue 2, pp. 129–133). American Statistical Association.

https://doi.org/10.1080/00031305.2016.1154108

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231–240. https://doi.org/10.1519/15184.1

Whitehead, A. L., Julious, S. A., Cooper, C. L., & Campbell, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, *25*(3), 1057–1073. https://doi.org/10.1177/0962280215588241

World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. (2013). *JAMA*, *310*(20), 2191–2194. https://doi.org/10.1001/JAMA.2013.281053

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine*, *31*(29), 3972–3981. https://doi.org/10.1002/sim.5466

Material Suplementario 1: Ejemplos Prácticos de Cálculos de Tamaño Muestral

El Cálculo del Tamaño Muestral en Ciencias de la Salud: Recomendaciones y Guía Práctica

Correspondencia:

Rubén Fernández Matías, Fisioterapeuta

ruben.fernanmat@gmail.com

EJEMPLO PRÁCTICO PARA ENSAYOS CLÍNICOS ALEATORIZADOS

Ejemplo 1: Diferencia de medias ajustada de un ANCOVA

Se desea realizar un estudio para comparar dos intervenciones de ejercicio en pacientes con dolor relacionado con el manguito rotador. Se estableció el seguimiento principal en un año, con el cuestionario *Shoulder Pain and Disability Index* como variable principal. Se asumió una desviación estándar de 23 puntos igual para ambos grupos y una correlación con la medición basal de 0.5. Además, se estableció nivel de confianza del 95%, un valor de MoE de 4 puntos y una probabilidad de obtener un MoE igual o inferior a 4 del 80%.

El cálculo se realizará con la función *'ss.aipe.c.ancova()'* del paquete de R *'MBESS*',(Lai & Kelley, 2012) que requiere de la especificación de los siguientes parámetros:

error.var.anova	Varianza del error del ANOVA. En nuestro caso al asumir que ambos grupos la misma desviación estándar, este valor es igual a $23^2 = 529$.
rho	Correlación entre la medición basal y la post-tratamiento. En nuestro caso de 0.5.
c.weights	Vector para especificar el contraste del ANCOVA, al tratarse de dos grupos se especifica como *c(-0.5, 0.5).*
width	Valor de MoE deseado.
conf.level	Nivel de confianza del intervalo, en nuestro caso 0.95.
assurance	Probabilidad deseada de obtener el valor especificado de MoE, en nuestro caso 0.80.

El código de R quedaría definido como:

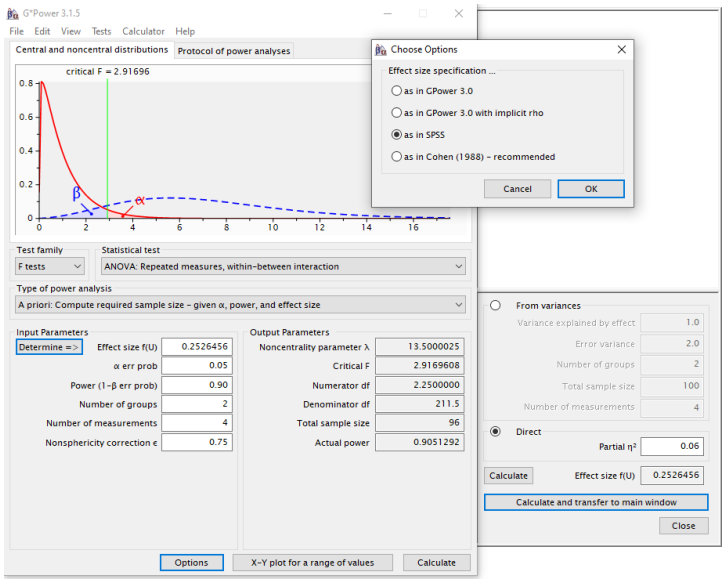
```
require(MBESS) # Cargar el paquete de R MBESS

ss.aipe.c.ancova(error.var.anova = 529, rho = 0.5, c.weights = c(-0.5,0.5), width = 4, conf.level = 0.95, assurance = 0.80) # Calcular el tamaño muestral
```

La función devolvería el tamaño muestral necesario por grupo para el estudio, que en este caso es de 203 sujetos por grupo. Asumiendo un 15% de pérdidas, la muestral final quedaría constituida por 203/0.85 = 238.82 ≈ 239 sujetos por grupo.

Ejemplo 2: Cálculo basado en potencia para un ANOVA en G*Power

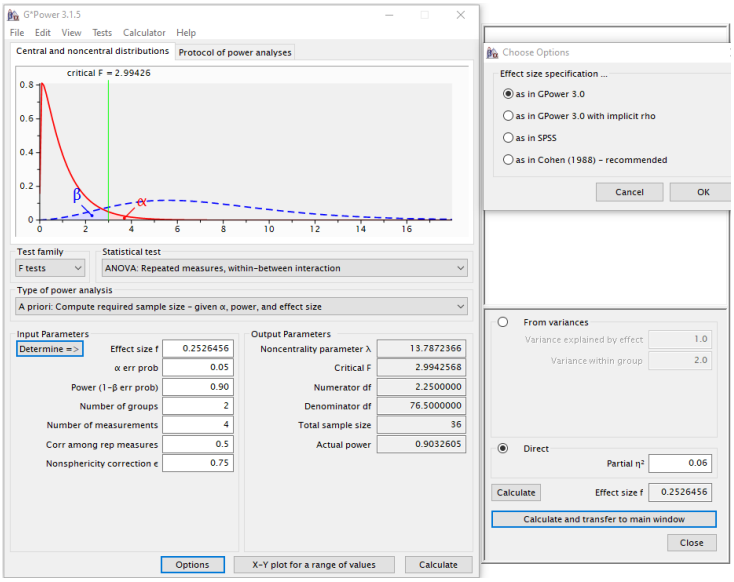
Se desea realizar un cálculo de tamaño muestral para un ANOVA mixto con 2 grupos y 4 mediciones. Se asume un eta cuadrado parcial de 0.06 en base a literatura previa publicada, especificándose el método según SPSS en G*Power, así como una potencia esperada del 90% y una corrección para el incumplimiento de la asunción de esfericidad de 0.75.



El tamaño muestral resultante es de 96 sujetos (48 por grupo). Asumiendo un 15% de pérdidas, el tamaño muestral final quedaría constituido por 96/0.85 = 112.9 ≈ 114 sujetos (57 por grupo).

Nota: Ejemplo de una práctica mal realizada con G*Power

Si, asumiendo los datos anteriores, se mantiene la especificación por defecto del tamaño del efecto según G*Power, el tamaño muestral resultante (asumiendo una correlación de 0.5) sería de solo 36 sujetos (18 sujetos por grupo), que asumiendo un 15% de pérdidas quedaría en 44 sujetos (22 por grupo), suponiendo una considerable infraestimación de la muestra necesaria.



EJEMPLO PRÁCTICO PARA ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

Ejemplo 1: Cálculo basado en precisión

Se desea realizar un cálculo de tamaño muestral para la para la precisión de los coeficientes de regresión estandarizados parcialmente de un análisis de regresión lineal múltiple con 4 parámetros predictores. Se asume como aceptable un MoE de 0.10, con una potencia para el mismo del 80% y un nivel de confianza del 95%. Además, se asumen los siguientes valores de correlación de los predictores con la variable dependiente y entre ellos, en función de literatura previa publicada:

Vector de correlaciones de los predictores con la variable dependiente:

$$Vector = [0.24, 0.30, 0.30, 0.30]$$

Matriz de correlaciones entre los predictores:

$$Matriz = \begin{bmatrix} 1 & & & \\ 0.24 & 1 & & \\ 0.24 & 0.30 & 1 & \\ 0.24 & 0.30 & 0.30 & 1 \end{bmatrix}$$

El cálculo se realizará con la función *'ss.aipe.reg.coef()'* del paquete de R *'MBESS'*,(Kelley & Maxwell, 2003) que requiere de la especificación de los siguientes parámetros:

Parámetros específicos en caso de introducir directamente los valores de R^2 y R^2_{XXj}	
Rho2.Y_X	Valor esperado de R^2.
Rho2.j_X.without.j	Valor esperado de R^2_{XXj}.
p	Número de parámetros predictores.
b.j	Valor del coeficiente de regresión para el predictor de interés (da igual el valor que se especifique ya que no influirá en el tamaño muestral, pero hay que especificarlo para que se realice el cálculo).
Parámetros específicos en caso de introducir el vector y matriz de correlaciones	
RHO.XX	Matriz de correlación entre los predictores.
Rho.YX	Vector de correlaciones de los predictores con la variable dependiente.
which.predictor	Posición (en la matriz de correlaciones) del predictor de interés para el que se desea realizar el cálculo. Debe especificarse el de mayor correlaciones en *RHO.XX*.
Parámetros genéricos a especificar en ambos casos	
width	Amplitud total del intervalo deseada (doble del MoE).
conf.level	Nivel de confianza del intervalo.
assuarence	Probabilidad deseada de obtener un valor igual o inferior al esperado del MoE

El código de R quedaría definido como:

```
require(MBESS) # Cargar el paquete de R MBESS

Rxx <- matrix(ncol=4, nrow = 4, data = c(1, 0.24, 0.24, 0.24, 0.24, 1, 0.30, 0.30, 0.24, 0.30, 1, 0.30, 0.24, 0.30, 0.30, 1)) # Crear la matriz de correlaciones entre predictores

pyx <- c(0.24, 0.30, 0.30, 0.30) # Crear el vector de correlaciones con la variable dependiente

ss.aipe.reg.coef(RHO.XX = Rxx, Rho.YX = pyx, width = 0.20, conf.level = 0.95, assurance = 0.80, which.predictor = 2) # Calcular el tamaño muestral
```

El tamaño muestral resultante sería de 408 sujetos, que asumiendo un 15% de pérdidas ascendería finalmente a 480 sujetos. Como se ha comentado, hay que especificar en *which.predictor* el predictor de mayor correlación en *RHO.XX*. Esto se debe a que, de no ser así, se infraestimará el valor esperado de R^2_{XXj} y por tanto el tamaño muestral. Por ejemplo, si se especifica el primer predictor como el de interés en la anterior función, la muestra resultante es de 387 sujetos, en comparación a los 408 necesarios para asegurar una precisión mínima de 0.20 en todos los predictores.

Por otro lado, si quisiéramos introducir directamente los valores de R^2 y R^2_{XXj}, como por ejemplo 0.30 y 0.40 respectivamente, entonces el código de R quedaría definido como:

```
require(MBESS)

ss.aipe.reg.coef(Rho2.Y_X = 0.30, Rho2.j_X.without.j = 0.40, p = 4, b.j = 1, width = 0.20, conf.level = 0.95, assurance = 0.80) # Calcular el tamaño muestral
```

Devolviendo la función una muestra mínima necesaria de 485 sujetos, que ascendería con un 15% de pérdidas a 571 sujetos.

Ejemplo 2: Cálculo basado en potencia

Se desea realizar un cálculo de tamaño muestral basado en potencia para los coeficientes de regresión de un análisis de regresión lineal múltiple con 4 parámetros predictores. Se asume como aceptable una potencia del 90% y un nivel alfa del 0.05. Además, se asumirá que la correlación de Pearson más pequeña entre alguno de los predictores y la variable dependiente es de 0.30, así como que las correlaciones entre los distintos predictores son iguales con un valor de 0.40.

Los cálculos se realizarán según la propuesta de Hsieh y cols. de 1998.(Hsieh et al., 1998) Se ha creado una función en R asumiendo que las correlaciones entre los distintos predictores son iguales, y que requiere de la especificación de los siguientes parámetros:

rj	Mínima correlación de Pearson entre los predictores y la variable dependiente.
rx	Correlación estimada entre predictores.
param	Número de parámetros predictores.
alfa	Umbral de significación estadística deseado (por defecto 0.05)
beta	1 – potencia estadística deseada (por defecto 0.10, 90% potencia)

El código de la función es el siguiente:

```
SampleSizePowerMLR <- function(rj,rx,param,alfa=0.05,beta=0.10){
 if (param==1) {
 Cr <- 0.5*log((1+rj)/(1-rj)) # Calcular la transformacion de Fisher
 n1 <- ((qnorm(1-alfa/2) + qnorm(1-beta))^2)/Cr^2 + 3 # Calcular n1
 return(paste("La muestra minima necesaria requerida es de ",ceiling(n1),"sujetos"))
 }
 else
 Cr <- 0.5*log((1+rj)/(1-rj)) # Calcular la transformacion de Fisher
 n1 <- ((qnorm(1-alfa/2) + qnorm(1-beta))^2)/Cr^2 + 3 # Calcular n1
 p <- c(rep(rx,param-1)) # Crear el vector de correlaciones entre predictores
 dimmatrix <- (param-1)*(param-1) # Crear la dimension de la matriz de predictores
 matrixx <- matrix(ncol=param-1, nrow=param-1, data = c(rep(rx,dimmatrix))) # Crear la matriz
 matrixx[col(matrixx)==row(matrixx)] = 1 # Especificar valor = 1 en la diagonal
 R2.xx <- t(p)%*%solve(matrixx)%*%p # Calcular R2xxj
 np <- n1/(1-R2.xx) # Calcular el tamaño muestral
 return(paste("La muestra minima necesaria requerida es de",ceiling(np),"sujetos"))
}
```

Con dicha función creada, el código de R para el cálculo de tamaño muestral quedaría definido como:

```
SampleSizePowerMLR(rj = 0.3, rx = 0.4, param = 4, alfa = 0.05, beta = 0.10)
```

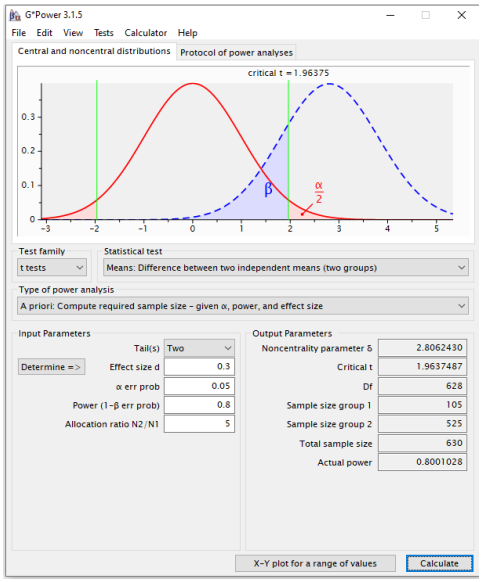
La función devolvería un tamaño muestral mínimo requerido de 154 sujetos, que ascendería asumiendo un 15% de pérdidas a 181 sujetos.

EJEMPLO PRÁCTICO PARA ANÁLISIS DE REGRESIÓN LOGÍSTICA BINARIA MÚLTIPLE

Ejemplo 1: Cálculo basado en una variable independiente continua

Se desea realizar un cálculo de tamaño muestral para un análisis de regresión logística binaria múltiple, donde el predictor de interés es una variable continua. Se asumió un tamaño del efecto medio (*d* de Cohen de 0.30) para dicha variable continua, un valor de R^2_{XXj} estimado de 0.30 y una prevalencia estimada del factor de interés de 0.20. Además, se asumieron como aceptables una potencia del 80% y un valor de alfa de 0.05.

El cálculo se realizará según la propuesta de Hsieh y cols. de 1998.(Hsieh et al., 1998) El primer paso será calcular la muestra necesaria para una prueba t-Student para dos muestras independientes, ajustando el ratio de asignación en función de la prevalencia estimada (1/0.2 = 5) en G*Power:



El segundo paso sería corregir dicho tamaño muestral para la correlación entre los predictores (R^2_{XXj}), dividiendo el tamaño muestral de G*Power entre $1-R^2_{XXj}$, es decir 630/0.7 = 900 sujetos. Finalmente, corrigiendo para un potencial 15% de pérdidas, el tamaño muestral final quedaría constituido por 1.059 sujetos.

Ejemplo 2: Cálculo basado en una variable independiente dicotómica

Se desea realizar un cálculo de tamaño muestral para un análisis de regresión logística binaria múltiple, donde el predictor de interés (X) es una dicotómica. Se asumió una prevalencia del factor de la variable predictora de 0.30, con una prevalencia del factor de la variable dependiente en X=0 de 0.10 y en X=1 de 0.20, además de un valor de R^2_{XXj} estimado de 0.10. Por último, se consideraron aceptables una potencia del 80% y un valor de alfa de 0.05.

Los cálculos se realizarán según la propuesta de Hsieh y cols. de 1998.(Hsieh et al., 1998) Se ha creado una función en R, que requiere de la especificación de los siguientes parámetros:

B	Prevalencia del factor de la variable predictora de interés.

P0	Prevalencia del factor de interés en X=0
P1	Prevalencia del factor de interés en X=1
R2.XXj	Valor esperado de R^2_{XXj}.
alfa	Umbral de significación estadística deseado (por defecto 0.05)
beta	1 – potencia estadística deseada (por defecto 0.20, 80% potencia)

El código de la función es el siguiente:

```
SampleSizePowerLogReg_binary <- function(B, P0, P1, R2.XXj, alfa = 0.05, beta = 0.20){
  P <- (1-B)*P0 + B*P1 # Calcular la prevalencia general del factor de interes
  z.alfa <- qnorm(1-alfa/2) # Calcular valor Z para alfa
  z.beta <- qnorm(1-beta) # Calcular valor Z para beta
  N <- ((z.alfa*(P*(1-P)/B)^0.5 + z.beta*(P0*(1-P0) + P1*(1-P1)*(1-B)/B)^0.5)^2)/(((P0-P1)^2)*(1-B)) # Calcular muestra inicial
  N.adj <- N/(1-R2.XXj) # Calcular muestra final corregida
  return(paste("La muestra minima necesaria requerida es de",ceiling(N.adj),"sujetos"))
}
```

Con dicha función creada, el código de R para el cálculo de tamaño muestral quedaría definido como:

```
SampleSizePowerLogReg_binary(B = 0.30, P0 = 0.10, P1 = 0.20, R2.XXj = 0.10)
```

La función devolvería un tamaño muestral mínimo requerido de 501 sujetos, que asumiendo un 15% de pérdidas quedaría finalmente constituido por 590 sujetos.

EJEMPLO PRÁCTICO PARA ANÁLISIS DE FIABILIDAD

Se desea calcular el tamaño muestral para un coeficiente de correlación intraclase (ICC), bajo la asunción de un modelo ICC(2,1), para un análisis de fiabilidad inter-examinador con cinco evaluadores y una única medición por sujeto. Se asumió una fiabilidad esperada de 0.85 en función de literatura previa, con un MoE aceptable de 0.05. Además, se asumió como aceptable una probabilidad del 80% para obtener un MoE igual o inferior al valor esperado y una confianza del 95%.

El cálculo se realizará adaptando la propuesta de Gwet, (Gwet, 2021) que requiere de la especificación de los siguientes parámetros:

p	Valor esperado del coeficiente de correlación intraclase.
r	Número de evaluadores.
IC	Amplitud total del intervalo de confianza al 95% (doble del valor del MoE).
assurance	Probabilidad deseada de obtener un valor igual o inferior al esperado del MoE (por defecto 0.80).

La función creada es una adaptación de las fórmulas de Gwet para el intervalo de confianza del ICC(2,1), que consiste en la iteración de cálculos de dicho intervalo para múltiples tamaños muestrales hasta obtener uno igual o inferior al deseado, con una potencia esperada (adición propia ya que Gwet no tiene este factor en consideración). Dado que la función se basa en muestreos aleatorios con el proceso iterativo, el tamaño muestral calculado puede variar ligeramente de una vez a otra (un sujeto arriba o abajo aproximadamente). El código de la función es el siguiente:

```
ICC21ic95 <- function(p, r, n, assurance){ # Calcular el intervalo de confianza del ICC(2,1)
  IC95 <- replicate(10000, {
    FSR <- rf(1, (r-1), ((r-1)*(n-1)))
    a <- r*p/(n*(1-p))
    b <- 1 + (r*(n-1)*p/(n*(1-p)))
    v <- (a*FSR + b)*(a*FSR + b)/(((a*FSR)*(a*FSR)/(r-1)) + (b*b/((r-1)*(n-1))))
f <- rf(1, (n-1), v)
FSS <- f*(a*FSR + b)
F1 <- qf(0.025, (n-1), v)
F2 <- qf(0.975, (n-1), v)
LCB <- (n*(FSS - F2))/(n*FSS + F2*(r*FSR + r*n - r - n))
UCB <- (n*(FSS - F1))/(n*FSS + F1*(r*FSR + r*n - r - n))
Diferencia <- UCB - LCB
Diferencia
```

```
  })
IC9580 <- mean(IC95) + qnorm(assurance)*sd(IC95)
c(n, IC9580)
}
SampleSizeICC21inter <- function(p,r,IC,assurance){ # Calcular el tamaño muestral
  for (i in c(10:1000)){
    SampleSizeICC21Results <- ICC21ic95(p, r, i, assurance)
    if(SampleSizeICC21Results[2] < IC){break}}
  return(paste("El tamaño muestral minimo necesario es de",SampleSizeICC21Results[1],"sujetos, con un IC al 95% esperado de",round(SampleSizeICC21Results[2],3)))
}
```

Con dicha función creada, el código de R para el cálculo de tamaño muestral quedaría definido como:

```
SampleSizeICC21inter(p = 0.85, r = 5, IC = 0.10, assurance = 0.80)
```

El tamaño muestral resultante es de 85 sujetos. La amplitud del IC al 95% (límite superior para un 80% de potencia) para este tamaño muestral siempre será ligeramente inferior a la especificada en la función y variará ligeramente de un cálculo a otro por el proceso iterativo, pero siempre será inferior al MoE especificado, de modo que el tamaño muestral no se infraestimará. Asumiendo finalmente un 15% de pérdidas, el tamaño muestral quedaría constituido por 100 sujetos.

EJEMPLO PRÁCTICO PARA MODELOS PREDICTIVOS MULTIVARIABLES

Se desea calcular el tamaño muestral necesario para elaborar un modelo predictivo multivariable del grado de discapacidad medida con el cuestionario *Shoulder Pain and Disability Index* (SPADI), al año de seguimiento, tras la aplicación de un programa de ejercicio terapéutico en pacientes con dolor relacionado con el manguito rotador. Se asumieron los siguientes valores:

- MMOE igual o inferior a 1.1 en la estimación de la desviación estándar residual.
- Un factor de *shrinkage* igual o superior a 0.90.
- Una diferencia entre el coeficiente R^2 y el coeficiente R^2 ajustado igual o inferior a 0.05.
- Un valor de R^2 esperado conservador de 0.30 en base a literatura previa publicada.
- Una media en el SPADI al año de seguimiento de 35, en función de literatura previa.
- Una desviación estándar en al año de seguimiento de 23 en el SPADI, en función de literatura previa.

Se valora incluir en el modelo 10 variables potenciales. Dos de dichas variables son multicategóricas, con 3 categorías, de modo que ello incrementa el número de parámetros potenciales a 12. Además, se valora la inclusión de asociaciones no lineales en 2 de las variables continuas, mediante el uso de polinomios fraccionarios, con un grado 2 de complejidad, así como la inclusión de un término de interacción entre otras dos variables continuas asumidas con relación lineal, de modo que el número final de parámetros potenciales sería de 15.

El cálculo se realizará con la función *'pmsampsize'* del paquete de R *'pmsampsize'*,(Riley et al., 2020) que require de la especificación de los siguientes parámetros:

type	Tipo de análisis de regresión ("c" = lineal, "b" = logística, o "s" = hazards proporcionales).
rsquared	Valor esperado de R^2.
parameters	Número de parámetros predictores potenciales.
shrinkage	Valor del factor de *shrinkage* (por defecto 0.90).
intercept	Media estimada de la variable dependiente.
sd	Desviación estándar estimada de la variable dependiente.
mmoe	Margen de error multiplicativo para la desviación estándar residual (por defecto 1.1).

El código de R quedaría definido como:

```
require(pmsampsize) # Cargar paquete pmsampsize

pmsampsize(type = "c", rsquared = 0.30, parameters = 15, shrinkage = 0.90, intercept = 35, sd = 23, mmoe = 1.1) # Calcular el tamaño muestral
```

El tamaño muestral resultante es de 322 sujetos. Sin embargo, dado que se incluirá una interacción, se decide incrementa en 100 sujetos más la muestra. Finalmente, asumiendo un 15% de pérdidas, la muestra estimada sería de 497 sujetos.

REFERENCIAS

Gwet, K. L. (2021). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Volume 2: Analysis of Quantitative Ratings* (5th ed.). AgreeStat Analytics.

Hsieh, F., Bloch, D., & Larsen, M. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, *17*(14), 1623–1634.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321. https://doi.org/10.1037/1082-989X.8.3.305

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: sample size planning via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology*, *65*(2), 350–370. https://doi.org/10.1111/J.2044-8317.2011.02029.X

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & Van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ (Clinical Research Ed.)*, *368*. https://doi.org/10.1136/BMJ.M441

Material Suplementario 2: Herramientas para el Cálculo del Tamaño Muestral

El Cálculo del Tamaño Muestral en Ciencias de la Salud: Recomendaciones y Guía Práctica

Correspondencia:

Rubén Fernández Matías, Fisioterapeuta

ruben.fernanmat@gmail.com

HERRAMIENTAS Y ENLACES PARA CÁLCULO DE TAMAÑO MUESTRAL

Herramienta	Tipo de cálculo	Enlace
Paquete de R *'MBESS'*	Precisión con probabilidad deseada	*CRAN – Package MBESS* Referencias: (Kelley, 2007, 2008; Kelley et al., 2018, 2019; Kelley & Lai, 2011; Kelley & Maxwell, 2003; Kelley & Rausch, 2006, 2011; Lai & Kelley, 2011, 2012; Terry & Kelley, 2012)
Paquete de R *'presize'*	Precisión	*CRAN – Package presize*
Paquete de R *'pwr'*	Potencia	*CRAN – Package pwr*
Paquete de R *'pwrss'*	Potencia	*CRAN – Package pwrss*
G*Power	Potencia	Enlace al Software Manual de utilización
Aplicación Shiny de Schoemann, Boulton & Short	Potencia para análisis de mediación	App Shiny Referencias: (Schoemann et al., 2017)

REFERENCIAS

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*(4), 755–766. https://doi.org/10.3758/BF03192966

Kelley, K. (2008). Sample Size Planning for the Squared Multiple Correlation Coefficient: Accuracy in Parameter Estimation via Narrow Confidence Intervals. *Multivariate Behavioral Research*, *43*(4), 524–555. https://doi.org/10.1080/00273170802490632

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2019). Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods*, *24*(4), 492–515. https://doi.org/10.1037/MET0000203

Kelley, K., & Lai, K. (2011). Accuracy in Parameter Estimation for the Root Mean Square Error of Approximation: Sample Size Planning for Narrow Confidence Intervals. *Multivariate Behavioral Research*, *46*(1), 1–32. https://doi.org/10.1080/00273171.2011.543027

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321. https://doi.org/10.1037/1082-989X.8.3.305

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. https://doi.org/10.1037/1082-989X.11.4.363

Kelley, K., & Rausch, J. R. (2011). Sample size planning for longitudinal models: accuracy in parameter estimation for polynomial change parameters. *Psychological Methods*, *16*(4), 391–405. https://doi.org/10.1037/A0023352

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*(2), 127–148. https://doi.org/10.1037/a0021764

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: sample size planning via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology*, *65*(2), 350–370. https://doi.org/10.1111/J.2044-8317.2011.02029.X

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. *Social Psychological and Personality Science*, *8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Terry, L., & Kelley, K. (2012). Sample size planning for composite reliability coefficients: accuracy in parameter estimation via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology*, *65*(3), 371–401. https://doi.org/10.1111/J.2044-8317.2011.02030.X