

# Análisis de poder estadístico y cálculo de tamaño de muestra en *R*: Guía práctica



Opciones gratuitas y abiertas, con énfasis en los paquetes {pwr} y {Superpower} para R

Juan David Leongómez\*

08 septiembre, 2020 (revisión 06 septiembre, 2023)

## Descripción

Esta guía práctica acompaña la serie de videos [Poder estadístico y tamaño de muestra en R](#), de mi canal de YouTube [Investigación Abierta](#), que recomiendo ver antes de leer este documento. Contiene una explicación básica del análisis de poder estadístico y cálculo de tamaño de muestra, centrándose en el procedimiento para realizar análisis de poder y tamaños de muestra en *jamovi* y particularmente en *R*, usando los paquetes [pwr](#) (para diseños sencillos) y [Superpower](#) (para diseños factoriales más complejos). La sección dedicada a [pwr](#) está ampliamente basada en [este video](#) de Daniel S. Quintana (2019).

**Fuentes y citas:** Con la intención de sustentar claramente, pero de forma sencilla, la información presentada, incluyo varias citas a lo largo del documento que, creo, podrían servir a estudiantes, docentes e investigadores para explorar un tema particular con mayor profundidad, o soportar una decisión en un proyecto de investigación. Las referencias completas de todas las citas (incluyendo hipervínculos a las fuentes), están al final del documento.

**Elementos interactivos:** Este documento tiene vínculos para facilitar la navegación. En [rojo](#), aparecen los vínculos a citas, notas al pie de página y secciones dentro del texto (referencias cruzadas), y en [azul](#), los vínculos a páginas y documentos externos.

## Contenidos

<b>1</b>	<b>Análisis de poder y tamaño de muestra</b>	<b>2</b>
1.1	¿Qué es potencia o poder estadístico?	2
1.2	Cómo determinar el tamaño de muestra necesario	3
1.3	Cómo estimar un tamaño del efecto esperado para calcular el tamaño de muestra	3
1.3.1	Por qué no es buena idea usar las definiciones de Cohen (tamaños de efecto “pequeños”, “medios” o “grandes”)	3
1.3.2	Técnicas comunes y sus limitaciones:	5
1.3.3	Alternativas	5
<b>2</b>	<b>G*Power</b>	<b>6</b>
<b>3</b>	<b><i>jpower</i> (módulo de <i>jamovi</i>)</b>	<b>6</b>
3.1	Tamaño del efecto en <i>jpower</i> para <i>jamovi</i> : $\delta$	7

---

\*Laboratorio de Análisis del Comportamiento Humano, Facultad de Psicología, Universidad El Bosque, Bogotá, Colombia. Información y datos de contacto disponibles desde mi sitio web [jdleongomez.info](http://jdleongomez.info).

<b>4</b>	<b>Para paquetes de R: instalación de R y RStudio</b>	<b>7</b>
<b>5</b>	<b>Datos usados en los ejemplos de paquetes de R</b>	<b>8</b>
<b>6</b>	<b>Paquete <i>pwr</i> para R (Diseños sencillos)</b>	<b>8</b>
6.1	Instalación y carga de <i>pwr</i>	8
6.2	Niveles de referencia de los tamaños del efecto. Función <i>cohen.ES</i>	9
6.3	Análisis de poder para correlaciones. Función <i>pwr.r.test</i>	9
6.3.1	Si hay hipótesis claras, pre-especificadas ( <i>a priori</i> ), se pueden hacer análisis de una cola	10
6.4	Análisis de poder para pruebas- <i>t</i> . Función <i>pwr.t.test</i>	11
6.4.1	Muestras independientes	12
6.4.2	Una muestra	12
6.4.3	Medidas repetidas o pareadas	13
6.4.4	Análisis de una cola	13
6.5	Análisis de poder para ANOVAs de una vía. Función <i>pwr.anova.test</i>	15
<b>7</b>	<b>Paquete <i>Superpower</i> para R (diseños factoriales complejos)</b>	<b>16</b>
7.1	Diseños factoriales	16
7.2	Instalación y carga de <i>Superpower</i>	17
7.3	Acerca de comparaciones <i>post-hoc</i>	17
7.3.1	Cómo controlar la tasa de errores al hacer pruebas <i>post-hoc</i> . Correcciones de <i>Bonferroni</i> y <i>Holm-Bonferroni</i>	18
7.4	ANOVA factorial de medidas independientes	19
7.4.1	Ejemplo ANOVA factorial de medidas independientes ( $2 \times 2$ )	19
7.5	ANOVA factorial de medidas repetidas	22
7.5.1	Ejemplo ANOVA factorial de medidas repetidas ( $2 \times 2$ )	23
7.6	ANOVA factorial mixto	26
7.6.1	Ejemplo ANOVA factorial mixto ( $2 \times 2$ )	26
7.7	Extra: Cómo estima <i>Superpower</i> el poder estadístico con base en simulaciones de bases de datos	29
<b>8</b>	<b>Referencias</b>	<b>32</b>
	<b>Agradecimientos</b>	<b>33</b>
	<b>Acerca de este trabajo</b>	<b>33</b>
	<b>APÉNDICE: Paquetes de R usados en la creación de este documento</b>	<b>34</b>

# 1 Análisis de poder y tamaño de muestra

## 1.1 ¿Qué es potencia o poder estadístico?

Básicamente es la probabilidad de encontrar un resultado estadísticamente significativo, dado un tamaño de muestra y un valor  $\alpha$  (alfa, o nivel de significación estadístico).

Por ejemplo, si yo hipotéticamente tuviera un poder de 0.8 (80%, que es el mínimo aceptado en la literatura), e hiciera 100 estudios, en promedio 80 de éstos serían significativos, y 20 serían no significativos.

Esto es muy importante, pues en áreas como las neurociencias el poder suele estar alrededor de 0.1 (10%) o 0.2 (20%) (ver [Button et al., 2013](#)). Y, lastimosamente, la situación no es mucho mejor en las ciencias del comportamiento en general, ni en otras disciplinas. Al usar un poder demasiado bajo, la probabilidad de error es muy alta, lo que genera conocimiento poco confiable, por dos razones:

1. Al tener un bajo poder estadístico, la probabilidad de Error Tipo II aumenta (alta tasa de falsos negativos). Es decir, muchas veces se rechaza la hipótesis alternativa erróneamente.
2. Al tener tamaños de muestra ( $n$ ) pequeños, los modelos estadísticos tienden a ser inestables. En otras palabras, los hallazgos son poco replicables pues los resultados varían mucho (o incluso se contradicen) entre diferentes

muestras. Este es uno de los factores de lo que en ciencia se conoce como *Crisis de replicación* (ver e.g. Baker, 2016; Loken & Gelman, 2017).

■ **Es probable que la literatura, en muchas disciplinas científicas, esté llena de resultados falsos.**

## 1.2 Cómo determinar el tamaño de muestra necesario

Existen cuatro parámetros esenciales y relacionados entre sí:

1.  **$n$ :** tamaño de muestra.
2.  **$\alpha$  (alfa o nivel de significación estadística):** es la probabilidad de Error Tipo I (falso positivo<sup>1</sup>), que es detectar un efecto que no existe (es decir, un efecto que no existe, ¡pero obtengo un resultado significativo!). En otras palabras, con un  $\alpha = 0.05$ , acepto que el 5% de los resultados sean falsos positivos. Es entonces el valor bajo el cual consideramos que un valor  $p$  es significativo, y asumimos que hemos detectado un efecto (rechazando la hipótesis nula). Aunque típicamente se usa un  $\alpha = 0.05$ , este valor puede ser ajustado. En algunas disciplinas, por ejemplo, se usan valores de  $\alpha$  distintos, y se ha argumentado que el  $\alpha$  debe ser justificado (Lakens, Adolphi, et al., 2018), o el que se usa por defecto, debería ser cambiado a, por ejemplo, 0.005 (Benjamin et al., 2018).
3. **Tamaño del efecto:** medida estandarizada de la fuerza de un fenómeno o magnitud de una relación entre variables (por ejemplo,  $r$ ,  $d$  de Cohen,  $f$  de Cohen o  $\eta^2$ ).
4. **Poder estadístico (también llamado potencia estadística):** se define como  $1 - \beta$ , donde  $\beta^2$  es la probabilidad de Error Tipo II (falso negativo), que es no detectar un efecto que sí existe (es decir, estudio un efecto que sí existe, ¡pero obtengo un resultado no significativo!). Como es una probabilidad, con valores entre 0 y 1, se puede convertir en porcentaje multiplicando por 100. Entonces, si yo considero que es aceptable que en el 10% de los casos yo no detecte un error que sí existe ( $\beta = 0.1$ ), mi poder ( $1 - \beta$ ) sería de 0.9, o 90%. En este caso, tendría 90% de probabilidades de detectar, como significativo, un efecto, si es éste existe.

Si se saben tres de estos cuatro valores, ya que están interrelacionados, se puede calcular el otro *a priori* (es decir, antes de hacer un estudio). Hacerlo es importante pues promueve que mis resultados sean confiables, evita que yo haga estudios con muy bajo poder que me lleven a conclusiones erróneas, y permite estimar la duración y costo de un estudio.

■ **Entonces, si conozco (1) el  $\alpha$ , (2) el tamaño del efecto esperado, y (3) el poder estadístico deseado ( $1 - \beta$ ), puedo calcular el tamaño de muestra necesario.**

## 1.3 Cómo estimar un tamaño del efecto esperado para calcular el tamaño de muestra

Normalmente, determinar el  $\alpha$ , y el poder estadístico deseado no es problemático. Se suele usar un  $\alpha = 0.05$ , y un poder mínimo de 80% ( $1 - \beta = 0.8$ ), aunque siempre sería mejor usar 90% ( $1 - \beta = 0.9$ ) si esto es posible, y teniendo en cuenta que esto aumentará el tamaño de muestra necesario. Entonces, si ya tengo el  $\alpha$ , y el poder estadístico deseado, el único problema es saber qué tamaño del efecto usar (Figura 1). Para esto, sin embargo, no hay una única respuesta, ni una que carezca de limitaciones.

### 1.3.1 Por qué no es buena idea usar las definiciones de Cohen (tamaños de efecto “pequeños”, “medios” o “grandes”)

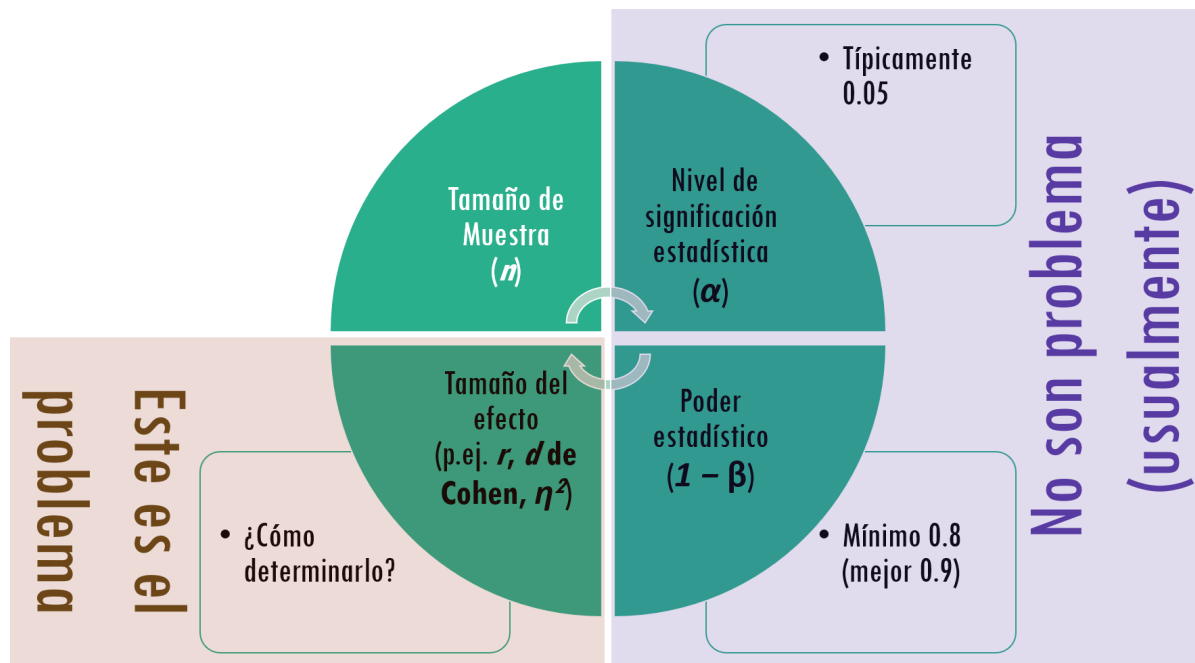
Cohen (1988, 1992) propuso unas definiciones para las medidas de tamaños del efecto (“*tamaños de camiseta*”: efectos “pequeños”, “medios” y “grandes”). Consciente de las limitaciones, él mismo advierte que su uso debe ser cuidadoso y que su utilidad es relativa. Cohen planteó estas definiciones solo como último recurso, cuando no hay evidencia previa que permita al investigador estimar el tamaño de un efecto que va a estudiar. Sin embargo, el uso indiscriminado y poco reflexivo de estas definiciones es terriblemente frecuente.

El uso indiscriminado de las definiciones de las “*camisetas*” es problemático por dos razones fundamentales: las definiciones son arbitrarias y son inconsistentes.

A pesar de que la mayoría de los tests estadísticos comunes (pruebas- $t$ , correlaciones, regresiones lineales simples y múltiples, ANOVAs y demás), hacen parte del mismo grupo de tests (modelos lineales generales), existe una

<sup>1</sup>Es triste decir esto, pero le ruego, especialmente a los colombianos, no confundir este concepto investigativo con la horrenda y mal llamada práctica de *falsos positivos* (asesinato de civiles para luego clasificarlos como delincuentes), que se dio en el marco del conflicto armado colombiano.

<sup>2</sup>No confundir con el coeficiente estandarizado de una regresión.



**Figura 1.** Parámetros necesarios para calcular el tamaño de muestra necesario, para obtener un poder estadístico deseado.

variedad de tamaños de efecto estandarizados. Por ejemplo, típicamente se usa  $d$  de Cohen para pruebas  $t$ ,  $r$  para correlaciones y regresiones simples,  $f$  de Cohen o  $\eta^2$  para ANOVAs, y  $f^2$  para regresiones múltiples<sup>3</sup>. Esta variedad de tamaños de efecto hace que sea difícil su comparación y su comprensión<sup>4</sup>, y cada una tiene definiciones de “pequeño”, “medio” y “grande” no congruentes (ver **Tabla 1**), que incluso llevan a investigadores a calcular diferentes tamaños de muestra para el mismo estudio, dependiendo de la prueba estadística que se va a usar<sup>5</sup> (Correll et al., 2020). En la **Tabla 1**, se ven las diferencias entre tamaños de efecto comunes, de acuerdo con el porcentaje de varianza explicada.

**Tabla 1.** Diferencias en la varianza explicada por las definiciones de Cohen para diferentes tamaños de efecto.

Definición de Cohen	Prueba- $t$		ANOVA/ANCOVA		Correlación/Regresión Simple		Regresión múltiple	
	$d$ de Cohen	Varianza explicada	$f$ de Cohen	Varianza explicada	$r$ de Pearson	Varianza explicada	$f^2$ de Cohen	Varianza explicada
<b>Pequeño</b>	0.2	1%	0.10	1%	0.1	1%	0.20	2%
<b>Medio</b>	0.5	6%	0.25	6%	0.3	9%	0.15	13%
<b>Grande</b>	0.8	16%	0.40	14%	0.5	25%	0.35	26%

*Nota:*

Las diferencias en la varianza explicada muestran que, aunque todas estas pruebas están relacionadas al ser todas modelos lineales generales, las definiciones propuestas por Cohen como tamaños de efecto "pequeños", "medios" o "grandes" difieren. Esto termina en que, para unos mismos datos, un análisis de poder termina por sugerir tamaños de muestra distintos, según el tamaño del efecto usado.

Recientemente, Correll et al. (2020) han hecho una profunda e interesante revisión de los análisis de poder y cálculo

<sup>3</sup>El  $f^2$  puede ser calculado de manera sencilla a partir del  $R^2$  de una regresión, con la ecuación  $f^2 = \frac{R^2}{1-R^2}$  (ver Selya et al., 2012).

<sup>4</sup>Correll et al. (2020) han sugerido usar un único tamaño de efecto para todas estas pruebas, y sugieren que sea  $\eta^2$  (eta al cuadrado) dada su relativa facilidad para ser interpretado, su generalizabilidad, y que puede ser aplicado tanto a predictores continuos (problemáticos para medidas como  $d$  y  $f$ ), como a predictores categóricos (problemáticos para  $r$ ).

<sup>5</sup>Es común que unos datos puedan ser analizados de más de una manera; por ejemplo, puedo hacer una prueba- $t$  o una regresión, que me darían el mismo resultado para los mismos datos, pero las definiciones de “pequeño”, “medio” y “grande” me sugerirían un tamaño de muestra diferente para obtener el mismo poder estadístico. De hecho, por ejemplo, si quiero obtener un efecto “medio” con una prueba- $t$  ( $d$  de Cohen = 0.5), el análisis de poder me sugeriría un  $n = 128$ , muestras que una regresión ( $r = 0.30$ ), me sugeriría un  $n = 82$ . ¡Para analizar los mismos datos!

de tamaño de muestra, resaltando las limitaciones de las técnicas comunes, así como sus alternativas.

### 1.3.2 Técnicas comunes y sus limitaciones:

1. **Definiciones de Cohen (tamaños de efecto “pequeños”, “medios” o “grandes”).** Fueron diseñados por Cohen (1988, 1992) solo como una guía cuando definitivamente no se tiene idea del tamaño del efecto que se va a estudiar. Sin embargo, son terriblemente problemáticos (ver Correll et al., 2020), y lastimosamente suelen usarse de manera casi automática. Además de la inconsistencia entre definiciones para diferentes medidas de tamaño del efecto (Tabla 1), su principal y más obvia limitación es que lo que se considera un efecto “pequeño” o “grande” cambia substancialmente entre disciplinas. Las definiciones propuestas por Cohen son percentiles (o cuartiles) de tamaños de efecto (25% percentil, 50% percentil o mediana, y 75% percentil, como tamaños de efecto “pequeños”, “medios” o “grandes”, respectivamente). Si uno mira la literatura publicada, estos valores tienden a ser más o menos ciertos, pero dado que la literatura tiene sesgos, y que muchos se basan de por sí en estas definiciones, es difícil saber qué tan acertado es este método en un caso particular. En todo caso, estas definiciones no tienen por qué ser relevantes para un estudio que yo esté diseñando. Por esto, no se deben usar de manera indiscriminada, pues es muy posible que no sean válidos para el estudio que estoy diseñando. De hecho, recientemente, se ha argumentado que su uso no solo debe evitarse, sino que debe ser descartado completamente (Correll et al., 2020).
2. **Usar el tamaño del efecto de un único estudio previo.** Tiene un gran potencial de ser sesgado, dados los sesgos de publicación (los resultados nulos tienden a no ser publicados).
3. **Hacer un estudio piloto.** Esta técnica tiene sesgos implícitos y tiende a crear tamaños de efecto inflados, como ha sido demostrado (Albers & Lakens, 2018). Sin embargo, si no hay estudios previos, podría ser una última opción para tener una indicación del poder estadístico esperado, pero asumiendo y reconociendo sus múltiples limitaciones.

### 1.3.3 Alternativas

1. **Ver la distribución de tamaños para un efecto particular:** Daniel S. Quintana, investigador en Psiquiatría Biológica de la Universidad de Oslo en Noruega propuso, cuando sea posible, ver la distribución de tamaños del efecto en un campo de estudio (Quintana, 2017). En su artículo, Quintana analizó casi 300 estudios (y tamaños de efecto), para el campo de variabilidad de la frecuencia cardíaca, y calculó la distribución con base en percentiles (25%, 50% o mediana, y 75%, a la manera de Cohen, pero aplicado directamente a su campo de estudio). Sin embargo, como él mismo menciona en este video, esta técnica, aunque tiende a ser menos sesgada que basarse en un único estudio previo, es todavía sujeta al sesgo que tengan los artículos en los que se basa, al igual que ocurre con meta-análisis.

Entonces, aunque tiene limitaciones, esto es mejor que usar un único estudio publicado, usar un estudio piloto, o las definiciones de Cohen para estimar el tamaño del efecto que estoy estudiando. Sin embargo, puede no ser posible, cuando se trata de campos de estudio donde pocos estudios, o ninguno, han mirado el efecto que quiero estudiar, con diseños comparables.

De nuevo, no hay una solución sencilla. Idealmente, el efecto que quiero usar para calcular el tamaño de muestra, debería ser el mismo (o muy cercano) al que de hecho encuentre al analizar mis datos.

2. **Determinar el “menor tamaño de efecto de interés”:** El “menor tamaño de efecto de interés” (en inglés, *smallest effect size of interest* o *SESOI*) es el tamaño mínimo de un efecto que se consideraría tiene importancia real. Esto se puede hacer tanto de manera objetiva, como subjetiva (Lakens, Scheel, et al., 2018). En ese caso, se deberían rechazar efectos menores a ese mínimo justificado<sup>6</sup>.

Esta es posiblemente la mejor opción, pues tiene en cuenta no solamente cuál es el tamaño de efecto esperado, sino también cuál es el mínimo tamaño de efecto que pueda resultar relevante para un efecto particular.

Por otra parte, es importante saber que muchos tamaños de efecto pueden ser comparables. Correll et al. (2020) proponen usar siempre el eta al cuadrado ( $\eta^2$ ) como tamaño del efecto, y presentan las ecuaciones, bastante sen-

<sup>6</sup>Esto, sin embargo, no permite encontrar soporte para una hipótesis nula para lo cual se deben hacer tests de equivalencia (ver Lakens, 2017), algo que excede el enfoque de este documento.

cillas<sup>7</sup>, para transformar medidas comunes de tamaños de efecto propuestos por Cohen, desde o hacia  $\eta^2$  (Tabla 2).

**Tabla 2.** Conversión entre  $\eta^2$  (eta al cuadrado) y medidas de tamaño de efecto de Cohen

	Tamaño de efecto de Cohen a $\eta^2$	$\eta^2$ a tamaño de efecto de Cohen
<b>Prueba-<i>t</i></b>	$\eta^2 = \frac{d^2}{d^2+4}$	$d = \frac{2\sqrt{\eta^2}}{\sqrt{1-\eta^2}}$
<b>Correlación/Regresión Simple</b>	$\eta^2 = r^2$	$r = \sqrt{\eta^2}$
<b>ANOVA/ANCOVA</b>	$\eta^2 = \frac{f^2}{1+f^2}$	$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$
<b>Regresión Múltiple</b>	$\eta^2 = \frac{f^2}{1+f^2}$	$f^2 = \frac{\eta^2}{1-\eta^2}$

*Nota:*

Las ecuaciones de esta tabla reproducen las presentados en la Tabla 2 de Correll et al. (2020).

A continuación, presentaré algunas opciones de Software gratuito para análisis de poder estadístico y tamaño de muestra.

## 2 G\*Power

Probablemente la opción más común para hacer análisis de poder estadístico y cálculos de tamaño de muestra es **G\*Power** (Faul et al., 2007, 2009), un software gratuito y relativamente sencillo. Sin embargo, la terminología y documentación del programa son confusas y se prestan para errores. De hecho, Correll et al. (2020) afirman que G\*Power fomenta el uso de las definiciones de Cohen, pues permite al investigador seleccionar una definición de tamaño con mínima consideración de los temas relevantes, y sin tener en cuenta sus numerosas y demostradas limitaciones (para una discusión de las limitaciones de las definiciones de Cohen, ver sección 1.3.1 **Por qué no es buena idea usar las definiciones de Cohen**). Esto, sin embargo, no es problema siempre y cuando el usuario del programa entienda los problemas de usar las definiciones de Cohen, y evite o justifique muy bien su uso.

Adicionalmente, aunque menos importante, G\*Power tiene limitaciones en cuanto a los diseños para los cuales se pueden hacer análisis, en especial cuando se trata de diseños factoriales, para los cuales sólo se puede hacer análisis para efectos principales, o para una interacción, pero solo para un efecto a la vez<sup>8</sup>.

Sin embargo, como ventaja, permite hacer análisis para una variedad de pruebas estadísticas.

## 3 *jpower* (módulo de *jamovi*)

Como alternativa, en *jamovi* existe el paquete *jpower*, extremadamente claro y fácil de usar pero, hasta el momento, solo permite análisis para diseños tipo prueba-*t* (aunque está en permanente desarrollo, y se está trabajando en incluir análisis de poder para otros tipos de diseño). Adicionalmente, aunque hace unos reportes muy claros y profesionales, que bien podrían ser adjuntados a cualquier proyecto, por ahora solo los produce en inglés.

Si quieres ver mis videos acerca de *jamovi*, y cómo instalarlo y usarlo (con énfasis en la transición desde *SPSS*), haz clic [acá](#).

<sup>7</sup>Son ecuaciones sencillas cuyos cálculos se pueden hacer en cualquier calculadora decente de celular o computador, o en *Excel*, por ejemplo.

<sup>8</sup>En contraste, el paquete **Superpower** es capaz de hacer análisis de poder para diseños más complejos, y permite ver en un solo análisis los tamaños de muestra necesarios para efectos principales e interacciones. El uso de este paquete se describe en la sección 7 **Paquete Superpower para R**

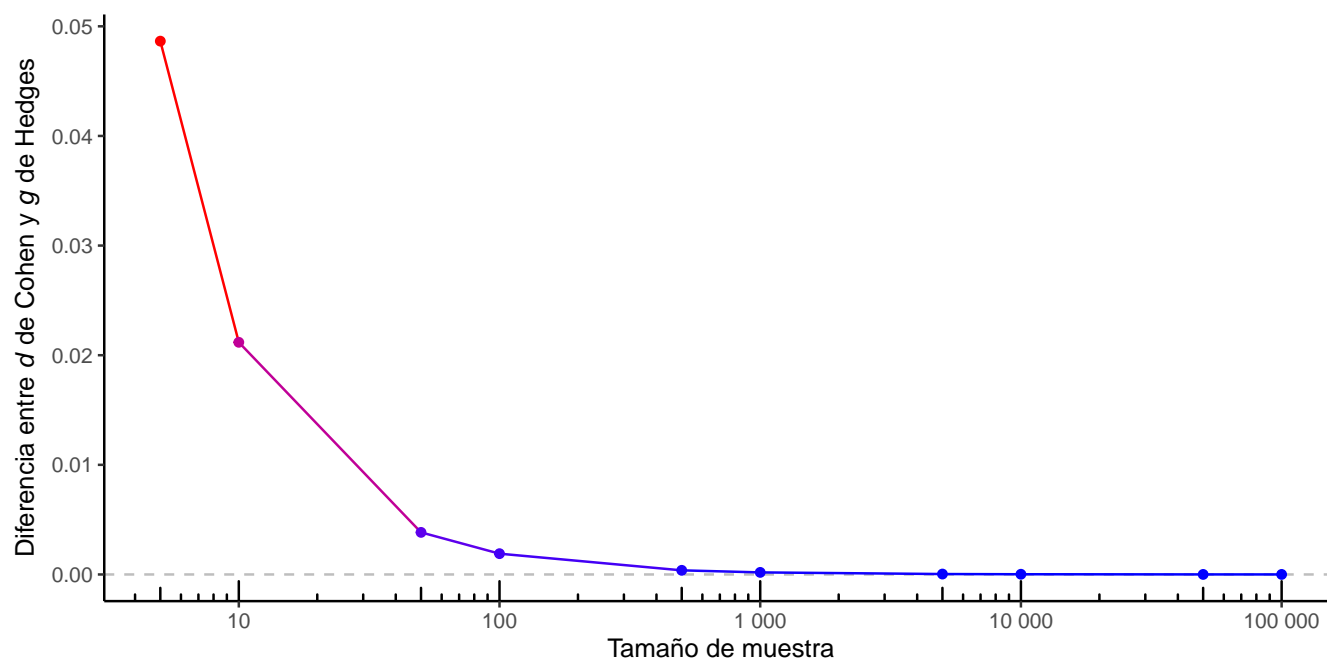
### 3.1 Tamaño del efecto en *jpower* para jamovi: $\delta$

*jpower* usa, como tamaño del efecto para pruebas-*t*, el  $\delta$  (delta) que es como se suele representar la diferencia estandarizada entre dos medias a nivel de una población entera.

Es completamente equivalente a la *d* de Cohen (o *g* de Hedges) pero, a diferencia de la *d* de Cohen, hace referencia al tamaño de efecto en la población entera, en vez de al efecto encontrado en una muestra.

Entonces, dado que el  $\delta$  es un tamaño de efecto completamente equivalente a la *d* de Cohen (o *g* de Hedges), pero que hace referencia a una población, *jamovi* nos está dando una pista: más que simplemente poner la *d* encontrada en un estudio anterior, sería ideal estimar el tamaño de efecto que esperamos (hipotetizamos) en la población entera, o el tamaño de efecto mínimo que esperamos que exista en la población entera, para estimar un tamaño de muestra que nos dé un poder estadístico suficiente para ese efecto (y, obviamente, un poder mayor si el efecto fuese mayor).

Es importante tener en cuenta que, aunque la *d* de Cohen es el tamaño del efecto más común para diferencias estandarizadas entre dos medias, tiende a proporcionar estimaciones sesgadas cuando se tienen tamaños de muestra pequeños<sup>9</sup> (**Figura 2**). Por este motivo, la *g* de Hedges es siempre una alternativa valiosa que encontrarás en muchos artículos (y que, sería mejor usar en todos los casos).



**Figura 2.** Ejemplo de la diferencia entre el efecto estimado como *d* de Cohen versus el efecto estimado como *g* de Hedges (eje Y), según el tamaño de muestra (eje X). En muestras pequeñas la *d* de Cohen estima un efecto superior a la *g* de Hedges (en rojo), pero a medida que el tamaño de muestra aumenta (en azul), la diferencia desaparece y tiende a 0 (línea gris). *Nota:* en este ejemplo, la diferencia entre las dos medias es 2 y la desviación estándar es de 4.

## 4 Para paquetes de R: instalación de R y RStudio

R puede parecer intimidante al principio, pues requiere comandos (o “scripting”), y a veces incluso algo de programación. Sin embargo, es probablemente la mejor herramienta que existe para hacer cualquier tipo de análisis de datos, y con certeza supera por mucho a los *software* comerciales (como *SPSS* y *Stata*), que además suelen ser costosos.

<sup>9</sup>Esto se debe a que tanto la *d* de Cohen como la *g* de Hedges agrupan las varianzas (asumen que los grupos tienen la misma desviación estándar), pero la *g* de Hedges agrupa utilizando  $n - 1$  para cada muestra en lugar de  $n$ , lo que proporciona una mejor estimación, especialmente cuanto menor es el tamaño de las muestras.



Primero, *R* es un software completamente **libre y gratuito** y, al ser abierto, sus avances y funciones no están limitados por lo que una compañía implemente (típicamente con fines comerciales y sujeto a las leyes de oferta y demanda), sino que depende directamente del trabajo colaborativo, y sin ánimo de lucro, de millones de personas alrededor del mundo; por esto, avanza más rápido que cualquier *software* comercial, y siempre tiene opciones para poder implementar las técnicas más novedosas<sup>10</sup>. Segundo, al ser en últimas un lenguaje de programación, permite hacer prácticamente cualquier cosa que se pueda hacer en un computador<sup>11</sup>, explotando realmente sus capacidades.

No puedo enfatizar de manera suficiente cuánto recomiendo, a cualquier persona que trabaje con datos, aprender a usar *R* (u otro lenguaje comparable, como *Phyton*), tanto por la calidad y eficiencia de los análisis, como por las posibilidades y ventajas laborales que esto da en el mundo de hoy.

Para usar los paquetes de *R*, se necesita por supuesto tener el programa instalado. Así mismo, recomiendo instalar y usar *R* a través de *RStudio*, pues facilita enormemente la interacción con *R*.

Para instalar *R* y/o *RStudio*, hay muchos tutoriales. Si tienes que hacerlo, te recomiendo buscar videos en YouTube, donde hay variedad de opciones (ninguna creada por mí hasta ahora) para todos los sistemas operativos, incluyendo Windows (por ejemplo [este video](#)), Mac (por ejemplo [este video](#)), y Ubuntu (por ejemplo [este video](#)).

En las siguientes secciones, el código de *R* está siempre resaltado sobre un fondo oscuro. Para correr estos códigos, puedes sencillamente copiarlos y pegarlos en un *script*<sup>12</sup>, o directamente en la consola de *R*.

## 5 Datos usados en los ejemplos de paquetes de *R*

Todos los datos usados en los ejemplos de cómo usar las funciones descritas en las siguientes secciones son generados de manera aleatoria, y en ningún caso buscan representar estudios realizados, hechos, o situaciones reales. En todos los casos, son datos creados y usados con el único propósito de ilustrar de manera clara y didáctica el proceso de realizar algunos análisis de poder, y calcular el tamaño de muestra necesario para estudios hipotéticos.

## 6 Paquete *pwr* para *R* (Diseños sencillos)

*pwr* ([Champely et al., 2020](#)) es un paquete flexible, claro, y muy confiable. Sin embargo, es importante señalar que, al menos por ahora, solo permite hacer análisis para diseños sencillos, incluyendo correlaciones y regresiones (simples o múltiples), pruebas-*t*, y ANOVAs de una vía (hasta el momento no permite hacer análisis para diseños factoriales).

A continuación, mostraré algunos ejemplos de funciones para hacer análisis de poder, y sus argumentos.

### 6.1 Instalación y carga de *pwr*

Para instalar y cargar *pwr*, se requiere correr las siguientes funciones:

1. Para instalar: `install.packages`, poniendo como argumento el nombre del paquete a instalar (en este caso, “*pwr*”) **entre comillas**. Una vez el paquete ha sido instalado, no es necesario volver a usar esta función, excepto si el paquete se quiere reinstalar, o actualizar en un computador.
2. Para cargar: `library`, poniendo como argumento el nombre del paquete a cargar. Acá no es necesario poner el nombre entre comillas (pero no hay problema si se hace).

En este caso, para instalar y cargar *pwr*, los comandos serían:

```
install.packages("pwr") #para instalar el paquete (no es necesario si ya ha sido instalado).
library(pwr) #para cargar el paquete una vez instalado.
```

<sup>10</sup>Prácticamente cada *nerd* que desarrolla una nueva prueba o técnica de análisis de datos, hace, además de una publicación, un paquete para *R*.

<sup>11</sup>Solo como un ejemplo, este documento en PDF, [mi sitio web](#), y [mi hoja de vida](#), fueron todos creados en *R*.

<sup>12</sup>[Este video](#), hecho por [Juan Carlos Correa \(2020\)](#), es un ejemplo de introducción al uso de *scripts* de *R* pero por supuesto hay muchos disponibles.



## 6.2 Niveles de referencia de los tamaños del efecto. Función `cohen.ES`

El paquete *pwr* tiene una función que podría resultar útil cuando no hay información respecto al tamaño del efecto esperado, pero que debe ser usada con cautela y buen criterio<sup>13</sup>. Se trata de la función `cohen.ES`, que permite rápidamente saber cuáles son las definiciones de referencia para diferentes tamaños del efecto. Esta sencilla función solo tiene dos argumentos: *test* y *size*.

1. **test**: permite definir el tipo de tamaño del efecto que necesito. Las opciones incluyen “t”, para pruebas-*t*, “r” para correlaciones, “anov” para ANOVAs de una vía (usando *f* de Cohen como medida del tamaño del efecto), y “f2” para regresiones múltiples (en referencia al  $f^2$ ).
2. **size**: permite definir el valor de referencia que quiero obtener. Las opciones son “small”, “medium”, y “large”, para obtener el valor de referencia para el tamaño del efecto deseado pequeño, medio o grande, respectivamente.

Por ejemplo, si quisiera saber cuál es el valor de referencia para una correlación con un efecto pequeño, usaría:

```
cohen.ES(test = "r", size = "small")
```

Esto produce:

```
##
##      Conventional effect size from Cohen (1982)
##
##      test = r
##      size = small
##      effect.size = 0.1
```

Donde, como se ve, me dice que el valor de referencia es **0.1** (en el campo llamado “`effect.size`”).

## 6.3 Análisis de poder para correlaciones. Función `pwr.r.test`

Si, por ejemplo, quiero hacer un estudio, en el que espero determinar si existe una correlación entre dos variables, para saber el tamaño de la muestra (*n*) adecuado, necesito entonces determinar:

1. El tamaño del efecto (*r*) esperado (es decir, qué tan fuerte es la asociación que espero entre las variables).
2. El alfa ( $\alpha$ ) o nivel de significación estadística (típicamente 0.05.).
3. El poder estadístico deseado (ver Sección 1.1 *¿Qué es Potencia o Poder Estadístico?*).

Por ejemplo, si quiero tener 80% de probabilidades de detectar como significativa una correlación con un tamaño del efecto de  $r = 0.3$ , con un  $\alpha = 0.05$ , puedo usar la siguiente función (en este caso, “*guardé*”<sup>14</sup> el resultado de la función en un objeto que llamé `pcorr`).

```
pcorr <- pwr.r.test(r = 0.3,
  sig.level = 0.05,
  power = 0.9) #0.8 es típico, pero algunas disciplinas usan 0.9, o incluso 0.99
```

Para ver el resultado, solo necesito *correr* en R el nombre de mi objeto.

```
pcorr
```

Lo que produce:

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##      n = 111.8068
```

<sup>13</sup>Como discutí extensamente (ver sección 1.3.1 *Por qué no es buena idea usar las definiciones de Cohen*), Cohen (1988, 1992) propuso definiciones de referencia para los tamaños de efecto (efectos “pequeños”, “medios”, y “grandes”), pero hay múltiples problemas con su uso indiscriminado.

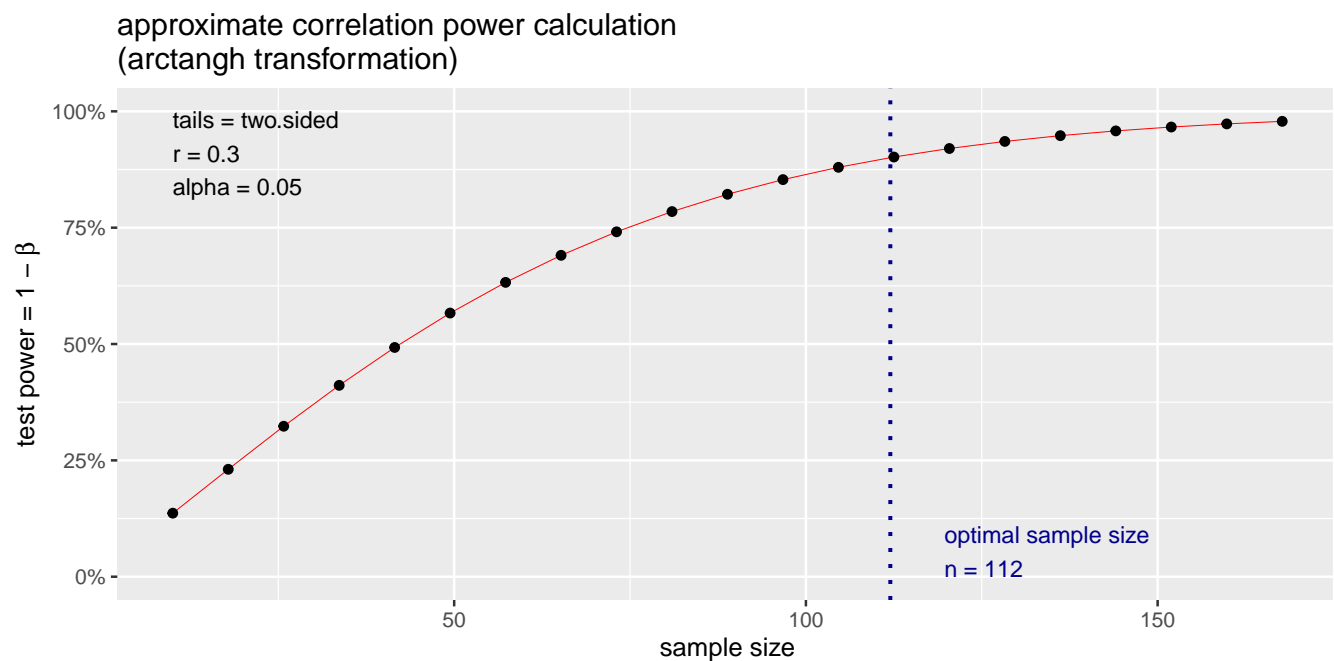
<sup>14</sup>En realidad, como cualquier usuario habitual de R sabe, lo que he hecho no es realmente “guardar”, sino *asignar* el resultado de la función a un objeto. Uso la palabra “guardar” para simplificar la explicación para personas que no estén familiarizadas con R, u otros lenguajes de programación, ya que los objetos suelen estar ocultos en la mayoría de los programas comerciales, y no requieren que el usuario los manipule directamente.

```
##           r = 0.3
##       sig.level = 0.05
##           power = 0.9
##   alternative = two.sided
```

Como se ve, en el campo llamado “n”, me dice que la muestra necesaria para obtener el poder deseado de 90% para detectar como significativa una correlación con un  $r$  de 0.3, es de 111.8068. Por supuesto, yo no puedo tener un número no entero de observaciones o participantes, por lo que el valor  $n$  (tamaño de muestra), debe aproximarse al siguiente número entero superior (en este caso, 112).

Finalmente, puedo ver la asociación entre el tamaño de la muestra y el poder estadístico gráficamente con la función `plot`<sup>15</sup>, introduciendo como argumento mi objeto (en este caso `pccorr`) puedo ver una figura de este análisis (Figura 3).

```
plot(pccorr)
```



**Figura 3.** Asociación entre el tamaño de la muestra y el poder estadístico para una correlación de 0.3, con un  $\alpha$  de 0.05, producida con la función `plot` del paquete `pwr`. Como se puede ver, sugiere un  $n$  de 112 (línea azul), para alcanzar un poder de 0.9 (90%).

### 6.3.1 Si hay hipótesis claras, pre-especificadas (*a priori*), se pueden hacer análisis de una cola

Cuando tengo una hipótesis precisa, como que la correlación que estudio será positiva, puedo hacer análisis de una cola, lo que reduce substancialmente el tamaño de muestra requerido. Sin embargo, esto solo se debe hacer cuando tengo una hipótesis clara, con sólido fundamento teórico, o empírico (por ejemplo, en el caso de una replicación).

Para hacerlo, el argumento `alternative` debe ser definido:

1. **Para correlaciones positivas:** `alternative = "greater"`.
2. **Para correlaciones negativas:** `alternative = "less"` (en cuyo caso el argumento `r` debe ser negativo; por ejemplo -0.3).

Por ejemplo, en este caso, al especificar que espero que el tamaño del efecto sea positivo, y **al menos** de  $r = 0.3$ , el  $n$  se reduce de 112 a 92 observaciones o participantes.

<sup>15</sup>Las figuras que produce la función `plot` para análisis hechos en el paquete `pwr`, son objetos de clase `ggplot`, por lo que alguien familiarizado con el paquete `ggplot2` puede modificar la figura para que, por ejemplo, los ejes, título y anotaciones estén en español, o para cambiar el tema y colores.

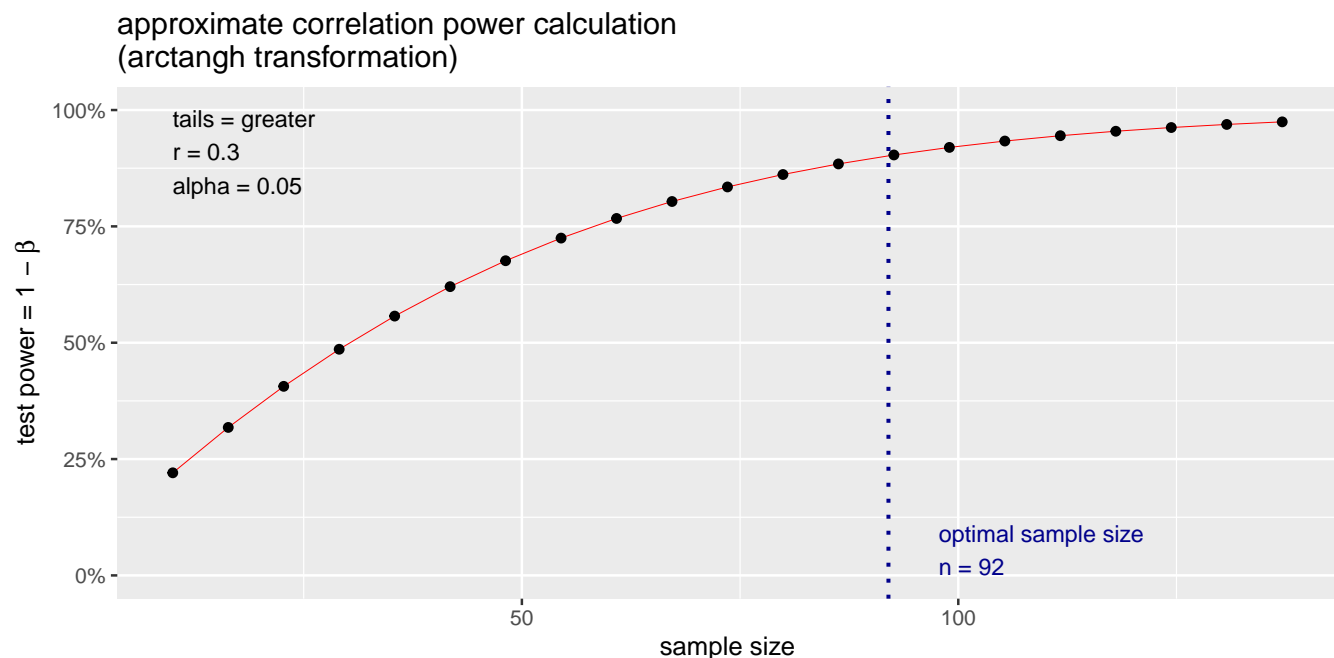
```
pwr.greater <- pwr.r.test(r = 0.3,
  sig.level = 0.05,
  power = 0.9,
  alternative = "greater") #Dependiendo de si es positivo o negativo
pwr.greater
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##      n = 91.41024
##      r = 0.3
##      sig.level = 0.05
##      power = 0.9
##      alternative = greater
```

❗ Es importante tener en cuenta que este análisis de poder de una cola únicamente tendrá sentido si mi correlación es, de hecho, positiva.

Al igual que antes, con la función `plot`, puedo ver una figura de este análisis (Figura 4).

```
plot(pwr.greater)
```



**Figura 4.** Asociación entre el tamaño de la muestra y el poder estadístico para una correlación **positiva** (de una cola), de **al menos** 0.3, con un  $\alpha$  de 0.05, producida con la función `plot` del paquete `pwr`. Como se puede ver, sugiere un  $n$  de 67 (línea azul), para alcanzar un poder de 0.9 (90%).

## 6.4 Análisis de poder para pruebas-t. Función `pwr.t.test`

Existen tres tipos de pruebas-t: (1) pruebas-t de muestras independientes (o no relacionadas), que se usan por ejemplo cuando comparo dos grupos de personas; (2) pruebas-t de medidas repetidas, relacionadas o pareadas, que se usan cuando comparo, por ejemplo, al mismo grupo de personas en dos momentos (e.g. antes y después de un tratamiento), o en dos condiciones (e.g. tomando una medicina vs tomando un placebo); y (3) pruebas-t de una muestra, que se usan cuando comparo una serie de datos (por ejemplo la estatura de un grupo particular de personas) con un valor ya conocido (e.g. la media de estatura establecida para país), para saber si los miembros de mi grupo tienden a ser más bajos, más altos, o de altura similar a la población general.

La función `pwr.t.test` asume por defecto que se trata de un análisis de medidas independientes (`type = "two.sample"`), a menos que se defina qué tipo de prueba-*t* responde a mi diseño, usando el argumento `type`:

1. Para pruebas-*t* de muestras independientes: `type = "two.sample"`.
2. Para pruebas-*t* de medidas repetidas: `type = "paired"`.
3. Para pruebas-*t* de una muestra: `type = "one.sample"`.

#### 6.4.1 Muestras independientes

En este caso, “guardé” el resultado de la función en un objeto que llamé `ptInd`.

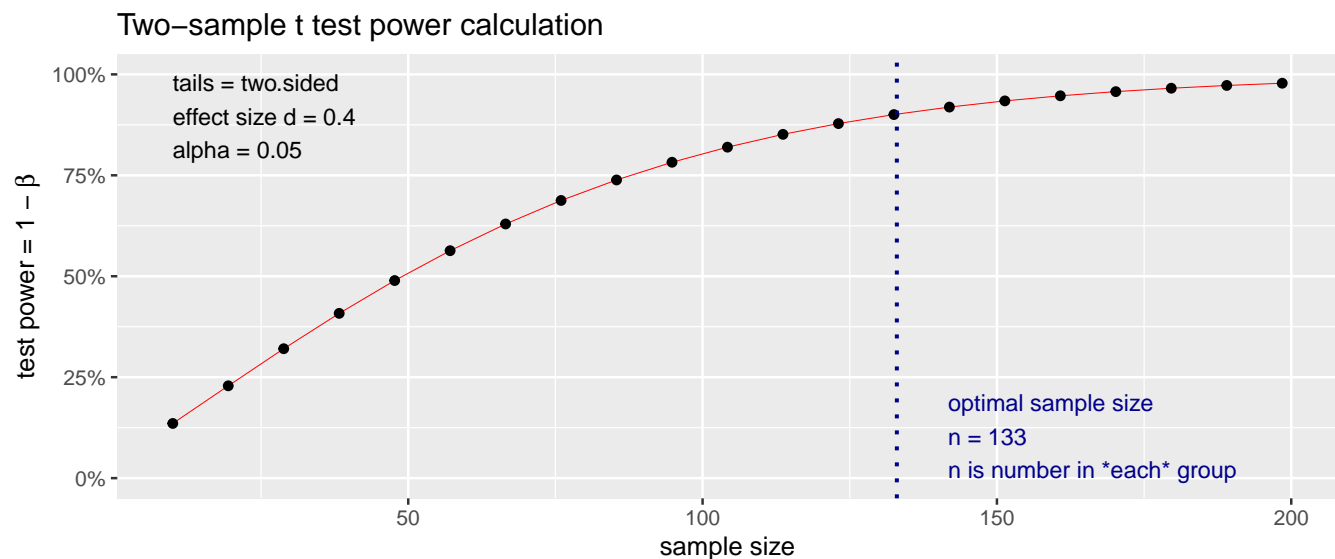
```
ptInd <- pwr.t.test(d = 0.4,
  sig.level = 0.05,
  power = 0.9,
  type = "two.sample")
ptInd #los resultados son MUY claros en cuanto a que el n es para cada grupo.
```

Lo que produce:

```
##
##      Two-sample t test power calculation
##
##              n = 132.3105
##              d = 0.4
##      sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

Al igual que con las correlaciones, con la función `plot`, puedo ver una figura de este análisis (Figura 5).

```
plot(ptInd)
```



**Figura 5.** Asociación entre el tamaño de la muestra y el poder estadístico para una prueba-*t* de 0.4, con un  $\alpha$  de 0.05, producida con la función `plot` del paquete `pwr`. Como se puede ver, sugiere un *n* de 133 por grupo (línea azul), para alcanzar un poder de 0.9 (90%).

#### 6.4.2 Una muestra

Para este ejemplo, “guardé” el resultado de la función en un objeto que llamé `ptUna`.

```
ptUna <- pwr.t.test(d = 0.4, sig.level = 0.05,
                   power = 0.9,
                   type = "one.sample")
ptUna
```

```
##
##      One-sample t test power calculation
##
##              n = 67.62138
##              d = 0.4
##      sig.level = 0.05
##      power     = 0.9
##      alternative = two.sided
```

### 6.4.3 Medidas repetidas o pareadas

Es importante tener en cuenta que, de ser posible, suele ser mejor hacer diseños de medidas repetidas, con respecto a medidas independientes, pues en general esto representa un mejor uso de recursos: como son análisis más poderosos estadísticamente, tienen un  $n$  más pequeño comparados con análisis equivalentes de medidas independientes.

Sin embargo, por supuesto no todas las preguntas pueden ser respondidas con un análisis de medidas repetidas. Si por ejemplo quiero mirar diferencias entre perros y gatos en la cantidad de atención prestada a los humanos, no puedo tener a mis participantes una vez como perros, y otra como gatos. En este ejemplo, no existe la opción de hacer medidas repetidas, como tampoco existiría si quisiera comparar la capacidad pulmonar (o cualquier otra variable) de personas nacidas en Uruguay y personas nacidas en Panamá, pues por supuesto las personas solo nacen una vez, y en un solo país.

En este caso, por ejemplo, mientras un análisis de medidas independientes como el presentado antes (sección 5.4.1 **Muestras independientes**) requería un  $n$  de 133 personas por grupo (es decir, 266 en total), el mismo análisis pero con un diseño de medidas repetidas requiere un  $n$  de 68 personas (con dos observaciones por persona).

```
ptRep <- pwr.t.test(d = 0.4,
                   sig.level = 0.05,
                   power = 0.9,
                   type = "paired")
ptRep #los resultados son MUY claros en cuanto a que el n es para pares de observaciones.
```

```
##
##      Paired t test power calculation
##
##              n = 67.62138
##              d = 0.4
##      sig.level = 0.05
##      power     = 0.9
##      alternative = two.sided
##
## NOTE: n is number of *pairs*
```

### 6.4.4 Análisis de una cola

De nuevo, cuando tenemos una hipótesis precisa, como que la diferencia entre grupos estudiados será en una dirección particular (e.g. Grupo 2 > Grupo 1), se pueden hacer análisis de una cola, tanto para pruebas- $t$  de medidas independientes, relacionadas, o de una muestra, lo que reduce substancialmente el tamaño de muestra requerido.

Para hacerlo, el argumento `alternative` debe ser definido:

1. **Para diferencias positivas (donde Gr 2 > Gr 1):** `alternative = "greater"`.
2. **Para correlaciones negativas (donde Gr 2 < Gr 1):** `alternative = "less"` (en cuyo caso el argumento `d` también debe ser negativo; por ejemplo -0.3).

Por ejemplo, en este caso, al especificar que espero que el tamaño del efecto de medidas repetidas (sección 5.4.3 **Muestras repetidas o pareadas**) sea positivo, y con un  $d$  de Cohen de **al menos** 0.4, el  $n$  se reduce de 68 a 55 pares de observaciones (es decir, 55 participantes, cada uno medido 2 veces).

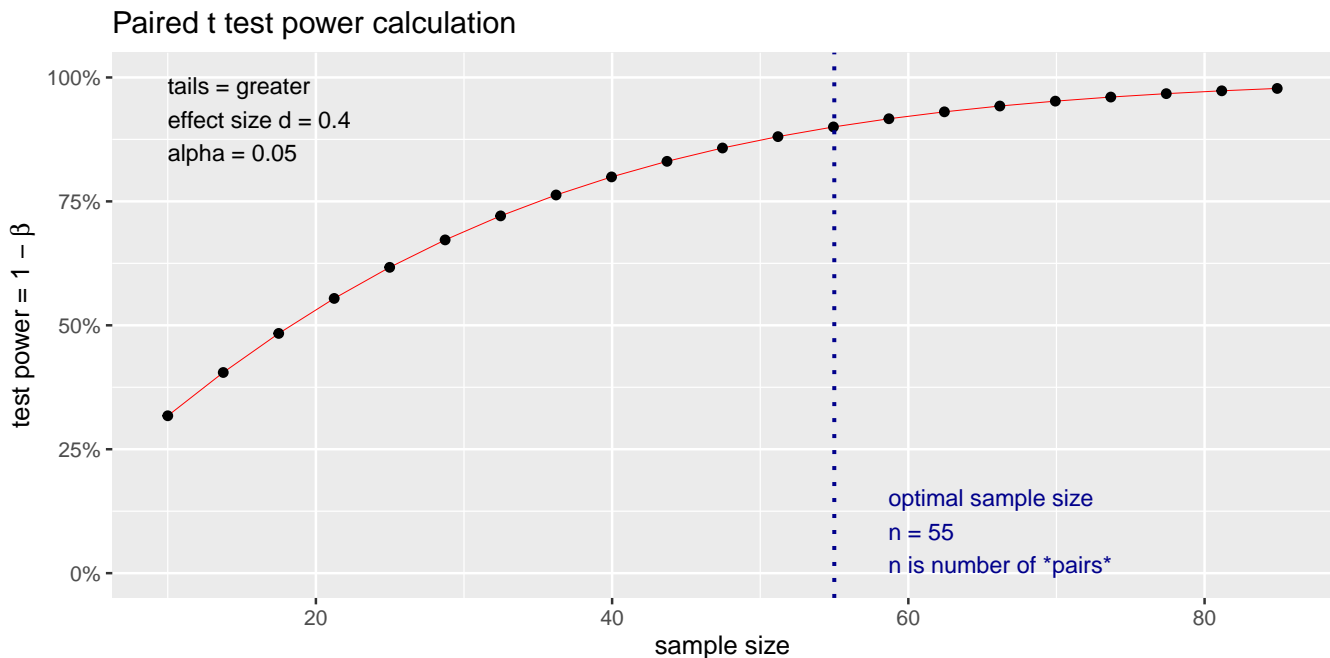
```
ptRepGreat <- pwr.t.test(d = 0.4,
  sig.level = 0.05,
  power = 0.9,
  type = "paired",
  alternative = "greater")
ptRepGreat
```

```
##
##      Paired t test power calculation
##
##              n = 54.90553
##              d = 0.4
##      sig.level = 0.05
##      power = 0.9
##      alternative = greater
##
## NOTE: n is number of *pairs*
```

❗ Es importante tener en cuenta que este análisis de una cola únicamente tendrá sentido si mi correlación es, de hecho, positiva.

Al igual que antes, con la función `plot`, puedo ver una figura de este análisis (**Figura 6**).

```
plot(ptRepGreat)
```



**Figura 6.** Asociación entre el tamaño de la muestra y el poder estadístico para una prueba- $t$  de **al menos** 0.4 (una cola), con un  $\alpha$  de 0.05, producida con la función `plot` del paquete `pwr`. Como se puede ver, sugiere un  $n$  de 55 (línea azul), para alcanzar un poder de 0.9 (90%).

## 6.5 Análisis de poder para ANOVAs de una vía. Función `pwr.anova.test`

Por lo menos hasta ahora<sup>16</sup>, `pwr` solo permite hacer análisis de poder para ANOVAs de una vía, balanceados (mismo número de participantes u observaciones por grupo), y de medidas independientes (es decir, comparando un número  $k$  de grupos)<sup>17</sup>.

Los argumentos de esta función son similares a los de las funciones `pwr.r.test` y `pwr.t.test`:

1. `k`: número de grupos a comparar en mi ANOVA de una vía.
2. `f`: tamaño del efecto ( $f$  de Cohen) esperado (es decir, qué tan fuerte es la asociación que espero entre la variable independiente, que divide mis grupos, y la variable dependiente).
3. `sig.level`: alfa ( $\alpha$ ) o nivel de significación estadística (típicamente 0.05.).
4. `power`: poder estadístico deseado (ver sección 1.1 *¿Qué es Potencia o Poder Estadístico?*).

Los argumentos `sig.level` y `power` son iguales, y se usan de manera idéntica en esta función que en las funciones `pwr.r.test` y `pwr.t.test`.

El tamaño del efecto también se usa de manera similar, pero con la diferencia de que para la función `pwr.anova.test` se define un tamaño  $f$ , mientras que para `pwr.r.test` y `pwr.t.test` se usaban tamaños del efecto  $r$  y  $t$ , respectivamente.

Sin embargo, la función `pwr.anova.test` tiene algunos cambios:

1. `k` es un argumento nuevo, que se usa para definir el número de grupos a comparar. Este argumento no se usaba en la función `pwr.t.test` para pruebas- $t$  de este paquete.
2. `alternative` y `type`, argumentos usados en la función `pwr.t.test` para pruebas- $t$  de este paquete, no se usan para esta función.

Entonces, teniendo esto en cuenta, si por ejemplo quiero alcanzar un poder  $1 - \beta$  de 0.9 (90%), con un  $\alpha$  de 0.05, para detectar un  $f$  de 0.25, al comparar 4 grupos, es necesario un  $n$  de 58 participantes en cada grupo (o 232 participantes en total, dado que el diseño propuesto tendría 4 grupos).

```
panova <- pwr.anova.test(k = 4,
                        f = 0.25,
                        sig.level = 0.05,
                        power = 0.9)
panova
```

Lo que produce:

```
##
##      Balanced one-way analysis of variance power calculation
##
##              k = 4
##              n = 57.67309
##              f = 0.25
##      sig.level = 0.05
##      power    = 0.9
##
## NOTE: n is number in each group
```

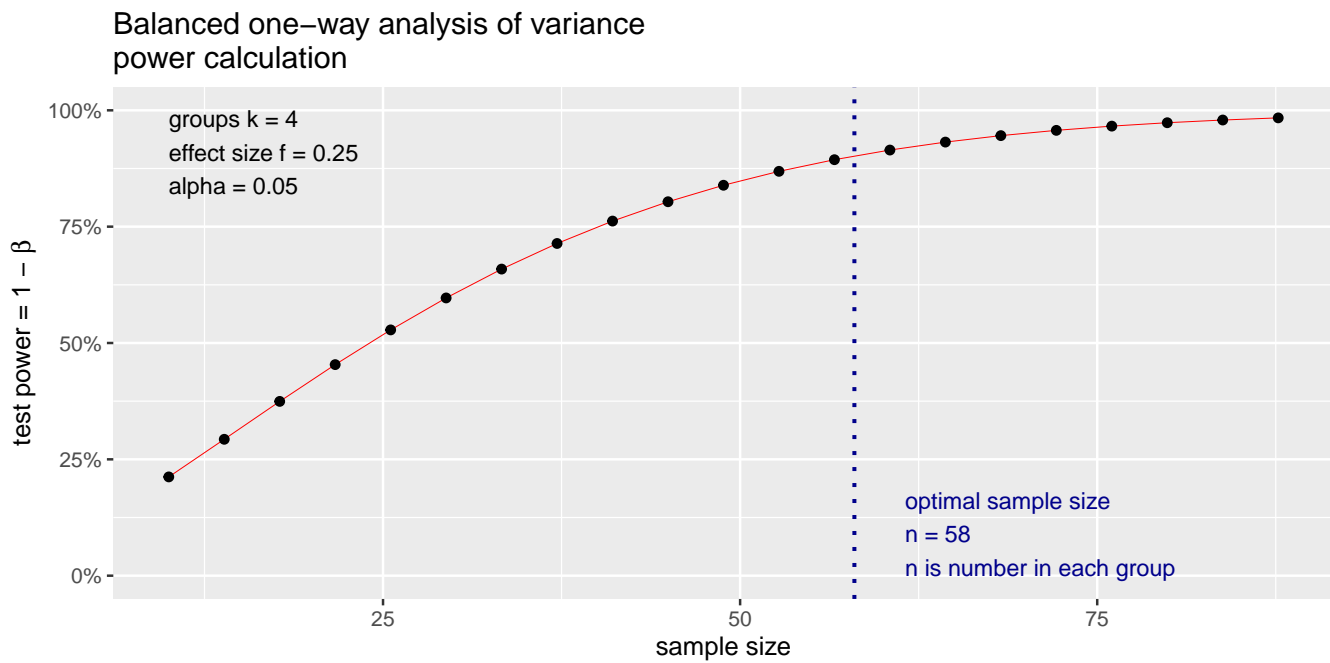
Al igual que con las demás funciones para calcular el poder estadístico del paquete `pwr`, con la función `plot`, puedo ver una figura de este análisis.

```
plot(panova)
```

<sup>16</sup>La versión actual de `pwr` es la 1.3-0.

<sup>17</sup>Para diseños más complejos, incluyendo diseños factoriales de medidas independientes, repetidas, o mixtas (tanto factores de medidas independientes como repetidas), la mejor opción actualmente es el paquete `Superpower`, descrito en la siguiente sección.





**Figura 7.** Asociación entre el tamaño de la muestra y el poder estadístico para un ANOVA de 0.4, con un  $\alpha$  de 0.05, producida con la función `plot` del paquete `pwr`. Como se puede ver, sugiere un  $n$  de 58 participantes por grupo (línea azul), para alcanzar un poder de 0.9 (90%).

## 7 Paquete *Superpower* para R (diseños factoriales complejos)

El paquete *Superpower* (Caldwell et al., 2020; Caldwell & Lakens, 2020; Lakens & Caldwell, 2020), está específicamente diseñado para permitir hacer análisis de poder para diseños de una vía o factoriales (con más de un factor o variable independiente nominal), incluyendo diseños complejos, de manera empírica. En esta guía, me centraré en los diseños factoriales.

### 7.1 Diseños factoriales

En los diseños factoriales, normalmente hay más de un efecto de interés. Por ejemplo, si quiero saber el efecto del género (hombre, mujer) y el máximo nivel educativo (pregrado, postgrado), en el salario de los médicos, podría hacer un diseño  $2 \times 2$ . Esto quiere decir que tengo dos factores (variables independientes nominales): el género y el nivel educativo, cada uno con dos niveles (1. hombre y 2. mujer, para el caso del género; y 1. pregrado y 2. postgrado, para el caso del nivel educativo).

En este caso, yo obtendría 3 valores  $p$ :

1. Efecto principal del género, que me diría si hay diferencias en el salario entre hombres y mujeres.
2. Efecto principal del nivel educativo, que me dice si hay diferencias entre el salario de médicos con pregrado y médicos con postgrado como máximo nivel educativo alcanzado.
3. Interacción entre género y nivel educativo, que me dice si el salario depende simultáneamente del género y el nivel educativo. Por ejemplo, si entre las mujeres ganan más dinero las personas con postgrado, pero entre los hombres ganan más las personas con pregrado<sup>18</sup>.

Por supuesto, yo podría tener diseños más complejos, como un  $2 \times 4$ , donde tengo, de nuevo, 2 factores, pero uno que tiene 2 niveles, y otro que tiene 4. O un diseño  $2 \times 2 \times 3$ , donde hay 3 factores, con 2, 2, y 3 niveles, respectivamente.

Es importante tener en cuenta que, al aumentar el número de factores, independientemente del número de niveles

<sup>18</sup>Obviamente este ejemplo muy seguramente no representa la realidad. Es solamente un ejemplo hipotético.

que tengan esos factores, el número de resultados (y valores  $p$  asociados) aumenta. Por ejemplo, mientras en el caso de un diseño con dos factores hay 3 resultados, en un diseño con tres factores (voy a llamarlos **A**, **B** y **C**), hay 7:

1. Efecto principal del primer factor (**A**).
2. Efecto principal del segundo factor (**B**).
3. Efecto principal del tercer factor (**C**).
4. Interacción entre el primer y segundo factor (**A**  $\times$  **B**).
5. Interacción entre el primer y tercer factor (**A**  $\times$  **C**).
6. Interacción entre el segundo y tercer factor (**B**  $\times$  **C**).
7. Interacción entre los tres factores (**A**  $\times$  **B**  $\times$  **C**).

Dada la complejidad de estos cálculos, y la multiplicidad de efectos (principales e interacciones) y resultados asociados, y dado que para detectar cada uno de esos efectos hay un poder estadístico distinto, la mayoría de los programas para análisis de poder solamente permiten analizar un efecto a la vez. Por ejemplo, G\*Power, aunque permite hacer análisis de poder para diseños complejos, solo calcula un efecto a la vez.

En contraste, el paquete *Superpower* permite hacer análisis, no solo para diseños más complejos, sino calculando simultáneamente el poder para cada efecto, y para posibles comparaciones *post-hoc*.

Este proceso es sumamente complejo, y no se puede solucionar de manera matemática *sencilla*<sup>19</sup>, por lo que *Superpower* usa una estrategia interesante: simula una base de datos para el diseño propuesto (dadas una serie de características de cada variable y su relación con las demás<sup>20</sup>), y empíricamente estima el poder, a partir de hacer muchas iteraciones (repeticiones aleatorias) de esta simulación.

Dadas estas complejidades, en ocasiones no es posible tener la información suficiente para hacer las simulaciones, pues se requiere de estudios previos (o pilotos) muy completos, con diseños idénticos, y que, o bien hayan reportado toda esta información, o hayan abierto libremente sus bases de datos para poder hacer estos cálculos<sup>21</sup>.

Por esto, la forma de usar *Superpower* y sus funciones, es muy distinta a la de *pwr*. En las siguientes secciones mostraré ejemplos de análisis de poder para diseños factoriales de medidas independientes, repetidas y mixtas, pero por simplicidad siempre usaré diseños  $2 \times 2$ . Para diseños más complejos la lógica es, en todo caso, la misma.

## 7.2 Instalación y carga de *Superpower*

Para instalar y cargar *Superpower*, se requiere correr las siguientes funciones:

```
install.packages("Superpower") #no es necesario si ya ha sido instalado.
library(Superpower) #para cargar el paquete una vez instalado.
```

## 7.3 Acerca de comparaciones *post-hoc*

Comúnmente, al hacer un ANOVA, bien sea de una vía o factorial, se deben hacer además comparaciones *post-hoc* [o, alternativamente, *contrastes planeados*; e.g. Chatham (1999)] para determinar entre qué grupos o condiciones están las diferencias. Por ejemplo, al hacer un ANOVA de una vía comparando cuatro grupos, si el resultado del ANOVA es significativo, es probable que quiera determinar si hay diferencias entre los grupos 1 y 2, 1 y 4, o 3 y 4, por ejemplo.

Es importante tener en cuenta que el número de comparaciones *post-hoc* posibles para un ANOVA aumenta rápidamente cuando tengo más factores, o los factores tienen más niveles (independientemente de que estos factores sean de medidas repetidas o independientes).

De hecho, el número de posibles comparaciones (que he denominado  $n_{comp}$ ) es el producto de los niveles de todos los factores al cuadrado menos el producto de esos mismos niveles, dividido por dos. Entonces:

Para un diseño  $2 \times 2$  (o un diseño con un solo factor con 4 niveles) hay 6 comparaciones posibles:

<sup>19</sup>No existe una única ecuación que permita hacer este cálculo.

<sup>20</sup>En particular, se necesita saber la media y desviación estándar de cada grupo y/o condición y, cuando hay factores de medidas repetidas, una matriz de correlaciones entre cada combinación de niveles.

<sup>21</sup>Esta es una de las razones por las cuales es muy importante que en cada artículo haya una sección de descriptivos muy buena y completa, e idealmente que los datos estén disponibles para poder re-analizarlos, o hacer este tipo de análisis descriptivos necesarios para hacer un análisis de poder

$$n_{comp} = \frac{(2 \times 2)^2 - (2 \times 2)}{2} = \frac{16 - 4}{2} = 6$$

Para un diseño  $3 \times 2$  (o un diseño con un solo factor con 6 niveles) hay 15 comparaciones posibles:

$$n_{comp} = \frac{(3 \times 2)^2 - (3 \times 2)}{2} = \frac{36 - 6}{2} = 15$$

Para un diseño  $2 \times 2 \times 2$ , hay 28 comparaciones posibles:

$$n_{comp} = \frac{(2 \times 2 \times 2)^2 - (2 \times 2 \times 2)}{2} = \frac{64 - 8}{2} = 28$$

Para un diseño  $3 \times 2 \times 2$ , hay 66 comparaciones posibles:

$$n_{comp} = \frac{(3 \times 2 \times 2)^2 - (3 \times 2 \times 2)}{2} = \frac{144 - 12}{2} = 66$$

Y para un diseño  $2 \times 2 \times 4$  (que tendría el mismo número de comparaciones que un diseño  $2 \times 2 \times 2 \times 2$ ), hay 120 comparaciones posibles:

$$n_{comp} = \frac{(2 \times 2 \times 4)^2 - (2 \times 2 \times 4)}{2} = \frac{256 - 16}{2} = 120$$

Esto es muestra de cómo la complejidad de los diseños, especialmente factoriales, aumenta exponencialmente al tener más factores, o más niveles por factor.

### 7.3.1 Cómo controlar la tasa de errores al hacer pruebas *post-hoc*. Correcciones de *Bonferroni* y *Holm-Bonferroni*

Hacer pruebas *post-hoc* o cualquier tipo de comparaciones múltiples sobre la misma base de datos, aunque importante, infla la tasa de errores Tipo I (falsos positivos), por lo que generalmente se hacen correcciones al  $\alpha$  (nivel de significación), para contrarrestar esta mayor posibilidad de encontrar diferencias que, en realidad, no existan.

En otras palabras, dado que si tengo un  $\alpha = 0.05$ , estoy aceptando una probabilidad de que el 5% (o 1 de cada 20) resultados sea falso, si hago dos análisis, la probabilidad de un falso positivo de dobla (10%), si hago 3 se triplica (15%), etcétera. Si hago 20 análisis, estoy probabilísticamente asegurando que obtendré un falso resultado positivo. Y si hago un ANOVA  $2 \times 2 \times 4$ , con sus 120 comparaciones *post-hoc*, probabilísticamente obtendría seis falsos positivos.

Para contrarrestar este problema, existen varias opciones, de las cuales probablemente la más conocida es la *corrección de Bonferroni* (Bonferroni, 1936), que consiste en reducir el  $\alpha$  (típicamente de 0.05), dividiéndolo por el número de comparaciones múltiples (o pruebas *post-hoc*) que se hagan.

Entonces, si por ejemplo hago dos comparaciones *post-hoc*,  $\alpha = \frac{0.05}{2} = 0.025$ , y si hago seis,  $\alpha = \frac{0.05}{6} = 0.0083$ . Esto, por supuesto, hace que sea más difícil encontrar resultados significativos (en efecto, reduciendo el poder estadístico, que es la probabilidad de detectar como significativo un efecto que sí existe). Por esto, aunque la *corrección de Bonferroni* controla muy bien la tasa de errores Tipo I (falsos positivos), infla la tasa de errores Tipo II (falsos negativos), dejando como única alternativa incrementar el tamaño de la muestra (y por consiguiente, el poder estadístico).

Sin embargo, existen alternativas más modernas y versátiles (para una revisión y comparación, ver Blakesley et al., 2009). De estas, una relativamente sencilla y popular, disponible en muchos paquetes estadísticos, es la *corrección de Holm-Bonferroni* (Holm, 1979). Esta alternativa, que personalmente me gusta mucho, es una suerte de *corrección de Bonferroni* pero aplicada secuencialmente.

Es decir, si por ejemplo hago seis comparaciones *post-hoc*,  $\alpha = \frac{0.05}{6} = 0.0083$  se aplicará para el efecto con el valor  $p$  más pequeño,  $\alpha = \frac{0.05}{5} = 0.01$  al segundo más pequeño,  $\alpha = \frac{0.05}{4} = 0.0125$  al tercero más pequeño,

$\alpha = \frac{0.05}{3} = 0.0167$  al cuarto más pequeño,  $\alpha = \frac{0.05}{2} = 0.025$  al quinto más pequeño, y  $\alpha = \frac{0.05}{1} = 0.05$  al sexto más pequeño (que sería el valor  $p$  más grande).

Al hacer esta corrección secuencial, a *corrección de Holm-Bonferroni* tiene la ventaja de limitar la inflación de la tasa de errores Tipo II (falsos negativos), en comparación a la *corrección de Bonferroni* (ver e.g. [Streiner, 2015](#)), sin dejar de controlar la tasa de errores Tipo I (falsos positivos) .

Como explicaré en las siguientes secciones, el paquete **Superpower** permite calcular en un solo análisis el poder estadístico, tanto de los efectos principales e interacciones, como también para las comparaciones *post-hoc*, implementando al tiempo correcciones de *Bonferroni*, *Holm-Bonferroni*, u otras opciones disponibles.

## 7.4 ANOVA factorial de medidas independientes

En todos los casos, los análisis de poder requieren **al menos** dos funciones:

1. `ANOVA_design` que permite especificar las características del diseño para el cual haré el análisis de poder.
2. Una función para obtener el análisis de poder con base en el diseño. Para esto, hay varias opciones, incluyendo<sup>22</sup>:
  - `ANOVA_power` que usa simulaciones para determinar el poder estadístico obtenido.
  - `plot_power` que muestra gráficamente el poder obtenido según el tamaño de la muestra (similar a la función `plot` de `pwr`), pero para todos los efectos principales e interacciones.

### 7.4.1 Ejemplo ANOVA factorial de medidas independientes ( $2 \times 2$ )

Entonces, si por ejemplo quiero estudiar los resultados de un examen de matemáticas entre estudiantes becados y no becados, de universidades públicas y privadas, tengo un diseño  $2$  (Beca: 1. Sí, 2. No)  $\times 2$  (Universidad: 1. Pública, 2. Privada).

Para hacer este análisis, debo primero definir el diseño y las características del mismo con la función `ANOVA_design`.

Para esto, requiero los siguientes argumentos:

1. **design**: este argumento esencial, define el tipo de diseño. En este caso, dado que tengo un diseño  $2 \times 2$ , donde ambos factores son de medidas independientes, debo ponerlo como "`2b*2b`", donde los números representan el número de niveles en cada factor, la letra **b** que ese factor es de medidas independientes (o *entre sujetos*, por lo cual usa la letra **b**, del inglés *between-subjects*).
2. **n**: el número de participantes que espero tener por condición (o, dicho de otro modo, por combinación de niveles entre mis factores; e.g. (1) estudiantes becados de universidades privadas; (2) estudiantes becados de universidades públicas, etcétera). Aunque, a diferencia de otros paquetes y programas, **Superpower** no calcula el  $n$  por mí, me permite cambiar el  $n$  hasta lograr el poder deseado.
3. **mu**: las medias para cada interacción entre los niveles de mis factores. En este caso, (1) la media de estudiantes becados de universidades públicas, (2) estudiantes becados de universidades privadas, (3) estudiantes no becados de universidades públicas, y (4) estudiantes becados de universidades privadas. Estos valores, como se verá en el código, deben estar *concatenados* usando la función `c`.
4. **sd**: la desviación estándar para la población (por lo cual es un solo valor. En este caso, la desviación estándar de las calificaciones del examen).
5. **labelnames** (Opcional): las etiquetas (nombres) de los factores y sus niveles. Al igual que con las medias, estas etiquetas<sup>23</sup> deben estar *concatenadas* usando la función `c`. Para definirlos, se deben poner en el siguiente orden:
  - Etiqueta del primer factor (en este caso "Beca")
  - Etiquetas de los niveles de ese factor (en este caso "Sí" y "No")
  - Etiqueta del segundo factor (en este caso "Universidad")
  - Etiquetas de los niveles de ese factor (en este caso "Pública" y "Privada")

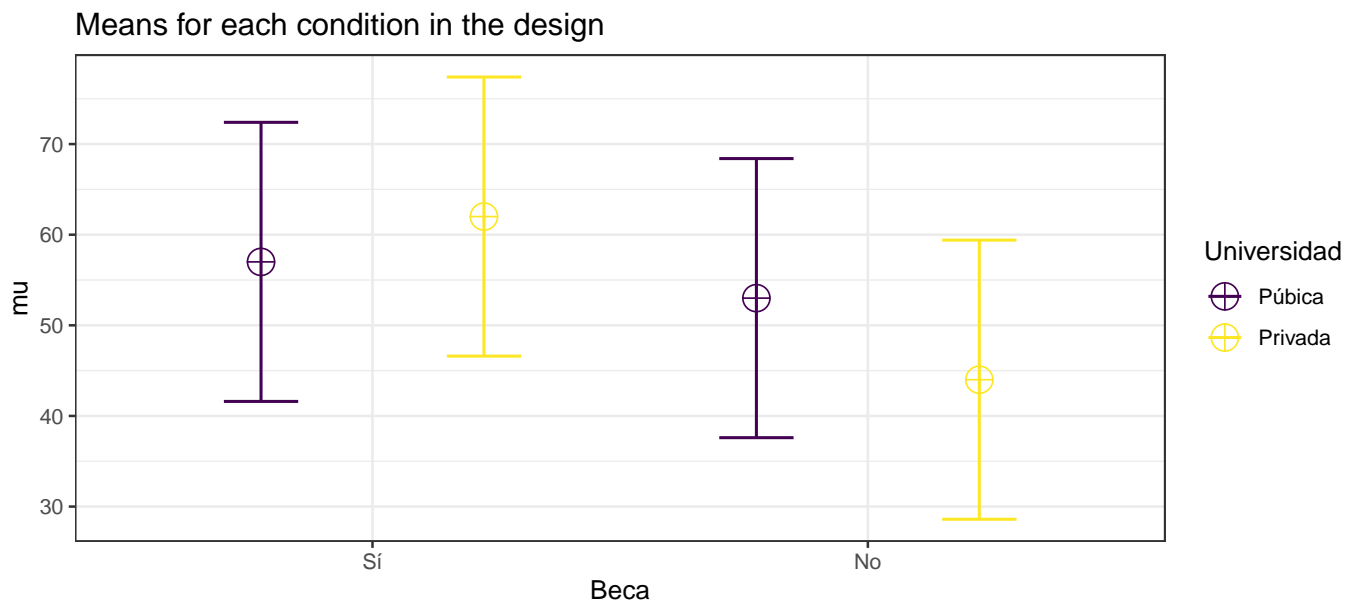
<sup>22</sup>El paquete **Superpower** tiene otras funciones muy útiles para análisis de poder estadístico, que por simplicidad no cubriré en este documento. Para conocerlas, recomiendo ver la [documentación del paquete](#), o la [introducción](#) hecha por los autores al mismo ([Lakens & Caldwell, 2020](#)).

<sup>23</sup>Estos nombres o etiquetas no deben contener espacios.

6. `plot` (Opcional): este es un argumento lógico que únicamente acepta como opciones `TRUE` y `FALSE`. Si incluyo `plot = TRUE` esta función creará además una figura (*plot*) mostrando las medias y sus intervalos de confianza (si no incluyo este argumento, por defecto la función asumirá la opción `FALSE` y no producirá esta figura.)

Entonces, el código para este diseño hipotético sería el siguiente, con 80 participantes en cada combinación de *Beca* y *Universidad* (320 en total), si los puntajes promedio ( $\mu$ ) del examen fuesen 57 (becados de universidad pública), 62 (becados de universidad privada), 53 (no becados de universidad pública), y 44 (no becados de universidad privada), con una desviación estándar ( $sd$ ) de 15.4<sup>24</sup>.

```
diseñoB <- ANOVA_design(design = "2b*2b",
                        n = 80, sd = 15.4,
                        mu = c(57, 62, 53, 44),
                        labelnames = c("Beca", "Sí", "No", "Universidad", "Pública", "Privada"),
                        plot = TRUE)
```



**Figura 8.** Ejemplo de la distribución de medias y sus intervalos de confianza para el diseño definido con la función `ANOVA_design`, al incluir el argumento `plot = TRUE`. Esto es muy útil para estar seguro de que las medias fueron *concatenadas* en el orden correcto.

Definido el diseño y las características de los datos (“guardando” este diseño en un objeto que llamé `diseñoB`), puedo ver el poder que obtendría con esos 80 participantes por cada combinación de *Beca* y *Universidad*, con la función `ANOVA_power`.

Esta función, además de requerir como argumento el diseño que acabo de definir (en este caso `diseñoB`), requiere que defina:

1. `alpha_level`:  $\alpha$  (nivel de significación) deseado. Típicamente 0.05.
2. `p.adjust`: si se debe hacer un ajuste para comparaciones *post-hoc* (por ejemplo correcciones de *Bonferroni* con la opción `"bonferroni"`, *Holm-Bonferroni* con la opción `"holm"`, o sin corrección, usando la opción `"none"`; para ver todas las opciones, recomiendo ver la documentación de la función `p.adjust`). En este caso definí que quiero hacer una corrección de “holm”, que se refiere a la corrección de *Holm-Bonferroni* (Holm, 1979).
3. `nsim`: número de simulaciones hechas para determinar el poder; acá es importante tener en cuenta que un número mayor de simulaciones me dará resultados más robustos y confiables, pero requerirá más tiempo. Los autores del paquete recomiendan usar mínimo 100 simulaciones (Lakens & Caldwell, 2020).
4. `seed` (opcional): Adicionalmente, para el siguiente ejemplo usé el argumento `seed` para que las simulaciones den siempre el mismo resultado<sup>25</sup>.

<sup>24</sup>Por supuesto, estos datos son hipotéticos y no representan ningún estudio real, ni diferencias entre estudiantes con y sin beca, ni entre universidades públicas y privadas. Solo son usados como un ejemplo.

<sup>25</sup>Dado que las bases de datos se simulan con base en las características definidas con la función `ANOVA_design`, pero de manera

El código entonces quedaría de la siguiente manera:

```
ANOVA_power(diseñoB,
             alpha_level = 0.05,
             p_adjust = "holm",
             seed = 2019,
             nsims = 1000)
```

Lo que produce:

```
## Power and Effect sizes for ANOVA tests
##               power effect_size
## anova_Beca      100.0    0.117825
## anova_Universidad  24.0    0.007927
## anova_Beca:Universidad 98.3    0.052368
##
## Power and Effect sizes for pairwise comparisons (t-tests)
##               power effect_size
## p_Beca_Sí_Universidad_Pública_Beca_Sí_Universidad_Privada  50.3    0.3180
## p_Beca_Sí_Universidad_Pública_Beca_No_Universidad_Pública  39.5   -0.2672
## p_Beca_Sí_Universidad_Pública_Beca_No_Universidad_Privada 100.0   -0.8617
## p_Beca_Sí_Universidad_Privada_Beca_No_Universidad_Pública  95.7   -0.5843
## p_Beca_Sí_Universidad_Privada_Beca_No_Universidad_Privada 100.0   -1.1752
## p_Beca_No_Universidad_Pública_Beca_No_Universidad_Privada  96.1   -0.5951
```

Este es un resultado muy interesante y completo, que incluye dos tablas; primero, una denominada "Power and Effect sizes for ANOVA tests" para los efectos principales e interacciones del ANOVA. Y segundo, una tabla denominada "Power and Effect sizes for pairwise comparisons (t-tests)" que muestra las comparaciones entre niveles de mis factores<sup>26</sup> (por ejemplo, la comparación entre estudiantes becados de universidad pública y estudiantes becados de universidad privada). Ambas tablas tienen columnas que muestran:

1. El poder estadístico obtenido con el tamaño de muestra propuesto, bajo la columna denominada **power**.
2. El tamaño del efecto para cada efecto principal o interacción (primera tabla), o para cada comparación *post-hoc* (segunda tabla), bajo la columna denominada **effect\_size**.

Por ejemplo, en la primera tabla, bajo el título "Power and Effect sizes for ANOVA tests", el resultado muestra que con el  $n$  y características propuestas, este estudio tendría un poder estadístico de 100%<sup>27</sup> (o 1) para detectar el efecto principal de *Beca*, si es que este existe (es decir, una diferencia en las calificaciones del examen entre becados y no becados), cuyo tamaño del efecto se calculó en  $\eta_p^2 = 0.116942$ . Así mismo, un poder de 0.22 (o 22%), que es muy bajo, para detectar el efecto principal del tipo de universidad (*Universidad*: Pública, Privada), pues el efecto es sumamente pequeño, y se estimó en  $\eta_p^2 = 0.007245$ . Y, finalmente, un poder de 0.99 (99%) para detectar la interacción entre *Beca* y *Universidad*, pues el efecto es muy grande ( $\eta_p^2 = 0.050779$ ).

■ **Es importante tener en cuenta que, a diferencia de *pwr*, que usa como medida del tamaño del efecto  $f$  de Cohen, el tamaño del efecto para ANOVAs usado por el paquete *Superpower* para efectos principales e interacciones es  $\eta_p^2$  (eta parcial al cuadrado), en línea con lo recomendado por Correll et al. (2020).**

Adicionalmente, diseños tipo ANOVA, bien sean de una vía o factoriales<sup>28</sup>, con frecuencia requieren de pruebas *post-hoc* o contrastes planeados, para comparar niveles específicos de los factores.

Por esto, en la segunda tabla, el resultado muestra la información relevante para cada comparación *post-hoc* (prueba-*t*) que se podría realizar, bajo el título "Power and Effect sizes for pairwise comparisons (t-tests)", con la corrección deseada [en este caso, usando la corrección de *Holm-Bonferroni*; ver Holm (1979)]. Acá también muestra tanto el poder (**power**), como el tamaño del efecto (**effect\_size**<sup>29</sup>). Por ejemplo, nos muestra que

**aleatoria**, cada vez que corra la función obtendré un resultado ligeramente distinto, especialmente si el número de simulaciones (**nsim**) es pequeño. Al darle una semilla (**seed**), que puede ser cualquier número, los datos simulados siempre serán los mismos, garantizando que la respuesta sea la misma.

<sup>26</sup>Ver sección 7.3 *Acerca de comparaciones post-hoc*.

<sup>27</sup>En realidad, el poder no puede llegar a ser 100%, pero se puede aproximar infinitamente.

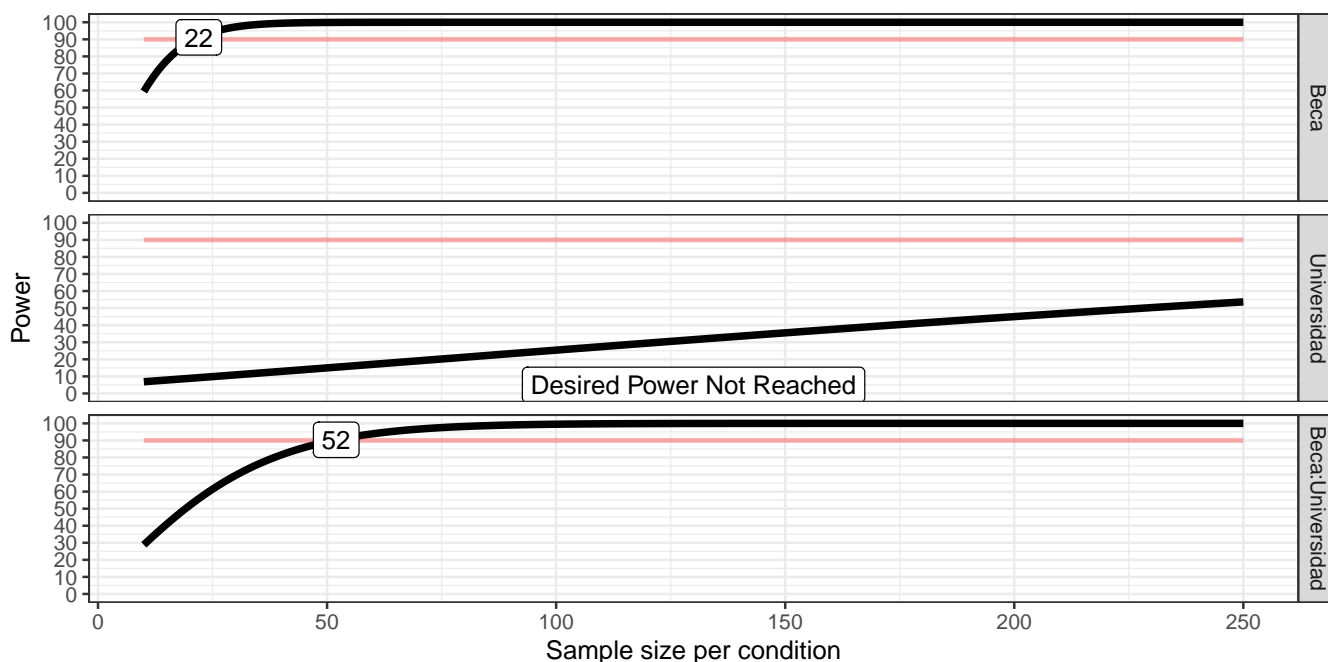
<sup>28</sup>Por supuesto, en el paquete *Superpower* se pueden hacer análisis de poder estadísticos para ANOVAs de una vía y, a diferencia de los análisis del paquete *pwr*, permiten estimar efectos para todas las comparaciones *post-hoc* en un solo análisis, sin pasos adicionales.

<sup>29</sup>Dado que estas comparaciones *post-hoc* serían típicamente analizadas con pruebas-*t*, el tamaño del efecto es  $d$  de Cohen.

el tamaño del efecto (diferencia) entre becados de universidad pública y becados de universidad privada, tiene un tamaño de  $d = 0.3247$ , y que con la muestra planteada de 80 participantes en cada combinación de *Beca* y *Universidad* (320 en total), tendríamos un poder  $1 - \beta$  de 0.49 (49%).

Finalmente, para ver gráficamente el poder estadístico a diferentes tamaños de muestra, puedo usar la función `plot_power`<sup>30</sup> (Figura 9), definiendo tanto el  $n$  mínimo (`min_n`) y el  $n$  máximo (`max_n`) que deseo incluir en mi figura:

```
plot_power(diseñoB,
  min_n = 10, max_n = 250,
  plot = TRUE)
```



**Figura 9.** Ejemplo de la asociación entre el tamaño de la muestra y el poder estadístico para un ANOVA  $2 \times 2$ , producida con la función `plot_power` del paquete `Superpower`. Como se puede ver, el poder obtenido según el tamaño de la muestra es diferente para cada efecto principal o interacción, pues el tamaño del efecto es diferente en cada caso, por lo que se requeriría un  $n$  diferente para alcanzar el mismo poder estadístico. En este ejemplo, un poder  $1 - \beta$  de 0.9 (90%) se alcanza para el efecto principal de la *Beca* (**panel superior**) con una muestra de apenas unos 24 participantes en cada combinación de *Beca* y *Universidad* (96 en total), mientras que ese mismo poder para detectar un efecto principal del tipo de *Universidad* (**panel intermedio**) no se logra ni siquiera con 250 participantes (1000 en total) por cada combinación de *Beca* y *Universidad* (de hecho, solo se logra al tener cerca de 600 participantes por condición, ¡o unos 2400 en total!). La interacción entre *Beca* y *Universidad* (**panel inferior**), logra un poder de 0.9 (90%) con unos 50 participantes (200 en total) por cada combinación de *Beca* y *Universidad*.

## 7.5 ANOVA factorial de medidas repetidas

En el paquete `Superpower` el análisis de poder para diseños factoriales de medidas repetidas es casi idéntico al de medidas independientes: se deben usar las funciones (1) `ANOVA_design` para especificar las características del diseño para el cual haré el análisis de poder, y (2) `ANOVA_power` y/o `plot_power` para el análisis de poder propiamente dicho.

Sin embargo, hay una diferencia importante: dado que tendremos factores de medidas repetidas, se debe especificar la correlación entre estos niveles, *concatenando* los  $r$  de Pearson con el argumento `r` de la función `ANOVA_design` (que no habíamos usado).

<sup>30</sup>Como sucedía con la función `plot` del paquete `pwr`, las figuras que produce la función `plot_power` del paquete `Superpower`, son objetos de clase `ggplot`, por lo que alguien familiarizado con el paquete `ggplot2` puede modificar la figura para que, por ejemplo, los ejes, título y anotaciones estén en español, o para cambiar el tema y colores.



El orden para ingresar estos coeficientes de correlación ( $r$  de Pearson), debe seguir un criterio específico. Este orden, que debe ser respetado, es equivalente al “*triángulo superior*” (resaltado en **amarillo**) de una matriz de correlaciones (**Tabla 4**).

**Tabla 4.** Orden para ingresar los coeficientes de una matriz de correlación entre niveles de un estudio  $2 \times 2$  de medidas repetidas

	A1 - B1	A1 - B2	A2 - B1	A2 - B2
A1 - B1	-	<b>1</b>	<b>2</b>	<b>3</b>
A1 - B2	1	-	<b>4</b>	<b>5</b>
A2 - B1	2	4	-	<b>6</b>
A2 - B2	3	5	6	-

*Nota:*

Los números representan el orden en el que deben ser ingresados los coeficientes de correlación en el argumento  $r$  de la función `ANOVA_design`. Se puede usar en triángulo superior (resaltado en amarillo).

### 7.5.1 Ejemplo ANOVA factorial de medidas repetidas ( $2 \times 2$ )

Por ejemplo, si tuviéramos un diseño donde medimos la ansiedad de un grupo de personas tras tomar una tasa de café con cafeína o descafeinado, en un día laboral y en un día de descanso, tendríamos un diseño  $2 \times 2$  con dos factores de medidas repetidas, cada uno con dos niveles:

- Factor 1: Cafeína (Sí, No)
- Factor 2: Día laboral (Sí, No)

En este caso, la ansiedad de cada participante sería medida cuatro veces, tras tomar:

1. Café con cafeína en un día laboral (Cafeína Sí; Día laboral Sí).
2. Café con cafeína en un día de descanso (Cafeína Sí; Día laboral No).
3. Café sin cafeína en un día laboral (Cafeína No; Día laboral Sí).
4. Café Sin cafeína en un día de descanso (Cafeína No; Día laboral No).

Lo importante es entonces especificar las correlaciones entre estas condiciones. Para esto, lo mejor es hacer una matriz de correlaciones (**Tabla 5**).

**Tabla 5.** Matriz de correlación hipotética entre niveles de un estudio  $2 \times 2$  de medidas repetidas

	Cafeína Sí; Día laboral Sí	Cafeína Sí; Día laboral No	Cafeína No; Día laboral Sí	Cafeína No; Día laboral No
Cafeína Sí; Día laboral Sí	1	<b>0.384</b>	<b>0.287</b>	<b>0.302</b>
Cafeína Sí; Día laboral No	0.384	1	<b>0.204</b>	<b>0.402</b>
Cafeína No; Día laboral Sí	0.287	0.204	1	<b>0.184</b>
Cafeína No; Día laboral No	0.302	0.402	0.184	1

*Nota:*

Los coeficientes de correlación resaltados en amarillo (triángulo superior), son los que se deben ingresar en el argumento  $r$  de la función `ANOVA_design`, siguiendo el orden especificado (e.g. Tabla 4). En gris está resaltada la diagonal de la matriz de correlaciones, que contiene la correlación de cada variable consigo misma.

Teniendo esto en cuenta, podemos especificar este diseño en la función `ANOVA_design`, incluyendo los coeficientes de correlación (**Tabla 5**) en el orden descrito (**Tabla 4**) en el argumento  $r$ . En total, los argumentos incluidos son:

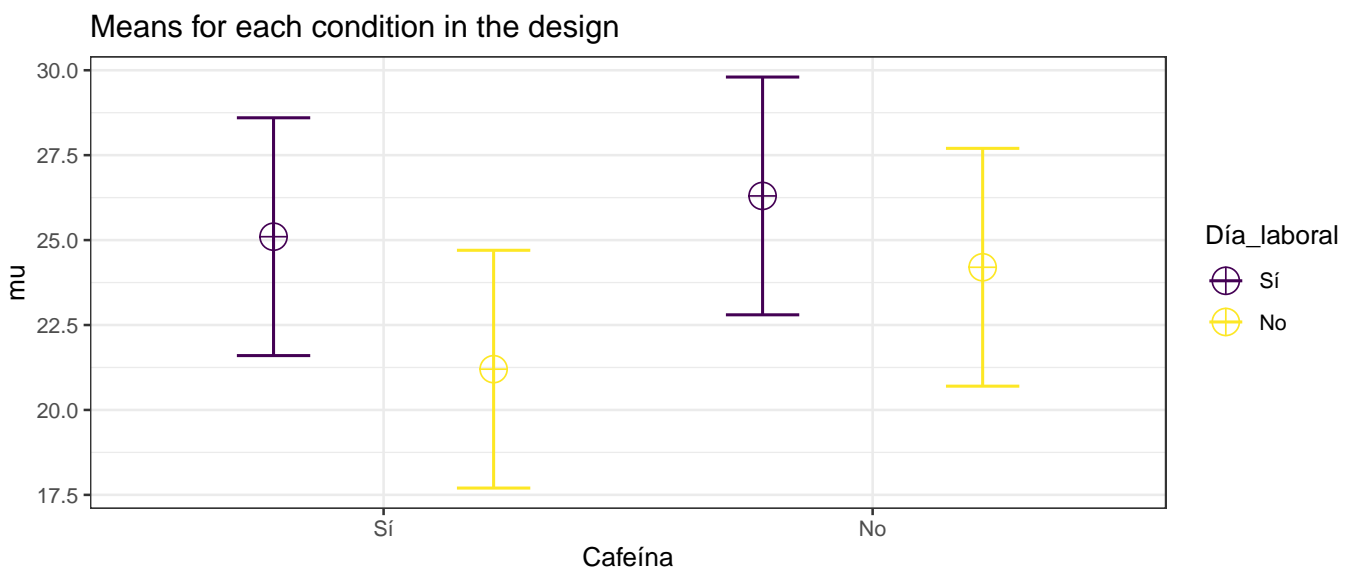
1. **design**: en este caso, dado que tengo un diseño  $2 \times 2$ , donde ambos factores son de medidas repetidas, debo ponerlo como "**2w\*2w**", donde los números representan el número de niveles en cada factor, la letra **w** que ese

factor en de medidas repetidas (o *intra sujetos*, por lo cual usa la letra **w**, del inglés *within-subjects*).

2. **n**: el número de participantes que espero tener; como mis factores son de medidas repetidas o intra-sujetos, lo importante es que para cada participante sea se hagan observaciones para cada condición (o, dicho de otro modo, por combinación de niveles de mis factores; e.g. (1) tras tomar café con cafeína en un día laboral; (2) tras tomar café con cafeína en un día no laboral, etcétera). Como lo mencioné anteriormente, a diferencia de otros paquetes y programas, **Superpower** no calcula el *n* por mí, pero me permite cambiar el *n* hasta lograr el poder deseado.
3. **mu**: las medias para cada interacción entre los niveles de mis factores. En este caso, la media de la ansiedad para participantes al tomar (1) café con cafeína en un día laboral, (2) café con cafeína en un día no laboral, (3) café descafeinado en un día laboral, y (4) café descafeinado en un día no laboral. Estos valores, como antes, deben estar *concatenados* usando la función **c**.
4. **sd**: la desviación estándar para la población (por lo cual es un solo valor. En este caso, la desviación estándar de los puntajes de ansiedad).
5. **r**: los coeficientes de correlación entre combinaciones de mis factores intra-sujetos o de medidas repetidas, en el orden correcto (como fue descrito en la **Tabla 4**).
6. **labelnames** (Opcional): las etiquetas (nombres) de los factores y sus niveles. Al igual que con las medias, estas etiquetas deben estar concatenadas usando la función **c**. Como expliqué en la sección **7.4 ANOVA factorial de medidas independientes**, para definirlos, se deben poner en el siguiente orden:
  - Etiqueta del primer factor (en este caso “Cafeína”)
  - Etiquetas de los niveles de ese factor (en este caso “Sí” y “No”)
  - Etiqueta del segundo factor (en este caso “Día\_laboral”, pues estos nombres NO pueden tener espacios).
  - Etiquetas de los niveles de ese factor (en este caso “Sí” y “No”).

En este caso, el diseño lo *guardaré* como un objeto llamado **diseñoW**, y usé la opción **plot = TRUE** para asegurarme de que las medias fueron *concatenadas* en el orden correcto (**Figura 10**).

```
diseñoW <- ANOVA_design(design = "2w*2w",
  n = 100, sd = 3.5,
  mu = c(25.1, 21.2, 26.3, 24.2),
  r <- c(0.384, 0.287, 0.302, 0.204, 0.402, 0.184),
  labelnames = c("Cafeína", "Sí", "No", "Día_laboral", "Sí", "No"),
  plot = TRUE)
```



**Figura 10.** Ejemplo de la distribución de medias marginales estimadas y sus intervalos de confianza para el diseño definido con la función **ANOVA\_design**, al incluir el argumento **plot = TRUE**. Esto es muy útil para estar seguro de que las medias fueron *concatenadas* en el orden correcto.

Del mismo modo, para asegurarme de que los coeficientes de correlación fueron *concatenados* en el orden correcto, puedo pedir una matriz de correlaciones usando el nombre del objeto que contiene el diseño (en este caso `diseñoW`) y agregando `$cor_mat`, y confirmar que los valores y su orden corresponden con la matriz original (en este caso, en la **Tabla 5**).

```
diseñoW$cor_mat
```

```
##      Sí_Sí Sí_No No_Sí No_No
## Sí_Sí 1.000 0.384 0.287 0.302
## Sí_No 0.384 1.000 0.204 0.402
## No_Sí 0.287 0.204 1.000 0.184
## No_No 0.302 0.402 0.184 1.000
```

Una vez definido el diseño y las características de los datos (“*guardando*” este diseño en un objeto que llamé `diseñoW`), puedo ver el poder que obtendría con esos 100 participantes por cada combinación de *Cafeína* y *Día laboral*, con la función `ANOVA_power`, de la misma manera que lo haría para un diseño de medidas independientes.

```
ANOVA_power(diseñoW,
             alpha_level = 0.05,
             p_adjust = "holm",
             seed = 1685,
             nsims = 1000)
```

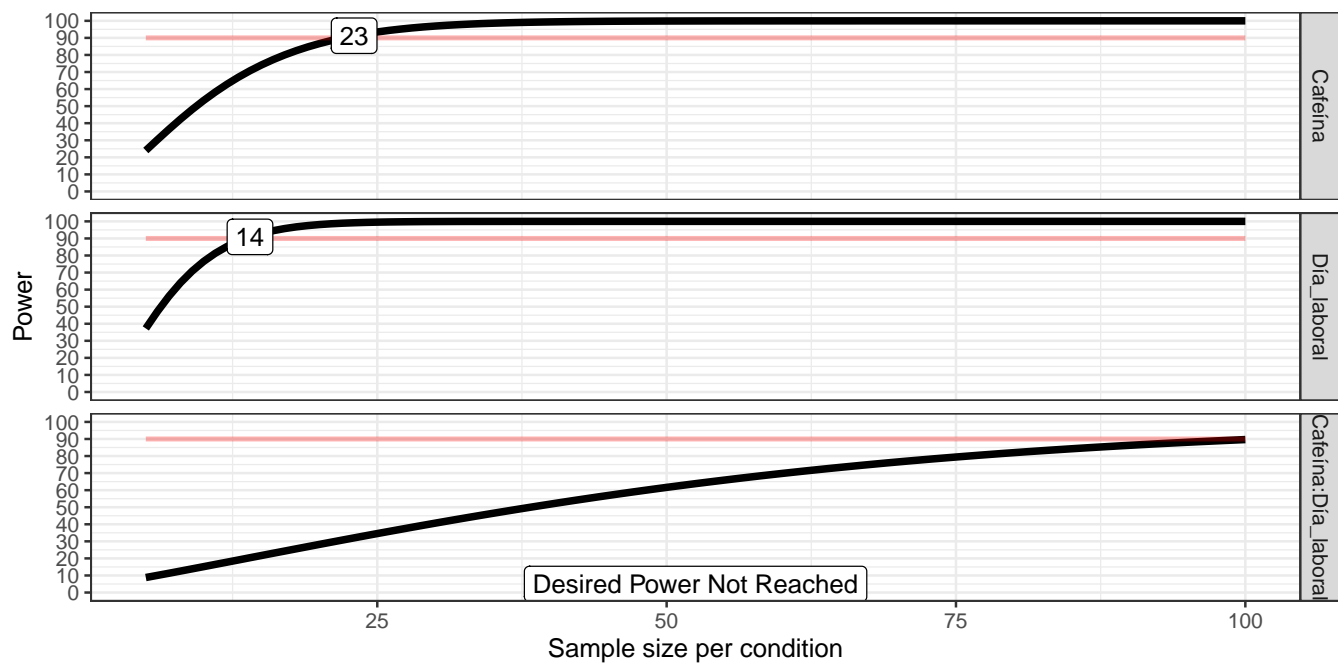
```
## Power and Effect sizes for ANOVA tests
##                                     power effect_size
## anova_Cafeína                      100.0      0.3469
## anova_Día_laboral                  100.0      0.4795
## anova_Cafeína:Día_laboral          89.9      0.1029
##
## Power and Effect sizes for pairwise comparisons (t-tests)
##                                     power effect_size
## p_Cafeína_Sí_Día_laboral_Sí_Cafeína_Sí_Día_laboral_No 100.0      -1.0095
## p_Cafeína_Sí_Día_laboral_Sí_Cafeína_No_Día_laboral_Sí  76.1       0.2875
## p_Cafeína_Sí_Día_laboral_Sí_Cafeína_No_Día_laboral_No  55.0      -0.2210
## p_Cafeína_Sí_Día_laboral_No_Cafeína_No_Día_laboral_Sí 100.0       1.1612
## p_Cafeína_Sí_Día_laboral_No_Cafeína_No_Día_laboral_No 100.0       0.7888
## p_Cafeína_No_Día_laboral_Sí_Cafeína_No_Día_laboral_No  98.4      -0.4731
```

Como se puede ver, el poder estadístico para detectar posibles efectos de *Cafeína* ( $1 - \beta = 100\%$ ), *Día laboral* ( $1 - \beta = 100\%$ ), y la interacción *Cafeína*  $\times$  *Día laboral* ( $1 - \beta = 89.9\%$ ), es más que suficiente con 100 participantes por condición (y casi el nivel deseado para la interacción), pues los tamaños del efecto estimado, son relativamente grandes ( $\eta_p^2 = 0.3469, 0.4795$ , y  $0.1029$ , respectivamente).

Del mismo modo, un  $n$  de 100 participantes por condición, da un poder suficiente ( $1 - \beta > 0.9$ ) para detectar casi todas las diferencias según fueron estimadas, excepto: (1) la diferencia, tal cual fue estimada, entre los niveles de ansiedad tras tomar café con cafeína vs café descafeinado en un día laboral ( $1 - \beta = 76.1\%$ ); y (2), la diferencia entre los niveles de ansiedad tras tomar café con cafeína en un día laboral, vs café descafeinado en un día no laboral ( $1 - \beta = 55\%$ ), que por supuesto tienen los tamaños de efecto más pequeños (más cercanos a 0), independientemente de su dirección ( $d = 0.2875$ , y  $-0.2210$ , respectivamente).

Sin embargo, para estimar el tamaño de muestra suficiente, puedo como antes usar la función `plot_power` (**Figura 11**), definiendo tanto el  $n$  mínimo (`min_n`) y el  $n$  máximo (`max_n`) que deseo incluir en mi figura:

```
plot_power(diseñoW,
           min_n = 5,
           max_n = 100,
           plot = TRUE)
```



**Figura 11.** Ejemplo de la asociación entre el tamaño de la muestra y el poder estadístico para un ANOVA  $2 \times 2$ , producida con la función `plot_power` del paquete `Superpower`. Como se puede ver, el poder obtenido según el tamaño de la muestra es diferente para cada efecto principal o interacción, pues el tamaño del efecto es diferente en cada caso, por lo que se requeriría un  $n$  diferente para alcanzar el mismo poder estadístico. En este ejemplo, un poder  $1 - \beta$  de 0.9 (90%) se alcanza para el efecto principal de la *Cafeína* (**panel superior**) con una muestra de apenas unos 20 participantes, mientras que ese mismo poder para detectar un efecto principal del tipo de *Día laboral* (**panel intermedio**) se logra con alrededor de 13 participantes, y la interacción entre *Cafeína* y *Día laboral* (**panel inferior**), logra un poder de 0.9 (90%) con unos 100 participantes. Si mi interés principal es la interacción entre estas variables, debo entonces usar una muestra de unos 100 participantes, a los cuales se les medirá la ansiedad en las cuatro condiciones en cada condición.

## 7.6 ANOVA factorial mixto

Si se entiende cómo hacer análisis de poder para diseños factoriales de medidas repetidas, y de medidas independientes en el paquete `Superpower`, hacer análisis para diseños mixtos es sencillo.

Básicamente, un diseño mixto es cuando tenemos factores (variables independientes nominales) tanto de medidas repetidas como independientes. Como tal, requiere combinar elementos de los análisis de poder de medidas independientes, con los de medidas repetidas, y se hace con las mismas funciones ya usadas: (1) `ANOVA_design` para especificar las características del diseño para el cual haré el análisis de poder, y (2) `ANOVA_power` y/o `plot_power` para el análisis de poder propiamente dicho.

Las diferencias son que, al especificar el diseño con la función `ANOVA_design` (argumento `design`), se debe especificar que hay niveles intra-sujetos (de medidas repetidas, que en el paquete se designado como **w**, del inglés *within-subjects*) y entre-sujetos (medidas independientes, **b**, del inglés *between-subjects*).

Adicionalmente, tenemos que especificar la correlación entre los niveles de factores intra-sujeto (tal cual como lo hicimos para diseños de medidas repetidas), pero no para factores de medidas independientes, pues en este caso se asume siempre que la correlación es 0<sup>31</sup>. De este modo, el número de coeficientes de correlación que debemos especificar (en el argumento `r` de la función `ANOVA_design`) es menor al de diseños equivalentes de medidas repetidas.

### 7.6.1 Ejemplo ANOVA factorial mixto ( $2 \times 2$ )

Por ejemplo, si quisiera medir qué tanto un ruido insoportable (por ejemplo, el sonido de la *fresa* de un odontólogo afecta el desempeño de jugadores aficionados de ajedrez<sup>32</sup>, en términos de partidas ganadas sobre un total de 20

<sup>31</sup>Por esto al hacer análisis de medidas independientes no se especifica ninguna correlación.

<sup>32</sup>Es solamente un ejemplo. Jamás pensaría en torturar a jugadores de ajedrez, ni a nadie, sometándolo a semejante tortura solo por curiosidad “científica”.

partidas jugadas, y quisiera saber si este efecto es diferente en personas según si son odontólogos o no, tendría un diseño mixto  $2 \times 2$ , pues mis factores serían:

1. Factor 1 (medidas independientes): Profesión (Odontólogo, Otro)
2. Factor 2 (medidas repetidas): Ruido (NO, Sí)

En este caso, tendría que someter a cada uno de mis participantes, odontólogos o no, a dos rondas de 20 juegos de ajedrez (una ronda con, y otra sin presencia del ruido).

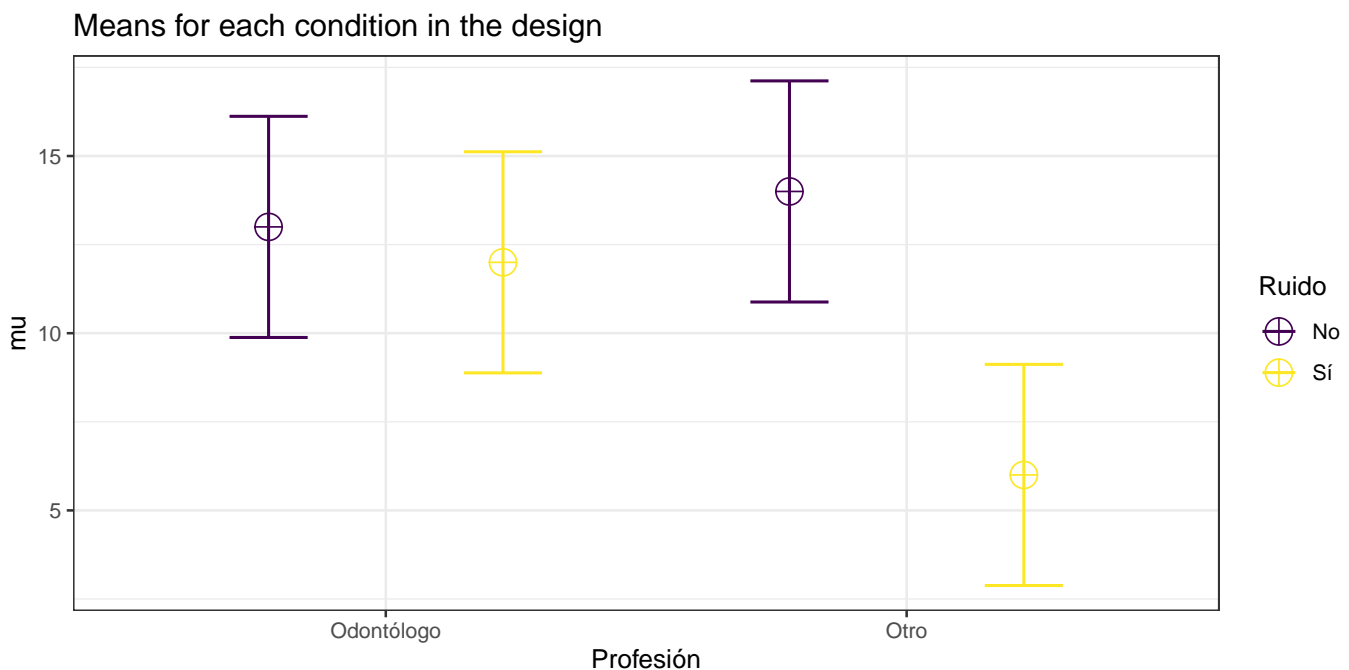
Como siempre, con la información clara, puedo definir los argumentos del diseño, con la función `ANOVA_design`, teniendo en cuenta que, como esta vez solo tengo un factor de medidas repetidas (*Ruido*), y este tiene solo dos niveles (sí, No), solo debo especificar ese coeficiente de correlación.

En este caso, el diseño lo “guardaré” como un objeto llamado `diseñoM`, y usé la opción `plot = TRUE` para asegurarme de que las medias fueron concatenadas en el orden correcto.

Como se puede ver en el código a continuación, y en la **Figura 12**, según mis datos (inventados), tanto odontólogos como no odontólogos ganan en promedio cerca de 12 partidas de 20 jugadas ( $\approx 60\%$ ), cuando las juegan sin ruido (en morado), pero el desempeño de personas con profesiones distintas a la odontología se ve muy afectado al jugar las partidas en presencia del ruido.

Dado que no sería fácil conseguir voluntarios para someterse a jugar, en total, 40 partidas de ajedrez, de las cuales 20 se jugarían con un ruido insoportable, pero que también espero un tamaño de efecto grande, voy a hacer el cálculo solo con 10 participantes por grupo (10 odontólogos, 10 no odontólogos).

```
diseñoM <- ANOVA_design(design = "2b*2w",
  n = 10,
  mu = c(13, 12, 14, 6),
  sd = 3.12,
  r = 0.3,
  labelnames = c("Profesión", "Odontólogo", "Otro", "Ruido", "No", "Sí"),
  plot = TRUE)
```



**Figura 12.** Ejemplo de la distribución de medias marginales estimadas y sus intervalos de confianza para el diseño definido con la función `ANOVA_design`, al incluir el argumento `plot = TRUE`. Esto es muy útil para estar seguro de que las medias fueron concatenadas en el orden correcto.

También, como siempre, puedo pedir una matriz de correlaciones usando el nombre del objeto que contiene el diseño

(en este caso `diseñoM` y agregando `$cor_mat`) para asegurarme de que los coeficientes de correlación están en el orden correcto.

```
diseñoM$cor_mat
```

```
##           Odontólogo_No Odontólogo_Sí Otro_No Otro_Sí
## Odontólogo_No           1.0           0.3      0.0      0.0
## Odontólogo_Sí           0.3           1.0      0.0      0.0
## Otro_No                0.0           0.0      1.0      0.3
## Otro_Sí                0.0           0.0      0.3      1.0
```

Definido el diseño y las características de los datos (que “*guardé*” en un objeto que llamé `diseñoM`), puedo ver el poder que obtendría con esos 10 participantes por grupo, con la función `ANOVA_power`, de la misma manera que lo haría para diseños de medidas independientes o repetidas.

```
ANOVA_power(diseñoM,
             alpha_level = 0.05,
             p_adjust = "holm",
             seed = 1985,
             nsims = 1000)
```

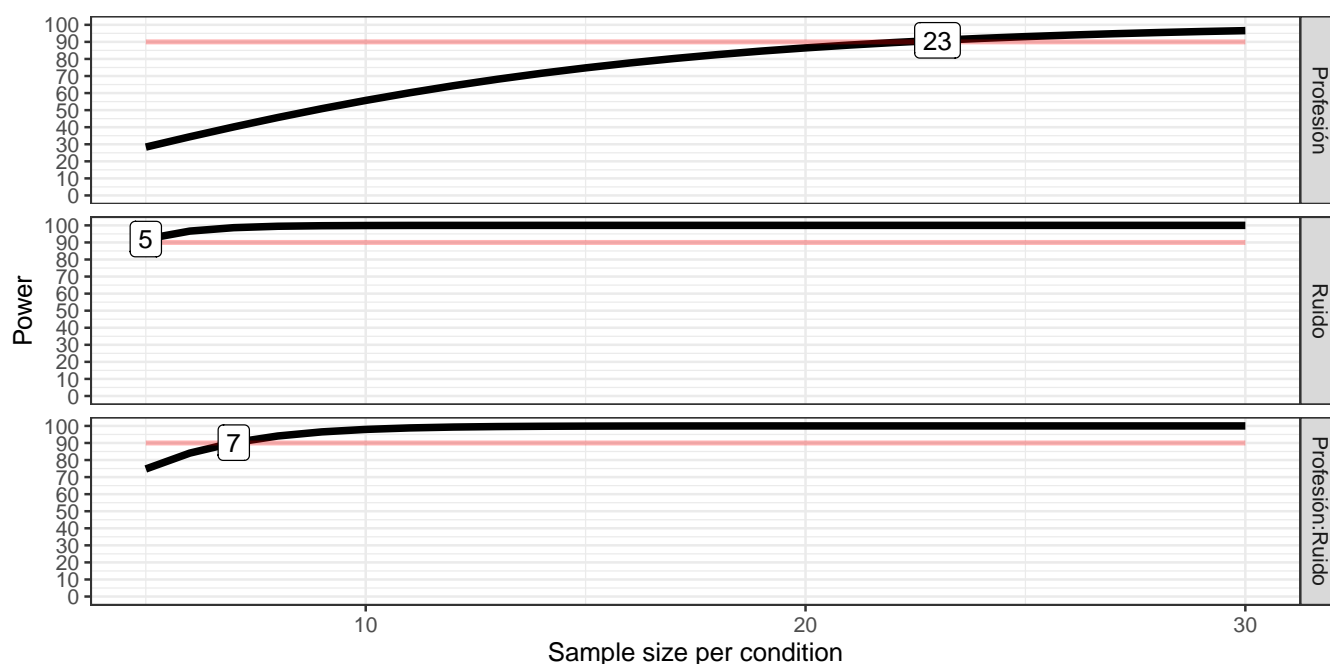
```
## Power and Effect sizes for ANOVA tests
##           power effect_size
## anova_Profesión      57.2      0.2368
## anova_Ruido         100.0      0.6248
## anova_Profesión:Ruido 97.2      0.5070
##
## Power and Effect sizes for pairwise comparisons (t-tests)
##           power effect_size
## p_Profesión_Odontólogo_Ruido_No_Profesión_Odontólogo_Ruido_Sí    5.0     -0.2934
## p_Profesión_Odontólogo_Ruido_No_Profesión_Otro_Ruido_No          6.1      0.3345
## p_Profesión_Odontólogo_Ruido_No_Profesión_Otro_Ruido_Sí         98.8     -2.3496
## p_Profesión_Odontólogo_Ruido_Sí_Profesión_Otro_Ruido_No         12.6      0.6522
## p_Profesión_Odontólogo_Ruido_Sí_Profesión_Otro_Ruido_Sí         93.9     -2.0133
## p_Profesión_Otro_Ruido_No_Profesión_Otro_Ruido_Sí              99.8     -2.3956
```

Como se puede ver, el poder estadístico para detectar posibles efectos de *Profesión* ( $1 - \beta = 57.2\%$ ) es algo bajo con solo 10 participantes por grupo, mientras que el poder para detectar un efecto de la condición de *Ruido* ( $1 - \beta = 100\%$ ), y la interacción *Profesión*  $\times$  *Ruido* ( $1 - \beta = 97.2\%$ ), es más que suficiente. Esto se da porque los tamaños de efecto son bastante grandes ( $\eta_p^2 = 0.2368, 0.6248$ , y  $0.5070$ , respectivamente), por lo cual nunca será necesaria una muestra demasiado grande (ni siquiera en el caso de la *Profesión*, como lo sugiere el hecho de que con apenas 10 participantes por grupo se obtenga un poder no muy lejano al deseado).

El poder para detectar comparaciones *post-hoc* es, sin embargo, muy bajo en varios casos. Solamente al comparar a los no odontólogos jugando con ruido vs odontólogos jugando con o sin ruido, o esos mismo no odontólogos jugando sin ruido, se logra un poder suficiente (lo que no es una sorpresa si se tienen en cuenta las medias esperadas; **Figura 11**).

Sin embargo, para estimar el tamaño de muestra suficiente, puedo como antes usar la función `plot_power` (**Figura 13**), definiendo tanto el  $n$  mínimo (`min_n`) y el  $n$  máximo (`max_n`) que deseo incluir en mi figura:

```
plot_power(diseñoM,
           min_n = 1,
           max_n = 30,
           plot = TRUE)
```



**Figura 13.** Ejemplo de la asociación entre el tamaño de la muestra y el poder estadístico para un ANOVA  $2 \times 2$  mixto, producida con la función `plot_power` del paquete **Superpower**. Como se puede ver, el poder obtenido según el tamaño de la muestra es diferente para cada efecto principal o interacción, pues el tamaño del efecto es diferente en cada caso, por lo que se requeriría un  $n$  diferente para alcanzar el mismo poder estadístico. En este ejemplo, un poder  $1 - \beta$  de 0.9 (90%) se alcanza para el efecto principal de la *Profesión* (**panel superior**) con una muestra de apenas unos 18 participantes por grupo, mientras que ese mismo poder para detectar un efecto principal del tipo de *Ruido* (**panel intermedio**) se logra con al apenas unos 8 participantes por grupo, y la interacción entre *Profesión y Ruido* (**panel inferior**), logra un poder de 0.9 (90%) con unos 6 participantes por grupo. Si me interesan tanto los efectos principales como la interacción entre estas variables, debo entonces usar una muestra de unos 18 participantes por grupo (18 odontólogos y 18 no odontólogos, para un total de 36 participantes), a los cuales se les medirá el número de partidas de ajedrez ganadas de 20 jugadas, en las dos condiciones de ruido.

## 7.7 Extra: Cómo estima **Superpower** el poder estadístico con base en simulaciones de bases de datos

El poder, como lo mencioné en la sección 1.1 *¿Qué es potencia o poder estadístico?*, es la probabilidad de detectar, como significativo (es decir, con un  $p < \alpha$ , que típicamente se establece en 0.05), un efecto, cuando este existe. Si aspiramos a tener un poder  $1 - \beta$  de 0.9 (90%), en el 90% de los casos deberíamos encontrar un  $p < \alpha$  (es decir, significativo).

Como lo mencioné brevemente antes, la función `ANOVA_power` simula un número de bases de datos (que definimos con el argumento `nsims`, para el que yo en todos los ejemplos he usado 1000 simulaciones). Estas bases de datos tienden a seguir las características que definí al usar la función `ANOVA_design`, pero varían aleatoriamente, como sucedería con datos reales, que difícilmente se ajustan exactamente a lo esperado.

Entonces, si al usar la función `ANOVA_power` pido 1000 simulaciones (`nsims = 1000`), se crearán 1000 bases de datos aleatorias. Para cada una se hace el ANOVA y las comparaciones *post-hoc*, y sus valores  $p$ . Entonces, se mira la probabilidad para cada uno de esos resultados de obtener un valor significativo (dado el  $\alpha$  definido con el argumento `alpha_level`, que suele definirse en 0.05). En otras palabras, ¿en cuántas de esas 1000 simulaciones se obtuvo un resultado significativo? Ese porcentaje, es el poder calculado empíricamente.

Ahora, ¿cuál es la distribución de los valores  $p$  para cada efecto principal, interacción, o comparación *post-hoc*? Por suerte, **Superpower** tiene opciones para ver esto gráficamente.

Si yo “guardo” cualquier análisis de poder con base en simulaciones creado con la función `ANOVA_power`, puedo usar el nombre del objeto en el que grabé ese análisis de poder, y agregar `$plot1` para ver la distribución de los valores  $p$  para efectos principales e interacciones, o `$plot2` para comparaciones *post-hoc*.



Por ejemplo, si “guardo” una de las simulaciones hechas (en este caso, usaré la simulación creada para el ANOVA factorial mixto), en un objeto, que ahora llamaré `simM`:

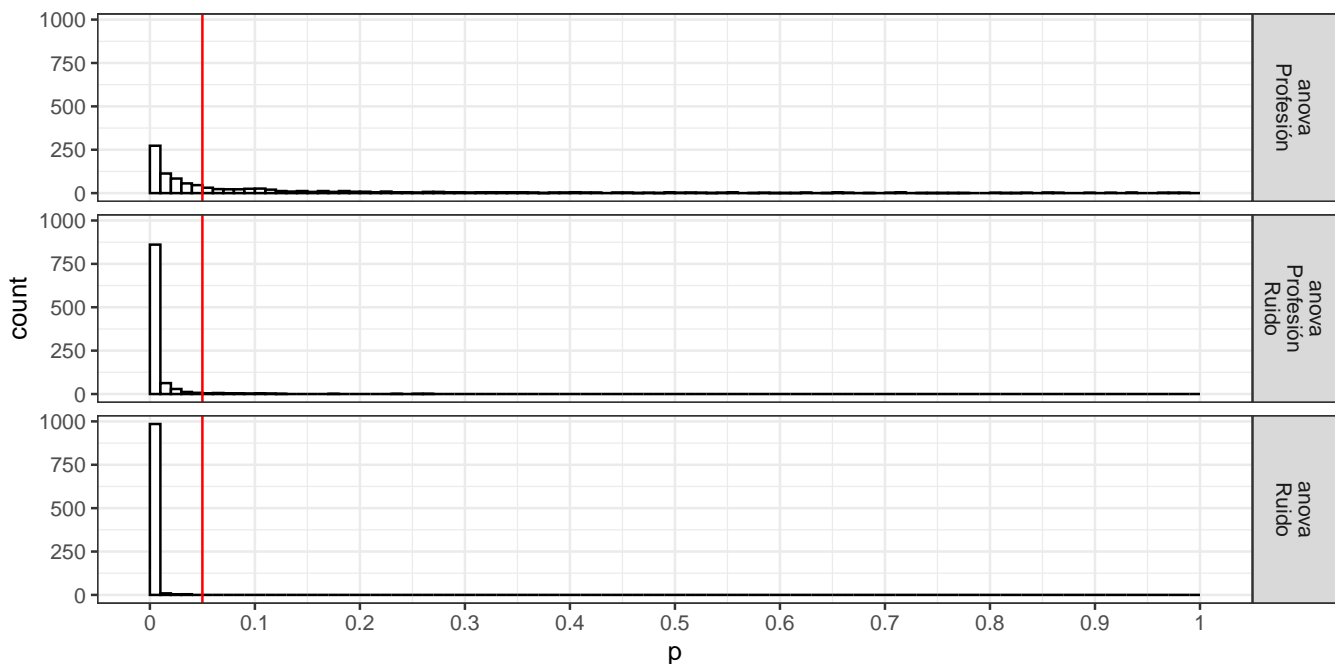
```
simM <- ANOVA_power(disenom,
  alpha_level = 0.05,
  p_adjust = "holm",
  seed = 1985,
  nsims = 1000)

## Power and Effect sizes for ANOVA tests
##               power effect_size
## anova_Profesión      57.2      0.2368
## anova_Ruido          100.0      0.6248
## anova_Profesión:Ruido 97.2      0.5070
##
## Power and Effect sizes for pairwise comparisons (t-tests)
##                                     power effect_size
## p_Profesión_Odontólogo_Ruido_No_Profesión_Odontólogo_Ruido_Sí    5.0      -0.2934
## p_Profesión_Odontólogo_Ruido_No_Profesión_Otro_Ruido_No           6.1       0.3345
## p_Profesión_Odontólogo_Ruido_No_Profesión_Otro_Ruido_Sí          98.8     -2.3496
## p_Profesión_Odontólogo_Ruido_Sí_Profesión_Otro_Ruido_No          12.6       0.6522
## p_Profesión_Odontólogo_Ruido_Sí_Profesión_Otro_Ruido_Sí          93.9     -2.0133
## p_Profesión_Otro_Ruido_No_Profesión_Otro_Ruido_Sí                99.8     -2.3956
##
##
## Within-Subject Factors Included: Check MANOVA Results
```

Puedo ver la distribución de los valores  $p$  para efectos principales e interacciones (**Figura 14**) con el comando:

```
simM$plot1
```

Que produce:

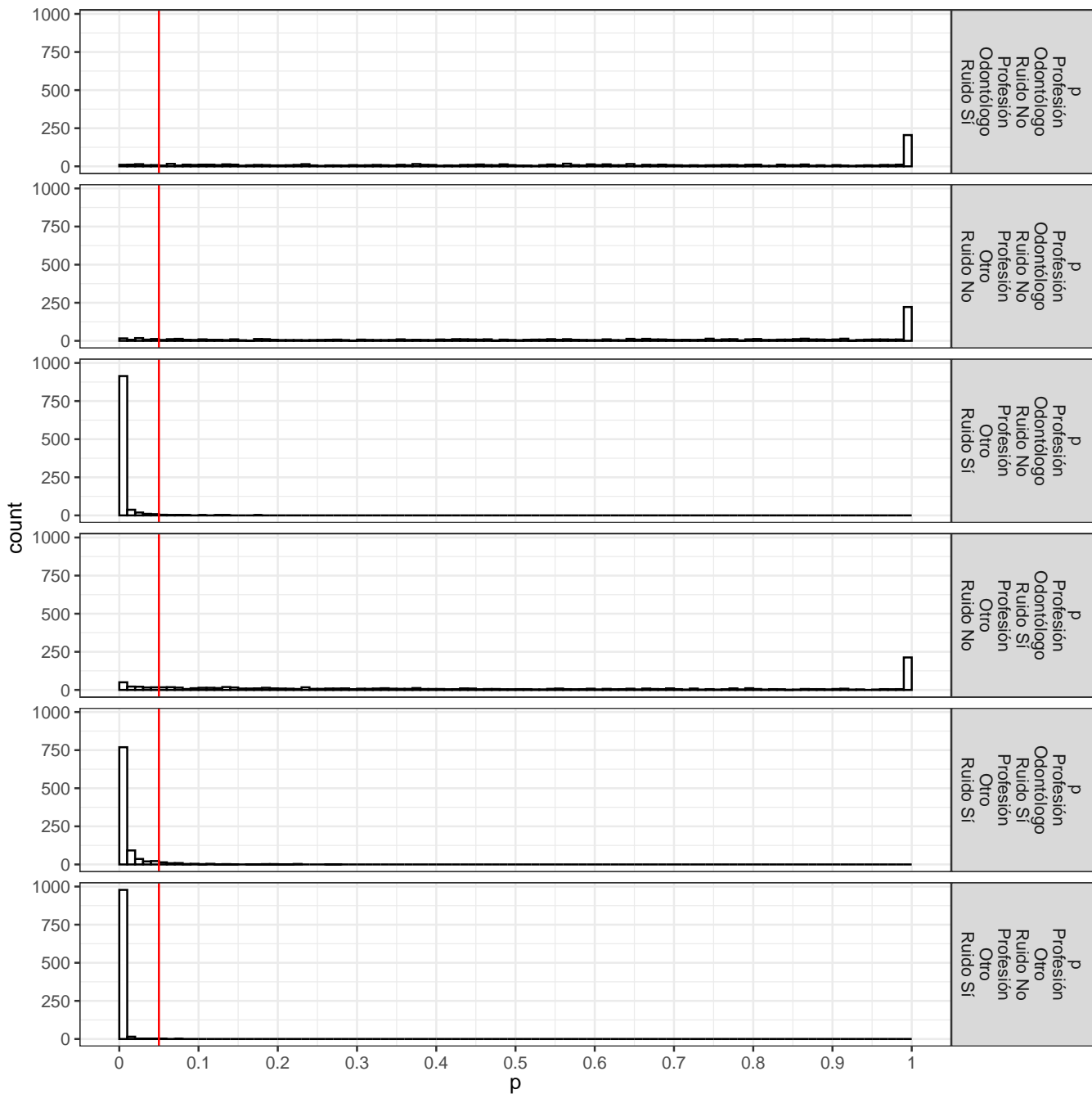


**Figura 14.** Ejemplo de distribución (histograma) de valores  $p$  para efectos principales e interacciones, producto de 1000 simulaciones hechas con la función `ANOVA_power` del paquete **Superpower**. La línea roja determina el nivel de significación estadística ( $\alpha$ ) definido (en este caso, el típico 0.05).

O la distribución de los valores  $p$  para las comparaciones *post-hoc* (**Figura 14**) con el comando:

```
simM$plot2
```

Que produce:



**Figura 14.** Ejemplo de distribución (histograma) de valores  $p$  para comparaciones *post-hoc*, producto de 1000 simulaciones hechas con la función `ANOVA_power` del paquete **Superpower**. Al contrario de las distribuciones de valores  $p$  para efectos principales e interacciones (**Figura 13**), que tienen una tendencia clara, para comparaciones *post-hoc* como estas la distribución puede verse extraña (aparecen, de repente, muchos unos) para comparaciones de efectos de tamaño pequeño, y que por ende tienen bajo poder, dada la corrección solicitada de *Holm-Bonferroni*. La línea roja determina el nivel de significación estadística ( $\alpha$ ) definido (en este caso, el típico 0.05).

## 8 Referencias

- Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, 74, 187–195. <https://doi.org/10.1016/j.jesp.2017.09.004>
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604), 452. <https://doi.org/10.1038/533452a>
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. <https://doi.org/10.1038/s41562-017-0189-z>
- Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds III, C. F., & Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*, 23(2), 255–264. <https://doi.org/10.1037/a0012850>
- Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze*.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Caldwell, A. R., & Lakens, D. (2020). *Power Analysis with Superpower*. <https://aaroncaldwell.us/SuperpowerBook/>.
- Caldwell, A. R., Lakens, D., DeBruine, L., & Love, J. (2020). *Superpower: Simulation-Based Power Analysis for Factorial Designs*.
- Champely, S., Ekstrom, C., Dalggaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *Pwr: Basic Functions for Power Analysis*.
- Chatham, K. (1999). *Planned Contrasts: An Overview of Comparison Methods*.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Correa, J. C. (2020). Scripts en R [Video]. In *YouTube*. <https://www.youtube.com/watch?v=ejQ0BS2gVJI>.
- Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen’s “Small,” “Medium,” and “Large” for Power Analysis. *Trends in Cognitive Sciences*, 24(3), 200–207. <https://doi.org/10.1016/j.tics.2019.12.009>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- 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. <https://doi.org/10.3758/BF03193146>
- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., Adolphi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., ... Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3), 168–171. <https://doi.org/10.1038/s41562-018-0311-x>
- Lakens, D., & Caldwell, A. R. (2020). Introduction to Superpower. In *The Comprehensive R Archive Network*. <http://shorturl.at/fnDX6>.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585. <https://doi.org/10.1126/science.aal3618>
- Quintana, D. S. (2017). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology*, 54(3), 344–349. <https://doi.org/10.1111/psyp.12798>
- Quintana, D. S. (2019). A non-technical guide to performing power analysis in R [Video]. In *YouTube*. <https://youtu.be/ZIjOG8LTTh8>.
- Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating

Cohen's  $f^2$ , a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, 3, 111. <https://doi.org/10.3389/fpsyg.2012.00111>

Streiner, D. L. (2015). Best (but oft-forgotten) practices: The multiple problems of multiplicity—whether and how to correct for many statistical tests. *The American Journal of Clinical Nutrition*, 102(4), 721–728. <https://doi.org/10.3945/ajcn.115.113548>

## Agradecimientos


Quiero agradecer especialmente a la Dra. [Milena Vásquez-Amézquita](#), investigadora del Laboratorio de Psicología Experimental de la Universidad El Bosque (Bogotá, Colombia) por la sugerencia de hacer un video acerca de este tema, que derivó además en la creación de este documento. Además, quiero agradecer especialmente a la Dra. [Maria Fernanda Reyes](#) por sus aportes críticos y comentarios a este trabajo.

## Acerca de este trabajo

Este trabajo está motivado por la triste escasez de fuentes de calidad, exhaustivas y actualizadas en español. Por eso, es de uso libre, pues mi única intención es apoyar a investigadores, docentes y estudiantes a entender la importancia de los análisis de poder para calcular un tamaño de muestra adecuado, y fomentar la generación de conocimiento confiable, con base en fundamentos estadísticos sólidos.

Como tal, puedes usarlo (junto con el video [Cómo calcular el tamaño de muestra para un experimento](#), de mi canal de YouTube [Investigación Abierta](#)) por ejemplo como material de clase, para tu propio aprendizaje, o como base para la planeación de un estudio.

Hacerlo, sin embargo, ha requerido de muchas horas de trabajo. Así que, si el trabajo te ha servido, te pido que me des el crédito debido.

Por esto, este trabajo está bajo una licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0) . Esta licencia te permite copiar y redistribuir este trabajo libremente, pero **debes dar crédito de manera adecuada**. Para más información, puedes ver el [resumen de la licencia CC BY 4.0](#).

Para cumplir con esto, por favor cita este documento correctamente. Por ejemplo, en algunos estilos comunes:

### APA (7a edición)

Leongómez, J. D. (2020). *Análisis de poder estadístico y cálculo de tamaño de muestra en R: Guía práctica*. Zenodo. <https://doi.org/10.5281/zenodo.3988776>

### MLA


Leongómez, Juan David. “Análisis de poder estadístico y cálculo de tamaño de muestra en R: Guía práctica”. Zenodo, Zenodo, agosto de 2020, [doi:10.5281/zenodo.3988776](https://doi.org/10.5281/zenodo.3988776).

### Chicago

Leongómez, Juan David. 2020. “Análisis de poder estadístico y cálculo de tamaño de muestra en R: Guía práctica”. Zenodo, agosto. <https://doi.org/10.5281/zenodo.3988776>.

Aunque he sido tan meticuloso como el tiempo me lo ha permitido, te ruego me hagas saber si encuentras un error, escribiéndome al correo [jleongomez@unbosque.edu.co](mailto:jleongomez@unbosque.edu.co). Trataré de corregirlo inmediatamente.

---

Juan David Leongómez, 2020 

---

# APÉNDICE: Paquetes de *R* usados en la creación de este documento

**R version 4.3.1 (2023-06-16 ucrt)**

**Platform:** x86\_64-w64-mingw32/x64 (64-bit)

**attached base packages:** *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

**other attached packages:** *pander*(v.0.6.5), *scales*(v.1.2.1), *faux*(v.1.2.1), *effectsize*(v.0.8.5), *reportr*(v.1.3.0), *kableExtra*(v.1.3.4), *lubridate*(v.1.9.2), *forcats*(v.1.0.0), *stringr*(v.1.5.0), *dplyr*(v.1.1.2), *purrr*(v.1.0.2), *readr*(v.2.1.4), *tidyr*(v.1.3.0), *tibble*(v.3.2.1), *tidyverse*(v.2.0.0), *ggplot2*(v.3.4.3), *Superpower*(v.0.2.0), *pwr*(v.1.3.0) and *knitr*(v.1.43)

**loaded via a namespace (and not attached):** *tidyselect*(v.1.2.0), *viridisLite*(v.0.4.2), *farver*(v.2.1.1), *fastmap*(v.1.1.1), *TH.data*(v.1.1-2), *bayestestR*(v.0.13.1), *digest*(v.0.6.33), *estimability*(v.1.4.1), *timechange*(v.0.2.0), *lifecycle*(v.1.0.3), *survival*(v.3.5-7), *magrittr*(v.2.0.3), *compiler*(v.4.3.1), *rlang*(v.1.1.1), *tools*(v.4.3.1), *utf8*(v.1.2.3), *yaml*(v.2.3.7), *labeling*(v.0.4.3), *plyr*(v.1.8.8), *xml2*(v.1.3.5), *multcomp*(v.1.4-25), *abind*(v.1.4-5), *withr*(v.2.5.0), *numDeriv*(v.2016.8-1.1), *grid*(v.4.3.1), *afex*(v.1.3-0), *datawizard*(v.0.8.0), *fansi*(v.1.0.4), *xtable*(v.1.8-4), *colorspace*(v.2.1-0), *emmeans*(v.1.8.8), *MASS*(v.7.3-60), *tinytex*(v.0.46), *insight*(v.0.19.3), *cli*(v.3.6.1), *mvtnorm*(v.1.2-3), *rmarkdown*(v.2.24), *generics*(v.0.1.3), *rstudioapi*(v.0.15.0), *httr*(v.1.4.7), *reshape2*(v.1.4.4), *tzdb*(v.0.4.0), *parameters*(v.0.21.1), *minqa*(v.1.2.5), *splines*(v.4.3.1), *rvest*(v.1.0.3), *parallel*(v.4.3.1), *vctrs*(v.0.6.3), *boot*(v.1.3-28.1), *webshot*(v.0.5.5), *Matrix*(v.1.6-1), *sandwich*(v.3.0-2), *carData*(v.3.0-5), *car*(v.3.1-2), *hms*(v.1.1.3), *systemfonts*(v.1.0.4), *glue*(v.1.6.2), *nloptr*(v.2.0.3), *codetools*(v.0.2-19), *stringi*(v.1.7.12), *gtable*(v.0.3.4), *lme4*(v.1.1-34), *lmerTest*(v.3.1-3), *munsell*(v.0.5.0), *ore*(v.1.7.3.1), *pillar*(v.1.9.0), *htmltools*(v.0.5.6), *R6*(v.2.5.1), *evaluate*(v.0.21), *lattice*(v.0.21-8), *highr*(v.0.10), *Rcpp*(v.1.0.11), *svglite*(v.2.1.1), *coda*(v.0.19-4), *nlme*(v.3.1-163), *xfun*(v.0.40), *zoo*(v.1.8-12) and *pkgconfig*(v.2.0.3)