FEIR 80: Potencia Estadística

Apuntes del curso FEIR3, curso 2014/15 actualizados. `Última actualización:` jueves 04 abril 2019, 09:29:58

Antonio Maurandi López

Índice

1. 1. Potencia estadística.

Existe una pregunta que persigue a cualquiera que haga análisis estadísticos, *"¿cuánta muestra necesito?"*, o *"¿merece la pena hacer tal estudio con estos n individuos que tengo?"*. Para responder estas preguntas necesitamos hacer "análisis de potencia" y este debería de ser indispensable en cualquier diseño experimental.

El análisis de potencia nos va a permitir:

- Determinar el tamaño muestral para descubrir efectos de un determinado tamaño con un cierto grado de confianza.
- Determinar la probabilidad de encontrar un efecto de un tamaño determinado, con un cierto nivel de confianza, con la muestra de la que actualmente dispones (dado un tamaño muestral).–> *si la probabilidad es muy baja quizás deberíamos dedicarnos a otra cosa en lugar de llevar ese experimento a termino, o deberíamos de aumenta la muestra,...*

El análisis de potencia hay que entenderlo desde el punto de vista del contraste de hipótesis, así que repasemos algunos conceptos.

1.1. Test de hipótesis. (null hypothesis significance testing).

En el test de hipótesis partimos de una hipótesis sobre un parámetro de una población, a lo que llamamos hipótesis **nula** (H_0). Entonces obtenemos una muestra de esa población, y con esa muestra construimos un *"estadístico"* (un *"estimador"*) que empleamos para hacer inferencias sobre ese parámetro poblacional (desconocido).

Ahora, asumiendo que la hipótesis nula es cierta, calculamos la probabilidad de observar el estadístico que calculamos con la muestra u otro aún mayor (diferencias como esas o mayores)

Un ejemplo: *Imaginamos que tenemos unas mediciones de tiempo que representan el tiempo que tarda un adolescente en responder a un estimulo visual con o sin usar unos cascos de música en los que suena a toda caña el "highway to hell" de los ACDC.*

La hipótesis nula será:

$$H_0 : \mu_1 = \mu_2 \qquad (\mu_1 - \mu_2 = 0)$$

(μ_1 es la media de los tiempos de respuesta de los adolescentes con cascos (tratamiento) y μ_2 es la media de respuesta en segundos de los adolescentes sin oír a los ACDC).

Si rechazamos la Hipótesis nula, nos quedamos ('aceptamos') con la Hipótesis alternativa (H_1 o H_a), de que los dos tiempos no son iguales,

$$H_a : \mu_1 \neq \mu_2$$

Tomamos dos muestras , una de cada condición experimental: jóvenes alienados con las música de ACDC y jóvenes sin oír musica. Se les plantean los estímulos visuales y se les cronometra. Basándonos en estas dos muertas calculamos el estadístico:

$$(\overline{X_1} - \overline{X_2}) \,/\, \left(\frac{S}{\sqrt{n}}\right)$$

(S es la desviación estándar de las dos muestras, y n es el número de participantes en cada experimento, lo suponemos de tamaños muestrales iguales).

Si la hipótesis nula es cierta y podemos suponer que los tiempos de reacción se distribuyen de forma normal, el estadístico que hemos construido sigue una distribución t con $2n - 2$ grados de libertad. Con esto podemos calcular la probabilidad de que el estadístico tome ese valor o valores superiores. Si la probabilidad es pequeña , más que un punto de corte que fijemos, (por ejemplo menos que 0.05)

		Decisión	
		Rechazamos H_o	**Aceptamos H_o**
Situación real	**H_o Cierto**	Error de tipo I	Correcto
	H_o Falso	Correcto	Error de tipo II

Figura 1: tabla-07-10

Si podemos suponer que que la Hipótesis nula es falsa, rechazarla y aceptar la alternativa. A este punto de corte, (usualmente 0.05) se le conoce como *"nivel de significación"* (o "significancia").

Usamos una muestra para hacer una inferencia sobre una población, las cuatro posibilidades que se nos pueden plantear son :

- La Hip nula es **falsa** y nuestro contraste nos invita a rechazarla: hemos llegado entonces a una **conclusión correcta**.
- La Hip nula es cierta y nuestro contraste **NO** nos invita a rechazarla: hemos llegado entonces a una **conclusión correcta**.
- La Hip nula es cierta y nuestro contraste nos invita a rechazarla: hemos llegado entonces a una **conclusión incorrecta**. Se dice que cometemos error de tipo I o α). (decimos que hay una diferencia cuando no la hay, y perjudicamos a los vendedores de cascos y a la productora de ACDC).
- La Hip nula es **falsa** y nuestro contraste NO nos invita a rechazarla: hemos llegado entonces a una **conclusión incorrecta**. Se dice que cometemos error de tipo II o β. (Decimos que hay NO hay una diferencia cuando si la hay, y perjudicamos a los jóvenes que piensan que reacciona de igual modo).

Así cuando planificamos un experimento hay que prestar atención a cuatro cosas principalmente: Tamaño muestral, nivel de significación, potencia y tamaño del efecto.

- Tamaño muestral: Número de observaciones en cada condición/grupo (condición experimental)del experimento.
- Nivel de significación (α): Probabilidad de cometer error de tipo I. (Cometemos **error de tipo I o** α, cuando '*afirmamos*' que si hay diferencias (p<0.05) y en verdad no las hay).
- Potencia (power): se define como 1 menos la probabilidad de cometer error de tipo II. Es "la probabilidad de encontrar un efecto que realmente existe". (Potencia estadística de un test a la capacidad de un test para revelar diferencias que realmente existen). Cometemos **error de tipo II o** β , cuando 'afirmamos' que no hay diferencias (p>0.05) y en verdad las hay.
- Tamaño del efecto: es la magnitud del efecto bajo la hipótesi alternativa.

El tamaño muestral y la significación están bajo el control del investigador, la potencia y los tamaños de los efecto son controladas de forma más indirecta. *Nosotros quisiéramos maximizar la probabilidad de encontrar un efecto real, y minimizar la probabilidad de encontrar un efecto que no es real, y a la misma vez mantener la muestra en un tamaño "razonable"*. Las cuatro magnitudes de las que estamos hablando están íntimamente relacionadas, cumpliendo se que , **fijadas 3 la cuarta puede ser determinada**.

1.1.0.1. El paquete pwr.

Vamos a emplear el paquete `pwr`(`???`), que tiene muchas funciones, empelamos solo algunas de ellas. En estas podemos especificar tres de las cantidades $(n, \alpha, 1-p, ES)$ y determinar la cuarta.

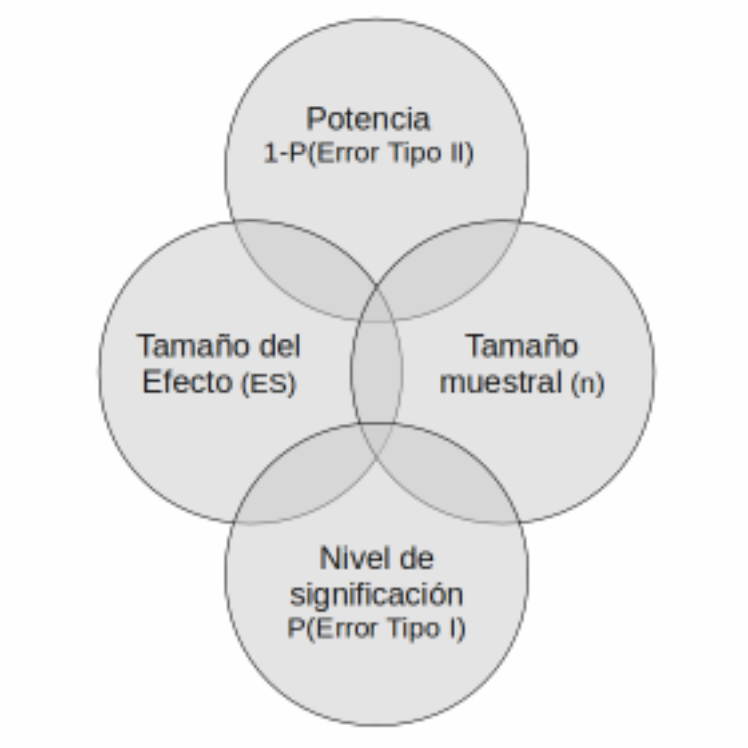


Figura 2: pics-07-20

Generalmente el tamaño del efecto es lo mas difícil de especificar, suele requerir conocimiento del tipo de datos, "experiencia" en investigaciones anteriores, etc... Hablaremos al final de este tema, para centrarnos en las funciones del paquete `pwr`.

Comenzamos instalando y cargando el paquete `pwr`.

```
#install.packages("pwr")
library(pwr)
```

1.2. 1.1. t-test.

Emplearemos la función `pwr.t.test()`:

```
pwr.t.test(n = ,d = ,sig.level = ,power = ,type = ,alternative=  )
* type = c("two.sample", "one.sample", "paired") * alternative = c("two.sided","less","greater")
```

Para `sig.level` '0.05' es el valor por defecto. `d` es el tamaño del efecto, y lo definimos como "la diferencia de medias estandarizada".

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

donde $\sigma^2 = varianza\ común$

Ejemplo. Retomamos el ejemplo de los adolescentes que oían ACDC a toda caña. Tenemos un experimento '`two-tailed`', y un test t de datos independientes.

- Tamaño del efecto: Imaginemos que sabemos que , la el tiempo de reacción de los jóvenes tiene una desviación estándar de 1,25 seg. Y también vamos a suponer que una diferencia de 1 seg es ya diferencia suficiente, importante. Con esto d=1/1,25=0,8 o mayor.

Función	Descripción
`pwr.2p.test()`	2 proporciones(n iguales)
`pwr.2p2n.test()`	2 proporciones(n desiguales)
`pwr.anova.test()`	ANOVA de una via, balanceada
`pwr.chisq.test()`	Chi-cuadrado
`pwr.f2.test()`	Modelo Lineal General (MLG)
`pwr.p.test()`	Proporciones (una muestra)
`pwr.r.test()`	Correlación
`pwr.t.test()`	t-tests (una y dos muestraa, datos apareados)
`pwr.t2n.test()`	t-test (dos muestras con n desiguales)

Figura 3: tabla-07-30

- Potencia. Queremos estar al 90 % seguros de que encontramos el efecto si este existe.
- Significancia: queremos tb, no cometer de un 5 % de error de tipo I, es decir estar al 95 % seguros de que los efectos que encontremos son reales. ¿Cuantos sujetos necesitamos para nuestro estudio?

```
pwr.t.test(d=.8, sig.level=.05, power=.9, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 33.8
##               d = 0.8
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Ejemplo. Cambiamos la pregunta. Supongamos que ahora queremos detectar un '0.5 de diferencias en desviaciones estándares entre las dos poblaciones' (tamaño del efecto). Queremos ampliar la seguridad de no cometer error de tipo I a 1/100=0,01.
Además supongamos que sólo podemos plantearnos el incluir 40 individuos en total (20 en cada grupo). ¿Cuál es la probabilidad de dar con un efecto significativo que sea real?

$d = 0{,}05 = 0{,}625/1{,}25 \qquad (\mu_1 - \mu_2 = 0{,}625)$

```
pwr.t.test(n=20, d=.5, sig.level=.01, type="two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 20
##               d = 0.5
##       sig.level = 0.01
##           power = 0.144
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Así que tendremos solo un 14 % de probabilidad de encontrar diferencias de medias de 0.625 o menores. Así que tendrás un 86 % de probabilidad de no encontrar un efecto real. (tienes una potencia paupérrima)

Hemos estado suponiendo que los tamaños muestrales son iguales. Si los tamaños los conocemos y son diferentes:

```
pwr.t2n.test(n1 =, n2= , d = , sig.level = 0.05, power = ,  alternative = ) * alternative
= c("two.sided",   "less","greater"))
```

```
pwr.t2n.test(n1=20,n2=30, d=.5, sig.level=.01,  alternative="two.sided")
```

```
##
##      t test power calculation
##
##              n1 = 20
##              n2 = 30
##               d = 0.5
##       sig.level = 0.01
##           power = 0.183
##     alternative = two.sided
```

Aumenta la potencia a 18 %.

1.3. 1.2. ANOVA.

Para ANOVA emplearemos la función `pwr.anova.test()`. `pwr.anova.test(k = NULL, n = NULL, f = NULL, sig.level = 0.05, power = NULL)`
Donde
k= número de grupos
n=es el tamaño de muestra común en cada grupo (la menor de las n's.)
f= tamaño del efecto para una ANOVA

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i \times (\mu_i - \mu)}{\sigma^2}}$$

donde
$p_i = \frac{n_i}{N_i}$
$n_i = número\ de\ observaciones\ en\ el\ grupo\ i$
$N = número\ total\ de\ observaciones$
$\mu_i = media\ del\ grupo\ i$
$\mu = media\ total\ (grand\ mean)$
$\sigma^2 = varianza\ entre\ grupos$

Ejemplo: Queremos comparar 5 grupos, y necesitamos saber el tamaño muestral necesario para cada grupo para tener un 80 % de potencia estadística, cuando el tamaño del efecto buscado es de 0.25 y el nivel de significaron permanezca en 0.05

```
pwr.anova.test(k=5, f=.25, sig.level=.05, power=.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 5
##               n = 39.2
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Necesitamos pues 39 individuos en cada grupo, es decir 39+5=195 individuos en total.

Ejemplo: Calcula el tamaño muestral si "buscáramos" efectos más grandes f=0.5.

1.4. 1.3. Correlación.

Para calcular la potencia en correlaciones, empleamos la función `pwr.r.test()`.
`pwr.r.test(n = NULL,r = NULL,sig.level = 0.05,power = NULL, alternative= )`
* `alternative = c("two.sided", "less","greater"))`

R es el tamaño del efecto (coeficiente de correlación)

Ejemplo: Estamos estudiando la relación entre dos variables "dinero que uno gana" y "número de amigos/contactos en Google +". Nuestra hipótesis es que se la correlación entre ambas variables es de 0.26 o mayor.

$$H_0 : \rho <= 0{,}25$$

$$H_a : \rho > 0{,}25$$

Entendemos que ρ es el coeficiente de correlación poblacional. Queremos un a potencia del 90 %, una significación del 0.05.. ¿Cuántas observaciones necesitamos para corroborar la hipótesis?

```
pwr.r.test(r=.25, sig.level=.05, power=.90, alternative="greater")
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 133
##               r = 0.25
##       sig.level = 0.05
##           power = 0.9
##     alternative = greater
```

Necesitamos 134 sujetos.

1.5. 1.4. Modelos lineales.

En modelos lineales, como la regresión lineal múltiple emplearemos la función `pwr.f2.test()`.
`pwr.f2.test(u = NULL ,v = NULL ,f2 = NULL ,sig.level = 0.05 ,power = NULL)`

- $f^2 = \dfrac{R^2}{1-R^2}$
  donde $R = coeficiente\ de\ determinación\ poblacional.$
- $f^2 = \dfrac{R^2_{AB} - R^2_A}{1-R^2_{AB}}$
  donde :
  $R^2_A = varianzaexplicitadaenlapoblaciónporelconjuntodevariablesA.$
  $R^2_{AB} = varianzaexplicitadaenlapoblaciónporelconjuntodevariablesAyBalavez.$
- u, v : *Grados de libertad del numerador y denominador.*

Emplearemos la primera formula cuando evaluamos el impacto de un conjunto de predictores en una variable dependiente.
La segunda formula es apropiada cuando evaluamos el impacto de un cto de predictores sobre otro conjunto de predictores.

Ejemplo: *Imaginemos que estamos interesados en saber cuando el estilo de mando ('estilo') de un jefe afecta la satisfacción laboral ('satisfacción'), más allá del 'sueldo' y el 'carácter' del empleado. (supongamos que todo esto lo miden los psicólogos es encalas cuantitativas).*

Sabemos que el "estilo""del jefe está relacionado con 4 variables y "suelo" y "carácter" con tres.

Supongamos que de estudios previos sabemos que "salario" y "carácter" explican el 30 % de la varianza en "satisfacción" laboral.

Pensamos que sería suficiente que "estilo" explicará al menos un 5 % de "satisfacción". Suponiendo una potencia del 90 %¿cuántos individuos se necesitarían para asegurar esa contribución con un 90 % de confianza?
$u = 3$, número de predictores, menos el número de predicadores en el conjunto B
$f = \dfrac{0{,}35 - 0{,}30}{1 - 0{,}35} = 0{,}00769$

```
pwr.f2.test(u=3, f2=0.0769, sig.level=0.05, power=0.90)
```

```
##
##      Multiple regression power calculation
##
##               u = 3
##               v = 184
```

```
##              f2 = 0.0769
##       sig.level = 0.05
##           power = 0.9
```

Los grados de libertad en MLG se calculan así N-k-1 (, N =numero total de obs, k=número de variables indep), N-7.1=185, así que la muestra requerida es N= 185+7+1= 193.

1.6. 1.5. Test de proporciones.

Emplearemos la función `pwr.2p.test()`.
`pwr.2p.test(h = NULL ,n = NULL ,sig.level = 0.05 ,power = NULL, alternative= )`
* `alternative = c("two.sided","less","greater")`
* h es el tamaño del efecto y n el tamaño muestral común.

$$h = 2arcsin\left(\sqrt{p_1}\right) - 2arcsin\left(\sqrt{p_2}\right)$$

Podemos emplear la función `ES.h(p1,p2)` para calcularlo.

Si los tamaños son diferentes emplearemos la función `pwr.2p2n.test()`.
`pwr.2p2n.test(h= ,n1= ,n2= ,sig.level=0.05 ,power= , alternative= )`

Ejemplo: *Sabemos que un tratamiento es efectivo en un 60 % de la población. Un nuevo tratamiento, nuevo y mucho más caro será autorizado si es efectivo en un 65 % de la población. ¿Cuantos sujetos necesitamos si queremos hacer un estudio comparando ambos medicamentos?*

```
pwr.2p.test(h=ES.h(.65, .6), sig.level=.05, power=.9,alternative="greater")
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##              h = 0.103
##              n = 1604
##       sig.level = 0.05
##           power = 0.9
##     alternative = greater
##
## NOTE: same sample sizes
```

Necesitamos 1604 individuos en cada condición experimental, es decir, ¡¡3210 sujetos!!.

Ejemplo: *Si la mejora del tratamiento 'asegura la farmacéutica que es de un 20 %' que individuos necesitamos para contrastarla?*

```
pwr.2p.test(h=ES.h(.80, .6), sig.level=.05, power=.9,alternative="greater")
```

Solución: tan sólo 88 individuos por grupo (N=176)

1.7. 1.6. Test χ^2

Emplearemos la función `pwr.chisq.test()`.
`pwr.chisq.test(w = NULL ,N = NULL ,df = NULL ,sig.level = 0.05 ,power = NULL)`
Donde w es el tamaño del efecto.

$$w = \sqrt{\sum_{i=1}^{m} \frac{(p0_i - p1_i)^2}{p0_i}}$$

G Étnico	Promociona	No promociona
Caucásico	0.42	0.28
Negro	0.03	0.07
Hispano	0.10	0.10

Figura 4: tabla-07-40

donde
$p0_i = probabilidad\ de\ la\ celda\ i-esima\ bajo\ la\ H_0$
$p1_i = probabilidad\ de\ la\ celda\ i-esima\ bajo\ la\ H_1$
$m = número\ de\ celdas\ de\ la\ tabla\ de\ contingencia.$

Ejemplo: Queremos ver la relación entre "grupo étnico" y promoción.

Esperamos que el 42 % de la población sea caucasicos que promocionan $(0{,}42 = 0{,}7 \times 0{,}6)$, el 7 % serán negros que no promocionan...
Fijamos $\alpha = 0{,}05$, ella potencia deseada en el 90 % (0,9).

Los grados de libertadad en una tabla de doble entrada es $(r-1) \times (c-1)$, donde r representa el número de niveles de la primera variable y c el de la segunda, $gl = (r-1) \times (c-1) = (3-1) \times (2-1) = 2 \times 1 = 2$
Calculamos el tamaño del efecto

```
P <- matrix(c(.42, .28, .03, .07, .10, .10), byrow=TRUE, nrow=3)
ES.w2(P)
```

```
## [1] 0.185
```

Calculamos ahora el tamaño muestral que necesitamos:

```
pwr.chisq.test(w=.1853, df=2, sig.level=.05, power=.9)
```

```
##
##      Chi squared power calculation
##
##               w = 0.185
##               N = 369
##              df = 2
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: N is the number of observations
```

Así que necesitamos 368 individuos para ddetectar una relación entre estas dos variables.

1.8. 1.7. Tamaños de los efectos.

Como ya hemos dicho lo más dificil suele ser determinar el tamaño del efecto. Suele requerir experiencia previa, haber realizado o estar en posesión de estudios previos en los que se tenga cierta seguridad de que repersantan los parametros que necesitamos. Pero ¿qué hacemos cuando nos enfrentamos a una situación nueva?

Test	Medida del tamaño del efecto	Valores sugeridos por Cohen(1988)		
		Pequeño	mediano	Grande
t-test	d	0.20	0.50	0.80
ANOVA	f	0.10	0.25	0.40
Modelos lineales	f2	0.02	0.15	0.35
Proporciones	h	0.20	0.50	0.80
Chi-2	w	0.10	0.30	0.50

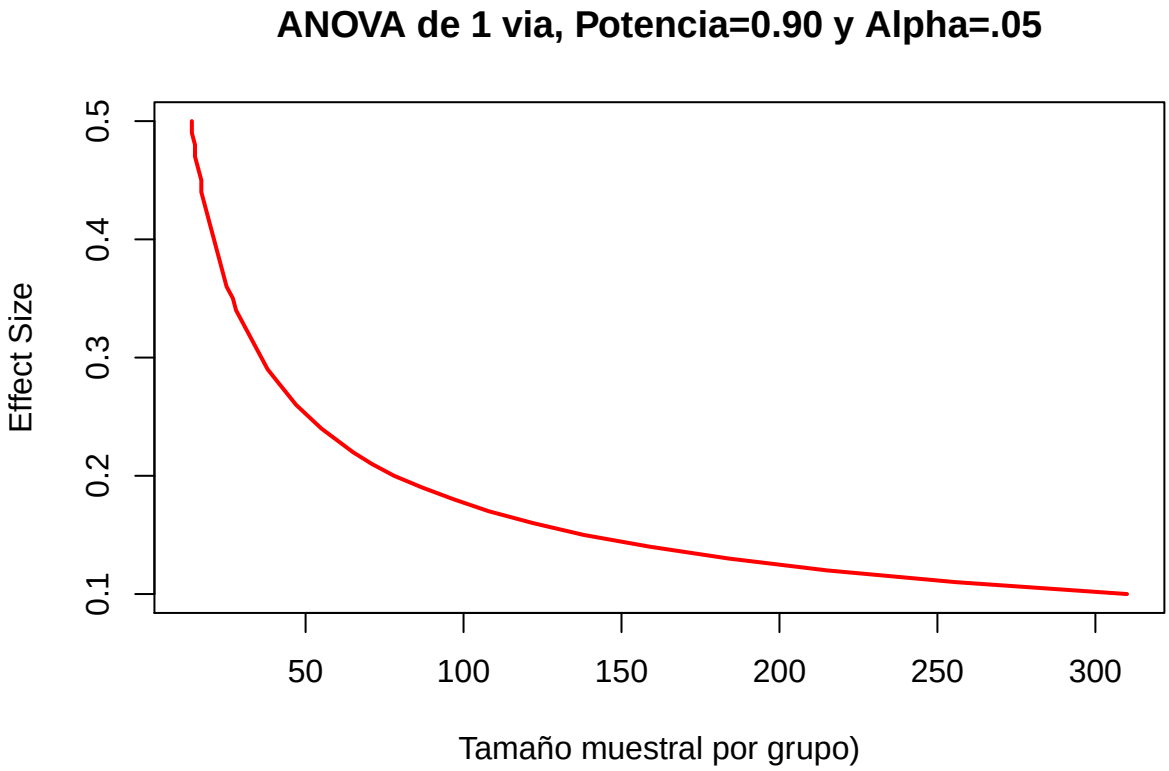
Figura 5: tabla-07-50

Para estudios en ciencias sociales Cohen(1988) porpuso unos limites para considerar efectos de diferentes tamaños: pequeños, medios y grandes.

Esta guia puede qu eno se adapte a un campo de estudio concreto. Lo que podemos hacer cuando no tenemos "ni idea" es variar los parámetros y observar que impacto tiene sobre los valores de potencia, tamaño de muetsras etc...

Podemos crera gráficos de potencia para analizar diferentes tamaños muestrales y relacionalos con el tamaño del efecto, por ejemplo para una anova:

```
es <- seq(.1, .5, .01)
nes <- length(es)
samsize <- NULL
for (i in 1:nes){
  result <- pwr.anova.test(k=5, f=es[i], sig.level=.05, power=.9)
  samsize[i] <- ceiling(result$n)
}
plot(samsize,es, type="l", lwd=2, col="red",
ylab="Effect Size",
xlab="Tamaño muestral por grupo)",main="ANOVA de 1 via, Potencia=0.90 y Alpha=.05")
```



Volver al índice del curso

Referencias y bibliografía

Kabacoff, R. (2011). *R in action* (1st ed.). Shelter Island, NY: Manning Publications.