Planificación del tamaño de la muestra para las evaluaciones de impactos

David Evans, Banco Mundial

Basada en transparencias de Esther Duflo (J-PAL) y Jed Friedman (Banco Mundial)

REGIONAL IMPACT EVALUATION WORKSHOP
Evaluating the Impact of Development Programs: Turning Promises into Evidence

Lima, Enero 2009

El tamaño de la muestra para las evaluaciones de impactos

- Pregunta general
  ¿De qué tamaño tiene que ser la muestra para *detectar un impacto de cierto tamaño*?

- ¿Qué quiere decir «detectar» aqui?
  La diferencia es debida al programa (y no al ruido)

- La aleatorización quita los sesgos pero no quita el ruido: Funciona por la ley de los grandes números…
  ¿Qué tan grande tiene que ser la muestra?

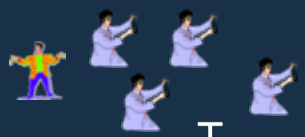
¿Qué tan grande?

- ¿2 personas seleccionadas de una forma aleatoria?


T


C

- ¿10 personas?


T


C

- ¡Muchas personas! ¿Cuántas son muchas?


T


C

Organización básica

- Al final del experimento, comparamos el resultado de interés en los grupos de tratamiento y de control
- Nos interesa la diferencia:

$$\frac{\text{Promedio del grupo de tratamiento} - \text{Promedio del grupo control}}{\text{Tamaño del efecto}}$$

- Por ejemplo

$$\frac{\text{Ingreso promedio de hogares que reciben CCTs} - \text{Ingreso promedio de hogares que no reciben CCTs}}{\text{Tamaño del efecto}}$$

La estimación

No tenemos suficiente dinero como para observar toda los hogares sino una muestra (ni lo tenemos que hacer).

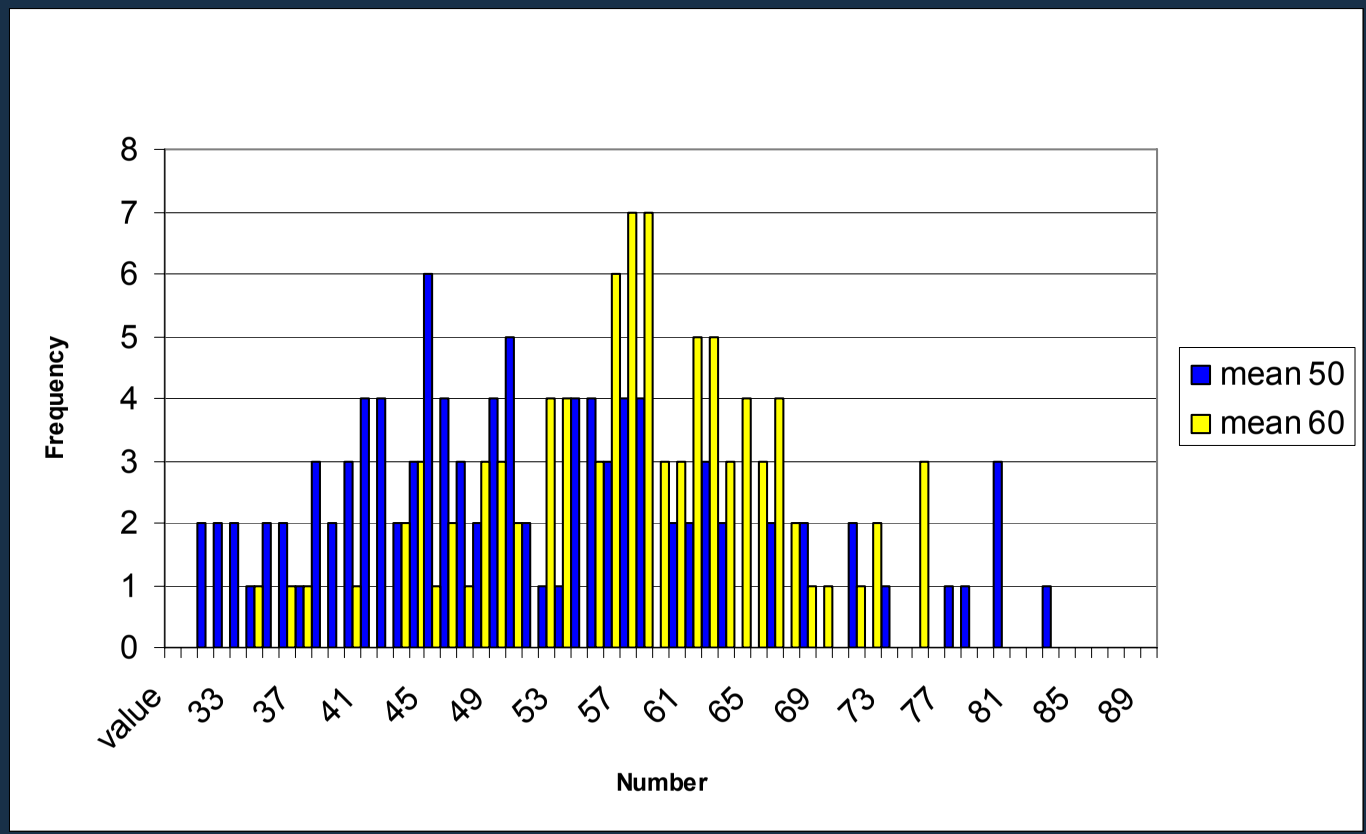
En cada hogar de la muestra, hay cierto nivel de ingreso. Puede estar más cerca o más lejos del promedio de toda la población, como función de los otros factores que afectan el ingreso.

Inferimos el ingreso promedio en la población utilizando el promedio en la muestra.

Si tenemos muy pocos hogares, los promedios estarán imprecisos. Si no vemos diferencias entre el promedio del grupo de tratamiento y de control, no sabemos si no hay efecto o si no hay potencia de detectar el efecto.

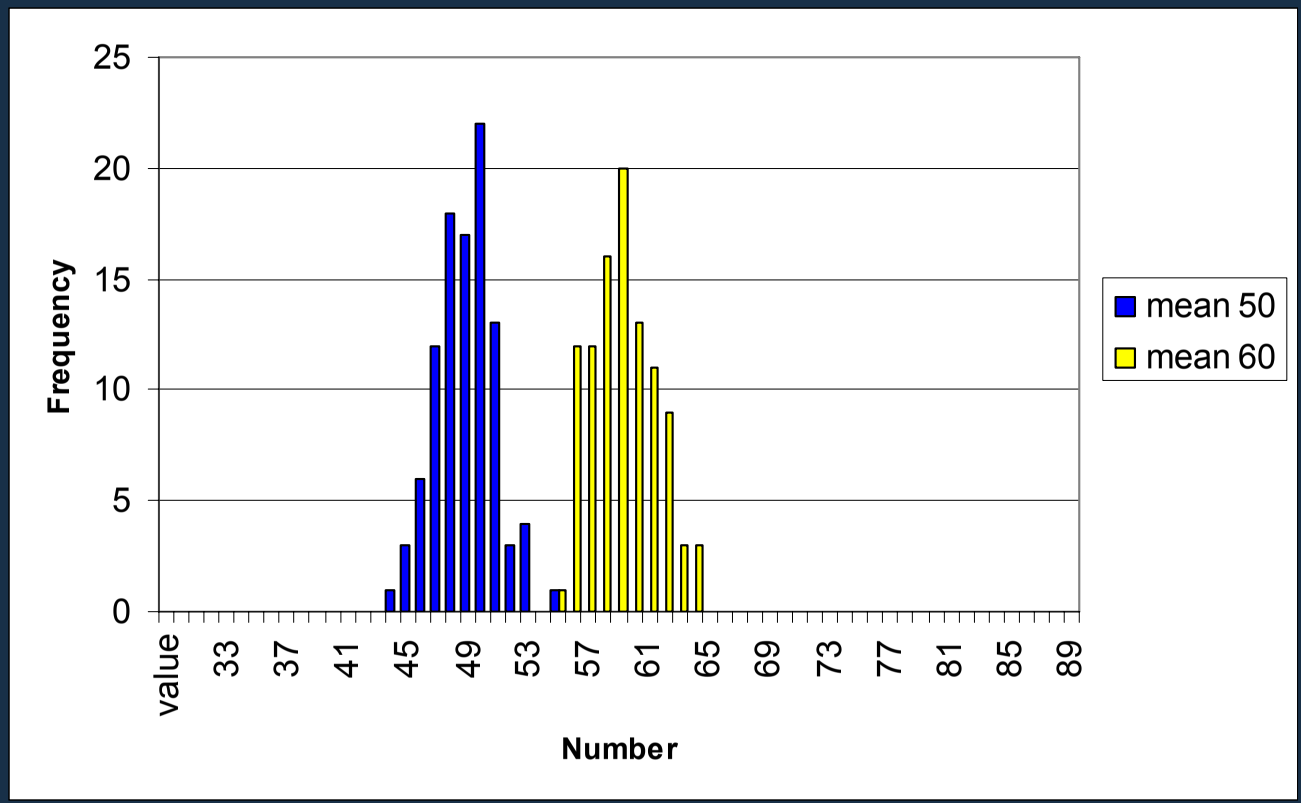
La variabilidad en el resultado que medimos

Si el resultado varía mucho dentro del grupo de tratamiento y de control, será difícil decir si fue el tratamiento que lo cambió



La variabilidad en el resultado que medimos

Si el resultado varía poco dentro de los grupos, es más fácil decir que fue el tratamiento



El error estándar

- El error estándar de la estimación en la muestra capta el tamaño de la muestra y la variabilidad del resultado
  - ↑ con una muestra pequeña
  - ↑ con un resultado muy variable
- Un intervalo de confianza de 95% para un efecto nos dice que, para 95% de las muestras que podríamos sacar de la misma población, el efecto estimado caería en este intérvalo.

$$Intervalo\,de\,confianza = efecto \pm 2\,errores\,estándares$$

Docimasia de Hipótesis

A menudo nos interesa probar el hipótesis que el tamaño del efecto es igual a cero (o sea, ¿mi programa no tiene ningún efecto? ¡esperamos que no!)

Queremos probar:

$$H_o : \text{Efecto} = 0$$

Contra:

$$H_a : \text{Efecto} \neq 0$$

Dos tipos de errores - I

- Primer tipo de error: Concluimos que hay un efecto cuando en verdad no hay efecto.

  El **nivel** de la prueba es la probabilidad que falsamente concluirá que el programa tiene un efecto cuando en verdad no lo tiene.

  Asi que con un nivel de 5%, podemos tener confianza de 95% en la conclusión que el programa tuvo un efecto

  Para la política, queremos tener mucha confianza en el impacto estimado: asi que ponemos el nivel bajo.

  Niveles comunes: 5%, 10%, 1%

La relación con intérvalos de confianza

- Si cero no está en el intérvalo de 95% del tamaño del efecto, podemos tener 95% certeza que el efecto no es cero (asi que hay un efecto).

- Asi que la regla general es que si el tamaño del efecto es más del doble del error estandar, puede concluir con más de 95% certeza que el programa tuvo un efecto.

Dos tipos de errores - II

Segundo tipo de error: no piensas que el programa no tuvo ningún efecto cuando en verdad tuvo efecto.

- La **potencia** de la prueba es la probabilidad de encontrar un efecto en el experimento si en verdad hay un efecto: más potencia es mejor porque es más probable que encuentre un efecto)
- La potencia es una herramienta de planificación. Nos dice la probabilidad de que vayamos a identificar un efecto significante con dado tamaño de muestra y de efecto.

El cálculo de la potencia

- Cuando planea una evaluación, con unas investigaciones preliminarias, podemos calcular la mínima muestra necesaria para:
  - Poner a prueba un hipótesis: el efecto del programa fue cero o no cero
  - Con un nivel especificado de antemano (p ej 5%)
  - Con un tamaño de efecto especificado de antemano (lo que piensas que el programa hará)
  - Para lograr cierta potencia
- Una potencia de 80% nos dice que, en 80% de los experimentos de este tamaño de muestra conducidos con esta población, si de verdad hay un efecto en la población, podremos decir que hay un efect en nuestra muestra con el deseado nivel de confianza.
- Más grande la muestra, más grande la potencia.

Niveles comunes de potencia: 80%, 90%

Ingredientes para el cálculo de potencia en un estudio sencillo

Lo que requerimos	Donde lo encontramos
Nivel de significancia	Esto a menudo se pone a 5% por convención. Más bajo que sea, más grande tiene que estar la muestra.
El promedio y la variabilidad del resultado en el grupo control	- De encuestas anteriores en contextos parecidos - Más grande que sea la variabilidad, más grande la muestra
El tamaño de efecto que queremos percibir	¿Qué es el efecto más chico que motivaría un cambio de política? Más chico el tamaño de efecto que queremos percibir, más grande la muestra que requeremos

Escoger un tamaño de efecto

- ¿Cuál es el efecto más pequeño que justificaría la adopción del programa:
  - Costo de este programa vs los beneficios que trae
  - Costo de este programa vs el uso alternativo del dinero
- Si el efecto es más chico que esto, para nosotros es lo mismo que ser cero: no nos interesa comprobar que un efecto muy pequeñito es distinto que cero
- Por otro lado, cualquier efecto más grande que aquel efecto justificaría adoptar el programa: queremos poder distinguirlo de cero
- Peligro común: escoger un tamaño de efecto que sea demasiado optimístico – el tamaño de muestra puede ser demasiado bajo!

Los tamaños de efecto estandardizado

- El tamaño de efecto que puedes detectar con cierta muestra depende de que tan variables son los resultados.
  - Ejemplo: Si todos los niños tienen niveles de aprendizaje muy parecidos sin el programa, un impacto muy pequeño será fácil de detectar
- La deviación estándar capta la variabilidad del resultado. Más variabilidad → una deviación estándar más alta
- El tamaño de efecto estandardizado es el tamaño del efecto dividido por la deviación estándar del resultado
  - **d** = tamaño del efecto / deviación estándar
- Tamaños comúnes de efectos:

  d=0,20 (chico), d =0,40 (mediano), d =0,50 (grande)

Los factores del diseño que influyen la potencia

- El nivel de la aleatorización
- La disponilidad de una línea de base (encuesta inicial)
- La disponilidad de variables control y de estratificación.
- El tipo de hipótesis que se quiere poner a prueba

El nivel de aleatorización
El diseño «cluster»

Los experimentos aleatorizados «cluster» son experimentos en que ciertas unidaded (o grupos) están asignados a los grupos de tratamiento y control, no los individuos

Ejemplos:

Tranferencias monetarias condicionales (CCTs)	Pueblos
Distribución de mosquiteros	Dispensas de salud
Tratamiento por malaria	Escuelas
Suplementación con hierro	Familia

Razones por adoptar la aleatorización «cluster»

- La necesidad de minimizar o quitar la contaminación
  - Ejemplo: En el program de quitar los parásitos, se escogía las escuelas porque los parásitos son contagiosos
- La viabilidad
  - Ejemplo: El programa PROGRESA no hubiera sido posible políticamente si unas familias pobres participaron y otras no en el mismo pueblo.
- La única posibilidad razonable
  - Ejemplo: Cualquier intervención que afecta una escuela entera, como el entrenamiento de profesores.

El impacto de «clustering»

- Los resultados para todos los individuos de un grupo pueden estar correlacionados
  - Todos de un pueblo reciben la misma lluvia para su maiz
  - Todos los pacientes tienen el mismo médico
  - Todos los estudiantes tienen el mismo director de escuela
  - El programa afecta todos los estudiantes a la vez
  - Los miembos de un pueblo se relacionan de día en día
- El tamaño de la muestra tiene que ser modificado por esta correlación
- Mayor correlación entre los resultados → Más necesidad de ampliar la muestra

Ejemplo del efecto de «clustering»

Número de pueblos, para potencia de 0.80

Correlación Intra-pueblo	Hogares en cada pueblo			
	10	50	100	200
0.00	23	7	5	4
0.02	25	10	8	8
0.05	30	16	15	13
0.10	40	25	23	22

Implicaciones

- Es sumamente importante aleatorizar un número adecuado de grupos
- El número de individuos en los grupos importa menos que el número de grupos
- La ley de los grandes números aplica solo cuando el numero de grupos aleatorizados aumenta
- ¡No se puede aleatorizar al nivel del distrito con un distrito de tratamiento y uno de control! [even with 2,000 hh per district or 10 l 100 vs 100 | 10]

Los factores del diseño que influyen la potencia

- El nivel de la aleatorización
- La disponilidad de una línea de base (encuesta inicial)
- La disponilidad de variables control y de estratificación.
- El tipo de hipótesis que se quiere poner a prueba

Disponibilidad de una línea de base

- Una línea de base:
  - Se puede verificar si los grupos de tratamiento y control eran parecidos o distintos antes del tratamiento
  - Se puede reducir el tamaño de la muestra, pero requiere que haga una encuesta antes de comenzar el tratamiento: el costo de la evaluación ↑ y el del tratamiento ↓
  - Se puede usar para estratificar y formar sub-grupos
- Para computar potencia con una línea de base:
  - Necesitas saber la correlación entre dos medidas subsiguientes del resultado (p ej: el consumo medido en dos años).
  - Más grande la correlación, más se aumenta la potencia.
  - Hay ganancias muy grandes para resultados muy continuos como PMO.

Los factores del diseño que influyen la potencia

- El nivel de la aleatorización
- La disponilidad de una línea de base (encuesta inicial)
- La disponilidad de variables control y de estratificación
- El tipo de hipótesis que se quiere poner a prueba

Variables Control

- Si tenemos más variables pertinentes (p. ej. población del pueblo, el distrito del pueblo), podemos controlar por su efecto
- Lo que importa para la potencia es la variación que queda después de controlar por estas variables
- Si los variables control explican mucha de la variabilidad, la precisión aumenta y la muestra requerida baja.
- Advertencia: las variables control solo pueden incluir las variables no influidas por el tratamento: las que fueron coleccionadas antes del tratamiento.

La muestras estratificadas

- La estratificación: crea BLOQUES según los variables control y aleatoriza el tratamiento entre cada bloque
- La estratificación asegura que los grupos de tratamiento y de control están balanceados con respecto a aquellas variables control.
- Reduce la variabilidad por dos razones:
  - Reduce la variabilidad del resultado de interés en cada estrato
  - La correlación de individuos dentro de grupos.
- Ejemplo: si quiere estratificar por distrito para un programa de agricultura
  - Se controla por condiciones climáticas
  - El efecto común dentro de cada distrito desaparece.

Los factores del diseño que influyen la potencia

- El nivel del aleatorización
- La disponilidad de una línea de base
- La disponilidad de variables control y de estratificación
- El tipo de hipótesis que se quiere poner a prueba

El hipótesis que se pone a prueba

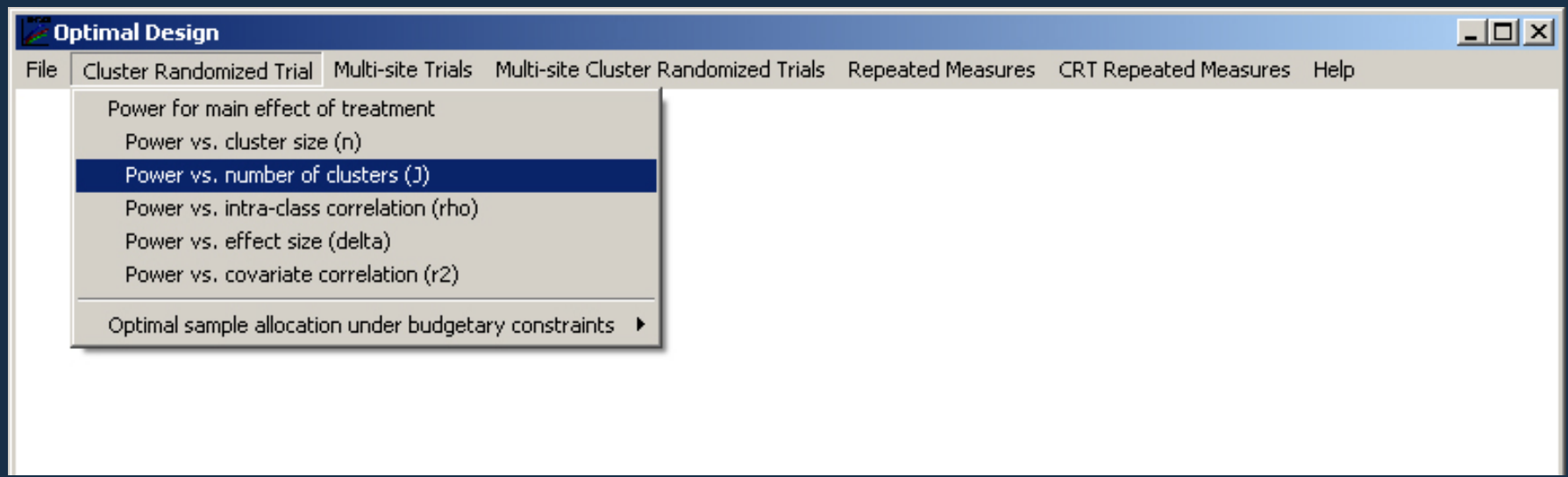
- Le interesa la diferencia entre dos tratamientos además de la diferencia entre tratamiento y control?
- Le interesa la interacción entre los tratamientos?
- Le interesa poner a prueba si el efecto es distinto entre sub-poblaciones distintas?
- El diseño solo incluye conformidad parcial con el tratamiento?

¡Para recordar!

- El número de grupos importa much más que el número de individuos
  - Escuelas vs estudiantes, pueblos vs hogares
- Dos tipos de errores
  - Tipo I: Piensas que hay efecto cuando no lo hay → nivel
  - Tipo II: Piensas que no hay efecto cuando lo hay → potencia
- Evitar los errores requiere una muestra suficiente → el cálculo de la potencia

Los cálculos de potencia usando el software grátis «Optimal Design»

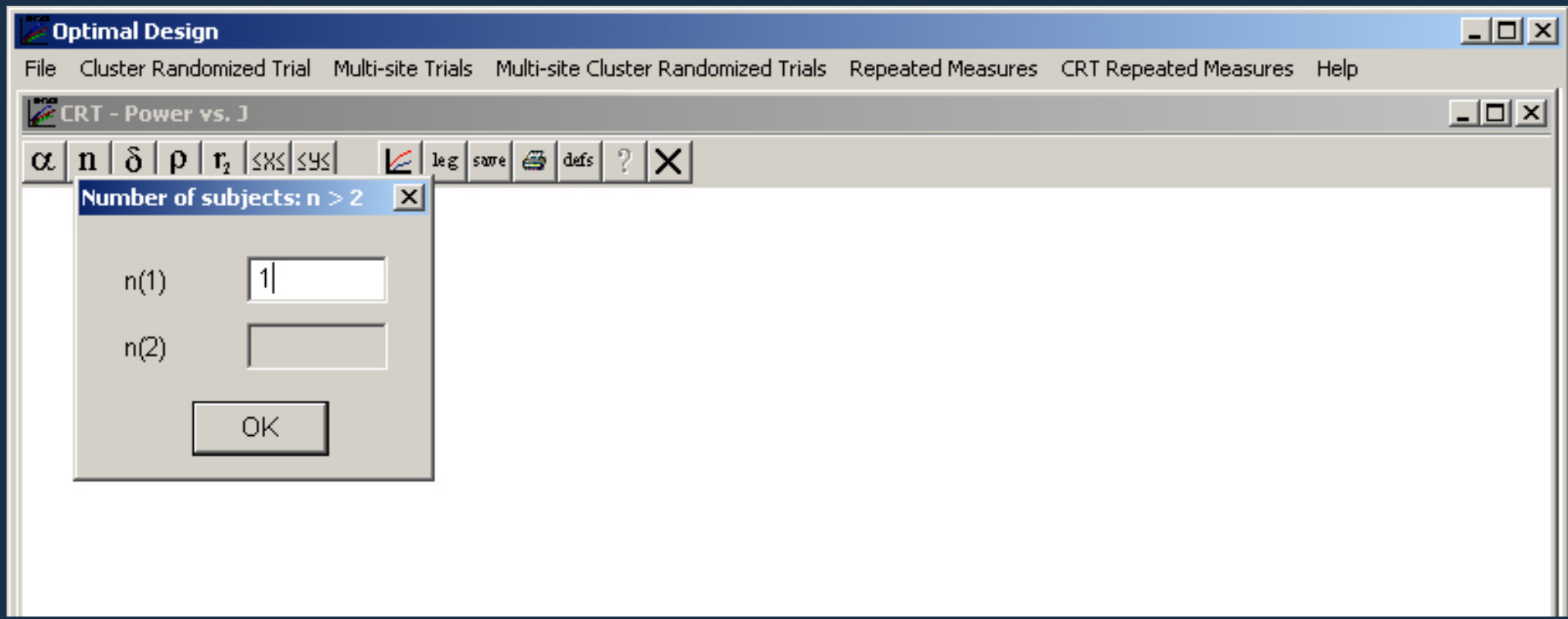
Escoge "Power v. number of clusters" (potencia vs número de grupos) en el menú "clustered randomized trials" (pruebas aleatorizadas y agrupadas)



http://sitemaker.umich.edu/group-based/optimal_design_software

Tamaño de cada grupos

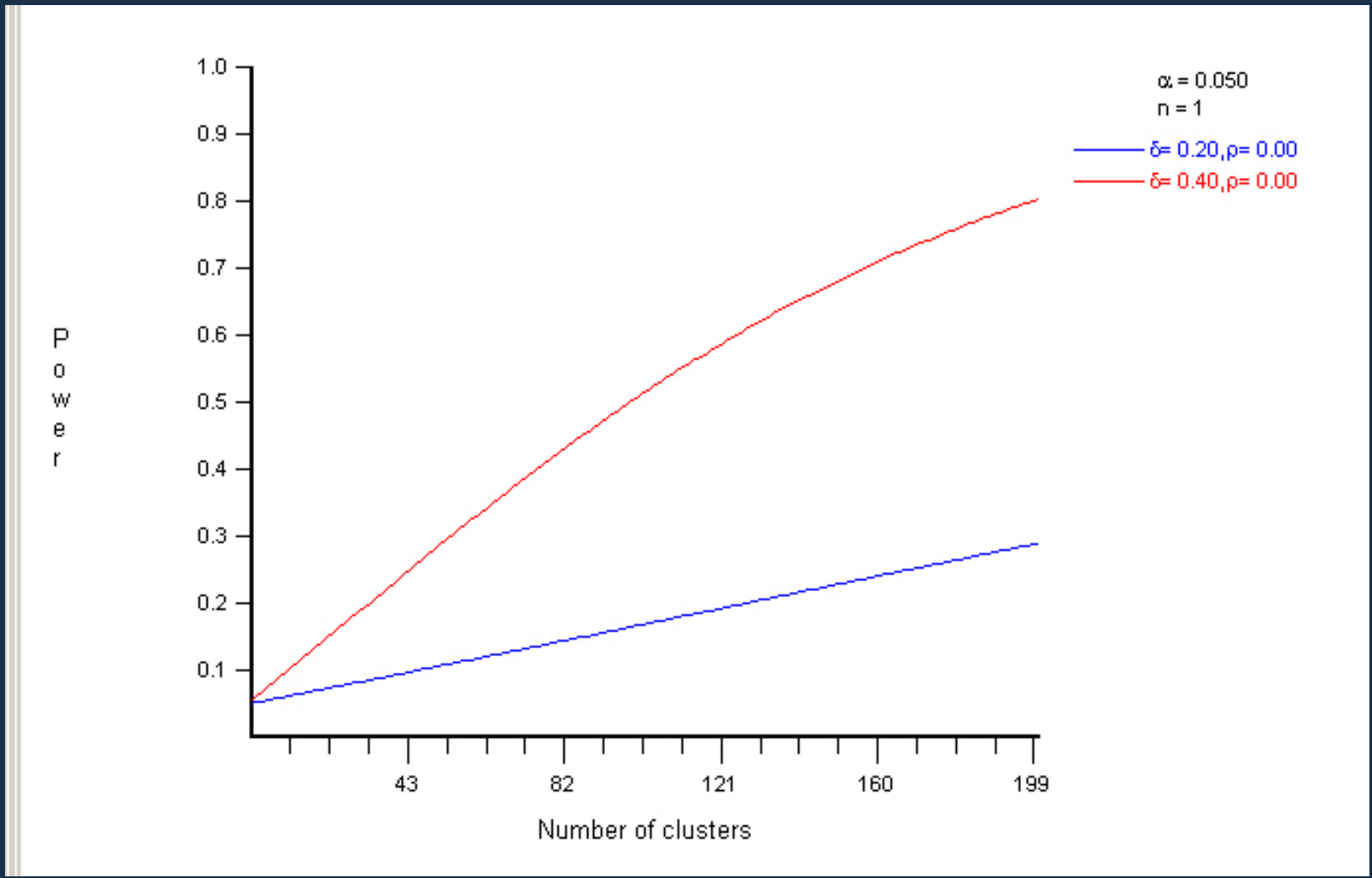
- Escoge "cluster size" (tamaño de grupo)



Escoge el nivel de significancia, el efecto de tratemento, y la correlación

- Escoge *a*: el nivel
  - Normalmente se escoge 0,05
- Escoge *d*:
  - Se puede experimentar con 0,20
- Escoge la correlación intra-clase (*rho*)
- Se obtiene el gráfico siguiente que muestra la potencia como función del tamaño de muestra

La potencia y el tamaño de la muestra



¡Para recordar!

- El número de grupos importa much más que el número de individuos
  - Escuelas vs estudiantes, pueblos vs hogares
- Dos tipos de errores
  - Tipo I: Piensas que hay efecto cuando no lo hay → nivel
  - Tipo II: Piensas que no hay efecto cuando lo hay → potencia
- Evitar los errores requiere una muestra suficiente → el cálculo de la potencia

Conclusiones: El cálculo de potencia en práctica

- Los cálculos de potencia requieren varias conjeturas.
- A veces no tenemos toda la información para hacerlos perfectamente
- Mientras tanto, es importante hacer lo mejor posible:
  - Evitar la iniciación de estudios que no tendrán ninguna potencia estadística: desperdicio de esfuerzo y de dinero
  - Dedica los recursos apropriados a los estudios que decide conducir (pero no demasiados recursos).