

حجم عينة التخطيط لعمليات التقييم العشوائية

جيد فريدمان، البنك الدولي
مأخوذة من عرض إيستر دوفلو، جي-بال

حجم عينة التخطيط لعمليات التقييم العشوائية

■ سؤال عام :

■ كم ينبغي أن يكون حجم العينة لاستكشاف حجم أثر معين بمصدافية؟

■ ماذا نعني "بمصدافية" هنا؟

هذا يعني أنه بإمكانني أن أكون متأكدا بصورة معقولة أن الفرق بين المجموعة التي تلقت البرنامج وتلك التي لم تتلقى البرنامج، سببه البرنامج.

■ التوزيع العشوائي يزيل التحيز، ولكنه لا يزيل الضجة: هو ينجح بسبب قانون الأرقام الكبيرة ... كم ينبغي أن يكون هذا الحجم الكبير؟

Basic set up

التشكيل الأساسي

- في نهاية تجربة ما، سنقارن نتيجة الإهتمام في مجموعات المعالجة ومجموعات المقارنة.
- نحن مهتمون بالفرق :
الوسيط في المعالجة – الوسيط في الرقابة = حجم الأثر
- مثلاً: وسيط عدد شبكات الفراش المعتمدة في قرى فيها توزيع مجاني مقابل وسيط عدد شبكات الفراش المعتمدة في القرى مع استرجاع التكاليف

التقدير

الا أننا لا نراقب كافة السكان، فقط **عينة منهم** .

وفي كل قرية من العينة، هناك عدد معين من شبكات الفراش. وهو نوعا ما قريب من الوسيط الفعلي لإجمالي السكان، كدالة لكل العوامل الأخرى التي تؤثر على عدد شبكات الفراش

ونقدر الوسيط عن طريق احتساب المتوسط في العينة

وإذا كان لدينا عدد قليل جدا من القرى، تكون المتوسطات غير دقيقة. وعندما نرى اختلافا في معدلات العينة، لانعرف ما إذا كانت تتأتى من أثر المعالجة او من أمر آخر.

التقدير

حجم العينة :

- مالذي يمكننا استنتاجه لو كان لدينا قرية خضعت للمعالجة وأخرى لم تخضع؟
- مالذي يمكننا استخلاصه إذا ما أعطينا دواء الملاريا (IPT) لصف واحد دون الآخر

- مع أنه لدينا حجم صف كبير؟
- المهم هو **حجم العينة الفعلي**، أي عدد **وحدات المعالجة ووحدات المراقبة** (مثل الصفوف **الدراسية**). وما هي الوحدة في حالة إعطاء IPT في الصف؟

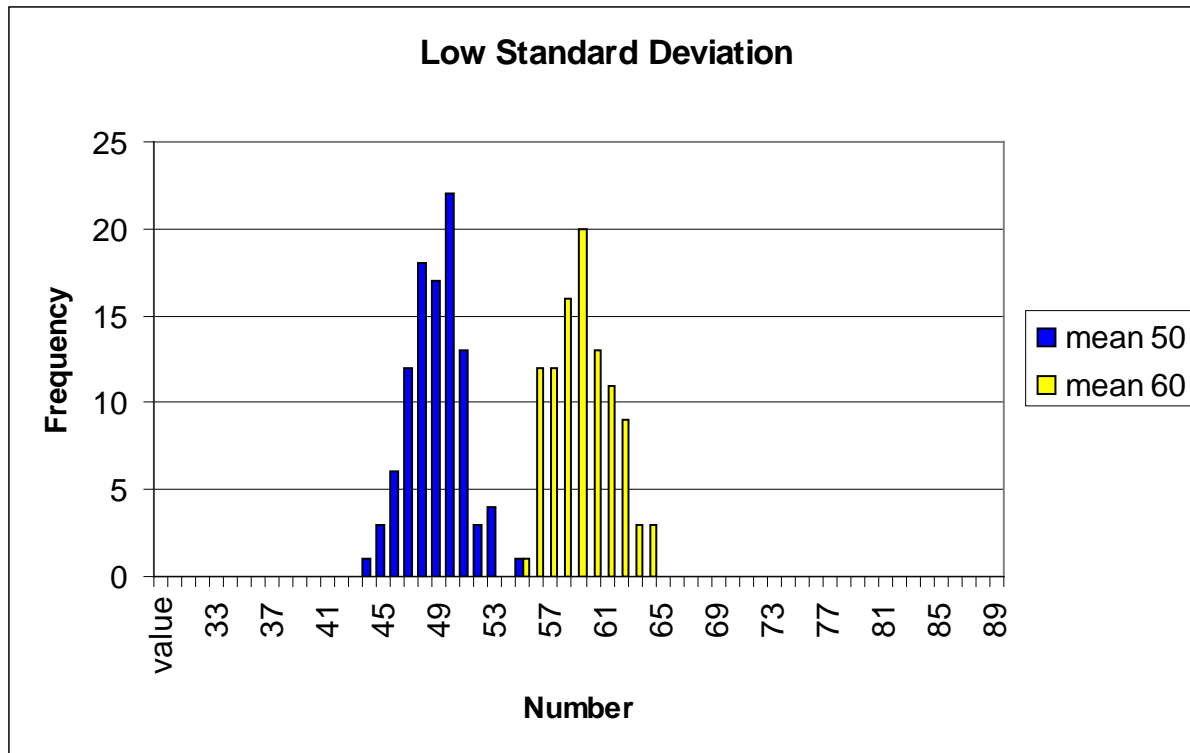
نحاول أن نقيس التنوع في النتيجة :

- إذا كان هناك العديد من الأمور الأخرى غير المقاسة التي تشرح نتائجنا، يكون من الأصعب معرفة ما إذا كانت المعالجة قد غيرته فعلاً.

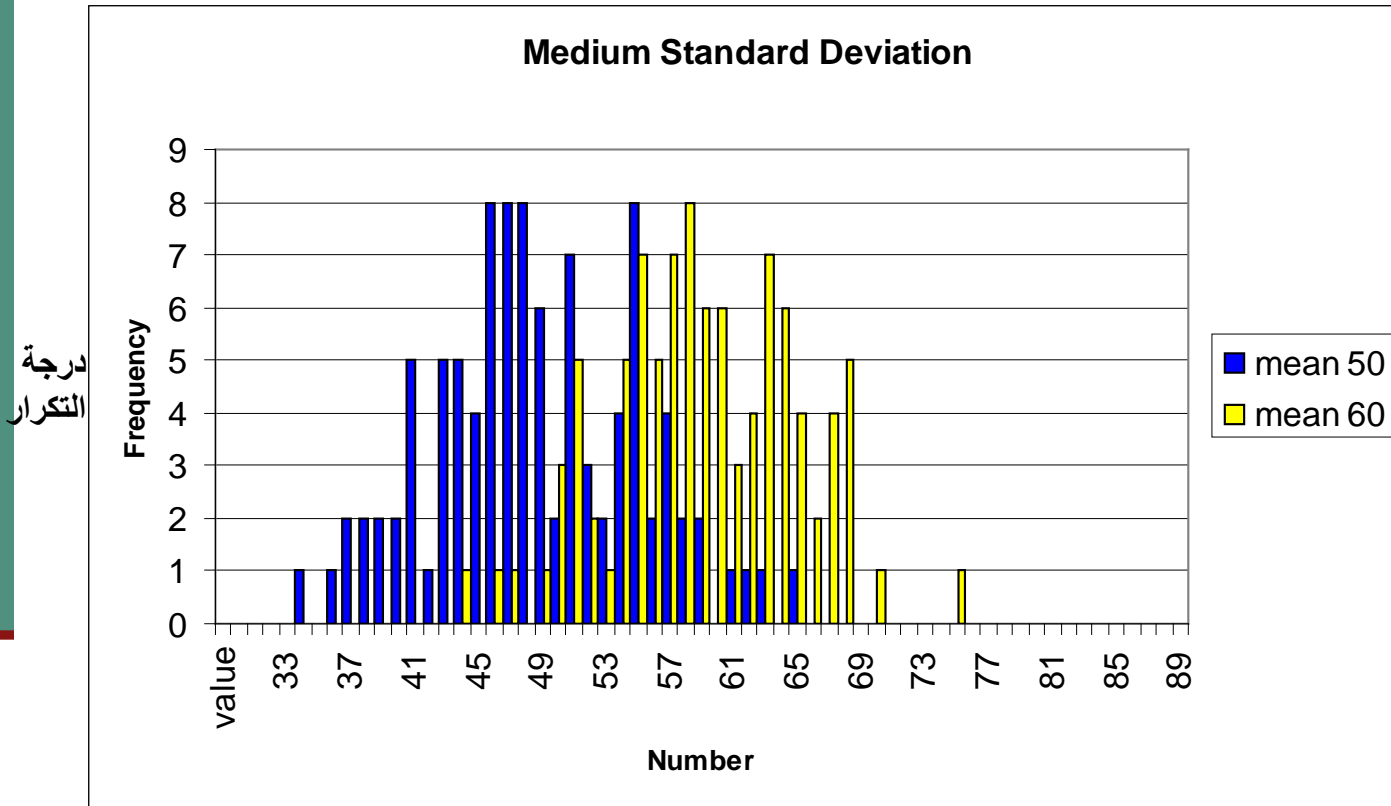
عندما تكون النتائج دقيقة للغاية

انحراف بمعيار منخفض

درجة التكرار

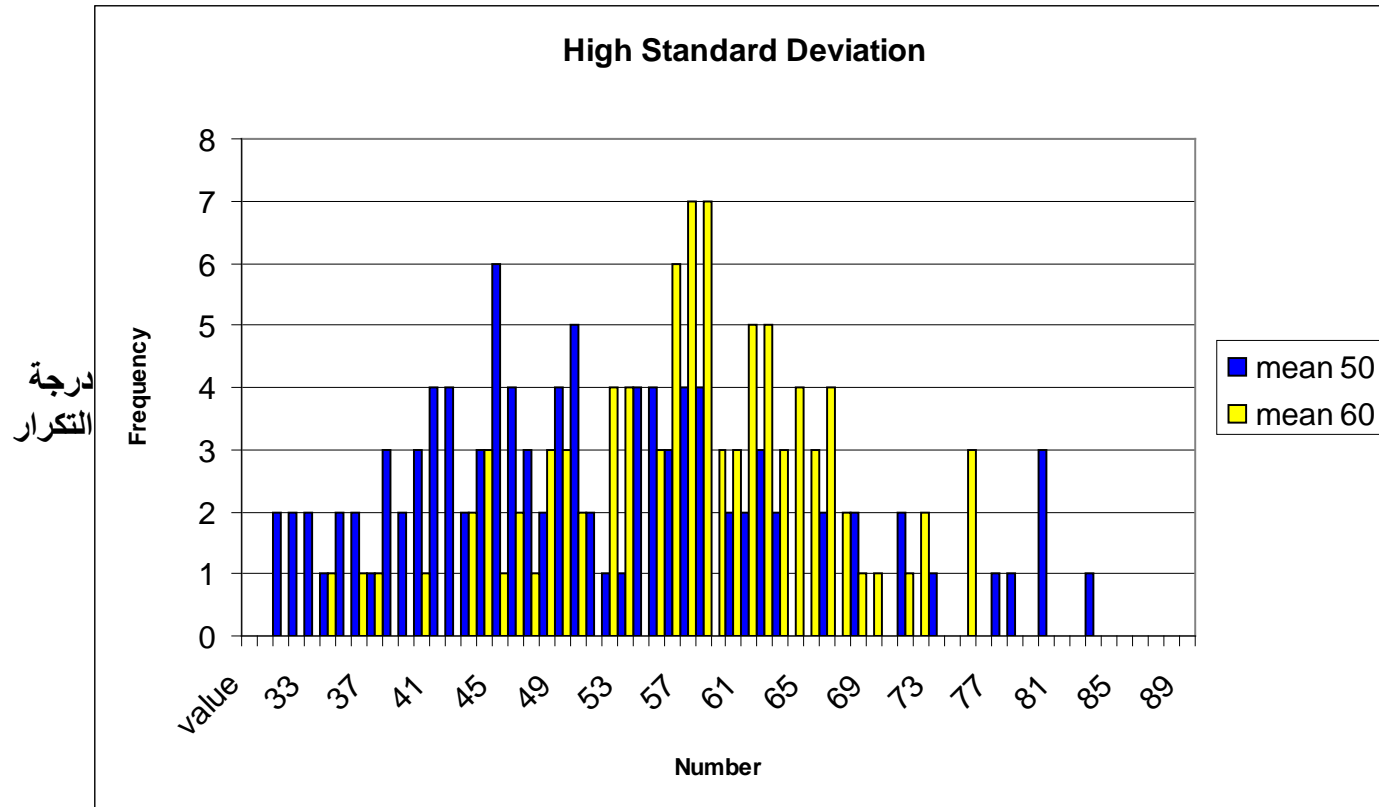


انحراف بمعيار متوسط



هل بإمكاننا أن نستنتج؟

انحراف بمعيار مرتفع



المراحل الفاصلة للثقة

■ ان حجم الأثر **المقدر** (الفرق في متوسطات العينة) قائم فقط لعينتنا. تعطي كل عينة جوابا مختلفا بعض الشيء. كيف نستخدم عينتنا لتقديم بيانات بشأن كامل السكان؟

■ فاصل ثقة قدره 95% لحجم الأثر يبين لنا أنه بالنسبة لـ 95% من كل العينات التي كان من الممكن أخذها من نفس السكان، فان حجم الأثر سيقع ضمن الفاصل المقدر.

■ **الخطأ المعياري (se)** لتقدير العينة يجسد كل من حجم العينة وتنوع النتيجة (تكون أكبر للعينة الصغيرة وبنتيجة متغيرة).

■ وبحكم التجربة : فاصل ثقة قدره 95%، هو تقريبا قيمة الأثر زائد او نقص خطئين معياريين.

اختبار الفرضية

كثيرا ما نكون مهتمين في اختبار الفرضية التي مفادها ان حجم الأثر يعادل صفر (نريد أن نكون قادرين على رفض الفرضية القائلة أن البرنامج لم يكن له أثر)

$$H_o : \text{Effect size} = 0$$

نريد أن نختبر :

$$H_a : \text{Effect size} \neq 0$$

مقابل :

نوعان من الأخطاء

■ أول نوع من الأخطاء : استخلاص ان هناك أثر، في حين أن في الواقع ليس هناك اثر.

مستوى اختبارك هو ان احتمالية أنك ستستخلص على سبيل الخطأ ان البرنامج له اثر، في حين أنه فعليا ليس له أثر.

وبالتالي، بوجود مستوى 5%، يمكنك أن تكون واثقا بنسبة 95% من صحة استنتاجك بأن البرنامج كان له أثر.

لأغراض السياسة العامة، تريد أن تكون واثقا للغاية من الإجابة التي تقدمها : سيحدد المستوى بصورة منخفضة نسبيا

مستوى شائع : 1%, 10%, 5% :

--

العلاقة مع فواصل الثقة

■ إذا لم يكن الصفر يقع ضمن فاصل الثقة 95% لحجم الأثر الذي قمنا بقياسه، يمكننا آنذاك ان نكون متأكدين بنسبة 95% من أن حجم الأثر ليس صفراً.

■ إذن بحكم التجربة، يكون حجم الأثر أكثر من الخطأ المعياري بمرتين، ويمكنك استنتاج بنسبة تأكد أعلى من 95% بأن البرنامج كان له اثر.

نوعان من الأخطاء

■ النوع الثاني من الخطأ : تعجز عن رفض ان البرنامج لم يكن له اثر، في حين أنه فعليا له اثر.

■ قوة الإختبار هي احتمالية أنني سأتمكن من إيجاد أثر مهم في تجربتي، إذا كان هناك فعليا أثر (القوة الأعلى أفضل بما أن الإحتمال الأكبر هو أن أبلغ عن وجود اثر فعلي).

■ القوة هي أداة تخطيط لتصميم دراسة. فهي تبين لي درجة احتمالية أن أجد أثرا مهما لحجم عينة معين.

■ القوة ناقص واحد هي احتمالية أن يخيب أمل المرء.

احتساب القوة

■ عند التخطيط لتقييم، وبعض التحريات الأولية، يمكننا احتساب العينة الدنيا التي نحتاجها لكي نقوم بالتالي :

■ اختبار فرضية محددة مسبقا : ما إذا كان أثر البرنامج صفر او ليس صفر

■ لمستوى محدد (مثل 5%)

■ في ضوء حجم اثر محدد مسبقا (ما تعتقد ان البرنامج سيقوم به)

■ لتحقيق قوة معينة

■ والقوة البالغة 80% تبين لنا أنه في 80% من التجارب على هذا الحجم من العينة، والتي تجرى على هؤلاء السكان، يمكننا القول أن هناك في عينتنا أثر على مستوى الثقة الرغوب به.

■ كلما كبرت العينة، كلما زادت القوة.

القوة الشائعة المستخدمة : 80%, 90%

مكونات احتساب القوة في دراسة بسيطة

أين نحصل عليه	ما نحتاج اليه
كثيرا ما يحدد هذا على 5%. وكلما زاد انخفاضه، كلما زاد حجم العينة المطلوبة لقوة معينة	مستوى الأهمية
- من السموات السابقة التي أجريت في أوضاع مشابهة - كلما زاد التنوع، كلما كبر حجم العينة لقوة معينة	الوسيط وتباين النتيجة في مجموعة المقارنة
ما هو أصغر أثر الذي ينبغي أن يثير ردا على السياسة؟ كلما صغر حجم الأثر الذي نريد استكشافه، كلما كبر حجم الأثر الذي نحتاجه لقوة معينة	حجم الأثر الذي نريد تقصيه

انتقاء حجم أثر

- ما هو أصغر أثر الذي ينبغي أن يبرر البرنامج الذي سيتم اعتماده:
- تكلفة هذا البرنامج مقابل المنافع التي يوفرها.
- تكلفة هذا البرنامج مقابل الإستخدام البديل للمال
- إذا كان الأثر سيكون أصغر من ذلك، فهو كأنه صفر: ولا يهمنا ان نثبت أن اثرا صغيرا جدا يختلف عن صفر.
- وبالمقابل، أي أثر أكبر من ذلك الأثر، من شأنه أن يبرر اعتماد هذا البرنامج: نريد أن نتمكن من أن نميزه عن الصفر.
- الخطر الشائع : انتقاء حجم الأثر متفائل للغاية – حجم العينة قد يكون حدد على مستوى منخفض أكثر مما ينبغي!

أحجام الأثر المعياري

- حجم الأثر الذي يمكنك استكشافه بعينة معينة يعتمد على درجة تنوع النتيجة
- مثال : لو كان لكل الأطفال مستوى تعلم مشابه جدا بدون برنامج، فإنه سيكون من السهل استكشاف أثر صغير جدا
- الإنحراف المعياري يجسد التنوع في النتيجة. كلما زاد التنوع، كلما زاد الإنحراف المعياري
- حجم الأثر المعياري هو حجم الأثر مقسم على الإنحراف المعياري للنتيجة
- $d = \text{effect size}/\text{St.dev.}$
- أحجام الأثر الشائعة :

(كبير) $d=0.50$ (وسط) $d=0.40$ (صغير) $d=0.20$

عوامل التصميم التي تؤثر على القوة

■ مستوى التوزيع العشوائي

■ توفر البيانات الأساسية

■ توفر متغيرات الرقابة والتقسيم الى طبقات.

■ نوع الفرضية التي يتم اختبارها.

مستوى التوزيع العشوائي للتصميم المجمع

ان محاولات التوزيع العشوائي المجمع هي تجارب يتم فيها عشوائيا توزيع الوحدات الإجتماعية أو التجمعات بدلا من الأفراد على مجموعات التدخل.

من الأمثلة على ذلك :

قرى	حوالات نقدية مشروطة
عيادات صحية	توزيع شبكات الفراش
مدارس	IPT
أسرة	دعم إجتماعي

سبب اعتماد التوزيع العشوائي للتجمعات

■ الحاجة للتقليل من التلوث للحد الأدنى أو إزالته بالكامل

■ مثال : في دراسة برنامج للتخلص من الديدان، تم اختيار المدارس على أنها الوحدة لأن الديدان معدية.

■ اعتبارات جدوى اساسية

■ مثال : برنامج PROGRESA لم يكن ليكون ذي جدوى سياسية لو تم إدخال بعض الأسر في قرية ما وليس الآخرين

■ فقط اختيار طبيعي

■ مثال : اي تدخل تعليمي يؤثر على على صف بأكمله (مثل الفلبيتشارت، وتدريب المدرسين)

Impact of Clustering

أثر

■ يمكن ربط النتائج لكل الأفراد ضمن وحدة ما

■ كل القرويين معرضين لنفس الطقس

■ هناك لكل المرضى موظف صحي واحد مشترك

■ لكل الطلاب مدير مدرسة واحد

■ يتفاعل سكان القرية مع بعضهم البعض

■ ينبغي تعديل حجم العينة لهذا الترابط

■ كلما زاد الترابط بين النتائج، كلما احتجنا لتعديل الأخطاء المعيارية

مثال عن مضاعفات أثر المجموعة

ما بين الطبقات الترايط	حجم المجموعة العشوائي			
	10	50	100	200
0.00	1.00	1.00	1.00	1.00
0.02	1.09	1.41	1.73	2.23
0.05	1.20	1.86	2.44	3.31
0.10	1.38	2.43	3.30	4.57

الآثار الناتجة

- من المهم للغاية توزيع عشوائيا مجموعة ملائمة من المجموعات
- كثيرا ما يكون عدد الأفراد ضمن مجموعات أقل أهمية من عدد المجموعات
- فكر أن "قانون العدد الكبير" ينطبق فقط عندما يزيد عدد المجموعات الموزعة عشوائيا
- لا يمكنك استخدام التوزيع العشوائي على مستوى المقاطعة، بمقاطعة واحدة خاضعة للمعالجة (مستفيدة)، ومقاطعة واحدة للرقابة!!!

توفر بيانات أساسية

■ للبيانات الأساسية ثلاثة استخدامات رئيسية :

- يمكن التحقق مما إذا كانت مجموعة المعالجة ومجموعة الرقابة متشابهة او مختلفة قبل المعالجة
- قم بتقليص حجم العينة المطلوبة، ولكن ذلك يتطلب إجراء مسح قبل البدء بالتدخل : آثار ذلك على التكلفة
- يمكن استخدام ذلك تقسيم المجموعات وتشكيلها في مجموعات فرعية

■ احتساب القوة ببيانات أساسية

- أنت بحاجة لمعرفة الترابط بين القياسين اللاحقين للنتيجة (مثلا: الإستهلاك المقاس في سنتين)
- كلما زاد الترابط، كلما زاد الربح
- أرباح كبيرة جدا لنتائج متواصلة مثل مشاركة القوة العاملة

متغيرات الرقابة

- إذا كان لدينا متغيرات إضافية ذات صلة (مثل سكان القرية، المنطقة التي تقع فيها القرية... الخ.) يمكننا أيضا أن نقوم بالرقابة لهم
- ما يهم الآن للقوة هو التنوع المتبقي بعد إجراء الرقابة لهذه المتغيرات
- إذا كانت متغيرات الرقابة تشرح جانبا كبيرا من التنوع، فإن الدقة ستزداد ومتطلب حجم العينة سينخفض
- تحذير: لا ينبغي أن تشمل متغيرات الرقابة سوى المتغيرات غير الواقعة تحت تأثير المعالجة : عادة المتغيرات التي تم تجميعها قبل التدخل.

العينات المقسمة الى طبقات

التقسيم الى طبقات : إنشاء مجموعات بموجب قيمة متغيرات الرقابة وقم بالتوزيع العشوائي ضمن كل مجموعة.

يضمن التقسيم الى طبقات أن مجموعات الرقابة والمعالجة متوازنة من حيث متغيرات الرقابة هذه.

ويقلل هذا من الإختلاف لسببين :

- يقلل من اختلاف نتيجة الإهتمام في كل طبقة
 - ترابط الوحدات ضمن التجمعات
- مثلا : إذا قمت بالتقسيم الى طبقات حسب المقاطعات، بشأن برنامج رش مضاد للناموس
- تتم السيطرة على العوامل المناخية الزراعية والعوامل الوبائية المصاحبه لها
 - يخفى "الأثر الحكومي للمقاطعة المشتركة".

عوامل التصميم التي تؤثر على القوة

■ Clustered design

■ التصميم المجمع

■ توفر بيانات أساسية

■ توفر متغيرات الرقابة والتقسيم الى طبقات.

■ نوع الفرضية التي يتم اختبارها.

نوع الفرضية التي يتم اختبارها

هل أنت مهتم بالفرق بين معالجتين بالإضافة الى الفرق بين المعالجة والرقابة؟

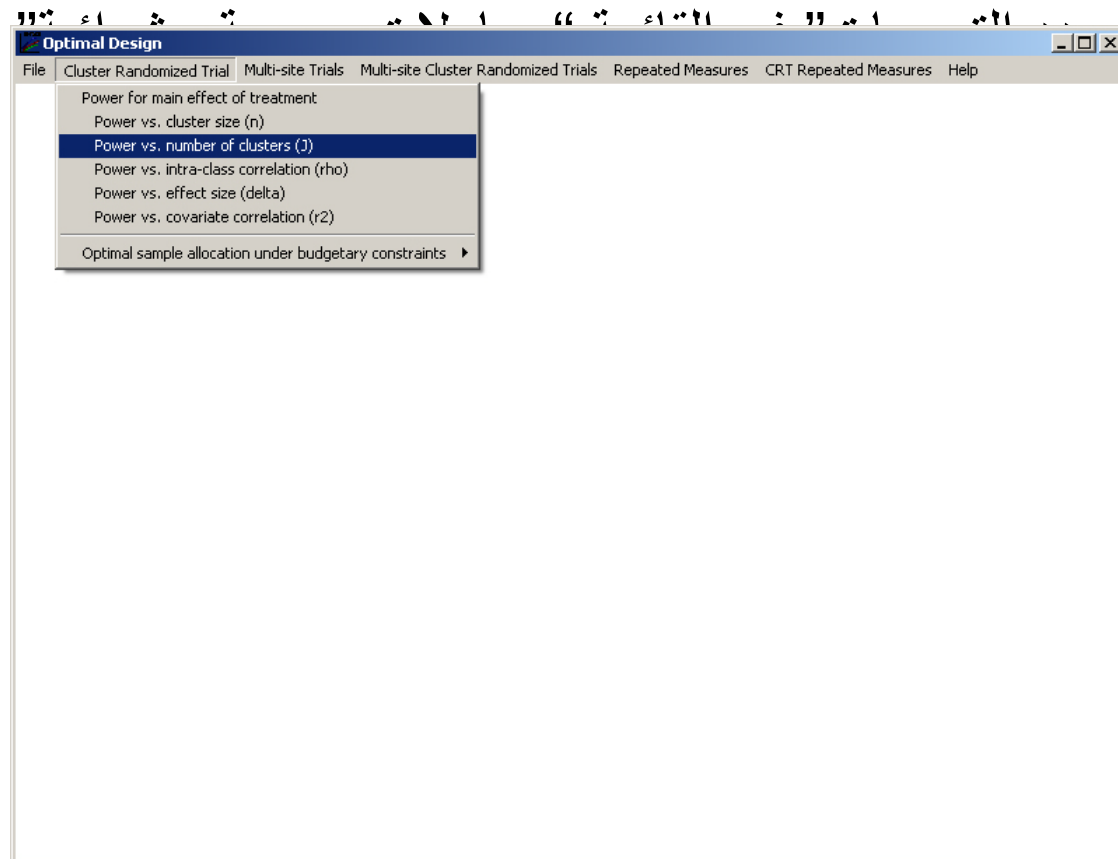
هل أنت مهتم بالتفاعل بين المعالجات؟

هل أنت مهتم باختبار ما إذا كان الأثر مختلفا في مجموعات السكان الفرعية؟

هل يتضمن تصميمك امتثال جزئي فقط؟ (مثل تصميم تشجيعي؟)

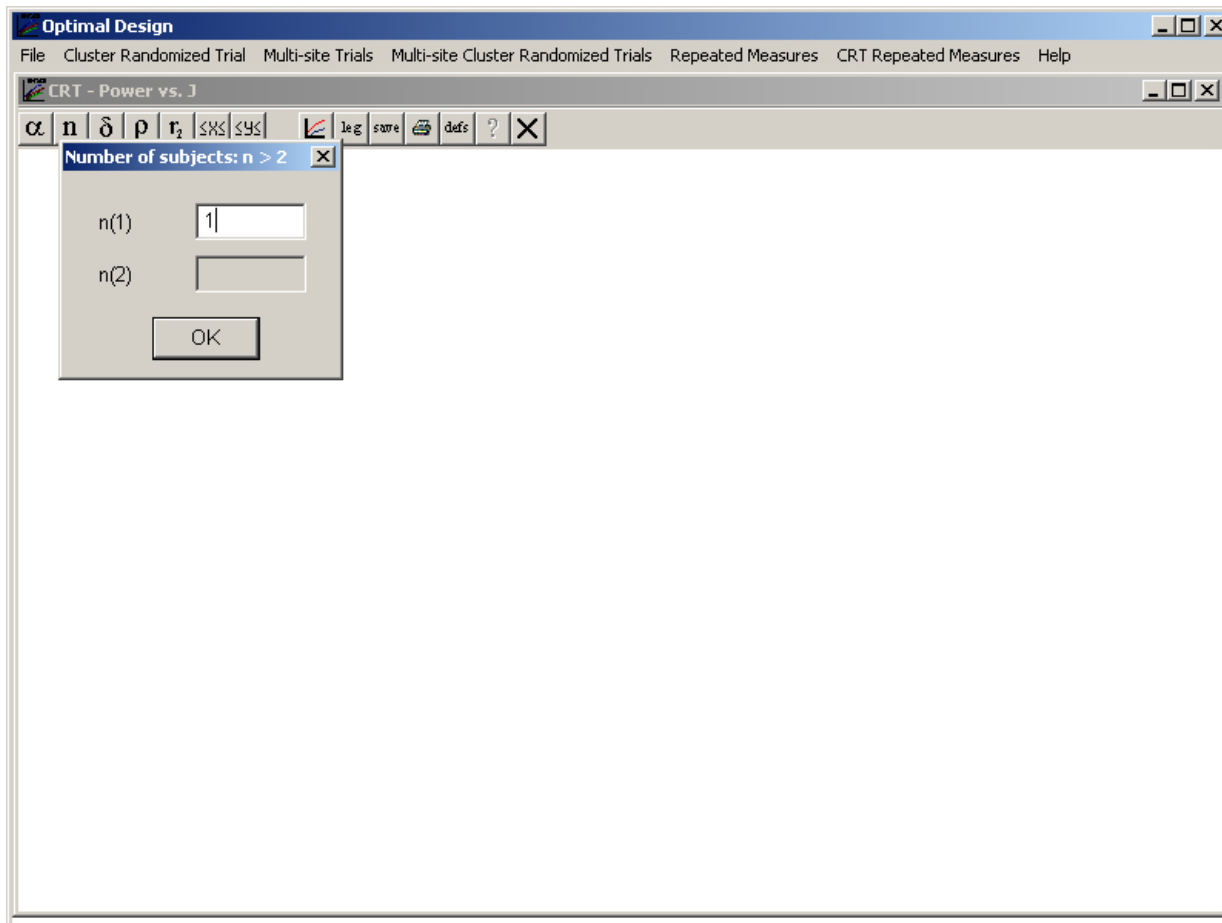
حسابات القوة باستخدام برمجية OD

- Choose “Power v. number of clusters” in the menu “clustered randomized trials”



حجم التجمع

■ اختر حجم التجمع



اختر مستوى الأهمية، والمعالجة والأثر، والترابط

■ اختر : مستوى

■ عادة تختار 0.05

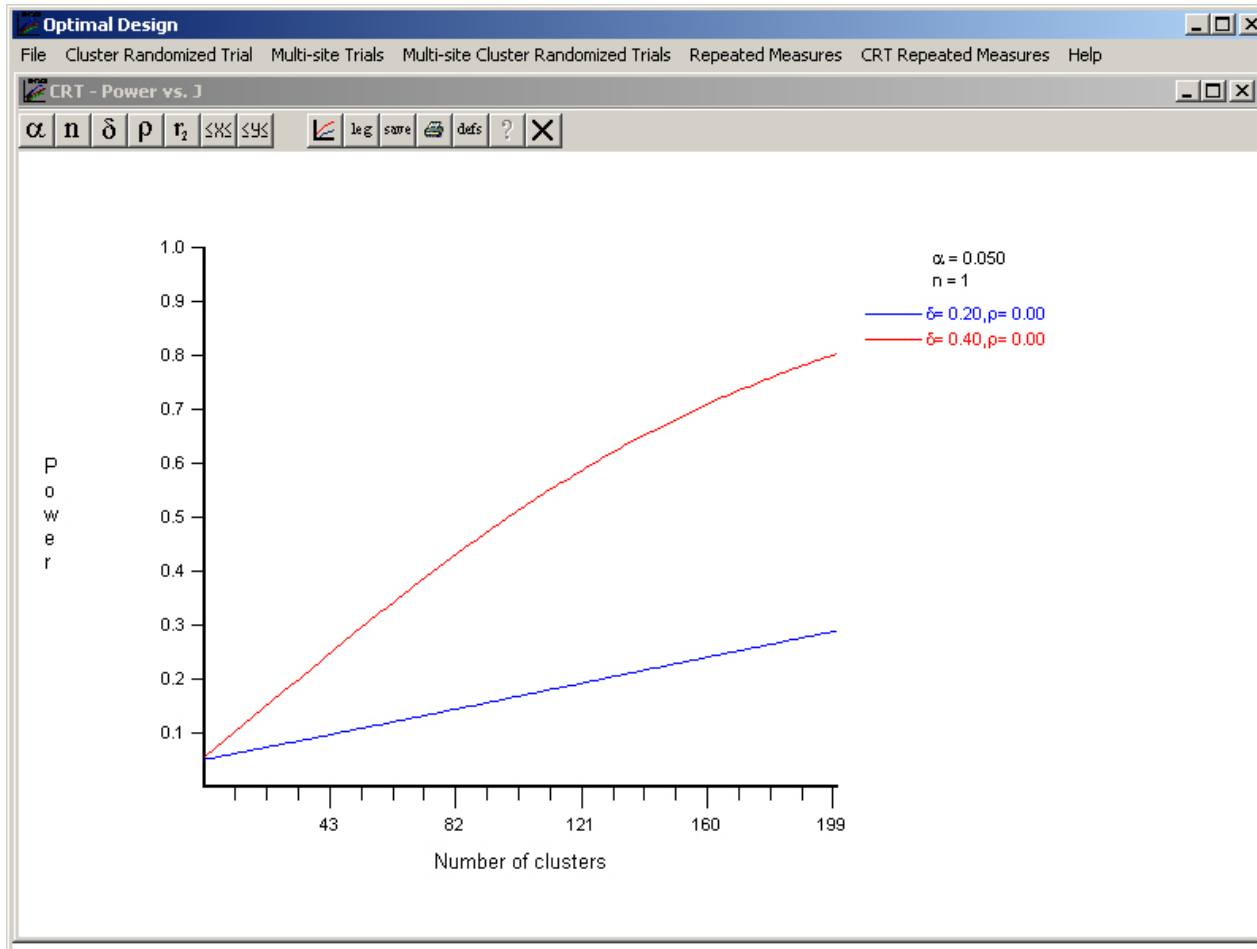
■ اختر: d:

■ يمكن الإختبار ب 0.20

■ انتقِ الترابط بين الطبقات (ρ)

■ تحصل على الرسم البياني الناتج وهو يظهر القوة كدالة لحجم العينة

القوة وحجم العينة



الخلاصات : احتساب القوة عمليا

- تتضمن حسابات القوة بعض التخمين.
- أحيانا لا يكون لدينا المعلومات الصحيحة لإجراء ذلك بالصورة الصحيحة.
- الا أنه من المهم ان نبذل جهودا للقيام بذلك :
- تفادى إطلاق الدراسات التي لن يكون لها قوة على الإطلاق: مضيعة للوقت والمال
- خصص الموارد الملائمة للدراسات التي تقرر إجرائها (وليس أكثر مما ينبغي).