



BMAHREC

Sample size for regression



สมหญิง ทิพย์มงคล พยาบาลวิชาชีพชำนาญการพิเศษ
ศรสวรรค์ คิ้ววิจิตร นักวิชาการสาธารณสุขปฏิบัติการ
วรารณ เนตรพราว นักวิชาการสถิติปฏิบัติการ



การคำนวณขนาดกลุ่มตัวอย่างสำหรับ Logistic Regression

Logistic Regression เป็นเทคนิคทางสถิติที่ใช้วิเคราะห์ตัวแปรตามที่มีลักษณะเป็น Binary Outcome (0/1) เช่น มีโรค/ไม่มีโรค หรือ ติดเชื้อ/ไม่ติดเชื้อ การกำหนดขนาดตัวอย่างที่เหมาะสมมีความสำคัญอย่างยิ่งต่อความแม่นยำและเสถียรภาพของตัวแบบ



หลักการพื้นฐานของ Logistic Regression

1 ลักษณะของ Logistic Regression

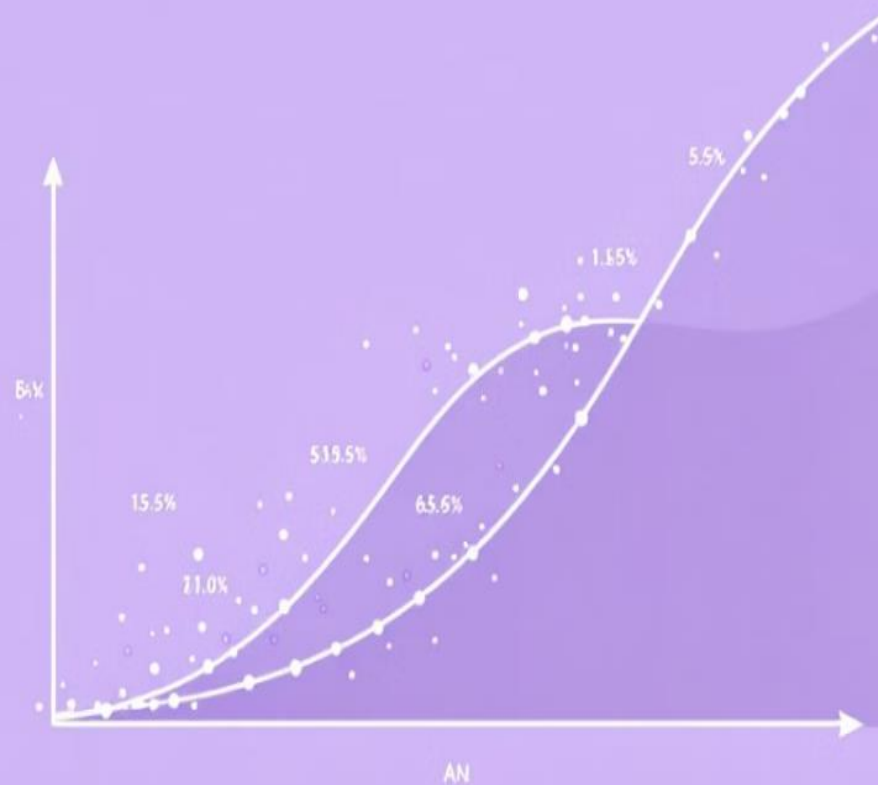
Logistic Regression ใช้สำหรับวิเคราะห์ตัวแปรตามที่เป็น Binary Outcome (0/1) เช่น มีโรค/ไม่มีโรค หรือ ติดเชื้อ/ไม่ติดเชื้อ โดยสามารถทำนายความน่าจะเป็นของการเกิดเหตุการณ์จากตัวแปรอิสระหลายตัว

2 ความสำคัญของขนาดตัวอย่าง

ขนาดตัวอย่างที่เพียงพอช่วยให้การวิเคราะห์มีความแม่นยำ และเสถียรภาพของตัวแบบ หากขนาดตัวอย่างไม่เพียงพอ อาจทำให้เกิดปัญหา overfitting หรือ underfitting ซึ่งส่งผลต่อความน่าเชื่อถือของผลการวิเคราะห์

3 ปัจจัยที่มีผลต่อขนาดตัวอย่าง

จำนวนตัวแปรอิสระ ความชุกของเหตุการณ์ที่สนใจ (prevalence) ขนาดของผลกระทบที่ต้องการตรวจจับ (effect size) และระดับความเชื่อมั่นทางสถิติ ล้วนมีผลต่อการกำหนดขนาดตัวอย่างที่เหมาะสม



Rule of Thumb (Peduzzi's Rule)

หลักการของ Peduzzi's Rule (1996) แนะนำว่าควรมีอย่างน้อย 10 เหตุการณ์ (Events) ต่อ 1 ตัวแปรอิสระ (Predictor) กรณีตัวแปรอิสระเป็น Binary เป็นกฎอย่างง่ายที่นิยมใช้ในการคำนวณขนาดกลุ่มตัวอย่างเบื้องต้น

สูตรการคำนวณ
$$N = n = (\text{cases} \times k) / p$$
 โดยที่ N = ขนาดกลุ่มตัวอย่าง, k = จำนวนตัวแปรอิสระ, p = อัตราส่วนของกรณีที่เกิดเหตุการณ์ (prevalence) ซึ่งมีค่าระหว่าง 0 ถึง 1

ตัวอย่างการคำนวณ

ถ้ามี 5 ตัวแปรอิสระ และคาดว่า 10% ของประชากรจะเกิดเหตุการณ์ ($p = 0.1$) จะได้ $N = 10(5)/0.1 = 500$
ดังนั้น ควรมีอย่างน้อย 500 ตัวอย่างในการศึกษา



Rule of Thumb สำหรับตัวแปรต้นที่เป็นกลุ่ม

ระบุตัวแปร Dummy

คำนวณจำนวนตัวแปร dummy ที่เกิดจากการแปลงตัวแปรเป็นกลุ่ม

ปรับขนาดตัวอย่าง

เพิ่มขนาดตัวอย่างหากมีตัวแปรกลุ่มจำนวนมาก



ใช้สูตร

$n = (b-1) \times 10$ cases per predictor
b คือ จำนวนการแบ่งกลุ่มตัวแปร
เช่น เพศ แบ่งเป็น 2 กลุ่ม ชาย หญิง
 $n = (2-1) \times 10 = 10$ เคส

คำนวณขนาดตัวอย่าง

แทนค่าในสูตรเพื่อหาขนาดกลุ่มตัวอย่างที่เหมาะสม

เมื่อมีตัวแปรต้นที่เป็นกลุ่ม เช่น เพศ (ชาย/หญิง), กลุ่มอายุ, ประเภทของโรคประจำตัว จะต้องพิจารณาจำนวนตัวแปร dummy ที่เกิดจากการแปลงตัวแปรเป็นกลุ่ม ซึ่งจะส่งผลต่อขนาดกลุ่มตัวอย่างที่ต้องการ

Rule of Thumb สำหรับตัวแปรต้นที่เป็นตัวแปรเชิงปริมาณ

- 30 per predictor [Cohen&Cohen, 1975]



ตัวอย่างการคำนวณสำหรับตัวแปรกลุ่ม

ตัวแปร	จำนวนระดับ	จำนวนตัวแปร Dummy
เพศ (ชาย/หญิง)	2	1
กลุ่มอายุ (<30, 30-40, 41-50, >50)	4	3
โรคประจำตัว (ไม่มี, เบาหวาน, ความดันสูง)	3	2
ตัวแปรอิสระที่เป็นตัวเลข	-	3
รวม	-	9

จากตาราง เมื่อคำนวณจำนวนตัวแปร dummy ทั้งหมดได้ 9 ตัวแปร และสมมติว่าอัตราการเกิดเหตุการณ์ (p) เท่ากับ 0.05 (5% ของประชากรเกิดเหตุการณ์) เราสามารถคำนวณขนาดกลุ่มตัวอย่างได้ดังนี้: $= 10 \times 9 / 0.05 = 1,800$ ตัวอย่าง

การคำนวณโดยใช้ Power Analysis

เริ่มต้นด้วยการกำหนดค่า Alpha (α) ซึ่งมักใช้ 0.05, Power ($1-\beta$) มักใช้ 80% หรือ 90%, Effect Size (Odds Ratio หรือ R^2) ที่ต้องการตรวจจับ และ Prevalence ของ Outcome ในประชากร

สูตรพื้นฐานคือ $N = (Z_{\alpha/2} + Z_{\beta})^2 / [p(1-p) \cdot (\log(OR))^2]$ โดยที่ $Z_{\alpha/2}$ และ Z_{β} คือค่า Z-score ที่สอดคล้องกับระดับนัยสำคัญและกำลังการทดสอบ ตามลำดับ

ตัวอย่างการคำนวณ

ถ้า Alpha = 0.05, Power = 80%, Odds Ratio (OR) ที่สนใจ = 1.5, และ Prevalence (p) = 20% สามารถแทนค่าในสูตรหรือใช้โปรแกรมช่วยคำนวณเพื่อหาขนาดตัวอย่างที่เหมาะสม

ปรับขนาดตัวอย่าง

หลังจากได้ขนาดตัวอย่างเบื้องต้น อาจต้องปรับเพิ่มเพื่อรองรับการสูญหายของข้อมูล (dropout) หรือปัจจัยอื่นๆ ที่อาจทำให้ต้องการขนาดตัวอย่างเพิ่มขึ้น



วิธีของ Vittinghoff & McCulloch

- 1 Vittinghoff & McCulloch (2007) เสนอว่าในบางกรณี 5-9 เหตุการณ์ต่อ 1 ตัวแปรอิสระ อาจเพียงพอ ซึ่งน้อยกว่ากฎของ Peduzzi ที่แนะนำ 10 เหตุการณ์ต่อตัวแปร วิธีนี้ช่วยลดขนาดตัวอย่างลงได้ในบางสถานการณ์
- 2 วิธีนี้เหมาะสำหรับกรณีที่มีตัวแปรอิสระจำนวนน้อย และ Prevalence ของ Outcome สูง นอกจากนี้ ควรมีความสัมพันธ์ที่ชัดเจนระหว่างตัวแปรอิสระและตัวแปรตาม และไม่มีปัญหา multicollinearity
- 3 ต้องระวังปัญหา overfitting โดยเฉพาะเมื่อมีตัวแปรจำนวนมาก หรือเมื่อ Outcome มีความชุกต่ำ ควรพิจารณาใช้เทคนิคเพิ่มเติม เช่น cross-validation

การใช้ซอฟต์แวร์ช่วยคำนวณ



G*Power



R (pwr หรือ epiR package)



Stata

โปรแกรมฟรีที่ใช้งานง่าย
สามารถเลือก Logistic
Regression และตั้งค่าตัวแปร
ต่างๆ เช่น Alpha, Power,
Odds Ratio และ Prevalence
เพื่อคำนวณขนาดตัวอย่างได้
อย่างแม่นยำ เหมาะสำหรับผู้
ที่ไม่คุ้นเคยกับการเขียนโค้ด

สำหรับผู้ที่คุ้นเคยกับการเขียน
โค้ด R สามารถใช้แพ็คเกจ pwr
หรือ epiR ในการคำนวณขนาด
ตัวอย่าง มีความยืดหยุ่นสูงและ
สามารถปรับแต่งพารามิเตอร์ได้
หลากหลาย รวมถึงสร้างกราฟ
แสดงความสัมพันธ์ระหว่างขนาด
ตัวอย่างและ Power

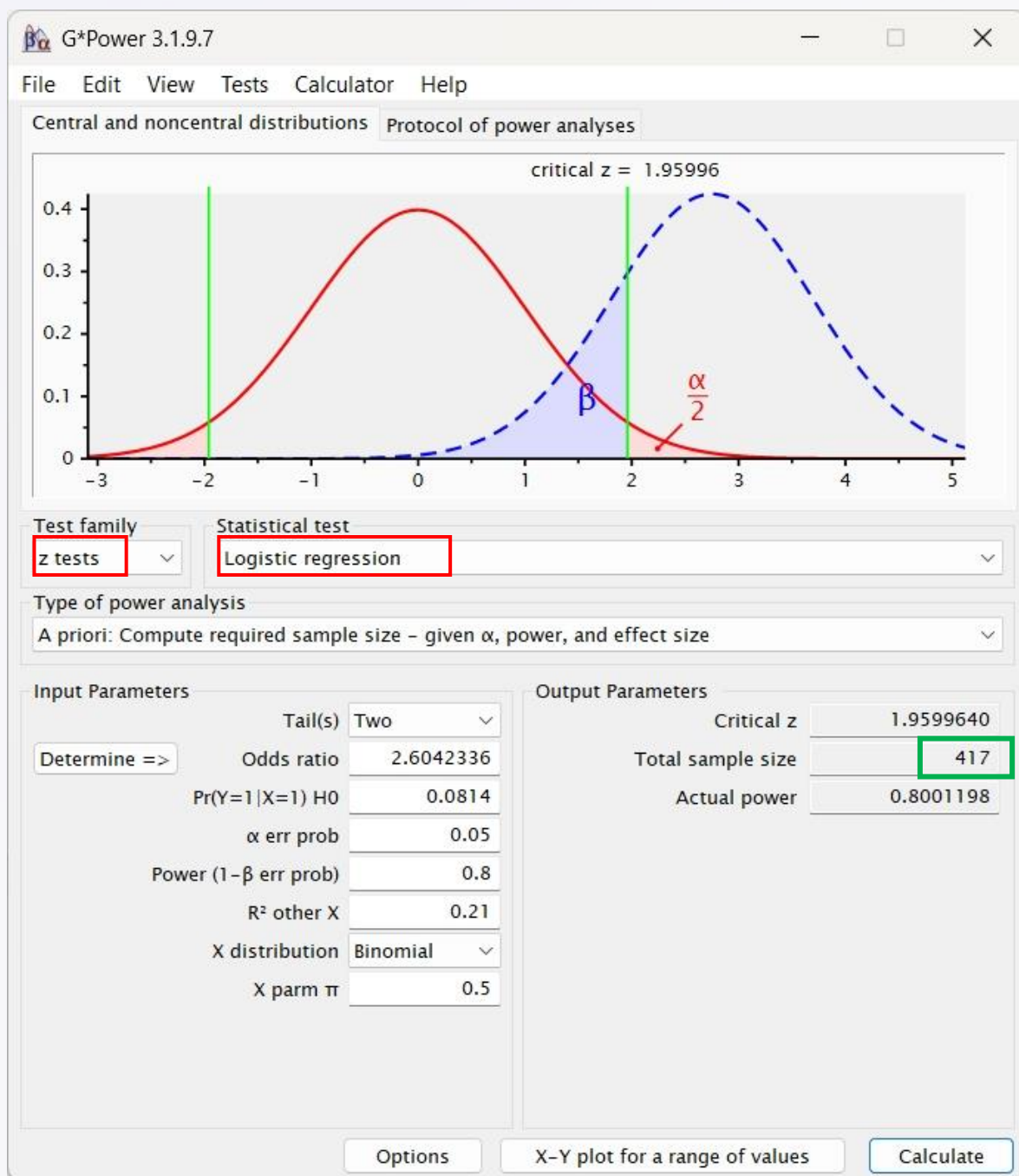
โปรแกรม Stata มีคำสั่ง
sampsi หรือ power
commands ที่ช่วยในการ
คำนวณขนาดตัวอย่างสำหรับ
Logistic Regression ได้
อย่างมีประสิทธิภาพ เหมาะ
สำหรับนักวิจัยที่ใช้ Stata
เป็นหลักในการวิเคราะห์
ข้อมูล



PASS Software

โปรแกรมเฉพาะทางสำหรับการคำนวณขนาดตัวอย่าง
มีฟังก์ชันครอบคลุมสำหรับ Logistic Regression และ
การวิเคราะห์ทางสถิติอื่นๆ มีอินเตอร์เฟซที่ใช้งานง่าย
และให้ผลลัพธ์ที่ละเอียด แต่เป็นซอฟต์แวร์เชิงพาณิชย์

ตัวอย่างการวิเคราะห์โดยใช้โปรแกรม G*Power



ตาราง Correlations

	age	sex	insur	weve	stay
age	1.0000				
sex	-0.0557	1.0000			
insur	-0.0530	-0.0104	1.0000		
weve	0.1007	0.1098	0.0365	1.0000	
stay	0.2109	-0.0289	0.0791	0.0843	1.0000

ตาราง two by two

sex	y		Total
	0	1	
1	104	24	128
2	158	14	172
Total	262	38	300

Pr(Y=1|X=1) H1: 0.1875

Pr(Y=1|X=1) H0: 0.0814

Calculate Odds ratio: 2.604234

Calculate and transfer to main window

Close

ตัวอย่างการวิเคราะห์โดยใช้โปรแกรม G*Power (ต่อ)

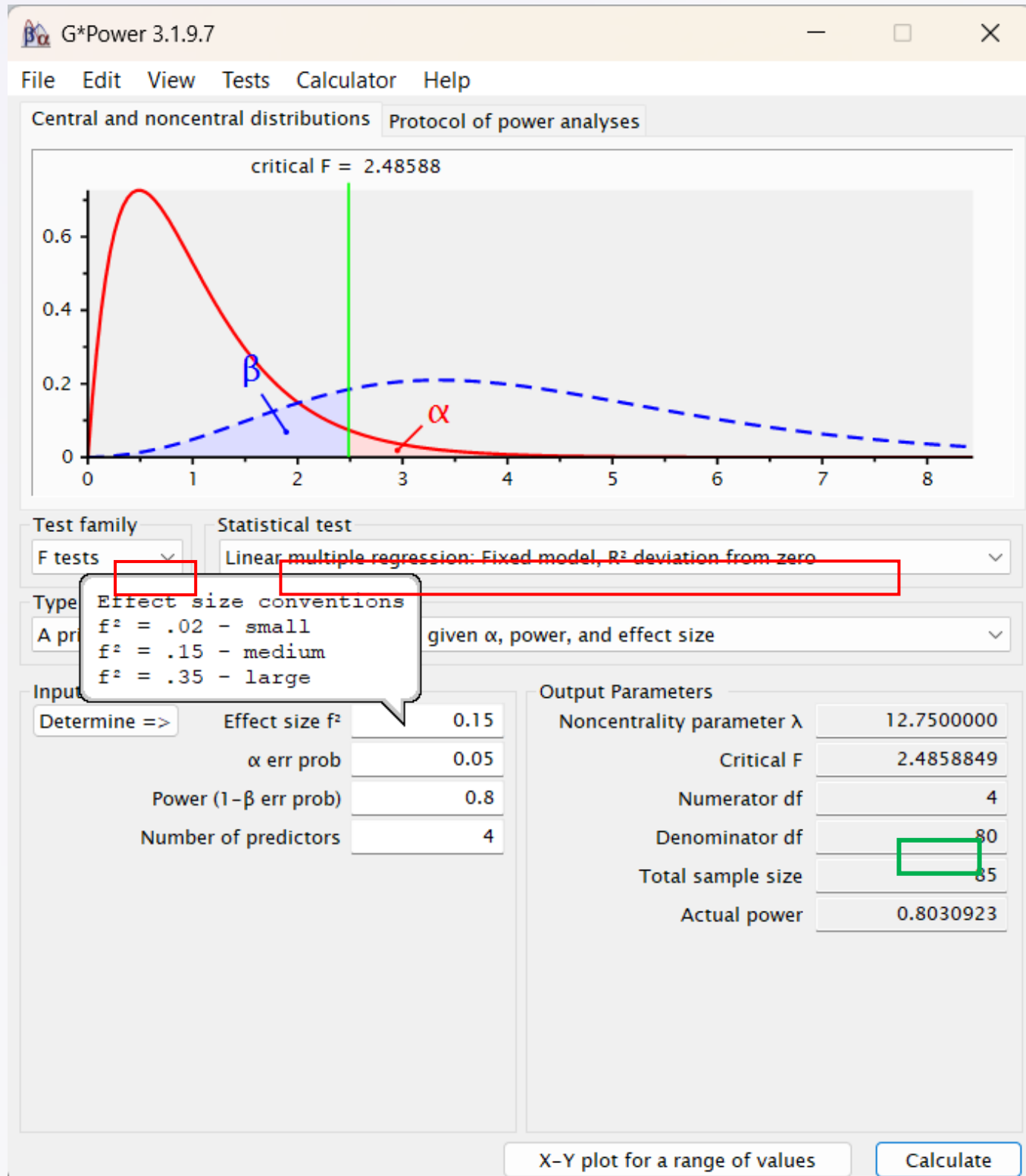
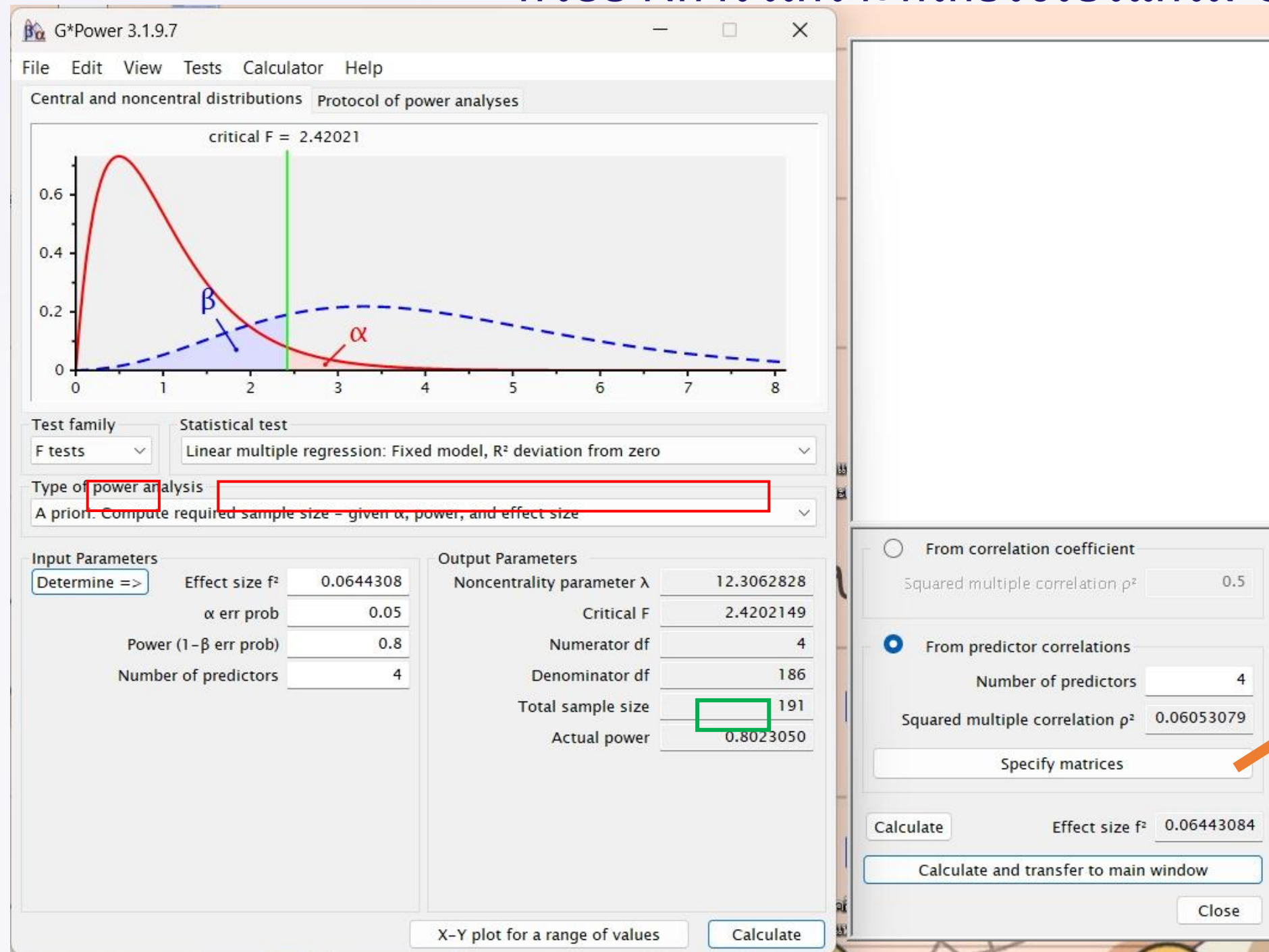


Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test variable	ES index	Effect size		
		Small	Medium	Large
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

อ้างอิง: Cohen J. A power primer. Psychol Bull. 1992 Jul;112(1):155-9. doi: 10.1037//0033-2909.112.1.155. PMID: 19565683.

ตัวอย่างการวิเคราะห์โดยใช้โปรแกรม G*Power (ต่อ)



ตาราง Correlations

	age	sex	insur	weve	stay
age	1.0000				
sex	-0.0557	1.0000			
insur	-0.0530	-0.0104	1.0000		
weve	0.1007	0.1098	0.0365	1.0000	
stay	0.2109	-0.0289	0.0791	0.0843	1.0000

The 'Input predictor correlations' dialog box is shown. The 'Corr between predictors and outcome' tab is active. The 'Number of predictors' is set to 4. The table below shows the correlations between predictors and the outcome Y:

predictor	P 1	P 2	P 3	P 4
corr with outcome Y	-0.0557	-0.0530	0.1007	0.2109

The 'Specify matrices' button is highlighted with an orange arrow. The 'Calculate' button is also visible. The 'Effect size f²' is 0.06443084. The 'Calculate and transfer to main window' button is at the bottom.

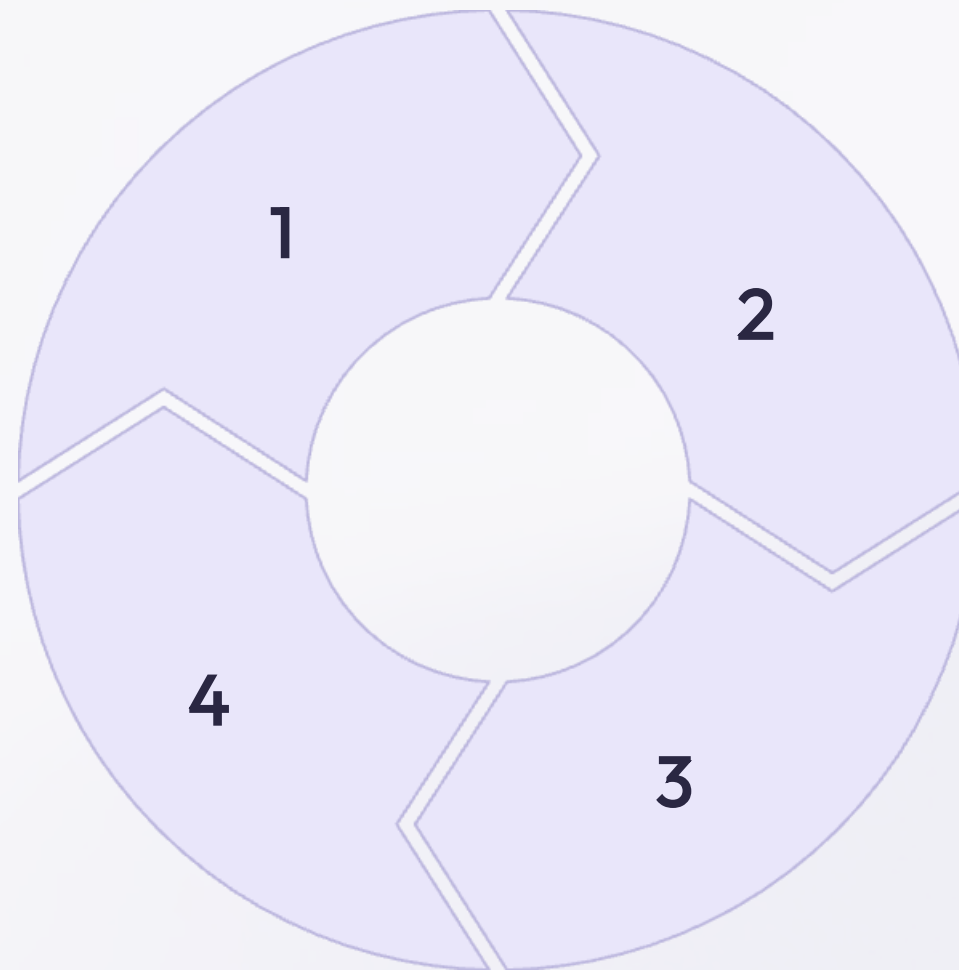
คำแนะนำสำหรับงานวิจัย

Outcome หายาก

ถ้า Outcome ที่สนใจหายาก (prevalence <10%) ควรเพิ่มขนาดกลุ่มตัวอย่างให้มากขึ้น เพื่อให้มีจำนวนเหตุการณ์เพียงพอต่อการวิเคราะห์

การทดสอบความไว

ควรทำ Sensitivity Analysis เพื่อประเมินว่าผลการวิเคราะห์จะเปลี่ยนแปลงอย่างไรหากมีการเปลี่ยนแปลงขนาดตัวอย่างหรือพารามิเตอร์อื่นๆ



ตัวแปรจำนวนมาก

ถ้ามีตัวแปรอิสระจำนวนมาก ควรใช้ Power Analysis แทน Rule of Thumb เพื่อให้ได้ขนาดตัวอย่างที่แม่นยำมากขึ้น และพิจารณาใช้เทคนิคการลดมิติข้อมูล

เทคนิคขั้นสูง

ถ้าใช้ Machine Learning หรือ Regularization Methods (LASSO, Ridge) อาจลดข้อจำกัดเรื่อง "10 Events per Variable" ได้ แต่ยังคงต้องมีขนาดตัวอย่างที่เพียงพอ

สูตรคำนวณกลุ่มตัวอย่างเพื่อ Dropout

หากต้องการปรับขนาดกลุ่มตัวอย่างให้เพียงพอเมื่อมีการ drop-out
สามารถใช้สูตรพื้นฐานที่ใช้กันอย่างแพร่หลายในงานวิจัยทางการแพทย์
และสาธารณสุข

สูตรคำนวณ

$$n_{\text{adjusted}} = n / (1 - d)$$

ตัวแปรในสูตร

n_{adjusted} = ขนาดตัวอย่างที่ต้องรับสมัครจริง

n = ขนาดตัวอย่างที่คำนวณได้ก่อนเพื่อ dropout

d = อัตราการ dropout ในรูปแบบสัดส่วน
(เช่น 20% ให้ใช้ 0.20)



ตัวอย่างการคำนวณกลุ่มตัวอย่าง

การคำนวณขนาดกลุ่มตัวอย่างที่เผื่อ dropout จะช่วยให้นักวิจัยวางแผนการรับสมัครผู้เข้าร่วมวิจัยได้อย่างเหมาะสม ตัวอย่างต่อไปนี้ จะช่วยให้เข้าใจวิธีการคำนวณได้ชัดเจนยิ่งขึ้น

สมมติว่าการศึกษาของคุณคำนวณขนาดตัวอย่างที่ต้องการไว้ที่ 1,000 คน

ประมาณอัตราการ dropout

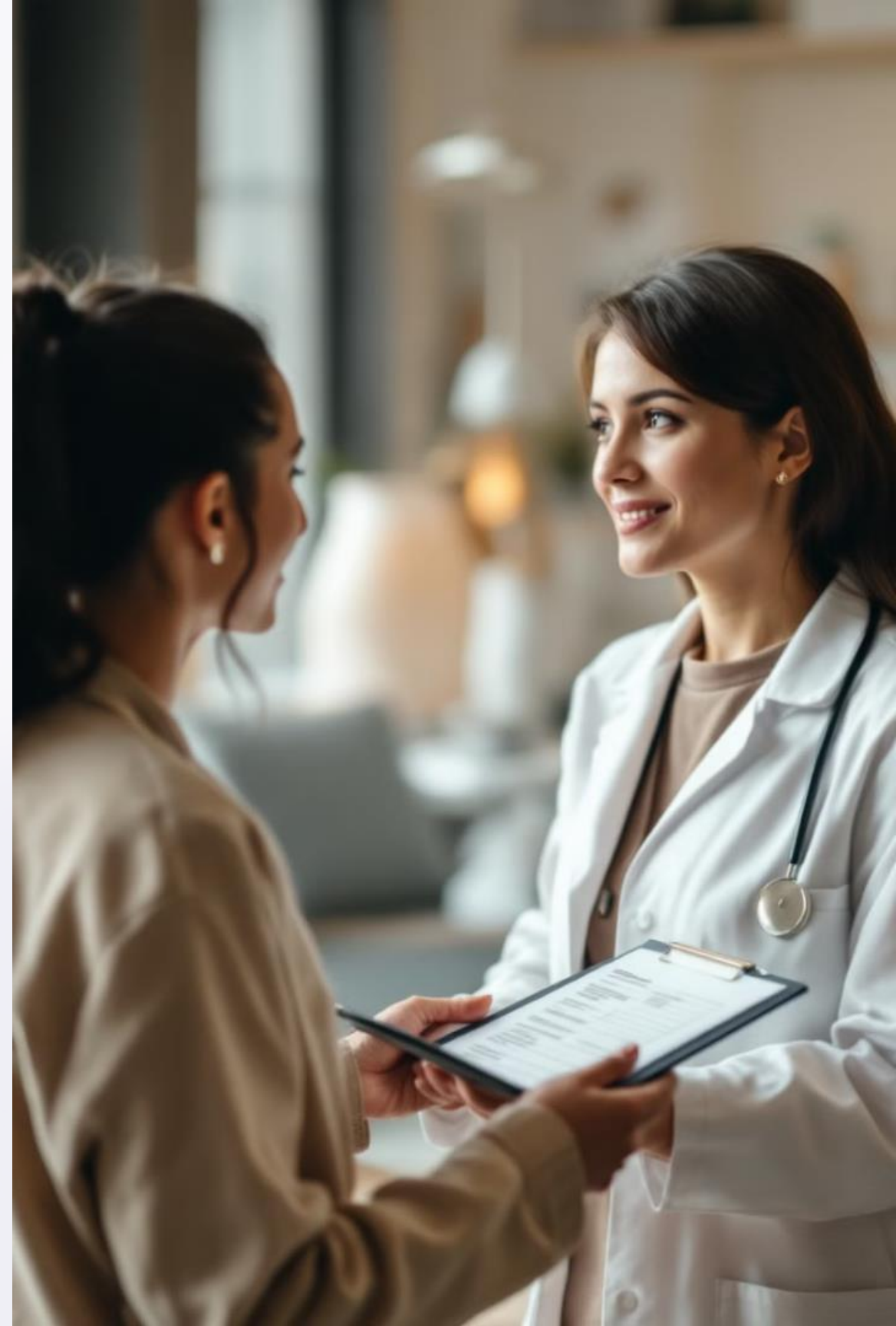
คาดว่าอัตราการ dropout จะอยู่ที่ 15% (0.15)

แทนค่าในสูตร

$$n_{\text{adjusted}} = 1000 / (1 - 0.15) = 1000 / 0.85 \approx 1177$$

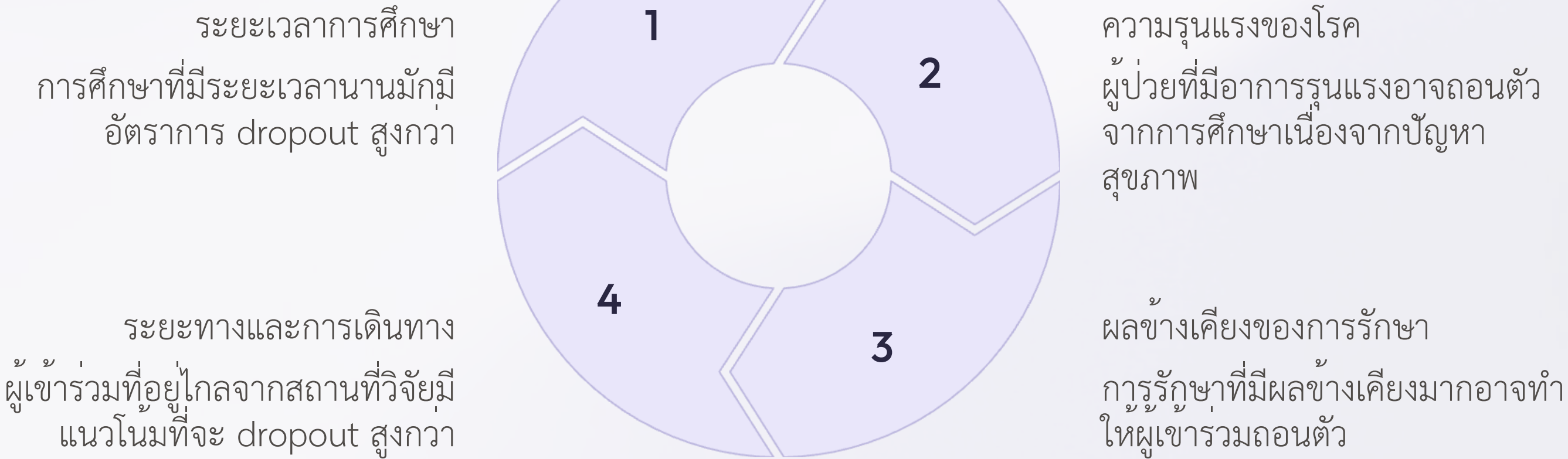
สรุปผลการคำนวณ

ต้องรับสมัครผู้เข้าร่วมอย่างน้อย 1,177 คน เพื่อให้เหลืออย่างน้อย 1,000 คนเมื่อสิ้นสุดการศึกษา



ปัจจัยที่มีผลต่ออัตราการ Dropout

การประมาณอัตราการ dropout ที่แม่นยำเป็นสิ่งสำคัญในการคำนวณขนาดกลุ่มตัวอย่าง นักวิจัยควรพิจารณาปัจจัยต่างๆ ที่อาจส่งผลต่ออัตราการ dropout ในการศึกษาของตน



สรุปการคำนวณขนาดกลุ่มตัวอย่างสำหรับ Logistic Regression



ไม่ว่าจะเลือกใช้วิธีใด สิ่งสำคัญคือต้องพิจารณาบริบทของการวิจัย ข้อจำกัดในการเก็บข้อมูล และวัตถุประสงค์ของการศึกษาประกอบกัน เพื่อให้ได้ขนาดตัวอย่างที่เหมาะสมที่สุดสำหรับงานวิจัยนั้นๆ



การคำนวณ sample size โดยใช้คำสั่ง pmsampsize

คำสั่ง pmsampsize ใน โปรแกรม STATA และ R เป็นคำสั่งสำหรับการคำนวณจำนวนขนาดตัวอย่างขั้นต่ำในการสร้างรูปแบบการทำนายแบบหลายตัวแปร (multivariable prediction model)

Logistic regression: ค่าที่จำเป็นต้องใช้วิเคราะห์ ได้แก่

1. ค่า pseudo R2 หรือ ค่าพื้นที่ใต้กราฟ ROC (AUC) ของ logistic regression model จากงานวิจัยอ้างอิงที่ทำการศึกษาคลายเรา
2. จำนวนตัวแปรอิสระที่ต้องการศึกษา (Predictor or Parameters)
3. สัดส่วนของการเกิด event หรือ outcome ของการศึกษาที่ผ่านมา หรือ จาก pilot study (Prevalence)

คำสั่ง

R2: pmsampsize, type(b) csrsquared() parameters() prevalence()

AUC: pmsampsize, type(b) cstatistic() parameters() prevalence()

Linear regression: ค่าที่จำเป็นต้องใช้วิเคราะห์ ได้แก่

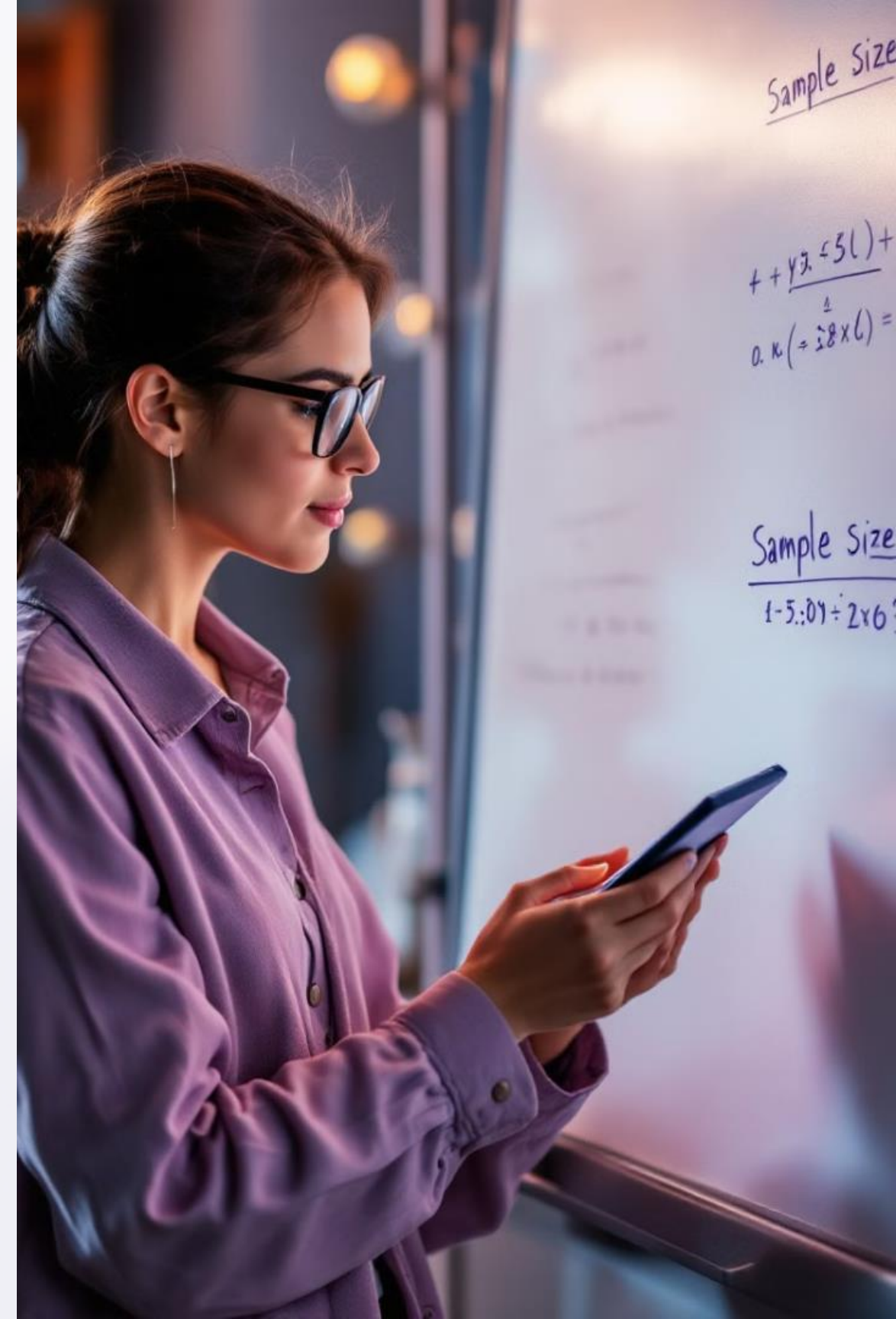
1. ค่า Adjusted R² ของ Linear regression model จากงานวิจัย
อ้างอิงที่ทำการศึกษาคลายเรา
2. จำนวนตัวแปรอิสระที่ต้องการศึกษา (Predictor or Parameters)
3. ค่าเฉลี่ย (mean)
4. ค่าส่วนเบี่ยงเบนมาตรฐาน (SD)

คำสั่ง

R2: pmsampsize, type(c) rsquared() parameters() intercept() sd()

เมื่อคำนวณแล้ว เลือกค่าที่มากที่สุดในตาราง sample size

อ้างอิง: BMJ 2020;368:m441 doi: 10.1136/bmj.m441 (Published 18 March 2020)



ตัวอย่างการวิเคราะห์โดยใช้โปรแกรม STATA

Stata/SE 15.1

File Edit Data Graphics Statistics User Window Help

Review

Filter commands here

Command _rc

1 pmsampsize, type(b) csrsquared...

. pmsampsize, type(b) csrsquared(0.288) parameters(24) prevalence(0.174)

NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared

NB: Assuming 0.05 margin of error in estimation of intercept

NB: Events per Predictor Parameter (EPP) assumes prevalence = .174

	Samp_size	Shrinkage	Parameter	CS_Rsq	Max_Rsq	Nag_Rsq	EPP
Criteria 1	623	.9	24	.288	.603	.477	4.52
Criteria 2	662	.905	24	.288	.603	.477	4.8
Criteria 3	221	.	24	.288	.603	.477	1.6
Final SS	662	.905	24	.288	.603	.477	4.8

Minimum sample size required for new model development based on user inputs = 662,
with 116 events (assuming an outcome prevalence = .174), and an EPP = 4.8

Criteria 1 - small overfitting defined as expected shrinkage of predictor effects by 10% or less

Criteria 2 - small absolute difference in model's apparent and adjusted Nagelkerke's R-squared

Criteria 3 - precise estimation of the average outcome risk in the population

.

Variables

Filter variables here

Name Label

There are no items to show.

Properties

Variables

Name

Label

Type

Format

Value label

Notes

Data

Filename

Label

Notes

Variables 0

Observations 0

Size 0

Memory 32M

Sorted by

Command

pmsampsize, type(b) csrsquared(0.288) parameters(24) prevalence(0.174)

$$n = \frac{P}{(s - 1) \ln \left(1 - \frac{R_{cs}^2}{S} \right)}$$



การคำนวณขนาดกลุ่มตัวอย่างสำหรับ Regression

Regression เป็นเทคนิคทางสถิติที่ใช้วิเคราะห์ตัวแปรตามที่มีลักษณะเป็น continuous เช่น อายุ BMI เป็นต้น

สูตร $n = \text{cases} * k$ (k คือปัจจัยที่สนใจ)

กรณีตัวแปรต้นเป็น Numerical (continuous) response:

- 30 per predictor

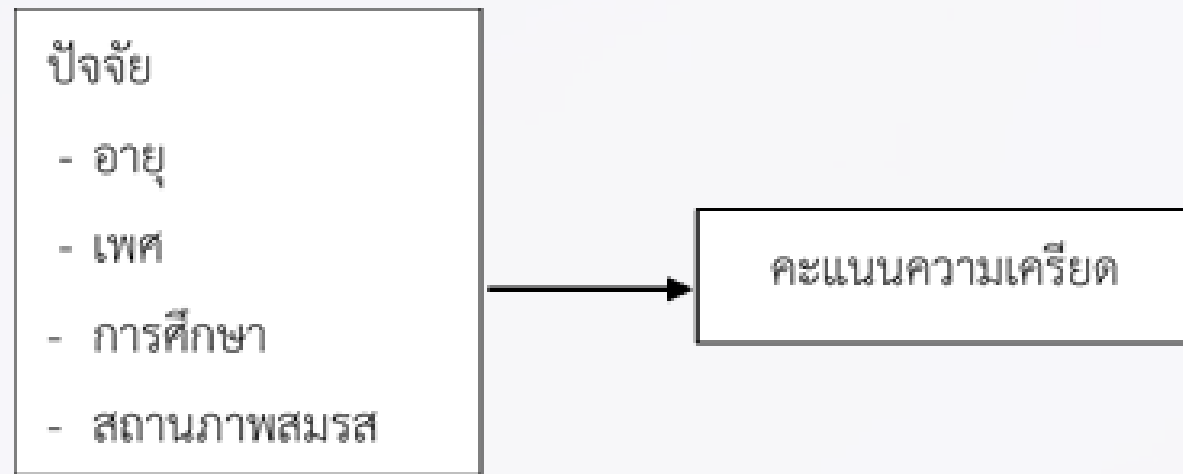
กรณีตัวแปรต้นเป็น Binary response

- 10 cases per predictor

กรณีตัวแปรต้นเป็น categorical

- $(b-1)*10$ cases per predictor โดยที่ b คือ จำนวนการแบ่งกลุ่มตัวแปร เช่น เพศ แบ่งเป็น 2 กลุ่มชายและหญิง จะใช้ตัวอย่าง $(2-1)*10 = 10$ cases

ตัวอย่างการคำนวณขนาดตัวอย่างของ Regression



สูตร $n = \text{cases} * k$, k คือจำนวนปัจจัยที่สนใจ (อ้างอิง Peduzzi P et al. A simulation study of the number of events per variable in logistic regression analysis. Clin Epidemiol. 1996 Dec;49(12):1373-9.)

แทนค่าสูตร $n = 30 + 10 + 30 + 20 = 90$ คน

โดยที่ $n =$ ขนาดตัวอย่างในการศึกษานี้

$k =$ จำนวนตัวแปรอิสระ ซึ่งมีจำนวน 4 ตัว

โดยที่ - ตัวแปรอายุ เก็บเป็น Numerical ; $n=30$ cases

- ตัวแปรเพศ แบ่งเป็น 2 กลุ่ม ชายและหญิง

; $n = (b-1)*10 = (2-1)*10 = 10$ cases

- ตัวแปรการศึกษา แบ่งเป็น 4 กลุ่ม ได้แก่ ประถม มัธยมต้น มัธยมปลาย
ปริญญาตรีขึ้นไป

; $n = (b-1)*10 = (4-1)*10 = 30$ cases

- ตัวแปรสถานภาพสมรส แบ่งออกเป็น 3 กลุ่ม ได้แก่ โสด สมรส หย่าร้าง

; $n = (b-1)*10 = (3-1)*10 = 20$ cases

สูตรการคำนวณขนาดตัวอย่าง เพื่อ Drop out

$$n_{adj} = n / (1 - p)$$

n = ขนาดกลุ่มตัวอย่างที่คำนวณได้

p = สัดส่วนของข้อมูลที่เราคาดว่าจะสูญหาย

จากตัวอย่างการคำนวณขนาดตัวอย่างของ Regression

เพื่อ Drop out 10 % จะได้ขนาดตัวอย่างที่ใช้ในการศึกษานี้เท่ากับ 100 คน

$$n_{adj} = 90 / (1 - 0.1) = 100 \text{ คน}$$

(จรณ์ิต แก้วกังวาน. (2562). คู่มือนักวิจัยมือใหม่: การประยุกต์ใช้ระบาดวิทยา และชีวสถิติในการวิจัยชีวเวชศาสตร์. กรุงเทพฯ: จี.เอส.เอ็ม.เทรดดิ้ง.)