Basic statistics for MT

How to choose appropriate methods for data analysis & sample size determination

Kanokwan Sanchaisuriya

Centre for Research and Development of Medical Diagnostic Laboratory (CMDL), Faculty of Associated Medical Sciences, Khon Kaen University, Thailand

เนื้อหา

ส่วนที่ 1: **Review** ☞ Role of statistics in research

☞ **Variables** in statistics

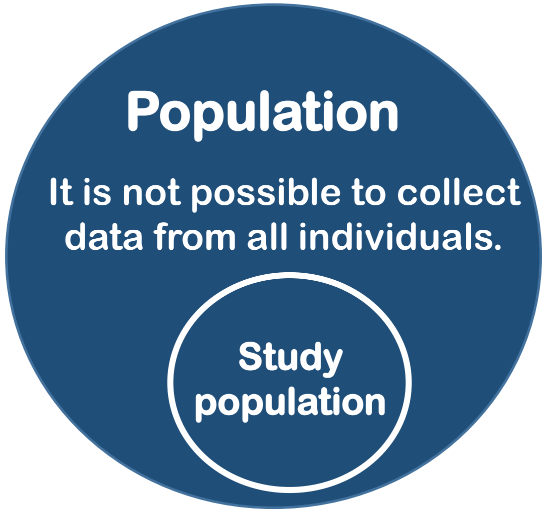
ส่วนที่ 2: Choosing appropriate statistical analysis for medical science research

ส่วนที่ 3: Choosing appropriate sample size formula

Roles of statistics in research

Aim: Use information from a sample of individuals to make some inference about the wider population of like individuals

- **Design**



How many subjects ?

↓

Sample size calculation & sampling technique

Sample should represent the target population; ideally by random sampling ← **not feasible**

- **Subjects' characteristic should be described so that their representative can be judged.**
- **Inclusion/Exclusion criteria**

- **Data Analysis**
- **Data interpretation**

Select appropriate statistics

How to choose appropriate statistics

Issues to consider:

- Type of outcome variable
- Objective of your study
- Number of study group
- Data independence
- Data distribution

วัตถุประสงค์เชิงสถิติ
- ☞ ประมาณค่า
- ☞ เปรียบเทียบ
- ☞ ความสัมพันธ์/ทำนาย

Outcome variable is "what the researcher is measuring or observing".

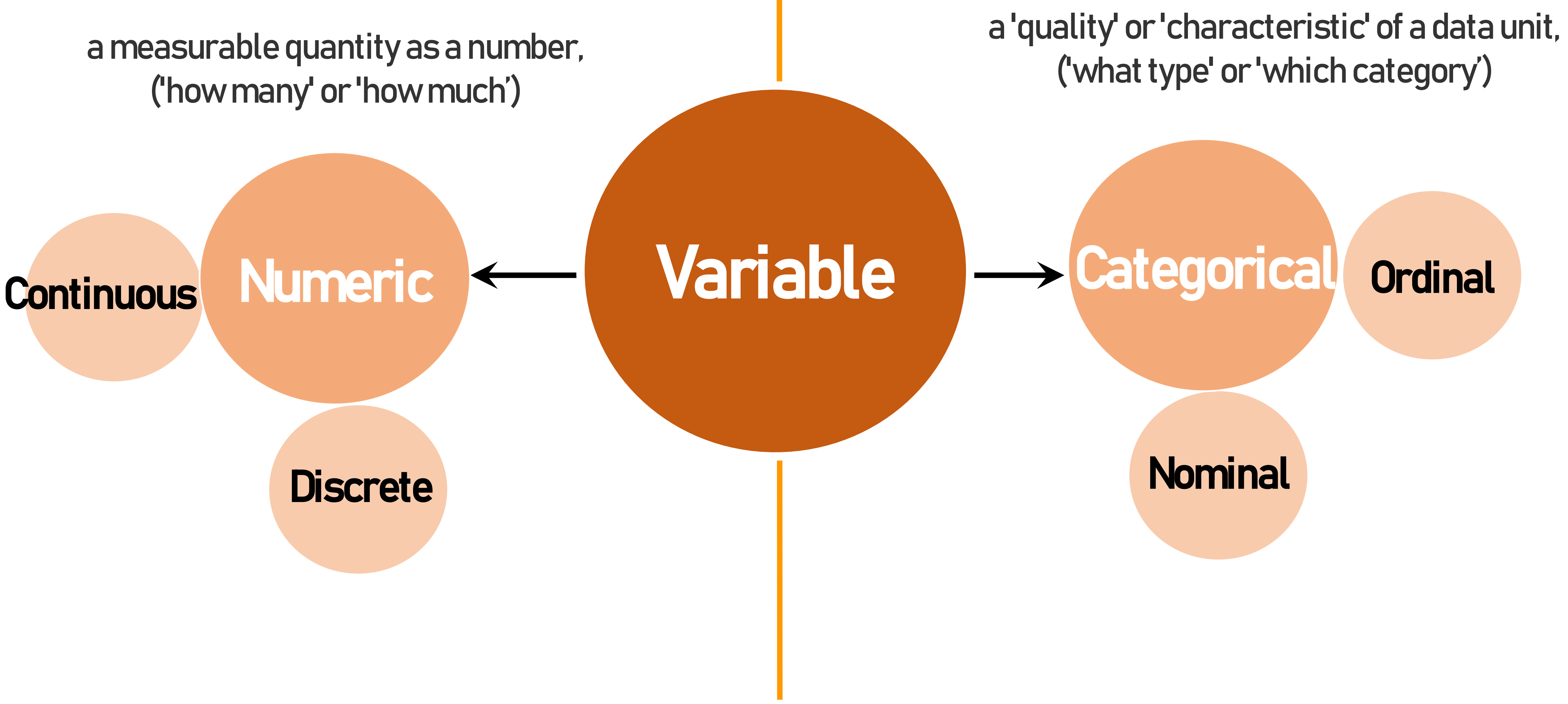
Variables in statistics

ตัวแปร

No.	Age	Sex	Ane	Rbc	Hb	MCV	MCH	RDW	SF	CRP	Hb-type	HbA2/E	α^0
1	29	F	Y	4.25	11.9	81.2	28.1	18.1	34.6	n	A2A	2.9	n
2	15	F	Y	3.88	11.3	83.5	29.1	18.2	47.1	P	EA	28	n
3	30	M	Y	4.35	11.6	80.9	26.6	18.4	56.8	n	A2A	3.1	n
4	32	M	Y	4.97	10.7	65.0	21.5	20.2	58.5	n	A2AH	1.3	P
5	24	F	N	4.74	12.3	80.6	25.9	17.5	229.4	n	EA	26	n
6	50	F	Y	5.49	10.4	59.7	18.9	21.2	17.4	n	A2A	2.1	P
7	1	F	Y	3.6	10.8	85.5	30.1	17.4	179.6	P	A2A	3.2	n
8	68	M	Y	3.63	10.9	86.0	29.9	17.8	48.3	n	A2A	2.7	n
9	30	M	N	4.21	14.3	83.1	29.1	16.7	48.8	n	A2A	2.9	n
10	17	F	Y	4.04	11.2	82.4	27.7	17.7	60.8	n	A2A	3.0	n

A variable is any characteristic, number, or quantity that can be measured or counted.

Types of variables

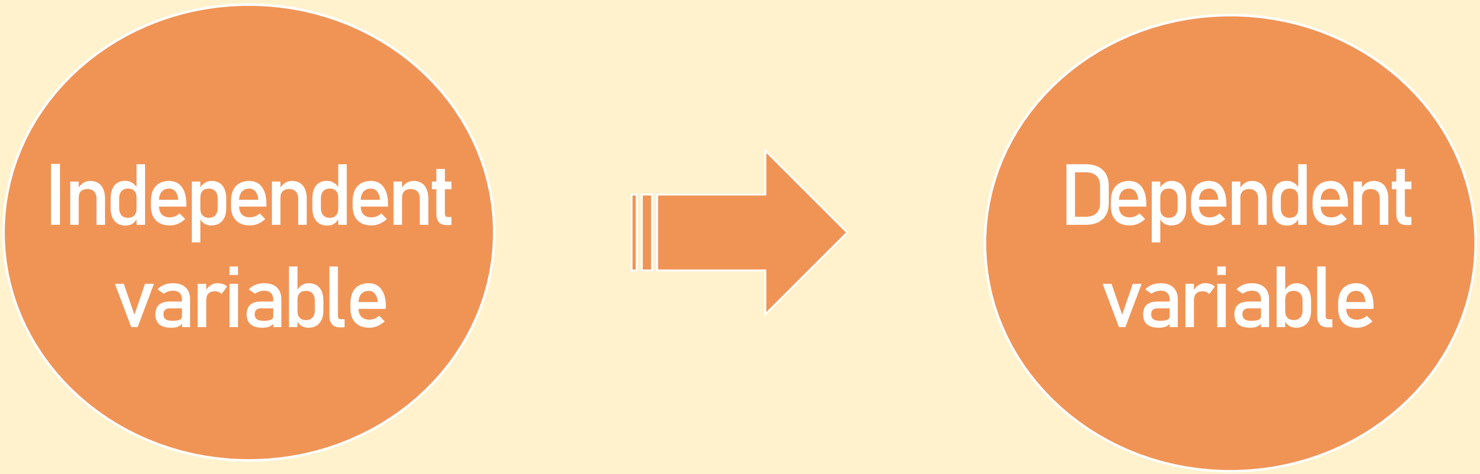


No.	Age	Sex	Ane	Rbc	Hb	MCV	MCH	RDW	SF	CRP	Hb-type	HbA2/E	α^0
1	29	F	Y	4.25	11.9	81.2	28.1	18.1	34.6	n	A2A	2.9	n
2	15	F	Y	3.88	11.3	83.5	29.1	18.2	47.1	P	EA	28	n
3	30	M	Y	4.35	11.6	80.9	26.6	18.4	56.8	n	A2A	3.1	n
4	32	M	Y	4.97	10.7	65.0	21.5	20.2	58.5	n	A2AH	1.3	**P**
5	24	F	N	4.74	12.3	80.6	25.9	17.5	229.4	n	EA	26	n
6	50	F	Y	5.49	10.4	59.7	18.9	21.2	17.4	n	A2A	2.1	**P**
7	1	F	Y	3.6	10.8	85.5	30.1	17.4	179.6	P	A2A	3.2	n
8	68	M	Y	3.63	10.9	86.0	29.9	17.8	48.3	n	A2A	2.7	n
9	30	M	N	4.21	14.3	83.1	29.1	16.7	48.8	n	A2A	2.9	n
10	17	F	Y	4.04	11.2	82.4	27.7	17.7	60.8	n	A2A	3.0	n

- Continuous variables: Age, Rbc, Hb, MCV, MCH, RDW, SF
- Categorical variables: Sex, Ane, SF, CRP, Hb-type, HbA2/E, α^0

Variables in studies dealing with <u>risk factor/prediction or experiment</u>

<u>Dependent</u> vs <u>independent</u> variable



The variable that is changed or controlled. ☞ **X**

The variable being tested or measured ☞ **Y** (outcome variable)

Questions to ask before choosing statistics

- What is **<u>the main objective?</u>** (in term of statistical point of view)
  - ☞ *Estimation*
  - ☞ *Comparison*
  - ☞ *Relationship/Prediction*
- What is **the <u>main outcome variable</u>?**
- What is **the <u>type of outcome variable</u>?**

แบบฝึกหัดที่ 1

วัตถุประสงค์เชิงสถิติ

ตัวแปรผลหลัก และ ประเภทตัวแปร

โครงการวิจัย	วัตถุประสงค์	ตัวแปรผลหลัก	ประเภทตัวแปร
1. อัตราการเกิดภาวะไตวายในผู้ป่วยเบาหวานชนิดที่ 2 ที่มารับบริการที่โรงพยาบาลรัฐ			
2. ความสัมพันธ์ระหว่างอายุครรภ์กับภาวะเลือดจางในหญิงตั้งครรภ์			
3. ค่าคะแนนความเครียดในนักศึกษาหญิงและนักศึกษาชาย มหาวิทยาลัยขอนแก่น			
4. ผลของอุณหภูมิต่อระยะเวลาการแข็งตัวของเลือด			
5. ความสัมพันธ์ระหว่างค่า BMI กับระดับ cholesterol			

ชื่อโครงการ	วัตถุประสงค์	ตัวแปรผลหลัก	ประเภทตัวแปร
6. UGT polymorphisms ในทารกแรกคลอดที่มีภาวะตัวเหลือง			
7. ความสัมพันธ์ระหว่าง UGT polymorphisms กับภาวะตัวเหลืองในทารกแรกคลอด			
8. ประสิทธิผลการใช้สารสกัดจากเปลือกมังคุดในการลดความดันโลหิต			
9. ปัจจัยที่สัมพันธ์กับการเกิดโรคมะเร็งท่อน้ำดี			
10. ประสิทธิผลค่ายปรับเปลี่ยนพฤติกรรมผู้ป่วยโรคเบาหวานที่ไม่สามารถควบคุมระดับน้ำตาลได้			

Choosing appropriate statistical analysis for medical science research

1. Continuous outcome variable

วัตถุประสงค์: <u>อนุมานค่าเฉลี่ย</u>ในประชากร ☛ <u>Normally distributed data</u>

One group: <u>Mean</u> with 95% Confidence interval (one-sample t-test)

วัตถุประสงค์: <u>เปรียบเทียบค่าเฉลี่ย</u>ระหว่างกลุ่ม

- 2 กลุ่ม
  - อิสระ → Independent t-test
  - ไม่อิสระ → Paired t-test
- มากกว่า 2 กลุ่ม
  - อิสระ → One-way ANOVA
  - ไม่อิสระ → Repeated measure ANOVA

วัตถุประสงค์: <u>อนุมานค่าเฉลี่ย</u>ในประชากร ☛ <u>Non-normally distributed data</u>

- <u>Median</u> with 95% CI (one-sample Sign or one-sample Wilcoxon)

การใช้งาน: <u>เปรียบเทียบค่า median</u>

2 กลุ่ม

- อิสระ → Mann-Whitney U test
- ไม่อิสระ → Wilcoxon-signed rank test

มากกว่า 2 กลุ่ม

- อิสระ → Kruskal-Wallis test
- ไม่อิสระ → Friedman test

วัตถุประสงค์: <u>ความสัมพันธ์</u>

<u>1.1 หาขนาดความสัมพันธ์</u>

- ☛ Pearson correlation; r (95% CI)
- ☛ Spearman rank correlation; r (95% CI)

<u>1.2 ทำนาย (prediction/risk factor)</u>

Y = 1; X = 1 (Bivariate analysis)	Y = 1; X > 1 (Multivariate analysis)
↓	↓
Simple linear regression	Multiple linear regression

Kendall–Theil Sen Siegel nonparametric linear regression

Relation between two continuous variables

Assumptions for Pearson's correlation

- ☞ Two variables are observed on a random sample of individuals.
- ☞ At least one of the variables have a normal distribution.
- ☞ No outliers in either variable

Common misuses of correlation

- ☞ **No background knowledge** on the link between two variables
  (large number of variables, then pick out just those which are statistically significant.)
- ☞ Mixed samples (different subgroups)
- ☞ Assessing agreement (one variable measured with 2 methods)

Misuse of correlation is so common that some statisticians have wished that the method had never been devised (Douglas G. Altman).

Understanding Bland Altman analysis

Davide Giavarina

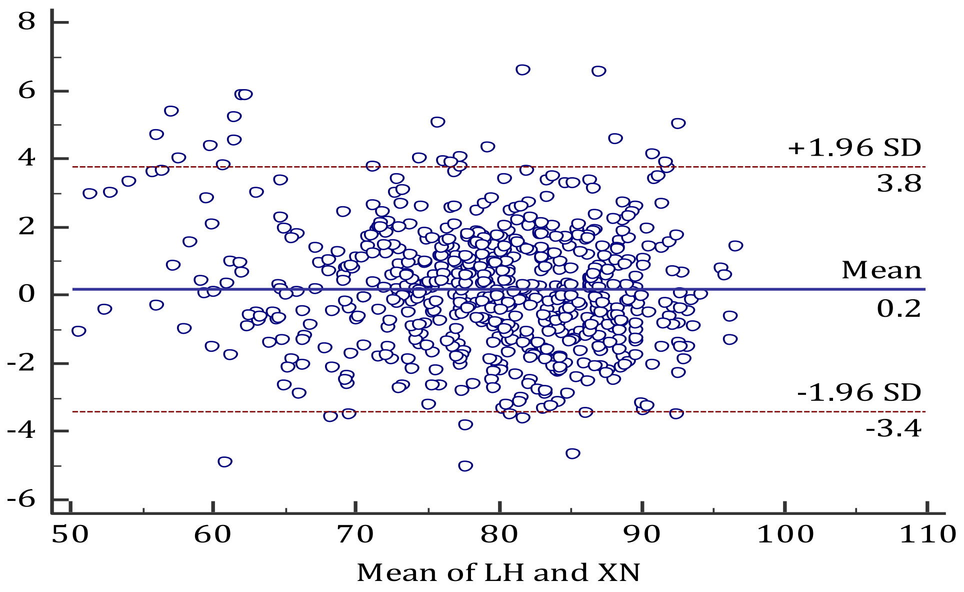
Biochemia Medica 2015;25(2):141–51

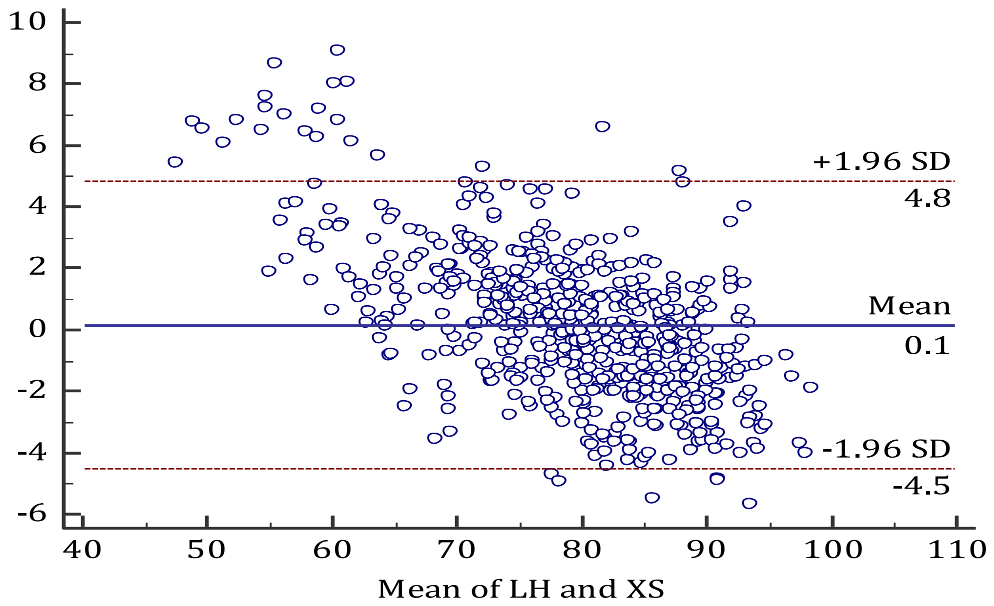
Summary and highlights

If you want to evaluate whether the differences between two measurements of the same substance are significant, study the differences, not the agreement. The correlation between methods is always misleading and should not be used for assessing the method comparability. The B&A plot analysis is a simple way to evaluate a bias between the mean differences, and to estimate an agree-

Bland-Altman plot for method comparison study

คำถามวิจัย: ค่า RBC indices ที่ตรวจวัดด้วยเครื่องวิเคราะห์เม็ดเลือดอัตโนมัติ 2 เครื่อง แตกต่างกันหรือไม่?





Chaitriphop et al. Clin Lab, 2016.

2. Categorical outcome variable

วัตถุประสงค์: อนุมานค่าสัดส่วนในประชากร

- ☛ Proportion (percent) with 95% Confidence interval (Z-test)

วัตถุประสงค์: เปรียบเทียบสัดส่วนระหว่างกลุ่ม

2 กลุ่ม

อิสระ

(Z-test or χ^2)

ไม่อิสระ

(Z-test or McNemar χ^2)

มากกว่า 2 กลุ่ม

อิสระ

χ^2 test

ไม่อิสระ

Generalized estimating equations: GEEs

วัตถุประสงค์: ความสัมพันธ์หรือปัจจัยเสี่ยง

Y = 1; X = 1

(Bivariate analysis)

อิสระ

RR or OR (95% CI);

p-value -> χ^2

ไม่อิสระ

OR (95% CI)

p-value -> Mcnemar χ^2

Y = 1; X > 1

(Multivariate analysis)

อิสระ

Unconditional logistic regression;

p–value –> χ^2

ไม่อิสระ

Conditional logistic regression;

p–value –> Mcnemar χ^2

Odds ratio; OR (95% CI)

Choosing appropriate sample size formula

Key concepts

- There are several formulas for sample size estimation.
- **One formula cannot be used in all designs.**
- Different designs (objectives) need different methods of sample size calculation

Issues to consider:

- Objective of your study
- Type of <u>outcome variable</u>
- Population size (finite or infinite population)
- Number of study group
- Data independence
- Data distribution

วัตถุประสงค์เชิงสถิติ

- ☞ ประมาณค่า
- ☞ เปรียบเทียบ
- ☞ ความสัมพันธ์/การทำนาย

Objective ☛ **Estimation**

Outcome variable	**Population size (N)**	**Sample size formula**; for
Categorical	Known →	Finite population **proportion**
	Not known →	Infinite population **proportion**
Continuous	Known →	Finite population **mean**
	Not known →	Infinite population **mean**

Objective ☛ **Comparison between 2 groups**

Outcome variable	**Data independence**	**Sample size formula; for**
Categorical	→ Dependence →	Two dependent **proportions**
	→ Independence →	Two independence **proportions**
Continuous	→ Dependence →	Two dependent **means**
	→ Independence →	Two independence **means**

การประมาณค่าสัดส่วน

กรณีทราบขนาดประชากร

$$n = \frac{Nz^2_{\alpha/2}p(1-p)}{e^2(N-1)+Z^2_{\alpha/2}p(1-p)}$$

N = ขนาดประชากร

Z = ค่าสถิติ z ที่ระดับ α ที่กำหนด

p = ค่าสัดส่วนของตัวแปรผลหลักในงานวิจัยที่เคยศึกษามาก่อนหน้านี้

e = ค่าความแม่นยำของการประมาณค่า

กรณีไม่ทราบขนาดประชากร

$$n = \frac{z^2_{\alpha/2}p(1-p)}{e^2}$$

Z = ค่าสถิติ z ที่ระดับ α ที่กำหนด

p = ค่าสัดส่วนของตัวแปรผลหลักในงานวิจัยที่เคยศึกษามาก่อนหน้านี้

e = ค่าความแม่นยำของการประมาณค่า

ที่มา: พงษ์เดช สารการ. เอกสารประกอบการสอน เรื่องการคำนวณขนาดตัวอย่าง. ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น.

การประมาณค่าเฉลี่ย

กรณีทราบขนาดประชากร

$$n = \frac{Nz^2_{\alpha/2}\sigma^2}{e^2(N-1)+Z^2_{\alpha/2}\sigma^2}$$

N = ขนาดประชากร

Z = ค่าสถิติ z ที่ระดับ α ที่กำหนด

σ^2 = ค่าความแปรปรวนของตัวแปรผลหลักในงานวิจัยที่เคยศึกษามาก่อนหน้านี้

e = ค่าความแม่นยำของการประมาณค่า

กรณีไม่ทราบขนาดประชากร

$$n = \frac{Z^2_{\alpha/2}\sigma^2}{e^2}$$

Z = ค่าสถิติ z ที่ระดับ α ที่กำหนด

σ^2 = ค่าความแปรปรวนของตัวแปรผลหลักในงานวิจัยที่เคยศึกษามาก่อนหน้านี้

e = ค่าความแม่นยำของการประมาณค่า

ที่มา: พงษ์เดช สารการ. เอกสารประกอบการสอน เรื่องการคำนวณขนาดตัวอย่าง. ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น.

การเปรียบเทียบค่าสัดส่วนระหว่าง 2 กลุ่ม

กรณีอิสระ

$$n/group = \frac{2p(1-p)(Z_\alpha + Z_\beta)^2}{(p_1 - p_2)^2}$$

Z_α = ค่าสถิติ z ที่ระดับ α ที่กำหนด

Z_β = ค่าสถิติ z ที่ระดับ β ที่กำหนด

p_1 = ค่าสัดส่วนกลุ่มที่ 1

p_2 = ค่าสัดส่วนกลุ่มที่ 2

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

กรณีไม่อิสระ

$$n/group = \frac{\left[\left(Z_\alpha\sqrt{\frac{b}{c}+1}\right) + \left(Z_\beta\sqrt{\left(\frac{b}{c}+1\right) - p\left(\frac{b}{c}-1\right)^2}\right)\right]^2}{p\left(\frac{b}{c}-1\right)^2}$$

Posttest

Pretest		+	-
	+	a	b
	-	c	d

$p = \frac{b}{n}$ เมื่อ $b < c$

$p = \frac{c}{n}$ เมื่อ $c < b$

ที่มา: พงษ์เดช สารการ. เอกสารประกอบการสอน เรื่องการคำนวณขนาดตัวอย่าง. ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น.

การเปรียบเทียบค่าเฉลี่ยระหว่าง 2 กลุ่ม

กรณีอิสระ

$$\boxed{n/group = \frac{2\sigma^2(Z_\alpha + Z_\beta)^2}{(\mu_1 - \mu_2)^2}}$$

Z_α = ค่าสถิติ z ที่ระดับ α ที่กำหนด

Z_β = ค่าสถิติ z ที่ระดับ β ที่กำหนด

$$\left.\begin{array}{l}\mu_1 = \text{ค่าเฉลี่ยกลุ่มที่ 1}\\ \mu_2 = \text{ค่าเฉลี่ยกลุ่มที่ 2}\end{array}\right\} \mu_1 - \mu_2 \quad \text{(Effect size)}$$

σ^2 = ค่าความแปรปรวนของงานวิจัยที่ผ่านมา

กรณีไม่อิสระ

$$\boxed{n/group = \frac{\sigma_d^2(Z_\alpha + Z_\beta)^2}{\mu_d^2}}$$

Z_α = ค่าสถิติ z ที่ระดับ α ที่กำหนด

Z_β = ค่าสถิติ z ที่ระดับ β ที่กำหนด

μ_d = ค่าเฉลี่ยของผลต่างในงานวิจัยที่ผ่านมา

σ_d^2 = ค่าความแปรปรวนของผลต่างในงานวิจัยที่ผ่านมา

ที่มา: พงษ์เดช สารการ. เอกสารประกอบการสอน เรื่องการคำนวณขนาดตัวอย่าง. ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น.

ประเด็นที่ต้องเข้าใจและวิเคราะห์จากโจทย์วิจัยให้ได้

1. วัตถุประสงค์ ในเชิงสถิติ
   - ☞ ประมาณ (อนุมาณ) ค่า/เปรียบเทียบ/หาความสัมพันธ์
2. อะไรคือ ตัวแปรผลหลักของการวิจัย (คำถามวิจัย)
3. ประเภทของตัวแปรผลหลัก (ข้อมูลที่เก็บ/วิเคราะห์)
   - ☞ แจงนับ/ต่อเนื่อง
4. ความเป็นอิสระของตัวแปร (ไม่อิสระ -> วัดซ้ำ)

Others ☛ search from the internet or consult statistical software

Use 'statistical test' as a keyword; e.g.

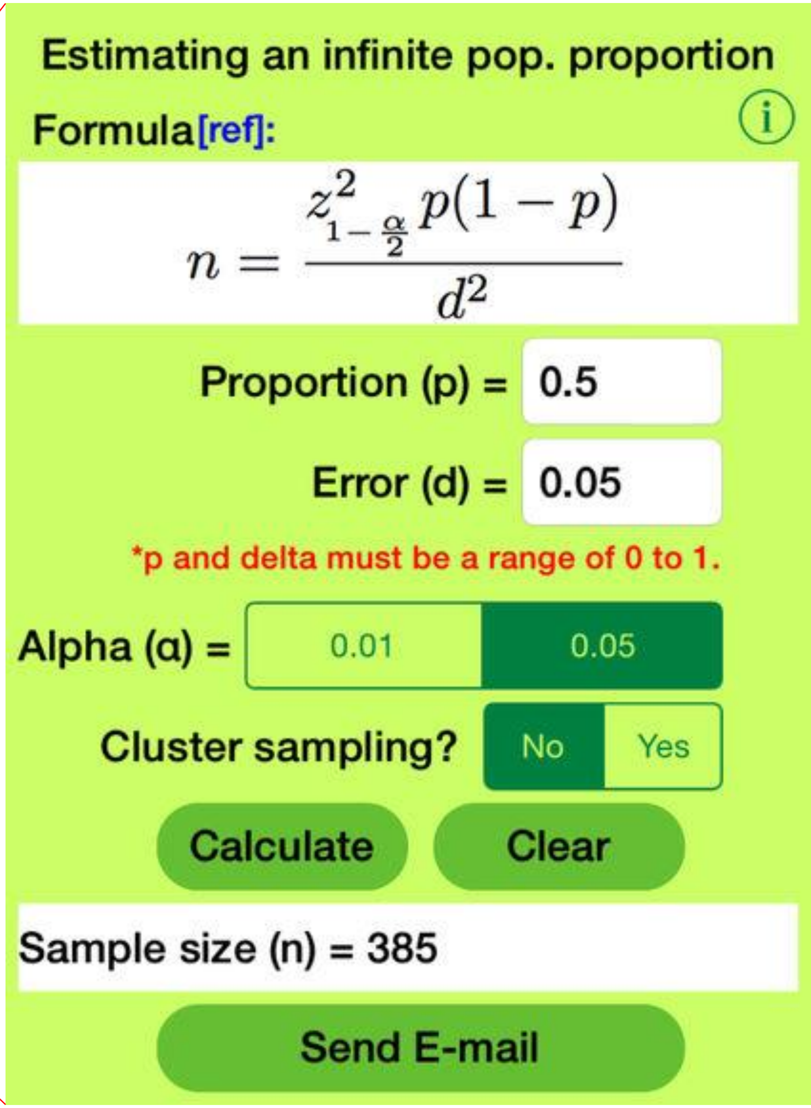
- Sample size for **ROC comparison**
- Sample size for **correlation coefficient**
- Sample size for **ANOVA**
- Sample size for **Chi-square test**
- .... etc. ....

Use 'study design' as a keyword; e.g.

- Sample size for **method comparison**
- Sample size for **case-control study**
- .... etc. ....

n4Studies: Sample Size Calculation for an Epidemiological Study on a Smart Device

Chetta Ngamjarus



แบบฝึกหัดที่ 2

วัตถุประสงค์การวิจัยในเชิงสถิติ

และการเลือกสถิติทดสอบ

โครงการวิจัย	ตัวแปรผลหลัก	ประเภทตัวแปร	วัตถุประสงค์	สถิติทดสอบ
1. อัตราการเกิดภาวะไตวายในผู้ป่วยเบาหวานชนิดที่ 2 ที่มารับบริการที่โรงพยาบาลรัฐ				
2. ความสัมพันธ์ระหว่างอายุครรภ์กับภาวะเลือดจางในหญิงตั้งครรภ์				
3. ค่าคะแนนความเครียดในนักศึกษาหญิงและนักศึกษาชาย มข.				
4. ผลของอุณหภูมิต่อระยะเวลาการแข็งตัวของเลือด				
5. ความสัมพันธ์ระหว่างค่า BMI กับ cholesterol				

โครงการวิจัย	ตัวแปรผลหลัก	ประเภทตัวแปร	วัตถุประสงค์	สถิติทดสอบ
6. UGT polymorphisms ในทารกแรกคลอดที่มีภาวะตัวเหลือง				
7. ความสัมพันธ์ระหว่าง UGT polymorphisms กับ ภาวะตัวเหลือง				
8. ประสิทธิผลการใช้สารสกัดจากเปลือกมังคุดในการลดความดันโลหิต				
9. ปัจจัยที่สัมพันธ์กับการเกิดโรคมะเร็งท่อน้ำดี				
10. ประสิทธิผลค่ายปรับเปลี่ยนพฤติกรรมผู้ป่วยโรคเบาหวานที่ไม่สามารถควบคุมระดับน้ำตาลได้	*	*	*	*

*ผู้วิจัยต้องการวัดผลด้วยอะไร? (กำหนดวัตถุประสงค์ที่จำเพาะอย่างไร?)

Selected references

1. Altman DG. Practical statistics for medical research. London: Chapman & Hall/CRC, 1991.
2. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015; 25(2):141-51.
3. Statistical methods.https://www.healthknowledge.org.uk/public-health-textbook/ , Accessed 21/06/2019.
4. Charan J, Biswas T. How to calculate sample size for different study designs in medical research. Indian J Psychol Med 2013; 35: 121–126.
5. Choosing Between a Nonparametric Test and a Parametric Test. https://blog.minitab.com/ Accessed 21/06/2019.
6. Ngamjarus C, Chongsuvivatwong V, McNeil E. n4Studies: Sample Size Calculation for an Epidemiological Study on a Smart Device. Siriraj Med J 2016; 68: 160-170.
7. บัณฑิต ถิ่นคำรพ. คู่มือปฏิบัติการชีวสถิติ. ขอนแก่น: โรงพิมพ์คลังนานาวิทยา. 2543.
8. พงษ์เดช สารการ. เอกสารประกอบการสอน เรื่องการคำนวณขนาดตัวอย่าง. ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น. 2560

Thank you for paying attention