Medical statistics

(QUANTITATIVE METHOD V)

หลักสูตรเวชศาสตร์ชุมชน

RACM 302

1

สถิติ (Statistics)

- วิชาสถิติเป็นวิชาที่เกี่ยวกับวิธีการรวบรวมข้อมูล จัดการ วิเคราะห์ นำเสนอ แปลผล และหาข้อสรุปจากข้อมูล

2

สถิติเชิงพรรณนา (Descriptive statistics)

- บรรยายและสรุปลักษณะของข้อมูล

เช่น ประชากรในชุมชนนี้ มีน้ำหนัก ส่วนสูง สัดส่วนเพศชายหญิง ระดับความดันเลือดฯลฯ เป็นเท่าไร

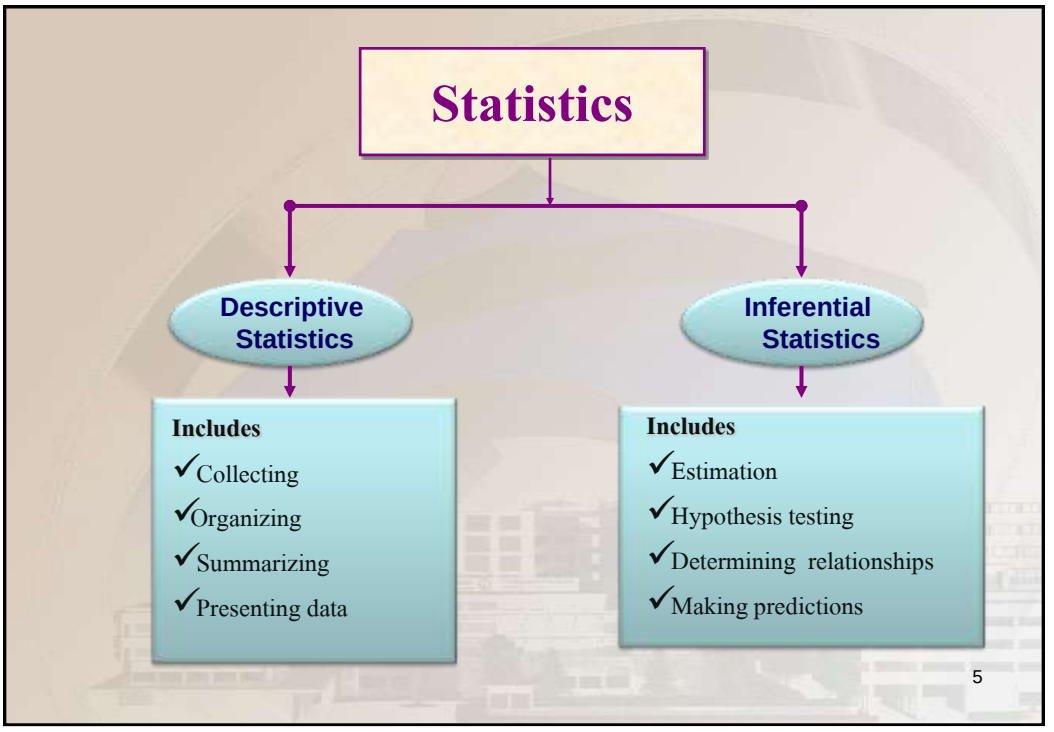
3

สถิติเชิงอนุมาน (Inferential statistics)

- หาข้อสรุปของประชากร จากข้อมูลที่เก็บได้ในกลุ่มตัวอย่าง

เช่น สุ่มตัวอย่าง นศพ.ชาย 50 คน ได้ส่วนสูงเฉลี่ยเท่ากับ 170 ซม. เพื่อต้องการหาส่วนสูงเฉลี่ยใน นศพ. ชายทั้งหมดว่า เป็นเท่าไร

4



Parameter & Statistic

- Parameter คือค่าของประชากรทั้งหมด
- Statistic คือค่าที่หาได้ในกลุ่มตัวอย่าง

6

ตัวแปร (Variable)

- หมายถึงลักษณะ (characteristic) ของข้อมูล เช่น อายุ เพศ น้ำหนัก ระดับน้ำตาล ความดันเลือด การหายหรือตายจากโรค ฯลฯ

7

ตัวแปร (**variable)**

- Qualitative (Discrete)
  - nominal
  - ordinal
- Quantitative (continuous)
  - interval
  - ratio

8

Scale of measurement

- **Nominal variable** (nominal = name) :

  บอกลักษณะ เช่น blood group, sex, cause of death

- **Ordinal variable** (ordinal = order) :

  บอกลำดับ เช่น severity of disease

(mild, moderate, severe), การสูบบุหรี่

(<= 10 มวน, 11-20 มวน, >20 มวนต่อวัน)

9

Scale of measurement

- **Ratio variable** :

  วัดต่อเนื่องได้ถึงทศนิยมไม่รู้จบ มีศูนย์ที่แท้จริง เช่น น้ำหนัก ส่วนสูง ระดับไขมัน องศาเคลวิน

- **Interval variable** :

  วัดต่อเนื่องได้ แต่มีศูนย์ที่ไม่แท้จริง เช่น อุณหภูมิ องศาเซลเซียส

10

Descriptive Statistics

- การวัดแนวโน้มเข้าสู่ส่วนกลาง

  (Measure of central tendency)

  – เป็นการบอกค่ากลางของชุดข้อมูลว่าเป็นเท่าใด

  – ตัวชี้วัดที่นิยมใช้ได้แก่ Mean, Median และ Mode

11

คำจำกัดความ

Mean = ค่าเฉลี่ยของชุดข้อมูล $\bar{x} = \frac{\sum x_i}{n}$

Median = ค่าที่อยู่กึ่งกลางของข้อมูลที่เรียงลำดับ

Mode = ค่าที่มีความถี่ (พบ) มากที่สุด

12

การวัดความแปรปรวน

(Measure of variability)

- เป็นการบอกถึงการกระจายตัวของชุดข้อมูลว่ามีมากน้อยเพียงไร
- ตัวชี้วัดที่นิยมใช้ได้แก่ Range, Standard deviation, Variance และ Interquartile range

13

คำจำกัดความ

Range = ค่ามากสุด ลบ ค่าน้อยสุด

Variance = $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

S.D. = Square root ของ variance

Interquartile range = $Q_3 - Q_1$ หรือ $P_{75} - P_{25}$

14

Coefficient of Variation (C.V.)

ตาราง เปรียบเทียบส่วนสูง ของ นักเรียนชั้นประถมปีที่ 6 และมัธยมศึกษาปีที่ 6

	ประถมปีที่ 6	มัธยมศึกษาปีที่ 6
อายุ (เฉลี่ย)	11	17
ส่วนสูง (เฉลี่ย)	140	170
S.D.	10	10

15

Coefficient of Variation (C.V.)

- สรุปไม่ได้
- น่าจะเทียบความแปรปรวนกับค่าเฉลี่ย

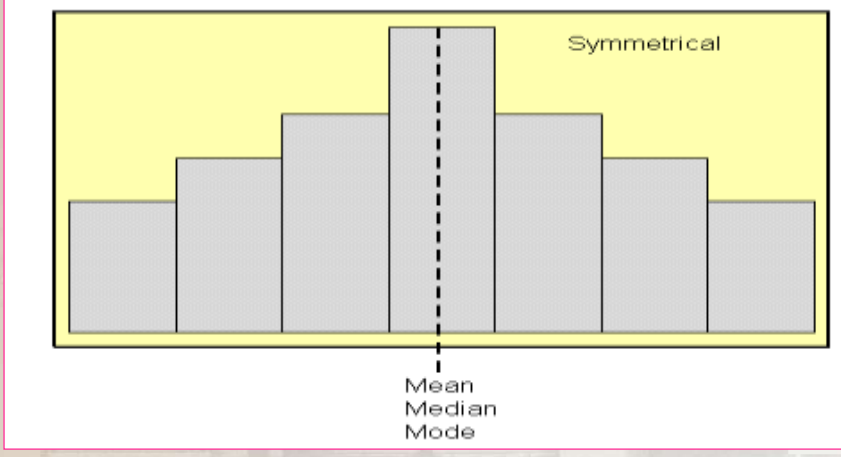
$$C.V. = \frac{S.D.}{\bar{X}} \; x \; 100\%$$

$$\therefore C.V. \text{ ประถมปีที่} 6 = \frac{10}{140} \times 100\% = 7.1\%$$

$$C.V. \text{ มัธยมปีที่ } 6 = \frac{10}{170} \times 100\% = 5.9\%$$

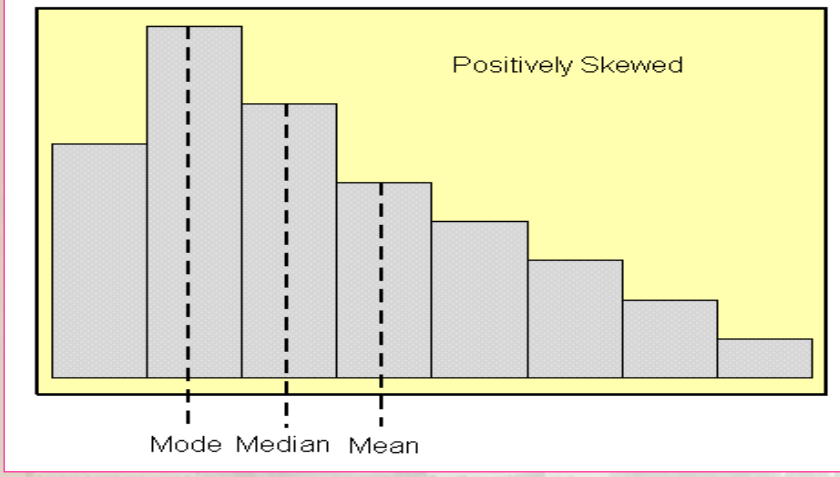
16

Distribution



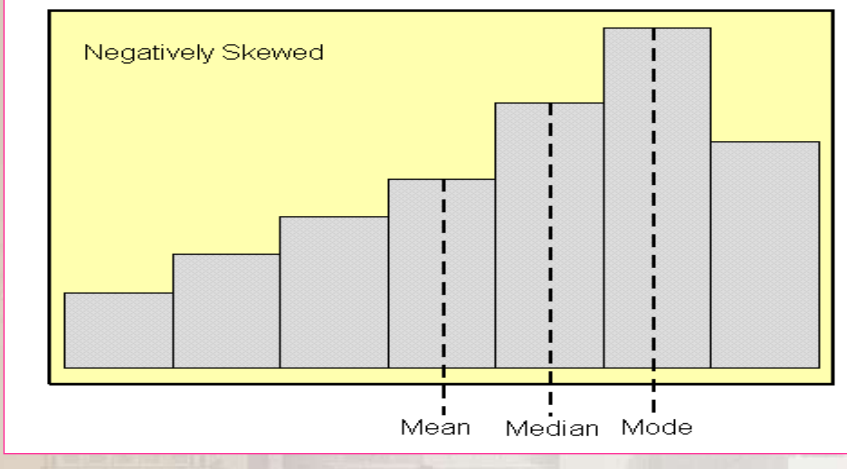
17

Distribution



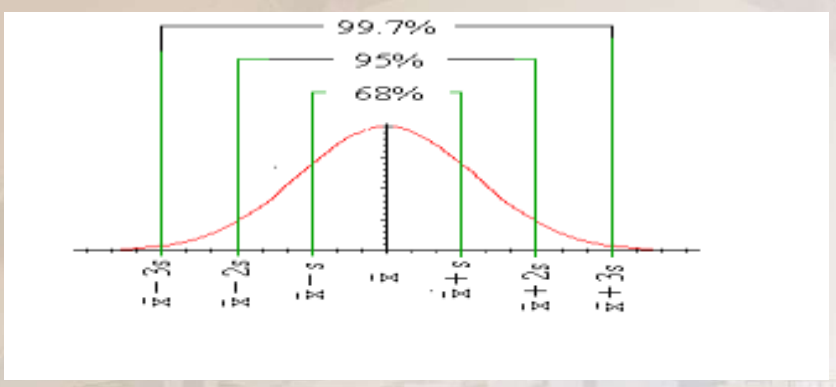
18

Distribution



19

การแปลผลค่า S.D.



$(\bar{x}-s, \bar{x}+s)$ contains approximately 68% of the measurements

$(\bar{x}-2s, \bar{x}+2s)$ contains approximately 95% of the measurements

$(\bar{x}-3s, \bar{x}+3s)$ contains virtually all of the measurements

20

Inferential statistics

- **Estimation** เป็นการประมาณค่าในประชากรจากข้อมูลในกลุ่มตัวอย่าง เช่น ความดันเลือดสูงในกลุ่มประชาชนอายุตั้งแต่ 40 ปีขึ้นไปในชุมชนเป็นเท่าไร

21

Estimation

- **การศึกษาในกลุ่มตัวอย่างย่อมได้ค่าในกลุ่มตัวอย่าง (sample estimate) จึงต้องหาค่าในประชากร (population estimate) ว่าเป็นเท่าใด**
- **วิธีการหาค่า population estimate ได้แก่ การหาค่า confidence interval**

22

การแปลผล confidence interval

Prevalence ของ HT ในชุมชน ก. เท่ากับ 30% (95% C.I. = 27%-33%) แปลผลว่า

Prevalence ของ HT ในชุมชน ก. น่าจะมีค่าอยู่ระหว่าง 27% - 33% ด้วยความเชื่อมั่น 95%

Sample estimate = 30%

Population estimate = 27% – 33%

23

Inferential statistics

- Hypothesis testing เป็นการพิสูจน์สมมติฐานว่ามีความสัมพันธ์ระหว่างตัวแปรหรือไม่

  เช่น มีความแตกต่างกันระหว่างกลุ่มหญิงและชายหรือไม่ในการมีภาวะความดันเลือดสูง

24

Hypothesis testing

- การศึกษาหรือวิจัยมักจะมีการตั้งสมมติฐานเพื่อทดสอบ
- สมมติฐานที่ตั้งไว้โดยทั่วไปจะมี 2 สมมติฐาน ได้แก่

1. Null hypothesis หรือสมมติฐานของความไม่แตกต่าง เช่น ไม่มีความแตกต่างของ ความชุกสำหรับภาวะ HT ระหว่างประชากรชายและหญิง (difference = 0)
2. Alternative hypothesis หรือสมมติฐานทางเลือก เช่น มีความแตกต่างของ ความชุกสำหรับภาวะ HT ระหว่างประชากรชายและหญิง (difference ไม่เท่ากับ 0)

25

ขั้นตอนของ hypothesis testing

1. ตั้งสมมติฐาน
2. กำหนดเกณฑ์ต่างๆ เช่น ระดับนัยสำคัญ
3. เก็บข้อมูลจากกลุ่มตัวอย่าง
4. หาค่าสถิติ (test statistics)
5. คำนวณค่า p (probability value)
6. สรุปรับหรือปฏิเสธสมมติฐาน

26

ขั้นตอนของ hypothesis testing

1. ตั้งสมมติฐาน
2. กำหนดเกณฑ์ต่างๆ เช่น ระดับนัยสำคัญ
3. เก็บข้อมูลจากกลุ่มตัวอย่าง
4. หาค่าสถิติ (test statistics)
5. คำนวณค่า p (probability value)
6. สรุปรับหรือปฏิเสธสมมติฐาน

27

ตารางเปรียบเทียบผลการตรวจคัดกรองความดันเลือดในกลุ่มชาย-หญิงอายุตั้งแต่ 40 ปีขึ้นไป (observed table)

เพศ	ความดันเลือดสูง	ความดันเลือดปกติ	รวม
ชาย	9	41	50
หญิง	21	29	50
รวม	30	70	100

Chi-Square **P-value**

28

$$\sum (o - e) = 0$$

$$Chi - Square = \sum \frac{(O - E)^2}{E}$$

Chi-Square = 3.84 p-value 0.05

degree of freedom=1

29

ตารางเปรียบเทียบผลการตรวจคัดกรองความดันเลือด
ในกลุ่มชาย-หญิงอายุตั้งแต่ 40 ปีขึ้นไป

(expected table)

เพศ	ความดันเลือดสูง		ความดันเลือดปกติ	รวม
ชาย	15	a b	35	50
หญิง	15	c d	35	50
รวม	30		70	100

Chi-Square **P-value**

30

ตารางเปรียบเทียบผลการตรวจคัดกรองความดันเลือด ในกลุ่มชาย-หญิงอายุตั้งแต่ 40 ปีขึ้นไป

เพศ	ความดันเลือดสูง	ความดันเลือดปกติ	รวม
ชาย	9	41	50
หญิง	21	29	50
รวม	30	70	100

Chi-Square 6.86 **P-value 0.0088**

31

การแปลผล hypothesis testing

- ระดับนัยสำคัญตั้งไว้ที่ระดับใด

  .05 หรือ .01 หรือ .1

- คำนวณค่า p ได้เท่าไร

  น้อยกว่าระดับนัยสำคัญที่ตั้งไว้หรือไม่

- ถ้าน้อยกว่า แสดงว่ามีนัยสำคัญทางสถิติ

32

การแปลผล hypothesis testing

- ถ้าเท่ากับหรือมากกว่า ก็ไม่มีนัยสำคัญทางสถิติ
- P-value คือ โอกาสที่จะได้ค่าสถิติตามที่พบในกลุ่มตัวอย่างหรือมากกว่าถ้า null hypothesis เป็นความจริง

33

การแปลผล

ถ้าความชุกของความดันเลือดสูงในเพศชายและหญิงไม่แตกต่างกัน โอกาสจะเห็นความแตกต่างเช่นนี้หรือมากกว่าโดยบังเอิญมีค่า น้อยกว่า 0.0088 หรือ .88%

เราจึงเลือกสรุปว่าแตกต่างกัน

เพราะโอกาสสรุปผิดมีน้อยกว่า 5%

34

โอกาสสรุปผิด

การศึกษาหรือวิจัยมีโอกาสสรุปผิดได้ 2 ทาง

1. สรุปผลว่ามีความแตกต่างในขณะที่ความจริงในประชากรไม่มี เรียกว่า เกิด ***Type I (alpha) error***
2. สรุปผลว่าไม่มีความแตกต่างในขณะที่ความจริงในประชากรมี เรียกว่า เกิด ***Type II (beta) error***

35

True Reality

Your Decision	Ha	Ho
Ha	correct conclusion	Type I error
Ho	Type II error	correct conclusion

36

Power & Sample size

- กรณีไม่พบนัยสำคัญทางสถิติ ต้องสงสัยว่าการศึกษานี้มี power เพียงพอหรือไม่
- Power คืออำนาจการตรวจสอบความแตกต่างที่อาจมีอยู่จริงในประชากร
- Power มีค่าเท่ากับ 1- beta (1 – type II error)
- Power จะขึ้นอยู่กับ Sample size เป็นหลัก

37

วิธีการทางสถิติที่ใช้บ่อย ในการทดสอบสมมติฐาน

1. **t-test ทดสอบความแตกต่างของ continuous variable ระหว่าง 2 กลุ่ม**
2. **F-test ทดสอบความแตกต่างของ continuous variable ระหว่าง 3 กลุ่มขึ้นไป**
3. **Chi-square ทดสอบความสัมพันธ์ของ discrete variable ตั้งแต่ 2 กลุ่มขึ้นไป**
4. **Correlation ทดสอบความสัมพันธ์ระหว่างตัวแปรที่เป็น continuous variable**
5. **Regression ทำนาย (predict) ค่า dependent variable จากค่า independent variables**

38

ตัวอย่าง t-test

- ค่า uric acid ในเด็ก mongolism แตกต่างจากเด็กปกติหรือไม่
- Null hypothesis : ค่า uric acid ในเด็ก mongolism เท่ากับเด็กปกติ
- กำหนดระดับนัยสำคัญที่ .05
- เก็บข้อมูลในกลุ่มตัวอย่างเด็กทั้งสองกลุ่ม
- นำข้อมูลดิบเข้าโปรแกรมคอมพิวเตอร์โดยใช้วิธีการทางสถิติ t-test
- ผล : ค่า p-value = 0.02 จึงปฏิเสธ null hypothesis
- สรุปผลว่า ระดับ uric acid ในเด็กทั้งสองกลุ่มแตกต่างกัน

39

ตัวอย่าง correlation analysis

- ค่า FBS สัมพันธ์กับค่า HbA_1C หรือไม่ในผู้ป่วยเบาหวาน
- Null hypothesis : ค่า FBS ไม่สัมพันธ์กับค่า HbA_1C ในผู้ป่วยเบาหวาน
- กำหนดระดับนัยสำคัญที่ .05
- เก็บข้อมูลค่า FBS และค่า HbA_1C ในกลุ่มตัวอย่างผู้ป่วยเบาหวาน
- นำข้อมูลดิบเข้าโปรแกรมคอมพิวเตอร์โดยใช้วิธีการทางสถิติ correlation
- ผล : ค่า r = 0.8 ค่า p-value = 0.03 จึงปฏิเสธ null hypothesis
- สรุปผลว่า ค่า FBS กับค่า HbA_1C มีความสัมพันธ์กันในระดับดี

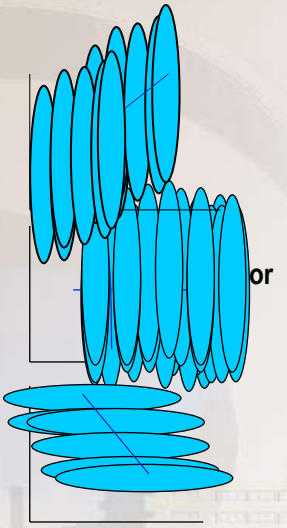
40

r =

+1 Strong positive linear relationship

0 No linear relationship

-1 Strong negative linear relationship



or

41

ภาคผนวก

42

ทำไมต้อง Confidence interval

- ถ้าอยากรู้ว่าความสูงเฉลี่ยของนักเรียนชายมัธยมปีที่6ทั่วประเทศ ปีการศึกษา 2550 มีค่าเท่าไร ? ในอดีตเคยมีการศึกษาในปี การศึกษา 2530 มีค่า population mean=165 ซ.ม. population standard deviation 10 ซ.ม. และความสูงของนักเรียนชายมัธยมปีที่6 มีการกระจายแบบโค้งปกติ

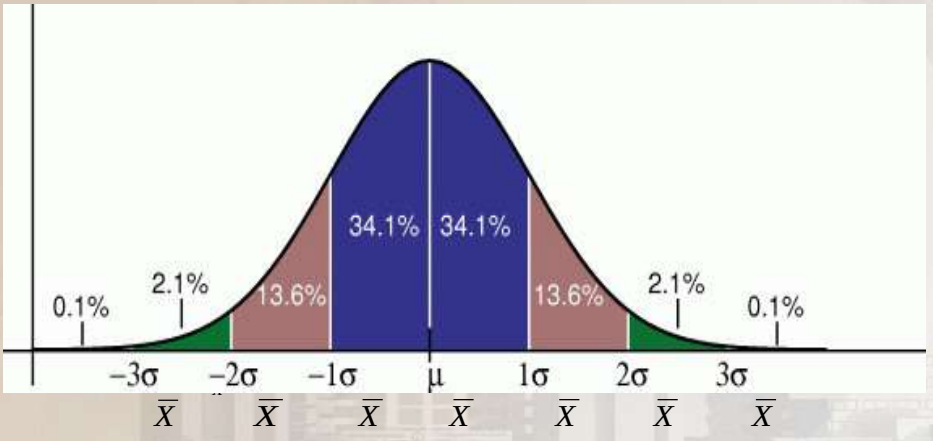
43

Sample estimation

- ถ้าความสูงเฉลี่ยของนักเรียนชายมัธยมปีที่6 ปีการศึกษา 2553 จำนวน 100 คนที่ได้จากการสุ่มตัวอย่าง มีค่า sample mean = 170 ซ.ม. S.D.= 8 ซ.ม. ท่านคิดว่าค่า population mean มีค่าเท่าไร

44

distribution of sample means



$$\pm 1.96\sigma_{\bar{x}} = 95\%$$

45

Population estimation

95% confidence interval ของ μ

$$= \bar{x} \pm 1.96\sigma_{\bar{x}}$$

$$= 170 \pm 1.96 * \frac{10}{\sqrt{100}}$$

= 168-172 ซ.ม.

46

95% CI สำหรับค่า Prevalence

- ถ้าอยากรู้ว่านักเรียนแพทย์ชั้นปี3 ทั่วประเทศเลือก ส.ส.ระบบสัดส่วนสำหรับ พรรค ก.ร้อยละเท่าไร ?
- ถ้าทำการสุ่มนักเรียนแพทย์ชั้นปี3 ทั่วประเทศ มา 100 คน พบว่า เลือกพรรค ก. 40% ซึ่งเป็น sample estimate
- การหา Population estimate ทำได้โดยใช้สูตร

95% CI ของ Prevalence (P) = $p \pm 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

47

95% CI สำหรับค่า Prevalence

- 95% CI สำหรับการเลือกพรรค ก.มีค่า
- = $0.4 \pm 1.96 \frac{\sqrt{0.4*(1-0.4)}}{\sqrt{100}}$
- = (0.304 -0.496)

48

การประยุกต์ค่า 95%CI สำหรับ sample size

จากสูตร 95% CI ของ P = $p \pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

interval = $\pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

$\pm d$ = $\pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

$$n = Z^2_{(1-\alpha/2)} * \frac{p(1-p)}{d^2}$$

49