Descriptive statistic
Measure of central tendency
Measure of frequency

Type of variables

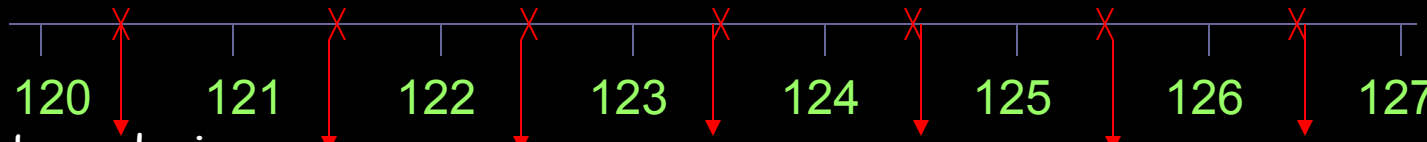
- **Qualitative: values can be categorized only**
  - **Dichotomous: 2 categories**
  - **Polytomous: more than 2 categories**
- **Quantitative: values can be measured**
  - **Discrete: takes only certain values. There are "gaps" between the values**
  - **Continuous – infinite number of possible values. There are no "gaps".**

Quantitative (numeric, metric) variables are classified as

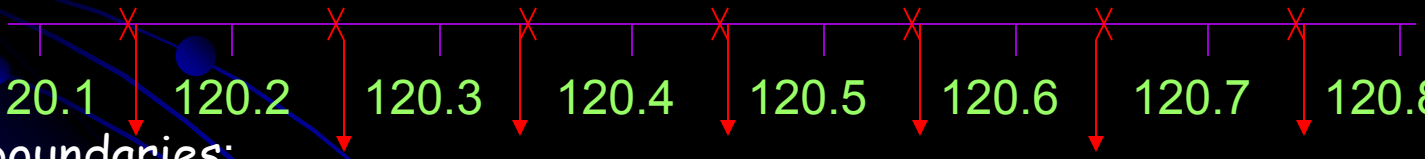
continuous It can take all values in an interval
e.g. weight, temperature, etc.

discrete It can take only certain values (often integer value)
e.g. parity, number of sex partners, etc.

Continuous data can be categorised into groups, which one needs to define "upper boundary" and "lower boundary" of a value (or a class)



boundaries:
120.5, 121.5, 122.5, 123.5, 124.5 ...



boundaries:
120.15, 120.25, 120.35, 120.45, 120.55 ...

boundaries:
120.115, 120.125, 120.135, 120.145, 120.155 ...

Measurement scales of variables

- **Qualitative variables**
  - **Nominal scale (group classification only)**
  - **Ordinal scale (classification with ordering / ranking)**
- **Quantitative variables**
  - **Interval (constant distance between points, no true zero value)**
  - **Ratio (constant distance between points + true zero)**

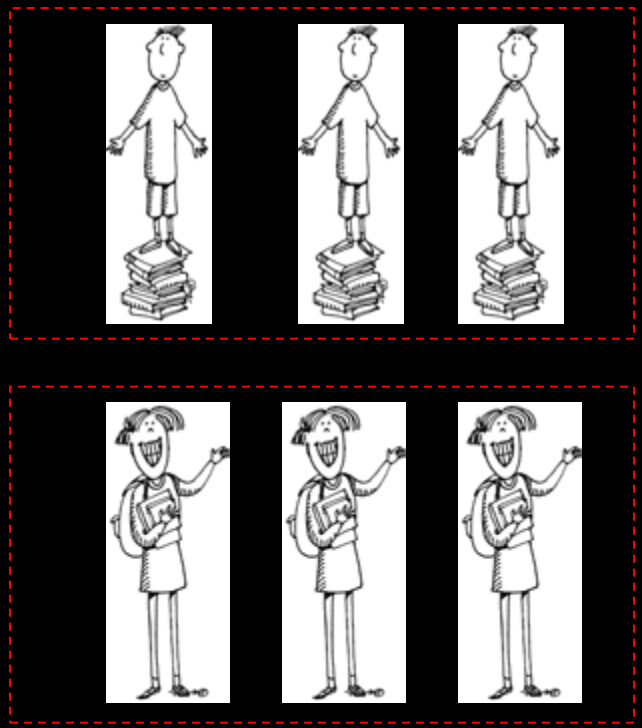
Example: Nominal scale

- **Gender, marital status, illness, death**

Male → 1

Values have no meaning

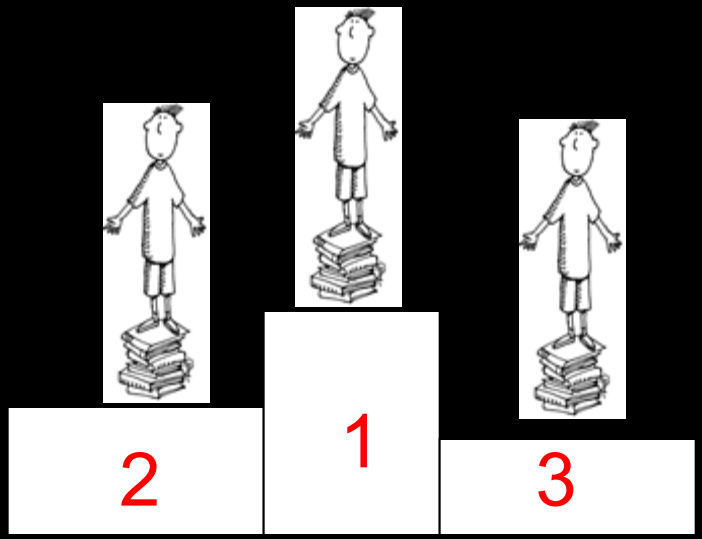
Female → 2



Example: Ordinal scale

- **Olympic winner, Education, Disease severity**

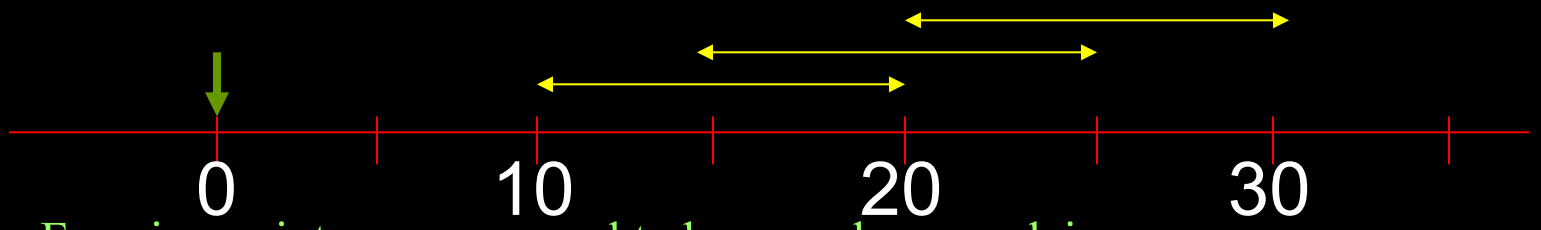
Equal distance between points does not reflect equal interval value.



1st rank → 2nd rank ≠ 2nd rank → 3rd rank

Example: Interval scale

- **Temperature**



Freezing point was supposed to be zero degree celcius

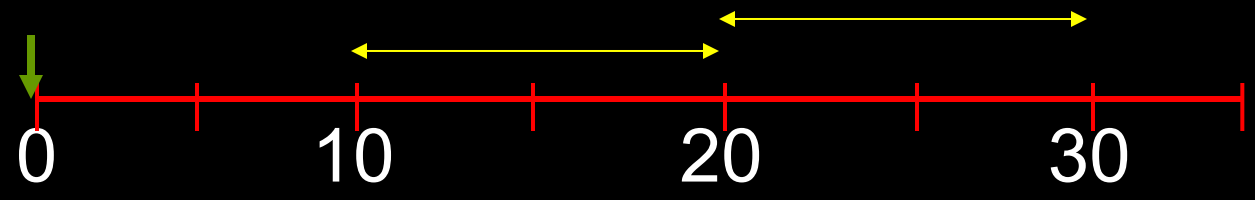
Not the true ZERO temperature (no heat)

Equal distance between points means equal interval value

10 ^{o}C →20 ^{o}C = 20 ^{o}C →30 ^{o}C

Example: Ratio scale

- **Weight, Height, Body mass index (BMI)**



True ZERO (nothing here)

Equal distance between points means equal interval value

10 →20 = 20 →30

Practice: variable type & measurement scale

- **Smoking**
  - **Yes/No:**
    - **Dichotomous, nominal**
  - **Number of cigarette**
    - **Discrete, ratio**
  - **Category of number of cigarette (0, 1 – 20, 21 – 40, >40)**
    - **Polytomous, ordinal**
- **Age**
  - **Young/old**
    - **Dichotomous, nominal**
  - **Year of age**
    - **Continuous, ratio**
  - **Age group**
    - **Polytomous, ordinal**

Variable	Result of measurement	Measurement scale
Marital status	single/married/divorced	nominal
gender	male/female	nominal
smoking	yes/no	nominal
smoking	nonsmoker/ light smoker/ moderate smoker/ heavy smoker	ordinal
smoking	number of cig/day	ratio
feeling of pain	yes/no	nominal
feeling of pain	none/light/moderate/high	ordinal
feeling of pain	0 ---------> 10	ordinal
attitude toward selective abortion	strongly agree/ agree/ not sure/ disagree/ strongly disagree	ordinal
blood pressure	mmHg	ratio
temperature	degree celcius	interval
weight	gram	ratio
tumor stage	I, II, III, IV	ordinal

Descriptive statistics - a way to summarize a dataset (a group of measurement)

Example: Height of 100 children, 10-12 years of age.

140	140	140	136	141	123	125	134	125	129
123	161	142	155	129	130	139	129	134	130
140	132	138	142	155	125	136	129	136	153
151	141	138	125	123	134	135	135	135	130
155	130	134	146	135	139	134	142	139	149
147	155	158	135	141	136	136	147	139	132
134	140	141	153	142	127	147	142	146	127
151	140	151	140	141	147	139	134	140	149
132	140	141	142	165	153	146	134	151	151
134	141	138	130	141	132	140	138	127	129

What are values that best describe the height of these 100 persons?

Summary: Descriptive Statistics

Type of variable	Measurement scale	Statistics	Presentation
Numerical discrete	**Interval** **Ratio**	**Mean, or Median, or Mode (rare), Range, SD, IQR**	**Histogram** **scatter plot** **line graph**
Numerical continuous	**Interval** **Ratio**	**Mean, or Median, or Mode (rare), Range, SD, IQR**	**Histogram** **scatter plot** **line graph**
Categorical dichotomous	**Nominal** **Ordinal**	**Proportion (%)**	**Table, Pie,** **Bar chart**
Categorical polychotomous	**Nominal** **Ordinal**	**Proportion (%)**	**Table, Pie,** **Bar chart**

การวัดค่ากลางของข้อมูล
Central Tendency

ค่าเฉลี่ย (Mean)

- ชื่อเต็ม คือ Arithmetic mean หรือ ค่าเฉลี่ยเลขคณิต
- เป็นผลรวมของค่าจากทุกสมาชิกในกลุ่มตัวอย่าง หารด้วยจำนวนสมาชิก (n)
- สัญลักษณ์ของ Mean ของกลุ่มตัวอย่าง คือ $\bar{x}$ (อ่านว่า “x-bar”)
- สูตรการคำนวณ คือ

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

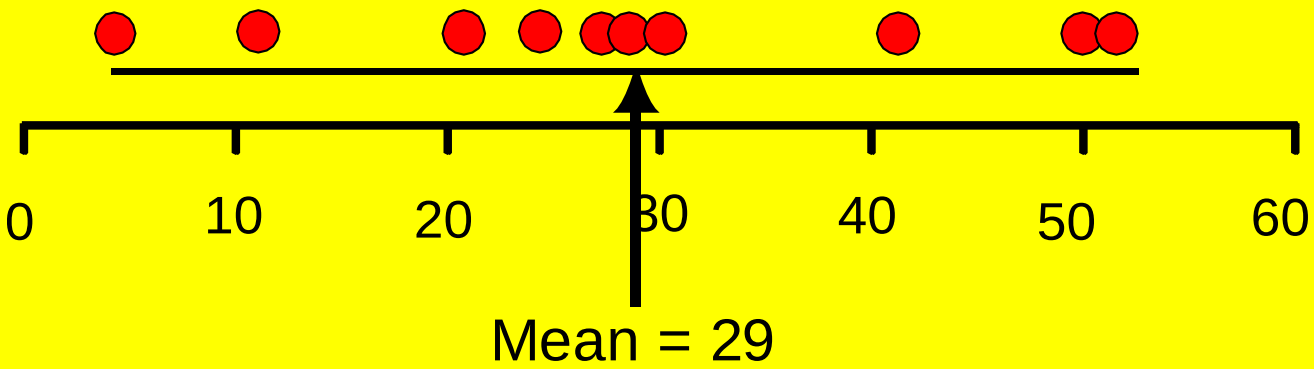
Mean: ตัวอย่าง

อายุ (ปี) ของชาวบ้าน 10 คน ดังนี้

11 28 42 5 50 30 24 52 21 27

$\Sigma x_i = 11+28+\ldots+27 = 290$ **n = 10**

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}(290) = 29.0$$



Mean: คุณสมบัติ

- ได้มาจากกระบวนการคำนวณที่ไม่ซับซ้อน และแสดงด้วยสัญลักษณ์ทางคณิตศาสตร์ได้
- ใช้กับข้อมูลเชิงปริมาณ (quantitative data) เท่านั้น
- Only one mean per data set
- ใช้ค่าของทุกสมาชิกในการคำนวณ
- ได้รับผลกระทบจากสมาชิกที่มีค่าสูงหรือต่ำ ที่ต่างมากๆจากสมาชิกส่วนใหญ่ (extreme observation) โดยเฉพาะในกลุ่มตัวอย่างขนาดเล็ก
  - เรียกว่า "*sensitive*" to outliers

มัธยฐาน(Median)

Median เป็นค่าของ สมาชิกที่อยู่ตรงกลาง (middle observation) เมื่อสมาชิกทั้งหมดถูกนำมาจัดเรียงลำดับตามค่าของมัน จากน้อยไปมาก (หรือมากไปน้อย)

–When *n* is odd (เลขคี่),

median คือ ค่าของสมาชิกตัวที่ $[(n+1)/2]^{th}$

–When *n* is even (เลขคู่),

median คือ ค่าเฉลี่ยเลขคณิตของ สมาชิกสองตัวตรงกลางได้แก่ สมาชิกตัวที่$[n/2]^{th}$ กับ $[(n/2)+1]^{th}$

Median: ตัวอย่าง

อายุ (ปี) ของชาวบ้าน 10 คน ดังนี้

11 28 42 5 50 30 24 52 21 27

Step 1: เรียงอายุจากน้อยไปมาก

5 11 21 24 27 28 30 42 50 52

Step 2: n=10 (เลขคู่): ดังนั้นต้องหาสองตัวที่อยู่ตรงกลาง,

$[n/2]^{th} = 5^{th}$ and $[(n/2)+1]^{th} = 6^{th}$ observations

Step 3: คำนวณค่าเฉลี่ยของสองตัวนั้น

Median = (27+28)/2 = 27.5

Median: ตัวอย่างอื่นๆ

- Example A: 2 4 6
  Median = 4
- Example B: 2 4 6 8
  Median = 5 (average of 4 and 6)
- Example C: 6 2 4
  Median ≠ 2

(Values must be ***ordered*** first)

Median: คุณสมบัติ

- เท่ากับ 50^{th} percentile
- ใช้กับข้อมูลเชิงปริมาณ (quantitative data) เท่านั้น
- Only one median per data set
- ไม่ได้รับผลกระทบจาก extreme observations
  - เรียกว่า “robust” to outliers

Median is "Robust"

The median is more resistant to skews and outliers than the mean; it is more *robust*.

1362 1439 1460 1614 1666 1792 1867

This data set has a mean of 1636

Here's the same data set with a data entry error **"outlier"**

1362 1439 1460 1614 1666 1792 ***9867***

This data set has a mean of 2743

The median is 1614 in both instances, demonstrating its robustness in the face of outliers.

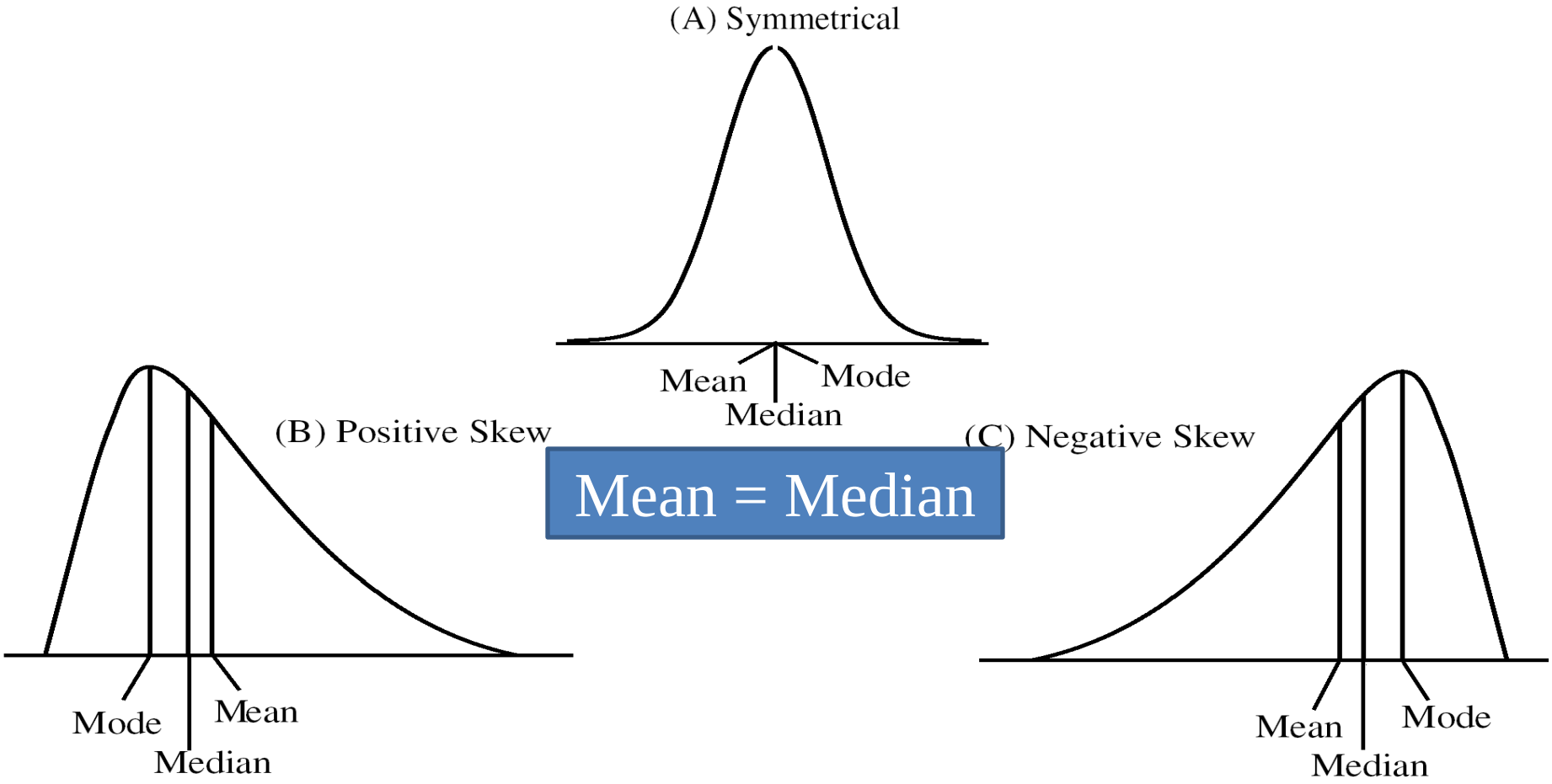
ฐานนิยม (Mode)

- Mode คือ ค่าที่ซ้ำกันมากที่สุด (the most frequent value in the dataset)
- {4, 7, 7, 7, 8, 8, 9} >>> mode of 7
- {4, 6, 7, 8} >>> has no mode
  - (each value appears only once)
- {4, 7, 7, 7, 8, 8, 8, 9} >>> mode of 7 and 8

Mode: คุณสมบัติ

- ใช้ได้ทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ (quantitative and qualitative variables)
- May be >1 mode per data set
- May have no mode for some data sets
- ไม่ได้รับผลกระทบจาก extreme observations
- The mode is useful only in large data sets with repeating values

Comparison of Mean, Median, Mode



a skew on the mean, median,

การวัดทางระบาดวิทยา

Part I:
หลักการพื้นฐาน

How epidemiologists work?

1. การนับจำนวน

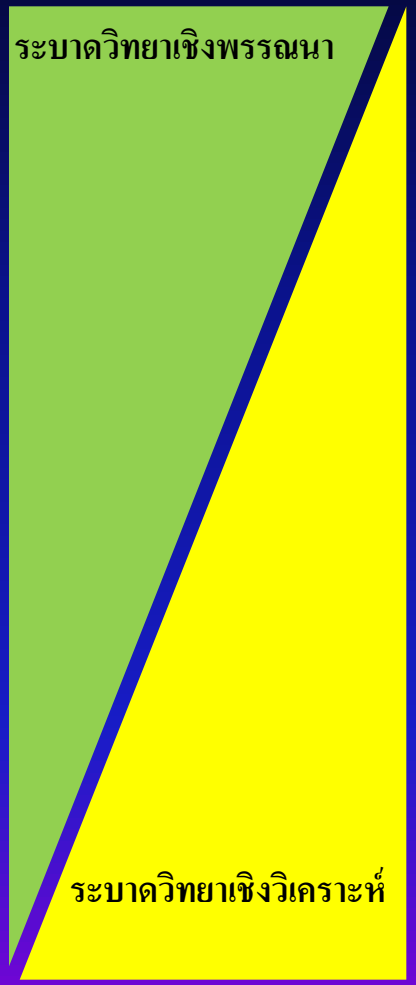
Counts cases or health events, and describes them in terms of time, place, and person

2. การหาร

Divides the number of cases by an appropriate denominator to calculate “rates”

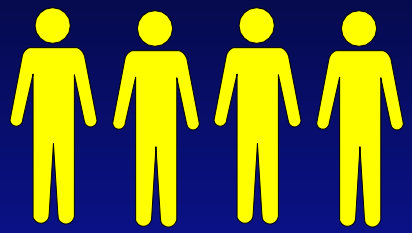
3. การเปรียบเทียบ

Compares these “rates” over time or for different groups of people



*คำว่า rate ในที่นี้ใช้ในความหมายทั่วไป หมายถึง การหารตัวเลขหนึ่งด้วยตัวเลขอีกตัวหนึ่ง

การนับจำนวน



- จำนวนผู้ป่วย

  “มีผู้ป่วยเบาหวาน 4 ราย”

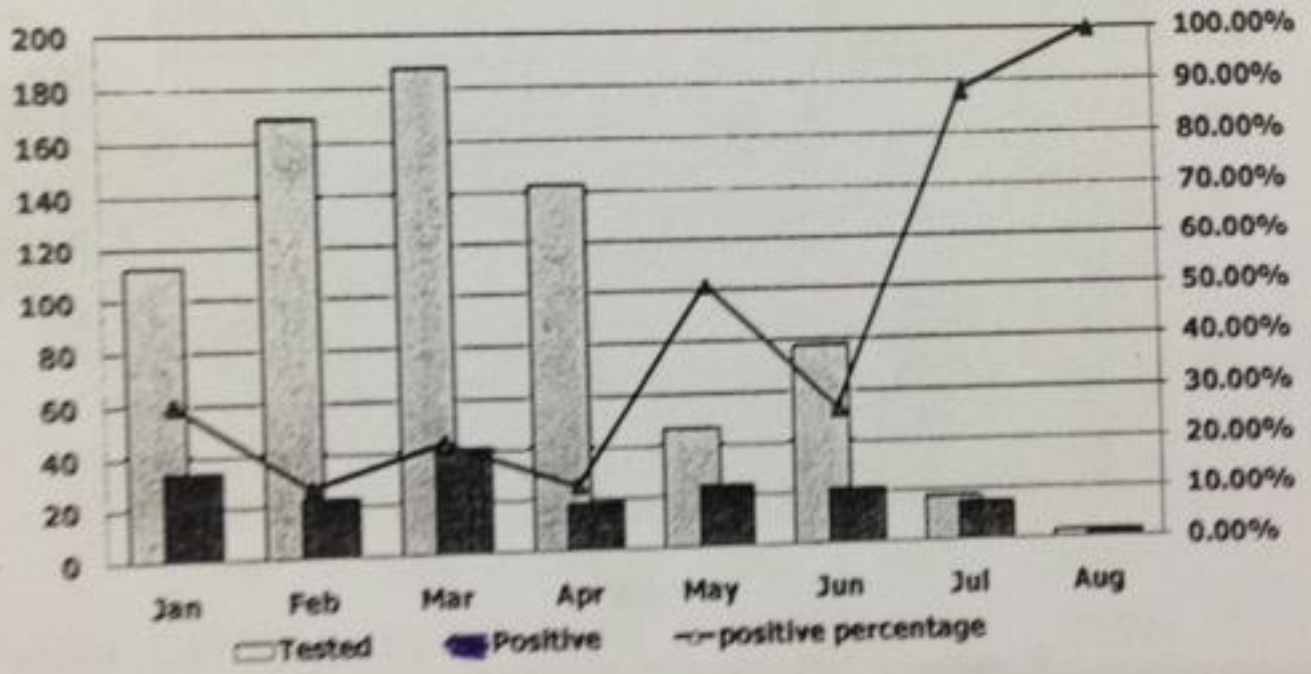
ง่ายแต่ไม่ค่อยบอกอะไร

เป็นปัญหามากหรือไม่เมื่อเทียบกับพื้นที่อื่น

ต้องการตัวหารเพิ่มเติม!!

Rabies in animal (Jan – 20 Aug 10)

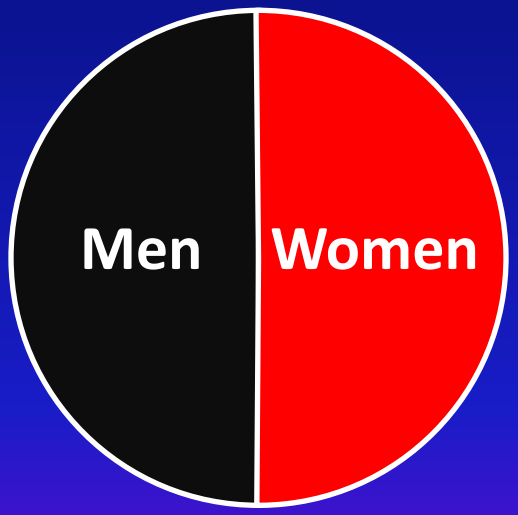
- 765 sample were tested; 182 confirmed (23.8%)



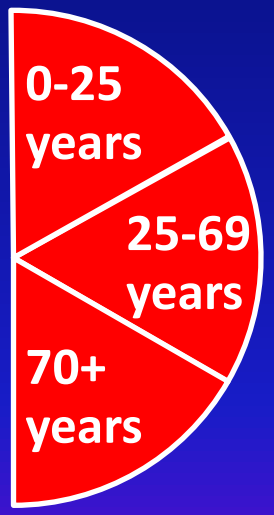
Source: P. Srisai, 2010

ตัวหารที่เหมาะสมของโรคมะเร็งปากมดลูก?

Total population



Population at risk



All of them still have cervix?

Source: Basic Epidemiology, 2nd Ed., R Bonita, 2006

อำเภอ ก. ประชากร 100 คน

มีหญิงอายุ 1 – 25 ปี	20 คน
มีหญิงอายุระหว่าง 25 – 69 ปี	50 คน
มีหญิงอายุมากกว่า 69 ปี	30 คน
มีผู้ป่วยมะเร็งปากมดลูกรายใหม่	15 คน

อำเภอ ข. ประชากร 100 คน

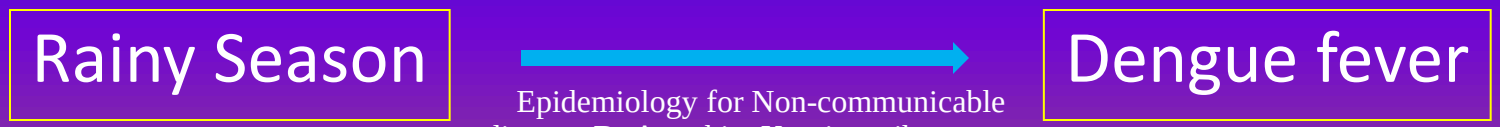
มีหญิงอายุ 1 – 25 ปี	50 คน
มีหญิงอายุระหว่าง 25 – 69 ปี	20 คน
มีหญิงอายุมากกว่า 69 ปี	30 คน
มีผู้ป่วยมะเร็งปากมดลูกรายใหม่	10 คน

อำเภอใดมีปัญหามากกว่ากัน

อัตราส่วน VS. สัดส่วน VS. อัตรา

- **Ratio** (อัตราส่วน)
  - เท่ากับ A / B โดยที่ A จะเป็นหรือไม่เป็นส่วนหนึ่งของ B ก็ได้ และไม่จำเป็นต้องมีหน่วยเดียวกัน
  - เช่น การตายมารดาต่อแสนการเกิดมีชีพ หรือ อัตราส่วนจำนวนเตียงผู้ป่วยต่อจำนวนประชากร
- **Proportion** (สัดส่วน)
  - เท่ากับ A / B โดยที่ A เป็นส่วนหนึ่งของ B และมีหน่วยเดียวกัน
  - มีค่าอยู่ในช่วงตั้งแต่ 0 ถึง 1 (แต่อาจจะแสดงเป็น ต่อร้อย ต่อหมื่น ต่อแสน ฯลฯ)
  - เช่น สัดส่วนของเพศชายในประชากรทั้งหมด
- **Rate** (อัตรา)
  - ในทางระบาดวิทยา หมายถึง เท่ากับ A / เวลาที่ทำการสังเกตุหรือศึกษา

Exposure and Disease



ชนิดของการวัด

1. การวัดขนาด (Measure of frequency)
   - วัดขนาดของโรคหรือภาวะทางสุขภาพ
   - เช่น ผู้ป่วยโรคอาหารเป็นพิษในจังหวัด ก มีจำนวนเท่าใดในปี 2552
2. การวัดความสัมพันธ์ (Measure of association)
   - การวัดความสัมพันธ์ระหว่าง “ปัจจัยที่ศึกษา” และ “โรค”
   - เช่น เพศกับการเกิดหัวใจมีความสัมพันธุ์กันหรือไม่
3. การวัดผลกระทบ (Measure of impact)
   - การวัดผลกระทบของการมีหรือไม่มีปัจจัยที่ศึกษาต่อการเกิดโรค
   - เช่น วัคซีนหัดมีประสิทธิภาพต่อการป้องกันโรคหัดดีแค่ไหน

Part II:
Measures of Frequency

การวัดขนาด

- Prevalence (ความชุก):
  - การวัดขนาดของโรค “ที่มีอยู่” ใน ณ เวลาที่กำหนด
  - นับรวมทั้งรายใหม่และรายเก่า
  - ประโยชน์ของการใช้ ความชุก?
- Incidence (อุบัติการณ์):
  - การวัดขนาดของโรคที่ “เกิดใหม่” ในช่วงเวลาที่กำหนด
  - สนใจเฉพาะรายใหม่
  - ประโยชน์ของการใช้ อุบัติการณ์?

EX1: แสดงข้อมูลโรคเบาหวานในอำเภอ ก และ ข

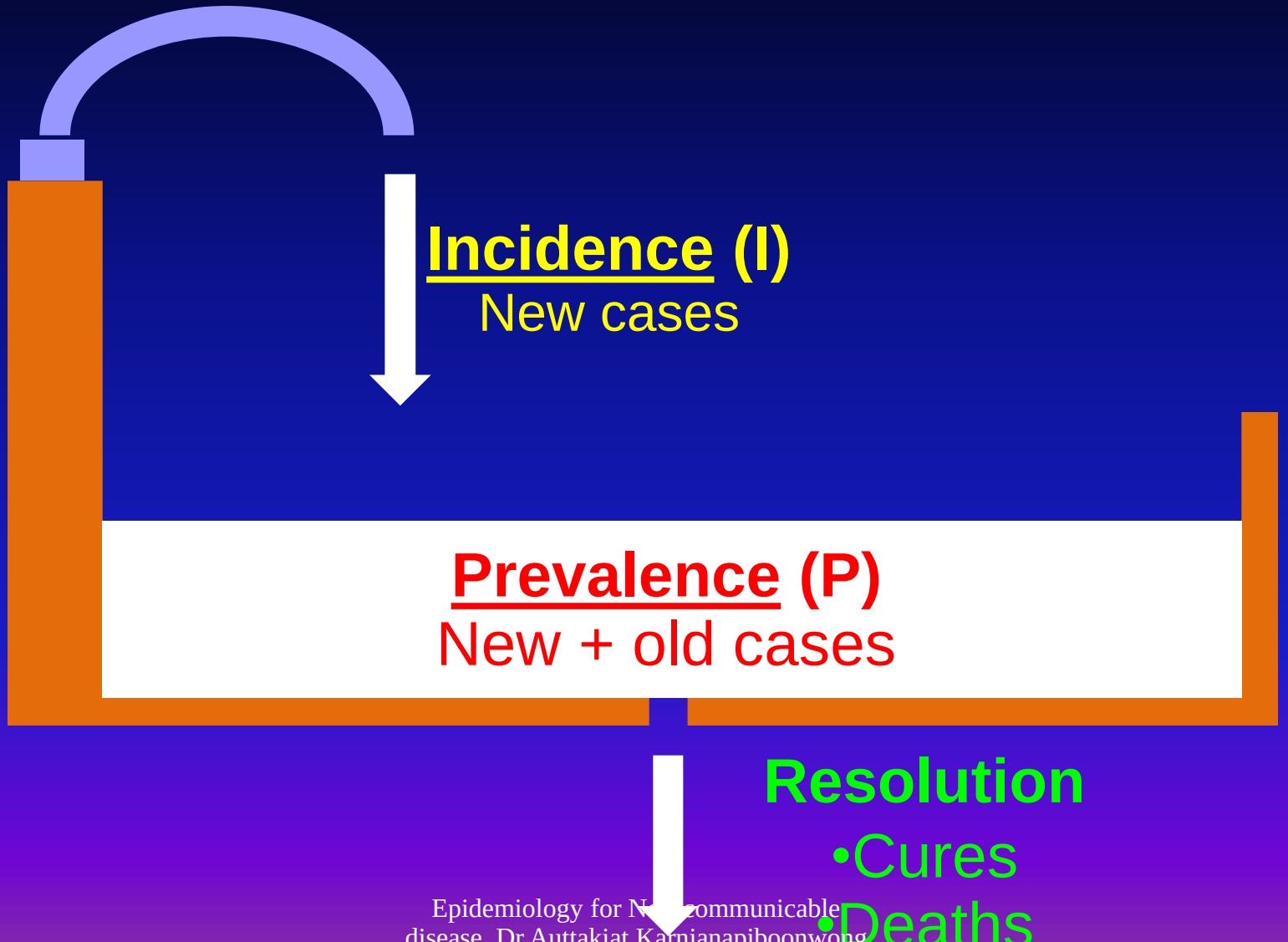
ความชุก

	2552	2553	2554	2555
อำเภอ ก	20%	22%	23%	24%
อำเภอ ข	20%	20%	19%	19%

อัตราป่วยตาย

	2552	2553	2554	2555
อำเภอ ก	1%	0%	1%	1%
อำเภอ ข	6%	5%	9%	10%

ความชุก และอุบัติการณ์



EX2:ข้อมูลโรคเบาหวาน ความดันโลหิตสูง

- ที่รายงานอยู่เป็น prevelence หรือ incidence
- จะทราบได้อย่างไรว่ามีผู้ป่วยที่เพิ่งเป็นโรคใหม่จริงๆกี่ราย

EX3: Prevalence *or* Incidence

- ร้อยละของคนที่ใส่แว่นตาในห้องนี้
- จำนวนผู้ที่เป็นโรคความดันโลหิตสูงในห้องนี้
- อัตราตายด้วยโรคมะเร็งในบุคลากรสาธารณสุขในปี 2545-2554
- จำนวนผู้ติดเชื้อเอ็ชไอวีรายใหม่ปี 2555
- จำนวนผู้เสพยาบ้าในกรุงเทพมหานครในวันที่ 1 มิ.ย. 2554
- อัตราป่วยโรคไข้ชิคุนกุนยาในปี 2555
- จำนวนคนที่นั่งหลับในขณะนี้

Prevalence

1. Point prevalence
   - ความชุก ณ จุดเวลา (ในทางปฏิบัติ คือ ช่วงเวลาที่แคบมากๆ)
   - เช่น อัตราความชุกของผู้ป่วยเบาหวาน ณ วันที่ 1 ตุลาคม 2555

$$= \frac{\text{จำนวนผู้ป่วยเบาหวาน ณ วันที่ 1 ตุลาคม 2555}}{\text{จำนวนประชากร ณ วันที่ 1 ตุลาคม 2555}} \times 100{,}000$$

Population at risk

2. Period prevalence
   - ความชุก ณ ช่วงเวลา
   - เช่น อัตราความชุกของผู้ป่วยโรคความดันโลหิตสูงในปี 2555

$$= \frac{\text{จำนวนผู้ป่วยเมื่อ 1 ม.ค. 2555} + \text{ผู้ป่วยใหม่ในปี 2555}}{\text{จำนวนประชากรกลางปี 2555}} \times 100{,}000$$

Population at risk

Interpretation of Prevalence

- เนื่องจากความชุกนั้นขึ้นกับ อุบัติการณ์ และระยะเวลาดำเนินโรค จึงไม่สามารถใช้ความชุกเพื่อหาสาเหตุของโรคได้
- มีประโยชน์ในการบอกภาระโรค ในประชากร การจัดสรรทรัพยากร

EX4: แปลผล ความชุก VS อุบัติการณ์

อำเภอ ก

- ความชุกเบาหวาน 30%
- อุบัติการณ์เบาหวาน 5%

อำเภอ ข

- ความชุกเบาหวาน 20%
- อุบัติการณ์เบาหวาน 10%

Incidence

1. **Incidence proportion** ชื่ออื่นๆ ได้แก่
   - Incidence risk
   - Cumulative incidence
   - Attack rate (ใช้กับการระบาด)
2. **Incidence rate** ชื่ออื่นๆ ได้แก่
   - Incidence density
   - Force of morbidity/Force of mortality
   - Hazard rate
   - Person-time rate
   - Disease intensity

Incidence Rate

เป็นการวัดว่าการเกิดโรคในกลุ่มคนที่ยังไม่มีโรคนั้น เกิดขึ้นเร็วหรือช้าเพียงใด จึงเป็นการวัดอัตราการเกิดโรคโดยใช้ระยะเวลาของการเสี่ยงต่อการเกิดโรคของทุกๆ คนที่อยู่ในการศึกษามาเป็นฐานของการคำนวณ

การติดตาม N คน เป็นเวลา t ปี

แต่ละคนมีระยะเวลาที่ติดตามจริงเท่ากับ t_i ปี โดย i=1, 2, 3, …, N

พบผู้ป่วย x รายในช่วงระยะเวลาดังกล่าว

Incidence Rate = x / ผลรวมของทุก t_i

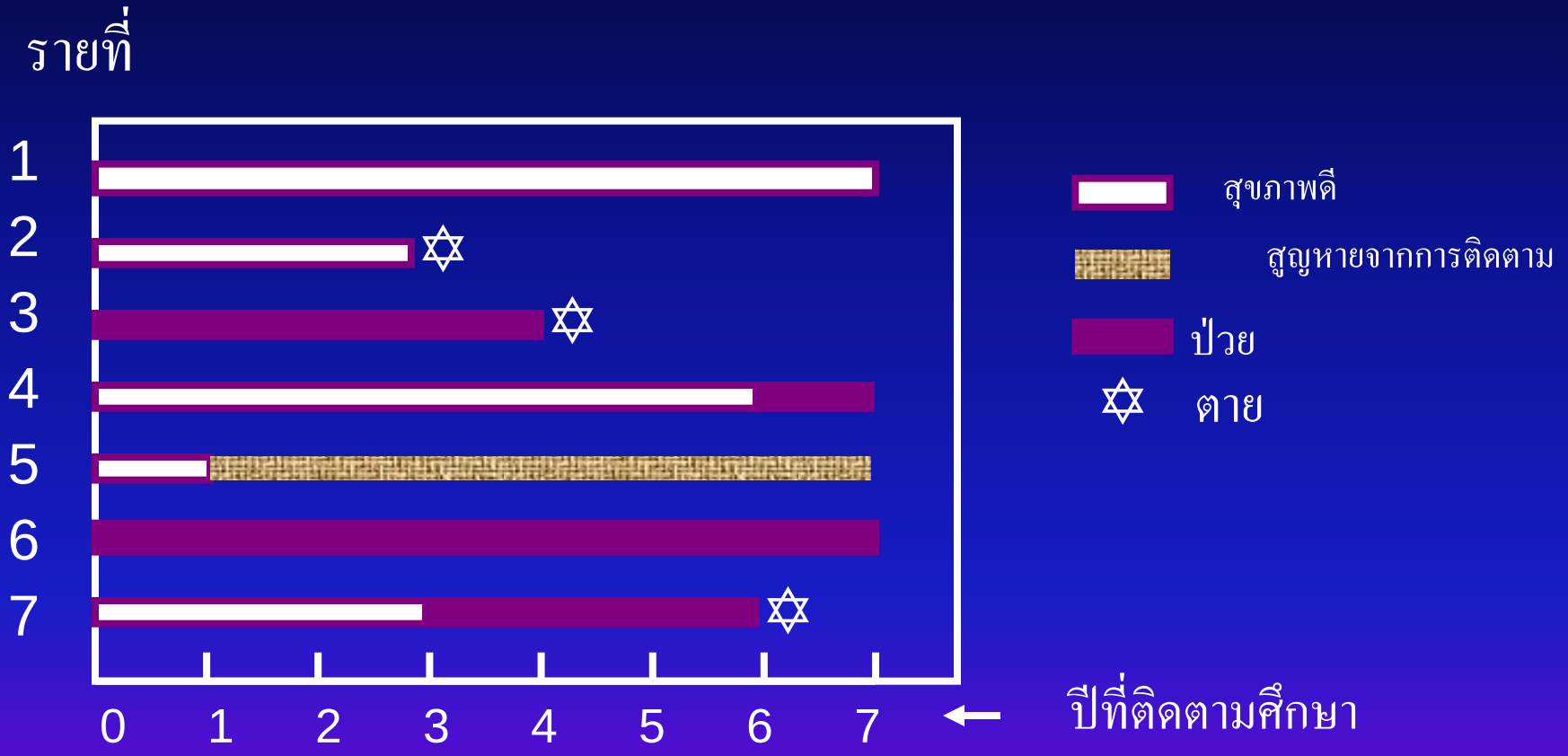
$= x/(t_1 + t_2 + t_3 + \ldots + t_N)$

Person-time at risk

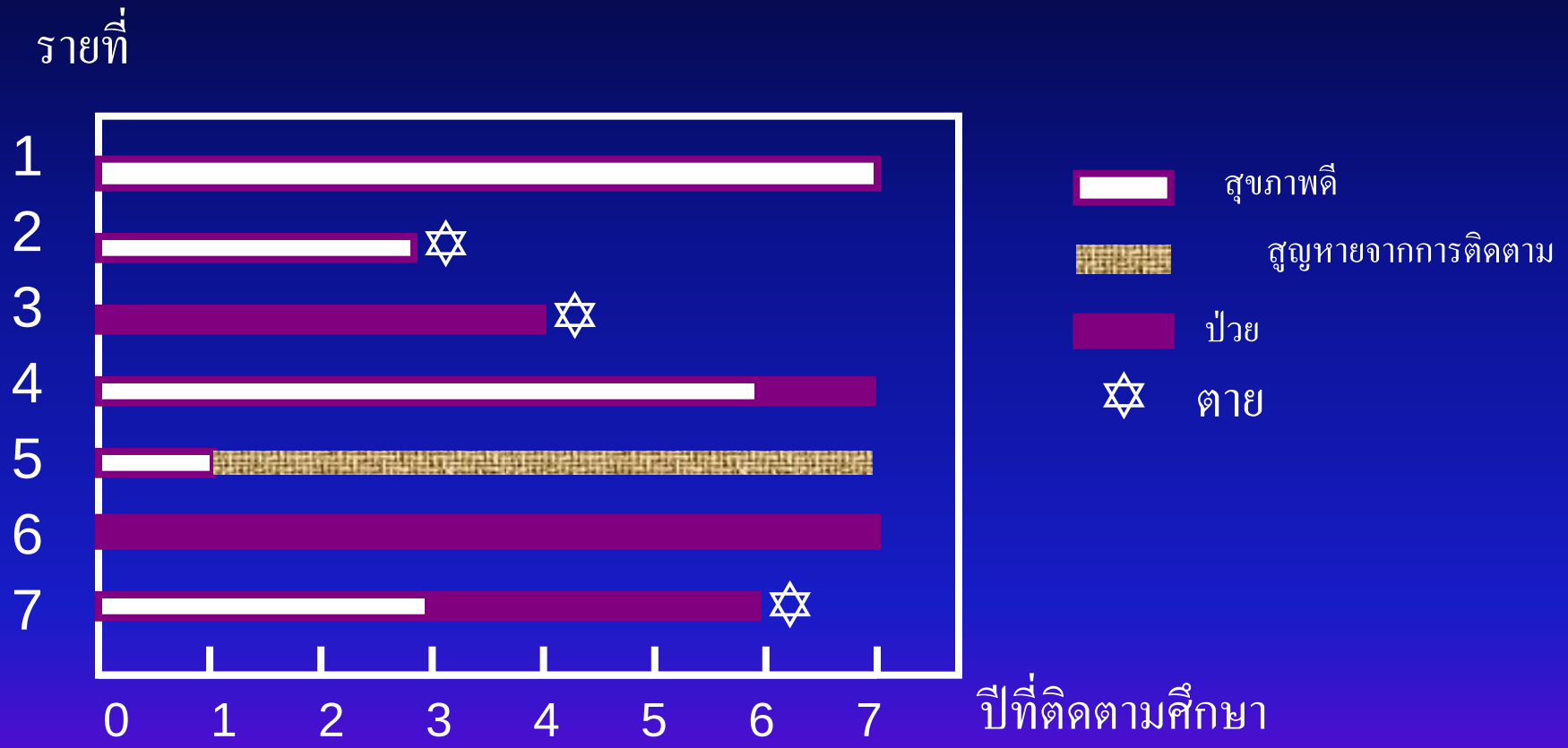
Interpretation of Incidence Proportion

- Three assumptions required when calculating incidence proportion:
  - ไม่มีการเคลื่อนไหวของประชากร, ไม่มีการเสียชีวิตจากสาเหตุอื่น
  - ประชากรทั้งหมดเป็นประชากรกลุ่มเสี่ยง ที่ได้รับการติดตามอย่างต่อเนื่อง
  - ที่จุดเริ่มต้น ทุกคนต้องมีโอกาสที่จะเกิดเหตุการณ์ที่กำลังศึกษาอยู่
- The only way to interpret incidence proportion, or "risk", is to know the length of the time period over which the risk applies

EX5: การติดตามประชากร 7 คนเป็นเวลา 7 ปี

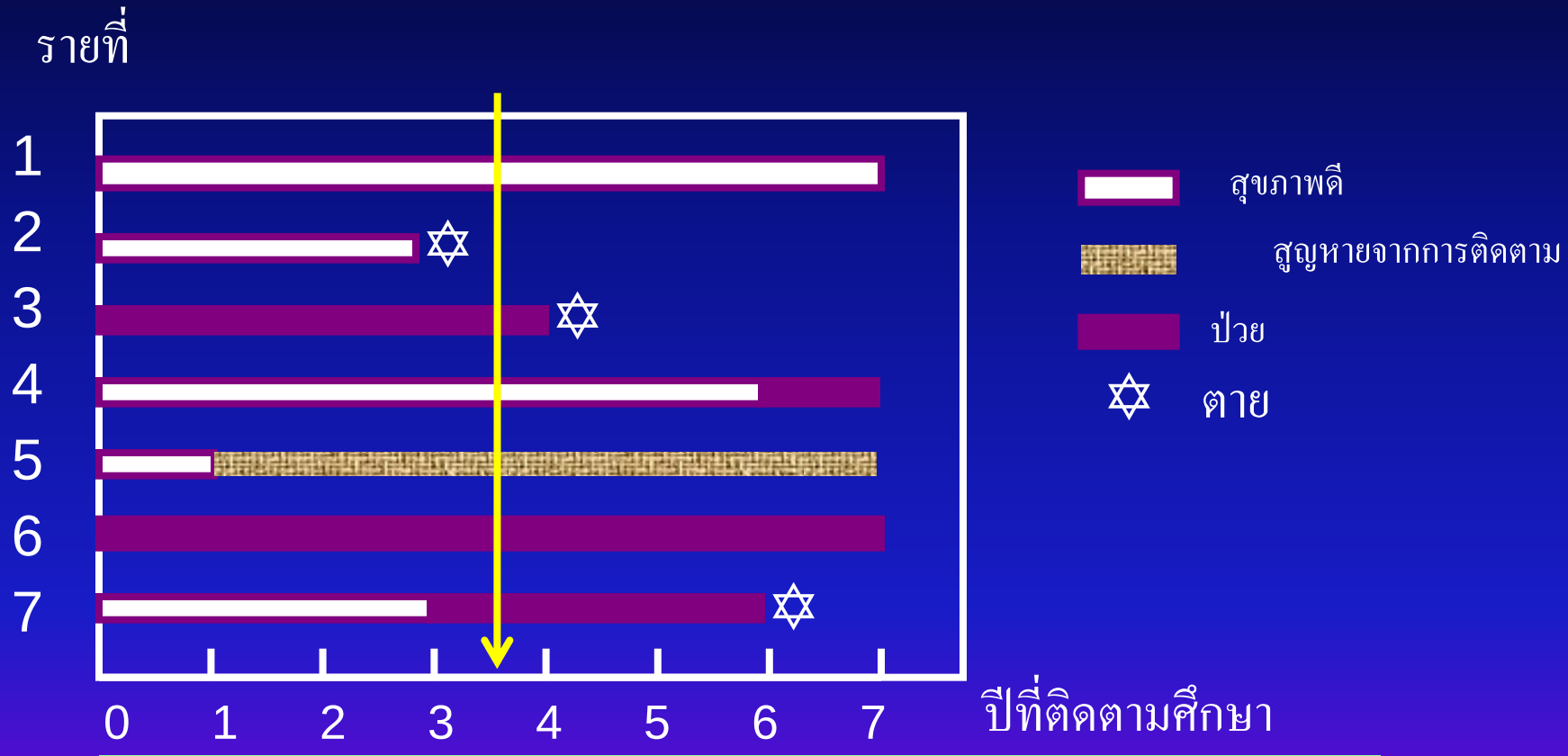


ตัวอย่างของ Point prevalence



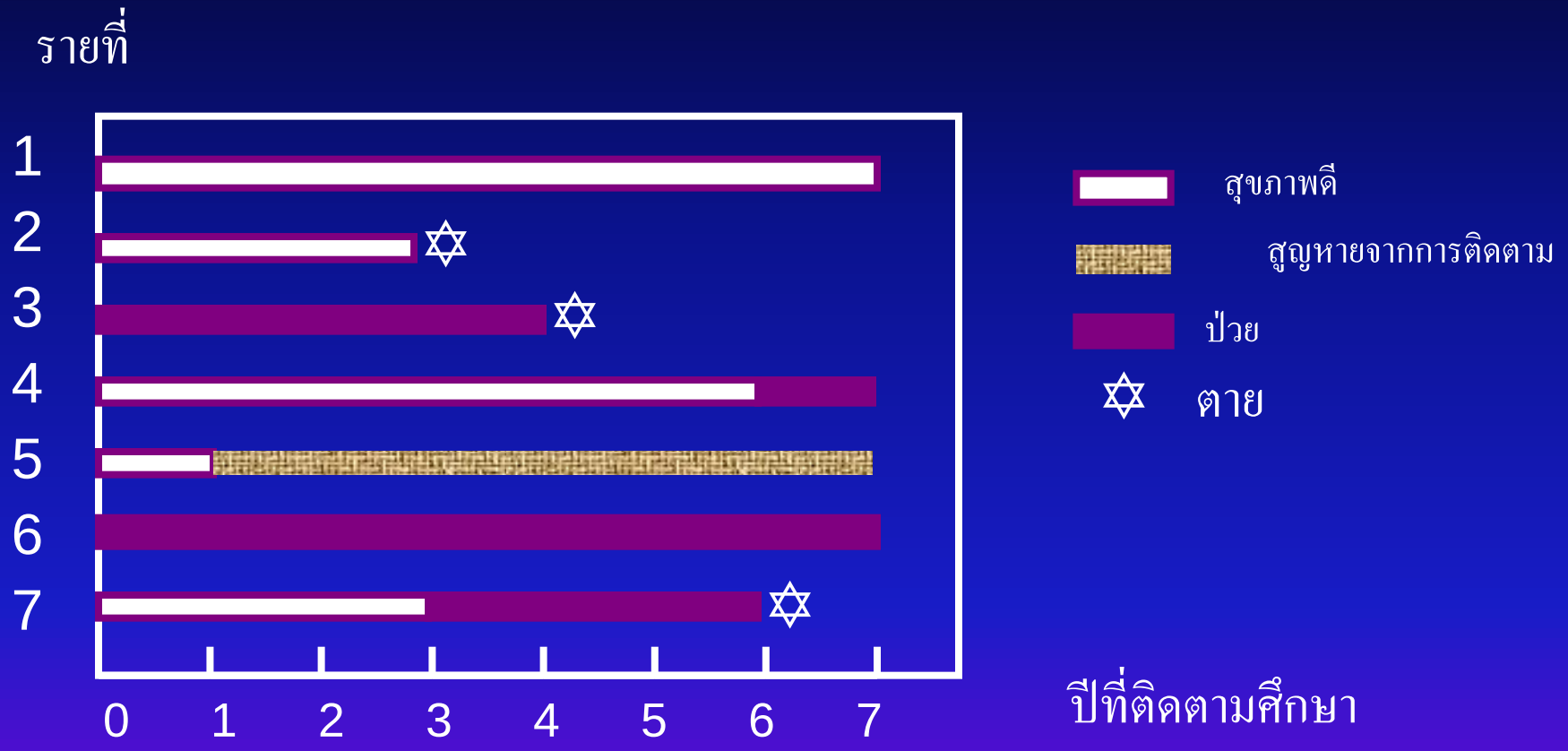
ความชุกของโรคเมื่อเริ่มต้น = ?

ตัวอย่างของ Period prevalence



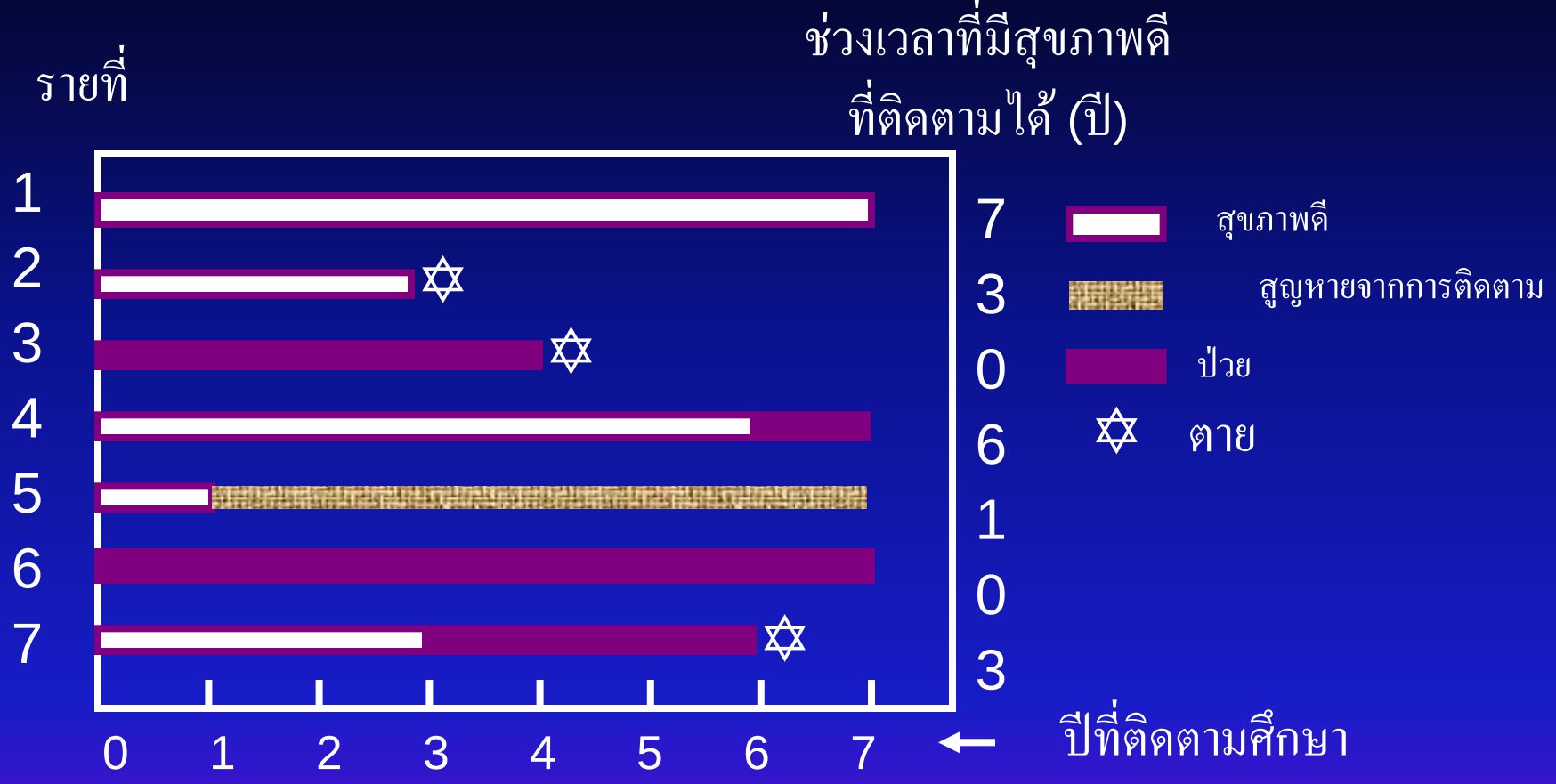
ความชุกของโรคใน 7 ปี = ?

ตัวอย่างของ Incidence proportion



ความเสี่ยงของการเกิดโรคใน 7 ปี = ?

ตัวอย่างของ Incidence rate



Person time = ?

Incidence rate = ?

Crude vs Specific Measures

- **Crude (overall) Incidence/Prevalence**
  - Incidence/prevalence among total population at risk
    = # of cases/deaths among total pop. at risk / # of total pop. at risk
  - Eg. National cancer incidence in 2009

- **Specific Incidence/Prevalence**
  - Incidence/prevalence among subpopulation

    = # of case/deaths among subpop. at risk/ # of that subpop. at risk
  - Eg.
    - Age-specific cancer incidence
    - Gender-specific DM prevalence
    - Food-specific attack rate

Differences among specific incidence/prevalence are important for generating hypotheses!

การใช้คำว่า “อัตรา” หรือ “Rate” ที่ไม่ตรงความหมายที่แท้จริง

- อัตราอุบัติการณ์ = **Incidence rate**

  หรือ อัตราป่วย (ระหว่างการระบาด) = **Attack rate**

  = # of new cases / # of pop at risk at the beginning

- อัตราความชุก = **Prevalence rate**

  = # of existing cases at a point of time / # of pop at risk at that point of time

- อัตราตาย = **Mortality/Death rate**

  = # of deaths / # of pop at the beginning

- อัตราป่วยตาย = **Case-fatality rate**

  = # of deaths from a disease / # of cases of that diseases at the beginning

ทั้งหมดที่กล่าวมานี้ ที่จริงแล้วเป็น “สัดส่วน” หรือ “Proportion”

ข้อพึงระวัง

- ใช้การวัดผิด (Prevalence or Incidence?)
  - เช่น ความพิการแต่กำเนิด
- ใช้ตัวหารผิด
  - ไม่ได้นำเฉพาะประชากรกลุ่มเสี่ยงมาหาร
- เรียกชื่อผิด
  - ใช้คำว่าอัตราไม่เหมาะสม

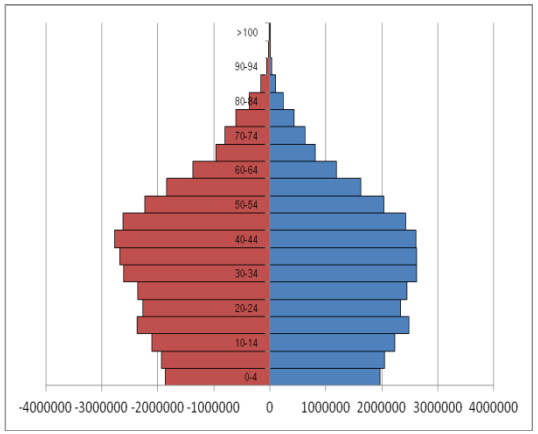
Descriptive Epidemiology: Describing the Burden and Distribution in Populations

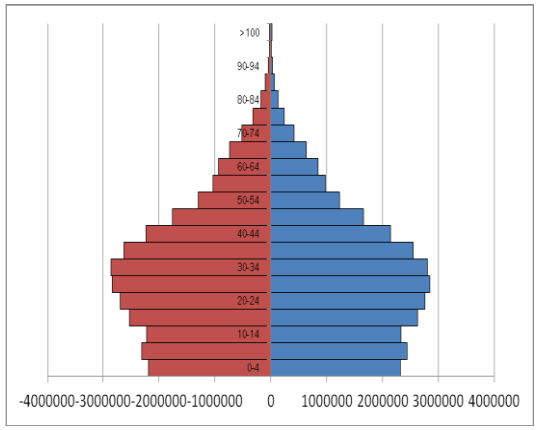
Calculating rates in populations

- Cumulative incidence:
- Prevalence:
- Incidence rate:

Comparing rate across populations

- age-specific rates
- Age-standardized





Different age structure

Example: Age-standardized mortality rate

age	Standard population	Age-specific mortality rate: White	Number of death: white	Age-specific mortality rate: Black	Number of death: black
<1	700	23.9	167.3	31.3	219.1
1 - 4	32000	0.7	25.3	1.6	51.2
5 - 17	113000	0.4	45.2	0.6	67.8
18 - 44	145000	2.5	362.5	4.8	59.6
45 - 64	81000	15.2	1231.1	22.6	1830.6
≥65	50000	69.3	3465.0	75.9	3795.0
total	428,000	-	5296.5	-	6659.7

Example: age-standardized mortality rate

- Crude mortality rate (per 1000 populations)
  - White: = 14.3
  - Black: = 10.2
- Age-standardized mortality rate (per 1000 populations)
  - white: = (5296.5/428000) x 1000 = 12.3/1000
  - Black: = (6659.7/428000) x 1000 = 15.6/1000

What is DALY's

- DALYs = Disability Adjusted Life Years
  - The sum of years of potential life lost due to premature mortality and the years of productive life lost due to disability.
- DALYs = YLL + YLD
  - YLD = Years Lived with Disability
  - YLL = Years of life lost

Originally developed by Harvard University for the World Bank in 1990

ธรรมชาติของโรค

- ประชากรแต่ละกลุ่ม เช่น เพศ อายุ มีความแตกต่างกันในแง่ของอุบัติการณ์ของป่วยหรือเสียชีวิต อายุคาดหวังเฉลี่ย
- โรคแต่ละโรคมีความแตกต่างกันหลายระดับความรุนแรงขึ้นอยู่กับธรรมชาติของโรค

$$DALY_{รวม} = DALY_1 + DALY_2 + DALY_3 + \ldots$$

$$DALY_1 = YLL_1 + YLD_1$$
$$DALY_2 = YLL_2 + YLD_2$$
$$\ldots$$
$$\ldots$$
$$DALY_n = YLL_n + YLD_n$$

Table1: Rank of diseases with DALYs by gender, Thailand 2547 (B.E.)

Male (Year(B.E.) 2547)			Female (Year(B.E.) 2547)		
Rank	Diseases	DALYs (x100,000)	Rank	Diseases	DALYs (x100,000)
1	HIV/AIDS	6.5	1	Stroke	3.2
2	Traffic accident	5.9	2	HIV/AIDS	2.9
3	Stroke	3.4	3	Diabetes	2.9
4	Alcohol dependence/harmful use	3.3	4	Depression	1.9
5	Liver cancer	2.8	5	Ischemic heart disease	1.4
6	Ischemic heart diseases	1.8	6	Osteoarthritis	1.3
7	COPD	1.8	7	Traffic accidents	1.3
8	Diabetes	1.8	8	Liver cancer	1.3
9	Cirrhosis	1.5	9	Deafness	1.1
10	Depression	1.4	10	COPD	1.1

source: International Health Policy Program Thailand, Ministry of Public Health

Male (Year(B.E.) 2552)			Female (Year(B.E.) 2552)		
Rank	Diseases	DALYs (x100,000)	Rank	Diseases	DALYs (x100,000)
1	Alcohol related disease	5.1	1	Diabetes	3.8
2	Traffic accident	5.0	2	Stroke	3.5
3	Stroke	3.7	3	Depression	2.4
4	HIV	2.6	4	Ischemic heart disease	1.8
5	Liver cancer	2.6	5	HIV/AIDs	1.6
6	Ischemic heart diseases	2.5	6	Cataract Osteoarthritis	1.5
7	Diabetes	2.2	7	Osteoarthritis	1.4
8	COPD	2.1	8	Traffic accident	1.3
9	Cirrhosis	1.8	9	Anemia	1.2
10	CA lung	1.3	10	Liver cancer	1.1

Thank You!