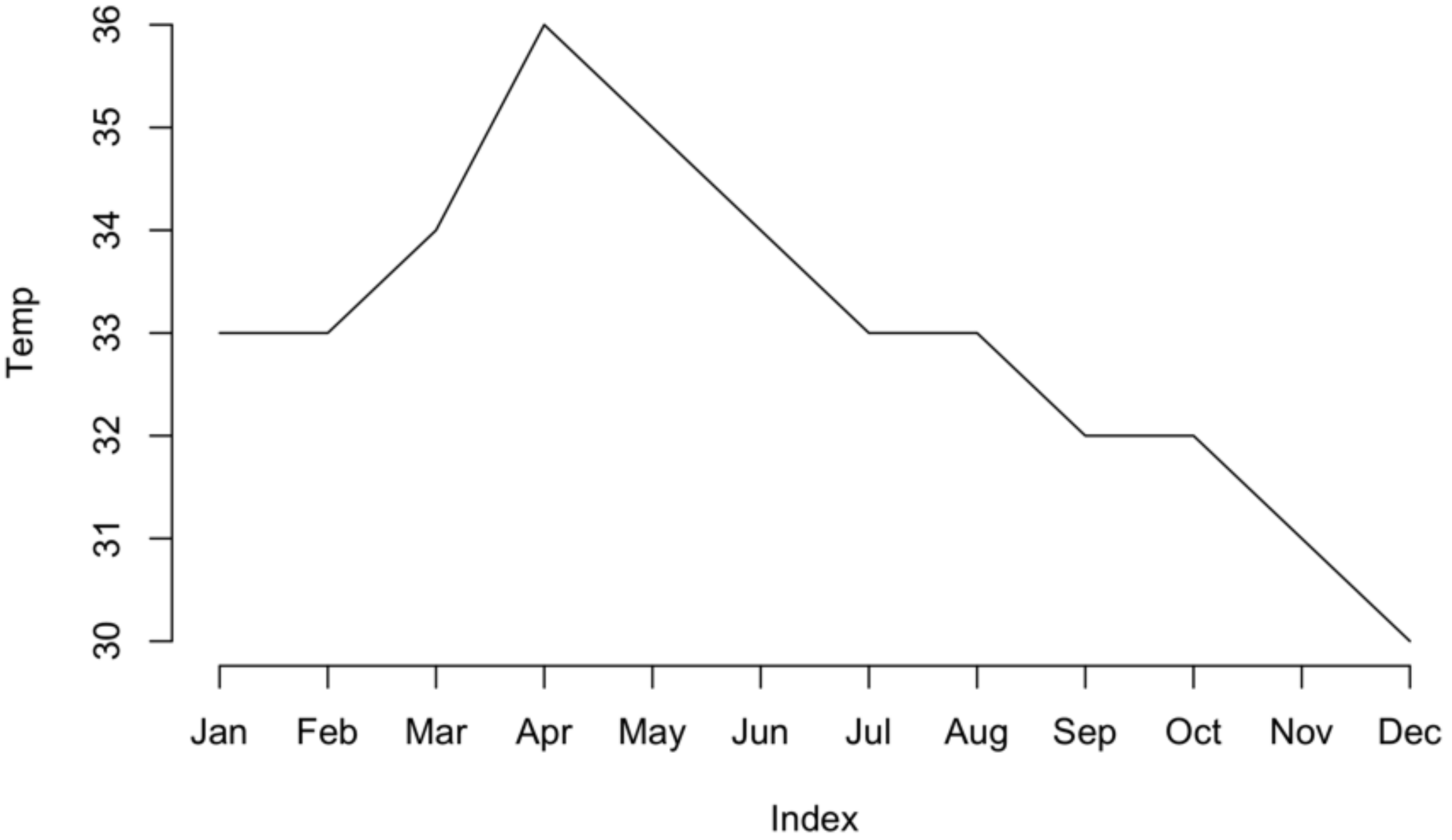
Basic Statistics

- สถิติเป็นศาสตร์ที่ใช้วิธีการทางคณิตศาสตร์ไปเก็บรวบรวม จัดระบบ ประมวล และวิเคราะห์ผลข้อมูล
- การแบ่งประเภททางสถิติ แยกเป็น descriptive กับ inferential
- 1. Descriptive สถิติแบบพรรณา เป้าหมายเพื่อลดจำนวนข้อมูลดิบให้ง่ายต่อการอธิบาย / ตีความ ให้เห็นภาพออกมาเป็นตัวเลข
  - 1.1. Data distribution ดูการกระจายของข้อมูล ความถี่ frequency, cumulative frequency
  - 1.2. Summary Statistics
  - 1.2.1. central tendency หาค่าที่เป็นตัวกลางของ sample ที่ทำ

R for statistics

https://cran.r-project.org/

- >Temp = c(33,33,34,36,35,34,33,33,32,32,31,30)
- >plot(Temp)
- >plot(Temp,type="l")
- > axis(1,at=1:length(Temp),
  labels=c("Jan","Feb","Mar","Apr","May","Jun","J
  ul","Aug","Sep","Oct","Nov","Dec"))


Temp
30
31
32
33
34
35
36
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Index

- 1.2.1. central tendency หาค่าที่เป็นตัวกลางของ sample ที่ทำ
- mode = ตัวที่เกิดมากสุด, mean ค่าเฉลี่ย,
  median = ค่าที่อยู่ตรงกลางของข้อมูล
  The mean is the sum of all the scores divided by the number of scores.
  The median is the middle of a distribution: half the scores are above the median and half are below the median.
  The mode is the most frequently occurring score in a distribution
- 1.2.2. dispersion/variability ดูลักษณะการกระจายของข้อมูลด้วย เพราะการเทียบ mean อย่างเดียวไม่พอ ข้อมูลที่มี mean เท่ากันอาจมีการกระจายตัวไม่เหมือนกัน
- variance, SD
  SD = sqrt(variance)

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

- ----------

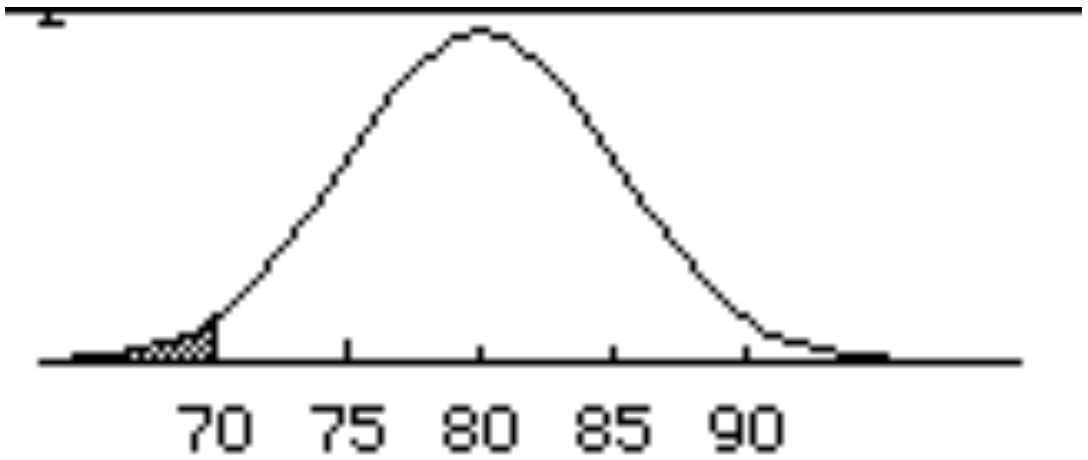
-2S -1S Mean +1S +2S

Mean +/- 1S = 68%

Mean +/- 2S = 95%

- **เมื่อรู้ค่า** Mean **กับ** SD **ก็สามารถคำนวณหา** percentile rank **ของคะแนนที่จุดต่างๆได้**
- **ตย. ค่า** mean = 80, sd = 5, percentile **ของคนที่ได้น้อยกว่า** 70 = 2.5%

(80 - 2 * 5 = 70)



R for statistics

```
> scoreA = c(23,24,23,21,17,15,25,21,18,23,15,27)
> mean(scoreA)
[1] 21
> sd(scoreA)
[1] 3.931227
> var(scoreA)
[1] 15.45455
> median(scoreA)
[1] 22
```

- 1.3 Effect statistics : one variable has an effect on another
- 1.3.1 Differences in means

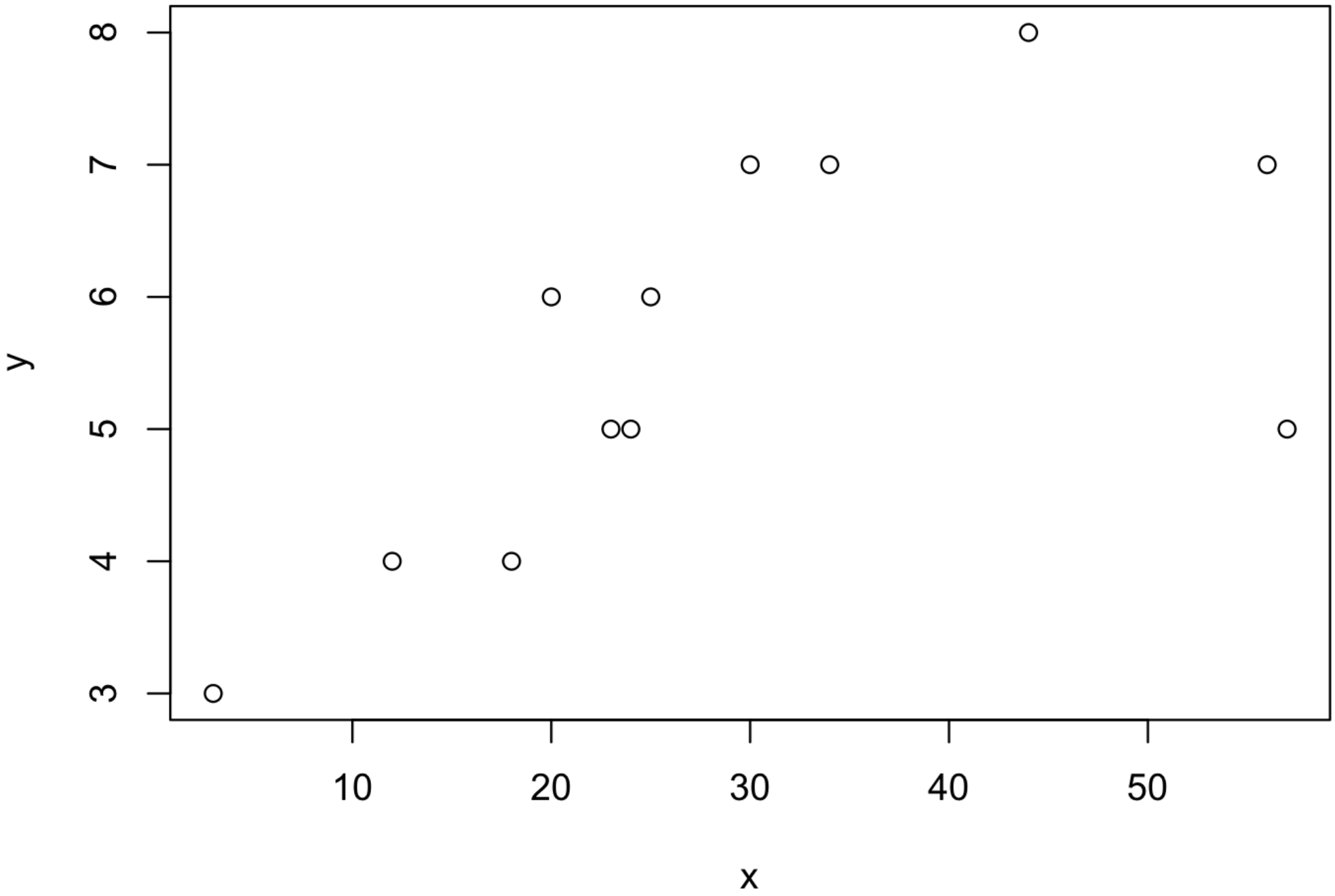
  ปกติ บอกความต่างเป็น % เช่น กลุ่ม 1, 2 ค่าเฉลี่ยน้ำหนักต่างกัน

  67.5 – 63.2/ 63.2 = 6.8%
- 1.3.2 Correlation coefficiency

  indicates the extent to which the pairs of numbers for these two variables lie on a straight line. มีค่า 0 ถึง 1

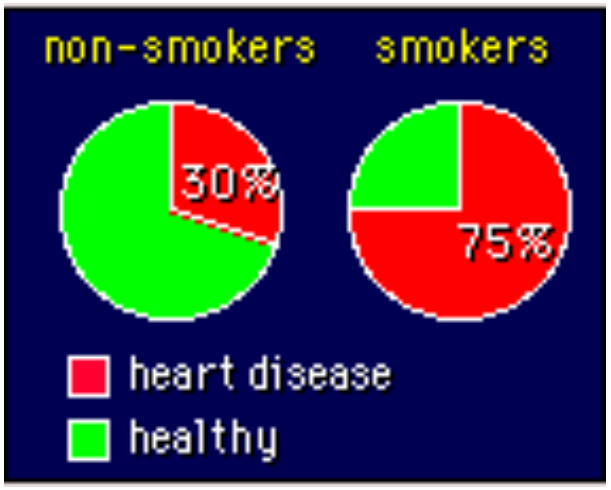
R for statistics

```
> x = c(12,23,24,25,3,34,56,44,30,57,18,20)
> y = c(4,5,5,6,3,7,7,8,7,5,4,6)
> cor(x,y)
[1] 0.6544815
>plot(x,y)
```


y
x
3
4
5
6
7
8
10
20
30
40
50

– 1.3.3 Relative frequency

the relative risk of developing heart disease for smokers is 2.5 (75/30)



- Odds ratio

  when the groups are **cases** and **controls**

  heart-disease group vs healthy group

  the odds of being a smoker in the heart-disease group are 75/25 = 3.

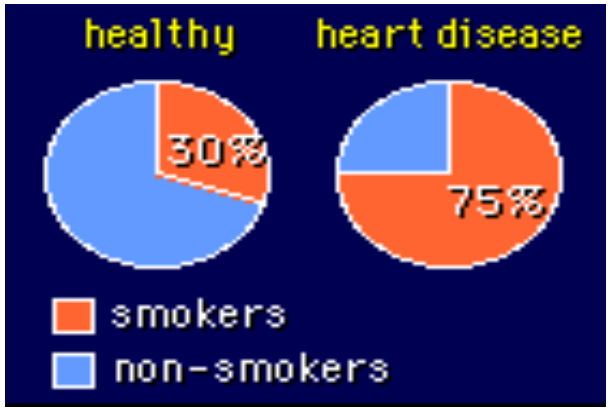
- the odds of being a smoker in the healthy group are 30/70 = 0.43.

  The odds ratio is therefore 3/0.43 = 7.

  Interpret this statistic as "seven people with heart disease smoke for every healthy person who smokes".

  พบคนสูบในกลุ่มคนเป็นโรคมากกว่าในกลุ่มคนปกติ

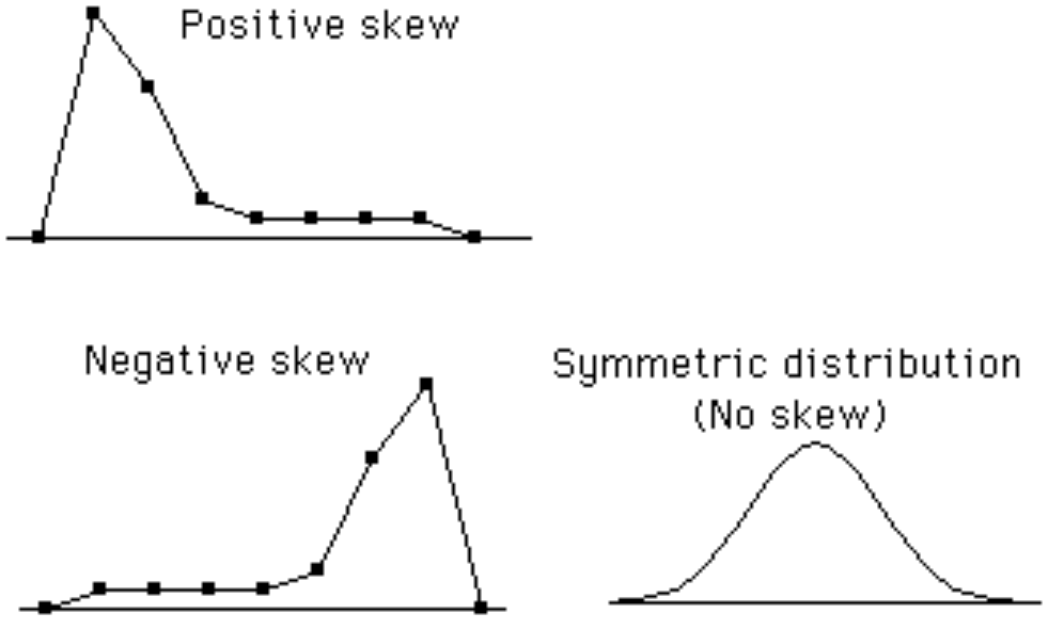
  http://www.sportsci.org/resource/stats/relfreq.html



- II. Inferential สถิติแบบอ้างอิง ศึกษาเพื่อนำไปอธิบายพฤติกรรมของ population โดยการศึกษาจาก sampling data วิธีการที่ใช้ใน inferential statistics คือ parameter estimation และ hypothesis testing
- 2.1 parameter estimation :
  - เป็นการใช้ sample เพื่อ estimate parameter ที่ต้องการทราบค่า
  - ในการ estimation คือ เราทำคล้ายแบบ descriptive คือหา mean, SD ของ sample data จากนั้นจึง infer ว่านี่เป็นพฤติกรรมของ population
  - mean ของ sample ที่มากพอจะใกล้กับ mean ของ population
  - ใช้ mean และ sd จาก sample เป็น point estimator ของ population (เป็น point estimator เพราะ estimate เป็น single number)

$$\sigma^2 = \Sigma(X-\mu)^2/N$$
$$s^2 = \Sigma(X-M)^2/(N-1)$$

- Mean จาก sample เป็น point estimator ได้ถ้าข้อมูลมี normal distribution
- ถ้าเป็น skewed distribution ใช้ไม่ได้



- แต่เนื่องจากเป็นการประมาณ ที่ assume normal distribution และเป็นการประมาณการ เราจึงต้องพูดถึง confidence intervals เช่น ค่า mean ที่ 5.9 จะมีค่าที่ 95% confidence อยู่ระหว่าง 4.7 ถึง 7.9
- ยิ่งถ้าให้มี confidence interval สูงขึ้น range จะกว้างขึ้น

- การคำนวณหา confidence interval มีสูตรเฉพาะคำนวณในหลายแบบ ตัวอย่าง
  - A 95% confidence interval for a single population mean is

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm 1.96 \sqrt{\frac{s^2}{n}}$$

    (1.96 คือ ค่า t-value ที่ 5% significant, infinite degree of freedom)
  - A 95% confidence interval for the difference between two population means is

$$(\bar{x} - \bar{y}) \pm 1.96 \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

- 2.2 hypothesis testing
  - คือการหาความสัมพันธ์ของตัวแปร โดยดูจาก sample data ที่ได้มา เพื่อสรุปความสัมพันธ์ของตัวแปรเหล่านั้น ใน population
  - วิธีการที่ใช้ทั่วไป คือ ตั้ง null hypothesis ขึ้นมาเพื่อที่จะพยายาม reject hypothesis นี้ (null hypothesis จึงตรงข้ามกับที่ต้องการ)
  - เช่น H0 เป็น null hypothesis ว่าไม่มีความแตกต่างระหว่างการเรียนด้วยวิธี X และวิธี Y alternative hypothesis H1 จะเป็นตรงกันข้ามกัน จากนั้นคำนวณค่าทางสถิติเพื่อ reject null hypothesis นี้โดยตั้งค่า statistical significance ไว้ เช่น ที่ .05 หรือ .01 (5% หรือ 1%)
  - ถ้าค่าที่คำนวณได้มี significant level น้อยกว่าหรือเท่ากับ 0.05 นี้ แสดงว่าเรามั่นใจได้อย่างน้อย 95% ที่จะ reject null hypothesis
  - แยกเป็น 2 ประเภท Non-parametric testing (ไม่ require special distribution) กับ Parametric testing

- หลังจากหาค่า mean สามารถตรวจว่าเป็น mean ของ population ด้วย
- H0: population mean = sample mean
- ถ้า H1 : population mean <> sample mean เป็น two-tailed test

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

- ถ้า H1: population mean > sample mean หรือ population < sample mean เป็น one-tailed test
- ดูค่า significant ที่ต้องการจากตารางสถิติ t-score ตาม df
- ใช้ t-test ตรวจสอบว่า sample นั้นมาจาก population ที่ต้องการหรือไม่
- ตย. มีข้อมูล population mean นร.ผ่านระดับ A เป็น 80 มีการเปลี่ยนการสอนในปีนั้น สุ่มตัวอย่างนร.มา 10 คนหา mean ได้ 71.5
  H0: mean นร.นี้มาจาก population ที่มี mean 80 (การเปลี่ยนการสอนไม่มี effect)
  H1: mean นร.นี้มาจาก population ที่มี mean < 80
  คำนวณ s, t-value เพื่อดู one-tailed test < 5% (< -1.83, df=9)

- Testing for differences between two populations
  - เป็น independent sample, test ความต่างของ mean, df = n1+n2-2
    assume มี sd เหมือนกันทั้ง 2 กลุ่ม = s
    ตย. เทียบผลนร. 2 กลุ่มที่ถูกสอนด้วยวิธีต่างกัน

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \qquad s = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

  - independent samples, เปรียบ variances (เพื่อ verify assumption ข้างบน)
    F = larger sample variance / smaller sample variance
    ดูค่าจากตาราง F-distribution ที่ df = n1(ตัวตั้ง), n2 (ตัวหาร)
  - paired samples: เทียบระหว่าง 2 กลุ่ม, df = n
    ตย. Pre-test vs post-test นร.กลุ่มเดิม
  -

$$t = \frac{\overline{d}}{s / \sqrt{n}} \qquad \overline{d} = mean\ of\ the\ observed\ differences$$

```
> group1 = c(26,22,27,15,24,27,17,20,17,30)
> group2 = c(22,18,26,17,19,23,15,16,19,25)
> t.test(group1,group2)

  Welch Two Sample t-test

data:  group1 and group2
t = 1.2423, df = 16.633, p-value = 0.2314
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.753072  6.753072
sample estimates:
mean of x mean of y
    22.5     20.0
```

```
> group1 = c(26,22,27,15,24,27,17,20,17,30)
> group2 = c(22,18,26,17,19,23,15,16,19,25)
> t.test(group1,group2,paired=TRUE)

  Paired t-test

data:  group1 and group2
t = 2.9531, df = 9, p-value = 0.01614
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5849444 4.4150556
sample estimates:
mean of the differences
                   2.5
```

- สถิติเพียงแค่ชี้ให้เราเห็นว่ามี cause-effect ในตัวแปรที่ศึกษา แต่ต้องมีคำอธิบายถึงความสัมพันธ์นั้น
- สถิติชี้ให้เห็นความเกี่ยวข้องการเป็นมะเร็งปอดกับการสูบบุหรี่
- แต่คำอธิบายอยู่ที่การหาสาเหตุที่บุหรี่ทำให้เกิดมะเร็งปอด
- ก่อนใช้สถิติ ต้องมี goal ที่ชัดเจนว่าต้องการศึกษาอะไร
- paired data : Two outcomes are paired when they are measured on the same observational unit.
- Pairing ทำให้เปรียบเทียบได้ดีกว่าการเทียบ mean เมื่อจำนวน subject น้อย ตย. เทียบคะแนน นร. Pre test, post test เป็นรายคน
- degree of freedom : จำนวนของ independent information ที่ใช้ในการ estimate parameter = จำนวน n - จำนวน parameter ที่ใช้ estimate ระหว่างทาง

  ตย. Variance = Sum (diff ^2) / df

  ใช้ mean ในการ estimate df จึง = N-1

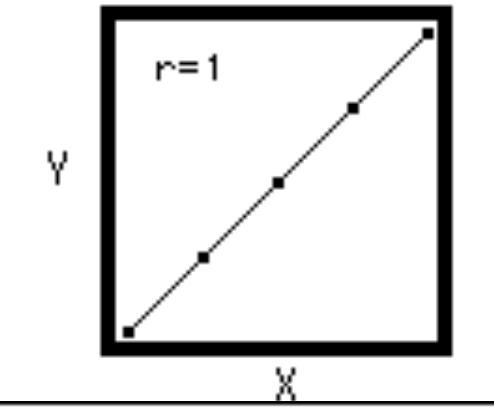
$$\sigma^2 = \Sigma(X-\mu)^2/N$$
$$s^2 = \Sigma(X-M)^2/(N-1)$$

- ชนิดของ data
  - nominal คือ data ที่สามารถ categorize ได้ว่าเป็นอะไร เช่น คำตอบว่า Yes - No nominal เป็นข้อมูลที่แยก category ต่างๆชัดเจน เช่น เพศ
  - ordinal เป็นข้อมูลที่มีการเรียงลำดับจากน้อยไปมาก แต่ตัวเลขไม่ได้มีค่าแท้จริง เช่น scale 1-5 อาจใช้เป็น 0-4 ก็ได้ ช่วงห่างระหว่าง 1-2, กับ 2-3 ไม่ได้มีนัยยะว่ามีความแตกต่างเท่ากัน
  - interval มีลักษณะของการเป็น scale ที่แต่ละช่วงห่างมีความหมายเท่าๆกัน เพียงแต่ว่าค่า ตัวเลขที่เป็นศูนย์ไม่ได้มีความหมายเป็น absolute zero ตัวอย่างเช่น scale วัดอุณหภูมิ ตัวเลข 30 องศาไม่ได้มีความหมายว่าร้อนเป็นสองเท่าของ 15 องศา
  - ratio คือค่า scale ของตัวเลขที่มีค่า absolute zero ตัวอย่างเช่น scale ของการวัดอุณหภูมิที่มีหน่วยเป็น kelvin คะแนนสอบของนักเรียน
  - หรือมอง 2 กลุ่ม เป็น label (nominal, ordinal) => non-parametric test หรือ numeric (interval,ratio) => parametric test

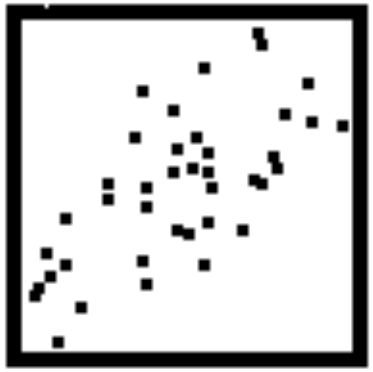
- **ชนิดของ** variable
  - dependent variable **เป็นตัวแปรที่เราสนใจ เพราะคิดว่าได้รับผลจากตัวแปรอื่น**
  - independent variable **เป็นตัวแปรต้นที่ส่งผลกระทบต่อตัวแปรที่สนใจ บางครั้งเรียก** predictor
  - **เช่น** IQ is affected by education
    IQ <= education (dependent <= independent)
  - **ชนิดข้อมูลตัวแปรตามตัวแปรต้นมีผลต่อการเลือก** stat **ที่ใช้**

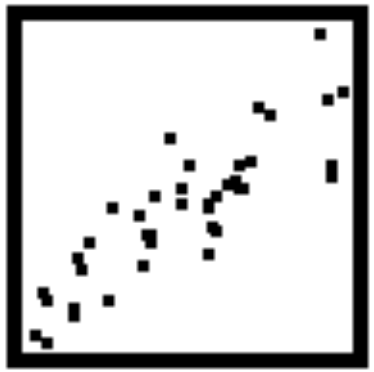
numeric <= numeric	Linear Regression
numeric <= nominal	T Test and One-Way ANOVA
nominal <= nominal	Contingency Table
nominal <= numeric	Categorical Modeling

- Linear Regression
  - หาความสัมพันธ์ระหว่างสองตัวแปร (numeric <= numeric)
  - สิ่งที่สำคัญคือ r หรือ slope ที่แสดงความสัมพันธ์ระหว่างตัวแปร



$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$


r=.67


r=.86


r=-.55


r=-.85

- T-test / One way ANOVA
  - หาความสัมพันธ์ระหว่างสองตัวแปร (numeric <= nominal) เช่น

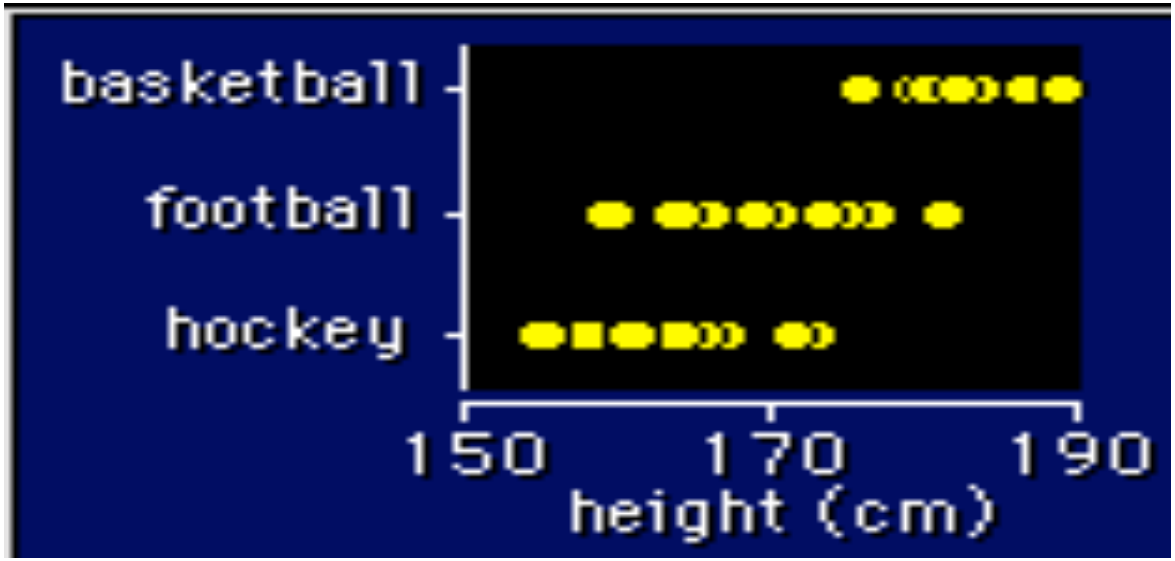
    ความสูง <= เพศ

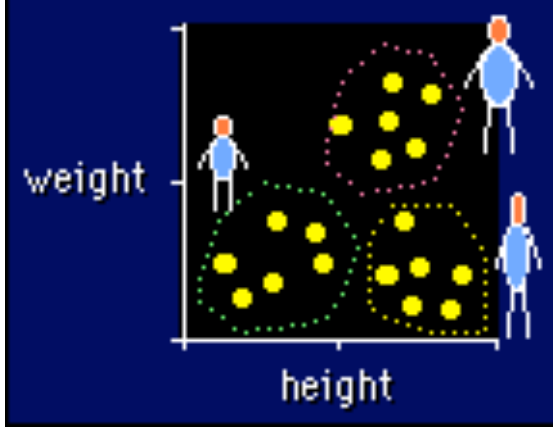
$$H_0: \mu_1 = \mu_2 = \ldots = \mu_a$$

  - กรณี nominal มี 2 group เช่น เพศ ใช้ t-test ได้
  - กรณี nominal มี > 2 group ใช้ one way ANOVA
  - ANOVA ตัดสินว่า sample ที่เลือกมา (>=2) มาจาก population เดียวกัน เหมือนกับ t-test แต่ใช้กับ sample มากกว่าสองกลุ่ม (ไม่สามารถทำโดยหา t-test ของ sample ทีละคู่ได้ เพราะมีปัญหาทางสถิติ เรื่องของ Type I Error ที่สูงขึ้น [reject H0 แต่ H0 ถูก])

- Two way ANONA
  - ดู independent variable 2 ตัวแปร เช่น
    ชั่วโมงออกกำลังกาย <= เพศ ประเภทกีฬา
    คะแนน <= วิธีสอน เพศ
- Three way Anona
  - ดู independent variable 3 ตัวแปร
- Manova (Multivariate analysis of variance)
  - กรณีที่มี dependent variable หลายตัว
  - ผลสัมฤทธิ์ ทัศนคติ <= วิธีการสอน

- Contingency Table
  - nominal <= nominal **เช่น** sport <= sex
  - Chi-Squared Test, Log Linear Analysis
- Categorical Modeling
  - nominal <= numeric **เช่น** sport <= height
  - **หรือเรียก** discriminant function analysis you end up with a function of height that allows you to predict which sport a person belongs in.





- Cluster Analysis
  - จัดกลุ่มข้อมูล อาจมีมากกว่า 2 ตัวแปรได้
  - เลือกได้ว่าต้องการกี่กลุ่ม หรือให้โปรแกรมจัด
- Principle Componential and Factor Analysis
  - ลด dimension หรือจำนวนตัวแปรลง อะไรที่ไปทิศทางเดียวกันรวมเป็น factor เดียวกัน อาจเหลือเป็น single score เช่น ดัชนีหุ้น ดัชนีผู้บริโภค
  - PCA แปลง var X1 X2 X3 X4 X5 … => Y1 Y2 Y3 Y4 Y5 … principle component แรกมีค่า VAR(Y1) มากสุด ไล่ไปเรื่อยๆ VAR รวมของ var เดิม = VAR รวม var ใหม่ ทำให้เลือก component ที่สำคัญน้อยจากเดิม
  - PCA ใช้ reduce variable PFA ใช้ detect structure
  - งานของ Biber วิเคราะห์ variation ใน text type ต่างๆ โดยใช้หลายๆตัวแปร ลดลงเหลือไม่กี่ dimension

- Anova numeric <= nominal
- Linear regression model numeric <= numeric
  - Linear regression model **เป็นการมองความสัมพันธ์ของ** independent variable **หนึ่งตัว** (y = a+b*x) **แต่ถ้ามีตัวแปร** independent variable **หลายตัว** model **จะเป็นแบบ** multiple regression model (y = a + b*x1 +c*x2 +d*x3 ...)
- Log-linear **และ** logistic regression model **ทั้งสองแบบนี้อยู่ภายใต้** statistical model **ที่เรียกว่า** generalized linear model (GLM)
- Logistic regression model nominal <= 0/1 Yes/No M/F
- Log-linear model nominal <= nominal

- งานวิจัยทางสังคมศาสตร์ใช้สถิติในกลุ่ม regression นี้มาก เพราะข้อมูลที่เก็บมีหลายตัวแปร ซึ่งเราไม่รู้ว่าตัวแปรไหนมีผลต่อตัวแปร dependent และไม่รู้ว่าแต่ละ independent variable มีผลต่อกันหรือไม่ด้วย
- จึงต้องใช้สถิติแนวนี้หา model ที่ fit กับข้อมูลที่รวบรวมมาได้ดีที่สุด แต่การที่จะใช้สถิติแบบนี้คำนวณต้องศึกษาให้เข้าใจหลักการและวิธีการตีความค่าตัวเลขต่างๆ ที่เกี่ยวข้อง ตลอดจนวิธีการปรับเปลี่ยน model เพื่อให้ได้ผลที่ดีที่สุด เช่น ต้องสร้าง model แบบที่ independent variable ไม่มีผลต่อกัน และสร้าง model แบบที่คิดว่า independent variable มีผลต่อกัน จากนั้นเปรียบเทียบ anova ของทั้งสองแบบว่ามีความแตกต่างอย่างมีนัยยะสำคัญหรือไม่ และในบรรดา independent variable ก็ต้องปรับ model เพื่อกันตัวแปรที่ไม่มีผลต่อ dependent variable ออกไป

- การใช้สถิติแบบนี้ จึงไม่สามารถทำการคำนวณได้ง่ายๆ แบบ chi-square หรือ anova
- ผู้ที่จำเป็นต้องใช้สถิติแบบนี้ จึงต้องศึกษาเพิ่มเติมอย่างมากเพื่อทำความเข้าใจวิธีการใช้และวิเคราะห์ model ต่างๆ เมื่อเข้าใจดีแล้ว ก็สามารถใช้คำสั่งในโปรแกรม R เพื่อช่วยคำนวณได้
- ผู้สนใจสามารถหาอ่านเพิ่มเติมได้ที่ multivariate statistics แยกเป็นสองกลุ่มใหญ่ๆ กลุ่ม analysis of variance กับกลุ่ม regression ซึ่งมีลักษณะคล้ายกันที่ใช้ดูว่าตัวแปรต้นไหนมีผลต่อ
- analysis of variance จะดีตรงที่ดู interaction effect ได้ดี รู้ว่าตัวแปรไหนมีผลต่อกัน ส่วน regression จะดีในแง่ของ prediction มากกว่า

Chi-square

- เป็น non-parametric test เพื่อหาว่าตัวแปรที่ศึกษามีความสัมพันธ์กันหรือไม่ เช่น หาว่ามีความสัมพันธ์กันหรือไม่ระหว่าง biological sex of American undergraduates at a particular university และ footwear preferences
- สุ่มตัวอย่างนิสิตชาย 50 คน หญิง 50 คน แล้วถามว่า เขาขอบสวมรองเท้าชนิดไหน โดยให้เลือก sandals sneakers, leather shoes, boot or something else
- ผลที่สำรวจได้จะแสดงออกมาในรูปของ bivariate table ซึ่งประกอบด้วย dependent variable (choice of shoes in this example) และ independent variable (sex in this example)

- คำถามคือการเลือกประเภทรองเท้า (dependent var) แปรตามเพศ (independent var) หรือไม่
- ข้อมูลแสดงในรูปของตาราง โดยที่ค่าของ independent variable จะจัดวางตามแกนตั้งและค่าของ dependent variable จะจัดวางตามแกนนอน เพื่อที่จะให้อ่านได้ตามแนวนอน

	Sandals	Sneakers	Leather shoes	Boots	Other
Male	6	17	13	9	5
Female	13	5	7	16	9

- Chi-square เป็นการทดสอบทางสถิติว่าความถี่ที่ได้จากการสุ่มตัวอย่างแสดงถึงความสัมพันธ์ของ 2 ตัวแปร ไม่ได้เกิดจากเหตุบังเอิญ

- Requiremnets for using chi square
  - 1.The sample must be randomly drawn from the population.
  - 2.Data must be reported in raw frequencies (not percentages);
    ส่วนหนึ่งของการคำนวณ chi square เป็นการ standardize data อยู่แล้ว จึงไม่ต้องทำ data ให้เป็นเปอร์เซ็นต์ก่อน (ซึ่งถือว่าเป็นการ standardize data แบบหนึ่ง)
  - 3.Measured variables must be independent;
    observation ที่ได้ต้อง independent คือ ไม่มีคำตอบใดที่ได้อิทธิพลจากคำตอบอื่น
  - 4.Values/categories on independent and dependent variables must be mutually exclusive and exhaustive;
    ค่าที่ observed ต้องตกลงใน category ใด category หนึ่ง เช่น subject ต้องตอบว่าชอบรองเท้าแบบไหนมากที่สุด ซึ่งจะมีคำตอบเดียว
  - 5.Observed frequencies cannot be too small.

- การคำนวณค่า chi square
  - เป็นการเปรียบเทียบระหว่าง observed frequency กับ expected frequency เพื่อดูว่า ค่าที่ observed ได้ไม่ได้มาจากความบังเอิญ (test against null hypothesis) ดังนั้น ก่อนอื่นจึงต้องคำนวณค่า expected frequency ในแต่ละ cell ก่อน
  - Ei = (row total X column total) / grand total

$$\chi^2 = \sum \frac{(|O - E|)^2}{E}$$

  - เทียบค่าที่ได้จากตาราง chi-square ที่ df = (no.row-1) x (no.colum-1)

    เช่น ที่ df =2, significant level = .05 (95% confidence level) มีค่า 3.84

- การคำนวณค่า chi square

สมมติว่าค่าที่ observe ได้ของการศึกษาความสัมพันธ์ระหว่าง tense and aspect เป็นดังข้างล่าง

Aspect \ Tense	Past tense	present tense	Total
Progressive	308	476	784
Non-progressive	315	297	612
Total	623	773	1396

ค่า expected frequency ของแต่ละ cell ก็จะเป็นดังข้างล่างนี้

Aspect \ Tense	Past tense	present tense	Total
Progressive	349.9	434.1	784
Non-progressive	273.1	338.9	612
Total	623	773	1396

E1 = (784 X 623) / 1396 = 349.9

E3 = (784 x 773) / 1396 = 434.1

E2 = (612 x 623) / 1396 = 273.1

E4 = (612 x 773) / 1396 = 338.9

- ได้ค่า 20.67 ซึ่ง > 3.87 จึง reject null hypothesis ได้

R for statistics

```
> data = matrix(c(308,315,476,297), nrow=2)
> chisq.test(data)

 Pearson's Chi-squared test with Yates' continuity
correction

data:  data
X-squared = 20.1602, df = 1, p-value = 7.122e-06
```

การหา collocation โดยสถิติ

- Collocation ที่ได้ต่างจาก collocation ที่นักภาษาศาสตร์พิจารณา
- บางครั้งเรียก probabilistic collocation vs linguistic collocation
- หลักการ => ดูสถิติในการปรากฏร่วมกันของคำ
- ดูจากความถี่ของ word1 - word2 ได้ไหม
- มีปัญหากรณีคำที่ปรากฏมากบังเอิญมาคู่กัน เช่น in the, of the, etc.
- ต้องคำนวณโดยดูนัยยะสำคัญทางสถิติ ไม่ใช่ความถี่
- มีวิธีคำนวณหลายแบบ เช่น chi-square, t-test, log likelihood, etc.

การหา collocation โดยสถิติ

- การใช้ความถี่
  - frequency อย่างเดียวไม่พอ eg. in the,
  - Justeson and Katz (1995) ใช้ POS กำหนด sequence ที่เป็นได้ eg. N-N, Adj-N
  - ปัญหา ไม่พบ collocation ที่มีความถี่ต่ำ
- การใช้ mean, variance
  - ใช้ collocation ที่ไม่จำเป็นต้องติดกัน eg. knock - door
    She knocked on his door They knocked at the door
    100 women knocked on Donaldson's door
    a man knocked on the metal front door

– หา mean, variance เพื่อดูระยะระหว่างคำทั้งคู่

– sd, var = 0 แสดงว่าระยะห่างจะคงที่ ใกล้ 0 น่าจะเป็น collocation

SD	Mean	Count	Word1	Word2
0.43	0.97	11657	New	York
0.48	1.83	24	Previous	Games
0.15	2.98	46	Minus	Points
0.49	3.87	131	Hundreds	Dollars
4.03	0.44	36	Editorial	Atlanta
4.03	0.00	78	Ring	New
3.96	0.19	119	Point	Hundredth
3.96	0.29	106	Subscribers	By
1.07	1.45	80	Strong	Support
1.13	2.57	7	Powerful	Organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	Said

ตาราง 6.1 การหาคำปรากฏร่วมโดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน

(Manning and Schutze 1999: 151)

Statistical Collocation : Hypothesis testing

- Null Hypothesis : คำ w1 w2 ปรากฏร่วมกันโดยบังเอิญ
- ดูจากข้อมูลจริงว่า w1 w2 ปรากฏร่วมกัน > เหตุบังเอิญหรือไม่
- ถ้า w1 w2 ปรากฏร่วมกันโดยบังเอิญ
  Prob(w1 ~ w2) = Prob(w1) * Prob(w2)
- Chisquare

	W1 = new	W1 ≠ new
W2 = companies	8 new companies	4667 e.g. old companies
W2 ≠ companies	15820 e.g. new machines	14287181 e.g. old machines

ตาราง 6.3 การเกิดร่วมกันของ new, companies, และคำอื่น

Statistical Collocation : Hypothesis testing

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- คำนวณออกมาได้ ≈ 1.55
- แต่ค่า chi-square ที่มีนัยยะสำคัญ 0.05 จะอยู่ที่ 3.84
- สรุปว่า new company ไม่น่าจะเป็น collocation
- สถิติแบบอื่น ๆ เช่น Log likelihood, t-test ก็ใช้หลักการทำนองเดียวกัน แต่รายการผลที่ได้อาจแตกต่างกันไป

- T-test
  - ถ้า corpus = 14,307,668 คำ มีคำว่า new = 15,828 company = 4,675
  - H0 : P(new company) = 15,828/14,307,668 x 4,675/14,307,668 = 3.615 x 10-7
  - หากคิดว่ามี process ที่จะสร้างตัวเลข 1 เมื่อพบ bigram "new company" และสร้างตัวเลขค่า 0 เมื่อพบ bigram อื่นๆ
    mean ของ process นี้ = 3.615 x 10-7 = population mean
  - แต่จากข้อมูล P(new company) จริง = 8/14,307,668 = 5.59110 x 10-9 ซึ่งถือได้ว่าเป็นค่าเฉลี่ย x-bar
  - ส่วนค่าความแปรปรวนนั้นสามารถคำนวณได้ว่าเท่ากับ p * (1-p) ซึ่งประมาณว่าเท่ากับ p ได้ในกรณีของ bigram เนื่องจากค่า p มีค่าต่ำมาก
  - ดังนั้นจึงสามารถคำนวณค่า t-test ในตัวอย่างนี้ได้ < 2.576 (sig .005)

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

- T-test
  - Church and Mercey (1993) ใช้ t-test มาใช้เพื่อช่วยเปรียบเทียบว่าในคำสองคำที่กำหนด มี collocation แตกต่างกันอย่างไร
  - ตย. strong, powerful
  - เนื่องจากค่า varaince ≈ p * (1-p) ซึ่งประมาณว่าเท่ากับ p ได้เพราะค่า p มีค่าน้อยมาก
  - จึงสามารถคำนวณค่า t ของคำใดใดที่ปรากฏร่วมกับ strong และ powerful ได้ดังนี้

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t \approx \frac{P(powerful\ \ w) - P(strong\ \ w)}{\sqrt{\frac{P(powerful\ \ w) + P(strong\ \ w)}{N}}} = \frac{\frac{C(powerful\ \ w)}{N} - \frac{C(strong\ \ w)}{N}}{\sqrt{\frac{C(powerful\ \ w) + C(strong\ \ w)}{N^2}}}$$

$$= \frac{C(powerful\ \ w) - C(strong\ \ w)}{\sqrt{C(powerful\ \ w) + C(strong\ \ w)}}$$

- T-test

t	C(w)	C(strong w)	C(powerful w)	word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition
3.9000	802	18	1	showing
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

ตาราง 6.2 การหาคำปรากฏร่วมโดยใช้ t-test

(Manning and Schutze 1999: 157)

- Berry-Rogghe's z-score
  - Berry-Rogghe (1973) ใช้วิธีดูความน่าจะเป็นที่คำๆหนึ่ง จะเกิดร่วมกับคำอื่นๆ ภายในขอบเขต (span) ที่กำหนด ถ้ากำหนดให้
  - Z เป็นจำนวนคำทั้งหมดภายในเอกสาร
  - A เป็นคำที่ต้องการศึกษาซึ่งปรากฏ Fn ครั้งในเอกสาร
  - B เป็นคำที่เป็น collocation ของ A ซึ่งปรากฏ Fc ครั้งในเอกสาร
  - K เป็นจำนวนครั้งของการเกิดร่วมกันของ B และ A
  - S เป็นขนาดของขอบเขต (span) หรือคือจำนวนคำที่อยู่ข้างใดข้างหนึ่งของคำที่ต้องการ
  - ขั้นแรก คำนวณความน่าจะเป็นที่จะพบ B ปรากฏร่วมกับ A เป็นจำนวน K ครั้ง โดยสมมติว่า B จะปรากฏแบบสุ่ม (random) ค่าที่ได้นี้ = ค่าความน่าจะเป็นที่คาดว่าจะพบ B ปรากฏร่วมกับ A จากนั้นจึงหาความแตกต่างระหว่าง expected number กับ observed number

- Berry-Rogghe's z-score
  - คำนวณค่าความน่าจะเป็นที่จะพบ B ปรากฏในตำแหน่งใดใดที่ A ไม่ปรากฏอยู่ ซึ่งจะ = p = Fc / (Z - Fn)
  - คำนวณจำนวนครั้งที่จะพบ B ปรากฏร่วมกับ A ภายในขอบเขต S ซึ่งได้เท่ากับ E = p * Fn * S
  - คำนวณว่า ค่าที่พบจริงๆ (observed) กับค่าที่คำนวณได้ (expected) มีความแตกต่างอย่างมีนัยยะสำคัญทางสถิติหรือไม่ ซึ่งสามารถคำนวณได้จาก z-score ดังนี้
    โดยที่ q = 1-p $z = (K - E) / \sqrt{E\,q}$
  - ที่ระดับ 1 % ค่า z-score จะต้องมากกว่า 2.576 และเมื่อรวมค่า z-score ของคำทุกคำที่คำนวณเทียบกับคำ A ผลที่ได้ควรจะมีค่าเป็น 0
  - ข้อจำกัดของวิธีทดสอบแบบนี้ คือ ไม่ยอมให้คำใดใดเป็น collocation ของตัวเอง

- Berry-Rogghe's z-score
  - Berry-Roggle ใช้วิธีนี้หา collocation ของ house ในหนังสือ A Christmas Carol, Each in his own Wildernes (แต่งโดย Doris Lessing) และ Everything in the Garden (แต่งโดย Giles Cooper)
  - ทดลองกำหนด span ต่างๆ จาก 3 ถึง 6
  - ใช้ค่า span 3 จะได้คำปรากฏร่วมของ house เป็น sold, commons, decorate, this, empty, buying, painting, opposite, loves, outside, lived, family, remember, full, my, into, the, has
  - ใช้ค่าเป็น 6 ได้ sold, commons, decorate, fronts, cracks, this, empty, buying, painting, opposite, loves, entered, black, near, outside, remember, lived, rooms, God, stop, garden, flat, every, big, my, into, family, Bernard, whole
  - Berry-Rogghe พบว่าค่าขอบเขตที่ควรจะใช้โดยทั่วไปคือ 4 ยกเว้นกรณีที่ต้องการหาคำปรากฏร่วมของคำคุณศัพท์ซึ่งควรจะใช้ค่าขอบเขตเป็น 2

- Berry-Rogghe's z-score
  - 1974 Berry-Rogghe ใช้วิธีการนี้ศึกษา phrasal verb เช่น look after, give in ซึ่งทำหน้าที่เหมือนเป็นคำๆ เดียว
  - ศึกษาโดยเอา particle เป็นตัวตั้ง หาคำกริยาที่ปรากฏหน้า particle นั้น
  - วิธีนี้สะดวกกว่าการเอาคำกริยาเป็นตัวตั้งแล้วหา particle ที่ปรากฏร่วมกับคำกริยานั้นเพราะว่าจำนวนคำที่เป็น particle มีน้อยกว่าจำนวนคำกริยา
  - ตัวอย่าง เมื่อหา collocation ของคำกริยาที่เกิดหน้า in ก็จะได้ interested, versed, lived, believe, found, live, ride, living, dropped, appeared, travelled, die, sat, died, interest, life, rode, stood, walk, find, house, arrived, came
  - การตัดสินว่า กริยาตัวไหน เมื่อเกิดร่วมกับ in แล้วเป็น idiomatic phrasal verb อาศัยหลักที่ว่า ความหมายของ idiomatic phrasal verb ควรจะต่างจากความหมายของส่วนประกอบย่อย

- Berry-Rogghe's z-score
  - ดูได้จาก collocation ที่ต่างกัน ตัวอย่างเช่น
    collocation ของ hot dog จะเป็น eat, mustard, stall
    collocation ของ hot จะเป็น weather, air, water
    collocation ของ dog จะเป็น bark, tail
    hot dog จึงมีลักษณะเป็นหน่วยคำเดียวมากกว่าจะเป็นสองคำ
  - Berry-Rogghe หาคำปรากฏร่วมของคำที่เกิดทางขวาของ in ได้รายการของคำที่เป็นคำปรากฏร่วมทางขวาของ in เดี่ยวๆ
  - จากนั้น นำคำที่ได้มาก่อน คือ คำกริยาที่สงสัยว่าจะรวมกับ in เป็น phrasal verb เช่น interested in, versed in, live in มาหาคำปรากฏร่วมทางขวาของ phrasal verb เหล่านี้ด้วย
  - คำนวณหาค่า R score โดยที่กำหนดให้ R = a/b โดยที่ a เป็นจำนวนคำที่เป็นคำปรากฏร่วมของ in และ verb+in ส่วน b เป็นจำนวนคำที่เป็นคำปรากฏร่วมของ verb+in

- Berry-Rogghe's z-score
  - ตัวอย่าง versed in มีคำที่เป็นคำปรากฏร่วมที่สำคัญอยู่ 3 ตัว คือ politics, history, Greek แต่ไม่มีคำไหนเลยที่เป็นคำปรากฏร่วมของ in ดังนั้น R = 0/3 = 0

    live in มีคำที่เป็นคำปรากฏร่วมที่สำคัญอยู่ 11 คำ คือ hut, house, *town, *country, *London, *room, *world, *place, family, happiness, ignorance ในจำนวนนี้มีอยู่ 6 คำที่เป็นคำปรากฏร่วมของ in ด้วย คือคำที่มีเครื่องหมาย * ใส่ไว้ข้างหน้า ดังนั้น R ของ live in = 6/11 = 0.54
  - ยิ่งตัวเลขที่ได้ต่ำเท่าไร phrasal verb ที่สงสัยนั้นก็มีแนวโน้มที่จะเป็น idiomatic มากขึ้น
  - z-score ใช้ในโปรแกรม SARA, TACT

- Log-likelihood
  - Hypothesis 1 : P(w2|w1) = p = P(w2| not w1) (w2 **เป็นอิสระจาก** w1)
  - Hypothesis 2 : P(w2|w1) = p1 <> p2 = P(w2|not w1) (w2 **ขึ้นกับ** w1)
  - p = C(w2)/N, p1 = C(w1 w2)/C(w1), **และ** p2 = (C(w2) - C(w1 w2)) / (N-C(w1))
  - assume **ว่าข้อมูลมีการกระจายตัวแบบ** binomial distribution

$$\log \lambda = \log \frac{L(H1)}{L(H2)} = \log \frac{b(c12, c1, p)b(c1 - c12, N - c1, p)}{b(c12, c1, p1)b(c1 - c12, N - c1, p2)}$$

$$= \log L(c12, c1, p) + \log L(c2 - c12, N - c1, p) - \log L(c12, c1, p1) - \log L(c2 - c12, N - c1, p2)$$

$$where \;\; L(k, n, x) = x^{k}(1 - x)^{n-k}$$

  - **คูณค่า** log likelihood **ด้วย** -2 **ดู** critical value **ของตาราง** chi-square **ค่านัยยะสำคัญที่** 0.005 (df = 1) **ต้องมากกว่า** 7.88 **จึงจะล้มสมมติฐานต้นว่าคำทั้งสองนั้นเป็นอิสระจากกันได้**

- Mutual Information
  - เป็นค่าที่วัดจากการเปรียบเทียบความน่าจะเป็นของการพบคำสองคำเกิดด้วยกัน หารด้วยความน่าจะเป็นที่จะพบคำทั้งสองโดยอิสระ

$$I(x',y') = \log_2 \frac{P(x'y')}{P(x')P(y')}$$

  - ข้อเสียกรณี perfect dependent คำมีความถี่น้อย MI มีค่าสูงกว่า
  - I(X,Y) = log2 P(X Y)/P(X)P(Y) = log2 P(X)/P(X)P(Y) = log2 1/P(Y) ค่าที่ได้จะขึ้นอยู่กับ P(Y)
  - ส่วนในกรณีที่การปรากฏของ X ไม่ขึ้นอยู่กับ Y เลย เราจะได้ = log2 P(X Y)/P(X)P(Y) = log2 P(X)P(Y)/P(X)P(Y) = log2 1 = 0
  - บางคนหลีกเลี่ยงปัญหาอันนี้โดยเลือกคำนวณค่า MI เฉพาะกับคำที่มีความถี่มากกว่าที่กำหนด เช่น 3 เป็นต้น หรือชดเชยค่าโดยนำเอาค่าความถี่มาคำนวณด้วยเป็น C(X Y) * I(X,Y)

- Mutual Information

I(x;y)	fxy	fx	Fy	x	y
10.47	7	7809	28	strong	northerly
9.76	23	7809	151	strong	showings
9.30	7	7809	63	strong	believer
9.22	14	7809	133	strong	second-place
9.17	6	7809	59	strong	runup
9.04	10	7809	108	strong	currents
8.85	62	7809	762	strong	supporter
8.84	8	7809	99	strong	proponent
8.68	15	7809	208	strong	thunderstorm
8.45	7	7809	114	strong	odor

ตาราง 6.4 คำปรากฏร่วมของ strong เรียงตามค่า MI

I(x;y)	fxy	fx	fy	x	y
8.66	7	1984	388	powerful	legacy
8.58	7	1984	410	powerful	tool
8.35	8	1984	548	powerful	storms
8.32	31	1984	2169	powerful	minority
8.14	9	1984	714	powerful	neighbor
7.98	9	1984	794	powerful	Tamil
7.93	8	1984	734	powerful	symbol
7.74	32	1984	3336	powerful	figure
7.54	10	1984	1204	powerful	weapon
7.47	24	1984	3029	powerful	post

ตาราง 6.5 คำปรากฏร่วมของ powerful เรียงตามค่า MI

- **การทดสอบแบบอื่นๆ**

	$W_1 = L_i$	$W_1 \neq L_i$
$W_2 = L_i$	a	c
$W_2 \neq L_i$	b	d

Simple matching coefficient (SMC) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$SMC = \frac{a+d}{a+b+c+d}$$

Kulczinsk coeffcient (KUC) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$KUC = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$$

Ochiai coefficient (OCH) ซึ่งให้ค่าระหว่าง 0 ถึง 1

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Frager and McGowan coefficient (FAG) ซึ่งให้ค่าตั้งแต่ - infinity ถึง 1

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{(a+b)}}$$

Yule coefficient (YUL) ซึ่งให้ค่าระหว่าง -1 ถึง 1

$$YUL = \frac{ad-bc}{ad+bc}$$

- การทดสอบแบบอื่นๆ

	$W_1 = L_i$	$W_1 \neq L_i$
$W_2 = L_i$	a	c
$W_2 \neq L_i$	b	d

Phi-squared coefficient ซึ่งให้ค่าระหว่าง 0 ถึง infinity

$$\Phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

Cubic association ratio (MI3)

เนื่องจากค่า Mutual Information ให้ความสำคัญกับเหตุการณ์ที่มีความถี่ต่ำมากไป จึงมีผู้เสนอ cubic association ratio (MI3) เพื่อให้น้ำหนักกับเหตุการณ์ที่เกิดบ่อยมากขึ้น

$$MI3 = \log_2 \frac{a^3 N}{(a+b)(a+c)}$$

- Collocation ระหว่างภาษา
  - ถ้ามี parallel corpus ที่ align ระหว่างประโยค
  - หา collocation ระหว่างคำในแต่ละภาษา ผลที่ได้เป็นคำที่น่าจะเป็นคำแปลในอีกภาษา ตย ใช้ MI $$I(e,f) = \log_2 \frac{p(e,f)}{p(e)p(f)}$$
  - p(e,f) เป็นค่าความน่าจะเป็นที่จะพบคำ e และ f ในประโยคที่ถูกจับคู่กัน p(e) เป็นค่าความน่าจะเป็นที่จะพบคำ e ใน ประโยคภาษาอังกฤษ p(f) เป็นค่าความน่าจะเป็นที่จะพบคำ f ในประโยคภาษาฝรั่งเศส
  - ตัวอย่าง prime มีคำภาษาฝรั่งเศสที่น่าจะเป็นคำแปลของคำนี้ คือ
  - sein MI=5.63 bureau MI=5.63 trudeau MI=5.34
    premier MI=5.25 residence MI=5.12 intention MI=4.57
    no MI=4.53 session MI=4.34
  - ตัดคำที่มีค่า MI สูงกับคำภาษาอังกฤษคำอื่น จะเหลือเพียง premier
  - จากการศึกษา พบว่าได้ผลถูกต้อง 65% 25% หาคำแปลภาษาฝรั่งเศสไม่ได้ 10% เลือกคำแปลมาผิด

- Collocation ของกลุ่มคำ >2
  - Silva and Lopes, 1999 เสนอวิธีคำนวณ collocation ของคำ >2
  - มอง n-gram เป็น pseudo bigram eg. w1-w2 + w3, w1 + w2-w3 แล้วมองหาค่า average ของทั้งกลุ่ม
- MI = log (P(x,y) / P(x)P(y))
  - log (P(w1w2w3) / P(w1)P(w2w3) or log(P(w1w2w3) / P(w1w2)P(w3)
  - Log(P(w1w2w3) / (P(w1)P(w2w3)+P(w1w2)P(w3))/2

$$SI\_f((w_1...w_n)) = \log(\frac{p(w_1...w_n)}{Avp}) \quad (4.4)$$

Where:

$$Avp = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n-1} p(w_1...w_i) \cdot p(w_{i+1}...w_n) \quad (4.5)$$

- Collocation ของกลุ่มคำ >2
- Chi2
  - 2 words
  - ถ้าแยกตำแหน่ง n-1

$$\phi^2((x,y)) = \frac{[f(x,y)\cdot N - f(x)\cdot f(y)]^2}{f(x)\cdot f(y)\cdot (N-f(x))\cdot (N-f(y))} \quad (4.6)$$

$$\phi^2(((w_1 \ldots w_{n-1}),w_n)) = \frac{[f(w_1 \ldots w_n)\cdot N - P]^2}{P\cdot (N-f(w_1 \ldots w_{n-1}))\cdot (N-f(w_n))} \quad (4.7)$$

Where

$$P = f(w_1 \ldots w_{n-1})\cdot f(w_n)$$

  - เฉลี่ยหา n-words association

$$\phi^2\_f((w_1 \ldots w_n)) = \frac{[f(w_1 \ldots w_n)\cdot N - Avp]^2}{Avp\cdot (N-Avx)\cdot (N-Avy)} \quad (4.8)$$

where Avp is determined according to (4.5)

$$Avx = \frac{1}{n-1}\cdot \sum_{i=1}^{i=n-1} f(w_1 \ldots w_i)$$

$$Avy = \frac{1}{n-1}\cdot \sum_{i=2}^{i=n} f(w_i \ldots w_n) \quad (4.9)$$

- Loglikelihood n-gram ก็คิดโดยหลักการเดียว