สถิติเบื้องต้นในแบบของวิทยาการข้อมูล

โดย อ.ดร.เสฏฐวิทย์ เกิดผล

Outline

* Descriptive Statistics
  * Central Tendency
  * Dispersion
  * Correlation
* Inferential Statistics สถิติเชิงอนุมาน
  * Population and Sample
  * Estimation
  * Hypothesis testing

Descriptive vs. Inferential Statistics

* **Descriptive Statistics** (สถิติเชิงพรรณา) คือ กระบวนการอธิบายหรือสรุปข้อมูลเชิงปริมาณและนำไปใช้งาน

* **Inferential Statistics** (สถิติเชิงอนุมาน) คือ กระบวนการคาดการณ์คุณสมบัติของสิ่งที่เราไม่รู้จากสิ่งที่เราวัดได้หรือเก็บข้อมูลมาแล้ว

* สถิติเชิงอนุมานจึงมี**ความไม่แน่นอน**และ**ความน่าจะเป็น**มาเกี่ยวข้องด้วย

Descriptive Statistics

* การสรุปข้อมูลมีหลัก ๆ 3 แบบ
  * ค่ากลางของข้อมูลชุดเดียว
  * การกระจายตัวของข้อมูลชุดเดียว
  * ความสัมพันธ์ระหว่างสองชุดข้อมูล

ค่ากลางของชุดข้อมูล

* ค่ากลางเปรียบเสมือนเป็น**ค่าปกติ**หรือ**ตัวแทนของกลุ่มข้อมูล**ทั้งหมด
* มาตรวัดที่ใช้เป็นค่ากลางได้
  * ค่าเฉลี่ย (mean / average)
  * มัธยฐาน (median)
  * ฐานนิยม (mode)

ค่ากลาง

* **ค่าเฉลี่ย:** บวกค่าทุกค่าเข้าด้วยกันแล้วหารด้วยจำนวนข้อมูล
  * ใช้กับค่าที่เป็นตัวเลข
  * อ่อนไหวต่อค่าที่สูงมากหรือต่ำมาก (outliers)
* **มัธยฐาน:** ค่าตรงกลางที่มีจำนวนข้อมูลที่มากกว่าและน้อยกว่าค่านี้เท่ากัน
  * ใช้กับค่าที่เป็นตัวเลขหรือลำดับก็ได้
  * ทนทานต่อค่าที่สูงมากหรือต่ำมาก

ค่ากลาง

* **ฐานนิยม:** ค่าที่เกิดขึ้นบ่อยที่สุด
    * ใช้ได้กับทั้งค่าตัวเลข ลำดับ และประเภท (category)
* ใน **pandas** มีคำสั่ง **mean()**, **median()**, **mode()** ในการหาค่าเฉลี่ย มัธยฐาน และฐานนิยม
* ใน **Excel/Sheets** จะใช้ **average()**, **median()** และ **mode()**

การกระจายตัว

* หรือ **dispersion** เป็นการวัดว่าข้อมูลกระจายตัวรอบค่ากลางอย่างไร
* มาตรวัด
  * ความแปรปรวน (variance)
  * ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)
  * พิสัย (range)
  * Inter-quartile range (IQR)

ความแปรปรวนกับส่วนเบี่ยงเบนมาตรฐาน

* **deviation** หรือ **error** คือผลต่างระหว่างค่าจริงกับค่าเฉลี่ย
* **variance** คำนวณจาก **deviation** ของทุกข้อมูลมายกกำลังสองแล้วหาค่าเฉลี่ย
  * การยกกำลังสองทำให้ **deviation** กลายเป็นค่าบวก
  * ยังเป็นการให้น้ำหนักกับค่า **deviation** ที่สูงกว่ามากขึ้น
* **standard deviation** คือรากที่สองของ **variance**
  * ถอดรากที่สองทำให้หน่วยของ **standard deviation** นั้นเป็นหน่วยเดียวกันกับข้อมูล เข้าใจง่ายกว่า **variance**

var() และ std()

* ใน **pandas** จะใช้คำสั่ง **var()** และ **std()** ในการคำนวณหาความแปรปรวนและส่วนเบี่ยงเบนมาตรฐาน
  * ทั้งสองคำสั่งใช้ **option ddof (degree-of-freedom)** ได้
    * **ddof=0** เมื่อคำนวณค่าของ **population** และ **1** เมื่อคำนวณค่าของ **sample**
* ใน **Excel / Sheets** จะใช้ **var.p()** และ **std.p()**
  * **p** ในฟังก์ชัน **var** และ **std** หมายถึง **population**
  * **s** ในฟังก์ชัน **var** และ **std** หมายถึง **sample**

มาตรวัดเกี่ยวกับลำดับ

* ถ้าเราเรียงลำดับข้อมูลจากน้อยไปมาก ข้อมูลลำดับแรกแน่นอนว่าคือค่าน้อยสุด ข้อมูลลำดับสุดท้ายคือค่ามากสุด
* พิสัย **(range)** คือค่ามากสุดลบค่าน้อยสุด
* **rank** ของข้อมูลคือ**ลำดับของข้อมูล**นั้นเมื่อเรียงข้อมูลทั้งหมดจากน้อยไปมาก
* **percentile** คือค่าที่จะทำให้ข้อมูลนั้นไปอยู่ใน **rank** ที่ระบุ (เป็นเปอร์เซ็น)
  * ดังนั้นค่าที่ **percentile** ที่ **35** จึงหมายความว่าจะมีข้อมูล **35%** ที่มีค่าน้อยกว่าหรือเท่ากับค่านี้

Inter-quartile range

Percentile	ความหมาย	quartile
0th	ค่าต่ำสุด	
25th		1th
50th	มัธยฐาน	2th
75th		3th
100th	ค่าสูงสุด	

* Quartile ที่ 1, 2, 3 คือค่าที่แบ่งข้อมูลเป็น 4 ส่วนเท่า ๆ กัน ตรงกับ percentile ที่ 25, 50, 75 ตามลำดับ

* **inter-quartile range** คือค่าของ quartile ที่ 3 ลบด้วยค่าของ quartile ที่ 1

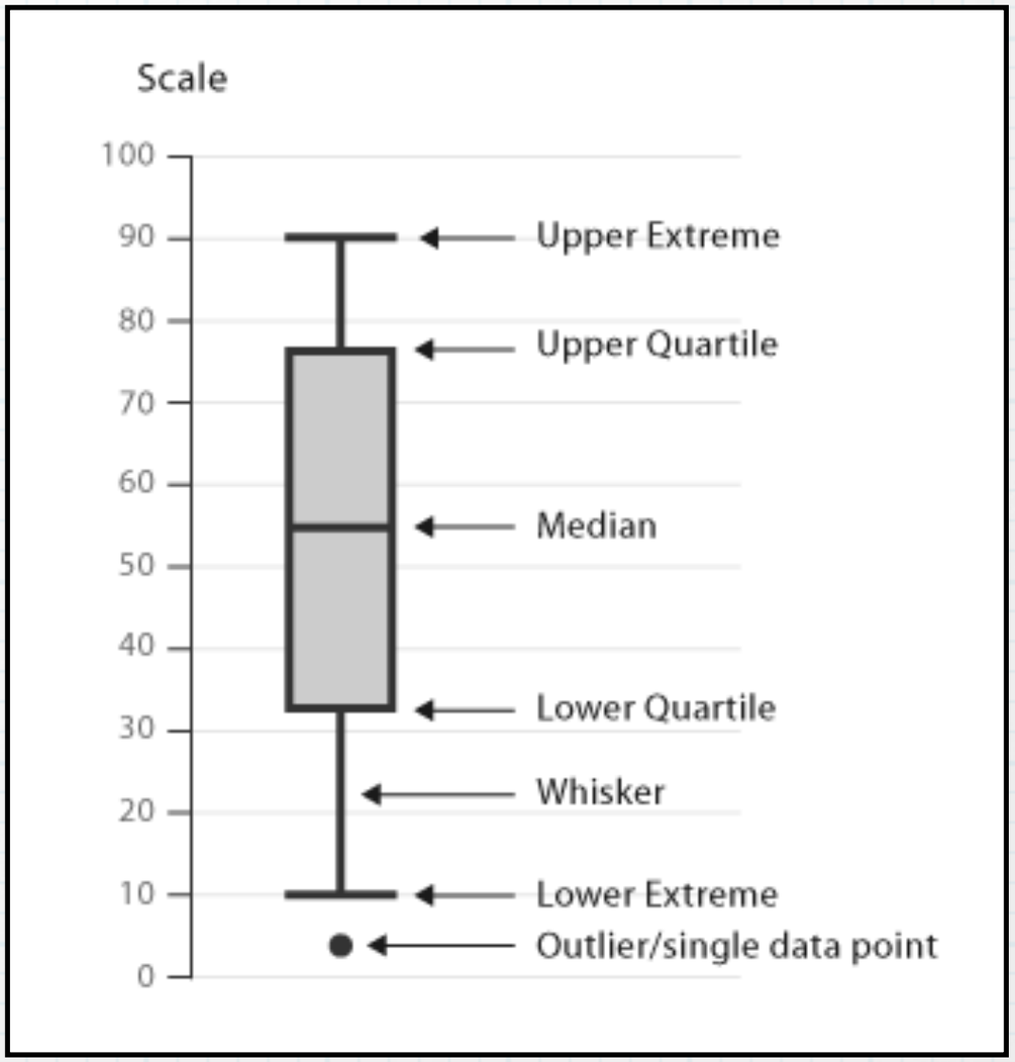
  * ใช้วัดการกระจายตัวของข้อมูล 50% ช่วงตรงกลาง

quantile() and percentile()

* ใน **pandas** จะใช้คำสั่ง **quantile()** ในการคำนวณค่าของ **percentile** โดยใส่เป็นเลขจุดทศนิยม
  * ระวังตัวสะกดของ **quantile()**
  * ไม่มีคำสั่งคำนวณ **quartile** โดยตรง ให้คำนวณ **quantile** ที่ตำแหน่ง **0.25, 0.5, 0.75** แทน
* ใน **Excel / Sheets** จะมีคำสั่ง **percentile()** และ **quartile()** แยกกัน

Box plot

* แสดงการกระจายตัวของข้อมูลในภาพเดียว
* **Upper/Lower Quartile** คือ **Q3/Q1**
* เส้น **Upper/Lower Extreme** จะอยู่ห่างจาก **Q3/Q1** เท่ากับ **1.5*IQR**
  * ยกเว้นถึงค่า **max/min** ก่อนจะอยู่ที่ค่า **max/min**
* ถ้ามี **outliers** จะ **plot** เป็นจุดในแผนภาพเลยเส้นไป



https://datavizcatalogue.com/methods/box_plot.html

การวาด box plot

* ใน seaborn ใช้คำสั่ง boxplot()
    * สามารถแยกวาด box plot ตาม category ได้ด้วย โดยกำหนดแกน x และ y ว่าจะวาดตามค่าใด

* ใน Excel เวอร์ชันใหม่ 2013+ มี box plot chart ให้ได้ใช้เลย
* ใน Google Sheets ยังไม่มี ต้องใช้เว็บอื่น เช่น http://shiny.chemgrid.org/boxplotr/

Histogram

* แสดงจำนวนข้อมูลของแต่ละประเภทหรือที่อยู่ในแต่ละช่วง
* ใช้แสดงการกระจายตัวของข้อมูลแบบตัวเลขได้เช่นกัน
* จำเป็นต้องคำนวณจำนวนและขอบเขตของแต่ละช่วง **(bin)** ที่จะจัดกลุ่มข้อมูล
* **seaborn** จะคำนวณขอบเขต **bin** ให้ตอนสร้างแผนภาพ หรือจะระบุเองก็ได้

การวาด Histogram

* ใน seaborn ใช้คำสั่ง distplot() ซึ่งจะวาดทั้ง histogram และ probability density function (PDF) ด้วย
  * กำหนดจำนวน bin หรือขอบเขตของ bin ได้ด้วย option bins
  * วาดหลาย histogram ซ้อนกันได้โดยเรียกคำสั่ง distplot() ต่อกันไป
* ใน Excel และ Google Sheets ก็มี histogram ให้ใช้เช่นกัน

การตีความ box plot และ histogram

* ทั้งสองแผนภาพใช้แสดงการกระจายตัวของข้อมูล
* "กล่อง"ของ **boxplot** แสดงข้อมูล **50%** ตรงกลาง และเส้น **median** แบ่งข้อมูลนี้ออกเป็น **25%** อีกสองส่วน
  * ใช้สังเกตได้ว่าข้อมูลเอนเอียงไปทางไหน
* เส้นหนวด **(whisker)** แสดงการยืดออกไปของ"หาง"ของข้อมูล ยิ่งยืดไปไกลหมายถึงข้อมูลของหางด้านนั้นยิ่งกระจาย **(long tail)**
  * ถ้าหางสองข้างยาวไม่เท่ากัน จะเรียกว่าข้อมูลเบ้(ขวา/ซ้าย) ไปทางหางที่ยาวกว่านั้น
* **histogram/pdf** แสดง"ความหนาแน่น"ของข้อมูลว่ากระจุกตัวอยู่ในช่วงไหน จะเห็นได้ชัดเจนว่าข้อมูลเบ้ไปทางด้านไหนหรือไม่

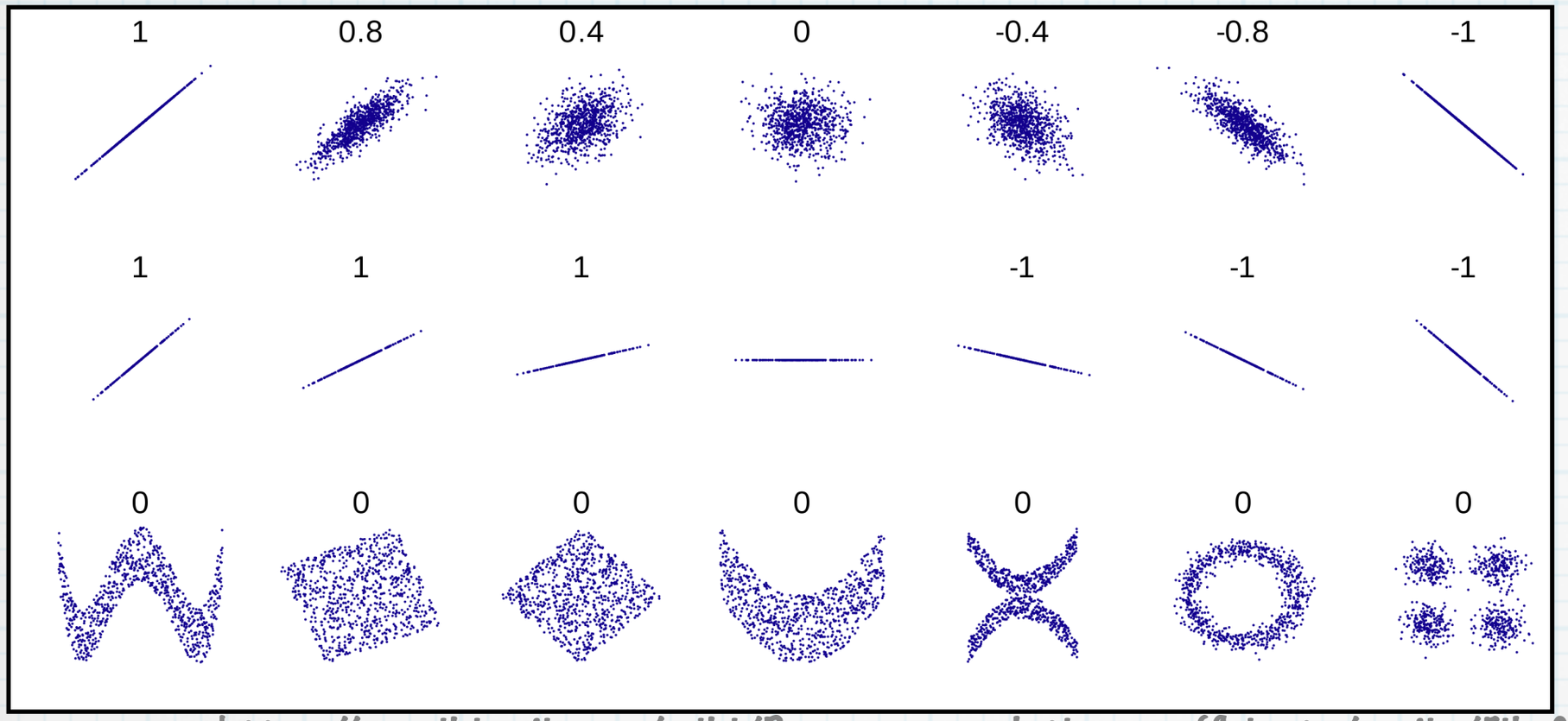
Correlation

* สหสัมพันธ์ คือ ความสัมพันธ์ทางสถิติระหว่างตัวแปรสองตัว
  * เมื่อตัวแปรหนึ่งเปลี่ยนไป ตัวแปรอีกตัวจะเปลี่ยนไปอย่างไร
  * คู่ตัวแปรที่มีค่าสหสัมพันธ์สูงสามารถใช้เป็นฟีเจอร์ในการสร้างโมเดลอธิบายหรือทำนายกันได้
* มาตรวัด
  * **Pearson's Product-Moment Correlation Coefficient (r)**
  * **Spearman's Rank Correlation Coefficient (ρ : rho)**
  * ใน **pandas** ใช้ฟังก์ชัน **corr()** ซึ่งกำหนด **option** ว่าจะคำนวณ **r** หรือ ρ ได้

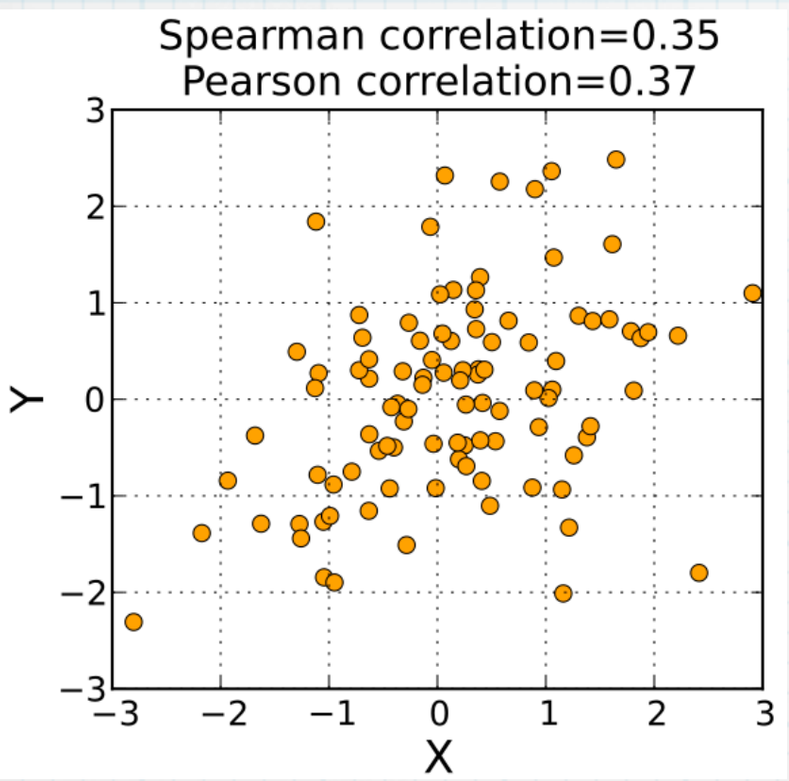
Pearson and Spearman

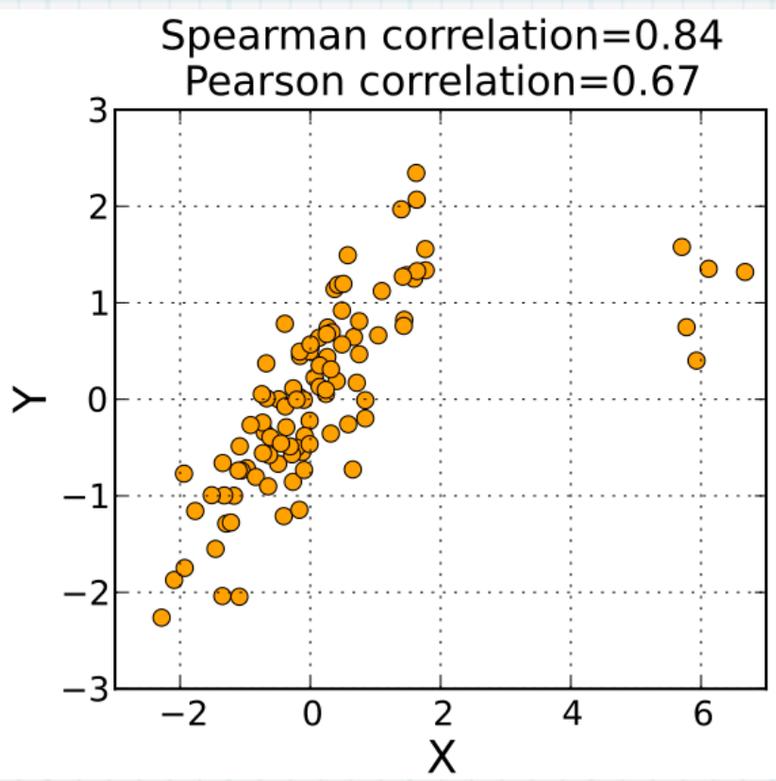
* Pearson's r
  * ความสัมพันธ์ระหว่างตัวแปรเป็นเส้นตรงหรือไม่
  * ค่าระหว่าง -1 (เส้นตรงความชันลบ) กับ 1 (เส้นตรงความชันบวก)
  * ค่า 0 คือตัวแปรสองตัวไม่สัมพันธ์กันเป็นเส้นตรง
* Spearman's rho
  * ความสัมพันธ์ระหว่างตัวแปรเป็นไปในทิศทางเดียวกันหรือไม่
  * ค่าระหว่าง -1 (ค่าของตัวแปรเปลี่ยนสวนทางกันเสมอ) กับ 1 (ค่าของตัวแปรเปลี่ยนไปในทางเดียวกันเสมอ)
  * ค่า 0 คือค่าของตัวแปรไม่เปลี่ยนตามกัน

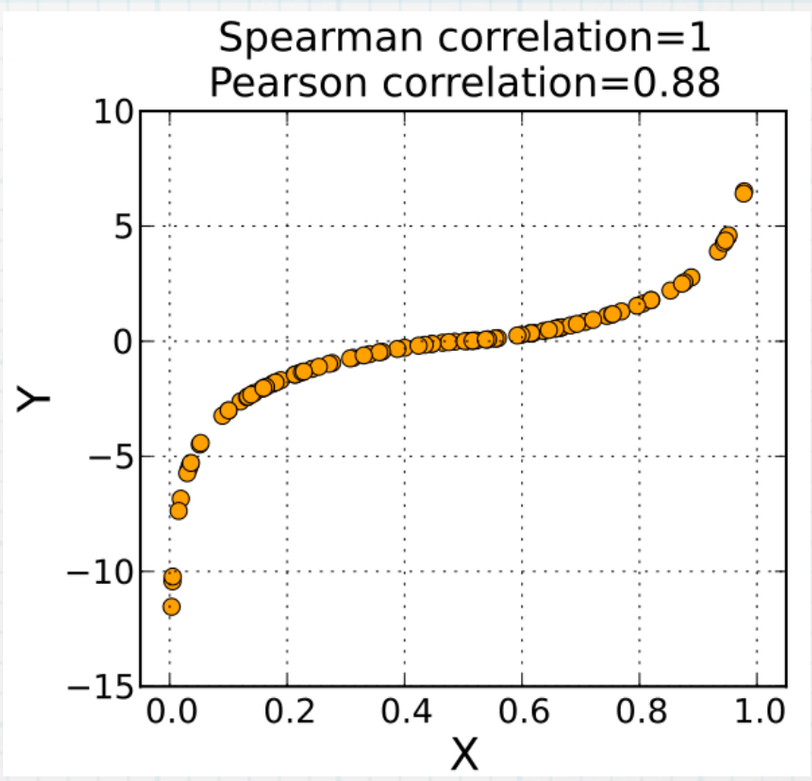
Pearson's r



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_examples2.svg







https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

สถิติเชิงอนุมาน

* เป็นการ**ประมาณค่าของสิ่งที่ไม่รู้**จากข้อมูลส่วนย่อยที่รู้
* ข้อมูลที่เรารู้ก็คือข้อมูลที่เราเก็บมาหรือสุ่มตัวอย่างมา เรียกว่า **sample**
  * เช่นคะแนนสอบของนักเรียนป.4 ที่สุ่มมาจำนวน 50 คน
* คุณลักษณะที่เราไม่รู้คือข้อมูลภาพรวมของประชากรที่เราสุ่มตัวอย่างมา เรียกว่า **population**
  * ค่าเฉลี่ยคะแนนสอบของนักเรียนป.4 ทั่วประเทศ
  * ค่าเฉลี่ยคะแนนสอบของนักเรียนป.4 ในปีหน้า

sample vs. population statistic

* การสุ่มตัวอย่างนั้นมี**ความไม่แน่นอน**เข้ามาเกี่ยวข้องด้วย
  * การสุ่มแต่ละรอบจะได้กลุ่มตัวอย่างที่ไม่เหมือนกัน
  * ดังนั้นค่าสถิติของแต่ละรอบก็ไม่เท่ากันด้วย
* เราจะประเมินความไม่แน่นอนที่ปนมากับค่าสถิติที่คำนวณจากกลุ่มตัวอย่างได้อย่างไร
  * จะ**ประมาณค่าของประชากรจากสถิติของกลุ่มตัวอย่าง**ได้ไหม
  * ค่าสถิติของกลุ่มตัวอย่างนี้**คลาดเคลื่อน**ไปจากค่าของประชากรแค่ไหน
* การประมาณค่าด้วยสถิติ(เช่นค่าเฉลี่ย)แบบนี้เรียกว่าการประมาณค่าแบบจุด **(point estimate)**

Law of large numbers

* ความไม่แน่นอนมีผลต่อกลุ่มตัวอย่างจำนวนน้อยมากกว่ากลุ่มจำนวนมาก
* กฎ **law of large numbers** กล่าวว่าถ้าเรา**สุ่มกลุ่มตัวอย่างมากขึ้นไปเรื่อย ๆ ค่าสถิติของกลุ่มตัวอย่างจะเข้าใกล้ค่าของประชากรมากขึ้นเรื่อย ๆ**
* แต่อาจไม่เป็นจริงในการกระจายตัวของประชากรบางรูปแบบ เช่น การกระจายตัวที่หางยาวและหนา (เบ้มาก)
* หรือการสุ่มตัวอย่างที่มีการเอนเอียง **(bias)**

ตัวอย่างความผิดพลาด

1936 Presidential Election Poll (prediction)

https://en.wikipedia.org/wiki/The_Literary_Digest#/media/File:LiteraryDigest-19210219.jpg

* 10 million surveys
* 2.4 million response
* predict Landon wins by 57%

VS

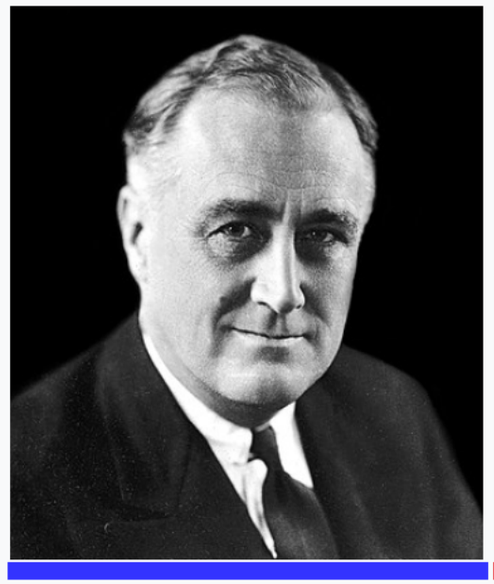
GALLUP

https://en.wikipedia.org/wiki/Gallup_(company)#/media/File:Logo_Gallup.svg

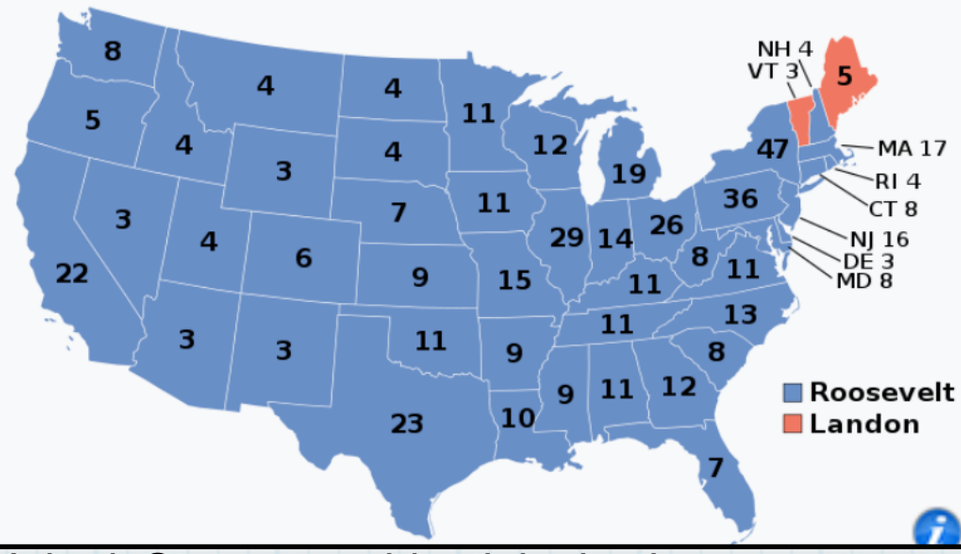
* 50,000 sample size
* predict Roosevelt wins
* also predict Literary Digest would mispredict

Result

* Digest ผิดไปเกือบ 20%
* ส่งแบบสอบถามไปยังสมาชิกและรายชื่อผู้ใช้รถและโทรศัพท์
  * selection bias เลือกเฉพาะคนฐานะปานกลางและรวย
  * nonresponse bias คนตอบแค่ 25%
* กลุ่มตัวอย่างไม่ใช่ตัวแทนประชากร



	Franklin D. Roosevelt	Alf Landon
Nominee		
Party	Democratic	Republican
Home state	New York	Kansas
Running mate	John Nance Garner	Frank Knox
Electoral vote	523	8
States carried	46	2
Popular vote	27,747,636	16,679,543
Percentage	60.8%	36.5%



https://en.wikipedia.org/wiki/1936_United_States_presidential_election

การประมาณค่าที่เอนเอียง

* **biased estimator** หมายถึงค่าสถิติที่คำนวณจากกลุ่มตัวอย่างแล้วจะเบี่ยงเบนไปจากค่าของประชากร แม้ว่าจะเพิ่มจำนวนตัวอย่างไปเรื่อย ๆ
* สถิติที่ไม่เอนเอียง เช่น ค่าเฉลี่ย
* สถิติที่เอนเอียง
  * เช่น **variance** และส่วนเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่างจะประมาณค่าของประชากรต่ำกว่าความเป็นจริง
  * สามารถปรับค่า **variance** ให้ไม่เอนเอียงได้ด้วย **degree of freedom (ddof** ในคำสั่งของ **pandas, var.s()** ใน **excel)**

สรุป point estimate

ค่า	กลุ่มตัวอย่าง	ประชากร	note
เฉลี่ย	$\bar{x}$	μ	unbiased estimator
variance	s^2	σ^2	bias ปรับแก้ได้
standard deviation	s	σ	bias ปรับแก้ยาก

การกระจายตัวของสถิติของกลุ่มตัวอย่างหลายกลุ่ม

* ถึงแม้จะสุ่มตัวอย่างโดยไม่มีการเอนเอียง เราก็ต้องการวิเคราะห์**ความคลาดเคลื่อน**ว่าสถิติของกลุ่มตัวอย่างใกล้เคียงกับค่าของประชากรแค่ไหน
  * เพราะค่าของสถิติของกลุ่มตัวอย่างสุ่มมาแต่ละครั้งไม่เท่ากัน
* ในเชิงทฤษฎี เราสามารถ**สุ่มตัวอย่างมาหลาย ๆ กลุ่ม** แล้วคำนวณค่าสถิติของแต่ละกลุ่ม
* การกระจายตัวของค่าสถิติของแต่ละกลุ่มมักจะเป็นแบบ **Normal distribution**
  * **Central Limit Theorem** สำคัญต่อทฤษฎีทางสถิติอย่างมาก
  * ในทาง **data science** นั้นไม่สำคัญมากนัก

Standard error

* เป็นมาตรวัดความน่าเชื่อถือของสถิติของกลุ่มตัวอย่าง ในการสุ่มครั้งนั้น
  * วัดว่าค่าสถิตินี้เบี่ยงเบนจากค่าจริงของประชากรแค่ไหน
* คำนวณจาก **standard deviation** ของสถิติของกลุ่มตัวอย่างหลายกลุ่ม
* ถ้าสถิติคือค่าเฉลี่ย คำนวณ **SE** ได้จาก **standard deviation** ของประชากร หารด้วยรากที่สองของจำนวนตัวอย่าง $\frac{\sigma}{\sqrt{n}}$
  * แต่หมายความว่าเราต้องรู้ **sd** ของประชากร
  * ถ้าไม่รู้ก็ประมาณเอาได้จาก **sd** ของกลุ่มตัวอย่าง $\frac{s}{\sqrt{n}}$

Confidence Interval

* ช่วงความเชื่อมั่น เป็นการประมาณค่าแบบช่วงเพื่อระบุความคลาดเคลื่อนของสถิติของกลุ่มตัวอย่าง
* เมื่อเราสุ่มหลายกลุ่มตัวอย่างและคำนวณค่าสถิติแล้ว ให้หา **percentile** ที่ **95** และ **5** ของค่าสถิติ
* จะระบุถึง **confidence interval** ที่ระดับความเชื่อมั่น **90%**
* ความหมายคือค่าสถิติ **90%** ตรงกลางของสถิติจากหลายกลุ่มตัวอย่าง จะอยู่ในช่วงนี้
* ถ้าต้องการความเชื่อมั่น **95%** ก็หาค่าที่ **percentile 97.5** และ **2.5**

Population

Sample 1

Sample 2

Sample n-1

Sample n

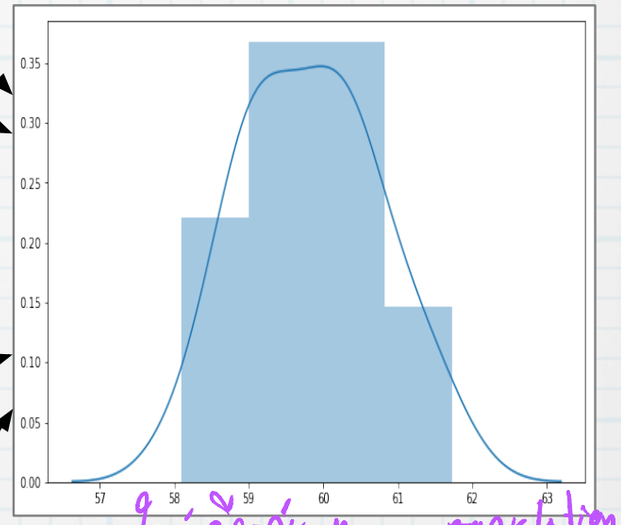
mean 1

mean 2

mean n-1

mean n

Sampling Distribution



* mean of means
* standard deviation of means is standard error
* 95th/5th percentile is confidence interval (90% confidence level)

จบประเด็น?

* เนื้อหาที่อธิบายไปใช้วิธีสุ่มตัวอย่างหลาย ๆ กลุ่ม ซึ่งอาจทำไม่ได้จริงในทางปฏิบัติ
* จะคำนวณ **point estimate** กับ **confidence interval** จากกลุ่มตัวอย่างเดียวได้อย่างไร

* **point estimate** ก็คือคำนวณสถิติของกลุ่มตัวอย่างเดียวแล้วนำมาใช้เลย
* **confidence interval (95% confidence level)** ใช้สูตรได้
  * $\pm 1.96 \times SE$ จากค่าเฉลี่ยของกลุ่มตัวอย่าง (บนสมมติฐานว่าประชากรกระจายตัวแบบ **normal distribution**)
  * ถ้าการกระจายตัวเป็นแบบอื่น ก็เลือก **distribution** ที่เหมาะสมได้

Bootstraping

* หนึ่งในเทคนิคที่เรียกว่า **resample**
* สร้าง"ประชากรจำลอง"ขึ้นมาจากกลุ่มตัวอย่าง แล้วสุ่มกลุ่มตัวอย่างใหม่หลาย ๆ กลุ่มจากประชากรจำลองนี้
* ในทางปฏิบัติ จะใช้การสุ่มข้อมูลจากกลุ่มตัวอย่างแบบมีการใส่ข้อมูลคืน
* ความถูกต้องของการทำ **resample** จะขึ้นกับกลุ่มตัวอย่างตั้งต้นว่าเป็นตัวแทนของประชากรได้ดีแค่ไหน
* ใช้พลังการคำนวณอย่างมาก ต้องทำด้วยคอมพิวเตอร์
* ข้อดีคือไม่จำเป็นต้องตั้งสมมติฐานใดเกี่ยวกับการกระจายตัวของประชากร

Simulated Population

Sample

Resample 1

Resample 2

Resample n-1

Resample n

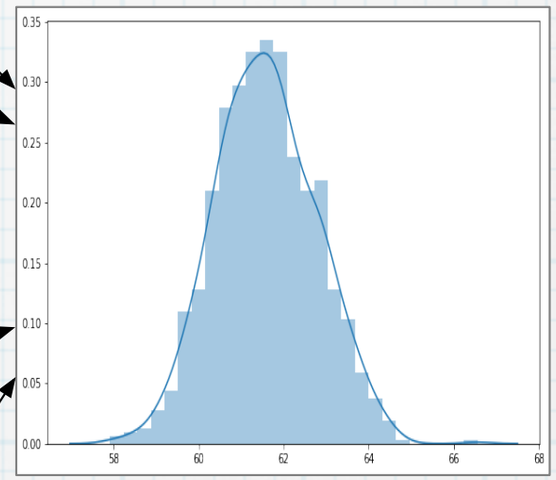
mean 1

mean 2

mean n-1

mean n

Sampling Distribution



* mean of means
* standard deviation of means is standard error

ความหมายของ Confidence Interval

* ค่าเฉลี่ยอยู่ที่ **58.44-61.34** ที่ความเชื่อมั่น **90%** ?
* การตีความที่ผิด: มีโอกาส **90%** ที่ **population mean** จะอยู่ในช่วงนี้
  * เพราะค่าเฉลี่ยประชากรนั้นคงที่
* การตีความที่ถูก:
  * ค่าเฉลี่ยของกลุ่มตัวอย่าง **90%** จะอยู่ในช่วงนี้
  * ถ้าสุ่มตัวอย่างและสร้าง **confidence interval** ไปเรื่อย ๆ **90%** ของ **CI** ที่สร้างขึ้นจะคร่อมค่าเฉลี่ยของประชากร

การทดสอบสมมติฐาน

* สิ่งที่เห็นเป็นเพียง**ความบังเอิญ**หรือไม่ — เช่น โยนเหรียญ 10 ครั้ง ได้ $n=7, \tau=3$ แล้วเหรียญอันนี้เที่ยงตรงหรือไม่ ที่โยนได้แบบนี้ชั่งเอิญหรือไม่
  * ความบังเอิญเกิดจากการสุ่มตัวอย่างทั้งจากการสังเกตหรือการทดลอง
* เทียบกับ **confidence interval** ที่ตอบคำถามว่า ความบังเอิญในการสุ่มตัวอย่างทำให้เกิดความคลาดเคลื่อนในการประมาณค่าแค่ไหน
* ทั้งสองวิธีเป็นการประเมินความไม่แน่นอนแต่มองคนละมุม
  * ใช้หลักการ **resample** หรือสูตรคำนวณได้เหมือนกัน

ตัวอย่าง

* บริกรเก็บข้อมูลการให้ทิปของลูกค้า
* ลูกค้าผู้ชาย **157** คน ผู้หญิง **87** คน
* พบว่าค่าเฉลี่ยของทิปจากลูกค้าผู้ชายสูงกว่าจากลูกค้าผู้หญิงอยู่ **$0.256**
  * เป็นความบังเอิญจากลูกค้าที่บริกรคนนี้ให้บริการพอดี
  * หรือเป็นความแตกต่างระหว่างเพศจริง ๆ

กระบวนการทดสอบสมมติฐาน

หัวเกิดขึ้น 7 ครั้งเป็นเรื่องบังเอิญ!

* ตั้งสมมติฐานก่อนว่าสิ่งที่เห็นเป็นความบังเอิญ (H_0: null hypothesis)

ไป หัวก้อย 7 ครั้ง ซึ่งโอกาสมากกว่า α ดังนั้นมันจึง

  * H_1 (alternate hypothesis) เป็นสมมติฐานแย้ง ว่าสิ่งที่เห็นเป็นจริง

เกิดขึ้นได้ปกติน่ะ

  * ถ้าเราโต้แย้ง H_0 ได้เราจะยอมรับ H_1

* กำหนดระดับนัยสำคัญ (α) เป็นเปอร์เซ็น ← α เราต้องตั้ง

  * เป็นความเสี่ยงที่ยอมให้เกิดขึ้นจากการยอมรับสมมติฐานว่ามันเป็นจริงทั้งที่เป็นเหตุบังเอิญ
  * กำหนดไว้ก่อนที่จะทดสอบสมมติฐาน
  * ค่าที่นิยมใช้เช่น **0.1** , **0.05** , **0.01**

10% 5% 1%

} ใส่ค่า α ได้ง่ายใจก่อน

ทดสอบสมมติฐาน

* แนวคิดคือเราสมมติว่าสิ่งที่เห็นเป็นความบังเอิญ
* แล้วทดสอบว่าสิ่งที่เห็นอย่างน้อย**มีโอกาสเกิดขึ้นบ่อยครั้งแค่ไหน ถ้ามันเป็นความบังเอิญ**
  * ค่านี้เรียกว่า **p-value**
  * ทดสอบโดยใช้สูตรหรือ **resample** ก็ได้
* ถ้าโอกาสนั้นต่ำกว่าระดับนัยสำคัญที่ตั้งไว้ เราจะปฏิเสธ H_0 และยอมรับ H_1

ทดสอบแบบ resample

* **คละคนทุกเพศเข้าด้วยกัน** (เพราะเราสมมติว่าความแตกต่างในการให้ทิประหว่างเพศเป็นความบังเอิญ)
* **จัดแบ่งกลุ่มใหม่โดยไม่สนเพศ** เป็นกลุ่ม **A 157** คนและกลุ่ม **B 87** คน ขั้นตอนนี้คือการ **resample**
* คำนวณความแตกต่างระหว่างค่าเฉลี่ยการให้ทิปของกลุ่ม **A** และ **B** แล้วบันทึกไว้
* ทำซ้ำหลาย ๆ รอบ เช่น **1000** รอบ
* วิเคราะห์ว่าค่าของ **1000** รอบนั้นมีสัดส่วนเท่าไหร่ที่มากกว่าหรือเท่ากับ **0.256** สัดส่วนนี้คือ **p-value**

คำนวณ p-valve จากสูตร

* ทางสถิติมีการสร้างสูตรประมาณการกระจายตัวของค่าในรูปแบบต่าง ๆ ไว้แล้ว
    * เช่น ความแตกต่างระหว่างค่าเฉลี่ยจากสองกลุ่มจะกระจายตัวแบบ t
* เราสามารถคำนวณค่าของ t แล้วหา p-value ได้
    * ใน pandas สามารถคำนวณได้จาก scipy.stats.ttest_ind() ซึ่งคำนวณแบบ two-tail
* ถ้าเราทดสอบค่าแบบอื่น เช่น variance หรือสัดส่วน อาจต้องใช้การทดสอบแบบอื่น

การตีความ p-value

* p-value บ่งบอกถึงความน่าจะเป็นที่เหตุการณ์ที่มากขนาดที่เห็นหรือมากกว่า เกิดเพราะเหตุบังเอิญ

* เช่น ถ้าการให้ทิปที่แตกต่างกันระหว่างเพศชายกับเพศหญิงเป็นเหตุบังเอิญ โอกาสที่จะเห็นความแตกต่างที่ผู้ชายให้ทิปมากกว่าผู้หญิงอย่างน้อย $0.256 คือ (ค่า p-value ที่ได้)

สรุปผล

* ถ้า **p-value** ต่ำกว่าค่า α (ระดับนัยสำคัญที่ตั้งไว้) หมายความว่า
    * โอกาสที่จะเกิดเหตุการณ์นั้นน้อยกว่าค่าที่ตั้งไว้ แสดงว่าเป็นเหตุการณ์ที่ค่อนข้างเกิดน้อย ดังนั้นไม่น่าจะเป็นความบังเอิญ
    * เราจึงปฏิเสธสมมติฐานว่าสิ่งที่เห็นเป็นเหตุบังเอิญ และยอมรับสมมติฐานว่าความแตกต่างนั้นเป็นจริง
* ถ้า **p-value** ไม่ต่ำกว่าค่า α แสดงว่าเหตุการณ์นี้ไม่ได้เป็นเรื่องที่แปลกประหลายแต่อย่างใด อาจเกิดขึ้นเพราะความบังเอิญได้
    * เราจะไม่สามารถปฏิเสธสมมติฐานว่าสิ่งที่เกิดเป็นความบังเอิญ

เปรียบเทียบ resample กับ สูตร

* การใช้สูตรนั้นทำได้อย่างรวดเร็ว แต่ผู้ใช้จำเป็นต้องรู้ว่าจะใช้การทดสอบด้วยสูตรไหนในกรณีใด
* สูตรมีความซับซ้อนและมีสมมติฐานเบื้องหลังมากมาย อาจทำให้เกิดความผิดพลาดในการประมาณค่า
  * เช่น ประชากรกระจายตัวอย่างไร ค่า **variance** ของสองกลุ่มประชากรเท่ากันไหม จำนวนตัวอย่างมากน้อยแค่ไหน
* การ **resample** ใช้เวลาและพลังการคำนวณเยอะ แต่ **เข้าใจง่าย และไม่มีสมมติฐานเกี่ยวกับประชากร**

ประเด็นเกี่ยวกับค่า α

* α เป็นความเสี่ยงที่ยอมให้เกิดขึ้นจากการยืนยันสมมติฐานว่ามันเป็นจริงทั้งที่เป็นเหตุบังเอิญ
* ค่า α ควรตั้งไว้ที่เท่าไหร่
* ขึ้นกับความแน่นอนในแต่ละสายงาน
* เช่นสายงานด้านวิศวกรรมอาจตั้งค่า α ไว้ต่ำมาก เช่น **0.0001** เพราะทดสอบกับสิ่งที่สร้างเอง
* สายงานทางสังคมอาจตั้งค่า α ไว้ที่ **0.1** เพราะมีองค์ประกอบที่ควบคุมไม่ได้มาเกี่ยวข้องเยอะ
* สายงานด้านสุขภาพต้องการความแน่นอนมากกว่าด้านสังคม อาจตั้งค่า α ที่ **0.05** หรือ **0.01**

ความผิดพลาดในการยอมรับสมมติฐาน

* **Type I error:** ยอมรับว่ามีผลกระทบจริงทั้ง ๆ ที่จริง ๆ แล้วเป็นเหตุบังเอิญ
  * **False Positive**
  * โอกาสเกิด **Type I error** ระบุด้วยค่า α
* **Type II error:** ยอมรับว่าผลเกิดจากความบังเอิญ ทั้ง ๆ ที่มีสาเหตุเบื้องหลังจริง
  * **False Negative**
* การปรับค่า α จะเป็นการแลกเปลี่ยน **error** สองประเภทนี้

การตีความผลการทดลอง

* ค่า **p-value** ในปัจจุบันใช้เป็นแนวทางในการแสดงความไม่แน่นอนของสมมติฐานและความเสี่ยงที่จะเกิด **Type I error**
  * ไม่ควรใช้เป็นเกณฑ์ในการตัดสินนัยสำคัญทางสถิติแบบเด็ดขาด
* นั่นคือไม่ควรตัดสินทันทีว่าถ้า $p < 0.05$ แล้วสมมติฐาน **H1** จะถูกต้อง ควรมองค่า **p** เป็นระดับความเสี่ยงแทน
* **p-value** ไม่ได้บ่งบอกถึงความรุนแรงของผลกระทบถ้าสมมติฐานเป็นจริง
  * เช่นถ้าผู้ชายให้ทิปมากกว่าผู้หญิงจริง จะให้มากกว่าเท่าไหร่
* ใช้ประกอบกับ **Confidence Interval** เพื่อวิเคราะห์อย่างละเอียด

Reference

1. Peter C. Bruce, "Introductory Statistics and Analytics: A Resampling Perspective", John Wiley & Sons, December 2014.

2. "1936 United States presidential election", wikipedia.com, retrieved on Nov 26, 2019.