การวิเคราะห์โดยใช้สถิติ

Statistical Analysis

Preparing Data for Statistical analysis

- Cleaning up data
  - ค้นหาจุดที่คิดว่าอาจมีข้อผิดพลาด เช่น ข้อมูลที่เป็นไปได้ หรือ ความคงเส้นคงวาข้อมูล (เพศ : M:0:1:ช)
- Coding Data ในบางกรณีต้องมีการแปลงข้อมูล เช่น ชาย → 1 หรือ ช่วงเงินเดือน 15000 – 30000 -→ 1 เป็นต้น
- Organizing Data
  - การจัดรูปแบบข้อมูลสำหรับการประมวลผลสำหรับแต่ละเครื่องมือ เช่น การใช้ SPSS เพื่อเปรียบเทียบความแตกต่างข้อมูลสองชุด ข้อมูลทั้งสองชุดต้องเก็บในคอลัมน์เดียวกัน

Descriptive Statistics

- เพื่อให้เข้าใจพื้นฐานข้อมูล
- ตัวอย่าง Means, medians, variances, standard deviation และ ranges
- การวัดความเข้าสู่ศูนย์กลาง (Measured of central tendency) : ค่าเฉลี่ย (mean/average) ค่ามัธยฐาน (Median) ฐานนิยม (mode)
- วัดการกระจาย (Measures of spread) : range, variance, sd
- รูปแบบข้อมูลการกระจาย : normal distribution
  - กรณีไม่ใช่ ND, ต้องแปลง หรือ ใช้ non-parametric

การเปรียบเทียบค่าเฉลี่ย

- Significance test
  - t test : two independent sample
  - ANOVA test มีเงื่อนมากกว่าสอง : one-way, factorial, repeated , split-plot

Experiment desing	Independent Var(IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-sample t test
	1	3 or more	one-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-sample t test
	1	3 or more	Repeated measure ANOVA
	2 or more	2 or more	Repeated measure ANOVA
Between-and-within-group	2 or more		Split-plot ANANO

T tests

- ใช้ในการทดสอบความแตกต่างของ ค่าเฉลี่ย
- ตัวอย่างข้อมูลที่นำมาทดสอบต้องเป็น อิสระจากกัน หรือไม่มีความสัมพันธ์กัน
- ตัวอย่าง : ไม่มีความแตกต่างกันระหว่างเวลาแล้วเสร็จระหว่างผู้ใช้งาน SW และ ผู้ไม่ใช้งาน
- Independent-samples t test
  - การทดสอบความแตกต่างระหว่างกลุ่มสองกลุ่ม เช่น กลุ่มหนึ่งมีการใช้เครื่องมือ อีกกลุ่มไม่มีการใช้เครื่องมือ
- Paired-sample t test
  - กลุ่มทดสอบกลุ่มเดียวกัน แต่วิธีการต่างกัน

T tests (ต่อ)

- การตีความผลการคำนวณ t test
  - การคำนวณ t test ได้คือคืนมาที่เรียกว่า t value ถ้ายิ่งมีค่ามาก null hypothesis มีความน่าจะเป็นสูงว่าเป็นเท็จ (ไม่สำเร็จ หรือไม่สามารถ claim) หรือ ถ้าเป็นการทดสอบค่า mean จะมีความแตกต่างกัน
  - ต้องกำหนดช่วงความเชื่อมั่น (95%) ดังนั้นค่า t value ที่ได้ถ้าสูงว่า t ที่เปิดจากตาราง แสดงว่าการทดสอบมีความแตกต่าง
  - Two-tailed and one-tailed t test ขึ้นอยู่ประเด็นการทดสอบ
    - มีความแตกต่างระหว่ากลุ่มผู้ใช้ sw กับไม่ใช้
    - กลุ่มผู้ใช้ sw มีความเร็วในการพิมพ์เร็วกว่าผู้ไม่ใช้

การวิเคราะห์ความแปรปรวน (Analysis of variance)

- ANOVA ใช้ในการเปรียบเทียบ mean มากกว่าสองกลุ่ม
- หรือบางทีเรียกว่าเป็น F test
- One-way Anova
  - ทดสอบความแตกต่างระหว่างกลุ่ม
  - ตรวจสอบเฉพาะตัวแปรอิสระเพียงหนึ่งตัวแปร แต่มีหลายๆ เงื่อนไง
  - ตัวอย่างเช่น ต้องการทดสอบประสิทธิผลของการบันทึกข้อมูลด้วยวิธีการปกติ (Standard) แบบมีตัวช่วย (word-prediction) และ แบบ speech-based (บันทึกด้วยเสียง)

ตัวอย่างข้อมูล

Group	Participants	task completion time
standard	Participant 1	245
standard	Participant 2	236
standard	Participant 3	321
standard	Participant 4	212
standard	Participant 5	267
standard	Participant 6	334
standard	Participant 7	287
standard	Participant 8	259
prediction	Participant 1	246
prediction	Participant 2	213
prediction	Participant 3	265
prediction	Participant 4	189
prediction	Participant 5	201
prediction	Participant 6	197
prediction	Participant 7	289
prediction	Participant 8	224
speech-based dictation	Participant 1	178
speech-based dictation	Participant 2	289
speech-based dictation	Participant 3	222
speech-based dictation	Participant 4	189
speech-based dictation	Participant 5	245
speech-based dictation	Participant 6	311
speech-based dictation	Participant 7	267
speech-based dictation	Participant 8	197

ลองหา ANOVA ด้วย Excel

- ดาวน์โหลดข้อมูลจากเว็บรายวิชา
- ทดลองหา ความแปรปรวนระหว่างกลุ่ม และ ภายในกลุ่มโดยใช้ Excel (ให้ลองศึกษาจากคู่มือ) ให้เวลาประมาณ 30 นาที นศ. ที่ทราบสามารถให้ความรู้เพื่อน ๆ ได้
- เปิดตาราง F (2, 21) = ???

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7842.25	2	3921.125	2.173781	0.138667	3.4668
Within Groups	37880.375	21	1803.827			

Factorial ANOVA

- ศึกษาความแตกต่างระหว่างกลุ่ม
- มีตัวแปรอิสระมากกว่าหนึ่งตัวแปร
- ตัวอย่าง
  - จากตัวอย่างเดิมที่ต้องทดสอบประสิทธิภาพการทำงานด้วยเครื่องมือแต่ละประเภท เราต้องการศึกษาเพิ่มเติมว่างานประเภทใดที่เหมาะสมกับการใช้เครื่องมือ
  - สมมุติเพิ่มงานอีก 2 ประเภท เงื่อนไขการทดสอบจะมีทั้งสิ้น 6 แบบ (2*3)
  - (หาข้อมูลไม่ได้)

Repeated measures ANOVA

- กรณีต้องการประเมินความแตกต่างภายในกลุ่ม และมีหลายเงื่อนไข ทำให้ต้องจัดเตรียมข้อมูลเพิ่ม
- ตัวอย่างเช่น ประเภทการบันทึกข้อมูล (Data entry type) - 3 และตามประเภทของงาน (task type) – 2 ต้องจัดเตรียมข้อมูล 6 ชุด ถ้าแต่ละเงื่อนไขของการทดสอบต้องการข้อมูล 12 รายการ ทำให้ต้องจัดเตรียมข้อมูลทั้งสิ้น 6*12 = 72 ในงานบางงานไม่สามารถจัดเก็บข้อมูลได้โดยเฉพาะที่เกี่ยวข้องกับมนุษย์
- ในการออกแบบให้แต่ละกลุ่ม ทำการทดลองให้ครบทุกประเภท

ตัวอย่าง

- ต้องการทราบความแตกต่างของเครื่องมือ แต่ให้ทุกคนทำการทดสอบ

	Standard	Prediction	Speech
Paticipant 1	245	246	178
Paticipant 2	236	213	289
Paticipant 3	321	265	222
Paticipant 4	212	189	189
Paticipant 5	267	201	245
Paticipant 6	334	197	311
Paticipant 7	287	289	267
Paticipant 8	259	224	197

ต้องใช้ two-way ANOVA without rep.

มีตัวแปรมากว่าสองตัว และ หลายระดับ

- Within-group

	Standard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composiotn	Group 1	Group 1	Group 1

ANOVA for split-plot design

- ต้องการศึกษาทั้ง between-group และ within-group
- ตัวอย่าง กลุ่มตัวอย่าง 2 กลุ่ม แยกการทดลองแต่ละงาน แต่ทุกเครื่องมือ

	Keyboard	prediction	Speech
Transciption	Group 1	Group 1	Group 1
Composition	Group 2	Group 2	Group 2

ข้อกำหนดของการใช้ t test และ F test

- มีข้อกำหนดอยู่สามอย่างคือ
  - ข้อผิดพลาดของข้อมูลทุกจุดต้องเป็นอิสระจากกัน
    - ตัวอย่าง ต้องการใช้เครื่องมือ โดยมีข้อกำหนดในการอบรมผู้เข้าอบเพียงให้ความรู้พื้นฐาน ถ้าเกิดมีกลุ่มใดกลุ่มหนึ่งให้การอบรมเกินกว่าข้อกำหนด ถือว่าเป็นเป็นอิสระจากกัน
  - รูปแบบการกระจายข้อผิดพลาด ต้องมีค่าความแปรปรวนไม่แตกต่างกัน ถ้ามีความแตกต่างกัน ต้องมีการแปลงรูปข้อมูล เช่น ถอดราก, log
  - ข้อผิดพลาดต้องมีการกระจายแบบปกติ

Correlation

- เป็นหนึ่งในรูปแบบของการหาความสัมพันธ์ของตัวแปร

	C Exp.	Standard	Prediction
Participant 1	12	245	246
Participant 2	6	236	213
Participant 3	3	321	265
Participant 4	19	212	189
Participant 5	16	267	201
Participant 6	5	334	197
Participant 7	8	287	289
Participant 8	11	259	224

	C Exp.	*Standard*	*Prediction*
C Exp.	1		
Standard	-0.72264	1	
Prediction	-0.46753	0.324856	1

Regression

- ความสัมพันธ์ระหว่างตัวแปรไม่อิสระ หนึ่งตัวแปร กับตัวแปรอิสระมากกว่าหนึ่งตัวแปร
- นำมาใช้
  - Model construction
  - Model prediction

Nonparametric statistical tests

- Chi-square test
  - ข้อมูลในตารางสรุป (Contingency table) ต้องเป็นอิสระจากกัน
  - ใช้ไม่ได้ดี กรณีข้อมูลขนาดเล็ก (ควรมากกว่า 20)

การประเมินประสิทธิผลของโมเดล

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation…

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 80%

Cost = 3910

Accuracy = 90%

Cost = 4255

Cost vs Accuracy

Count	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

Accuracy is proportional to cost if

1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

$N = a + b + c + d$

$\text{Accuracy} = (a + d)/N$

Cost	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	p	q
	Class=No	q	p

$$\begin{aligned}\text{Cost} &= p\,(a + d) + q\,(b + c)\\ &= p\,(a + d) + q\,(N - a - d)\\ &= q\,N - (q - p)(a + d)\\ &= N\,[q - (q-p) \times \text{Accuracy}]\end{aligned}$$

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

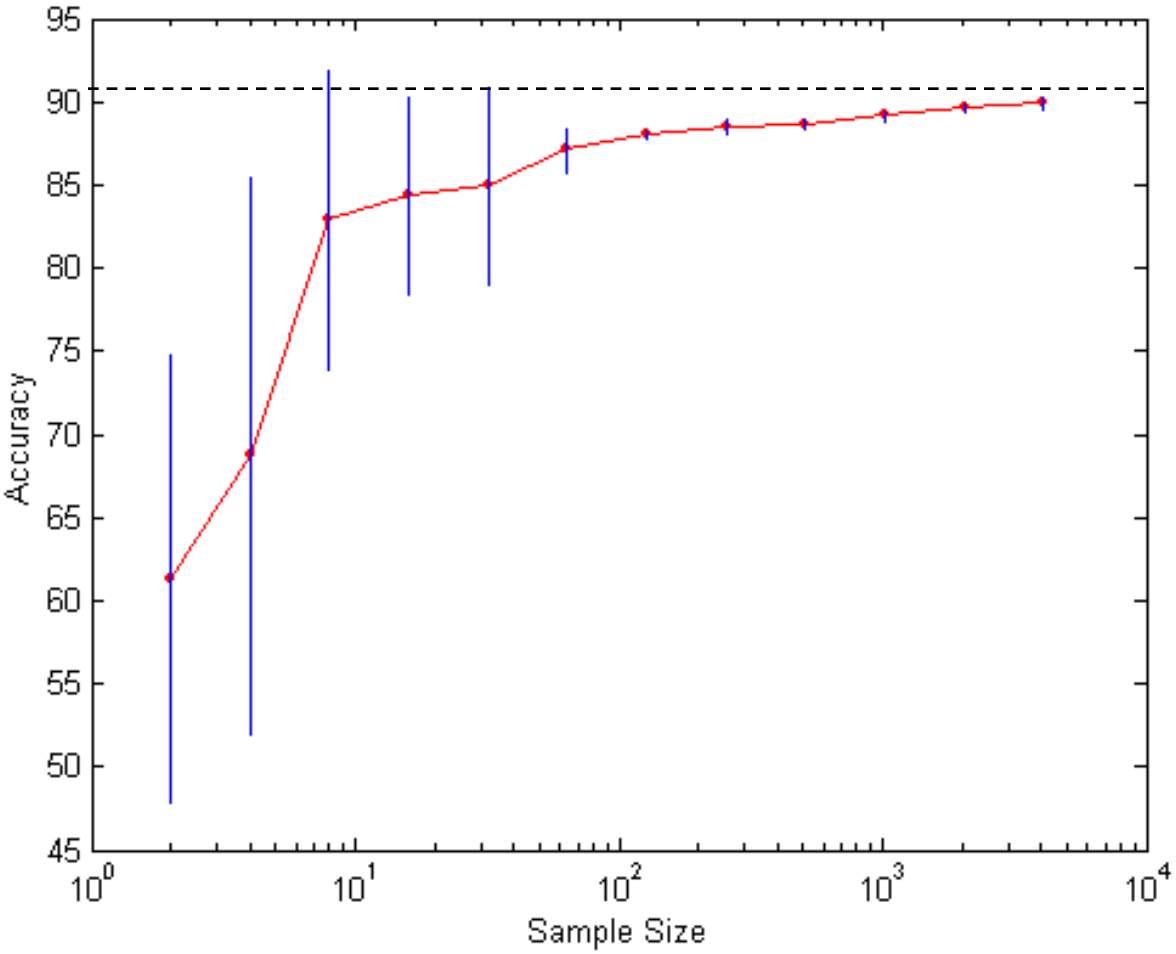
$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Methods for Performance Evaluation

- How to obtain a reliable estimate of performance?

- Performance of a model may depend on other factors besides the learning algorithm:
  - Class distribution
  - Cost of misclassification
  - Size of training and test sets

Learning Curve



- Learning curve shows how accuracy changes with varying sample size
- Requires a sampling schedule for creating learning curve:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)

Effect of small sample size:

- Bias in the estimate
- Variance of estimate

Methods of Estimation

Holdout

- Reserve 2/3 for training and 1/3 for testing

- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n
- Stratified sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling with replacement

Comparing Performance of 2 Models

- Given two models, say M1 and M2, which is better?
  - M1 is tested on D1 (size=n1), found error rate = e_1
  - M2 is tested on D2 (size=n2), found error rate = e_2
  - Assume D1 and D2 are independent
  - If n1 and n2 are sufficiently large, then

$$e_1 \sim N(\mu_1, \sigma_1)$$

$$e_2 \sim N(\mu_2, \sigma_2)$$

  - Approximate: $$\hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

Comparing Performance of 2 Models

- To test if performance difference is statistically significant: d = e1 – e2
  - d ~ $N(d_t, \sigma_t)$ where d_t is the true difference
  - Since D1 and D2 are independent, their variance adds up:

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$

$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

  - At (1-α) confidence level, $d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$

An Illustrative Example

- Given: M1: n1 = 30, e1 = 0.15
  M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1 (2-sided test)

$$\hat{\sigma}_d = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- At 95% confidence level, $Z_{\alpha/2}$=1.96

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> Interval contains 0 => difference may not be statistically significant

Comparing Performance of 2 Algorithms

Each learning algorithm may produce k models:

- L1 may produce M11 , M12, …, M1k
- L2 may produce M21 , M22, …, M2k

If models are generated on the same test sets D1,D2, …, Dk (e.g., via cross-validation)

- For each set: compute $d_j = e_{1j} - e_{2j}$
- d_j has mean d_t and variance σ_t
- Estimate:

$$\hat{\sigma}_t^2 = \frac{\sum_{j=1}^{k}(d_j - \overline{d})^2}{k(k-1)}$$

$$d_t = d \pm t_{1-\alpha,k-1}\hat{\sigma}_t$$

ตารางสถิติที่จำเป็น

- http://www.math.unb.ca/~knight/utility/t-table.htm
- http://www.watpon.com/table/