巨量資料與統計分析

政治大學統計系余清祥

2024年12月17日

第十四週：報告主題與格式

http://csyue.nccu.edu.tw

英文課的啟發

- 思維辯證的技巧在於大數據分析的過程中一再重複需要用到，包括定義問題、資料蒐集、EDA、分析推論等階段。
- 英文課的啟發

平權運動

- 美國平權運動在1980年代炒作得相當熱烈，公家機關與學校等機構保障聘任一定比例的女性、黑人。

→一對白人兄弟到公家機關求職，履歷表中填寫少數族裔，但人事單位質疑血統，這對兄弟拿出家譜證明外婆是百分之百的黑人。

→1/4的黑人血統算是黑人嗎？

- 可利用離散（是或否「0或1」）、連續資料（基因比例0%~100%）來定義族裔。

黑人認定問題

- 從美國平權運動的議題可知，血統認定本身存有不少爭議。

→台灣也有身障、原住民的保障名額。

- 分析大數據也是如此，即使給定相同題目（出發點）及素材（資料庫），因為**推理方向**和**思考角度**的差異，最後的分析結果大異其趣。（註：「戲法人人會變，各有巧妙不同」！）

歸納法、演繹法

歸納法和演繹法的比較

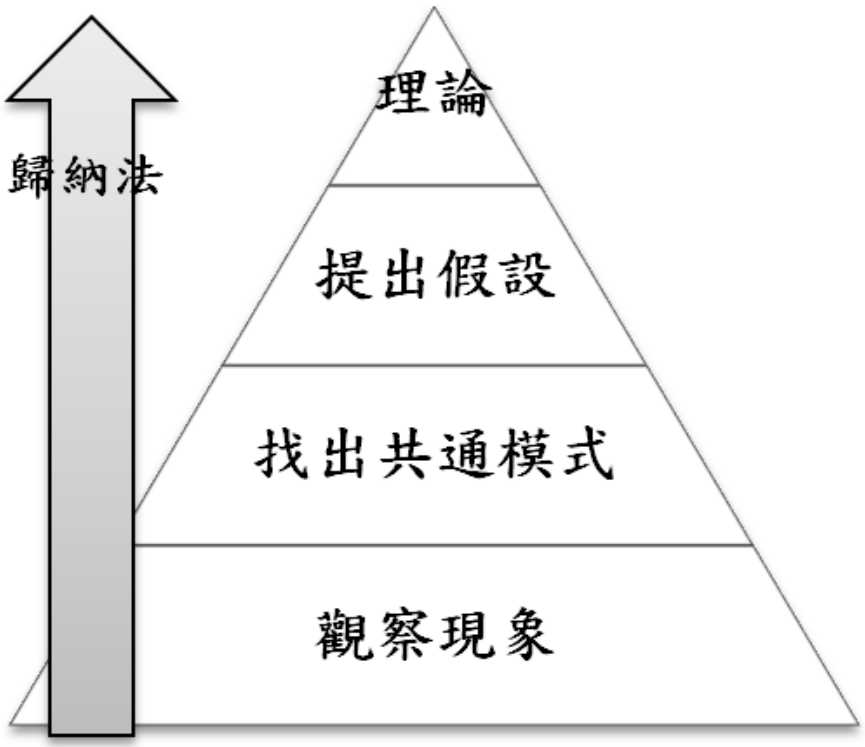
- **歸納法**：透過對個體進行一系列的觀察後，發展出可達一般性的論述或模式，足以代表既定事件的秩序。

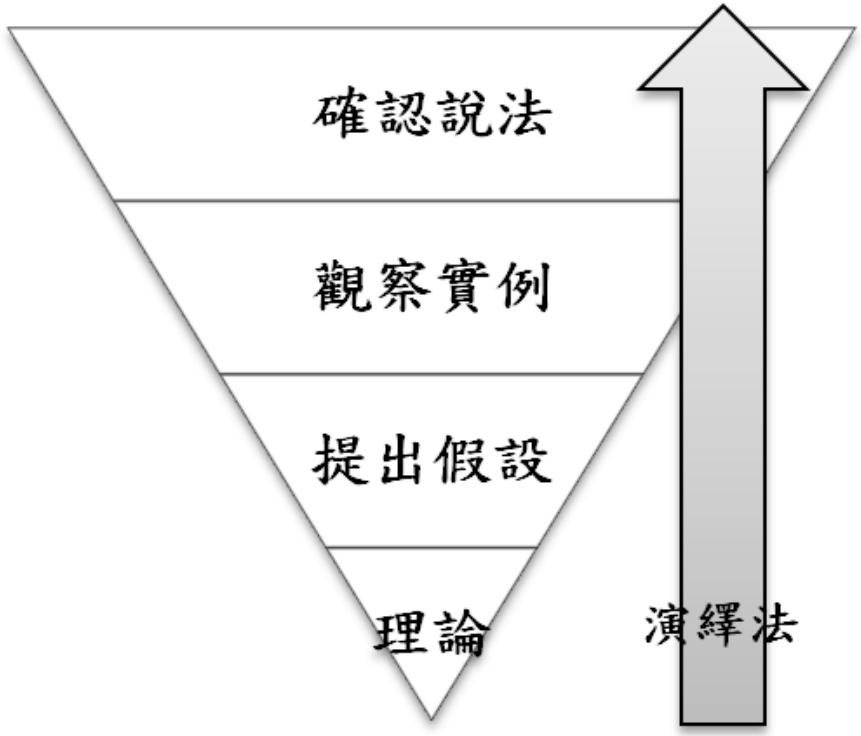
→如：孔子會死、國父會死、林肯總統會死➔人都會死！

- **演繹法**：從整體概念（假設、前提）出發建立獨特性，再把普遍的法則運用到特定個體，亦即舉一反三。

→如：人都會死➔孔子是人、國父是人，林肯總統是人，所以他們都會死！

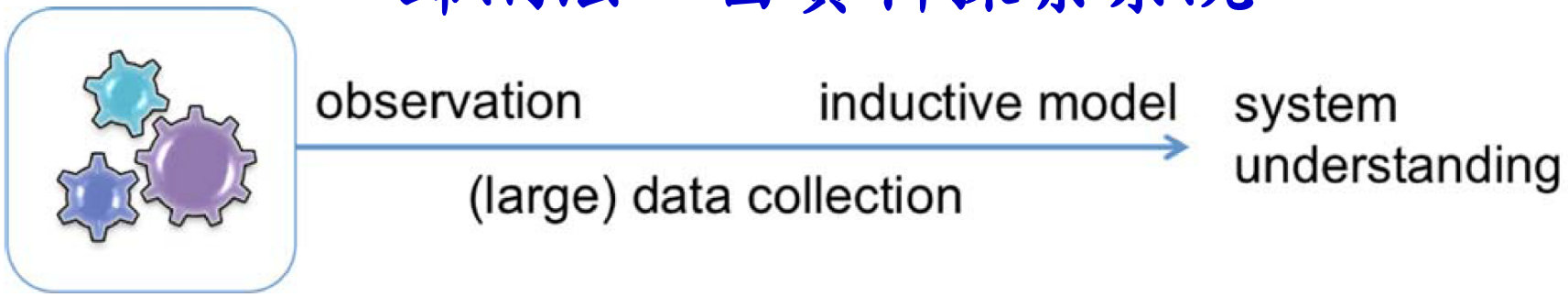
歸納法和演繹法的比較





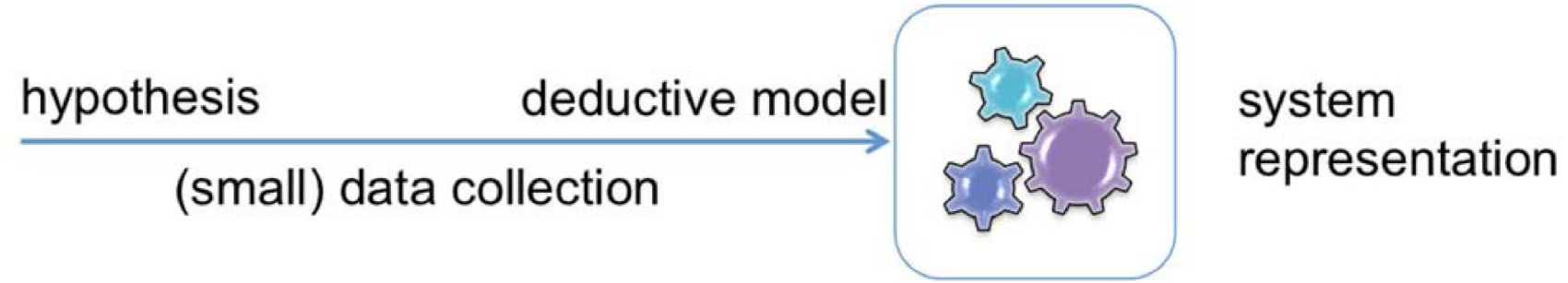
歸納法和演繹法的比較

歸納法：由資料探索系統



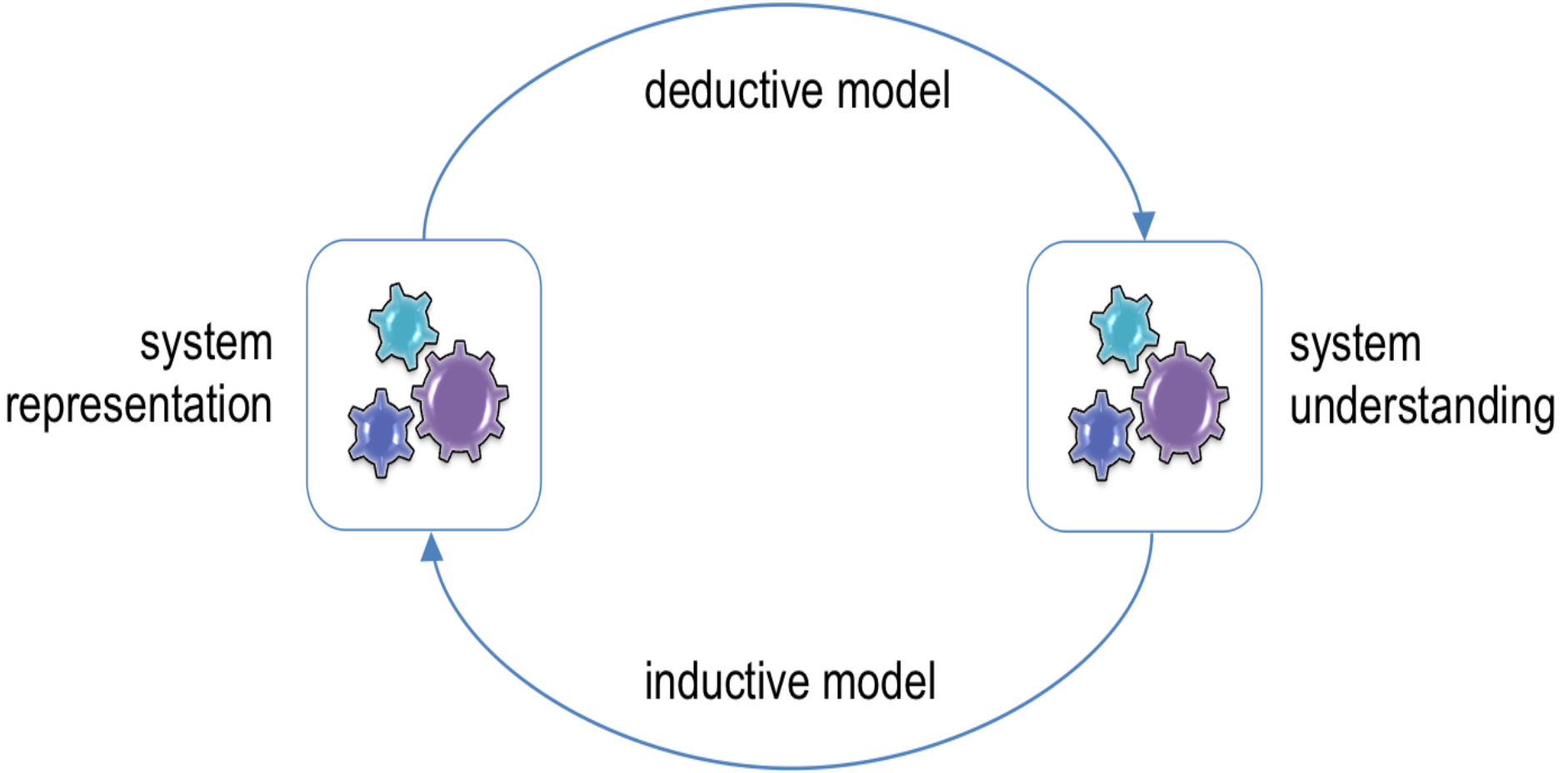
「資料」、「觀察」➔「推論」、「假設」➔➔「認知」

演繹法：由假設描述現象



由少量資料推論可能假設，再以大量資料驗證理論！

研究推論：歸納與演繹



參考文獻：Two Virtuous Models - The Power of Deduction and Induction (Aker, 2014)

歸納法的反例

- **歸納法**：資料不足、樣本代表性有疑慮等，都有可能得出似是而非的結果推論。（**論述未必是事實！**）

→如：裙擺指數、520行情、狐假虎威、吸毒指數、房地產只漲不跌等。

→某一年冠軍已連續六個聖筊勝出，但連續擲六次都可得到聖筊者，可以說是擲筊高手嗎？

→超級約分的謬誤：

$$\frac{16}{64}=\frac{16}{64}=\frac{1}{4}$$

$$\frac{26}{65}=\frac{26}{65}=\frac{2}{5}$$

驚！台灣8例「50多歲的怎麼了」

■網友在網路論壇 PTT Gossiping 板表示，「外出請戴口罩，保護自己保護別人，勤洗手做好自主健康管理，不必過度恐慌，N95 留給醫護人員」。

→網友們紛紛一面倒留言表示**「50多歲的這年代人是有什麼問題」！**

https://s.yimg.com/ny/api/res/1.2/XaMVHNjk0uHNAHGp0X00Q--~A/YXBwaWQ9aGlnaGxhbmRlcjtzbT0xO3c9MTI4MDtoPTk2MA--/https://media.zenfs.com/zh-tw/nownews.com/1231bde7563bd797886319b7e8035785

https://zh.wikipedia.org/wiki/%E5%A4%9A%E5%88%A9

演繹法的反例

https://i2.kknews.cc/SIG=17c0unf/5092000498q887q141p9.jpg

- **演繹法**：前提是否成立、數個前提是否抵觸或衝突等，會得出矛盾的結論。（邏輯！）（**前提未必成立！**）

→如：「自相矛盾」、所有符合「反證法」的結果。

（註：演繹法不能解決前提是否真實，前提的真實性要依靠科學方法和實證分析來檢驗。）

→複製羊桃莉可能推翻人都會死的理論！

→俗語說：「天下烏鴉一般黑」，烏鴉都是黑色嗎？同理，天鵝都是白色嗎？

http://photozou.jp/photo/photo_only/2870200/160820317?size=1024#content

歸納法與演繹法的磨合

- **歸納法**為「觀察值獲得可能論述」、**演繹法**為「將假設套入資料分析取得驗證」兩者出發點不同，但**兩者可以相輔相成、互補不足**，激盪出更多元的可能。
- 資料分析在各領域研究都是必要步驟，但進行方式、研究目標、使用名詞等不盡相同。

→如：質性研究的個案分析(Case Study)目的在於描述真實狀況，以文字方式書寫刺激讀者思考，探究問題原因、可行解決方法、提供預防措施等目標。

註：個案分析也可分為兩個類型：量化研究、深度詮釋（注重質性分析）。

尋找大數據資料的研究主題

分析大數據的可行步驟之一

- 如同數量化非結構資料，即使資料已經格式化，但整體分析方向仍可能毫無頭緒。

→例如：健保資料庫中變數有一定欄位，但研究目標及相關變數可能毫無頭緒。

- 敘述性統計、關連性之類的分析結果，多半只呈現某些變數的特性及關係，可能會有見樹不見林的疑慮。

→如何從分析結果、或是其他管道獲取資訊，整理出有系統的研究目標？

唐詩的代表色

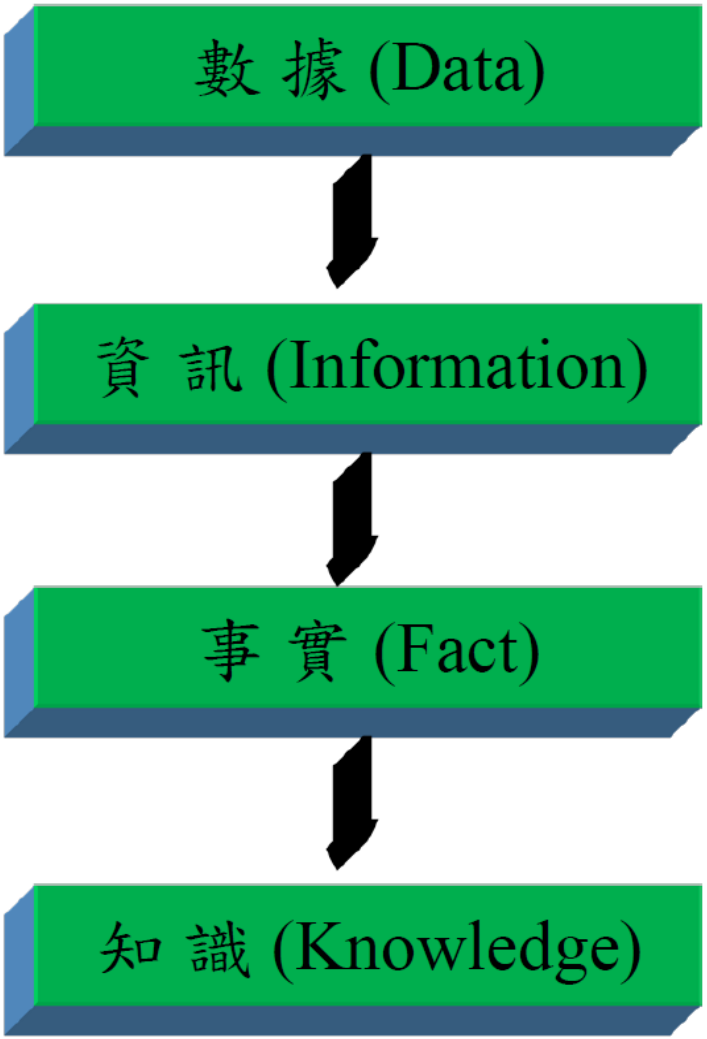
- 政治大學老師最近想以科學方法找出唐詩代表顏色
- **歸納法**：從幾首詩中可看出白色可以是形容詞、名詞，或以雪的意象(白)出現。

→李白的《秋浦歌》提到「白髮三千丈、緣愁似個長」、柳宗元《江雪》的「孤舟蓑笠翁，獨釣寒江雪」、王之渙的《登鸛鵲樓》的「白日依山盡，黃河入海流」

- **演繹法**：若唐詩為白色，請問白色的定義為何？

→白色的同義字很多，包括包括「**雪、皚、皓、皙、霜、素**」；參考實驗設計概念實驗組（Treatment）、對照組（Placebo）的作法，應把其他常見顏色利入考量（如紅色、黑色），以確認唐詩中出現白色的比例明顯較高。

知識累積的過程



他山之石？蒐集文獻與研究主題

- 演繹法的出發點為理論假設，和強調證據的歸納法頗多衝突，但因理論提供有系統的分析方向，或許可蒐集相關文獻，由既有結果中衍生出新的可能。

→例如：參考衛福部出版刊物、學術研究，獲取健保資料庫的可能分析方向。

- 關鍵：如何在大方向（理論）、可執行的子題（現象）兩者間取得平衡？

由相關文獻尋找研究方向

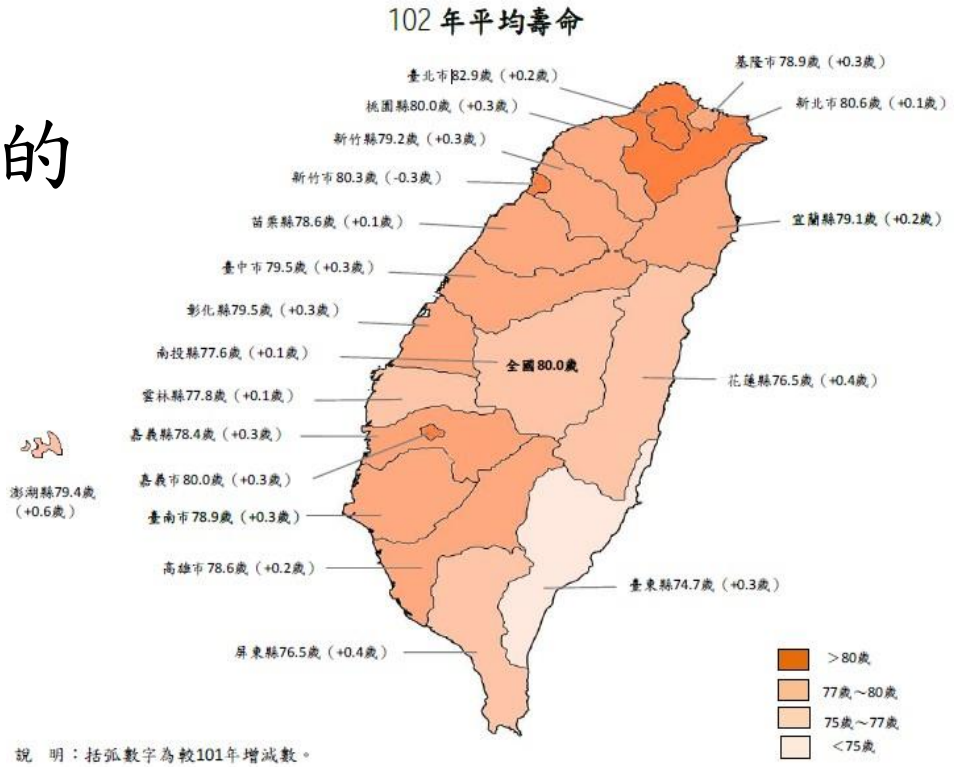
- 範例一、壽命的城鄉差距

→根據內政部統計，2013年台灣居民零歲平均餘命，各縣市間的差異非常大（超過八歲）。

- 由壽命的城鄉差距，可以推論出幾個可能的探索新方向。

→年齡別死亡率；

→各縣市居民的健康。



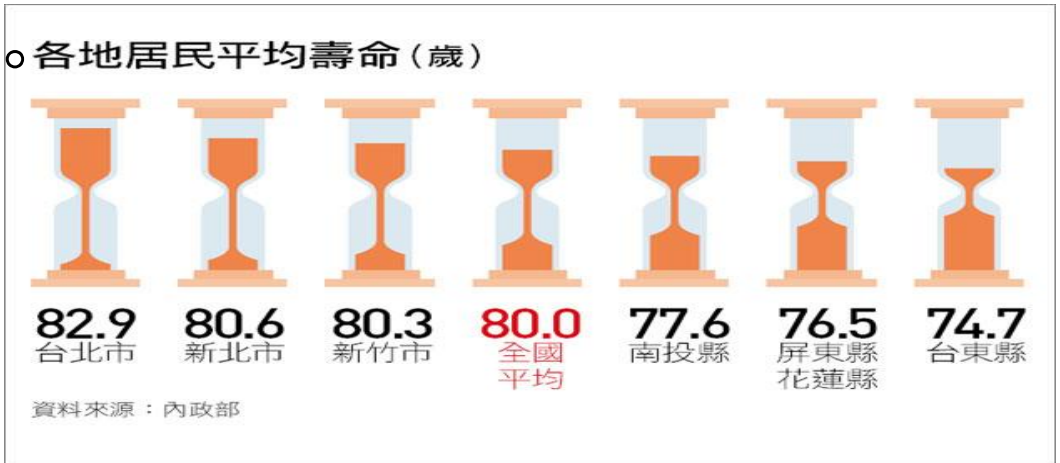
城鄉差距與健康、壽命

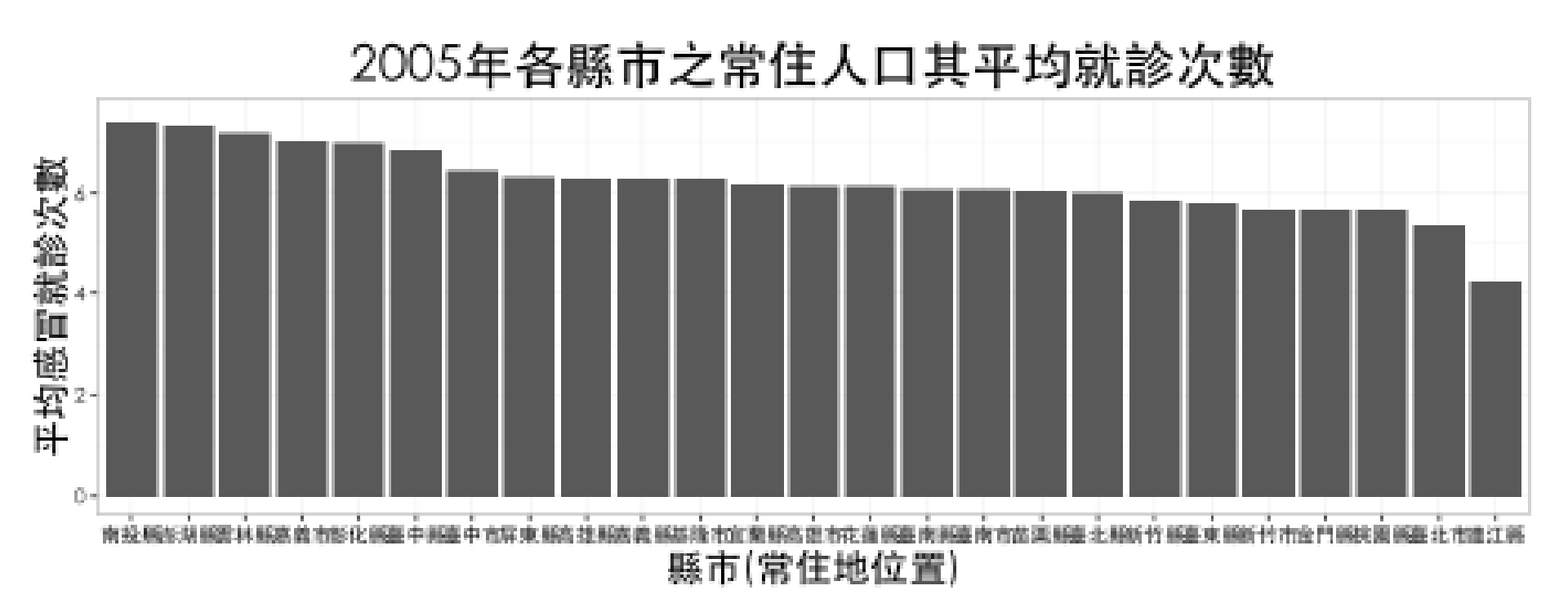
- 與區域、壽命等相關的探討議題：

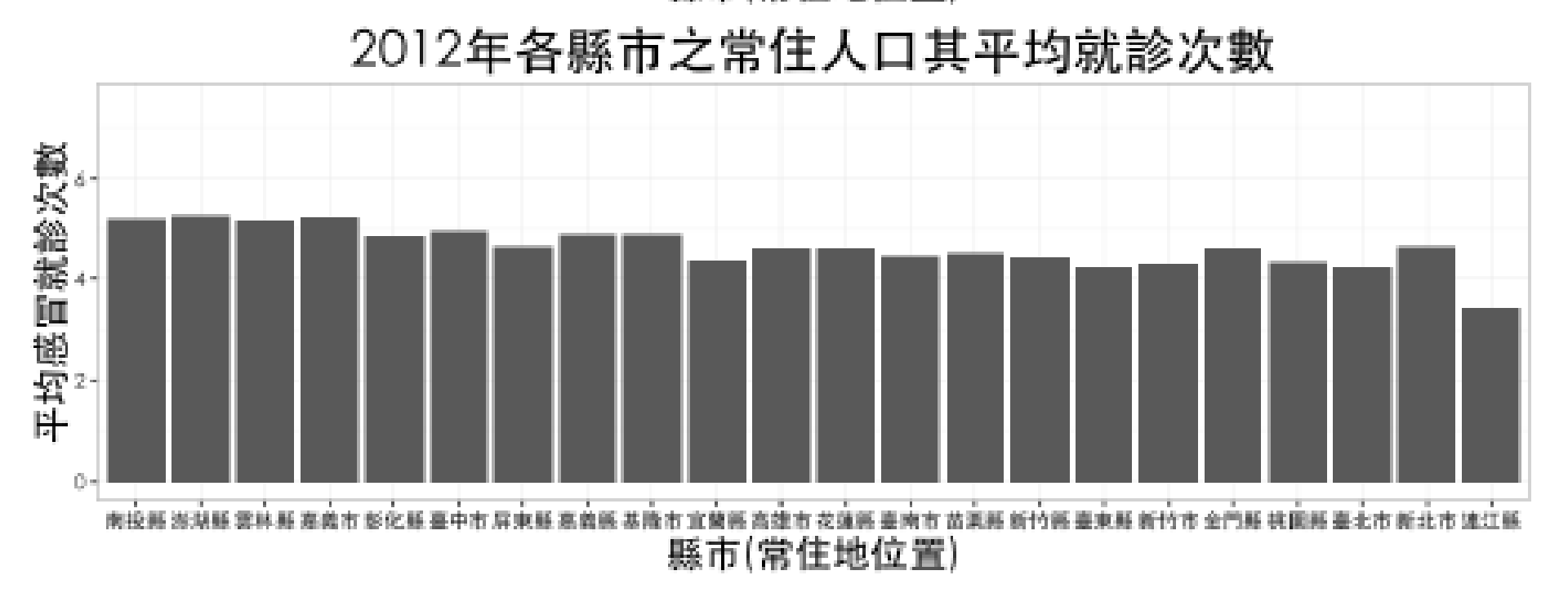
→台灣幅原不大，但壽命差異如此巨大，是否與城鄉（資源）差距有關？例如：比較縣市層級的醫療供給（醫療可近性）。

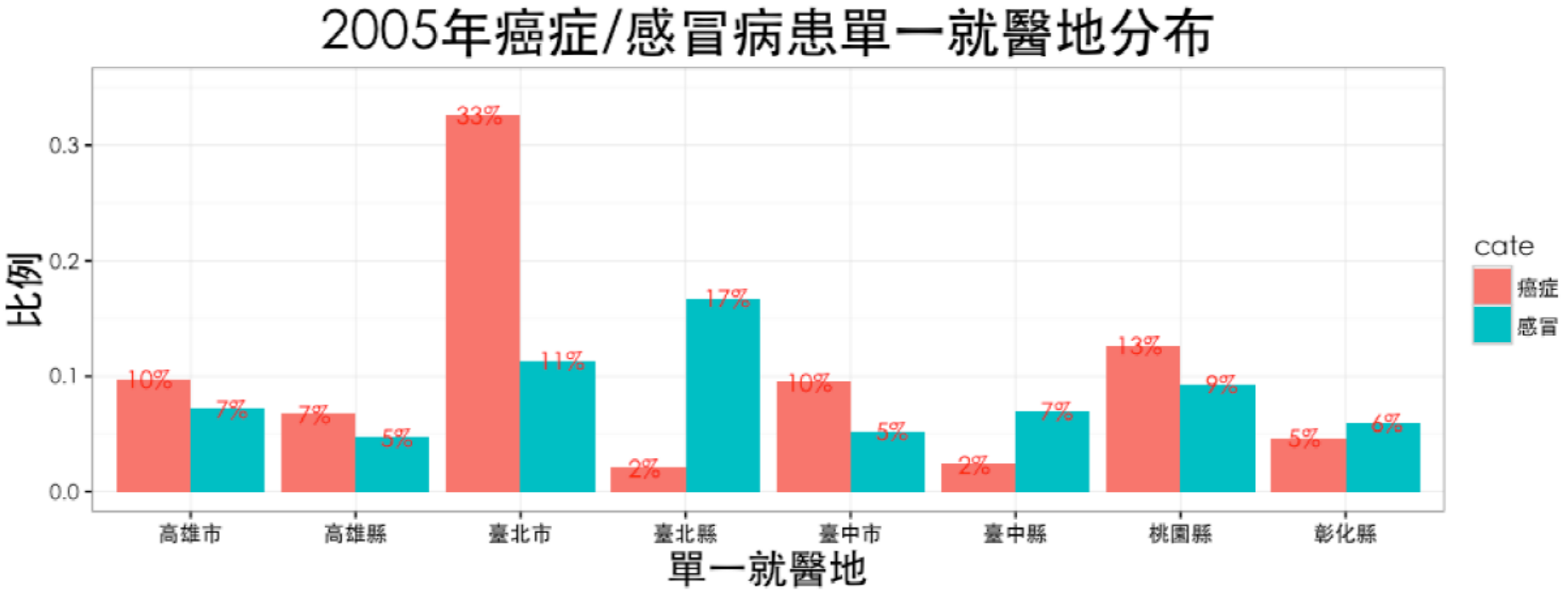
→壽命和健康之間的連結？例如：各縣市居民的醫療利用（次數、金額）、疾病類型（發生率、死亡率）等特性

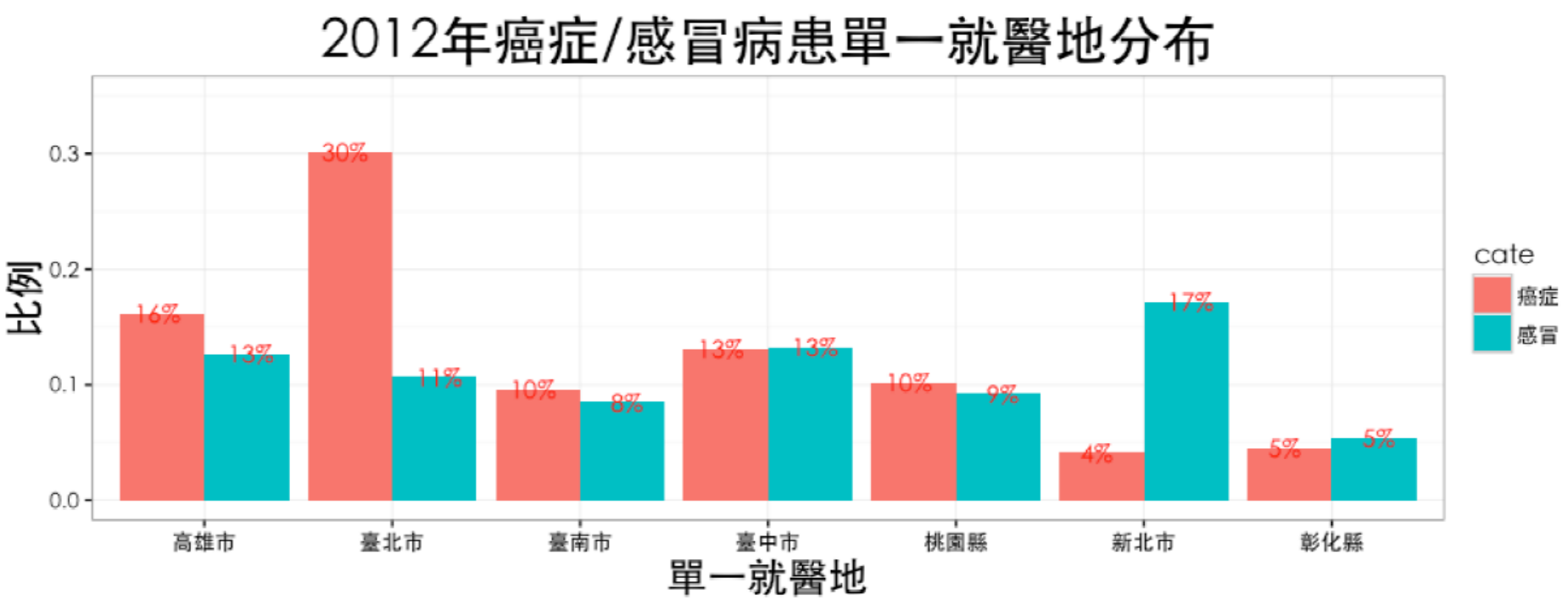
→各縣市重大傷病之相關研究。




2005年各縣市之常住人口其平均就診次數
平均感冒就診次數
0
2
4
6
縣市(常住地位置)


2012年各縣市之常住人口其平均就診次數
平均感冒就診次數
0
2
4
6
縣市(常住地位置)


2005年癌症/感冒病患單一就醫地分布
比例
0.0
0.1
0.2
0.3
10%
7%
7%
5%
33%
11%
2%
17%
10%
5%
2%
7%
13%
9%
5%
6%
高雄市
高雄縣
臺北市
臺北縣
臺中市
臺中縣
桃園縣
彰化縣
單一就醫地
cate
癌症
感冒

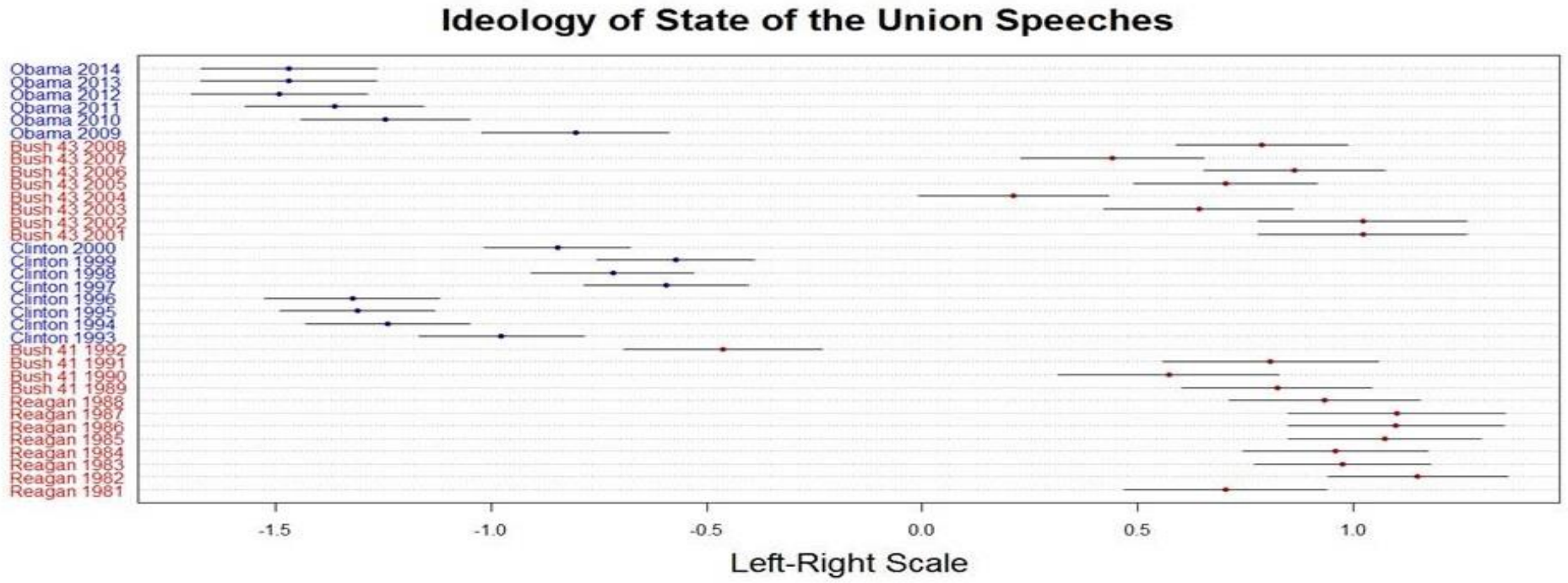

2012年癌症/感冒病患單一就醫地分布
比例
0.0
0.1
0.2
0.3
16%
13%
30%
11%
10%
8%
13%
13%
10%
9%
4%
17%
5%
5%
高雄市
臺北市
臺南市
臺中市
桃園縣
新北市
彰化縣
單一就醫地
cate
癌症
感冒

由相關研究尋找研究方向(續)

- 範例二、總統就職演講稿。

→比較最近兩任美國總統小布希、歐巴馬，兩者在國情咨文的立場。

(http://www.realclearpolitics.com/articles/2015/01/20/text_mining_the_state_of_the_union.html)



總統的就職演講稿分析

- 以類似想法套入蔡英文總統第一任就職演講稿，以下是《蘋果日報》的量化報導：

→*33分鐘就職演說近6千字，提及5次「中華民國」，41次「台灣」；獲得40次現場觀眾掌聲，其中掌聲還3度打斷蔡演說。總統府直播在9時宣誓典禮開始時湧入超過3萬人，10點40分表演節目進入《島嶼天光》時，線上觀看人數已超過5萬人，11時蔡英文開始演說後突破6萬人。*

註：兩岸、司法各14、13次。

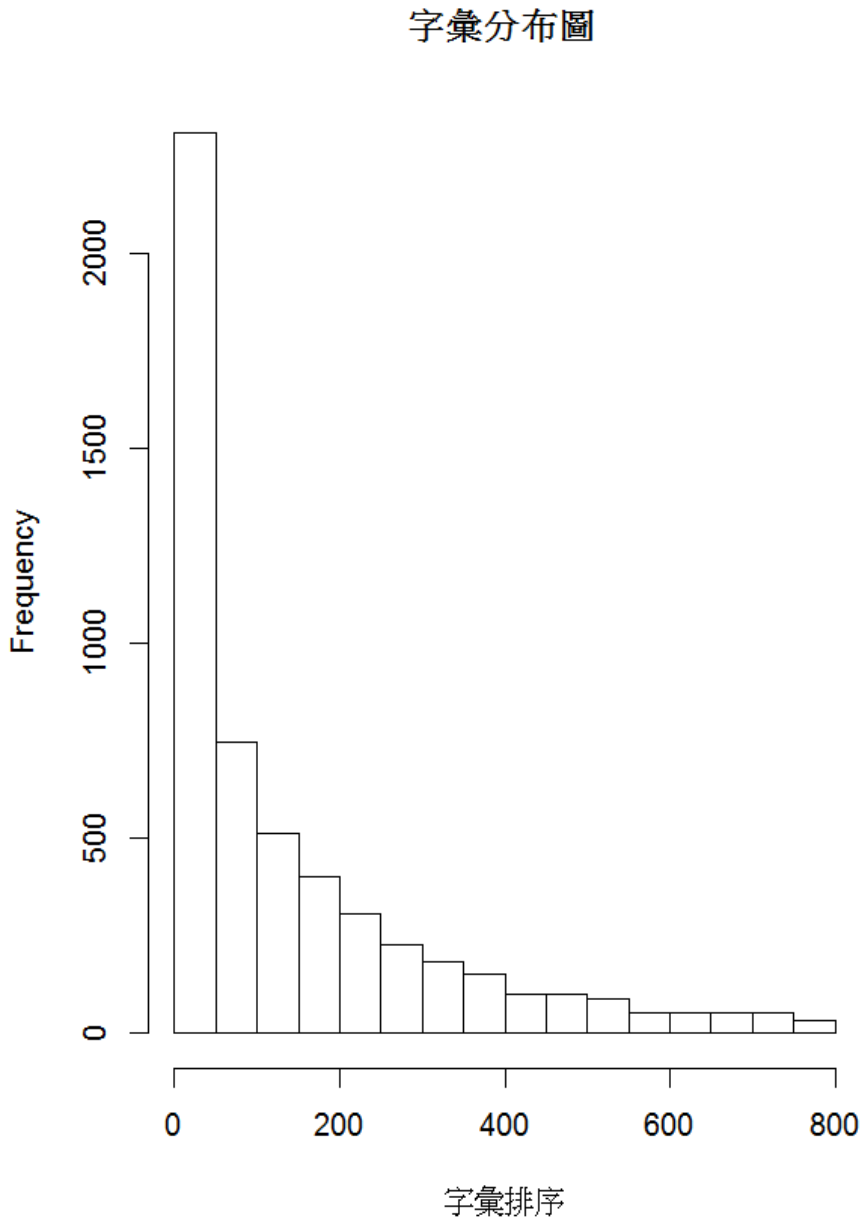
第十四屆總統就職演講稿分析(續)

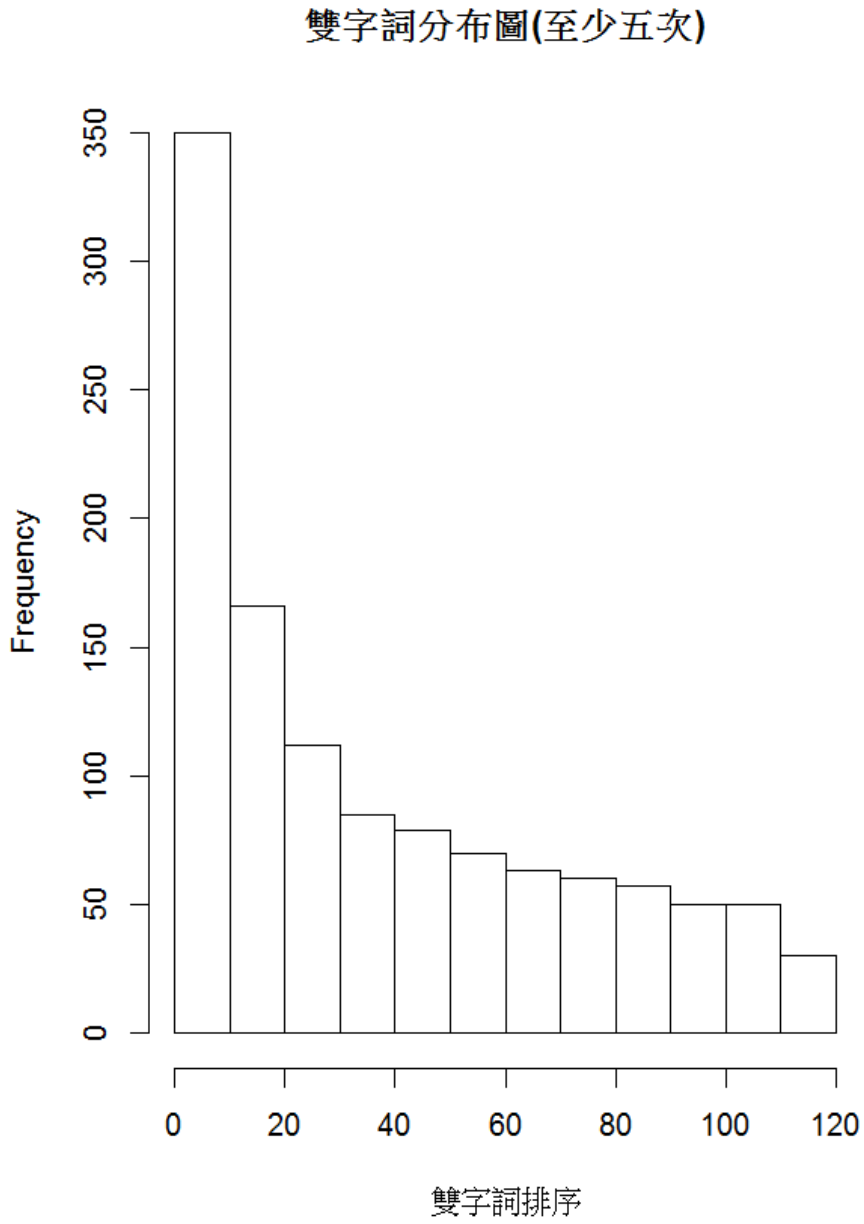
- 仿造《新青年》字彙的整理方式，整理單字、雙字詞的出現個數：

	個數	種類數	前十名比例	Simpson	Entropy
單字	5,345	780	18.02%	0.007441	5.79651
雙字詞	4,274	2,476	8.19%	0.001548	7.37551
雙字詞(至少4次)	1,432	181	24.44%	0.011767	4.89732
雙字詞(至少5次)	1,172	116	29.86%	0.016810	4.47884

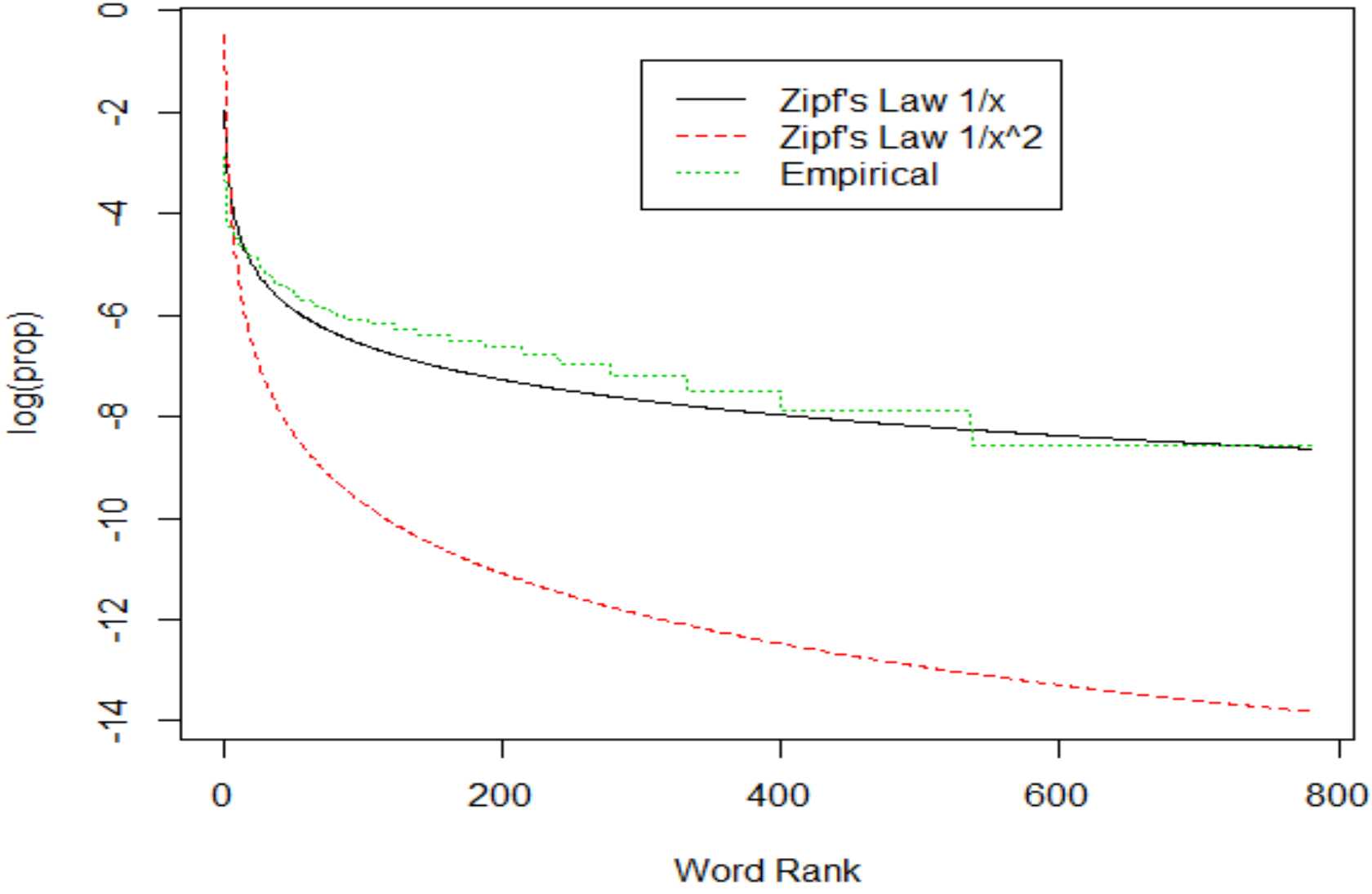
第十四屆總統演講稿前十個最常出現字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	的	293	5.48%	我們	86	2.012%
2	我	114	2.13%	台灣	41	0.959%
3	們	90	1.68%	政府	37	0.866%
4	一	75	1.40%	國家	32	0.749%
5	會	74	1.38%	一個	29	0.679%
6	是	70	1.31%	新政	27	0.632%
7	個	66	1,23%	經濟	27	0.632%
8	民	63	1.18%	這個	25	0.585%
9	人	59	1.10%	民主	24	0.562%
10	國	59	1.10%	社會	22	0.515%

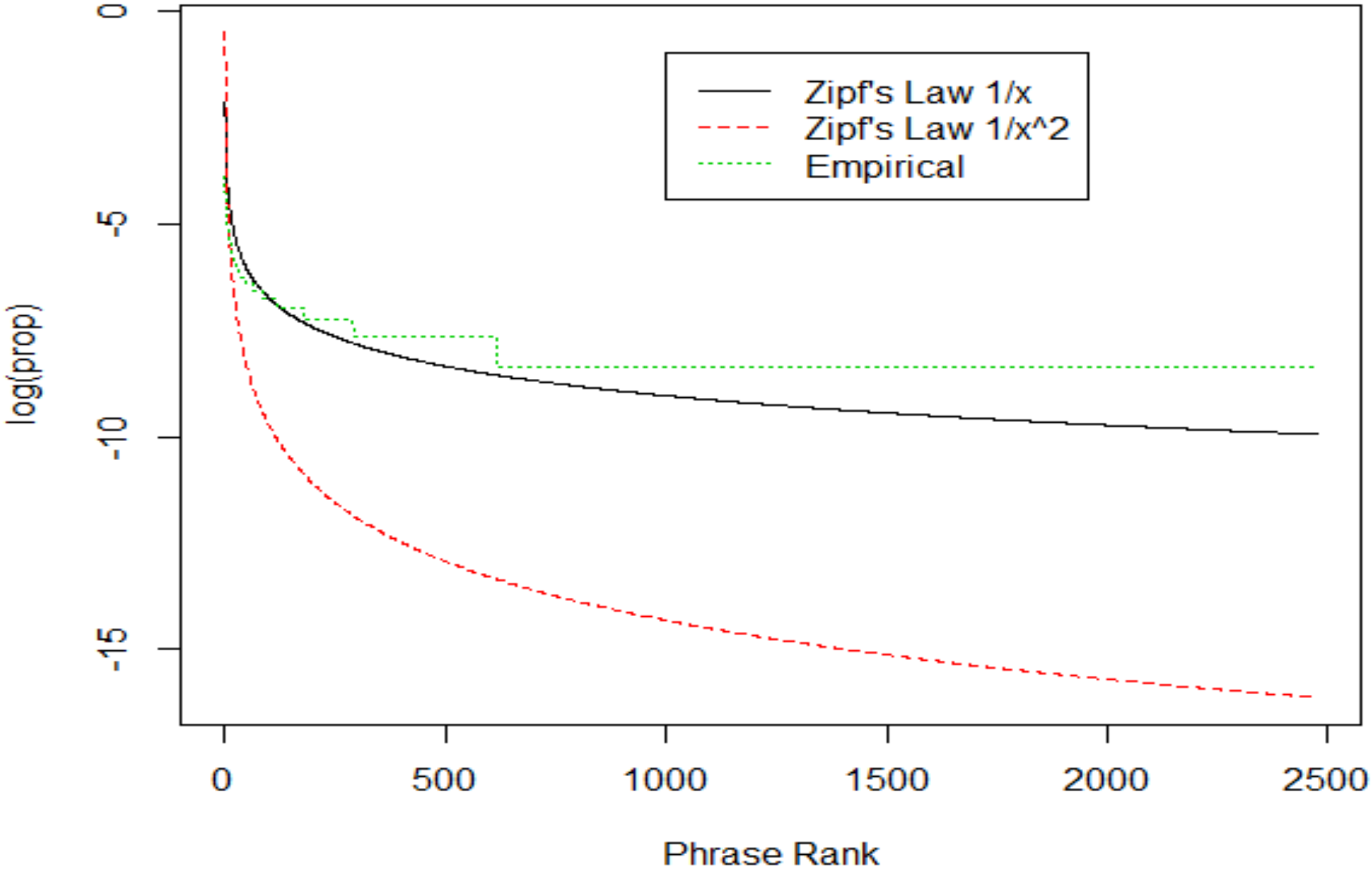




蔡英文總統第一任演講稿最常出現字詞的頻率



蔡英文第一任總統演講稿最常出現字彙與齊夫法則



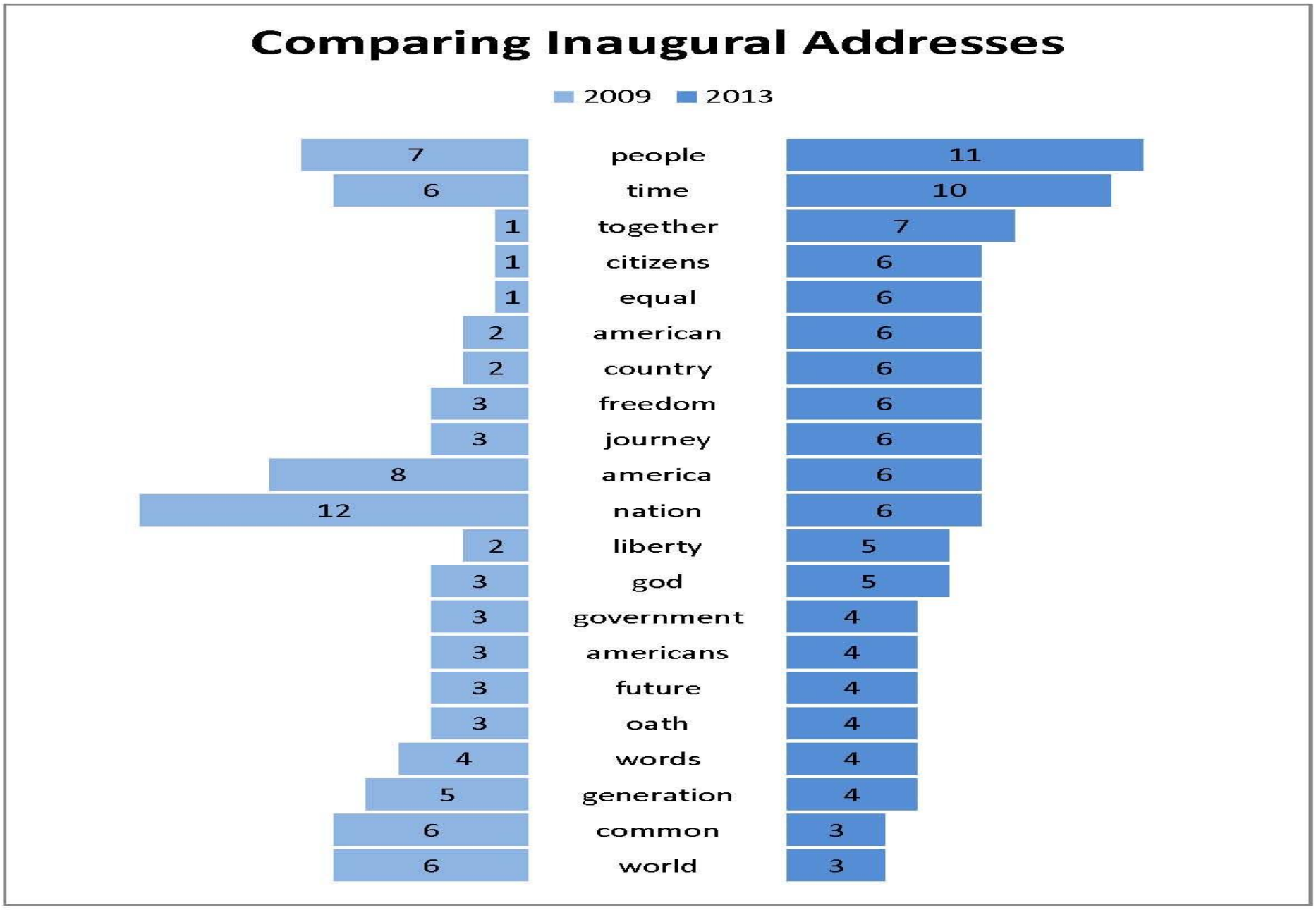
蔡英文第一任總統演講稿常見雙字詞與齊夫法則

第十四屆總統英文演講稿前十個常見字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	the	82	7.86%	of the	11	1.05%
2	of	43	4.12%	taiwanese people	8	0.76%
3	and	37	3.55%	the taiwanese	7	0.67%
4	to	35	3.36%	we will	7	0.67%
5	that	26	2.49%	in the	6	0.57%
6	will	26	2.49%	a government	5	0.48%
7	in	19	1.82%	government that	5	0.48%
8	a	18	1.73%	of china	5	0.48%
9	I	16	1.53%	republic of	5	0.48%
10	we	16	1.53%	the republic	5	0.48%

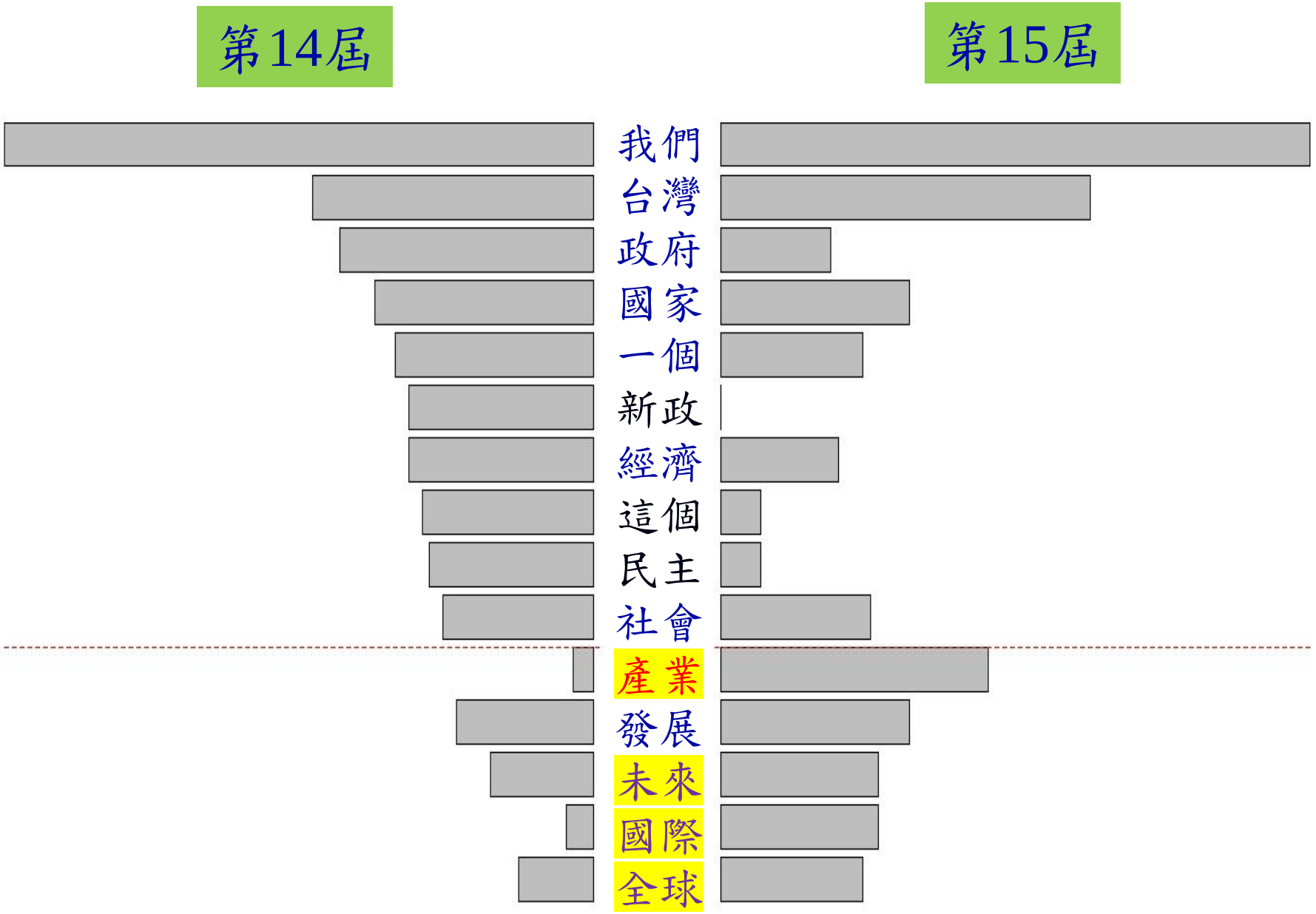
第十四屆總統英文演講稿前十常見現字詞(續)

排序	單字		雙字詞	
	類別	次數	類別	次數
1	people	15	taiwanese people	8
2	election	10	we will	7
3	taiwanese	10	a government	5
4	government	9	the republic of china	5
5	country	8	this election	5
6	democratic	8	we have	5
7	taiwan	7	the people	4
8	political	6	this country	4
9	thank	6	23 million people	3
10	china	5	our democratic	3



Analyzing the speech of President Obama (Textmining)

蔡英文總統就職演講稿常見雙字詞



國外通訊社對總統演講稿的反應

- 法新社：以和解口吻呼籲和中國大陸良性對話。
- 美聯社：未提及一個中國可能觸怒北京當局。
- 路透社：台灣將致力維持兩岸關係的和平穩定。
- CNN：儘管蔡總統獲得強力授權（行政與立法），但若與北京關係惡化可能削弱她內政目標的能力。
- BBC：就職演說提到台灣人有堅定信念「捍衛民主自由的生活方式」，可能會惹惱中國大陸。

註：理念解讀和「關鍵詞」有關。

理論假設 vs. 實證分析

- 理論的文字敘述與實證的分析項目，兩者間未必可直接劃上等號，經常得經過某種程度的轉換。（註：為什麼不是「接受H_0」？）

→例如：如何確定一組數字來自均勻（或常態）分配的亂數？（意即**均勻**＋**獨立性**！）

→通過不同類型的分析與檢定，也不能確認一組亂數符合某種分配。（**反證** vs. **證明**）

- 理論不等於實務，實際的觀察值會因量測等因素產生**誤差**，需加上邊際誤差的想法。

由問卷設計看大數據分析

- 理論及實務分析的關連，與問卷調查如何設計各問項問題類似，在整體大方向的導引下，將調查目標分割為幾個子題，如同拼圖，缺乏任一子題會有不完整之虞。

→需考量如何切割研究目標為幾個子題、子題之間的關連等問題。

→例如：高速公路局「102年元旦連續假期交通疏導工作」民眾滿意度調查，整體調查方向、各子題的目標為何？

來源：http://www.hpb.gov.tw/files/16-1000-928.php

一、請問您行駛高速公路的頻率如何？		
每週5次以上	95	15%
每週1至4次	159	25%
每月1至3次	281	44%
每月不及1次	97	15%
二、請問您本次旅程預計行駛距離？		
10-50公里	102	16%
51-100公里	221	35%
101-150公里	124	20%
151公里以上	185	29%
三、請問您對本次行旅時間與您估計的相差多少？		
短很多	51	8%
短一些	103	16%
差不多	351	56%
長一點	104	16%
長很多	23	4%
四、請問您對於元旦連續假期警察執勤率是否滿意？		
非常滿意	232	37%
滿意	351	56%
普通	48	8%
不滿意	1	0%
非常不滿意	0	0%
五、請問您對於元旦連續假期警察管制疏導是否滿意?		
非常滿意	223	35%
滿意	348	55%
普通	60	9%
不滿意	1	0%
非常不滿意	0	0%
六、請問您在旅程中，見到公警局員警的表現如何?		
非常滿意	270	43%
滿意	339	54%
普通	22	3%
不滿意	1	0%
非常不滿意	0	0%

蒐集問卷資料

- 仿造EDA及CDA的分類，問卷也可分成專家意見（包括焦點訪談、德菲法）、問卷調查（份數通常較多）。

→若對於問題缺乏共識或是相關文獻，可透過專家意見等方式蒐集可能的子議題。

→若對於問題已有清楚輪廓及整體方向，且問卷問項能清楚訴諸文字，則可逕行進入抽樣調查。

問卷內容與格式

- 建議在設計問卷時逐項檢查以下各點：

1. 確切的問題定義

→有哪些問項必須詢問、哪些個人問項與這些問項有關？

2. 調查母題

→母體可否與問題配合？

3. 調查方法

→考量資料品質、金錢、時間、人力等。

撰寫大數據分析報告

如何整理大數據的分析結果

- 分析大數據的策略與傳統資料分析並無不同，僅在於大量、時效（速度）上的需求及差異，特別需要資料所屬領域的知識支持，尤其是定義研究目的、量化目標變數（標的）。
- 藉由大數據探討某個議題，如同撰寫論文、報告，具備幾個必要的元素。

→即使研究主題相同，因為切入角度、研究素材（資料）、方法理論（研究者專業）等，使得研究方向、甚至研究結論會大異其趣。

報告撰寫的幾個關鍵要素

- 原則上，一篇研究報告包含至少三個要素：

動機與目標：問題背景及動機、問題的重要性及其影響、具體（或量化）的研究目標；

文獻探討：相關研究方法及參考文獻、現有方法的優勢及限制、本篇研究貢獻（市場區隔）；

本研究特色：本文研究方法及素材、主要研究發現（及其意涵）、本文適用時機及限制。

- 註：研究發現的價值除了創新之外，也希望能夠具有實質意涵及影響。

報告的格式要求

- 一篇完整報告至少包含以下幾項：

→標題 (Title)

→摘要(Summary或Abstract)

→報告主體。其中包括動機及背景、研究目的、資料來源與研究方法、分析結果說明、詮釋與結論

→參考文獻

→附錄（圖表、附件、...）

標題與摘要

- 建議標題另起一頁，與作者、日期等訊息獨立放在首頁。
- 摘要為一篇文章的濃縮，建議篇幅不多於一頁，以簡明扼要的原則闡述文章的撰寫動機（即背景）、目的、使用的資料及方法、大致的結論。

報告主體

- 報告最重要在於「定義問題」，詳細說明文章的動機以吸引讀者興趣，繼而提出與動機有關的探討目標（即問題）。

➔最後的結論需要回答這個問題！

- 定義「目標讀者」(Target Audience)

➔文章的用詞及整理，配合讀者及其閱讀目的，才能發揮文章的效果。

起承轉合

- 文章除了有一個中心主題前後貫穿，脈絡整理遵循主題外，也需考慮章節、段落、甚至是句子之間的連貫，切忌隨意轉換語氣、立場不一致、或是給出風馬牛不相及的評語。

→臺灣學生的報告容易在「轉」及「合」出現問題，尤其是文章結尾常有虎頭蛇尾的現象，草草給出結語，讓整篇文章功虧一簣。

報告需要注意的其他地方

- 用字遣詞、語句通順（敘述過於口語化）
- 標點符號、避免不必要的語助詞或虛詞
- 圖表與說明、圖表標示
- 參考文獻的引用
- 「專有名詞」定義不夠清楚（中英文名詞的對照）

引述與剽竊(Plagiarism)

- 寫文章最忌諱的是「文抄公」，將別人的文字原封不動地照抄，尤其網際網路發達的今日，抄襲更是普遍。

→如有需要引用別人的文字時，需標示哪些部分是「節錄」，又是從哪裡節錄。

→提出的觀點如果是引述別人的想法，需要詳加說明，文章中也應區隔哪些觀點是自己的創見。