漫談統計陷阱

洪志真
國立交通大學統計所

2012統計科學營
September 5, 2012
中央研究院統計科學研究所

統計思考 (Statistical Thinking)

- Statistics is the science of learning from data (資料, 數據)
- Data are numbers, but they are not "just numbers"
- 資料(data)+說明 (context)=資訊 (information)
- 例：50 (just a number)
  - 50公斤是可接受的體重
  - 50分則是不及格的分數
- 統計是將資料(數據)適當處理後，彙整成資訊的過程

Always Look at the Data

- 一般認知未必為真
- 真實資料才能提供正確的資訊
  - 某年美國黑人 vs. 美國白人 之比例
    - 白人 認為 23.8% vs. 49.9% (average)
    - 人口調查局 (Census Bureau): 11.8% vs. 74%
  - 交通大學的總學生數

資料勝過軼聞
Data Beat Anecdotes

軼事, 趣聞

- 軼聞 (anecdotes)是令人印象深刻的事件，多為特例，可能產生誤導。研究資料才能提供正確的結論
- 電纜線與白血病的案例
  - 傳聞：聽說電纜線產生的電磁場會引發白血病
  - 研究：五百萬美元經費歷經五年的研究顯示
    - 暴露在電纜線產生的電磁場與白血病沒有關聯
    - E. W. Campion, "Editorial: power lines, cancer and fear," *New England Journal of Medicine*, 337, No. 1 (1997).

留意隱藏變數
Beware the Lurking Variable

- 表面的資料未必可信
- 比較兩航空公司的班機延誤率:

	On time	Delayed	Rate of Late Flights
Alaska Airlines	3274	501	501/3775 =13.3%
America West	6438	787	787/7225 =10.9%

- America West 似乎較佳

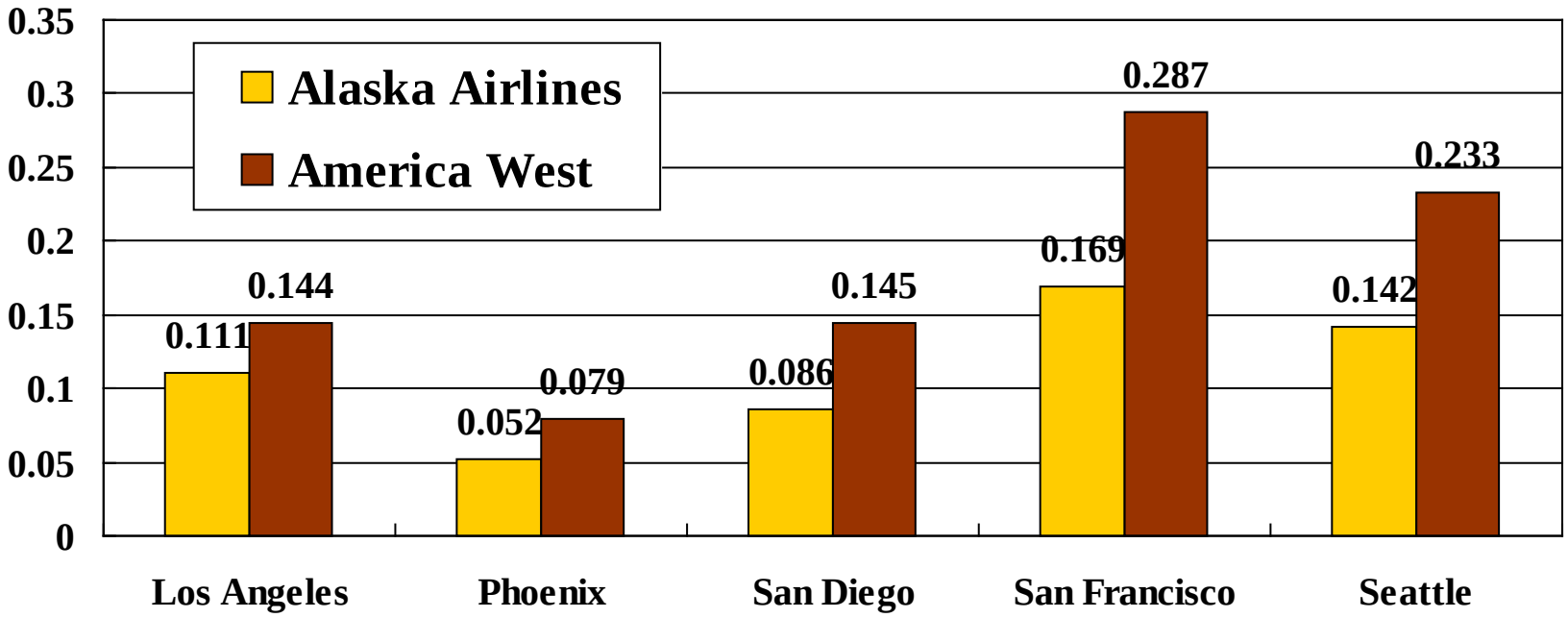
留意隱藏變數(續)

- 考慮隱藏變數：班機起飛城市

	Alaska Airlines		America West	
	On Time	Delayed	On Time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61
Total	3274	501	6438	787

留意隱藏變數(續二)

- 每一個班機起飛城市的班機延誤率都是Alaska Airlines 較低



留意隱藏變數(續三)

- Hint: The hub of Alaska Airlines is in Seattle and the hub of America West is in Phoenix.
- 資料來源
  - A. Barnett, "How numbers can trick you," *Technology Review*, October 1994, Department of Transportation.
- Simpson's paradox

辛普森悖論(Simpson's Paradox)

- 當研究兩個變數之間的關聯性，有可能存在一個**隱藏變數(lurking variable)**，而當隱藏變數被考慮時，兩個變數之間的關聯性方向剛好與隱藏變數沒有被考慮時**相反**
- 隱藏變數會將樣本分成子群，當沒有考慮到這個有不同群組的因素時，可能會對兩個變量之關聯性得到錯誤的結論

歧視？(Simpson's Paradox)

考慮以下兩組男性和女性申請大學的錄取率

counts	Accepted	Not accepted	*Total*
Men	198	162	*360*
Women	88	112	*200*
Total	*286*	*274*	*560*

percents	Accepted	Not accepted	
Men	55%	45%	
Women	44%	56%	

男性被接受的比例較高： 歧視？

歧視？ 
(Simpson's Paradox)

分開成申請商學院和藝術學院之人數

商學院

counts	Accepted	Not accepted	*Total*
Men	18	102	*120*
Women	24	96	*120*
Total	*42*	*198*	*240*

percents	Accepted	Not accepted	
Men	15%	85%	
Women	20%	80%	

在商學院有較高比例的**女性**被接受。

歧視？
(Simpson's Paradox)

藝術學院

counts	Accepted	Not accepted	*Total*
Men	180	60	*240*
Women	64	16	*80*
Total	*244*	*76*	*320*

percents	Accepted	Not accepted	
Men	75%	25%	
Women	80%	20%	

在藝術學院亦有較高比例的**女性**被接受。

歧視？(Simpson's Paradox)

- 因此，各學院內相對於男性有較高比例的女性被接受。沒有任何對女性的歧視！
- 這是辛普森悖論的一個例子。當潛藏變數（申請學院：商學院或藝術學院）被忽略時的數據似乎顯示出對女性的歧視。然而， 當學院因奇被考慮進來時，關聯性是相反的，而且反過來顯示存在對男性的歧視。

小心隱藏的變數

- 範例: 冥想和老化

(*Noetic Sciences Review*, Summer 1993, p. 28)

- 解釋變數:是否有作冥想的練習 (yes/no)
- 反應變數:與年齡有關的某酵素之測量值

- 一個人若很注意自己的健康也可能會影響此反應變數之結果
- 同時，也可能會想嘗試冥想

資料來源很重要
Where the Data Come from Matters

- 專欄作家安・蘭德斯 (Ann Landers) 以
  “如果可以重新再來，你是否還要孩子?”
- 調查其讀者的意見得到一個聳動的結論 ：
  - 70%的父母認為有小孩不值得 (約一萬封回信)
- 另一問卷調查給所有父母有相同表達機會，結果顯示：
  - 91%的父母認為有小孩很值得

資料來源很重要(續)

- Ms. Landers 的讀者多為親子關係有問題的父母， 調查結果自然偏頗
- 網路調查、街頭訪問也有類似的情形

Variation is Everywhere
變異處處可見

- 資料不可能一成不變
  - 個體變異 (如身高體重)
  - 量測誤差
- 統計幫助我們處理變異 (variation)

結論的不確定性
(Conclusions are not certain)

- 乳房X光攝影 (mammograms)是否可以降低乳癌死亡的風險？
- 由13個 臨床試驗資料顯示，乳房攝影可以使50~64歲女性死於乳癌的風險降低26%
- 風險降低率之95%信賴區間(confidence interval)為17%~34%
  - H. C. Cox, "Editorial: benefit and harm associated with screening for breast cancer," *New England Journal of Medicine*, 338, No. 16 (1998)
- Statistics gives us a language for talking about uncertainty that is used and understood by *statistically literate* people everywhere.

- 在大部分的時候，讓我們陷入困境的，並非我們不知道的事物，而是我們認為不會讓我們陷入困境的事物。

— 華德(Artemus Ward，美國幽默作家)

統計數字會撒謊

- 大仲馬的作品多曲折感人，而大仲馬又多私生子，所以，取笑譏諷他的人，往往把他的作品比作他的私生子。最使他頭痛的是巴黎統計學會的秘書長李昂納，這人是大仲馬的朋友，每次舉統計數字的例子，總是說大仲馬的情婦和私生子有多少。
- 有一年該統計學會開年會，大仲馬估計，李昂納又要大放厥詞，說他的壞話了。於是他請求參加年會，獲得了批准。果然不出大仲馬所料，李昂納又舉他的情婦和私生子的例子。
- 李昂納報告完畢，請大仲馬致詞，一向不願在大庭廣眾之下發表演講的大仲馬，這次卻破例登臺說：“所有統計數字都是撒謊的，包括有關本人的數字在內。”聽眾哄堂大笑。
- **（網路笑話大全）**

統計數字 vs. 謊言

- 世界上有三種謊言，就是謊言，天大的謊言，與**統計數字**
- There are three kinds of lies: lies, damned lies, and *statistics*.

-- Benjamin Disraeli (1804－1881，英國首相，議員，保守黨政治家和文學人物)

是誰在讓數字說話？

- 為什麼很多人會被統計數字騙了呢？
  - 有數據支持的論點，大家通常容易相信
  - 但讓數字「說話」的是人
  - 說實話？說謊話？
  - 故意操弄數據？還是只是對數據處理不當？
- 統計，在一個重視事實的文化中非常有用，但也有人利用它作為惡意誇大或簡化、甚至隱藏或曲解事實以達到其特定目的之工具
- 水能載舟也能覆舟

統計是必備知能

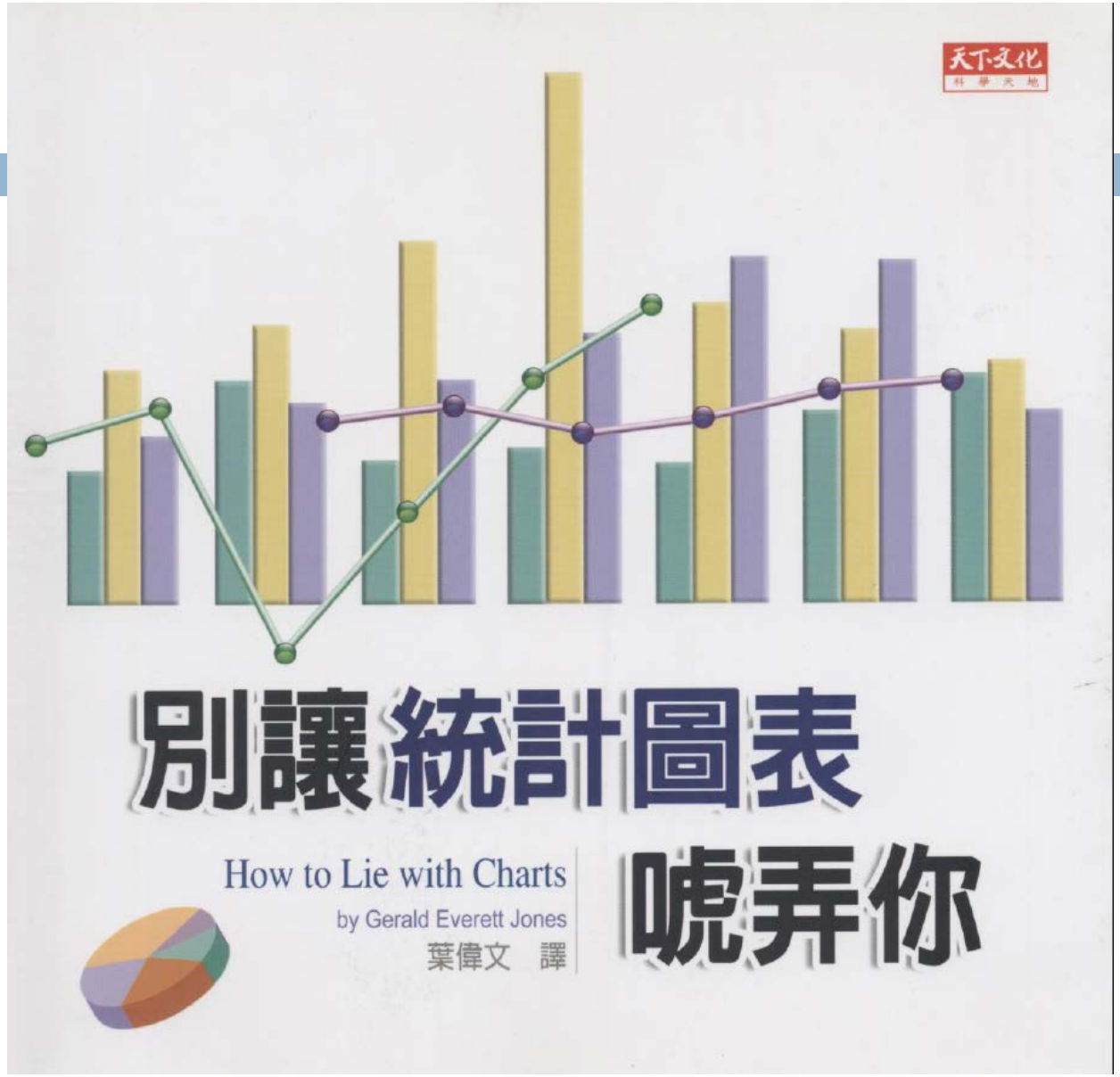
- 現今，在報告社會經濟趨勢、商業狀況、民意調查和普查的數據時，統計方法或者統計術語是不可少的。但如果作者不能正確理解並恰當地使用這些統計語言，而讀者又並不能真正了解這些術語的涵義，那麼，所敘述的統計結果對讀者毫無意義。
- “終有一天，統計思考會像閱讀與寫作能力一樣，成為公民不可不具備的能力。”

(原文：“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”)

— 威爾斯(H. G. Wells ， 1866－1946 ，英國著名科幻小說家)

HOW TO LIE WITH STATISTICS
別讓統計數字馬扁了你

【別讓統計數字騙了你】
Darrell Huff著（1954）、鄭惟厚譯（2005）
天下文化出版


天下文化
科學天地
別讓統計圖表
唬弄你
How to Lie with Charts
by Gerald Everett Jones
葉偉文 譯

學習目標

- 了解統計能如何騙人的招術
- 壞蛋們早就會了，為了自衛，老實人也必須學會

1. 有內建偏差的樣本

- 著名範例：美國總統選舉民調 (FDR Poll)
- 1936年
- Franklin D. Roosevelt (FDR) vs. Alf Landon
- 民調預測共和黨候選人Landon會大勝,結果 FDR 大勝
- 民調這麼不準嗎？ 統計沒用嗎？
- 原因：採用電話調查
- 雖然樣本數很大但有嚴重內建偏差(bias) — 只有有錢人才裝得起電話

有內建偏差的樣本

- <<時代雜誌>>1950年代在評論紐約<<太陽報>>的某篇報導時曾寫道：「1924年畢業的耶魯大學畢業生，平均年薪為25111美元。 」

  PS. 當時一般人的平均年收入低於10000美元

這份報導有哪些可能的誤導？

假設數字和抽樣上沒有任何造假……

- 數字的精確程度令人懷疑　（多報 or 少報）
- 樣本足以代表全體嗎?
- 問卷設計的適確性?

就算樣本夠大，問卷設計得宜，

但受訪者常常會想要給一個讓訪問員喜歡的答案，

訪問員也常挑選特定族群訪問

有時我們也許得「拐彎抹角」用別的方法而非直接提問

面子問題

- 曾有人挨家挨戶的訪問「你家讀什麼雜誌?」
- 結果顯示許多人喜歡高格調的<< Harper's >> ，讀八卦雜誌<<True story>>的人卻不多，但這卻和出版商的數據差異甚大……

有什麼替代的方法?

內建偏差之來源

- 有代表性的樣本，是指把各種偏差來源都排除的樣本。
- 電話民調有哪些可能的內建偏差?
- 在火車站、大賣場做民調又可能忽略了哪些族群?
- 若想知道大家平均每天刷幾次牙、洗幾次澡，抽樣得到的數據會準確嗎？
  - 若否，高估或低估？

內建偏差之來源

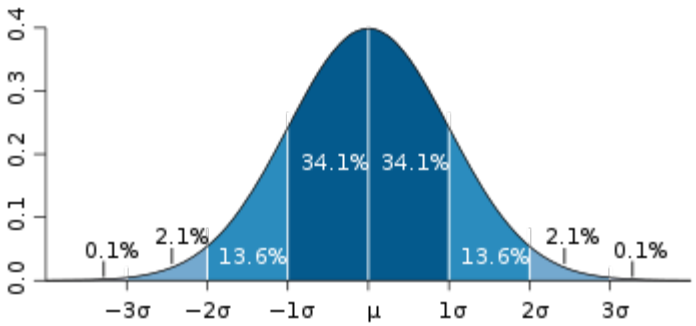
- 敏感性問題
  - 面子、金 錢、道德規範、法令規範 etc.
- 樣本涵蓋率不足
- 問卷低回收率
- 自發性回應

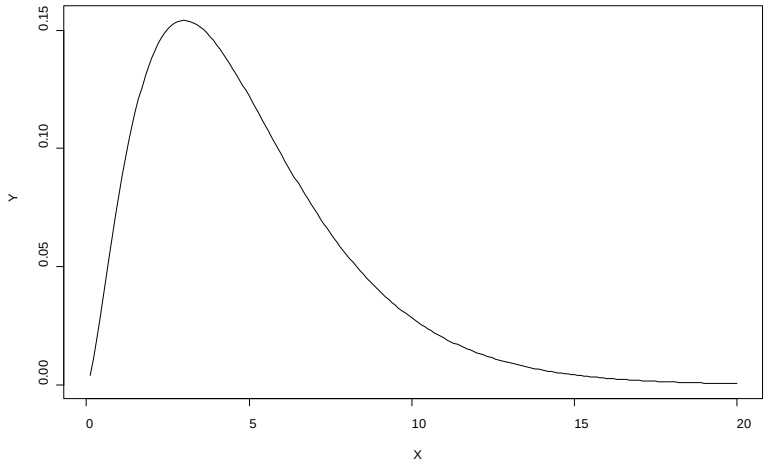
抽樣方法

- 簡單隨機抽樣（simple random sampling）
  - 由母體隨機抽樣而得
  - 符合許多統計理論的假設
  - 有些情況難以取得，或花費太高
- 分層抽樣（stratified random sampling）
  - 將母體依特性分成若干組(稱為「層」，「層」內同質性高），由每層依比例隨機抽樣
  - 民調、市場調查常用
  - 有些情況層與層之間可能難以辨認，每一層的比例也難以拿捏

2. 精心選擇的平均(Average)

- 三個常用的中心 (central tendency) 測度
  - 平均數（mean）
  - 中位數（median）
  - 眾數（mode）
- Average could mean any of them
- 對稱 vs. 偏斜分配





騙子往往根據目的挑最有利的平均

Incomes:

- $9000
- $9000
- $9000
- $12,000
- $120,000
- $85,000
- $15,000

Mean = $37,000

Median = $12,000

Mode = $9000

Each is a legitimate average but can serve conflicting purposes

平均數 vs. 中位數

- 當機率分佈有嚴重偏斜時（skewed），平均數會受到嚴重影響，像是「所得」這種數字就有嚴重右偏斜（right skewed）
  - 例:假設你跟台灣首富是國小同班同學，那麼你們班的平均月收入可能高得嚇人。但是月收入的眾數可能卻只是25 K不到，這種時候中位數可能最具代表性。
- 對「離群值」，中位數比平均數來得穩健 (robust)

用哪個平均？

- Person Money
- John 2
- Ann 3
- Bob 1
- Mary 10
- Sue 5
- Carol 2
- Ken 999

- Mean $146
- Median $3
- Mode $2

點估計 vs. 區間估計

- 對這些類型的點估計，要能了解它們有多少誤差
- 更好的方法是，利用區間估計來取代它們
- 信賴區間（Confidence Interval）
  - 例: 沙漠地區的日均溫可能看來舒適，但早晚溫差卻很大
  - 例: 在 95% 的信心水準下，25歲台灣人的平均月收入是 25k ± 3k

顯著水準和信賴水準

- 統計推論 (statistical inference)
  - 假設檢定(testing hypotheses)
    - 顯著水準（significance level）
  - 信賴區間（confidence interval）
    - 信賴水準 (confidence level)

例:醫院檢驗愛滋病，篩檢用的檢驗方法是 ELISA，此法使用顯著水準為千分之五之統計檢定來判定，亦即對任一次檢驗，產生HIV 假陽性(false positive)的機率不超過千分之五。

平均 (Average)

- **當你見到一個平均，除了要知道是那種平均外，還要搞清楚這是哪些東西的平均！**
  - 例:美國某鋼鐵公司曾聲稱員工的平均週薪在1948年和1940年比較增加了107%，聽起來加薪了很多。

    但事實上是1948年的正職人數較多，工時也較長。
  - 例:曾有份報紙寫道1949年一般美國家庭的收入是3100美元。

    但你最好別對這個數字太過認真，除非你了解這其中「家庭」的定義。

3. 隱藏起來的小數字

- 某牙膏廣告聲稱它能減少23%的蛀牙，而且這些結果來自令人信任的XX實驗室。

試用人數:12人

- 經過科學驗證，硬幣擲出正面的機率是80%。

但是我只有丟10次

- 某社區有450位孩童接種了小兒麻痺疫苗，680位沒有接種疫苗當作對照組。結果發現接種疫苗的孩童裡面一個小兒麻痺的病例都沒有。

  這足以說明疫苗對小兒麻痺有顯著性的效果嗎?

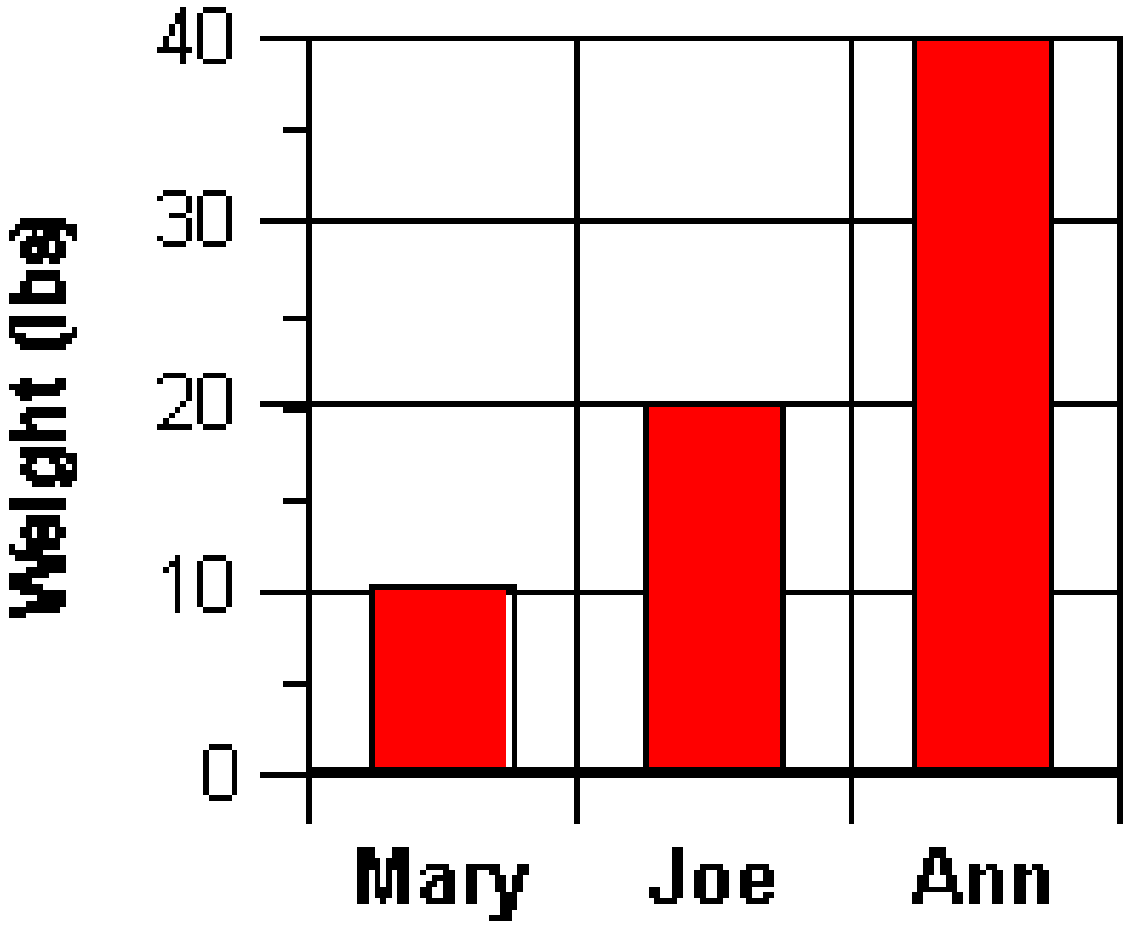
對照組也一個小兒麻痺的病例都沒有

4. 在圖上作文章

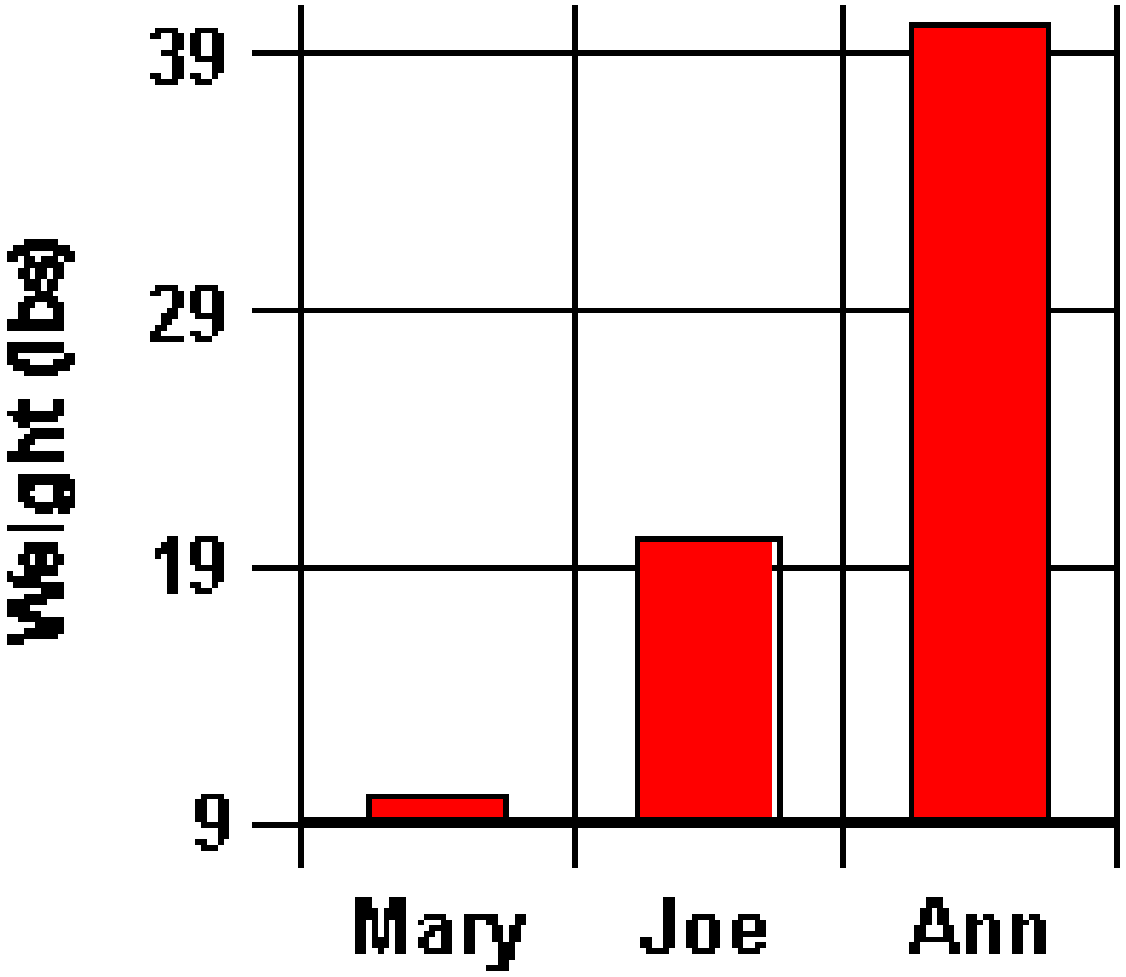
- 什麼是事實？
- **“Many of the truths we hold onto depend on our point of view”**

Ben Kenobi, Star Wars 星際大戰

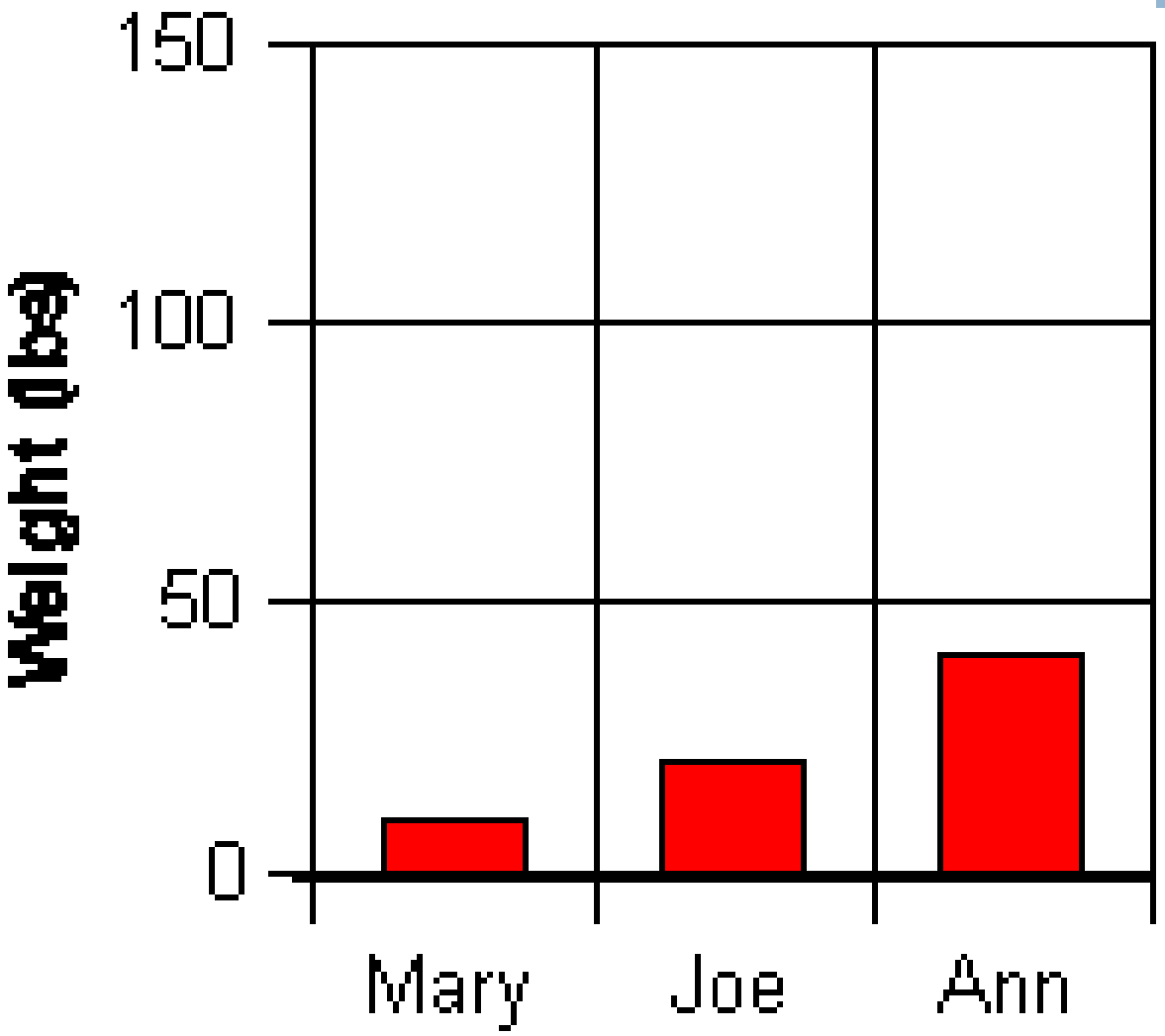
看這張圖



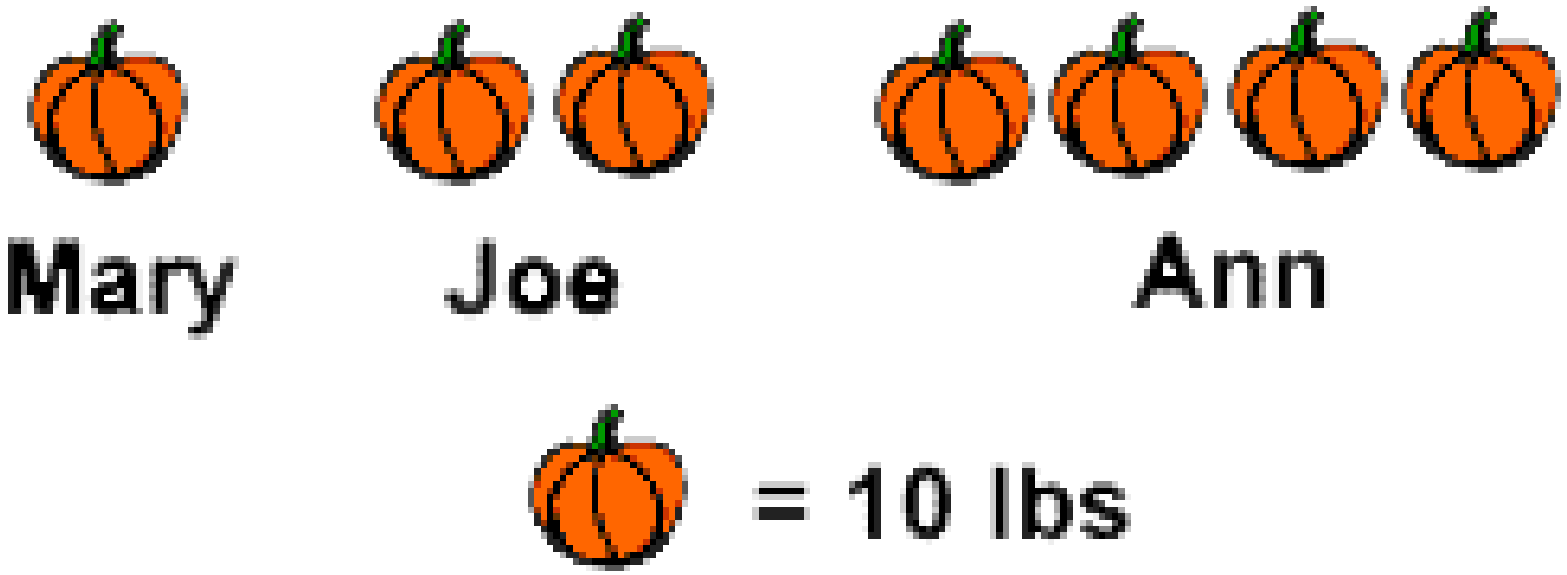
與這張圖比較



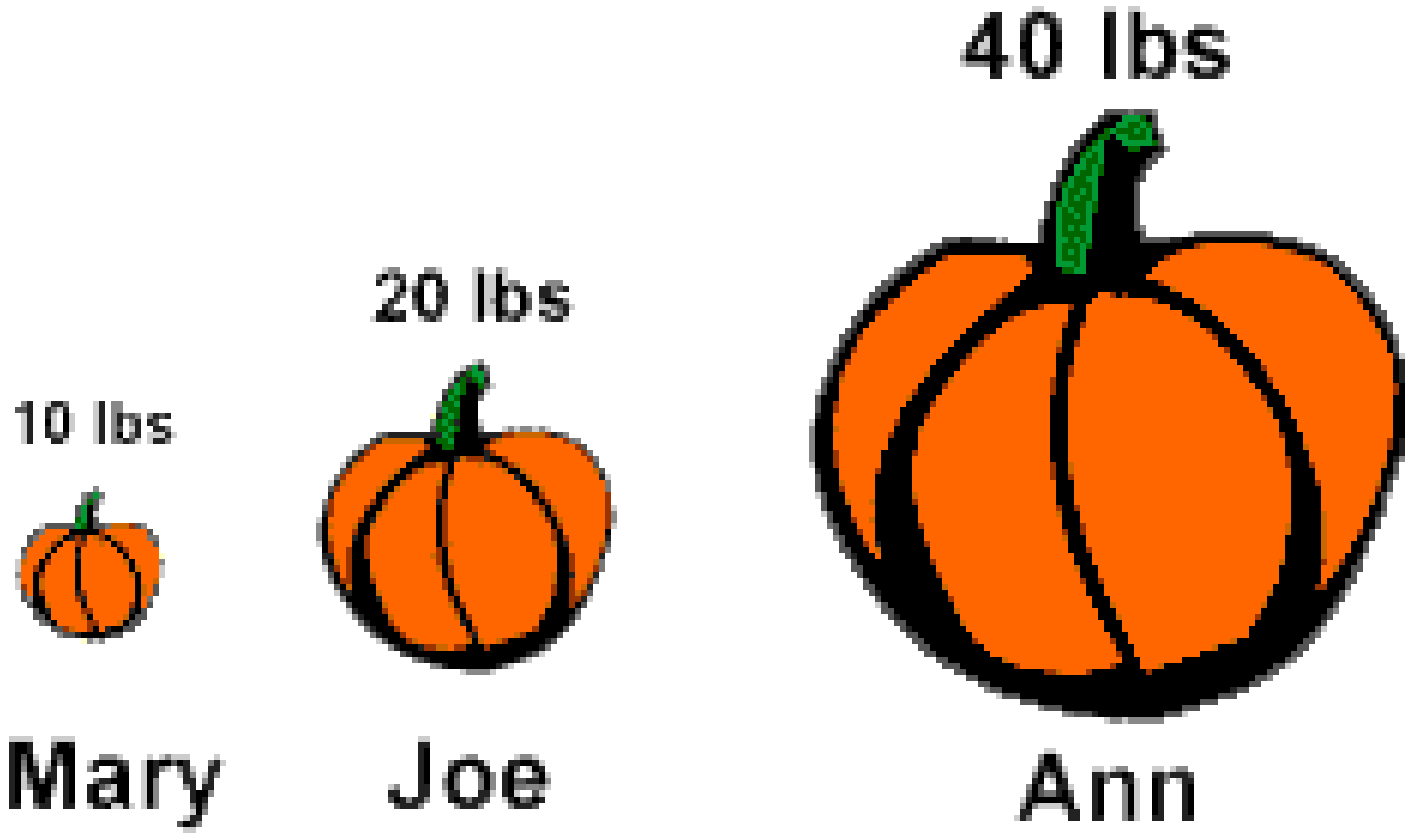
或與這張圖比較



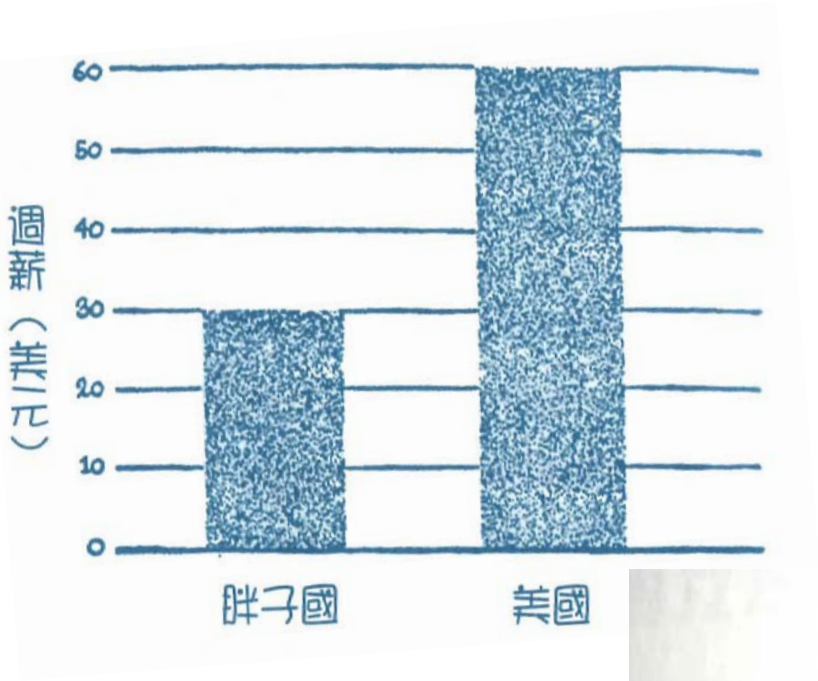
或更 fancy ， 用物件之圖案表示



這張圖有什麼問題？

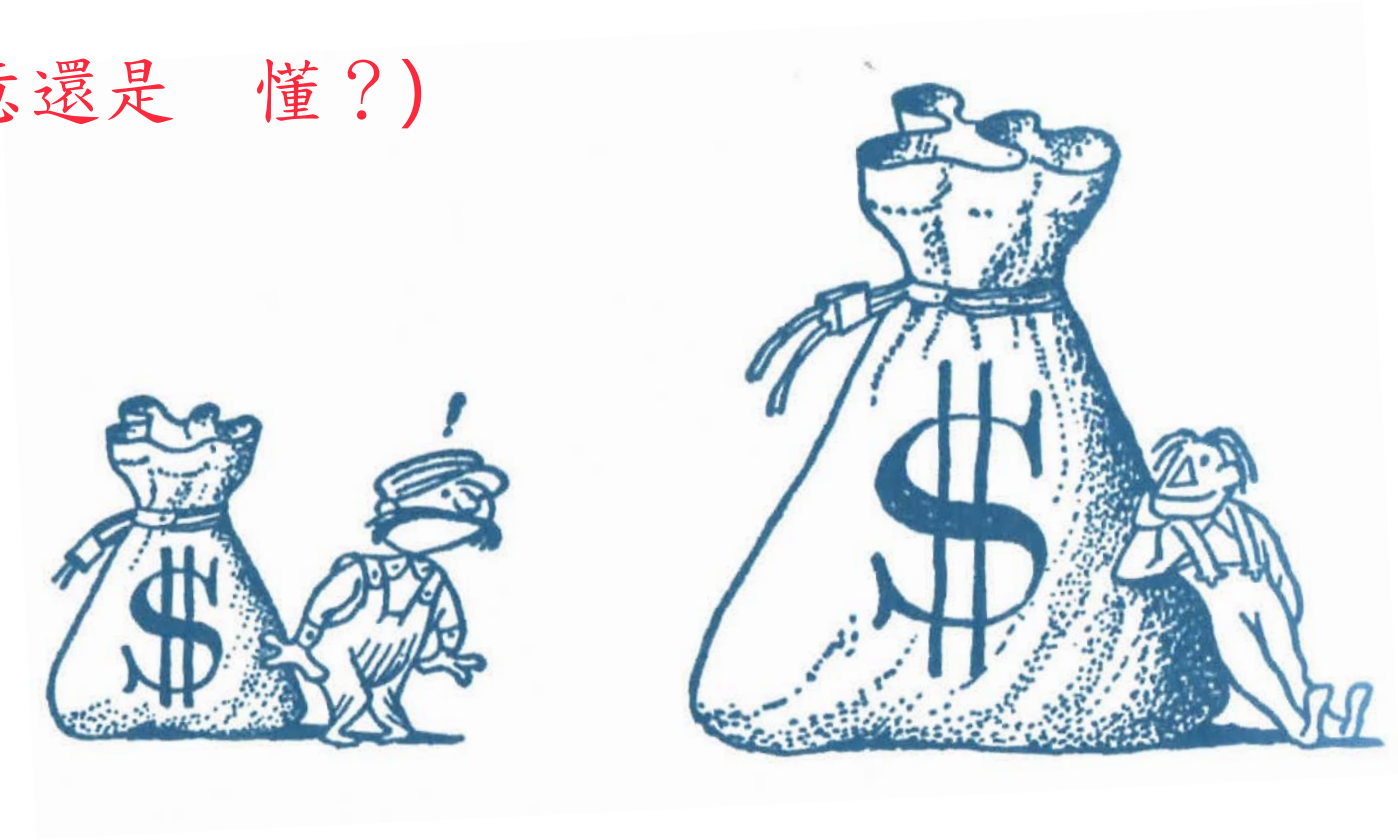


另例：美國週薪為胖子國的兩倍



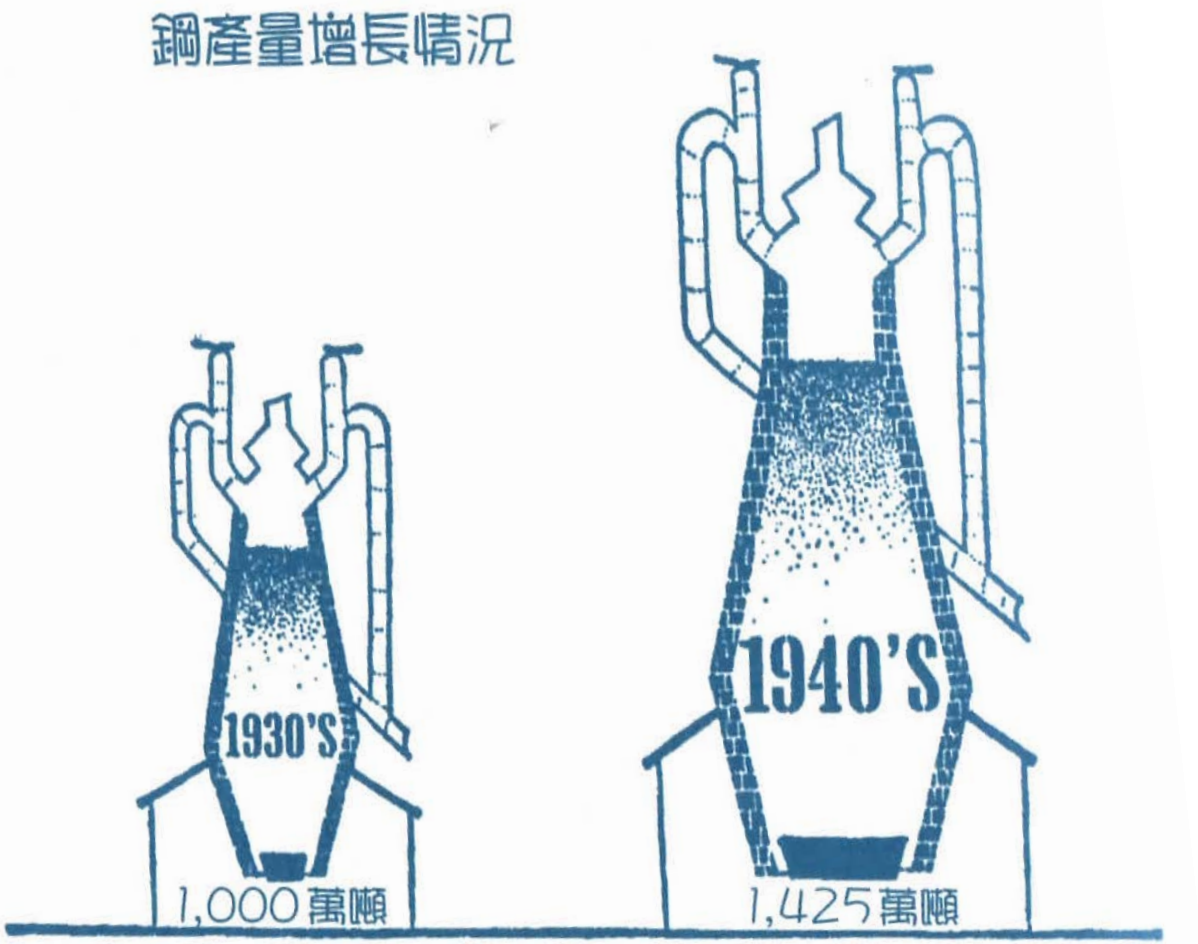
常犯的錯誤作圖法

(故意還是　懂？)



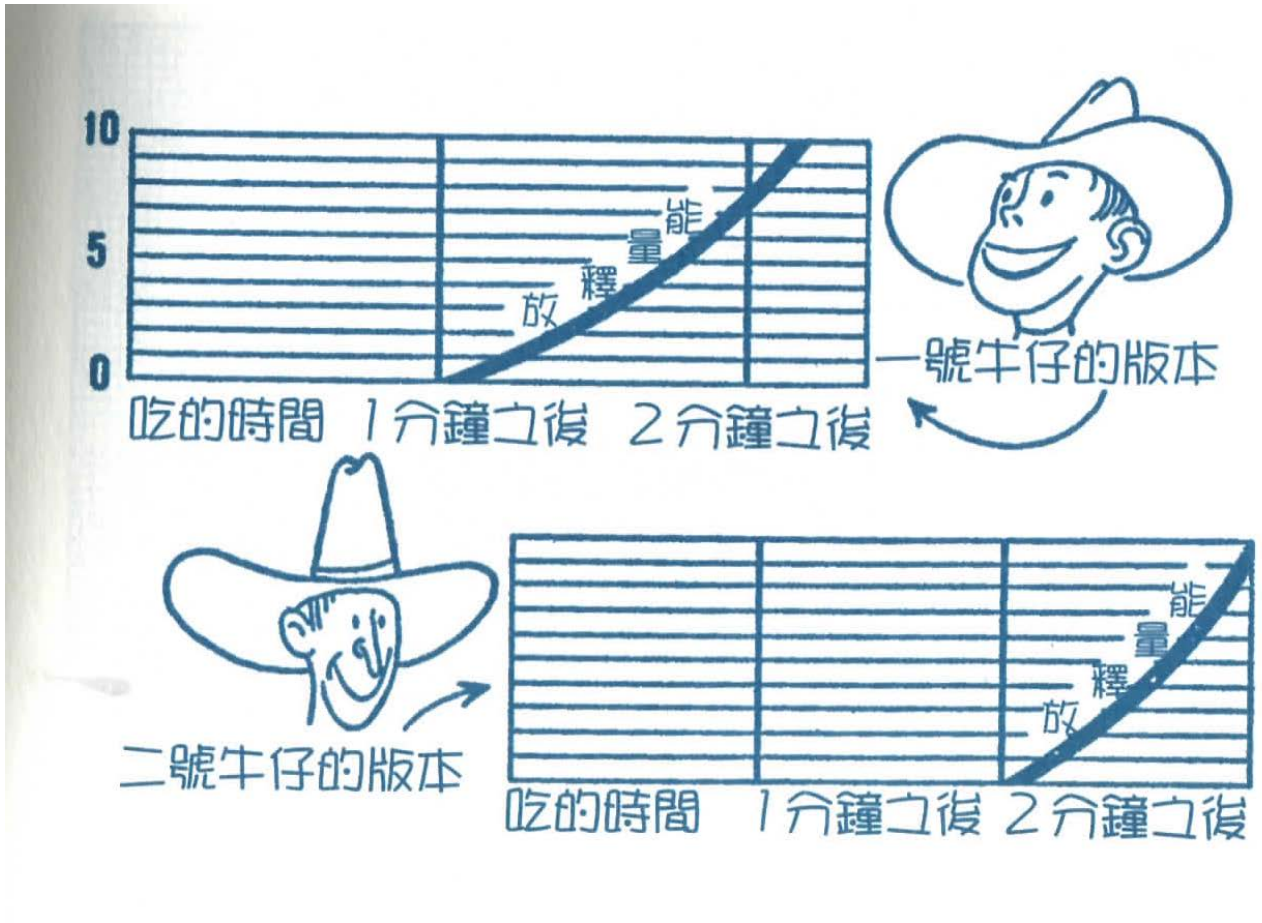
繪圖者太遜還是故意欺騙？

(原本42.5%的增長被畫成150%)



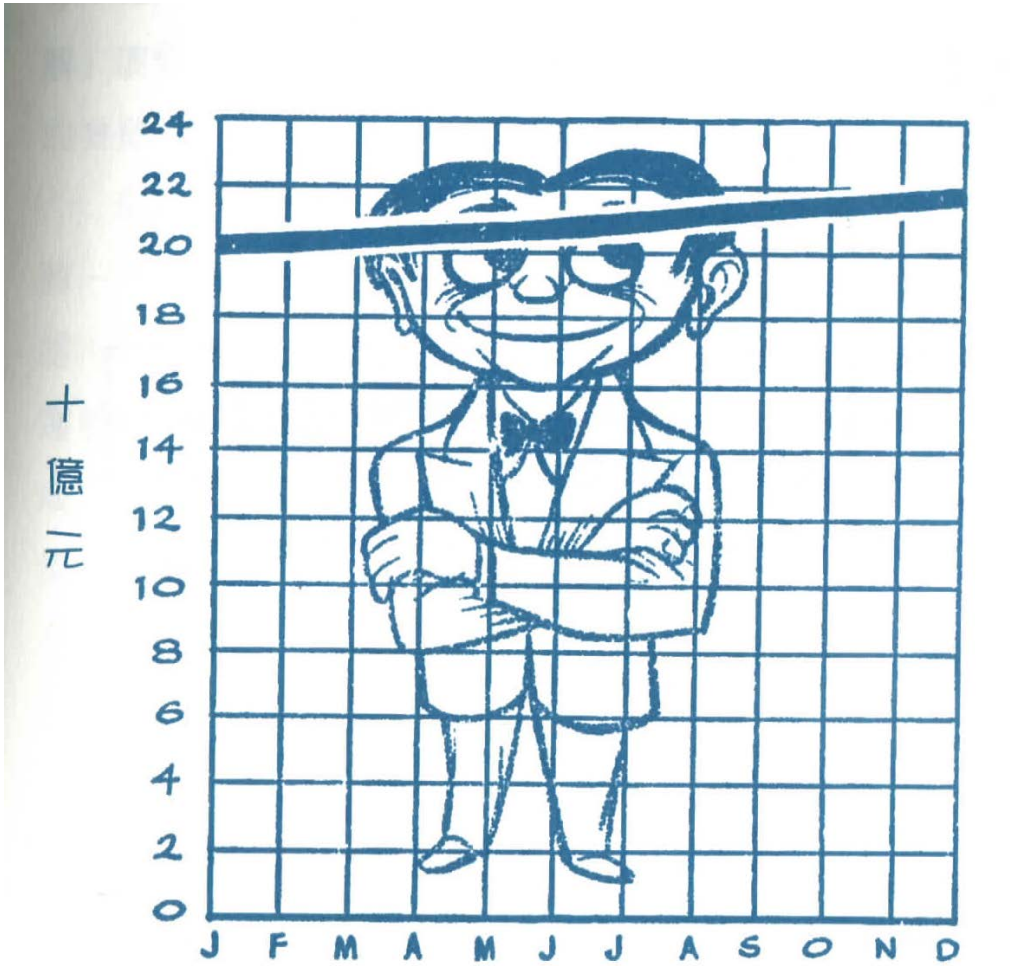
沒列的數字

- 該列卻未列出的數字，常會遭人忽略，結果形同欺騙!
  - 穀物早餐外盒圖案



常用來看趨勢之線圖

- 國民生產所得一年內增加10%

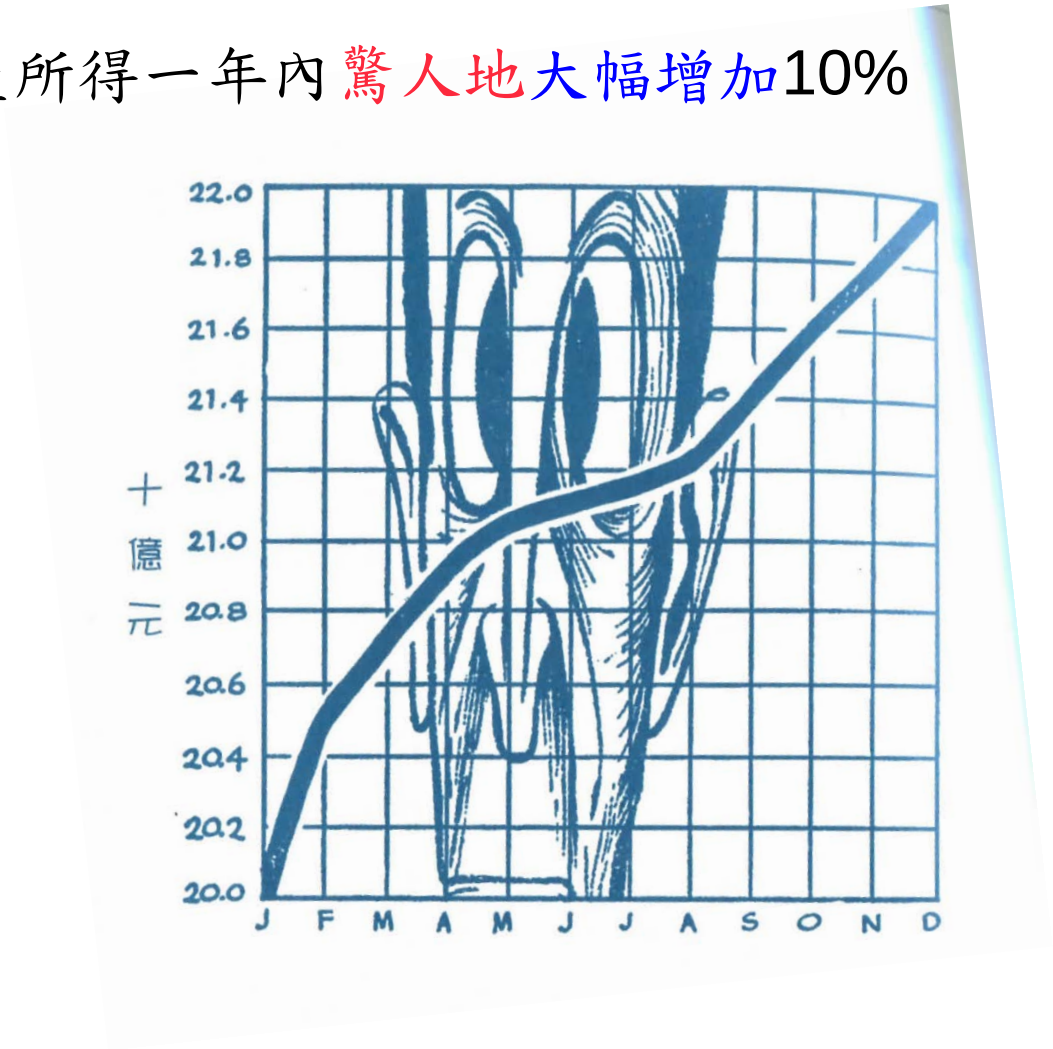


在圖上作文章 (縱軸切掉下面沒資料的部分)

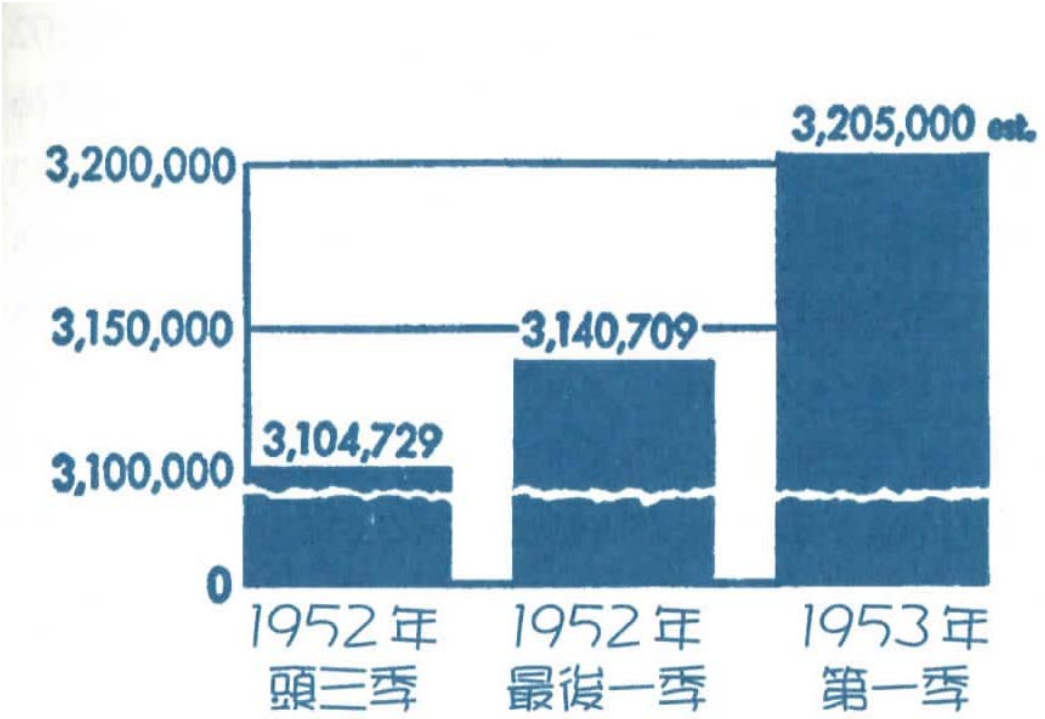
- 國民生產所得一年內大幅增加10%

在圖上作文章 (再加上改變尺度)

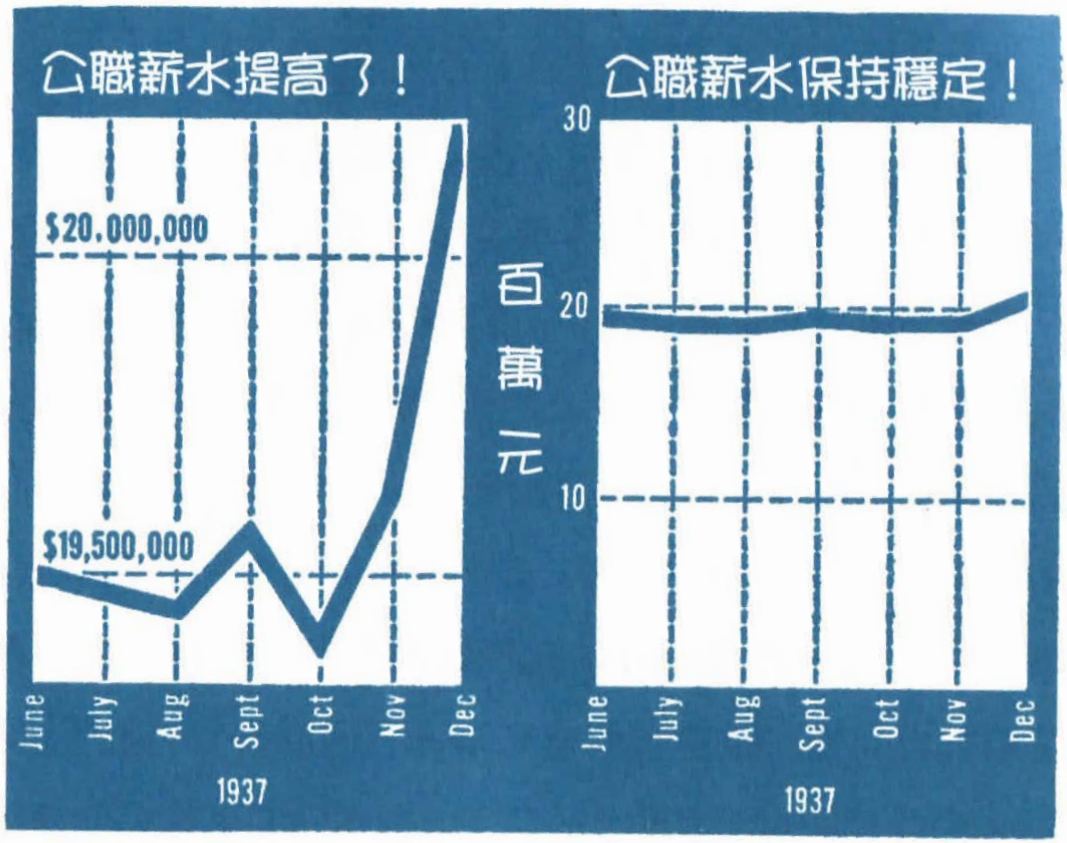
- 國民生產所得一年內驚人地大幅增加10%



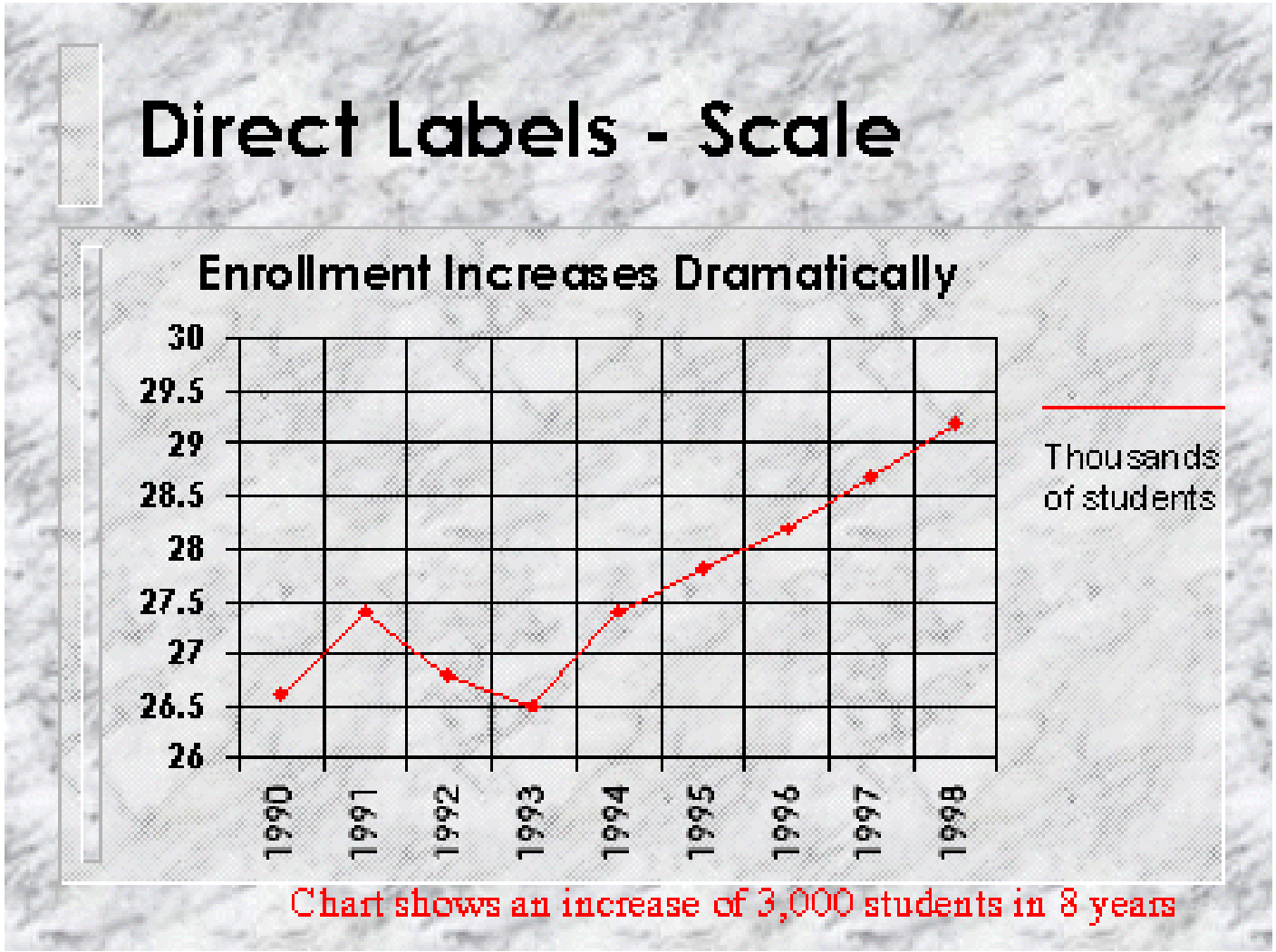
在圖上作文章(縱軸切掉中間的部分)



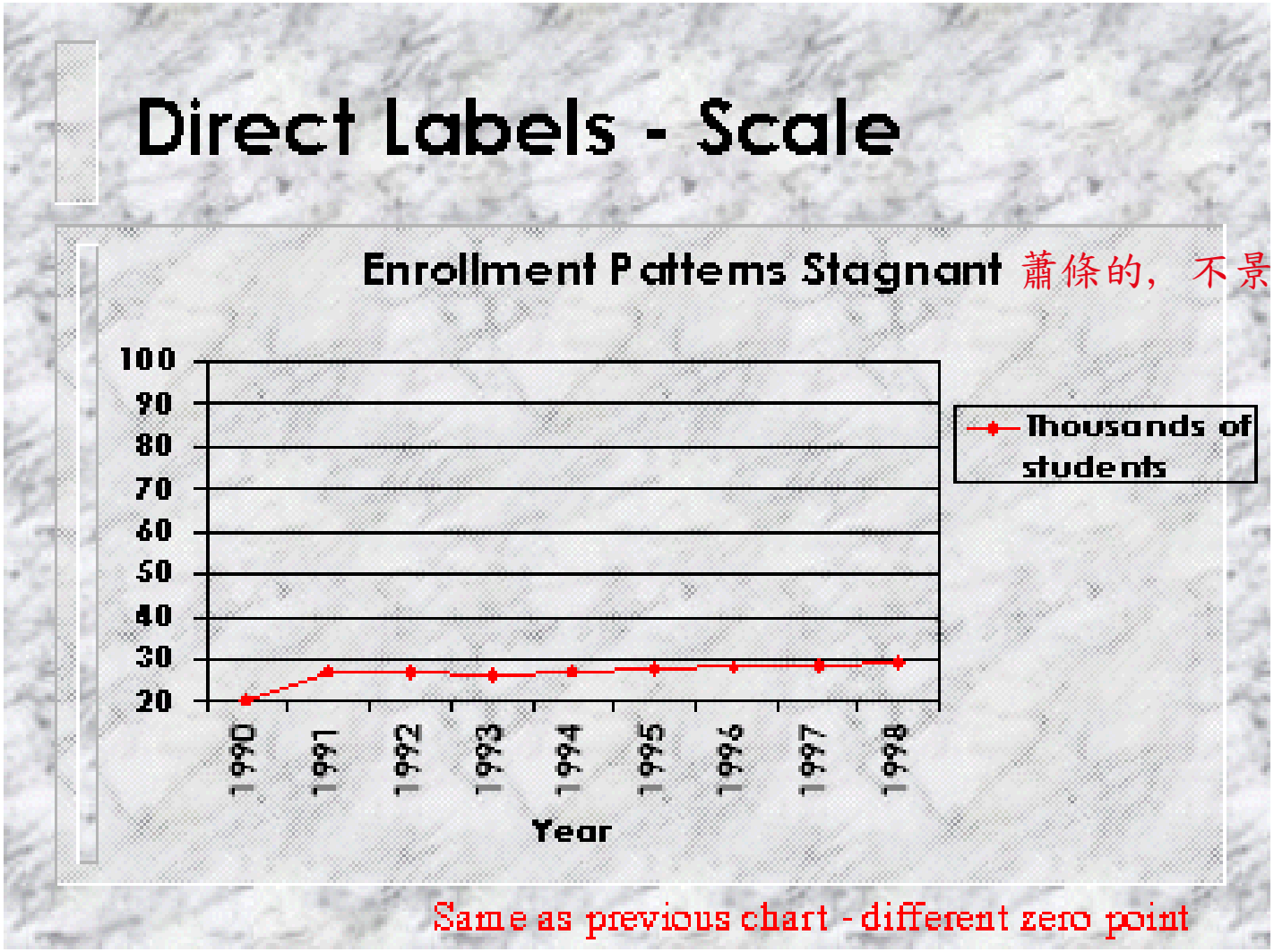
圖會說話－挑想說的畫


公職薪水提高了！
$20,000,000
$19,500,000
June
July
Aug
Sept
Oct
Nov
Dec
1937
百萬元
公職薪水保持穩定！
30
20
10
June
July
Aug
Sept
Oct
Nov
Dec
1937

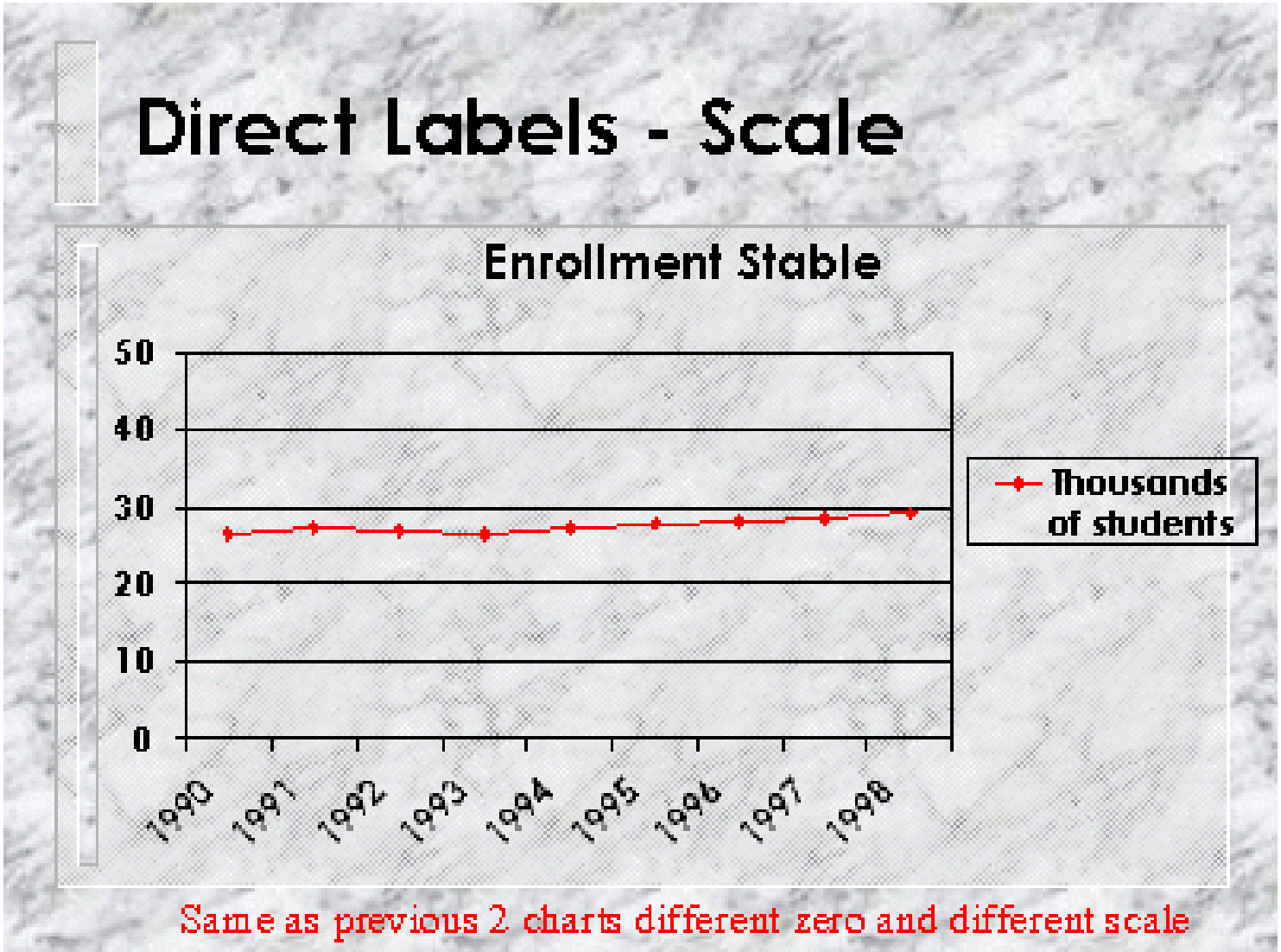
想說學生入學率急速上升



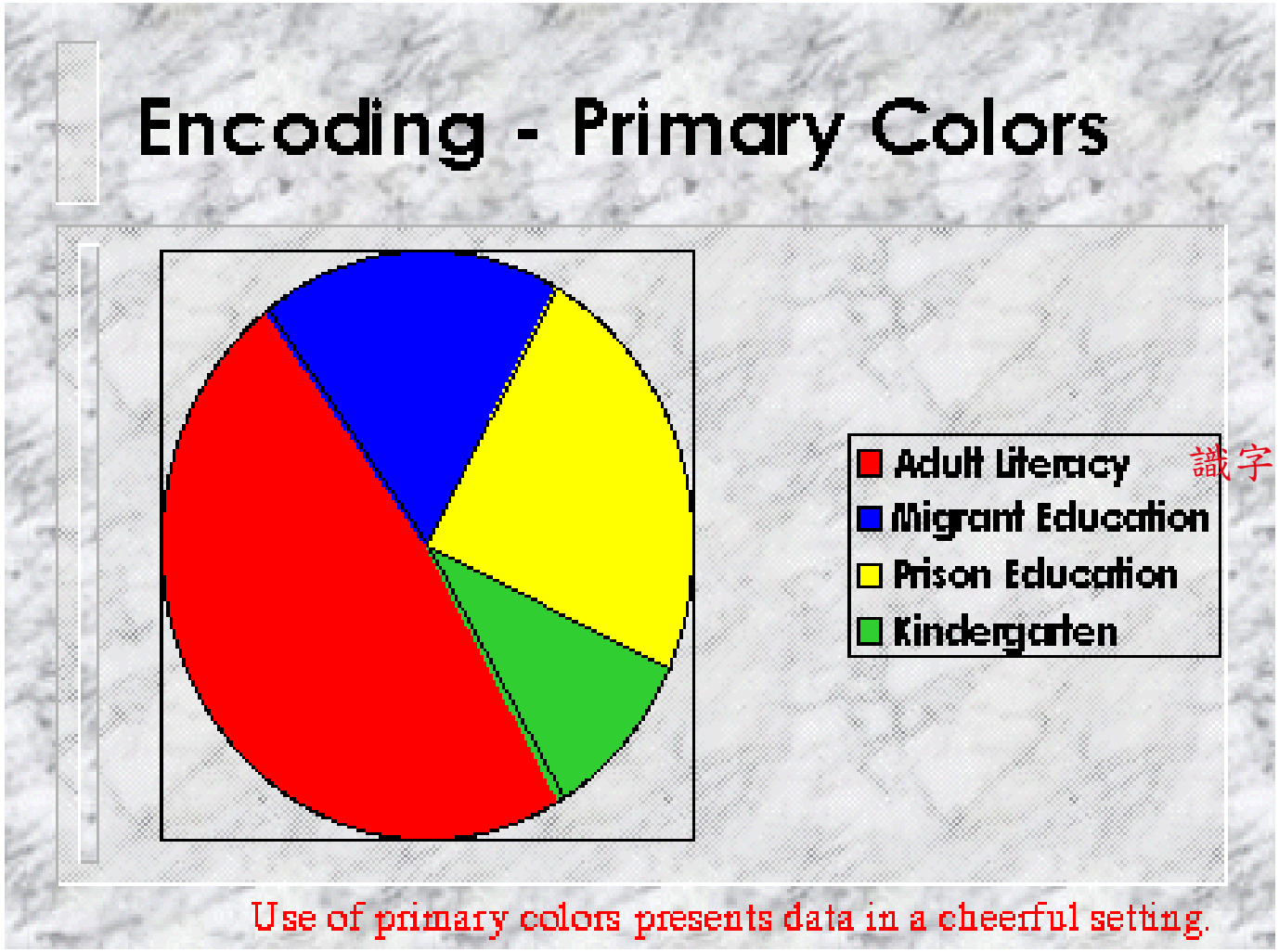
想說學生入學率很低


Direct Labels - Scale
Enrollment Patterns Stagnant 蕭條的，不景氣的
100
90
80
70
60
50
40
30
20
1990
1991
1992
1993
1994
1995
1996
1997
1998
Year
Thousands of students
Same as previous chart - different zero point

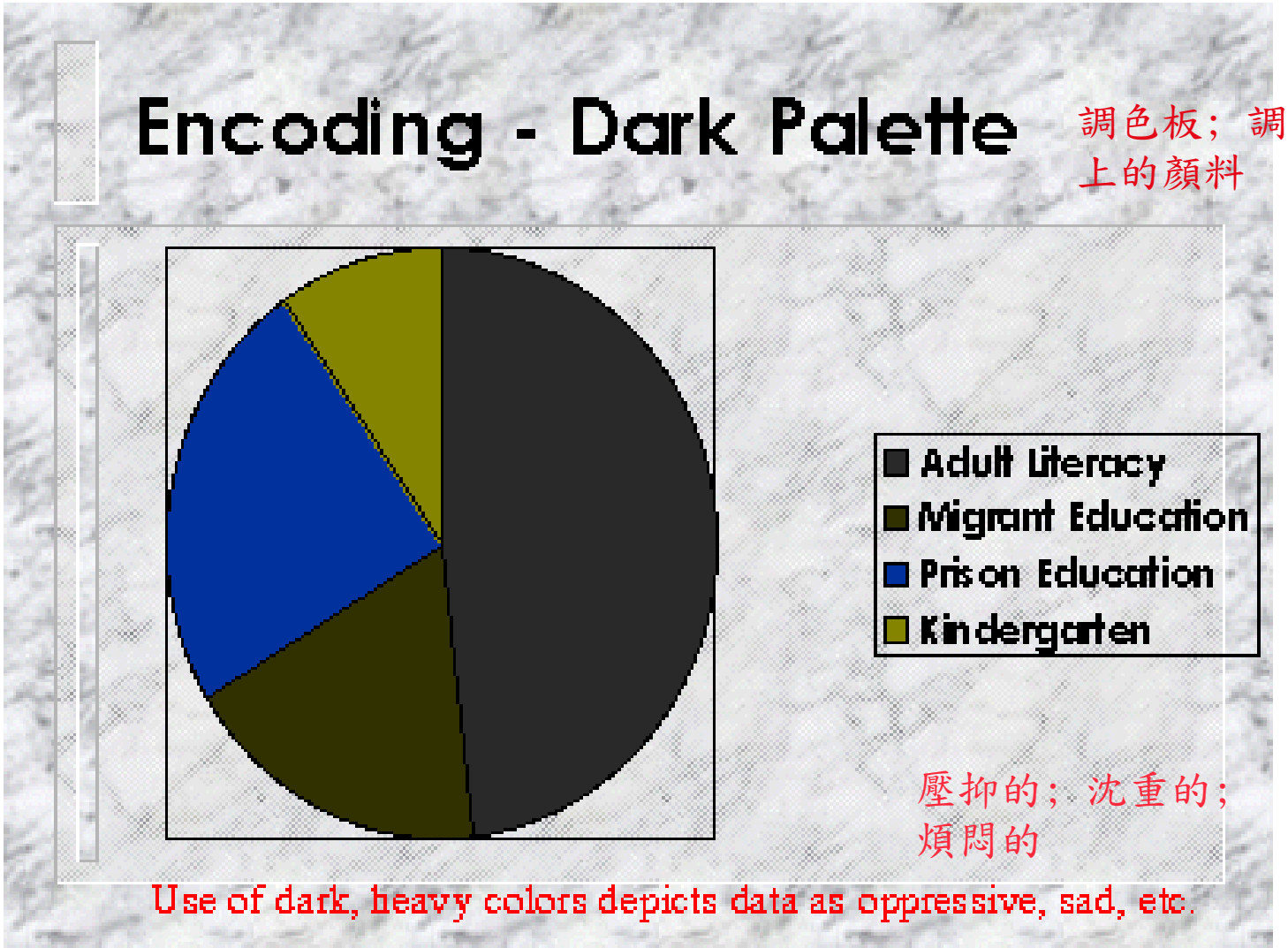
想說學生入學率很穩定


Direct Labels - Scale
Enrollment Stable
50
40
30
20
10
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
Thousands of students
Same as previous 2 charts different zero and different scale

顏色也有涵意：明亮—正面


Encoding - Primary Colors
Adult Literacy
Migrant Education
Prison Education
Kindergarten
識字；讀寫能力
Use of primary colors presents data in a cheerful setting.

暗色系－負面


Encoding - Dark Palette
調色板；調色板上的顏料
Adult Literacy
Migrant Education
Prison Education
Kindergarten
壓抑的；沈重的；煩悶的
Use of dark, heavy colors depicts data as oppressive, sad, etc.

5. 似相關而非相關的數字

- 經過著名實驗室證實，某感冒藥秘方只要放14g在試管裡，就可以在11秒內殺死31108個細菌

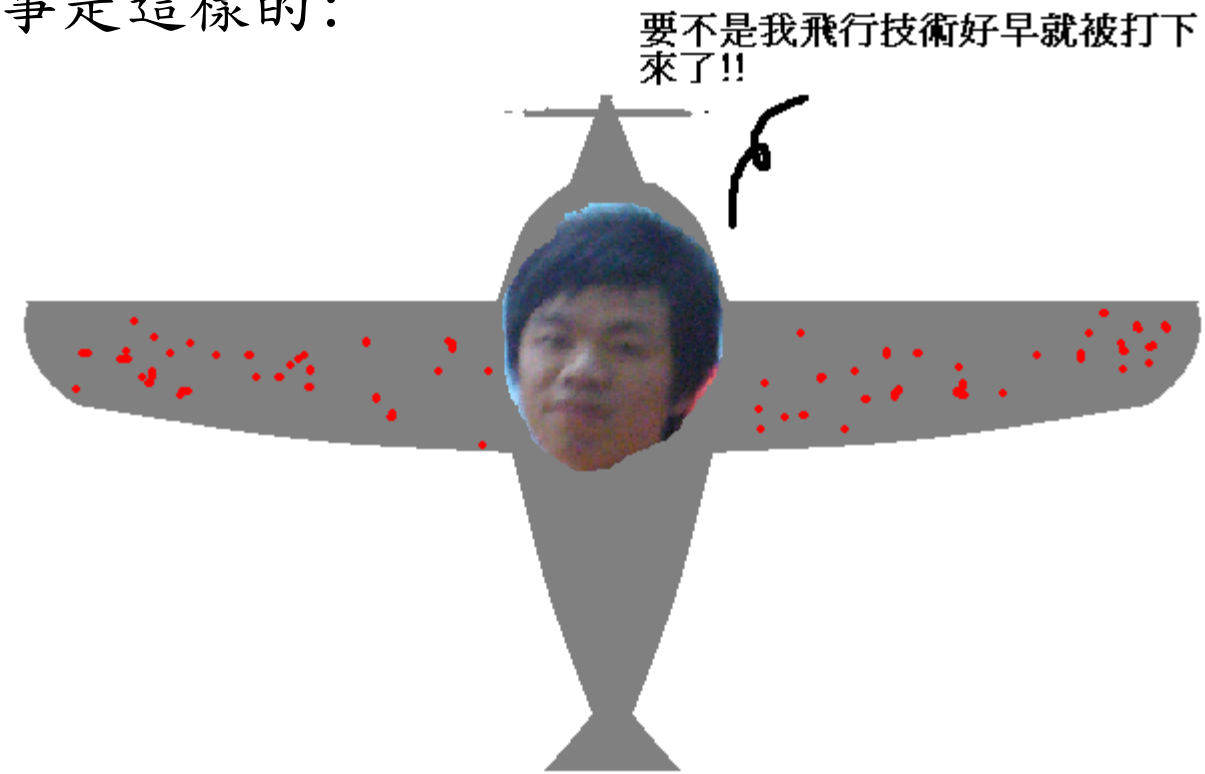
  這代表它治療感冒，快又有效?

  （感冒是病毒引起的）
- 為了瞭解種族歧視的情況是否有惡化，某人找了家信譽良好的民調公司對受訪者提問，問他們是否認為黑人找到工作的機會和白人一樣好。每隔一段時間重新調查，藉此了解種族歧視的問題是否變得更嚴重。

  (同情黑人者中有2/3說no；而歧視黑人者中有2/3說yes)
- 事情的真相常常不像表面上看到的那樣，民調結果尤其如此。

問題常常發生在看不見的地方

有一個故事是這樣的:



After you plot your data, think!

- The statistician Abraham Wald (1902-1950) invented some statistical methods that were military secrets during World War II. Here is one of his simpler ideas.
- Wald studied the location of enemy bullet holes in planes returning from combat. He plotted the locations on an outline of the plane. As data accumulated, most of the outline filled up.
- Put the armor in the few spots with no bullet holes, said Wald.
- That's where bullet hits the plane that didn't make it back.

似相關而非相關的數字（續）

- 一個很大樣本的知名醫師當中有 27% 抽的是利喉牌香菸，比任何其他牌子都多。
- 某型榨汁機宣稱自己能多榨出26%的果汁。
- 某年美國因火車而死亡的人數是4712人，這是否代表坐火車非常危險?
- 美西戰爭中海軍的死亡率是每千人中有9人，而在同時期紐約市的百姓中每千人有16人死亡。負責招募新兵的人就用這些數字來「證明」加入海軍比不加入海軍還安全。

似相關而非相關的數字（續）

- 在<<Harper's>>雜誌上有位讀者替A&P連鎖商店辯護，指出該商店的銷售淨利只有1.1%這麼低。並問：「會有任何美國公民……在每年投資的每1000元只賺10元的情況下，還擔心被社會大眾指控牟取暴利嗎?」
- 美國某新聞報導：1952年是美國醫學史上小兒麻痺最嚴重的一年，理由是因為該年的小兒麻痺病例創新高……但原因其實是?

6. 錯誤因果結論

- 調查發現吸菸者的大學成績比不吸菸者差。一定是抽菸把腦袋變鈍了！
- 調查發現3~6歲小孩的腳越大，閱讀能力就越強。所以小孩的腦袋應該長在腳上！
- 某樹林72%的烏鴉在松樹上築巢，因此可得結論為「烏鴉喜歡在松樹上築巢」。

  可是此樹林95%都是松樹！

6. 錯誤因果結論

- 當存在很多種合理的解釋時，你並沒有權利去選一個你喜歡的解釋，然後堅持它是對的。
- 很多時候，兩者間沒有因果，甚至互為因果。也有很多時候，兩者間強烈的關係是由於另外一個因素。
- 另外，有人在做結論的時候，會把根據數據得到的關聯性延伸(外插)到數據的範圍外去，這是要特別小心的。

關聯性 vs. 因果關係

- 假設有人根據調查證明了以下事情：

  高中畢業生的收入比中輟生多，而每多讀一年大學，收入就更多一些。

  所以我們得到了結論：書讀得越多，錢就賺越多？
- 研究發現，年長女性走路時兩腳掌之間的角度較大。

  所以外八造成了年長？還是年長造成了外八？

  都不是，真正的原因是：當年年長女性在成長年代被教導走路要腳尖向外；而年輕女性則否

關聯性 vs. 因果關係

- 如果我們會讓統計以及一堆數字和小數點擾亂了因果關係，那也沒比迷信好到哪去。

例: 島國萬那杜的島民曾經深信身上長蝨子會讓身體更健康，這是因為他們幾世紀以來觀察發現，健康的人身上通常有蝨子，而感冒發燒的人則沒有。

(Why?)

7. 如何對統計提出質疑

- 問題一：誰說的
- 問題二：他怎麼知道的
- 問題三：漏了什麼
- 問題四：是否有人改變了主題
- 問題五：這有道理嗎

問題一：誰說的

- 第一件該注意的事情，就是有沒有偏差存在。
- 實驗室為了支持一項理論、自己的名聲、或者因為收了費而必須證明某件事？報紙是否以寫出動人故事為目標？
- 要尋找蓄意的偏差，例如，做比較時，先用某一年做標準，而另一項比較卻換了標準。
- 不自覺的偏差更要注意，尤其在問卷或是民調。

問題一：誰說的(續)

- 某作者在他的文章中提到：

  康乃爾大學研究了1500位擁有學士學位的典型中年人，其中的男性有93%已婚，而全體中年男性的已婚比例是83%。但是中年女性大學畢業生當中，只有65%已婚，而全體中年女性的已婚比例是88%。

  最後他下了結論：「女性讀大學會妨礙結婚」。

- 雖然數據來自康大，但結論卻不是康大下的。

然而人們卻很可能因為康大的名聲而在腦袋留下「康乃爾大學說」的錯誤印象。

問題二：他怎麼知道的

- 樣本是否夠大？
- 樣本如何獲得？
- 樣本足以代表母體嗎？
- 得到的結論有統計上的顯著性嗎？
- 邏輯是否正確？

問題三：漏了什麼

- 有哪些數據被寫得很小？或是被刻意忽略了？
- 做比較時的標準是否一致、 公平？

例:為了爭取加薪，一個勞工組織曾經指出，在經濟大蕭條以後，利潤與產量的指數上升的比薪水指數快得多。

但這只不過是因為利潤才剛到達低點，所以計算百分比時用的分母較小。

例:曾有人公開過一些數字，指出該年四月份的營業額高過去年四月。

而他略過的事實是:前一年的復活節落在三月，而該年卻落在四月。

問題四：是否有人改變了主題

- 病例變多，不代表得這種病的人真的變多。

  例:近百年來，死於癌症的人數大增。

- 數據如果根據人們說什麼而得來的，就會出現許許多多的怪事。

例:英國的「他」比「她」更常洗澡。

例:中國某個區域曾經在統計後得知人口數是兩千八百萬，五年後，數字變成一億零五百萬。

問題五：這有道理嗎

- 某個對社會保險條例修正案的聽證會上，有人指出：「因為平均壽命差不多只有63歲，所以若要為65歲退休的人建立一套社會保險計畫，那根本就是騙局，因為幾乎每個人都還沒到65歲就死了。」
- 不加限制的外插法，常常會出現荒唐的結果。

例: 1947年到1952年，美國家庭的電視機數目增加了約10000%。把這樣的增長比例投射到接下去的五年裡，你會發現不久之後每個家庭有四十台電視。

例:由兒童的生長曲線用外插來推估30歲時的身高，則每個人都會是巨人。

總結

- Types of Lies 造成錯誤或誤導之原因
  - Intentional deceit 故意欺騙
  - Selective data use 選擇性使用資料
  - Extrapolation 外插
  - Creative graphics 在圖上作文章
  - Faulty assumptions 所使用的統計方法之假設不成立
  - Incompetence 統計能力太遜
- 害人之心不可有，防人之心不可無
- 建議大家要學防身術，才有能力自衛

參考資料

- "How to Lie with Statistics" by Darrell Huff (1954).
- 【別讓統計數字騙了你】Huff著、鄭惟厚譯(2005)、天下文化出版。
- "*The Basic Practice of Statistics*"（第五版）by David S. Moore (2010).
- Linda Tansil's powerpoint on "How to Lie with Statistics as in the book by Darrell Huff" (2003) from internet

謝謝大家！