繁亂的數字背後所呈現的真實面貌
–統計檢定p值的意義

台北榮民總醫院

郭英調

推論統計案例

- A病房過去每周約有一個院內感染個案，上週發生三個院內感染個案，要不要通報群突發?

原來/無差模式 Null Model

- ✓ 用機率模式 probability model解釋每周院內感染個案的個數變化
- ✓ 假設用Poisson distribution解釋每周院內感染個案的個數變化
- ✓ 本案用Poisson distribution每周一個來解釋其個數變化。

原來/無差模式 Null Model

- ✓ p值就是用Poisson distribution描述每周院內感染個案的個數變化時，發生三個以上院內感染個案的機率(0.08)
- ✓ 假設每周院內感染個案的發生率是每周一個的狀況(**原來模式**)。

代替/對立模式Alternative Model

- ✓ p 值太小時，要考慮其他的解釋方法(對立模式)。例如用群突發來解釋。
- ✓ 群突發的定義為每周有四個院內感染個案。

上周三個院內感染個案的解釋

- ✓原來模式 (沒有群突發): 看到三個院內感染個案的機率是 0.08 (p value).
- ✓代替模式(發生群突發): 看到三個院內感染個案的機率是 0.762 (power of study)

誤差的分類

	推論	
事實	未爆發流行 (無差模式) (Null model)	已爆發流行 (代替模式) (Alternative model)
未爆發流行	推論正確	第一類誤差(α)
已爆發流行	第二類誤差(β)	推論正確(檢力)

誤差的分類

	推論	
事實	無差異 (無差模式) (Null model)	有差異 (代替模式) (Alternative model)
無差異	推論正確	第一類誤差(α)
有差異	第二類誤差(β)	推論正確(檢力)

第一類誤差
(Type Ⅰ error)

- 所下的結論為「有效」時，錯誤的機率是多少。
- 結果為有效時，事實上這樣的差異是可以用「機率」來解釋。
- 也叫做阿爾發誤差（α error)或偽陽性誤差(false positive error)。
- 研究論文中常見到的 p值。

第二類誤差

(Type II error)

- 若研究結果為無效時，這樣的結果，可能事實上是有效，而因運氣不好(機率)的關係，發生無效的研究結果。
- 結論為「無效」時，錯誤的機率。
- 也叫做貝它誤差(β error)或偽陰性誤差(false negative error)
- 研究論文中常以檢力(power)呈現。
- Power= 1- β

誤差的分類

	推論	
事實	無差異 (無差模式) (Null model)	有差異 (代替模式) (Alternative model)
無差異	推論正確	第一類誤差(α)
有差異	第二類誤差(β)	推論正確(檢力)

P值的意義

- 在沒有相關性的假設下，獲得此研究結果之相關性的機率。

檢定方向

Direction of test

- 計算p 值的兩種方式
- 單尾檢定 vs 雙尾檢定

雙尾檢定(two-tailed test)

- 所考慮的情況是無差模式是否與代替模式「相同」時，稱為雙尾檢定(two-tailed test)或雙邊檢定(two-sided test)。
- 要同時考慮大於和小於無差模式的兩種情形(兩個方向)。
- 較單尾檢定保守(p值較大)

單尾檢定(one-tailed test)

- 所考慮的情形只有一個方向，只看發生率增加的情況，而不考慮減少的可能性，稱為單尾檢定 (one-tailed test)或單邊檢定(one-sided test)。
- 只考慮感染率增加，不考慮感染率降低的可能性。
- 較雙尾檢定大膽(p值較小)

單尾檢定 vs 雙尾檢定

- 若研究者的態度較保守時，採用雙尾檢定，計算出來的 p值較大。
- 若研究者有預設立場，大膽採用單尾檢定時，則會計算出較小的 p值。
- 不同的檢定方向會計算出不同的 p值。

# P<0.05

- 傳統上都是用p值是否小於0.05來判定是否有「統計上的意義」。
- 0.05其實是一種非常武斷的傳統。
- 一般相信，若發生機率小於二十分之一的事，不是偶然。

# P<0.05

- 連續拋一個銅板，若正面連續向上到第四次或第五次時(**P<0.05**)，大多數的人不再相信機率是唯一的因素，而較會相信是這個銅板有問題。
- 當發生機率(p值)小於0.05時，我們不接受原來的機率模式，作為解釋產生我們手上的這份資料的理由。

# P<0.05

- 將研究上複雜因素產生結果的機會，一律當作一般單純的機率問題來考慮，為免失之武斷。
- 若研究結果影響重大，p值可能要比0.05小很多才能取信於人。
- 若研究結果僅作參考，p值即使大一點也無所謂。

報告正確的p值

- 許多雜誌會要求論文作者寫出正確的p值，而讓讀者依對所探討的問題，對犯第一類誤差的忍受度，自行決定是否接受研究結論。
- 當p值小於0.001時，可以<0.001表示外，當寫出正確的p值至小數點下3位。

# P<0.05

- p值要多小才能接受，取決於錯誤的後果有多嚴重。
- Reasonably people might accept higher values or insist on lower ones depending on the consequences of a false decision in a given situation.

影響p值的因素

- 作用量(Effect Size)
  - 相關危險度、勝算比(類別資料)
  - 相關係數、平均值差(基數資料)
- 樣本數
- 統計方法

Sample Size

Small sample size prove nothing, large sample size prove anything.

研究結果的意義

- 研究結果的意義有統計意義和臨床意義兩方面。
- p值僅能表示研究結果的統計意義，無法表示其臨床意義。
- 信賴區間可表示其臨床意義，也可表示其統計意義。

統計意義

- 統計意義指是否能以機率來解釋所發現的相關性。
- 此研究結果是否適用在其他類似狀況？
- p值很小時，表示所發現的相關性，發生的機率太小，不能以機率來解釋，此相關性確實存在。

統計意義

- p值僅有統計上的意義，並不能呈現臨床上的重要性。研究結果的臨床意義，是無法由p值來判讀。
- 如在前述例子中，對於感染流行的嚴重程度，並無法由p值的大小來判斷。
- p值越小祇是我們對爆發流行的推論越有把握而已，並無法表示出流行率改變的大小。

臨床意義

- 臨床意義是指此結論對臨床作業有影響，臨床作業的改善幅度值得重視。
- 臨床意義所看重的，是其作用量(effect size)的大小。
- 作用量是指相關性的大小或是差異的程度。

Statistical vs. Clinical Significance

Example:

- Is supplement calcium reduce BP?
- 48 HT patients give 1000mg/day or not.
- Decrease SBP(5.6mmHg) and DBP(2.3mmHg) (Both $p<0.05$) after 8 weeks.
- May be an important clue to understanding the biology of hypertension.
- The magnitude of the reduction is not large enough for clinical HT treatment.

McCarron: Ann Intern Med 1985;103:825-831

信賴區間
(confidence interval)

- 在一定把握程度（即信賴程度）下，用一段區間來描述其範圍。
- 傳統上都是用 p值是否小於 0.05 來判定是否具有「統計意義」，因此一般以在95%信賴程度下的範圍來表示，因此稱為「九五信賴區間」。
- 樣本數越大，信賴區間越窄。

信賴區間

- 在信賴水準(95%)下，資料所在的範圍。
- 重覆做100次，有95次的結果會在此範圍內。
- 信賴水準越大，範圍越大。

信賴區間
(confidence interval)

- 使用口服降血糖藥後，血糖值降低平均為25mg/dl。九五信賴區間為 8 mg/dl 和 42 mg/dl。
- 我們有95%的把握，這個口服降血糖藥服用後的降糖效果，會在這個區間裡面。
- 也就是說，如果重複做了100次相同的研究，則其中有95次的結果會在 8 mg/dl和 42mg/dl的範圍之中。
- 8 mg/dl稱為其信賴下限(lower confidence limit)；42mg/dl 稱為其信賴上限(upper confidence limit)。

信賴區間
(confidence interval)

- 由於信賴區間中並沒有零出現，故不會產生無效的情形。
- 由信賴區間的範圍，我們便可得知這個藥的降血糖效果是確定的($p<0.05$)。
- 信賴區間不僅描述研究結果，由其區間的範圍，也同時描述了推論統計上的意義。
- 許多醫學雜誌要求論文作者將研究結果改用信賴區間的方式來表示，以便讓讀者清楚知道研究結果的顯著程度及其統計上的意義。

創新1號療效

創新1號　156/400 = 39%

傳統療法 128/400 = 32%

療效差距：7 % (0.4% to 14%)

p= 0.046

創新2號療效

創新2號　15/25 = 60%

傳統療法　7/25 = 28%

療效差距：32 % (6% to 58%)

p= 0.046

創新3號療效

創新3號　240/400 = 60%

傳統療法 112/400 = 28%

療效差距：32 % (25.5% to 38.5%)

p< 0.001

創新4號療效

創新4號　15/25 = 60%

傳統療法　9/25 = 36%

療效差距：24 % (-3% to 51%)

p= 0.157

創新系列療效

創新1號	7 %	(0.4% to 14%)	p= 0.046
創新2號	32 %	(6% to 58%)	p= 0.046
創新3號	32 %	(25.5% to 38.5%)	p <0.001
創新4號	24 %	(-3% to 51%)	p=0.157

P值的意義

- 在沒有相關性的假設下，獲得此研究結果之相關性的機率。