介紹

• CI（置信區間）提供了一種

估計我們的樣品是否良好

足以提供 真實的統計資料

實際人口。

• 在我們提到的許多例子中

早些時候，我們試圖找出平均值

樣本的值（大小有限）為

逼真的表示真值

人口值（在

大小）。

續

• 例如，平均值是否從

一個 n = 40 個樣本 離這裡不太遠

整個人口的真實？

• 有了這個n=40，我們可以估計一個範圍

即95%的抽樣手段

相同的大小將包含實際平均值。

n

s

x t 掃描電鏡 t

n

x 掃描電鏡

   

1.96 1.96 

（t 校正）

假設檢驗

• 另一種確定是否

您收集的樣品（有限的樣品）

大小）足夠好（代表

人口的真實統計）是做一個

hypothesis testing (假說檢定).

• 我們如何得出結論，一個

計算的平均值為

可接受作為實際

人口的平均值？

續

• 為了進行這樣的測試，我們從

聲稱人口均值等於

到一些假設的（假定的）值 0。

• 我們將有兩個假設：

– 一個稱為零假設，或

H0，

– 另一個稱為替代

假設，或HA。

續

• 例如，H0 ：  = 0 = 211 是

零假設，我們認為

計算的樣本均值為

與實際平均值“相同”。

• 第二種說法與

（矛盾） H0 ， 將是 HA ：   211.

• 兩種說法之一

必須為真。

續

• 如果出現以下情況，我們將否定原假設：

沒有足夠的證據

顯示計算的樣本均值

將與實際均值相同。

• 因此，這是一個“拒絕”或

不拒絕「空值 」。

假設。

實現

• 兩個語句（H0 和 HA）首先是

樺。

• 我們接下來繪製一個隨機的大小樣本

n 來自感興趣的人群。

• 我們計算此範例的平均值，

並稱之為X。

• 我們將樣本均值 X 與

假設均值 0

回想一下，我們希望看到「如果沒有足夠的證據

表明計算的樣本均值 X 將與

實際平均值 0。“

續

• 具體來說，我們想知道

這兩種手段之間的差異

（X 和 0） 太大而無法歸因

to chance alone (純屬巧合).

• 如果是這樣，則存在許多其他樣本

與實際均值相差甚遠的均值

（然後是 這個X）。

當前

意味著

我夠遠嗎？

續

• 另一種說法是看看是否

他們的差異不太可能很大。

• 在這些情況下，我們不會拒絕

零假設。也就是說，我們“認為”

呈現的樣本均值很好。

當前

意味著

我夠遠嗎？

簡要總結

• 要問的問題：差異是否

not likely to be big (差異大的機率很低),

或者我們可以很容易地找到其他樣品

意味著產生更顯著的差異

(容易找到差異更大的例子)?

• 答案是肯定的或否定的。

– 是：樣本均值 X 可以

表示實際均值（一個商品

樣本）。這意味著我們沒有

否定原假設。

（1）否定假設

• 假設 H0 為真，概率

獲得樣本意味著“極端”或

比「觀測值 X 更極端」。

（更 極端意味著離更遠

值 0 = 211） 足夠小。

• 如果 X 遠離 0 = 211 是一個糟糕的情況，

上面的陳述只是說它不太可能

我會得到一個比這更糟糕的估計。（不

太可能再得到這麼極端的平均值)

H0 ：  = 0 = 211

續

• 在這種情況下，數據不相容

使用原假設，我們認為

我們的樣本均值將等於

實際平均值。 所以我們 拒絕了這個假設。

• 結果現在會更多

支援房委會。也就是說，我們的採樣是

不太可能得到一個接近

實際平均值。

當前

意味著

我夠遠嗎？

這

當前

人口

均值為

這裡

我

樣本

均值為

這裡

有機會擁有

樣本均值 『作為極端

as“或”更極端”

此特定範例

平均值（下

右尾）。在這種情況下，

這個面積很小，

意味著一個人可以

幾乎找不到 其他 樣品

意味著比這更大

特定樣本 均值。

也就是說，這個

樣本均值為“遠”

從實際來看足夠了

卑鄙，我們會失敗

以滿足空值

假設。

標準正常

的分佈

採樣手段

（2） 不拒絕

假設

• 如果沒有足夠的證據來懷疑

原假設的有效性（“不是這樣-

例如，上一張幻燈片中的小尾巴），我們

無法拒絕此聲明。

• 但是，我們不說我們接受 H0。這

檢驗不能「證明」原假設。

• 實際平均值仍然很有可能

值不會是 211。例如，當

樣本量太小，我們可能不是

能夠輕鬆證明樣本均值遠

從實際平均值。

與正義的類比

System （司法系統）

法院

1. 從前提開始，這個人

是無辜的。（對面是

有罪。

2. 如果找到足夠的證據

超出合理懷疑的顯示

該人犯罪，

拒絕前提。（夠了

該人有罪的證據。

3. 如果發現的證據不足，

我們不拒絕這個前提。

（沒有足夠的證據

斷定一個人有罪。

統計假設檢驗

1. 從原假設開始。

（相反是另類

假設。

2. 如果大量

發現證據反駁（駁斥）

零假設，否定它。

（有足夠的證據得出結論

替代方案是正確的。

3. 如果發現的證據不足，

我們不拒絕空值。（不

有足夠的證據來反駁

空值。

小“如何足夠

小“？

• 在大多數情況下，0.05被認為是一個小

概率。[與95%的置信度相反。

• 因此，我們拒絕原假設，當

樣品可能具有的幾率

來自均值  = 0 的人口

= 211 小於或等於 5%。（該

樣品不顯著，或太

代表整體微不足道

人口。

續

• 這相當於說我們拒絕

錯誤地5%的時間。（我們是

小於5%的錯誤製作

這個電話。[回想一下，這被稱為“假”

負\*“。

• 檢視置信度的類比

95%?

\*在這裡，我們假設進行診斷的原假設。

成功拒絕一個應該被拒絕的是正確的

陰性。拒絕一個不應該被拒絕的人是錯誤的

陰性，依此類推。

顯著的測試水準

• 如果我們想，概率可以是0.01

更保守，或者我們想要0.1

不那麼保守。

• 這個值稱為有效水準

的測試，通常由

希臘字母 alpha （）  a pre-

確定是否剔除 H0 的值。

• 請注意，測試結果以這種方式處理

據說具有統計學意義。是的

不完全等同於臨床或

科學意義。

P 值

• 獲得均值的概率為

極端或更極端

觀測到的樣本均值 X，給定

原假設為真，稱為 p-

價值。（ 尾巴下方的區域）

• 如果 p 等於或小於

確定顯著性水準 ，我們拒絕

假設。

否則，我們不會否定假設。

（回想一下，我們不拒絕它不

意味著我們接受它。

示例 1

• 進行實驗以確定

擲硬幣是否公平（50%幾率

落在頭或尾上）或不公平

偏向，要麼朝向頭部（>50%

頭部著陸的機會）或朝向尾部

（<50%的幾率落地頭部）。（彎曲

硬幣產生有偏見的結果。

• 因為我們認為兩者都有偏見

替代方案，雙尾測試是

執行。

• 零假設是硬幣是

一般（50/50）。

續

• 假設實驗結果顯示

硬幣在20個總數中出現14次

翻轉。

• 此結果的 p 值將是機會

一枚公平的硬幣落在頭上至少14次

20次翻轉加上公平硬幣的機會

在 20 次翻轉中，有 14 次或更多次落在尾部\*。

• 在這種情況下，隨機變數 T 有一個

二項分佈。

\*我們問這枚硬幣是否公平。 所以 所有極端情況

（降落 在更多的頭上或降落在更多的尾巴上）必須考慮。這

導致雙面測試。稍後我們將看到另一種詢問

硬幣的公平性，導致片面的測試。

續

• 回想一下，在二項式分佈中（在

第7章）我們有公式

• 在這種情況下，n = 20，p = 0.5。 所以 對於

顯示14或更多

頭20個剪輯，我們想

計算 P（X=14）+ P（X=15）+

P（X=16）+...P（X=20）。

n x n x

P X x Cx p p

（  ）  （1 ） 

20 20

20

20

16

20

15

20

14

20 20

20

20 20 20 20

20

20 20

15

20 15 20 15

15

20 20

14

20 14 20 14

14

( ... ）0.5

（ 14） ...（ 20）

（ 20） 0.5 （1 0.5） 0.5

...

（ 15） 0.5 （1 0.5） 0.5

（ 14） 0.5 （1 0.5） 0.5

C C C C

P X P X

P X C C

P X C C

P X C C

    

   

   

   

   







溶液

• 從上面提供的網路計算機中，我們

有 C（20，14）+C（20，15）+...+C（20，20） =

38，760+15，504+4，845+1，140+190+20+1 =

60,460.

• 0.520 = 9.54107，我們有以下概率

在 20 次翻轉中獲得 14 個或更多的頭像

這兩個數位的倍數，或

60，4609.54107 = 0.0577。

• 通過對稱性，20 次翻轉的概率

硬幣將產生14個或更多的尾巴

（或者，6個或更少的頭）是相同的，

0.0577.

MATLAB 的 binopdf

>>binopdf（14，20，0.5）+雙魚（15，20，0.5）+雙魚（16，20

,0.5）+binopdf（17，20，0.5）+binopdf（18，20，0.5）+binopdf（

19，20，0.5）+黑白（20，20，0.5）

年 = 0.0577

>>

BINOPDF 二項式概率密度函數.

Y = BINOPDF（X，N，P） 傳回二項式

參數為 N 的概率密度函數

和 P 在 X 中的值處。請注意，密度

函數為零，除非 X 是整數。

MATLAB 的 binocdf

>>1-二元（13，20，0.5）

年 = 0.0577

>>

類似於其他累積密度函數

（CDF），可以很容易地使用BINOCDF

計算上一張幻燈片中的答案：

續

• 因此，硬幣的 p 值會上升

同一張臉（頭部或尾部）14

20 次總翻轉中的次數是

0.0577 + 0.0577 = 0.1154。

• 計算出的 p 值超過 0.05，

因此 ，觀察結果與

原假設 — 觀察到的結果

的 14 頭 20 翻轉可以是

僅僅歸因於機會 - 當它落下時

在將要發生的事情的範圍內

95%的時間。

續

• 在我們的示例中，我們無法拒絕 null

假設在5%的水準。

• 雖然硬幣沒有均勻地掉落，但

偏離預期結果只是

小到足以被報告為“不

在5%的水準上具有統計學意義」。

• 當然，就像我們之前說的，“不是

在 5% 時具有統計學意義

等級“ 並不一定意味著它是一個

公平的硬幣。

續

• 但是，有一個額外的頭

獲得，得到的 p 值（雙尾）

將是 0.0414 （4.14%）。[試試這個

你自己...]

• 這次是原假設 - 即

20次翻轉中15次頭部的觀察結果

可以歸因於偶然性 - 是

拒絕。

• 這一調查結果將被描述為

“在5%時具有統計學意義

級別“。

評論

• 雖然我們說過做假設

在顯著水準 0.05 的測試是

相當於估計95%的置信區間，它是

顯然，在這種情況下，一個人不能

真正計算這次擲硬幣中的CI

場景。（計算以前的置信區間

基於正態分佈。

• 這說明瞭使用的優勢

假設檢驗。

10.2 雙面測試

• 我們已經從上一個中看到

擲硬幣示例，以確定

硬幣是否公平。

• 在這種情況下，執行雙尾測試，

這意味著我們考慮

二項式下的兩個尾部區域

分佈為 p 值。

• 這正是雙面測試

方法。

Z 軸測試

• 假設連續隨機

變數 X 的均值 為 0 且已知

標準偏差 。

• 根據中心極限定理，

在規範化之後將給出

變數 Z 為標準正態分佈，如果

n 足夠大。

n

X

跟

/

0







續

• 我們現在可能考慮 Z 的值，即

作為極端或更極端

一個觀察到，比如說，一個平均值X，到

設置原假設並對其進行檢驗

就像我們在擲硬幣時所做的那樣

例。

• 因為它依賴於標準

正態分佈，這種檢驗

稱為 z 檢驗。

t 檢驗

• 當實際標準偏差（對於

一般人群）未知，我們

將替換樣本值

的標準差 s。這

或者給出以下公式。

• 如果基礎人群正常

分佈，隨機變數 t 有一個 t

自由度為 n1 的分佈。

s n

X

t

/

0



續

• 基於 t 上的 t 值

分佈（類似於 z 上的值

標準正態分佈，以及

回想一下，t分佈是一個“踩踏”

標準正態分佈），再次，我們

可以檢驗概率

獲得一個樣本均值，即更多

比觀察到的極端。

• 此過程稱為 t 檢驗。

示例 2

• 給予血清膽固醇分佈

美國成年男性的水準，他們是

高血壓和吸煙，  = 46。

• 要檢驗的原假設是

H0：  = 211，

其中 0 = 211 是所有 20- 到

74歲男性。

請注意，正在調查的受試者（高血壓和

吸煙）明顯不同於所有20至74歲的男性。

我們為什麼要建立這個零假設？

我們期待看到什麼？

示例 2 – 續

• 我們現在有一個12人的團體

平均x=217的高血壓吸煙者。

• 此範例是否可能意味著

（高血壓）將是任何

不同於 211 的平均值，從

所有成年男性？

– 如果答案是

是的？如果答案是否定的，那會是什麼？

• 這應該是 z 檢驗還是 t 檢驗？

單面還是雙面？

溶液

• 這應該是一個 z 檢驗，因為

STD人群被稱為46。

• 這應該是一個雙面測試，

因為遠大於 211 或

比211小得多

被視為“不同於

211".

溶液

0.45

46/ 12

217 211

/

0 



  

n

X

跟





Z = 0.45

陰影區域

右尾下方

將是 0.326。

這適用於

左尾小於

z=0.45 太。所以

作為的概率

極端或更多

極端，或 p-

值為 0.652。

Z = 0.45

>> 標準cdf（-0.45）

年 = 0.3264

>> 1-標準cdf（0.45）

年 = 0.3264

>>

續

• 所以 p 值為 0.625，這是很遠的

大於普通的0.05閾值。

• 因此，我們不拒絕空值

假設。

• 證據不足以得出結論

血清膽固醇水準

高血壓吸煙者群體

與 211 不同，後者是均值

從整個 20 歲到 -74 歲的美國

男性。[不能說區別是

重要！

續

• 換句話說，我們不能說，從

這個12人的抽樣和統計，

高血壓和吸煙

會導致血清升高

膽固醇水準。

• 如果我們提高樣本量呢？

提高到什麼大小會

結論不然？

與 CI 的類比

• 事實上，我們已經瞭解到該地區

當 z=±1.96 時，超過兩條尾巴為 5%。

• 也就是說，這在統計上經常使用0.05

顯著水準在數學上

相當於95%置信區間

從 z=1.96 到 z=1.96。

換句話說 ，假設檢驗和

CI估計基本相同

概念在判斷有多好

樣本統計數據與

人口統計。

續

• 按照前面的範例，95% 的

發現會議間隔為

217±1.96\*46/sqrt（12）=217±26=（191，

243).

• 所有患者的平均血清膽固醇水準

美國男性為211，在95%的置信區間內。

• 因此，我們不拒絕空值

假設這兩個平均值

是“接近”。

• 兩組膽固醇水平均值

具有可比性。

總結

• 雖然CI和假設檢驗都領先

我們得出同樣的結論，資訊

他們提供的有些不同。

• CI 提供一系列合理的值

平均值 ，並告訴我們一些事情

關於我們的點估計x̅中的不確定性。

• 假設檢驗有助於確定是否

假定的平均值 217 可能是

正確與否（至 211），具有特定的 p 值。

示例 3

• 給定兩個人群 - 所有嬰兒

接收含有鋁的抗酸劑，

和所有未接受抗酸劑的嬰兒。

• 他們的等離子鋁含量為

記錄。

• 所有接受抗酸劑的嬰兒都沒有

已知  和 。對於這些嬰兒

不接收抗酸劑，  = 4.13.

續

• 從所有嬰兒中選出10名兒童

接收抗酸劑，等離子鋁水準

x̅=37.20，s=7.13。

• 請注意，這兩種手段非常不同 （37.20

vs 4.13）。但是，STD=7.13，可能不是

小到足以證明這兩種手段都非常合理

不同。

• 我們樣本中的數據是否有可能

來自 =4.13 的人口？也就是說，我們

想要檢驗原假設

H0 ：  = 0 = 4.13

續

• 雙面測試，=0.05。

• 使用 t 檢驗而不是 z 檢驗。

• t=2.2622，得到t-中兩條尾巴的5%

分佈 df=10-1=9。我們的計算

t=14.67 正好進入右尾

結束。

• 預計 p 值非常小。

14.67

7.13/ 10

37.2 4.13

/

0 



  

s n

x

t 

>> 錫（0.975，9）

年 = 2.2622

>>

續

• 計算的 p 值小於

規定 =0.05。

• 因此，我們拒絕原假設。

• 即「平均等離子鋁水準」

接受抗酸劑的兒童“不是

等於「平均等離子鋁”

沒有接受的兒童水準

他們“。

• 服用抗酸劑會導致不同的

兒童等離子鋁水準。

>> 2\*（1-1-14.67， 9））

年 = 1.3683e-007

>>

評論

• 通常在典型應用中， 那些

不接收抗酸劑屬於”

Group” (控制組or 對照組).

• 服用抗酸劑的人被稱為

“Experiment Group”. (實驗組)

• 我們通常希望拒絕空值

假設（即

兩組參數

興趣），對於打擊的意義

這樣的實驗。

10-3 單側測試

• 類似於從估算中學到的知識

置信區間，這裏我們也有

假設的片面檢驗。

• 一個這樣的例子是測試

一枚硬幣是否偏向於

磁頭 – 只有大量磁頭

將很重要。

示例 4

• 投擲硬幣5次，全部落在

頭。

• 這是否足以說明硬幣有偏見

顯示更多頭部？

• 從統計學上講，我們問這是否

硬幣偏向於頭部的水準

=0.05？

• 原假設是

硬幣不偏向頭部（注意

它被允許偏向於

不過，尾巴）。

溶液

• 擁有5或的概率是多少

折騰5次后更多的頭 ？

• 此 p 值顯著（小於 0.05）

否定原假設。

>>雙魚（5，5，0.5）

年 =

0.0313

>>

評論

• 有趣的是， binopdf（4，4，0.5） 會有

值 0.0625，大於

0.05.

• 這意味著我們不能說硬幣是

偏向頭部，即使我們把它扔了 4

時間，都落在頭上。

?

10.4 錯誤類型

• 假設發生時出現兩種類型的錯誤

進行測試。

• I 型錯誤（或拒絕錯誤，或  錯誤）

• II 類錯誤（或接受錯誤，或  錯誤）

 = P（拒絕 H0 |H0 是真的）

 = P（不剔除 H0 |H0 為假）

續

• 顯然，I 型錯誤等效於

前面提到的“假陰性（FN）”

（ 如果您 考慮拒絕空值

假設為負數），並且II型錯誤

確實是“誤報（FP）」。。

結果

測試次數

假設

 = 0  ≠ 0

不要

正確剔除（TP 或 1-）

類型 II 錯誤

（FP 或 ）

拒絕類型 I 錯誤

（FN 或 ） 正確（TN 或 1-）

10.5 電源

• 測試的功效是

當 H0 時否定原假設

確實是假的。

• 它實際上是“真負”的情況

（TN）”。 因此 ，我們可能會獲得測試的力量

作為 1。

• 功率也可以被認為是

特定研究的可能性

能夠檢測到與零值的偏差

hypothesis given that one exists. (可正確

偵測出差異的能力)

示例 5

• 考慮 20 至 20 至的血清膽固醇水準

74歲的美國男性。此分佈的 =46

與 =211。

• 假設我們不知道 =211。相反

我們有20至24歲的美國數據

男性，和他們的 =180。

• 由於老年人往往具有更高的水準

年輕人，我們期望  大於

20至74歲的美國男性為180。

• 也就是說，我們認為存在差異

在180和實際平均膽固醇之間

20 至 74 歲美國男性的水準。

續

• 我們想對空值進行片面測試

假設

反對替代假設

• 我們預計 H0 會被拒絕。（因為舊

人們往往有較高的膽固醇水準）。

• 如果我們應該拒絕，但未能拒絕，

然後我們有一個II型錯誤（或假

陽性）。

H0 ：  ≦ 180

HA ：  > 180

錯誤地認為老人會有可比性

膽固醇水準 比年輕男性高。

續

• 假設我們有 25 個樣本。

也就是說，n=25。

• 我們已經知道 =46。

• 這意味著，原假設將是

如果樣本均值大於

比 195.1.

195.1

1.645

46/ 5

180

/

0



    

x

x

n

x

跟



 Z=1.645 切一個

5% 右手-

側尾為95%

單側 CI。

>> x=[140：250];

>> y1=normpdf（x，180，46/5）;

>>圖（x，y1）

x=195.1

地區

尾巴下

為 0.05。

20-24 歲的美國男性

>>規範（0.95）

年 = 1.6449

>> 180+歲\*46/5

年 = 195.1327

>>

請注意使用 46/sqrt（25） 而不是簡單的 46。

這是均值的抽樣分佈

n=25。

簡要總結

• 已知有25名年輕的美國男性（20-

至24歲）有平均膽固醇

值為 180。

• 讓另一個人口具有均值

膽固醇水準“顯著”大於

180，上限確定為

195.1327.

• 換句話說，我們“相信”真正的

20至74歲人群的膽固醇水準

美國男性將大於

195.1327.

• 考慮以

=211，代表血清的真實值

20 至 74 歲美國人的膽固醇水準

男性。

• 這應該是正確的分佈，

它被移到前一個的右側

20 至 24 歲美國人的分佈情況

男性。

x=195.1

下

這個綠色的尾巴是

0.042\* （z=-1.73）。

\* 獲得的機會

樣本意味著更少

比 195.1 從知道

真實人口

均值為 211，而不是 180。這

是 II 型錯誤  = 0.042

（誤報 ）。

20- 至 24-

一歲

美國男性

20- 至 74-

一歲

美國男性

>> z=（年-211）/（46/5）

z = -1.7247

>> normcdf（z）

年 = 0.0423

>>

續

• 所以 =0.042 對於單側檢驗

原假設 H0 ：  ≦ 180.

• 因此，此測試的功率為1-

= 0.958。（超過80%被認為是好的。

• 也就是說，對於在 0.05 處進行的測試

顯著性水準和使用樣本

大小為 25，有 95.8% 的幾率

否定原假設 H0 ： ≦ 180.

回想一下 ， 是一個接受錯誤，我們甚至沒有拒絕 H0

如果 H0 為假。

討論

• 可以看到功率 （1-） 可以是

在以下情況下增加：

–  增加（例如，從 0.05 增加到 0.1）。這

有效地移動截止邊界（

從透水滑梯的橙色垂直線）到

左邊，從而減小 了  的值。[請參見

下一張幻燈片]

–  如果替代

人口（以211為中心的人口）是

進一步向右移動，因此 左尾 較少

切。

>>規範（0.90）

年 = 1.2816

>> 180+歲\*46/5

年 = 191.7903

此區域下

綠色尾巴為 0.0184\*

（z=-2.088）。

\* 獲得的機會

樣本意味著更少

比 191.8 從知道

真正的總體平均值為

211，而不是180。這是類型

II 錯誤  = 0.0184 （假

陽性）。

20- 至 24-

一歲

美國男性

20- 至 74-

一歲

美國男性

>> z=（年-211）/（46/5）

z = -2.0880

>> normcdf（z）

年 = 0.0184

>>

結論

• 此權衡 介於  和  之間

類似於在

敏感性和特異性。

（敏感性可能增加

特異性下降）

•  和  之間的餘額，

然而，這是一個微妙的。

帶回家練習

• 如果我降低水平怎麼辦

顯著性  = 0.01？

• 新的 會是什麼？

• 測試的力量是什麼？

20- 至 24-

一歲

美國男性

20- 至 74-

一歲

美國男性

最好是100

有罪的人應該

逃脫比那一個

無辜的人

應該受苦

本傑明·佛蘭克林