

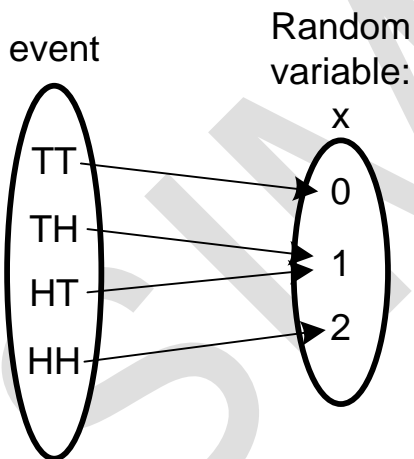
probability

前言

自然界中是沒有"機率"這個字眼的，我們會需要這個字眼是因為我們的瞭解太少了，例如：我們投擲一個銅板，我們很自然地就會說擲出正面和反面的機率都為0.5，但是如果我們能夠精準測出我們投擲出硬幣的力道和角度，也能掌控硬幣從離開指尖到落地的空氣阻力、重力、風力...等等的因素，我們其實是能100%知道投擲出來的結果的，不會有甚麼投擲出正反面的機率，但因為這些因素我們都無法掌握，所以我們才會有機率的發明，機率一詞換句話說，也就是用來包裝我們的無知的表象而已。

以下我們用投擲一枚公平的硬幣作為例子，並順便解釋各個名詞的意義

1. **trial(試驗)**: 投擲一枚硬幣
2. **outcome(試驗結果)**: head (H, 正面)、tail(T, 反面)
3. **U(字集)**: {H, T} // 為所有 outcome 所生成的集合，或稱為樣本空間
4. **event(事件)**: 為符合條件的字集(樣本空間)的子集合，舉例說，投擲一個骰子，擲出 4 點以上的事件為{4,5,6}
5. **random variable(隨機變數)**: 為一個 mapping function，方向是由 event 到 random variable
用投擲兩枚硬幣做為例子，我們定義 random variable 為投擲出正面的次數



由以上的定義，我們就能定義機率了

定義：

重複執行 trail 無限多次，可得到每個 outcome 的發生次數，可由各個 outcome 得到每個 random variable 的值，就是機率，但實際上我們沒辦法重複 trail 無限次，故通常我們會執行有限多，但足夠大量的 trail 來求得機率

e.g.

重複丟擲兩枚銅板 80 次，以下是這次實驗得到的數據

TT: 20 次 TH: 21 次 HT: 23 次 HH: 16 次

x	次數
0	20
1	21+23=44
2	16

$$P(x=0) = \frac{20}{80}$$

$$P(x=1) = \frac{44}{80}$$

$$P(x=2) = \frac{16}{80}$$

note: 所有 outcome 發生機率相加應為 1，故 $\sum_x P(x) = 1$ ，若 outcome 為連續分布，則 $\int_x P(x)dx = 1$

probability density function(pdf)

為描述 random variable 值的輸出值的 function

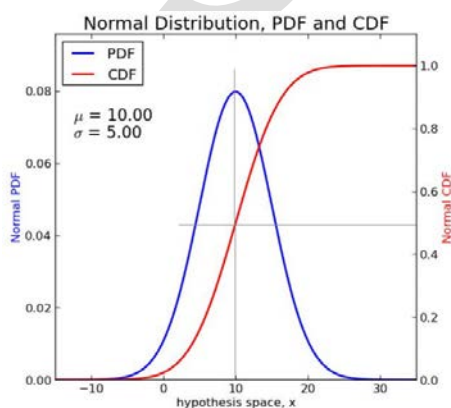
cumulative density function(cdf)

為到目前為止所有的 random variable 輸出值累計的 function

$$f(x_1) = \int_{-\infty}^{x_1} P(x)dx \quad // continuous$$

$$= \sum_{x=-\infty}^{x_1} P(x) \quad // discrete$$

e.g. (以連續函數為例)



在資訊領域來說，通常我們都是用 discrete 形式。

note:

在描述機率時，有時我們會將 random variable 省略

e.g. $P(x=H) \Rightarrow P(H)$

條件機率的問題(三門問題)

三個門中，有一扇門後面有車，另兩扇門後面有羊，選中車就算勝利。

而在選擇一扇門之後，主持人會從剩下兩扇門中，打開一扇後面為羊的門，再問你要不要更改原來的選擇，是換好還是不換好呢？

ans.

換比較好

我們將所有的可能都列出來

選擇/機率	開門後	是否更換/機率	結果	機率
車 1/3	1 羊 1 車	換 1/2	必為羊	1/6
		不換 1/2	必為車	1/6
羊 2/3		換 1/2	必為車	1/3
		不換 1/2	必為羊	1/3

是否更換選擇我們以 Y/N 代表，選擇到車/羊我們以 C/G 代表

若更換選擇，我們選到車的機率為 $P(C|Y) = \frac{\frac{1}{3}}{\frac{1}{6} + \frac{1}{3}} = \frac{2}{3}$

若不更換選擇，我們選到車的機率為 $P(C|N) = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{1}{3}$

參考網站(講的很清楚)：

<https://www.ptt.cc/bbs/Inference/M.1146115242.A.8AF.html>

也可以用這個觀點來看(我喜歡這種表示方式)

是否更換	選擇到/機率	開門後	目前選擇/其餘選擇	結果/機率	Joint probability
換	車 1/3	1 車 1 羊	車/1 羊	羊 1	1/3
				車 0	0
	羊 2/3		羊/1 車	羊 0	2/3
				車 1	0

不換	車 1/3		車/1 羊	車 1	1/3
	羊 2/3		羊/1 車	羊 1	2/3

如果換成 4 門 1 車

是否更換	選擇到/機率	開門後	目前選擇/其餘選擇	結果/機率	Joint probability
換	車 1/4	1 車 2 羊	車/2 羊	羊 1	1/4
				車 0	0
	羊 3/4		羊/1 車 1 羊	羊 1/2	3/8
				車 1/2	3/8
不換	車 1/4		車/2 羊	車 1	1/4
	羊 3/4		羊/1 車 1 羊	羊 1	3/4

如果更換選擇的話，選到車的機率為 $P(C|Y) = \frac{\frac{3}{8}}{\frac{1}{4} + \frac{3}{8} + \frac{3}{8}} = \frac{3}{8}$

若不更換選擇，我們選到車的機率為 $P(C|N) = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{3}{4}} = \frac{1}{4}$

有興趣的，可以去推推看 n 門 1 車的問題更換/不更換選擇時選到車的機率。

而如果你有去推 n 門 1 車的問題，你會發現無論有多少門，只要更換選擇，選到車的機率一定比沒有更換還要高，因為當我們一開始選擇時，我們的字集是四扇門。如果我們不更換選擇，我們的選擇就是等同於在原本的字集挑一扇門，但是如果我們更換了選擇，我們的 sample space 此時會縮小成三扇門，選到車的機率會上升。

note.(補充) 有些人可能覺得很奇怪，如果我不換，也有可能是我考量其他扇門後才選擇不換的，那我的 sample space 應該還是三扇門阿！但是注意，如果你真的是以 1/2 的機率在選門，然後選擇不換門，那 sample space 的確縮小了，機率也的確會上升。這時的問題應該是變成“我要不要在開了一扇門後再重新選擇一次”，那肯定的，如果重新選擇一次，一定是比“不重新選擇一次”的機率還高。但是這裡是直接就叫你選擇要不要換，你做了這個決定後，該決定的機率就變成 100%，這也是為何我把一開始表格中的“是否更換/機率”那格的機率拿掉，如果你是要比較是否會再選擇一次選到車的機率，那結果就換變成這樣

	是否更換	選擇到/機率	開門後	目前選擇/其餘選擇	結果/機率	Joint probability
會重新考慮要選哪扇門	換 1/2 (如果是 4 門 1 車，就是 2/3)	車 1/3	1 車 1 羊	車/1 羊	羊 1	1/6
		羊 2/3			車 0	0
				羊/1 車	羊 0	1/3
					車 1	0

	不換 1/2	車 1/3		車/1 羊	車 1	1/6
		羊 2/3		羊/1 車	羊 1	1/3
不會重新考慮要選哪扇門	x	車 1/3		車/1 羊	車 1	1/3
	x	羊 2/3		羊/1 車	羊 1	2/3

會重新考慮要選哪個門選到車的機率為 $\frac{\frac{1}{6} + \frac{1}{3}}{\frac{1}{6} + \frac{1}{3} + \frac{1}{6} + \frac{1}{3}} = \frac{1}{2}$ ，sample space 這時就真的變成 2 扇門選

1 扇門

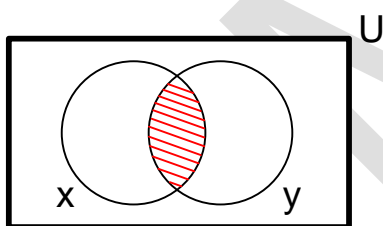
不會重新考慮要選哪扇門選到車的機率為 $\frac{\frac{1}{3}}{\frac{2}{3} + \frac{1}{3}} = \frac{1}{3}$ ，sample space 這時就真的變成 3 扇門選 1 扇

門

Joint probability

note: $P(x=1, y=2|U) \neq P(x=1|y=2)$

發生 x 又發生 y 的機率和發生 y 的前提下發生 x 的機率，意義上是不同的
以文氏圖來看就會一目了然



$$P(x, y|U) = \frac{\text{shaded intersection}}{\text{rectangle } U}$$

$$P(x|y) = \frac{\text{shaded intersection}}{\text{circle } y}$$

Baye's theorem

$$P(A, B) = P(A | B)P(B) = P(B | A)P(A)$$

$$\Rightarrow P(A | B) = P(B | A) \frac{P(A)}{P(B)}$$

貝氏定理告訴我們，即使我們不知道 $P(A | B)$ ，只要我們知道 $P(B | A)$ 、 $P(A)$ 、 $P(B)$ 就可以得知 $P(A | B)$

統計中的一些名詞解釋及由來

通常最基本的就是以下兩個名詞：

mean(平均值)

應該不需要贅述，就是所有資料總和的平均

$$E(x) = \frac{\sum_{i=1}^n x_i}{n} = \sum_{i=1}^n x_i P(x_i)$$

e.g. 擲一個公平骰子，平均擲一次骰子可得到

$$E(x) = \frac{1+2+3+4+5+6}{6} = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

而 $E(x) = \frac{\sum_{i=1}^n x_i}{n}$ 可成立的前提是各個 outcome 發生的機率皆相同，若不相同，只能使用

$$E(x) = \sum_{i=1}^n x_i P(x_i) \quad \text{note: 其實修正公式也可，但不重要就不贅述}$$

e.g. 擲一個不公平骰子，擲出 2 的機率是其他點數的機率的兩倍，則平均擲出一個骰子可得到

$$E(x) = 1 \times \frac{1}{7} + 2 \times \frac{2}{7} + 3 \times \frac{1}{7} + 4 \times \frac{1}{7} + 5 \times \frac{1}{7} + 6 \times \frac{1}{7} = 4.29$$

variance(標準差)

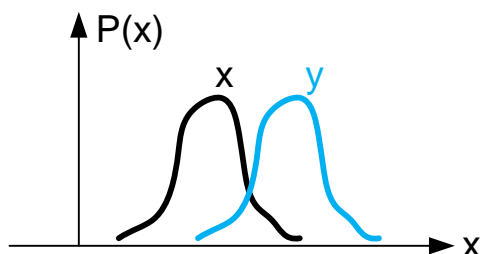
描述這些資料分布是集中還是分散

$$\text{定義: } \sigma^2 = \frac{1}{n} \sum_i (x_i - E(x))^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

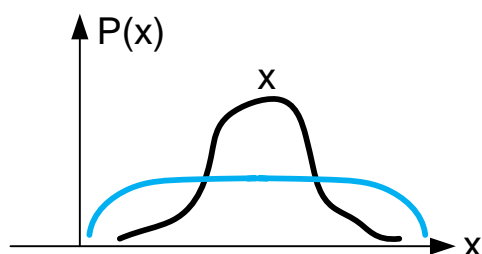
會採用平方的格式是因為平方可以微分，在許多數學性質上會有比較好的表現，是人為定義的。

mean 和 variance 對於描述一筆數據是不可或缺的，看下列圖就能明白了

1. 只用 variance 判別數據，可能造成相同 variance 但有不同的 mean，卻判斷為相同的資料:
e.g.



2. 只用 mean 判別數據



variance 公式推導：

$$\begin{aligned} E((x - \mu)^2) &= \frac{1}{n} \sum_i (x_i - \mu)^2 = \frac{1}{n} \sum (x^2 - 2x\mu + \mu^2) = \frac{1}{n} \sum x^2 - \frac{2}{n} \sum x\mu + \frac{1}{n} \sum \mu^2 \\ &= E(x^2) - 2\mu \frac{\sum x}{n} + \frac{1}{n} \mu^2 \cdot n = E(x^2) - 2\mu^2 + \mu^2 = E(x^2) - E^2(x) \end{aligned}$$

covariance(共變數)

為比較兩筆資料是否有相關性，若算出來的值為正，代表兩筆資料有正相關，若為負則代表負相關

$$\text{cov}(x, y) = E(x - E(x)) \cdot E(y - E(y))$$

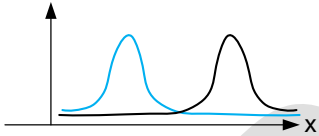
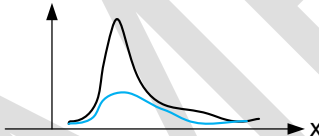
e.g. 我們想知道一個班上數學成績和國文成績是否有相關(是真的有理組腦，還是真的努力都可以考得好)

	數學	國文
student1	100	90
student2	40	60

student3	70	45
student4	65	80
student5	80	75
student6	75	60
mean	71.67	68.33
variance	322.22	222.22
covariance	866.67	

代表努力還是能讓成績變好的！

除了 mean 和 location 外，仍有其他變數會拿來評估數據

影響 location	Expectation	$E(x) = \frac{1}{n} \sum_{i=1}^n x_i$	first moment
影響 shape	Variance	$E((x - \mu)^2)$	second moment
	Skewness 	$E((x - \mu)^3)$	third moment
	Kurtosis 	$E((x - \mu)^4)$	forth moment

兩種不同的統計流派：frequentist vs. Bayesian

參考網站：

<https://read01.com/zh-tw/o ygB20.html#WcdxckQ2vTw>

<http://ithelp.ithome.com.tw/articles/10186897>

<http://ithelp.ithome.com.tw/articles/10186898>

<http://ithelp.ithome.com.tw/articles/10186900>

frequentist

frequentist 流派相信現實生活中的參數必定是不變的，並不會因為我們現在得到的事件為何影響這些參數。通常我們只要有足夠大的數據，就可以逼近這個參數(幾周後會說明 **bias** 的 **variance** 和取樣數目的關係，這邊有個感覺就好)，就像是一個銅板我們擲了 100000 次，正反面的機率都會接近 0.5，**frequentist** 流派相信這個 0.5 是永恆不變的真理。

note: **frequentist** 給我的感覺就好像是命運論，會發生甚麼事情都是命中註定，不管我現在做了甚麼努力都無法改變這個命運。

Bayesian

會利用到我們現在的已知(**prior**)來做推論，之後再利用推論的結果，找出最適合的模型作為結論，每增加一筆資料，結論可能就會更改，所以在資料量足夠大時，其實也能夠得到不錯的結論，但和 **frequentist** 不同的是，若我們一開始就能有不錯的 **prior**，或是我們能夠早點得到不錯的 **prior**，我們很早就可以得到不錯的結論，不需要大量的資料去 **train**，**bayesian** 流派認為任何參數都是有可能會改變的，例如車用久了就會容易壞、滴水也會穿石，任何參數都是有可能隨著環境、時間而改變，但是 **frequentist** 會認為這只是我們掌握的資訊不足而已，只要我們能掌握足夠關鍵的因素，這些問題也都可以模擬的出來，但是 **bayesian** 就順其自然，不考慮這些因素也沒關係，我只要與時俱進就好，**frequentist** 是超級固執的個性，**bayesian** 是隨波逐流個性(但也可調整要多隨波逐流)。所以，在資訊不足、掌握到的關鍵因素不足時，使用 **bayesian** 會得到就好的結果，因為 **bayesian** 能夠較貼近現在的環境、狀況。

note: **bayesian** 給我的感覺就好像是命運後天決定論，命運是由自己創造自己決定，並不是絕對的。

舉個小例子來說明此兩個不同流派的觀點：

我們想要知道一枚銅板是不是公平的

以 **frequentist** 的作法，假設我們重複投擲一枚銅板 2 次，得到 HH(兩次正面)

我們想要得到“哪種銅板是最可能發生 HH 這種情況”，但是因為我們使用 **frequentist**，銅板擲出正面這件事現在不是一個隨機變量，是已經確定的參數，只是我們不知道而已。所以我們不能使用“機率”來定義，而是採用描述參數的“**likelihood**”，簡寫為 **L**，也可以想成“從一個假設推出此現象發生的機率”

而 **likelihood** 的定義是 $L(\theta | \text{observed event}) = P(\text{observed event} | \theta)$

e.g.(補充，不用看也無傷大雅)

我們做一個擲兩次銅板的試驗

$$P(HH | P = 1) = 1$$

$$P(HH | P = 0.7) = 0.49$$

$$P(HH | P = 0.5) = 0.25$$

$$P(HH | P = 0.1) = 0.01$$

...

如果我們想知道在觀察到這個 event 時，背後參數為 $P=x$ 時的機率時，當然，我們可以將所有的可能加總後做 **normalize**，這又是另外一件事。後面會再提到，不過，就算我們要做 **normalize**，

$\frac{P(\theta_k | event)}{\sum_i P(\theta_i | event)}$ ，其分母都是相同的，只是對每個 $P(\theta_k | event)$ 做 **scaling** 而已，不影響彼此的比例關係，所以其實如果不代表“機率”這個概念的話，不做 **normalize** 是無傷大雅的。雖然有時候，我們可以藉由積分的方式做 **normalize**，像這裡我們可以藉由 $\int_0^1 P(\theta_k = x | event) dx = \int_0^1 x^2 dx = \frac{1}{3}$ 得到分母，但是更多時候 $P(\theta_k | event)$ 是很複雜是無法積分的。

所以 **likelihood** 就是直接取沒有 **normalize** 前的版本，也就是上述的公式

$$L(observed\ event | \theta) = P(\theta | observed\ event)$$

$$L(P = 1 | HH) = P(HH | P = 1) = 1$$

$$L(P = 0.7 | HH) = P(HH | P = 0.7) = 0.49$$

$$L(P = 0.5 | HH) = P(HH | P = 0.5) = 0.25$$

$$L(P = 0.1 | HH) = P(HH | P = 0.1) = 0.01$$

...

note:

因為 **likelihood** 並不是機率，故沒有遵從“機率加總=1”這個鐵則，在連續的分布底下，也沒有 $likelihood \leq 1$ 這件事，相信你們接下來就會看到這件事了

當然，我們沒辦法將所有的銅板的 **likelihood** 都算出來，所以其實這個問題在某些情況下是很難解決的。但在這個 **case** 下，我們是能確定最大的 **likelihood** 在 $P=1$ 的情形下，以 **frequentist** 的觀點來看，銅板擲出正面的機率只會是一個定值，不可能有其他種可能，故我們直接推論這枚銅板正面機率為 1，沒有其他正面機率的可能。

以 **bayesian** 的作法：

其實我們的大腦應該是 **bayesian** 的大腦，如果不相信，可以試試看等下的例子中你的思路是不是和我打的是一樣的，以丟擲銅板為例，如果是平常的我們，我們應該都會猜測銅板應該是公平的，但是我們仍不抹滅他們可能是不公平的銅板的可能，當我們擲出越多次連續的正面後，我們就會漸漸猜測銅板是不公平的，當在擲出更多次正面後，我們會在心裡增加這個銅板擲出正面的機率，若是真的太誇張，例如我們連續擲出了 1000 個正面，我們就會大膽假設，這個銅板只會擲出正面。所以若是我們使用 **bayesian** 來想推測一個事實發生的機率，我們會先有 **prior**，即事前我們會先猜測 **pdf** 分布。在由此分布來計算發生此事件後，**pdf** 分布應該會變成怎樣才比較合理。

e.g. 以我剛剛提出的例子為例，假設我們連續擲出 100 次正面，在這之間 bayesian 是如何變化的

假設我們的 prior 是正面擲出機率為 0.5 的機率為 0.7，，其他正面機率如下

$$P(\theta = 0.5) = 0.7, P(\theta = 0) = 0.05, P(\theta = 0.3) = 0.1,$$

$$P(0.7) = 0.1, P(\theta = 1) = 0.05$$

θ 為擲出正面的機率，也可稱為參數

我們也會需要用到這個式子的：

$$P(\theta | event) = \frac{P(event | \theta)P(\theta)}{P(event)}, P(event) = P(event, event_1) + P(event, not event_1)$$

$P(event | \theta)$ 我們稱為 likelihood probability，可以理解吧！likelihood 的想法是“從一假設推出此現象的機率”，要銘記在心！

$P(\theta)$ 取決於我們從小到大的經驗公式，例如我們認為銅板都應該是公平的

$P(event)$ 稱為 marginal probability

舉個例子應該就懂了

	20 歲以下	20-40 歲	40-60 歲	60 歲以上
胖	0.05	0.1	0.1	0.2
瘦	0.2	0.2	0.1	0.05

margin 就是不管其他事件發生與否，只針對我們有興趣的事件做計算

$$P(\text{胖}) = 0.05 + 0.1 + 0.1 + 0.2 = \sum_{\text{其餘因素}} P(\text{胖}, \text{其餘因素})$$

$$P(\text{瘦}) = 0.2 + 0.2 + 0.1 + 0.05 = \sum_{\text{其餘因素}} P(\text{瘦}, \text{其餘因素})$$

$P(\theta | event)$ 稱為 posterior probability，可想成我們建立的模型

至於這個式子

$$P(\theta | event) = \frac{P(event | \theta)P(\theta)}{P(event)}, P(event) = P(event, event_1) + P(event, not event_1)$$

$P(event | \theta)$ 是我們的 likelihood，我們認為這個參數 θ 多有可能會發生，就是我們的 prior $P(\theta)$ ，

那開始囉！

擲出第一次正面時，我們要算 $P(\theta | H)$ ：

$$P(\theta = 0, H) = P(H | P = 0)P(0) = 0$$

$$P(\theta = 0.3, H) = P(H | P = 0.3)P(0.3) = 0.3 \times 0.1 = 0.03$$

$$P(\theta = 0.5, H) = P(H | P = 0.5)P(0.5) = 0.5 \times 0.7 = 0.35$$

$$P(\theta = 0.7, H) = P(H | P = 0.7)P(0.7) = 0.7 \times 0.1 = 0.07$$

$$P(\theta = 1, H) = P(H | P = 1)P(1) = 1 \times 0.05 = 0.05$$

因為機率相加不為 1，我們修正一下各種 θ ，也就是 normalization

$$P(H) = P(\theta = 0, H) + P(\theta = 0.3 | H) + P(\theta = 0.5 | H) + P(\theta = 0.7 | H) + P(\theta = 1 | H)$$

$$= 0 + 0.03 + 0.35 + 0.07 + 0.05 = 0.5$$

$$P(\theta = 0 | H) = \frac{P(\theta = 0, H)}{P(H)} = 0$$

$$P(\theta = 0.3 | H) = \frac{P(\theta = 0.3, H)}{P(H)} = \frac{0.03}{0.5} = 0.06$$

$$P(\theta = 0.5 | H) = \frac{P(\theta = 0.5, H)}{P(H)} = \frac{0.35}{0.5} = 0.7$$

$$P(\theta = 0.7 | H) = \frac{P(\theta = 0.7, H)}{P(H)} = \frac{0.07}{0.5} = 0.14$$

$$P(\theta = 1 | H) = \frac{P(\theta = 1, H)}{P(H)} = \frac{0.05}{0.5} = 0.1$$

修正之後，我們再擲一次銅板，又是正面
我們整理如下

Θ	0	0.3	0.5	0.7	1
$P(\Theta H)$	0	0.018	0.35	0.098	0.1
修正過機率	0	0.032	0.618	0.173	0.177

若我們再擲完 10 次後再來計算：

Θ	0	0.3	0.5	0.7	1
$P(\Theta \text{HHHHHHHHHH})$	0	1.88E-07	6.04E-04	4.89E-03	1.77E-01
修正過機率	0	1.03E-06	0.003	0.027	0.970

可看到，隨著我們擲出越來越多的正面，我們猜測銅板為必定擲出正面的機率越來越上升。

baysian 還有個好處是，如果我們一開始就能猜到不錯的 Θ ，我們只需要少少筆數據就能夠達到不錯的猜測了。

Naïve Baye's classifier

我們用例子來說明：

假設我們想要觀察一個人是否會出門打網球，我們列出可能影響他會不會出門的因素有天氣(O)、氣溫(T)、濕度(H)和風力大小(W)，我們統整出來的結果如下

(圖片參考：<https://computersciencesource.wordpress.com/2010/01/28/year-2-machine-learning-naive->

bayes-classifier/)

Day	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

假設今天的狀況為 $O=\text{sunny}$, $T=\text{cool}$, $H=\text{high}$, $W=\text{strong}$

我們想求為

$$P(\text{play} = \text{yes} \mid O = s, T = c, H = h, W = s) = \frac{P(O = s, T = c, H = h, W = s \mid \text{play} = \text{yes})P(\text{play} = \text{yes})}{P(O = s, T = c, H = h, W = s)}$$

我們可經由上表得知 $P(\text{play} = \text{yes})$ ，但是我們無法得知 $P(O = s, T = c, H = h, W = s \mid \text{play} = \text{yes})$ 及 $P(O = s, T = c, H = h, W = s)$

我們無法得知 $P(O = s, T = c, H = h, W = s \mid \text{play} = \text{yes})$ 是因為我們的 data 太少且參數太多了，在這裡我們有 4 個參數，故至少要有 $3^4 = 81$ 個 case 才能涵蓋所有 case，但每種 case 的機率不同，故如果我們要觀察到所有的 case，我們需要有足夠多的 data 才能讓發生機率較低的事件都發生過，但有時我們很難做到。假設有一種 case 發生機率是 0.000001，理論上來說我們需要 1000000 筆資料才能看到這個 case。

故我們使用一個假設簡化此問題：假設各參數間都是獨立的(在已知某些隨機變數的情形下)，在這裡就是天氣、溫度、濕度、風力在已知是否打網球的情況下都是獨立的，雖然這個假設有些勉強，但事實上結果都還不錯，若我們的假設成立，我們可得到下列等式

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

$$\begin{aligned} P(A_2, A_3 \mid A_4) &= P(A_2 \mid A_4)P(A_3 \mid A_4) \\ \Rightarrow \frac{P(A_2, A_3, A_4)}{P(A_4)} &= \frac{P(A_2, A_4)}{P(A_4)} \frac{P(A_3, A_4)}{P(A_4)} \\ \Rightarrow \frac{P(A_2, A_3, A_4)}{P(A_3, A_4)} &= \frac{P(A_2, A_4)}{P(A_4)} \\ \Rightarrow P(A_2 \mid A_3, A_4) &= P(A_2 \mid A_4) \end{aligned}$$

$$\begin{aligned}
P(A_1, A_2 | A_3, A_4) &= P(A_1 | A_3, A_4)P(A_2 | A_3, A_4) \\
\Rightarrow \frac{P(A_1, A_2, A_3, A_4)}{P(A_3, A_4)} &= \frac{P(A_1, A_3, A_4)}{P(A_3, A_4)} \frac{P(A_2, A_3, A_4)}{P(A_3, A_4)} \\
\Rightarrow \frac{P(A_1, A_2, A_3, A_4)}{P(A_2, A_3, A_4)} &= \frac{P(A_1, A_3, A_4)}{P(A_3, A_4)} \\
\Rightarrow P(A_1 | A_2, A_3, A_4) &= P(A_1 | A_3, A_4)
\end{aligned}$$

根據上一段推導結果

$$\Rightarrow P(A_1 | A_2, A_3, A_4) = P(A_1 | A_4)$$

故

$$P(A_1, A_2, A_3 | A_4) = P(A_1 | A_2, A_3, A_4)P(A_2 | A_3, A_4)P(A_3 | A_4) = P(A_1 | A_4)P(A_2 | A_4)P(A_3 | A_4)$$

當 A 越多，以此類推下去

$$P(O = s, T = c, H = h, W = s | \text{play} = \text{yes}) = P(O = s | \text{play} = \text{yes})P(T = c | \text{play} = \text{yes})P(H = h | \text{play} = \text{yes})P(W = s | \text{play} = \text{yes})$$

雖然我們還是無法得知 $P(O = s, T = c, H = h, W = s)$ ，但其實這是不重要的項，因為我們只要知道 $P(\text{play} = \text{yes} | O = s, T = c, H = h, W = s)$ 和 $P(\text{play} = \text{no} | O = s, T = c, H = h, W = s)$ 的比值，因為兩者相加為 1，所以做完 normalize 後我們就可得到發生的機率 $P(\text{play} = \text{yes} | O = s, T = c, H = h, W = s)$ 講這麼多，實際算一次吧！

$$\begin{aligned}
P(\text{play} = \text{yes} | O = s, T = c, H = h, W = s) &= \frac{P(O = s, T = c, H = h, W = s | \text{play} = \text{yes})P(\text{play} = \text{yes})}{P(O = s, T = c, H = h, W = s)} \\
&= \frac{P(O = s | \text{play} = \text{yes})P(T = c | \text{play} = \text{yes})P(H = h | \text{play} = \text{yes})P(W = s | \text{play} = \text{yes})P(\text{play} = \text{yes})}{P(O = s, T = c, H = h, W = s)}
\end{aligned}$$

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{P(O = s, T = c, H = h, W = s)}$$

$$\begin{aligned}
P(\text{play} = \text{no} | O = s, T = c, H = h, W = s) &= \frac{P(O = s, T = c, H = h, W = s | \text{play} = \text{no})P(\text{play} = \text{no})}{P(O = s, T = c, H = h, W = s)} \\
&= \frac{P(O = s | \text{play} = \text{no})P(T = c | \text{play} = \text{no})P(H = h | \text{play} = \text{no})P(W = s | \text{play} = \text{no})P(\text{play} = \text{no})}{P(O = s, T = c, H = h, W = s)}
\end{aligned}$$

$$= \frac{\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}}{P(O = s, T = c, H = h, W = s)}$$

則

$$\begin{aligned}
\frac{P(\text{play} = \text{yes} | O = s, T = c, H = h, W = s)}{P(\text{play} = \text{no} | O = s, T = c, H = h, W = s)} &= \frac{\frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{P(O = s, T = c, H = h, W = s)}}{\frac{\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}}{P(O = s, T = c, H = h, W = s)}} = \frac{125}{486}
\end{aligned}$$

$$\Rightarrow P(\text{play} = \text{yes} \mid O = s, T = c, H = h, W = s) = \frac{125}{125+486} = 20\%$$

$$P(\text{play} = \text{no} \mid O = s, T = c, H = h, W = s) = \frac{486}{125+486} = 80\%$$

SIMPSON