Bernoulli distribution

當 sample space 只有兩種 outcome 時，(e.g. 擲一次銅板，只有正面和反面兩種可能、學期成績是否及格…)這樣的試驗稱為 Bernoulli trial，假設其隨機變數 X 只有 1 和 0 兩種，

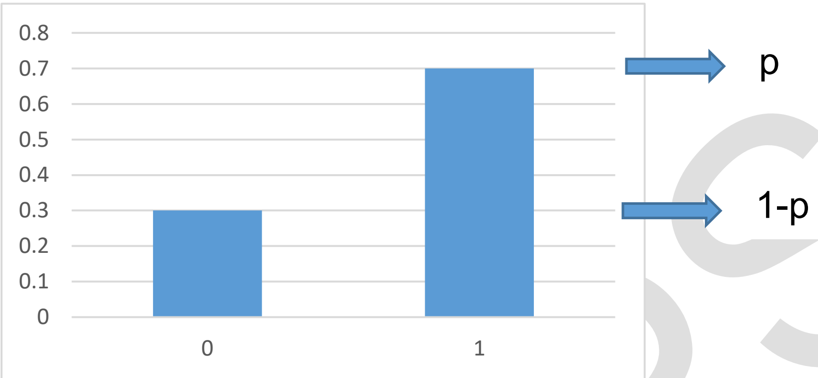
P(X=1)=p

P(X=0)=1-p

我們只須給定 p，重複試驗數次，觀察 X=1 的次數，這樣的分佈即為 Bernoulli distribution

e.g.

我們丟擲銅板 10 次，得到 7 個正面，其 Bernoulli distribution 如下圖



$$E(X)=1\cdot p+0\cdot(1-p)=p$$
$$Var(X)=E(X^2)-E^2(X)=(1^2\cdot p+0^2\cdot(1-p))-p^2=p-p^2=p(1-p)=pq \quad //q=1-p$$

假設執行 Bernoulli 試驗 N 次，第一次試驗結果為 X_1，第二次試驗結果為 X_1…

得到這樣結果 D 的機率為

$$P(D\,|\,\theta=p)=\prod_{i=1}^{N}p^{X_i}(1-p)^{1-X_i}$$

e.g. 擲銅板 N 次

假設結果為 $D=\{1,0,0,1,\ldots,0,1\}=\{X_1,X_1,\ldots X_N\}$

$\Rightarrow P(D\,|\,\theta=p)=p^{X_1}(1-p)^{(1-X_2)}(1-p)^{(1-X_3)}p^{X_4}\ldots(1-p)^{(1-X_{N-1})}p^{X_N}$

但這是在我們已知參數 $\theta=p$ 的前提下，但很多時候我們無法提前知道參數，(e.g. 在進賭場前，我們並不知道莊家在背後偷調的中獎機率為何)，當然，如果我們知道參數值，我們可以輕易得到目前情形的機率，但在參數不知道的前提下，對於目前發生的情形，我們只能試著找出”最可能發生這種狀況的參數”，例如我們投擲銅板 1000 次，得到 511 次銅板，如果我們完全客觀的來觀察這件事，我們會說擲出正面的機率是 0.511。回到原本的命題，若在執行試驗 N 次中，成功(X=1)的次

數是 k，那麼我們通常會猜成功的機率是 $\frac{k}{N}$ 。

　　但是，這樣的猜測真的是”最可能發生這種狀況”的參數嗎？以下我們就來證明其實我們的大腦還是很厲害的！

重複一次問題並寫成數學式，在事件 D 發生時，我們希望找出參數 $\theta = p$ 能使發生此事件的機率最大，即求

$$\arg\max_{p} \prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i}$$

也就是是求 $\prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i}$ 這個式子最大時的參數 p 為多少

然而，因為乘法的微分比較麻煩(如果有兩項相乘，就要做”前微後不微+前不微後微”)，若相乘項更多，會更複雜，我們試著將相乘項改為相加項，當然首選就是取 log 了

性質：

a > b　⇔　loga > logb

所以若 p_k 在所有的 p 中

$$\prod_{i=1}^{N} p_k^{X_i}(1-p_k)^{1-X_i} > \arg\max_{p} \prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i}\ ,\ for\ all\ \ p \neq p_k$$

$$\Updownarrow$$

$$\log(\prod_{i=1}^{N} p_k^{X_i}(1-p_k)^{1-X_i}) > \log(\arg\max_{p} \prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i})\ ,\ for\ all\ \ p \neq p_k$$

白話來說，若 p_k 有在所有 p 中最大的 $\log(\prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i})$ ，那麼這個 p_k 也是所有 $\prod_{i=1}^{N} p_k^{X_i}(1-p_k)^{1-X_i}$ 中最大的 p

所以我們就可以安心的取 log 了！

$$\log(\prod_{i=1}^{N} p^{X_i}(1-p)^{1-X_i}) = \sum_{i=1}^{N} \log(p^{X_i}(1-p)^{1-X_i}) = \sum_{i=1}^{N} \log(p^{X_i}) + \sum_{i=1}^{N} \log((1-p)^{1-X_i})$$

$$= \sum_{i=1}^{N} X_i \log(p) + \sum_{i=1}^{N} (1-X_i)\log(1-p)$$

我們想取這樣 p 函數的最大值，微分=0 的那一點即是

$$\frac{d}{dp}(\sum_{i=1}^{N} X_i \log(p) + \sum_{i=1}^{N} (1-X_i)\log(1-p)) = 0$$

$$= \sum_{i=1}^{N} X_i \frac{1}{p} - \sum_{i=1}^{N} (1 - X_i) \frac{1}{1-p} = 0$$

$$\Rightarrow \frac{\sum_{i=1}^{N} X_i}{p} = \frac{\sum_{i=1}^{N} (1 - X_i)}{1-p} = \frac{N - \sum_{i=1}^{N} X_i}{1-p}$$

$$\Rightarrow (1-p) \sum_{i=1}^{N} X_i = p(N - \sum_{i=1}^{N} X_i)$$

$$\Rightarrow p = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\text{成功次數}}{\text{總試驗次數}}$$

證明成功！

Binomial distribution

相較於 Bernoulli distribution，Binomial distribution 多給定了試驗次數 N 這個參數，其實推導都和 Bernoulli distribution 差不多，若成功次數總合為 m

$$\mathrm{P}(\mathrm{X}{=}m \mid p, N) = \binom{N}{m} p^m (1-p)^{(N-m)}$$

因為試驗 N 次，故期望值、變異數都是 Bernoulli distribution 的 N 倍

$E(X) = Np$

$Var(X) = Npq$

Maximum likelihood 仍為 $\frac{m}{N}$

Frequentist -> bayesian

之前提過這兩個的不同，frequentist 是看數據說話，例如，我們連續擲銅板三次，出現三個正面，以 frequentist 角度來看，他推得的參數 $\theta = p = 1$，完全不可能有擲出反面的機會，但是其實若是一個公平的銅板，擲出這樣的情形的機率為 1/8，其實還是有可能會發生的，再擲一次銅板，其實還是有可能會擲出反面的，但 frequentist 否定這種可能，由此可知，單純用 frequentist 去判斷一件事是不可靠的。但是用 bayesian 做又可能因為給一個太差的 prior 而使得結果離現實差很多。

這裡使用的方法是給予一個 distribution 給 prior，我們給予看似不太可能發生的是一個很低的機率，我們可以看到整體的分布而不是只有單點，我們再回來看這個式子

$$P(\theta \mid H) = \frac{P(event \mid \theta)P(\theta)}{P(event)}$$

我們給予一個 $P(\theta)$ 的 distribution 做為 prior，distribution 上的每點我們都可以算出其 likelihood，prior 和 likelihood 相乘後，再做 normalize 得到 posterous，lession3 筆記 p7~9 有例子可以參考，但是每次我們除了算 likelihood 外，還要 normalize，要算 marginal 很麻煩，我們希望有一種分布，可以表現出各式各樣的分布，同時給予此分布給 prior 後，可以直接用 prior 和給予的情形求出和 prior 有相同分布形式的 posterous，若有這種性質的分布，我們稱為 Conjugate。

Conjugate prior-posterior

我們會希望 prior 和 posterior 的分布是相同類型的，這樣的話我們可以將 posterior 直接視為 prior，在觀察到一個現象，繼續再做下一次 posterior 的預估，不需要再計算 marginal。

符合這樣形式的是 beta function，在講 beta function 前，需要先提到 gamma function

gamma function 定義如下：

$$\Gamma(x) = \int_0^\infty p^{x-1} e^{-p}\, dp$$

性質：

1.

$$\Gamma(x) = (x-1)\Gamma(x-1)$$

proof:

$$\Gamma(x) = \int_0^\infty p^{x-1} e^{-p}\, dp = -p^{x-1} e^{-p}\Big|_0^\infty + (x-1)\int_0^\infty p^{x-2} e^{-p}\, dp = (x-1)\int_0^\infty p^{x-2} e^{-p}\, dp = (x-1)\Gamma(x-1)$$

2.

$$\int_0^\infty p^{a-1}(1-p)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

proof:

$$\because \int_0^\infty \beta(p,a,b)\, dx = 1 = \int_0^\infty p^{a-1}(1-p)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^\infty p^{a-1}(1-p)^{b-1}\, dx$$

$$\Rightarrow \int_0^\infty p^{a-1}(1-p)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

代幾個值看看

$$\Gamma(1) = \int_0^\infty p^0 e^{-p}\, dp = -e^{-p}\Big|_0^\infty = 1$$

$$\Gamma(2) = \int_0^\infty p^1 e^{-p}\, dp = -p\, e^{-p}\Big|_0^\infty + \int_0^\infty e^{-p}\, dp = -e^{-p}\Big|_0^\infty = 1 = (2-1)\Gamma(1)$$

$$\Gamma(3) = 2\Gamma(2) = 2$$

故

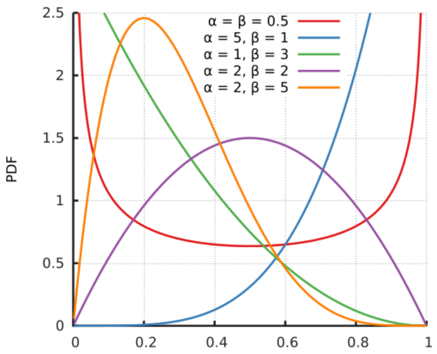
$$\Gamma(x)=\begin{cases}1 & ,x=1\,or\,2\\(x-1)! & ,otherwisw\end{cases}$$

而 beta function 是由 gamma function 定義而來

$$\beta(p\,|\,a,b)=p^{a-1}(1-p)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

如果仔細觀察可以發現，這樣的形式和 binomial distribution 很像，p 就是成功機率，a 是成功次數，b 是失敗次數，如果以 binomial distribution 顯示會是

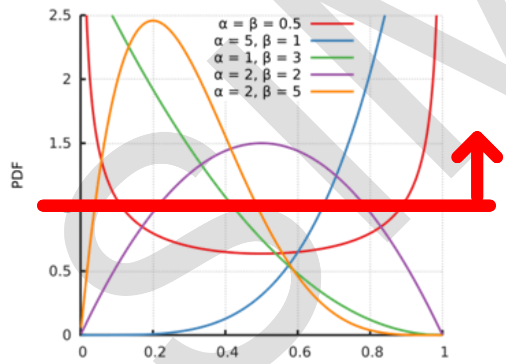
$$P(p\,|\,a,b=N-a)=p^{a-1}(1-p)^{b-1}\frac{(a+b)!}{a!b!}$$



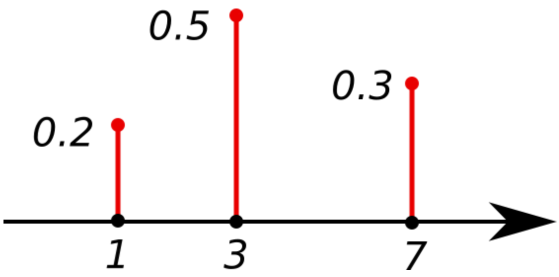
圖片來源：維基百科

note: 後記

剛念到這裡時沒甚麼注意就放他過去了，到了一兩個星期後我才發現 beta distribution 的 pdf 有很多地方都大於 1，發現是我觀念不清楚



這裡要注意的是不要和 PMF(probability mess function)搞混，PMF 是離散型的機率分布



$\sum_i p_i = 1$，故一定遵守每點機率均小於 1

但對於連續型的 PMF 就不是這回事，PDF 遵守 $\int_x p(x)dx = 1$，故可能有部分的值大於 1，但是機率加總仍為 1，我們沒辦法得到”某點”的機率值，我們只能得到”某個區間”發生的機率而已，例如，我們沒辦法得到 beta distribution(0.5)的值，但我們可以得到 beta distribution(0.4≤x≤0.5)的值

扯遠了，繼續吧！

$$E(X) = \int_0^\infty x \cdot x^{a-1}(1-x)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^\infty x^{(a+1)-1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} = \frac{a}{a+b}$$

$$Var(X) = E(X^2) - E^2(X)$$

$$E(X^2) = \int_0^\infty x^2 \cdot x^{a-1}(1-x)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^\infty x^{(a+2)-1}(1-x)^{b-1}dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+2+b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} = \frac{a(a+1)}{(a+b)(a+b+1)}$$

$$Var(X) = E(X^2) - E^2(X) = \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} = \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)}$$

$$= \frac{a((a^2+ab+a+b)-(a^2+ab+a))}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}$$

假設成功機率 p，我們的試驗是 binomial，試驗 N 次，成功次數 m，機率(likelihood)為

$\binom{N}{m} p^m (1-p)^{(N-m)}$，prior 假設為 beta function，試驗 a+b 次，成功次數 a，分布為

$$p^{a-1}(1-p)^{b-1} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

note: 不要讓 m 和 a,b 搞混了，m 是該次試驗成功的次數，a,b 視 prior 的成功、失敗次數，可想成是”在這之前成功了幾次、失敗了幾次”

其後驗機率(posterior)為 $\frac{\text{likelihood} \times \text{prior}}{\text{marginal}}$

$$P(\theta, event) = \frac{\text{likelihood} \times \text{prior}}{\text{marginal}} = \frac{\binom{N}{m} p^m (1-p)^{N-m} p^{a-1}(1-p)^{b-1} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}{\int_0^1 \binom{N}{m} \theta^m (1-\theta)^{N-m} \theta^{a-1}(1-\theta)^{b-1} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} d\theta} = \frac{p^{m+a-1}(1-p)^{N-m+b-1}}{\int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1} d\theta}$$

marginal 為所有參數θ=p'的 posterior 的加總

$$\int_0^1 \beta(\theta, m+a-1, N-m+b-1)d\theta = \int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1}\frac{\Gamma(a+N+b)}{\Gamma(m+a)\Gamma(N-m+b)}d\theta$$

$$=\frac{\Gamma(a+N+b)}{\Gamma(m+a)\Gamma(N-m+b)}\int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1}d\theta = 1$$

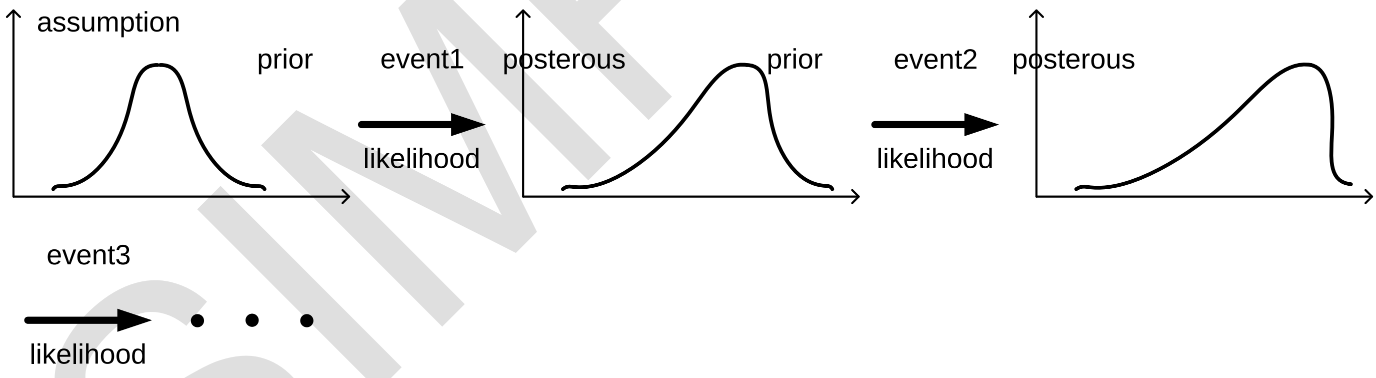
$$\Rightarrow \int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1}d\theta = \frac{\Gamma(m+a)\Gamma(N-m+b)}{\Gamma(a+N+b)}$$

$$\therefore P(\theta, event) = \frac{\text{likelihood}\times\text{prior}}{\text{marginal}} = \frac{p^{m+a-1}(1-p)^{N-m+b-1}}{\int_0^1 \theta^{m+a-1}(1-\theta)^{N-m+b-1}d\theta} = \frac{p^{m+a-1}(1-p)^{N-m+b-1}}{\frac{\Gamma(m+a)\Gamma(N-m+b)}{\Gamma(a+N+b)}}$$

$$=\frac{\Gamma(a+N+b)}{\Gamma(m+a)\Gamma(N-m+b)}p^{m+a-1}(1-p)^{N-m+b-1} = \beta(p, a+m, b+N-m)$$

得證，故我們可以對一件 binomial 試驗做出 beta distribution 的 prior，求 posterior 的過程很簡單，只需要在原來的 a 中加上 m-1，b 中加上 N-m-1，就馬上得到 posterior 的分佈，也是一個 beta distribution。。如果我們還有試驗繼續發生，我們可以將目前得到的 posterior 視為 prior，再求得此試驗後的 posterior…，這樣的過程我們稱為 online(sequential) learning。

例如目前我們已擲了 5 次正面，6 次反面，prior 為 $\beta(5,6)$ ，如果我們再擲三次銅板，為 1 次正面 2 次反面，posterior 為 $\beta(6,8)$ 。



事件一直來，beta function 就會一直更新

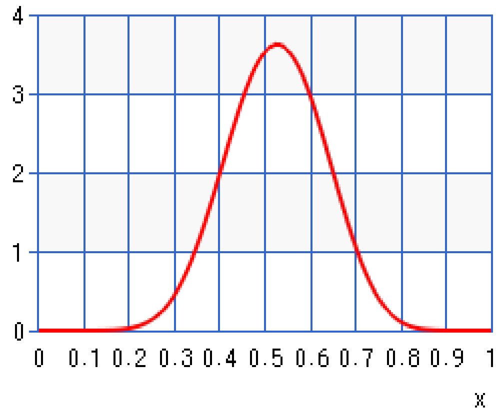
//後來寫作業後對於 prior, posterior, beta distribution 的感想

用個情境來想應該會比較能感受那些 a, b, a+m-1, b+N-m-1 在幹嘛，假設我們想知道手中的銅板擲出正面的機率 p 為何，我們的 prior 是假設 a, b 而不是 p，不知道會不會有人有疑惑，我假設 p=0.5 和 a=2, b=2 有不同嗎？當然有不同，將 p =0.5 是為 prior 的例子在 lession3 的筆記裡有，那麼 a=2, b=2 和 a=10, b=10 的 prior 是一樣的嗎？這裡的 prior 可以想成"在現在之前我們已經做過的試驗"，a=b=10 的 prior 當然比 a=b=2 的 prior 還要強，
假設我們再擲一次銅板得到正面，a=b=10 正面的 posterior MLE 為

$$\frac{總反面次數}{總正面次數+總反面次數} = \frac{10}{10+1+10} = \frac{10}{21} = 0.48$$

posterior 的 a 為 $a+m=10+1=11$, b 為 $a+N-m=10+1-1=10$

beta distribution 為 $\beta(11,10)$ 大概只會偏離中心一點而已



繪圖網站：http://keisan.casio.com/exec/system/1180573226

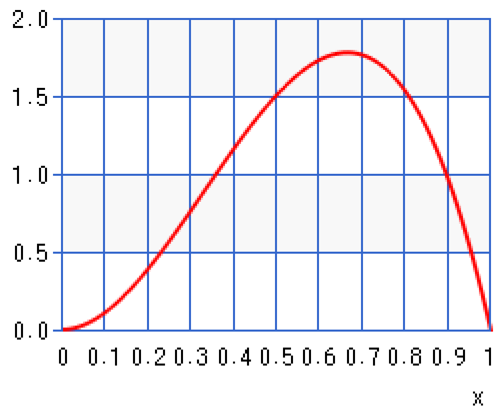
也可以想成，我們心中對於 p=0.5 這件事已經根深蒂固，很難因為一兩次試驗而改變我們的想法

而 a=b=2 正面的 posterior MLE 為

$$\frac{總反面次數}{總正面次數+總反面次數} = \frac{2}{2+1+2} = \frac{2}{5} = 0.4$$

posterior 的 a 為 $a+m=2+1=3$, b 為 $a+N-m=2+1-1=2$

beta distribution 為 $\beta(3,2)$ 就明顯和 $\beta(11,10)$ 差很多



變化：mutinomial

	binomial	multinomial
參數	m	m1,m2,m3,...

機率	$\binom{N}{m} p^m (1-p)^{(N-m)}$	$\binom{N}{m_1\ m_2\ m_3\ \ldots m_k} \prod_i p_i^{m_i}$
例子	銅板	骰子

和 binomial 的 beta distribution 對應的 distribution: Dirichlet distribution

$$Dir(a) = \frac{\Gamma(a_1 + a_2 + \ldots + a_k)}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)\ldots\Gamma(a_k)} \prod_i P_i^{a_k - 1}$$

仍有 conjugate 的關係