Probabilità delle probabilità

Saverio Napolitano

15 Aprile 2020

Supponiamo per un momento che tu possieda una fabbrica di auto, e prima di lanciare un nuovo modello decida di fare un piccolo controllo sulla qualità della produzione delle tue macchine. Decidi di effettuare un test con un gruppo di campioni limitato, diciamo 100. Supponiamo anche che la produzione di ogni macchina sia indipendente l'una dall'altra, ovvero che la qualità di una macchina non influisca su quella della successiva. Dopo un periodo di tempo raccogli i dati, e vedi che 2 macchine su 100 sono difettose, mentre il resto ha superato il test. L'approccio più ingenuo sarebbe quello di utilizzare le formule della probabilità classica, e dunque giungi alla conclusione che la probabilità che la fabbrica produca una macchina difettosa sia del 2%. Per confermare la tua ipotesi decidi di fare un altro test, questa volta con un insieme di campioni più ampio: 1000 macchine. Questa volta però le macchine difettose sono 30, e facendo i calcoli adesso la probabilità è del 3%. Dunque ti rendi conto che non è possibile conoscere la probabilità esatta, ma solamente un range di valori che potrebbero racchiudere il valore reale. Sarebbe però utile determinare il valore più prossimo possibile a quello vero, **il valore più probabile**. Grazie a questo esempio possiamo capire come non sempre nel mondo reale sia possibile avere delle probabilità certe, ma solamente delle *probabilità delle probabilità*, che sono un modo per analizzare i possibili valori più vicini a quell'ipotetico numero che stabilisce il rapporto tra due eventi binari (ovvero solo due possibilità). L'unico metodo per avvicinarci sempre di più a quel numero è aggiungere nuovi dati, e di conseguenza aggiornare la nostra ipotesi (attraverso per esempio il teorema di Bayes).

0.1 Distribuzione binomiale discreta

Facendo un esempio più classico, come quello della moneta, possiamo stabilire un modello probabilistico (stocastico) più facile da gestire matematicamente. Definiamo s la probabilità che lanciando una moneta esca testa. Per il momento stabiliamo che la moneta non è truccata: dunque $s = 0.5$, ovvero c'è il 50% di probabilità che lanciando la moneta esca testa. Definiamo n invece il numero di lanci che vogliamo eseguire, ad esempio 100. Infine definiamo k il numero di teste che sono uscite in seguito alla nostra sperimentazione. Utilizzando il linguaggio di programmazione `Python` possiamo realizzare algoritmi di simulazione e scalarli facilmente a numeri più grandi.

Supponiamo adesso di voler sapere quante volte (se ripetessimo l'esperimento dei 100 lanci) k equivale a 48 (nonostante il fatto che s equivalga a 0.5, è ragionevole pensare che il risultato non sarà solamente e sempre quello di 50 teste e 50 croci, conclusione tratta anche nel primo esempio). Consideriamo sempre le stesse condizioni di prima: il modello è **discreto** (non esite mezza testa o mezza croce, quindi k non può essere decimale), e ogni lancio è **indipendente l'uno dall'altro**.

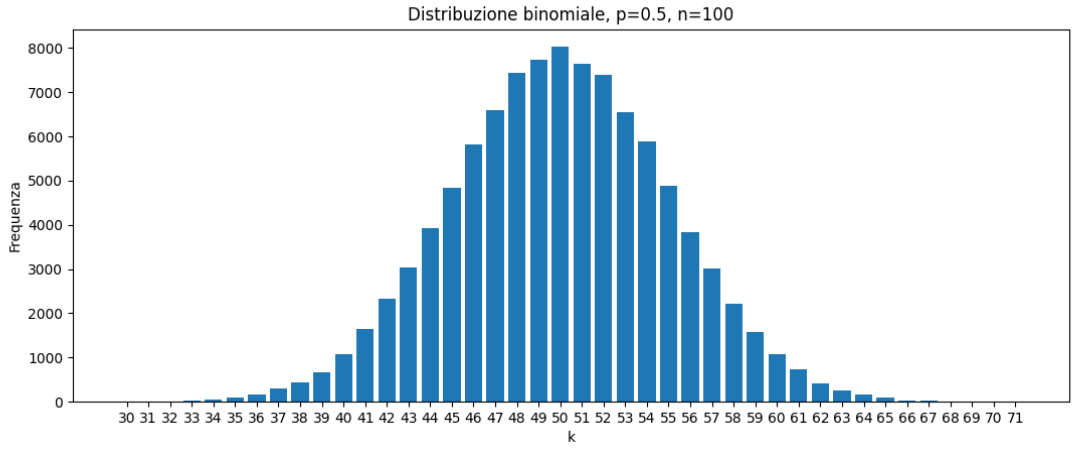


Figure 1: Insieme campione: 100000 simulazioni

Eseguita la simulazione, per calcolare la probabilità che k sia 48 basta prendere la frequenza del caso in cui escono 48 teste (7433) e dividerlo per il numero di simulazioni (100000). La probabilità risultante sarà $0,07433$, ovvero nel 7,4% dei casi.

Abbiamo visto come questo modello segue la **distribuzione binomiale**, e quindi per calcolare ogni probabilità di un evento con dati k, n e s, basterà utilizzare la formula apposita:

$$\binom{n}{k} s^k (1-s)^{n-k} \tag{1}$$

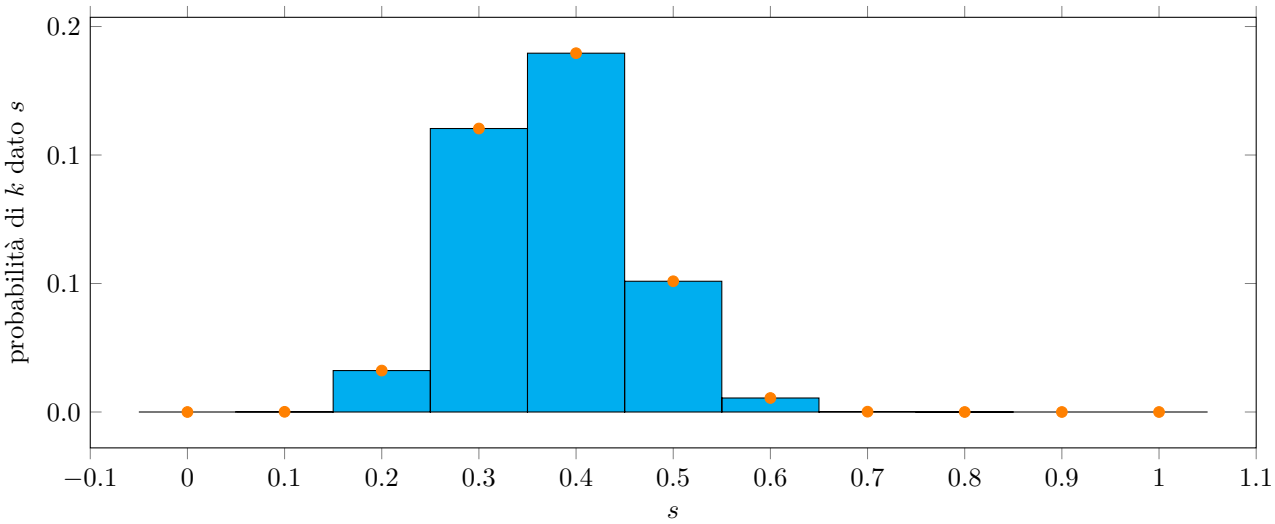
Questa formula può essere utilizzata anche con altri valori, ad esempio se volessimo computare la probabilità che su 30 lanci escano 10 teste, posso sostituire i dati nell'equazione (tenendo sempre conto però della moneta, che in questo caso non è truccata):

$$P(teste = 10) = \binom{30}{10} 0.5^{10}(1-0.5)^{20} = 0.02798160072416067 \tag{2}$$

0.2 Distribuzione binomiale continua

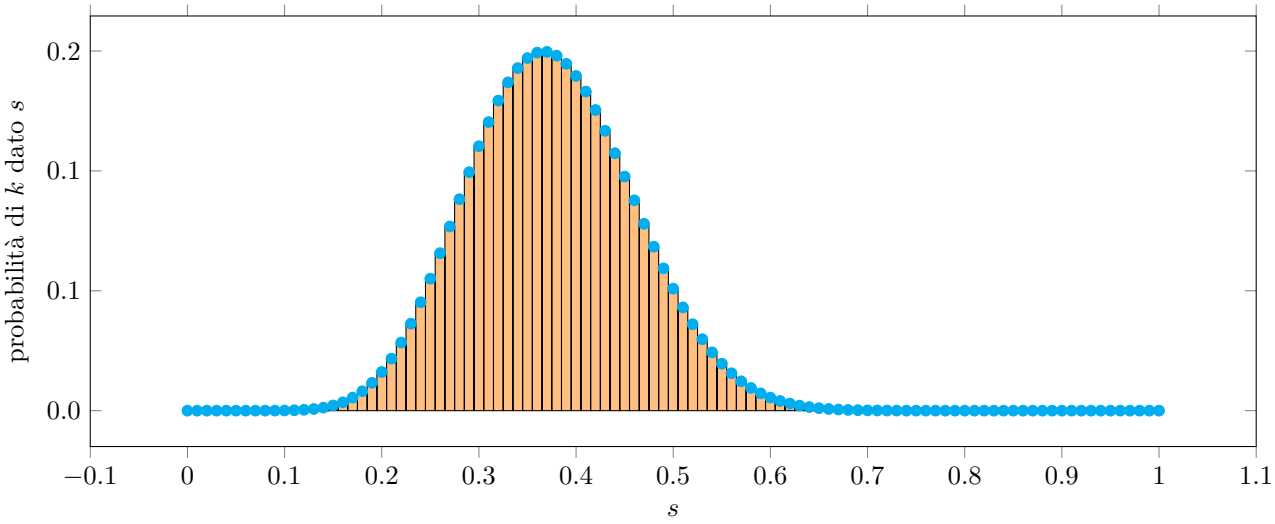
Prendiamo in esame un esempio più complicato (ma anche più vicino alla realtà). I valori n e k sono come sempre stabiliti perchè arbitrari, ma questa volta il valore s non è stabilito, ma diventa oggetto di indagine. Attraverso i dati ottenuti dai calcoli eseguiti con n e k si otterranno solo delle approssimazioni di s, e non un valore preciso. Consideriamo di nuovo l'esempio della moneta: in questo caso la moneta è truccata, e la percentuale s di ottenere testa è l'incognita. Possiamo solamente ricavare la previsione a priori, e poi aggiornarla con l'aggiunta di nuovi dati.

Il primo caso è proprio quello della moneta: $n = 30$ e $k = 11$. Ciò significa che ci sono 11 teste su 30. Come possiamo indovinare s? Un primo approccio potrebbe essere quello di dividere le possibili probabilità in intervalli di 0.1, e vedere con la formula della distribuzione binomiale quale è più probabile.



Procedendo con questo metodo possiamo ottenere una risposta approssimativa, ovvero quella più vicina al risultato: 0.4. Se però si riflette sul valore del numero, possiamo anche pensare ai numeri reali più vicini, ad esempio 0.40001 o 0.39999. Dunque per avere più precisione nelle nostre previsioni occorre rendere l'intervallo di ogni valore di s più piccolo.

In questo caso possiamo usare un intervallo di 0.01 per garantire una maggiore precisione.

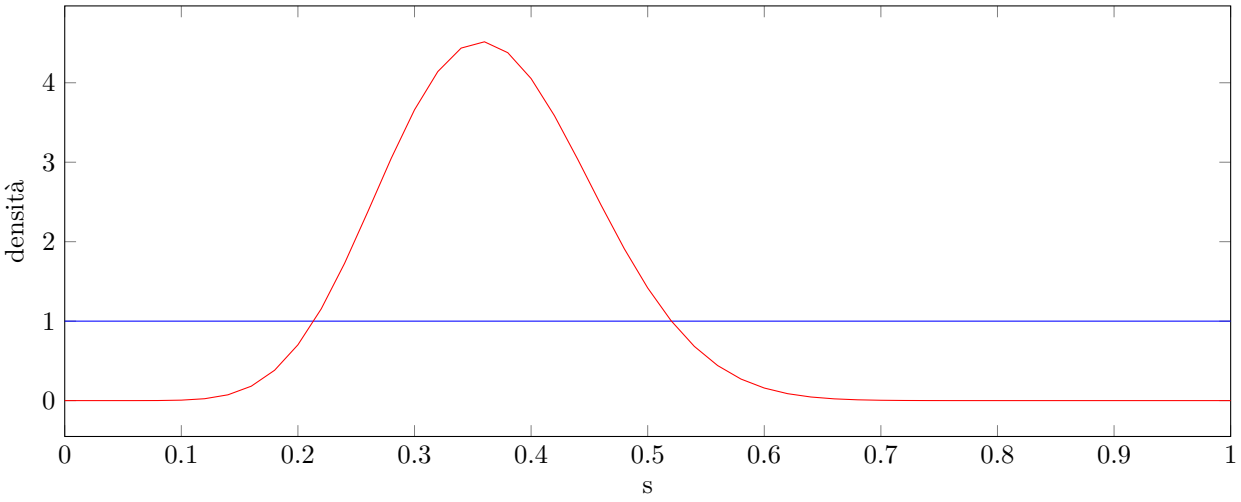


Grazie alla maggiore precisione possiamo meglio definire un possibile valore di s, che non è più 0.4, ma è un po' più piccolo. Dal grafico si può dedurre un valore di circa 0.367, ma se utilizziamo i dati n e k (che in questo caso possono essere utilizzati per calcolare l's approssimativo della distribuzione, in quanto abbiamo eseguito solo una simulazione), possiamo calcolare il valore preciso, che è $0.3\overline{6}$.

Di conseguenza s può essere un qualsiasi valore contenuto nell'insieme dei numeri reali, oltre ad essere compreso tra 0 e 1. Dunque il fatto che s sia un numero reale rende necessaria una sorta di distribuzione binomiale continua, ovvero senza interruzioni. La distribuzione che risolve questo problema è la *Distribuzione Beta* (detta anche *Funzione di densità della probabilità*), descritta dall'equazione:

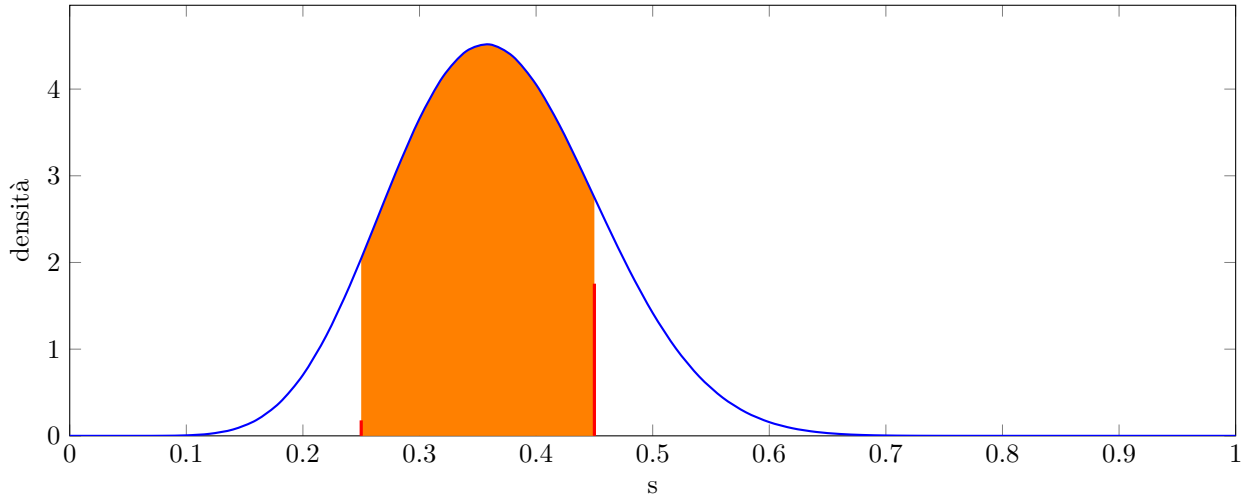
$$Beta(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{3}$$

Dove $\alpha = k$, $\beta = n - k$ e $B(\alpha, \beta)$ è una particolare funzione definita *Funzione Beta*.



Grazie a questo strumento matematico possiamo finalmente calcolare la probabilità che s abbia un certo valore. Prima però è importante porsi delle domande sul valore s stesso. La domanda da farsi è "può s assumere un valore preciso?" o più chiaramente, "Qual è la probabilità che s sia **esattamente** un valore x?".

Il buon senso ci dice che probabilmente non è molto alta: non può essere **esattamente** (ad esempio) 0.4 e non 0.399999 o 0.400001. Anche matematicamente ciò ha senso: possiamo pensare alla distribuzione beta come una distribuzione discreta con intervalli sempre più piccoli. Se la probabilità di un determinato valore s è rappresentata dall'area di una "barra", riducendo l'intervallo questa diventerà sempre più piccola, riducendo di conseguenza la probabilità che sia quel valore.
Ha più senso calcorare l'area di un determinato intervallo in cui s può cadere: per ottenere questo risultato basta quindi calcoare l'area sotto al grafico (definendo un intervallo). Ciò si può fare con gli integrali.



C'è da notare anche come l'area di tutto il grafico sia 1, mentre l'area di un valore preciso (che avrà un intervallo sulle ascisse infinitamente piccolo) sarà praticamente 0.
Per calcorare il valore s in quel intervallo basterà infine utilizzare gli integrali. Il valore (l'area sotto il grafico delimitato in quell'intervallo) sarà la probabilità che il valore s si trovi in quel range.

$$\int_{0.25}^{0.45} Beta(11, 19) \tag{4}$$