Generazione casuale uniforme di partizioni

Walter Mottinelli

2 marzo 2009

Sia dato un insieme, definito **supporto**, sul quale è stata definita una partizione. Come è noto, il numero di partizioni di un insieme costituito da n elementi è l'n-esimo numero di Bell (http://www.research.att.com/ njas/sequences/A000110):

n	b_n
0	1
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4140
9	21147
10	115975
11	678570
12	4213597
13	27644437
14	190899322
15	1382958545
16	1382958545
17	82864869804
18	682076806159
19	5832742205057
...	...
n	$\sum_{k=0}^{n-1} \binom{n-1}{k} b_k$

Poniamo ora di volere studiare una particolare proprietà delle partizioni. Al crescere della dimensione del supporto, aumenta il numero di possibili partizioni e quindi diventa sempre meno praticabile lo studio di questa proprietà su ogni singola partizione. In alternativa, si può usare un approccio statistico campionando nello spazio delle partizioni, analizzando quante partizioni del campione soddisfano la proprietà in esame e quindi

rapportando questo dato all'intero spazio delle partizioni.
Perché la stima sia non distorta il campionamento deve essere uniforme, quindi eseguito in modo tale che ogni partizione dello spazio abbia pari probabilità $(\frac{1}{b_n})$ di essere estratta.

Campionamento uniforme di partizioni: l'algoritmo RANEQU

L'algoritmo RANEQU, presentato in [NW78], permette di campionare uniformemente nello spazio delle partizioni di un dato insieme. Spieghiamo quali considerazioni conducono alla sua definizione.
Sia $\mathcal{P}_k$ una partizione data di $\{1, 2, \ldots, k\}$. Vogliamo estendere $\mathcal{P}_k$ a esattamente $\binom{n-1}{k}$ partizioni di $\{1, 2, \ldots, n\}$, quindi ci comportiamo come segue:

1. scegliamo un sottoinsieme S di k elementi da $\{1, 2, \ldots, n-1\}$;
2. etichettiamo i k elementi di $\mathcal{P}_k$ usando gli elementi di S;
3. affianchiamo alla partizione trovata tutti gli $n-k$ elementi rimanenti di $\{1, 2, \ldots, n\}$ in un'unica classe.

In questo modo otteniamo $\binom{n-1}{k}$ partizioni di $\{1, 2, \ldots, n\}$ da ciascuna delle b_k partizioni di $\{1, 2, \ldots, k\}$, o

$$\sum_{k=0}^{n-1} \binom{n-1}{k} b_k = b_n$$

partizioni di $\{1, 2, \ldots, n\}$.
Se dividiamo l'ultima relazione per b_n, otteniamo

$$\sum_{k=0}^{n-1} \binom{n-1}{k} \frac{b_k}{b_n} = 1$$

Possiamo riscriverla in maniera equivalente in questo modo:

$$\sum_{k=1}^{n} \binom{n-1}{k-1} \frac{b_{n-k}}{b_n} = 1$$

Diamo a questo risultato un significato probabilistico, e scriviamo quindi che l'argomento della sommatoria è la probabilità che un blocco abbia dimensione k in una qualsiasi partizione di un insieme di n elementi:

$$\mathrm{P}\,[k] = \binom{n-1}{k-1} \frac{b_{n-k}}{b_n}$$

Nota questa distribuzione di probabilità, possiamo quindi generare casualmente con probabilità uniforme una partizione in questo modo:

1. generiamo casualmente le dimensioni dei blocchi della partizione;

2. permutiamo gli elementi dell'insieme, così da redistribuirli casualmente nei blocchi di dimensione appena calcolata.

Ecco l'algoritmo completo:

Input: un insieme $\{1, \ldots, n\}$
Output: $(q_1, \ldots, q_n)$, dove q_i è il numero del blocco a cui l'elemento i appartiene

1: calcola e memorizza ogni b_i non ancora calcolato
2: $m \leftarrow n$
3: $l \leftarrow 0$
4: scegli k secondo la probabilità

$$\mathrm{P}[k] = \binom{n-1}{k-1} \frac{b_{n-k}}{b_n} \qquad (1 \leq k \leq m)$$

5: $l \leftarrow l + 1$
6: memorizza l in $q_{m-k+1}, \ldots, q_m$
7: $m \leftarrow m - k$
8: **if** $m > 0$ **then**
9: ritorna al passo 4.
10: **end if**
11: permuta casualmente $(q_1, \ldots, q_n)$

Verifica sperimentale: il test chi-quadro

Vogliamo ora verificare sperimentalmente il risultato promesso dell'algoritmo RANEQU: la partizione generata casualmente deve essere campionata con probabilità uniforme dallo spazio delle partizioni. Usiamo a questo scopo il test chi-quadro.
Consideriamo un insieme di 4 elementi, che avrà $b_4 = 15$ possibili partizioni. Usando l'algoritmo RANEQU, generiamo casualmente 750 partizioni di questo insieme e descriviamo i risultati dell'esperimento con le variabili casuali $X_1, \ldots, X_{750}$ a valori in $\{1, \ldots, 15\}$. Nella nostra ipotesi H_0 (**ipotesi nulla**) che il campionamento sia uniforme, supponiamo quindi che la variabile casuale X, rappresentante ogni X_j, abbia distribuzione uniforme

$$\mathrm{P}_{H_0}[X = i] = \frac{1}{15} \qquad i = 1, \ldots, 15$$

Per testare questa ipotesi, sia N_i, con $i = 1, \ldots, 15$, la variabile casuale che denota il numero di X_j uguali a i. Siccome ogni X_j è indipendente e identicamente distribuita, ne segue che N_i ha distribuzione binomiale con parametri $n = 750$ e $p = \frac{1}{15}$. Se la nostra ipotesi è valida, il suo valore atteso sarà quindi pari a

$$\mathrm{E}_{H_0}[N_i] = np_i = \frac{750}{15} = 50$$

Un primo indicatore della validità della nostra ipotesi è quindi la differenza tra la frequenza ottenuta N_i e quella attesa np_i, per ogni i, rapportata alla quantità np_i:

$$T = \sum_{i=1}^{15} \frac{(N_i - np_i)^2}{np_i}$$

Nel nostro esperimento avremo quindi

$$T = \sum_{i=1}^{15} \frac{(N_i - 50)^2}{50}$$

Se T sarà grande (cioè se i valori osservati si discosteranno troppo dai valori attesi), scarteremo l'ipotesi nulla H_0.
Nel nostro esperimento, osserviamo le seguenti frequenze N_i con $i = 1, \ldots, 15$:

$$[54, 37, 44, 63, 57, 55, 37, 55, 50, 51, 44, 51, 43, 57, 52]$$

Possiamo ora calcolare il valore di T (15.96), e verificare quanto inverosimilmente un valore così alto sarebbe stato ottenibile se la nostra ipotesi nulla fosse stata vera. Definiamo quindi la quantità p:

$$p = \mathrm{P}_{H_0}\left[T \geq 15.96\right]$$

Per valori di p bassi (inferiori a 0.05, o 0.01), si usa rifiutare l'ipotesi nulla, altrimenti si conclude che l'ipotesi sembra essere consistente con i dati sperimentali.
Per calcolare il valore di p, usiamo la seguente approssimazione: per grandi valori di n, T ha approssimativamente una distribuzione chi-quadro con $15 - 1$ gradi di libertà: $T \sim \chi^2_{14}$.
Calcoliamo di conseguenza il valore di p in questo modo:

$$\mathrm{P}_{H_0}\left[\chi^2_{14} \geq 15.96\right] = 1 - 0.684 = 0.316$$

Ne concludiamo che l'ipotesi che le frequenze siano estratte da una distribuzione binomiale, e che quindi le partizioni siano state estratte da una distribuzione uniforme, è consistente con i dati sperimentali.
Per verificare che non è stata una semplice coincidenza, ripetiamo l'esperimento 100 volte e otteniamo i seguenti valori (riordinati) della variabile casuale T:

3.24	5.44	5.6	5.64	5.8	6.08	6.28	6.28	6.56	7.08
7.16	7.16	7.8	7.8	8.0	8.48	8.6	8.68	8.8	8.88
9.16	9.52	9.84	9.92	9.96	10.36	10.76	10.96	10.96	11.0
11.08	11.24	11.32	11.48	11.56	11.72	11.76	11.88	11.92	12.08
12.72	12.84	12.96	13.0	13.04	13.4	13.52	13.52	13.72	13.72
13.8	13.84	13.88	14.0	14.32	14.4	14.84	14.84	14.88	15.0
15.4	15.6	15.68	15.8	15.84	16.08	16.08	16.2	16.24	16.28
16.4	16.48	16.56	16.56	16.64	16.64	16.68	16.68	16.96	17.0
17.04	17.68	17.72	18.04	18.28	18.52	19.12	19.4	19.52	19.56
19.92	20.0	21.12	21.16	21.76	24.2	24.32	24.52	24.6	24.96

Decidiamo di rifiutare l'ipotesi per valori di $p \leq 0.05$.
In un primo momento, decidiamo di individuare l'area critica del 5% nella coda destra della funzione di densità di probabilità di T: quindi rifiuteremo l'ipotesi nulla per valori

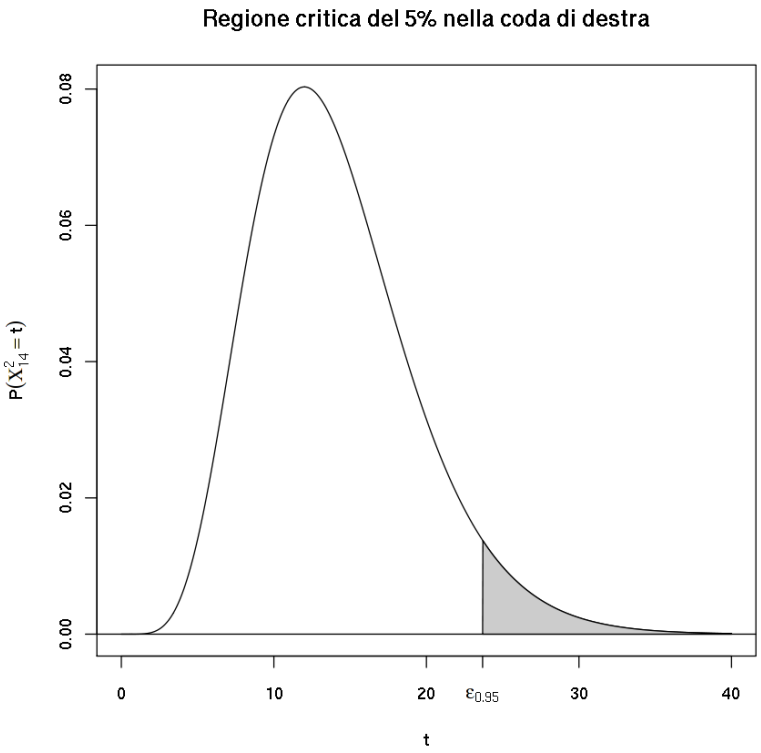

Figura 1: Distribuzione di probabilità di ${\chi^2}_{14}(t)$ con regione critica del 5% nella coda di destra

di $T \geq 23.7$, infatti il quantile 0.95-esimo di T ($\xi_{.95}$) è circa 23.7.

Nei 100 esperimenti eseguiti, l'ipotesi sarebbe stata rifiutata quindi 5 volte. Questo dato non è incoerente con quanto supposto: stabilendo un'area critica del 5%, accettiamo di ottenere un valore di $T \geq 23.7$ mediamente in 5 casi su 100.

Nelle procedure di verifica delle ipotesi si è soliti diffidare di valori di T troppo bassi, perché fanno nascere il sospetto che i risultati dell'esperimento siano stati "aggiustati" per supportare meglio l'ipotesi da verificare. Al fine di escludere questa eventualità, stabiliamo quindi che l'area critica del 5% sia ripartita equamente tra la coda di destra (valori di T alti) e la coda di sinistra (valori di T bassi).

Per soddisfare questa condizione, dovrà valere

$$\xi_{.025} < T < \xi_{0.975}$$

e quindi, approssimativamente,

$$5.63 < T < 26.12$$

Nei 100 esperimenti eseguiti, questa condizione è stata soddisfatta 97 volte.

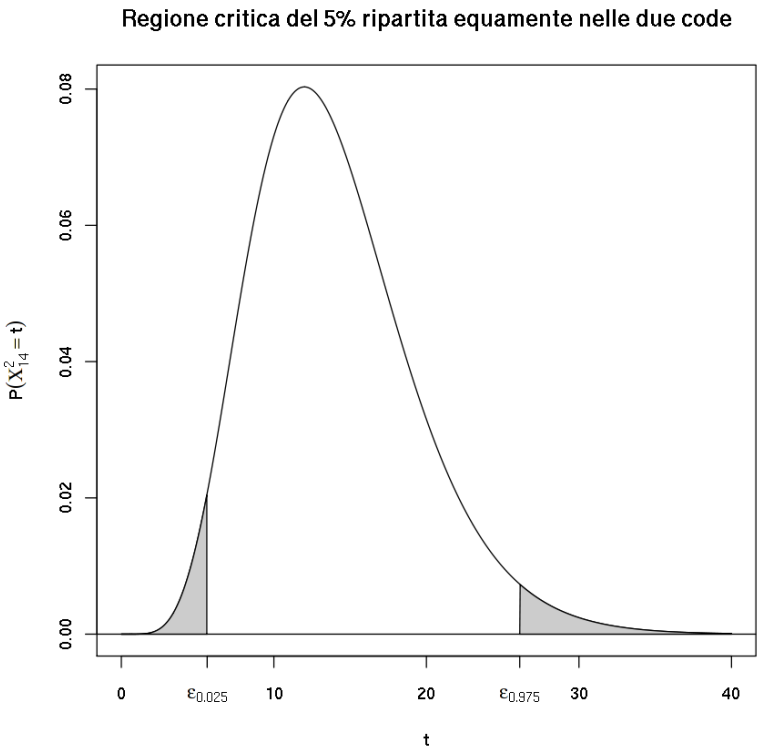

Figura 2: Distribuzione di probabilità di ${\chi^2}_{14}(t)$ con regione critica del 5% ripartita equamente nelle due code

Applicazione: stima del numero di partizioni regolari

Una partizione è definita **regolare** se tra i suoi blocchi è possibile stabilire un ordine, determinato dal poset stesso.
Poniamo di voler determinare quante sono le partizioni regolari di un poset.
Se il poset è costituito da un numero ridotto di elementi, è possibile generare tutte le sue partizioni e verificare quante di esse soddisfano la proprietà richiesta: il risultato ottenuto sarà esatto.
Nel caso di poset di dimensioni maggiori può invece essere impossibile, o comunque computazionalmente difficile, generare tutte le possibili partizioni. In alternativa, possiamo dare una stima del risultato a partire da un un campione rappresentativo estratto dallo spazio delle partizioni. La rappresentatività del campione è garantita dall'algoritmo RANEQU, che genera una partizione casuale con probabilità uniforme all'interno dello spazio delle partizioni.
Poniamo quindi di volere stimare il numero di partizioni regolari del poset avente le seguenti relazioni: $\{(a,b),(a,d),(d,c),(d,e),(f,g)\}$: delle $b_7 = 877$ partizioni possibili, 521 saranno regolari.

L'esperimento viene così definito: generiamo 100 partizioni casuali, verifichiamo quante

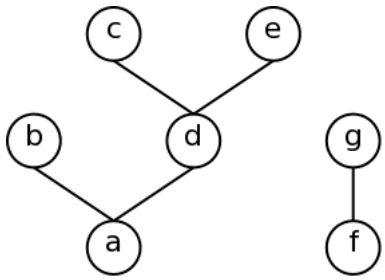

Figura 3: Poset $\{(a,b),(a,d),(d,c),(d,e),(f,g)\}$

di essere sono regolari ed infine proiettiamo il risultato sul totale delle partizioni.
Ripetiamo l'esperimento 100 volte, ottenendo quindi 100 stime delle quali calcoliamo media (520.938) e varianza (1269.06285).
Si noti che la scelta di utlizzare questa procedura per stimare il numero di partizioni regolari di un poset di 7 elementi è fatta solo a scopo esemplificativo. In termini di efficacia e di efficienza, per poset di dimensioni così ridotte non è certamente la soluzione migliore: abbiamo infatti generato complessivamente 10000 partizioni casuali, mentre evidentemente sarebbe stato più conveniente generare tutte le 877 partizioni distinte del poset, ed avremmo oltretutto ottenuto un risultato non più stimato ma esatto. A scopo esemplificativo della capacità di questa procedura di approssimare il risutato esatto, è invece interessante osservare come le stime assumano una distribuzione in media molto prossima al risultato esatto, con uno scarto quadratico medio di circa $\sqrt{1269} \approx 35.623$.
Ripetiamo ora l'esperimento altre 100 volte, lavorando però su un campione di 200 partizioni casuali: rispetto ai risultati precedenti la media non muta sensibilmente (520.7626), a differenza della varianza (778.40469574) che si riduce in maniera significativa, com'era lecito attendersi.
Ripetiamo infine l'esperimento 100 volte per dimensioni del campione sempre crescenti: 300, 400, 500, 600, 700, 800. Ovviamente la stima si riduce progressivamente quanto più il numero di partizioni campionate si avvicina al numero di partizioni totali (877):

dimensione del campione	media	varianza
100	520.938	1269.06285
200	520.7626	778.40469574
300	520.119466667	346.90794416
400	521.836925	281.837227232
500	519.78036	113.04473451
600	522.29735	81.6154828942
700	521.802471429	50.4955163614
800	521.7382625	18.0806604967

Mostriamo in un grafico tutti i risultati degli esperimenti condotti, e per riferimento anche il valore del risultato esatto (521) quando il campionamento estrae tutte le 877 partizioni dello spazio:

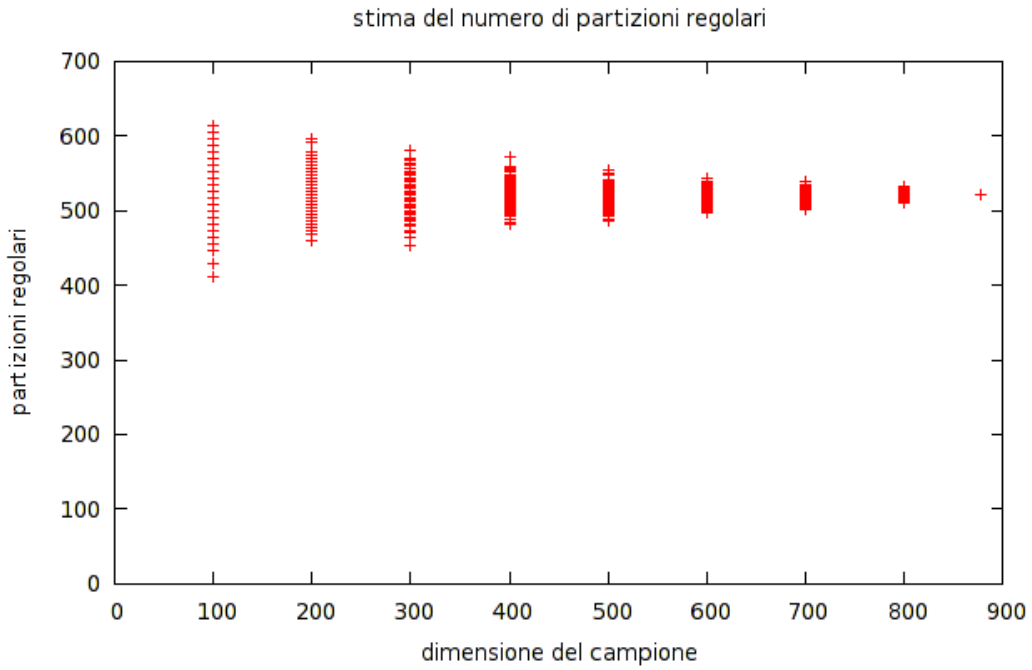

Figura 4: Stime del numero di partizioni regolari del poset $\{(a,b),(a,d),(d,c),(d,e),(f,g)\}$ in funzione dell'ampiezza del campione casuale

Riferimenti bibliografici

[NW78] A. Nijenhuis and H.S. Wilf. *Combinatorial algorithms*. Academic Press, ii ed. edition, 1978.