Esercizi proposti per il 7 e l'8 aprile

1. Consideriamo il seguente modello. Supponiamo di avere un alfabeto di n lettere, e di scrivere a caso una parola scegliendo k lettere a caso, con reimmissione delle lettere estratte. In alcuni punti occorre pensare che tali lettere siano ordinate da 1 a n. Si calcoli il valore atteso delle seguenti variabili aleatorie:

a) Numero delle volte che compare la lettera numero 1;

Scrivendo questa variabile aleatoria come $I_1 + ... + I_k$, dove I_j è la variabile indicatrice dell'estrazione della lettera 1 nella j-esima estrazione, e dato che $P(I_j = 1) = 1/n$, si ha che il valore atteso è k/n.

Le variabili addende sono indipendenti quindi la legge della variabile è binomiale $(k, 1/n)$ e la varianza è $\frac{k}{n}\left(1 - \frac{1}{n}\right)$. L'approssimazione di Poisson è appropriata quando n è grande e k/n tende ad una costante.

b) Numero delle lettere che non compaiono nella parola;

Scrivendo questa variabile aleatoria come $I_1 + ... + I_n$, dove I_j è la variabile indicatrice dell'evento "la lettera j non è stata estratta", dato che $P(I_j = 1) = (1 - 1/n)^k$, si ha che il valore atteso è $n\left(1 - 1/n\right)^k$.

In questo caso le variabili aleatorie non sono indipendenti neanche a coppie (il fatto che una lettera non è stata estratta fa diminuire la probabilità che una lettera diversa non venga estratta). Trascuriamo questo fatto, e chiediamoci come k debba comportarsi al crescere di n in modo che la media tenda a stabilizzarsi (e quindi $P(I_j = 1)$ tenda a 0), legittimando così l'approssimazione di Poisson. Teniamo conto che $\left(1 + \frac{a}{n}\right)^n$ tende ad a per $n \to \infty$ si ha che, se

$$k = k_n = n \ln(n/\lambda) \tag{1}$$

abbiamo che

$$n(1 - \frac{1}{n})^{k_n} \to \lambda \tag{2}$$

Si può dimostrare che, nonstante la dipendenza tra le variabili, con questa scelta di k la PMF della variabile in questione converge alla Poisson di media λ per $n \to \infty$.

c) Numero delle lettere che compaiono nella parola;

Scrivendo questa variabile aleatoria come la sottrazione dal numero n della variabile aleatoria al punto precedente e sfruttando la linearità si ha che il valore atteso è $n(1 - (1 - 1/n)^k)$.

Ciò equivale a scrivere la variabile come somma delle indicatrici degli eventi che ciascuna lettera compaia nella parola $1 - 1_{A_i}$, dove gli eventi A_i sono stati definiti al punto precedente, e applicare la linearità del valore atteso. Come già detto le variabili non sono indipendenti neanche a coppie ma trascuriamo questo aspetto e chiediamoci come scegliere $k = k_n$ per stabilizzare la media; intuitivamente mentre prima k doveva essere grande, stavolta deve essere piccolo. Dato che in ogni caso deve essere intero, prendiamo k costante: allora per la formula del binomio di Newton

$$n(1 - (1 - \frac{1}{n})^k) = k + O(\left(\frac{1}{n}\right)) \tag{3}$$

che quindi legittima l'approssimazione con una Poisson di media k della legge della variabile in questione, dato che si può di nuovo dimostrare che la dipendenza dalle variabili non inficia la convergenza della PMF.

d) Numero delle lettere che compaiono esattamente una volta nella parola;

Scrivendo questa variabile aleatoria come $I_1 + ... + I_n$, dove I_j è la variabile indicatrice dell'evento "la lettera j è stata estratta una sola volta", e dato che $P(I_j = 1) = \frac{k}{n}(1 - 1/n)^{k-1}$, si ottiene che il valore atteso è $k(1 - 1/n)^{k-1}$. Trascuriamo ancora una volta il fatto che le variabili I_j non sono indipendenti (neanche a coppie) e osserviamo che stavolta ci sono vari regimi possibili per k_n, al crescere di n, per la convergenza di $k(1 - 1/n)^{k-1}$: il caso più semplice è ovviamente k costante, ma in questo caso si può dimostrare che l'effetto della dipendenza non è trascurabile, quando $n \to \infty$. Per motivi che è troppo complicato qui spiegare la scelta opportuna (esprimendo n in funzione di k piuttosto che viceversa), è $n_k = k/\ln(k/\lambda)$, per far convergere la media a λ). Si noti che in questo caso n cresce più lentamente di k, in modo che l'informazione che la lettera j è stata estratta una volta sola non cambia di molto la probabilità che lo sia anche la lettera $h \neq j$.

e) Numero di lettere che compaiono più di una volta nella parola;

Scrivendo questa variabile aleatoria come differenza tra la variabile aleatoria al punto c) e quella al punto d) si ottiene che il valore atteso è

$$n - n(1 - \frac{1}{n})^k - k(1 - \frac{1}{n})^{k-1}. \tag{4}$$

Allo stesso risultato si perviene esprimendo la variabile come somma di indicatrici che, per ciascuna lettera, indicano se questa è stata estratta più di una volta. E' chiaro che anche in questo caso queste non sono indipendenti neanche a coppie; data la complicazione evitiamo di discuterne l'approssimazione di Poisson.

f) Numero delle volte in cui una lettera estratta è più grande delle lettere estratte precedentemente;

Questo esercizio è pensato nel caso in cui vengono estratte delle variabili continue, nel quale non ci possono essere *pareggi* tra le variabili. Ma nel nostro caso le variabili sono discrete e si può ovviamente estrarre due volte la stessa lettera. Comunque si scrive questa variabile aleatoria come $I_1 + ... + I_k$, dove I_j è la variabile indicatrice dell'evento "la lettera estratta alla j-esima estrazione è più grande di quelle estratte in precedenza". Il metodo più semplice per determinare $P(I_j = 1)$ è applicare la LOTP utilizzando la partizione associata ai valori della variabile aleatoria

$$Y_{j-1} = \max(X_1, ..., X_{j-1}). \tag{5}$$

dove X_i è il numero della lettera estratta per i-esima. Naturalmente $P(I_j = 1|Y_{j-1} = h) = 1 - \frac{h}{n}$, inoltre

$$P(Y_{j-1} = h) = P(Y_{j-1} \leq h) - P(Y_{j-1} \leq h - 1) = \left(\frac{h}{n}\right)^{j-1} - \left(\frac{h-1}{n}\right)^{j-1} \tag{6}$$

da cui

$$P(I_j = 1) = \sum_{h=1}^{n-1} (1 - \frac{h}{n}) \left[\left(\frac{h}{n}\right)^{j-1} - \left(\frac{h-1}{n}\right)^{j-1} \right] \tag{7}$$

espressione che può essere un pò manipolata per arrivare ad un'espressione (complicata) che dipende dalle somma delle potenze dei primi n interi.

g) Numero di coppie di estrazioni in cui è estratta la stessa lettera;

Si scrive questa variabile aleatoria come $I_{1,2} + ... + I_{k-1,k}$, dove $I_{i,j}$ è la variabile indicatrice dell'evento "nelle estrazioni i e j è uscita la stessa lettera" con $i < j$, e si nota che $P(I_{i,j} = 1) = 1/n$, per cui il valore atteso richiesto è $k(k-1)/2n$. Si noti ora che le variabili sono indipendenti a coppie perché sapere che $I_{i,j} = 1$ non cambia la probabilità che $I_{h,l} = 1$ se $\{i, j\} \cap \{h, l\} = \varnothing$, ma anche quando le due coppie di indici hanno un elemento in comune. Vedremo che questo implica che la varainza della somma è uguae alla soma delle varianze, in questo caso $\frac{k(k-1)}{2n}(1 - \frac{1}{n})$. Invece sapere che $I_{i,j} = 1$ e $I_{j,h} = 1$ ovviamente implica che $I_{i,h} = 1$, quindi le triple di variabili non sono tutte indipendenti. Per giustificare l'approssimazione di Poisson occorre che, quando $n \to \infty$, $k = k_n = \sqrt{2\lambda n}$, in modo che la media tenda a λ.

h) Numero di r-ple di estrazioni in cui sono estratte le stesse lettere.

Ovviamente è analogo al precedente, ci sono $\binom{k}{r}$ modi di scegliere i posti in cui può essersi verificata la coincidenza degli estratti e per uno qualunque di essi la probabilità che questa si verifichi è $n^{-(r-1)}$, per cui il valore atteso richiesto è $\binom{k}{r} n^{-(r-1)}$. Omettiamo per semplicità ulteriori discussioni.

2.N oggetti numerati da 1 a N vengono estratti uno dopo l'altro, senza reimmissione. Sia A_i l'i-esimo estratto, $i = 1, 2, \ldots, N$ (si tratta quindi di una permutazione aleatoria). Si calcoli il valore atteso delle seguenti variabili aleatorie:

a) Numero delle volte che $A_j = j$ (coincidenze), con $j = 1, ..., N$;

Si scrive questa variabile aleatoria come $I_1 + I_2 + ... + I_N$, dove I_j è la variabile indicatrice dell'evento $A_j = j$, che ha probabilità $1/N$, per cui il valore atteso è 1. Ovviamente le variabili aleatorie non sono indipendenti; un modo elegante per provarlo è che la variabile aleatoria in questione non può avere valore $N - 1$: se le variabili fossero indipendenti il numero delle coincidenze avrebbe distribuzione binomiale con N prove e media 1 e quindi necessariamente avrebbe come supporto l'insieme degli interi tra 0 e N. La dipendenza tra le variabili è debole al crescere di N, la probabilità che $A_j = j$ sapendo che $A_i = i$ è evidentemente $\frac{1}{N-1}$ invece che $\frac{1}{N}$ che è il valore ignorando questa informazione. La convergenza alla Poisson di media 1 è anche consistente con il fatto che la probabilità che non ci sia nessuna coincidenza tende a e^{-1} (vedi principio di inclusione-esclusione).

b) Numero di volte che $A_i > A_j$ per $i < j$;

Si scrive questa variabile aleatoria come $I_{1,2} + ... + I_{N-1,N}$, dove $I_{i,j}$ è la variabile indicatrice dell'evento $A_i > A_j$ che, per simmetria, ha probabilità $1/2$, quindi il valore atteso richiesto è $N(N-1)/4$. Le variabili $I_{i,j}$ e $I_{h,l}$ sono indipendenti quando $\{i, j\} \cap \{h, l\} = \varnothing$, non lo sono quando i due insiemi di indici ne hanno uno comune, infatti $P(I_{i,j} = 1, I_{j,h} = 1) = 1/6 \neq 1/4 = P(I_{i,j} =$

$1)P(I_{j,h}=1)$. In ogni caso un'approssimazione di Poisson è inappropriata (la media diverge quando N cresce), questo accadrà in tuti gli esempi fino al punto f).

c) Numero di massimi locali ($i=2,\ldots,N-1$ è massimo locale se $A_{i-1} <A_i$ e $A_i > A_{i+1}$, $i=2,\ldots,N-1$, mentre 1 è massimo locale se $A_1 > A_2$ e N è massimo locale se $A_N > A_{N-1}$);

La probabilità che i sia un massimo locale è $\frac{1}{3}$ (tutti i 6 ordinamenti dei valori A_{i-1}, A_i e A_{i+1} sono equiprobabili e 2 sono favorevoli all'evento in questione), per $i=2,...,N-1$, mentre la probabilità che 1 (ugualmente per N) sia un massimo locale è $1/2$ (punto b)), da cui scrivendo la variabile aleatoria come $I_1+I_2+...+I_N$, dove I_j è la variabile indicatrice dell'evento "j è un massimo locale" si ha che il valore atteso richiesto è $2\cdot\frac{1}{2}+\frac{N-2}{3}$ cioè $(N+1)/3$. Le variabili I_j e ${}_{I_{j+h}}$sono indipendenti quando $h\geq 3$, nonostante le variabili A_j non lo siano (qualunque siano i valori di A_{j-1}, A_j e A_{j+1} la probabilità di $I_{j+h}=1$ non cambia. Certamente non lo sono I_j e I_{j+1} (perché?), mentre vi lascio il noioso (ma istruttivo) calcolo di $P(I_j=1, I_{j+2}=1)$ al fine di determinare se I_j e I_{j+2} sono indipendenti.

d) Numero dei tratti monotoni massimali (un tratto è monotono di lunghezza $l-h$ quando $A_h< A_{h+1} < ... < A_l$ oppure $A_h> A_{h+1} > ... > A_l$ per qualche coppia $h<l$, ed è monotono massimale quando non può essere esteso ad un tratto monotono più lungo);

Questo numero è il numero dei massimi locali più il numero di minimi locali meno 1, quindi per simmetria la sua media è $2(N+1)/3-1=(2N-1)/3$. Qui si sfrutta il punto precedente.

e) Numero dei tratti monotoni di lunghezza l fissata;

Di possibili tratti di lunghezza l (che coinvolgono $l+1$ variabili A_i) ce ne sono $N-l$ e, condizionando all'insieme dei valori di questi, ci sono 2 ordinamenti su $(l+1)!$ che realizzano l'evento "il tratto da 1 a $l+1$ è monotono", per cui il valore atteso richiesto è $2(N-l)/(l+1)!$ (nota che per $l=1$ viene $N-1$, dato che tutte le coppie di interi consecutivi sono monotoni). Evidentemente quando i tratti sono disgiunti, cioè coinvolgono delle estrazioni differenti le variabili indicatrici che i due tratti siano monotoni sono indipendenti, mentre se i tratti si intersecano non sono indipendenti.

f) $\sum_{i=1}^N a_i A_i$, dove $a_i, i=1,...,N$ è una qualunque N-pla di costanti reali;

Per linearità, e per il fatto che $E(A_i)=\frac{N+1}{2}$, il valore atteso richiesto è $\frac{N+1}{2}\sum_{i=1}^N a_i$.

g) Numero delle trasposizioni (una trasposizione si ha quando, per qualche coppia $b_1\neq b_2$ si ha $A_{b_1}=b_2, A_{b_2}=b_1$);

Scrivendo questa variabile aleatoria come $I_{1,2}+...+I_{N-1.N}$, dove $I_{i,j}$ è l'indicatrice dell'evento "$A_i=j, A_j=i$" e tenendo conto che $P(I_{i,j}=1)=\frac{1}{N(N-1)}$ il valore atteso richiesto è $\binom{N}{2}\frac{1}{N(N-1)}=\frac{1}{2}$ (evidentemente per $N=2$, che è il minimo per cui la questione abbia senso, il risultato torna). Le variabili $I_{i,j}$ e $I_{j,h}$ non sono indipendenti perché $P(I_{j,h}=1|I_{i,j}=1)=0$ mentre quando $\{i,j\}\cap\{h,l\}=\varnothing$, $P(I_{i,j}=1|I_{h,l}=1)=\frac{1}{(N-2)(N-3)}$ che tuttavia differisce di poco quando N è grande da $\frac{1}{N(N-1)}$: si può dimostrare che questo basta per garantire la validità della convergenza alla PMF di Poisson di media $\frac{1}{2}$ quando $N \to \infty$._

h) Numero di cicli di lunghezza l (un ciclo di lunghezza l si ha quando esistono $b_1, ..., b_l \in \{1, ..., N\}$ tali che $A_{b_j} = b_{j+1}$ per $j = 1, ..., l$-1, dove $b_{l+1} = b_1$), una trasposizione è un ciclo di lunghezza $l = 2$;

La variabile in questione può essere ottenuta sommando una variabile aleatoria indicatrice di ogni ciclo distinto che può verificarsi. Dato un insieme $\{b_1, ..., b_l\}$ (che posso scegliere in $\binom{N}{l}$ modi diversi) ci sono $(l-1)!$ cicli distinti possibili e la probabilità di ciascuno di essi è $\frac{(N-l)!}{N!} = \frac{1}{N(N-1)...(N-l+1)}$, per cui il valore atteso richiesto è

$$\binom{N}{l} \frac{(l-1)!}{N(N-1)...(N-l+1)} = \frac{1}{l} \binom{N}{l} \frac{1}{\binom{N}{l}} = \frac{1}{l}. \tag{8}$$

i) Numero di cicli di lunghezza qualunque.

Per linearità $\frac{1}{2} + ... + \frac{1}{N}$ (non considerando cicli di lunghezza 1, quali potrebbero essere considerati i punti fissi).