Esercitazione del 29 aprile 2014

Esercizio 10.13 pg. 294
Complemento: Calcolare la probabilità che un negozio apra tra le sette e venti e le nove e quaranta del mattino.
Soluzione: Siccome non è nota la distribuzione della variabile casuale X che indica l'orario di apertura di un negozio del Comune, non è possibile dare una risposta esatta al quesito posto. Tuttavia, l'intervallo di cui si cerca la probabilità è simmetrico attorno al valore atteso $E(X)$ = 8:30 (otto e trenta) della popolazione. Pertanto, possiamo cercare di approssimare la probabilità cercata con la disuguaglianza di Chebychev:

$$P(|X - E(X)| \geq k\, \sigma(X)) \leq 1/k^2 .$$

Supponiamo, quindi, che X sia una variabile casuale assolutamente continua con valore atteso $E(X)$ = 8:30 e deviazione standard $\sigma(X) = 30$, in minuti. Si ha:

$$P(7{:}20 \leq X \leq 9{:}40) = P(|X - E(X)| \leq 70) = 1 - P(|X - E(X)| \geq 70) = 1 - P(|X - E(X)| \geq 70 \times 30 / 30) =$$

$$= 1 - P(|X - E(X)| \geq (7/3)\, \sigma(X)) \geq 1 - 1/(7/3)^2 = 1 - 9/49 = 40/49 = 0.816.$$

Osservazione. Se cerchiamo la probabilità che un negozio apra tra le 8:15 e le 8:45 la disuguaglianza di Chebychev non fornisce alcuna informazione di rilievo, infatti si ottiene
$P(8{:}15 \leq X \leq 8{:}45) = P(|X - E(X)| \leq 15) = 1 - P(|X - E(X)| \geq \sigma(X) / 2)$ per cui $k = ½$ e, quindi, l'applicazione della disuguaglianza porta solo a dire che tale probabilità è $\geq 1 - 4 = -3$, cioè ≥ 0.

Esercizio.
Un'urna di cui non sia possibile vedere l'interno contiene solo palline bianche e palline rosse. Sia p la frazione di palline rosse nell'urna. Consideriamo estrazioni successive, con re-immissione, di una pallina per volta dall'urna (*schema delle prove ripetute di Bernoulli*). Sia X la variabile che conta quante estrazioni si devono fare per ottenere la prima estrazione di una rossa.
1. Qual è la probabilità che la prima comparsa di una pallina rossa si abbia alla quarta estrazione?
2. Quale distribuzione ha la variabile casuale X ?
3. Per stimare p si ripete 15 volte l'esperimento di estrarre con re-immissione dall'urna una pallina alla volta fino all'uscita della prima pallina rossa, segnando per ciascuna delle 15 volte il risultato. Si ottengono i seguenti valori: 5, 2, 3, 5, 11, 1, 2, 4, 6, 7, 1, 3, 9, 10, 8. Usando questo campione, determinare la stima di p con il metodo della massima verosimiglianza.

Soluzione.
1. E' la probabilità che le prime tre estrazioni siano palline bianche e la quarta sia rossa.
Quindi: $(1-p)^3\, p$, perché le estrazioni sono indipendenti.

2. La variabile casuale X è una variabile discreta che può assumere un qualunque valore intero. Ragionando come al punto precedente si ha $P(X = k) = (1-p)^{k-1}\, p$, per k = 1, 2, … (la rossa esce per la prima volta alla estrazione k-ma se nelle prime k-1 estrazioni esce sempre bianca e poi esce rossa). Si tratta, quindi, della distribuzione ***geometrica*** di parametro p, che ha valore atteso dato da $1/p$ e varianza data da $(1-p)/p^2$.

3. Stimare p con il metodo della massima verosimiglianza significa cercare il valore di p che rende massima la probabilità di ottenere il campione osservato:
$P(X_1 = 5, X_2 = 2, X_3 = 3, X_4 = 5, X_5 = 11, X_6 = 1, X_7 = 2, X_8 = 4, X_9 = 6, X_{10} = 7, X_{11} = 1, X_{12} = 3, X_{13} = 9, X_{14} = 10, X_{15} = 8)$, ove $(X_1, \dots, X_{15})$ è un campione casuale estratto dalla distribuzione geometrica di parametro p.

Allora

$P(X_1 = 5, X_2 = 2, X_3 = 3, X_4 = 5, X_5 = 11, X_6 = 1, X_7 = 2, X_8 = 4, X_9 = 6, X_{10} = 7, X_{11} = 1, X_{12} = 3, X_{13} = 9, X_{14} = 10, X_{15} = 8) = [(1-p)^4 p]\,[(1-p)^1 p] \dots [(1-p)^7 p] = (1-p)^{62} p^{15}$

ove 15 è proprio la dimensione campionaria (n) mentre $62 = \sum_{i=1}^{15} (x_i - 1) = \sum_i x_i - n$, somma di tutti i dati (77) diminuita di n.

Si ha pertanto che la *verosimiglianza* è $L(p; x_1, \dots, x_{15}) = (1-p)^{62} p^{15}$, funzione reale di variabile reale, p, variabile in (0, 1). Determiniamo la *log-verosimiglianza*, considerando il logaritmo naturale (ln),

$$l(p; x_1, \dots, x_{15}) = 62 \ln(1-p) + 15 \ln p$$

e, tramite lo studio della derivata di $l(p; x_1, \dots, x_{15})$ rispetto a p, determiniamo la stima cercata. Si ha:

$$\frac{d\, l(p; x_1, \dots, x_{15})}{d\, p} = -62 \frac{1}{1-p} + 15 \frac{1}{p} = \frac{-62p + 15(1-p)}{p(1-p)} = \frac{-77p + 15}{p(1-p)}$$

Pertanto la derivata prima si annulla in $p = 15/77$, è positiva (e dunque crescente) per $p < 15/77$ e dunque 15/77 è l'unico punto di massimo della log-verosimiglianza $l(p; x_1, \dots, x_{15})$ e, dunque, della verosimiglianza $L(p; x_1, \dots, x_{15})$. Cioè, la stima cercata della frazione di palline rosse nell'urna è 15/77.
Dalla Figura 1, dove viene rappresentata la funzione di verosimiglianza $L(p; x_1, \dots, x_{15})$ in corrispondenza ai dati del testo, si vede che per valori di $p > 0.4$ la probabilità del campione è praticamente nulla, così come per $p < 0.05$.

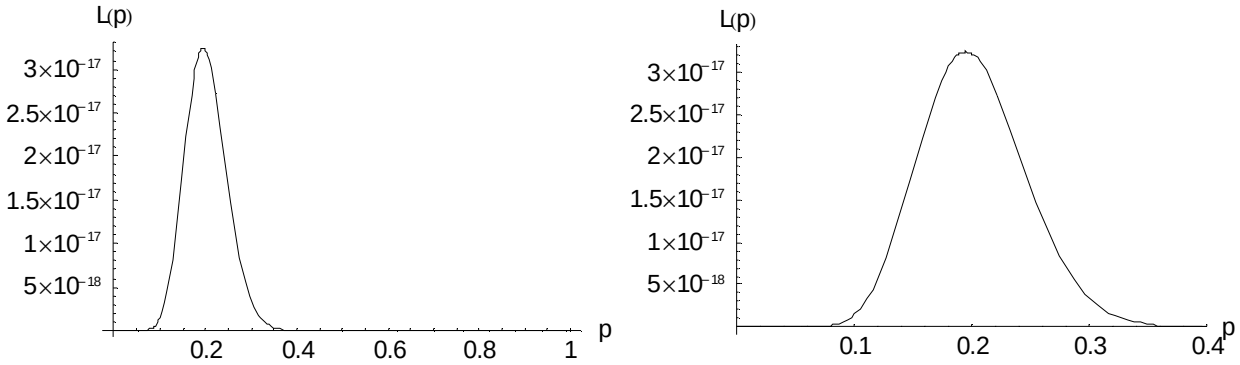

Figura 1 – A sinistra, grafico della funzione di verosimiglianza (e dunque, della probabilità del campione) al variare di p tra 0 e 1, a sinistra. In dettaglio attorno al massimo, a destra.

Esercizio 2 (parte seconda) dal tema d'esame del 25/06/2013.
Campione casuale $X_1, \dots, X_n$ da una popolazione con densità

$$f(x) = \frac{2}{\vartheta^2}\, x \exp\left[-(x/\vartheta)^2\right] \quad , \qquad x > 0$$

ove ϑ e' un parametro incognito > 0.

1. Determinare la funzione di log-verosimiglianza
2. Determinare l'espressione dello stimatore di massima verosimiglianza.
3. E' stato osservato il campione di 10 dati: 1.5, 2, 2.6, 1.4, 2.3, 3.1, 3.2, 1.6, 2.1, 2.2. Calcolare il valore assunto dallo stimatore di massima verosimiglianza in corrispondenza al campione.
4. Due esperti, Mario Rossi e Luigi Bianchi, hanno valutato il parametro incognito ϑ con i valori, rispettivamente, di 2 e 2.5. Sulla base del campione osservato, quale dei due esperti ha dato la valutazione più verosimile?

Soluzione

1. La funzione di verosimiglianza, per definizione, è data da:

$$L(\vartheta; x_1, \dots, x_n) = \prod_i \frac{2}{\vartheta^2}\, x_i \exp\left[-(x_i/\vartheta)^2\right]$$

e, quindi, quella di log-verosimiglianza è

$$l(\vartheta; x_1, \dots, x_n) = \sum_i \{\ln 2 + \ln x_i - (x_i/\vartheta)^2 - 2 \ln \vartheta\} = n \ln 2 + \sum_i \ln x_i - \left(\sum_i x_i^2\right) / \vartheta^2 - 2n \ln \vartheta \qquad (*)$$

2. La derivata prima della funzione di log-verosimiglianza rispetto a ϑ è

$$\frac{d\, l(\vartheta; x_1, \dots, x_n)}{d\,\vartheta} = \frac{2\left(\sum_i x_i^2\right)}{\vartheta^3} - \frac{2n}{\vartheta} = \frac{2\left(\sum_i x_i^2\right) - 2n\,\vartheta^2}{\vartheta^3}$$

che, quindi, si annulla per $\vartheta^2 = n^{-1}\left(\sum_i x_i^2\right)$. Siccome il denominatore è positivo per ipotesi ed il numeratore è positivo per $\vartheta^2 < n^{-1}\left(\sum_i x_i^2\right)$, si tratta proprio di un punto di massimo. Quindi, lo stimatore di massima verosimiglianza di ϑ è $\sqrt{\, n^{-1}\left(\sum_i x_i^2\right)}$.

3. In corrispondenza al campione assegnato, la stima di ϑ risulta essere 2.2786. La funzione di verosimiglianza $L(\vartheta; x_1, \dots, x_n)$ è rappresentata, per i dati assegnati, in Figura 2, dove si vede che valori di ϑ più grandi di 4 e minori di 1.5 circa rendono il campione praticamente impossibile.

4. Per rispondere calcoliamo l'espressione della log-verosimiglianza (*) in corrispondenza ai dati e per $\vartheta = 2$ (Rossi) e $\vartheta = 2.5$ (Bianchi) rispettivamente. Si ottiene:

$l(\vartheta = 2; 1.5, 2, 2.6, 1.4, 2.3, 3.1, 3.2, 1.6, 2.1, 2.2) = -12.39$
$l(\vartheta = 2.5; 1.5, 2, 2.6, 1.4, 2.3, 3.1, 3.2, 1.6, 2.1, 2.2) = -12.18 > -12.39$

dunque la valutazione più verosimile è quella di Luigi Bianchi, perché per $\vartheta = 2.5$ la log-verosimiglianza del campione risulta più alta (e, quindi, il campione "più probabile") che per $\vartheta = 2$.

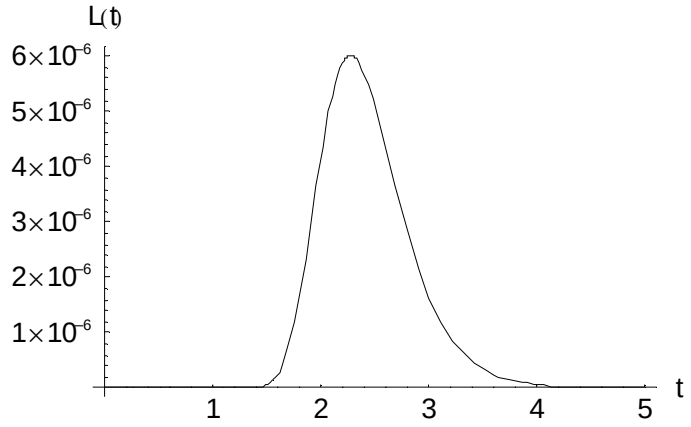

Figura 2 – Funzione di verosimiglianza per i dati dell'esercizio dal tema d'esame del 25/06/2013 in funzione del parametro ϑ, qui indicato con "t" per ragioni tipografiche.

Esercizio 2 (parte seconda) del tema d'esame del 18/09/2013

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

1. Stimare la percentuale di studenti universitari favorevoli al partito Arcobaleno. Fornire inoltre una stima della varianza dello stimatore.
2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 41. Si può ragionevolmente asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

Soluzione

Consideriamo un campione $X_1, \dots, X_n$ da una popolazione Bernoulliana di parametro p, cioè $X_i = 1$ se lo studente corrispondente è favorevole al partito, 0 altrimenti, e la probabilità che $X_i = 1$ è il parametro incognito p.

1. Allora, possiamo scrivere $P(X_i = x) = p^x(1-p)^{1-x}$ ove $x = 0$ o 1, e quindi la log- verosimiglianza come

$$l(p; x_1, \dots, x_n) = \sum_{i=1}^{n} \ln P(X_i = x_i) = \sum_{i=n}^{n} \{x_i \ln p + (1-x_i)\ln(1-p)\} = (\ln p)\sum_{i=1}^{n} x_i + \{\ln(1-p)\}\sum_{i=1}^{n}(1-x_i) =$$

$$= n\bar{x}\ln p + n\ln(1-p) - n\bar{x}\ln(1-p)$$

da cui

$$\frac{dl(p; x_1, \dots, x_n)}{dp} = \frac{n\bar{x}}{p} - \frac{n}{1-p} + \frac{n\bar{x}}{1-p} = 0 \Leftrightarrow n\bar{x}(1-p) - np + n\bar{x}p = 0 \Leftrightarrow p = \bar{x}$$

che si verifica essere un punto di massimo. Quindi, la media campionaria è lo stimatore di massima verosimiglianza. Nel caso del campione dell'esercizio, 0.40.

Sappiamo allora che la sua distribuzione deriva dalla distribuzione Binomiale(n,p) e quindi che la sua media è data proprio da p e la sua varianza da $p(1-p)/n$. Pertanto, una stima della varianza

dello stimatore è $\bar{x}(1-\bar{x})/n$. Per una media campionaria pari a 0.40, il caso del nostro campione, la stima della varianza è 0.4×0.6/100 = 0.0024.

2. Per rispondere a questa domanda, domandiamoci che probabilità c'è che in un campione di 100 studenti ci siano 41 o anche più favorevoli al partito Arcobaleno, quando la vera proporzione di favorevoli nell'intera popolazione è 0.40 (come nel 2012). Basta calcolare

$$P^*(\bar{X}_{100} \geq 0.41) = P^*(X_1 + \ldots + X_{100} \geq 41)$$

ove *P** indica che sto considerando un campione dalla distribuzione Bernoulliana di parametro 0.40. Il calcolo è teoricamente possibile perché sappiamo che $X_1 + \ldots + X_{100}$ è una variabile Binomiale(100, 0.40), ma in pratica ricorriamo all'approssimazione gaussiana:

$$P^*(\bar{X}_{100} \geq 0.41) = P^*\left(\frac{\bar{X}_{100} - 0.4}{\sqrt{0.0024}} \geq \frac{0.41 - 0.4}{\sqrt{0.0024}}\right) \approx P^*(Z \geq 0.20) = 1 - P^*(Z \leq 0.20) = 1 - 0.5793 = 0.42$$

Ne concludiamo che è abbastanza probabile ottenere 41 o più favorevoli al partito in un campione di dimensione pari a 100 quando la "vera" percentuale di favorevoli nell'intera popolazione è del 40%, e quindi non c'è ragione evidente per pensare ad un aumento del favore al partito nel 2013.

Se nel 2013 i favorevoli fossero stati 50, procedendo allo stesso modo avremmo ottenuto

$$P^*(\bar{X}_{100} \geq 0.5) = 0.0227$$

e quindi, al contrario, ne avremmo concluso che molto probabilmente il favore è aumentato, perché con una percentuale di favorevoli pari al 40% un risultati di almeno 50 favorevoli in un campione di 100 è piuttosto improbabile.

Esercizio 9.14 pg. 267.
Complemento: Calcolare P(X = 1 | Y = 1). Il risultato cosa ci dice a proposito della indipendenza delle variabili X ed Y ?