Università degli Studi di Salerno

Dipartimento di Informatica

Progetto di Statistica e Analisi dei Dati

Studio sulle variabili aleatorie

Docente

Amelia Giuseppina Nobile

Studenti

Francesco Abate 0522500993

Vincenzo De Martino 0522500966

Sommario

Capitolo 1

Variabile aleatoria discreta di Bernoulli

La variabile aleatoria di Bernoulli è detta discreta siccome assume un numero finito o numerabile di valori xi con rispettive probabilità pi; tale variabile aleatoria viene utilizzata in esperimenti in cui è possibile avere solo due risultati, cioè il successo e l'insuccesso, i quali si verificano rispettivamente con probabilità p e 1-p, dove p può assumere un qualsiasi valore compreso tra 0 e 1 inclusi. Segue la funzione di probabilità, la quale mostra la probabilità in base al valore dell'esperimento X.

$$p_X(x) = P(X = x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \\ 0, & altrimenti, \end{cases}$$

Definiamo ora il valore atteso E(X), il quale è la previsione del valore che ci si aspetta. Il valore atteso mette in relazione i valori di X con le associate probabilità, quindi avremo:

$$E(X) = p$$

Definiamo la varianza Var(X), la quale non è altro che un indice che fornisce una misura della variabilità dei valori assunti dalla variabile stessa.

$$Var(X) = E(X^2) - \{E(X)\}^2 = p - p^2 = p(1-p)$$

Un esempio pratico riguardante l'utilizzo della variabile aleatoria di Bernoulli è il lancio di un dado con la scommessa su un preciso numero i, compreso tra 1 e 6: se il lancio del dado riporta il numero scommesso si avrà un successo, altrimenti un insuccesso. Il successo si avrà sicuramente con probabilità p = 1/6, siccome si cerca di indovinare uno dei sei numeri in un unico lancio.

Variabile aleatoria discreta binomiale

Il problema è che Bernoulli, in sé, è limitato, siccome tende ad effettuare un esperimento composto da un'unica prova. È possibile, però, effettuare n prove di Bernoulli indipendenti considerando la variabile aleatoria binomiale.

Cosa significa, però, effettuare n prove indipendenti? Ogni prova effettuata nell'esperimento non deve influenzare la successiva. Poniamo di voler fare una scommessa sull'estrazione delle k biglie da una sacca: se viene estratta una biglia per una prova ed essa non viene reinserita, nella prossima estrazione verrà estratta una nuova biglia tra le k-1 rimanenti nella sacca. In questo caso, si parla di prove dipendenti, siccome l'estrazione influisce l'esito della prossima estrazione. Per rendere le prove indipendenti, sarebbe corretto inserire nuovamente la biglia estratta nella sacca.

A tal punto, X assume un qualsiasi valore pari al numero di successi in n prove. In sostanza, si sta ripetendo l'esperimento di Bernoulli n volte su n prove indipendenti. La funzione di probabilità è la seguente:

$$p_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & altrimenti, \end{cases}$$

Il valore medio, invece, è E(X) = n p, mentre la varianza è Var(X) = n p (1 - p).

Vediamo un esempio riguardante l'utilizzo della variabile aleatoria binomiale: in un gioco d'azzardo, un giocatore scommette su uno dei numeri compresi tra 1 e 6. Si lanciano tre dadi: per ogni volta in cui esce il numero scelto, si vince un euro; altrimenti, si perde un euro.
Calcolare la probabilità in cui si abbiano 3, 2, 1 e nessun match.

$P(X = 3) = \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^0 = \frac{1}{216}$

$P(X = 2) = \binom{3}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^1 = \frac{15}{216}$

$P(X = 1) = \binom{3}{1} \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^2 = \frac{75}{216}$

$P(X = -1) = \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^3 = \frac{125}{216}$

Perché per X = 2 e X = 1 si moltiplica un coefficiente binomiale? Esso stabilisce il numero dei modi con cui si possono scegliere i successi nella sequenza. Siano k i successi e sia n il numero degli elementi della sequenza, il coefficiente binomiale sarà il seguente:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Perché si moltiplica, quindi, il numero dei modi con cui possono essere disposti i successi? Ebbene, se non si moltiplicasse il coefficiente binomiale si indicherebbe che quel determinato caso si verificherebbe in un'unica disposizione (ad esempio, per X = 2, ciò varrebbe se i due successi escano solo nei primi due tentativi e in nessuna altra disposizione), mentre i successi sono rilocabili in più modi. Se ciò non è ancora chiaro, segue un esempio connesso al precedente in cui con "v" si indica un successo, mentre con "x" si indica un fallimento.

Per X = 2

1) v v x
2) v x v
3) x v v

Sono tre combinazioni che rispettano la condizione X = 2, difatti $\binom{3}{2} = 3$.

Lo stesso ragionamento, comunque, va fatto per i casi in cui si hanno tutti i successi e tutti i fallimenti, quindi in tal caso per X = -1 e X = 3: difatti il coefficiente binomiale viene moltiplicato ma non riportato siccome ha valore 1.

$$\binom{3}{0} = \binom{3}{3} = 1$$

Possiamo, infine, calcolare E(X), il quale ci permette di osservare se il gioco è equo o meno:

$E(X) = -1 * \frac{125}{216} + 1 * \frac{75}{216} + 2 * \frac{15}{215} + 3 * \frac{1}{216} = -0{,}079$, quindi si tende a perdere.

Definiamo, inoltre, la funzione di distribuzione, la quale definisce la probabilità che X assuma un determinato valore minore o uguale ad x.

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}, & k \le x < k+1 \qquad (k = 0, 1, \ldots, n-1) \\ 1, & x \ge n. \end{cases}$$

Il linguaggio R semplifica il tutto tramite la funzione dbinom(x, size, prob), dove x è l'insieme dei valori assumibili dalla variabile aleatoria binomiale, size è il numero di prove da svolgere e prob è la probabilità di successo in ogni prova: tale funzione non fa altro che calcolare le probabilità binomiali.

La funzione pbinom(x, size, prob, lower.tail = TRUE) viene utilizzata per il calcolo della funzione di distribuzione binomiale, dove x è l'insieme dei valori assumibili dalla variabile aleatoria binomiale, size è il numero di prove da svolgere, prob è la probabilità di successo in ogni prova e lower.tail calcola $P(X \le x)$ se TRUE, altrimenti $P(X > x)$: tale funzione non fa altro che calcolare la funzione di distribuzione binomiale.

R semplifica anche il calcolo del valore medio, della varianza, della deviazione standard e del coefficiente di variazione della distribuzione binomiale. Sia n il numero di prove e sia p la probabilità, tali componenti si calcolano nel seguente modo:

```
x <- 0:n
EX <- sum(x*dbinom(x, size=n, prob=p))
VarX <- (sum(x^2*dbinom(x, size=n, prob=p))) - EX^2
DevStd <- sqrt(VarX)
CoefVar <- sqrt(VarX)/EX
c(EX, VarX, DevStd, CoefVar)
```

R permette anche il calcolo dei quantili tramite la funzione qbinom(z, size, prob), dove z è l'insieme delle misure per i quantili, size è il numero di prove da svolgere e prob è la probabilità di successo in ogni prova.

Infine, è possibile simulare la variabile aleatoria binomiale generando una sequenza di numeri pseudocasuali mediante la funzione rbinom(N, size, prob), dove N è il numero di esperimenti da svolgere, size è il numero di prove da svolgere per ogni esperimento e prob è la probabilità di successo in ogni prova.

```
simulation <- rbinom(50, 20, 0.2)
simulation
table(simulation)
freq = (table(simulation)/length(simulation))
freq
```

```
> simulation <- rbinom(50, 20, 0.2)
>   simulation
 [1] 3 4 4 3 5 4 2 2 3 4 5 4 4 0 4 6 2 3 5 6 1 4 4 4 5 8 2 3 5 6 2 5 3 5 6 7
[37] 3 2 3 6 3 5 4 2 1 2 2 5 4 4
>   table(simulation)
simulation
 0  1  2  3  4  5  6  7  8
 1  2  9  9 13  9  5  1  1
>   freq = (table(simulation)/length(simulation))
>   freq
simulation
   0    1    2    3    4    5    6    7    8
0.02 0.04 0.18 0.18 0.26 0.18 0.10 0.02 0.02
```

Ciò che restituisce la funzione rbinom, precisamente, è una lista di numeri dove ogni numero rappresenta il numero di successi per ogni esperimento svolto. Nell'esempio soprastante vengono svolti 50 esperimenti, ognuno composto da 20 prove: nella posizione K dell'array simulation viene indicato il numero di successi per il K-esimo esperimento.

Ogni esecuzione di tale funzione restituirà valori diversi siccome, appunto, calcola il tutto casualmente. È possibile fissare la sequenza restituita tramite la funzione set.seed(inputSeed).

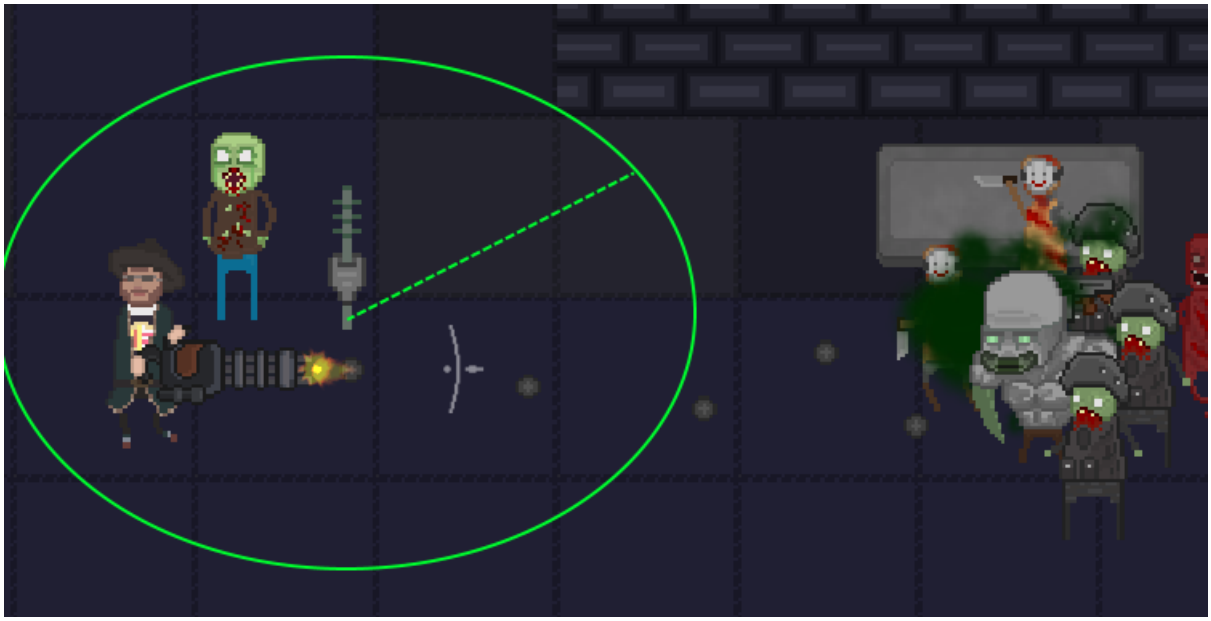
Capitolo 2

Introduzione all'applicazione della variabile aleatoria

Lo scopo di questa tesi consiste nello svolgere diverse stime utilizzando la variabile aleatoria binomiale. Al fine di poterla usare, è necessario un campione, detto anche popolazione, su cui lavorare. Tale campione farà riferimento ad una situazione riscontrata durante la progettazione di un videogioco, quindi ne vediamo subito la problematica.

Dungate è un videogioco di sopravvivenza in cui il giocatore deve resistere quanto possibile da un'invasione di mostri, difendendosi con armi e accessori in cooperazione con i propri amici. Attualmente, lo sviluppo del videogioco comprende l'introduzione di diverse novità, tra cui l'introduzione dei totem, quali sono dei gadget che il giocatore può utilizzare per sopravvivere al meglio.

Uno di questi gadget, detto "Stealer Totem", permette al giocatore, secondo una percentuale di successo e se egli si trova nell'area di effetto, di recuperare punti vita quando con l'arma colpisce qualche nemico.



Si è deciso di dare una percentuale di successo all'effetto del totem per bilanciare al meglio il videogioco, altrimenti tale gadget sarebbe risultato molto più conveniente rispetto agli altri che il gioco offre.

Si voglia supporre, quindi, una percentuale di successo del 14% per ogni colpo sparato; si considera l'arma con il rateo di fuoco più alto, quindi la Minigun (l'arma illustrata nell'immagine), la quale in un minuto spara ben 600 colpi (secondo il manuale del gioco, 10 colpi ogni secondo).

Si voglia testare il totem ben 50 volte, dove in ogni test l'arma spara 600 colpi in un minuto e per ogni colpo c'è il 14% di successo nell'attivazione dell'effetto del totem.

Testare 50 volte il totem sparando 600 colpi in un minuto nella stessa posizione è impossibile, ergo ci affideremo alla simulazione che il linguaggio R offre.

Dal capitolo precedente, consideriamo la funzione rbinom(N, size, prob), dove N è il numero di esperimenti da svolgere, size è il numero di prove da svolgere per ogni esperimento e prob è la probabilità di successo in ogni prova.
Nel caso ipotizzato: N = 50; size = 600; prob = 0.14.

```
########################################
# gvRaw <- rbinom(50, 600, 0.14)
# gvRaw
########################################
```

```
> data <- gvRaw
> data
 [1]  84  79  76  77  75  87  84  90  79  86  91  84  83  80  81
[16]  85  82  84  81  83  78 105  81  68  88  73  76  86  91  86
[31]  94  77  91  87  96  85  85  80  82  76  79  74  78  73  85
[46]  81  93  89  77  82
```

Al fine di evitare problematiche con il dataset generato, esso è stato trascritto valore per valore in un array in modo da non perderlo mai e il codice della generazione è stato commentato, al fine di renderlo inutilizzabile.

```
data = c(84,  79,  76,  77,  75,  87, 84,  90,  79,  86,  91,  84,  83,  80,
         81,  85,  82,  84,  81,  83,  78, 105,  81,  68,  88,  73,  76,  86,
         91,  86,  94,  77,  91,  87,  96,  85,  85,  80,  82,  76,  79,  74,
         78,  73,  85,  81,  93,  89,  77,  82)
data
```

I dati restituiti dalla funzione rbinom riguardano i numeri di successi per ogni esperimento eseguito. Ciò vale a dire che preso un determinato xi appartenente all'array data, xi sarà il numero di successi riguardante l'i-esimo esperimento.

Dati tutti questi esperimenti, è interessante comprendere le frequenze riguardanti i numeri di successi per ogni esperimento, ergo le ricaviamo tramite il seguente codice.

```
frequences = (table(data)/length(data))
sort(frequences, decreasing = FALSE)
```

```
> frequences = (table(data)/length(data))
> sort(frequences, decreasing = FALSE)
data
  68   74   75   88   89   90   93   94   96  105   73   78   80   83   87   76   77   79
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.04 0.04 0.04 0.04 0.04 0.06 0.06 0.06
  82   86   91   81   84   85
0.06 0.06 0.06 0.08 0.08 0.08
```

Concludendo, siccome ora si possiede un campione su cui lavorare, nonostante sia generato da R, è possibile effettuare delle stime sui dati.

Capitolo 3

Stima puntuale: campioni casuali e stimatori

Uno dei problemi dell'inferenza statistica consiste nello studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene uno o più parametri non noti. Se non esistono parametri non noti, allora la legge di probabilità è completamente specificata.

Per ottenere informazioni sui parametri non noti della popolazione, si possono effettuare delle misurazioni su un campione della popolazione, come quello ricavato nel capitolo precedente. Tale campione non è altro che un sottoinsieme dell'array data generato casualmente nel capitolo precedente, dove ogni item appartenente ad esso viene identificato come variabile aleatoria Xi, mentre il corrispettivo valore viene identificato come xi.

La funzione di distribuzione del campione casuale generato è la seguente.

$$F_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$$
$$= P(X_1 \le x_1)\, P(X_2 \le x_2) \cdots P(X_n \le x_n) = \prod_{i=1}^{n} F_X(x_i).$$

Considerato, quindi, il campione di grandezza n, definiamo uno stimatore come una funzione in cui i valori x1, …, xn vengono utilizzati per stimare un parametro non noto della popolazione. Un parametro non noto della popolazione è proprio il valore p, quindi la probabilità di successo per ogni prova. I principali metodi di stima puntuale dei parametri sono il metodo dei momenti e il metodo della massima verosimiglianza.

Attenzione: è bene sapere che al crescere dell'ampiezza del campione, la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione, ergo più sarà grande il campione e più precisa sarà la stima.

Metodo dei momenti

Il metodo dei momenti è uno dei più antichi e semplici metodi di stima dei parametri. Per illustrarlo è necessario definire i momenti campionari.

Un momento campionario r-esimo è la media aritmetica delle n osservazioni effettuate sulla popolazione elevate alla potenza r, in sostanza:

$$M_r(x_1,x_2,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

Da quel che possiamo notare, quindi, per r=1 avremo che il momento campionario di ordine 1 sarà uguale al valore dato dalla media campionaria, quindi varrà che:

$$M_1 = (x_1 + x_2 + \ldots + x_n)/n.$$

Nel generale esistono k parametri da stimare, quindi il metodo dei momenti consiste nell'uguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale. In sostanza, quindi, si risolve un sistema di k equazioni:

$$E(X^r) = M_r(x_1, x_2, \ldots, x_n) \qquad (r = 1, 2, \ldots, k).$$

I termini alla sinistra dell'equazione dipendono dal parametro non noto, nel nostro caso il valore della probabilità, mentre i termini a destra vengono calcolati partendo dal campione osservato.

A tal punto, quindi, riscriviamo l'equazione sapendo che E(X) = kp:

$$k\,\widehat{p} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \widehat{p} = \frac{\overline{x}}{k}.$$

dove p con cappelletto si indica la probabilità stimata.
Il metodo dei momenti fornisce quindi come stimatore del parametro kp la media campionaria.

Ricordiamo, come annunciato prima, che al crescere dell'ampiezza del campione, la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione, ergo più sarà grande il campione e più precisa sarà la stima.

Supponiamo, quindi, un campione composto da 10 elementi, otteniamo in subdata il sottoinsieme di 10 elementi, per poi ricavare la stima della probabilità dividendo la media campionaria del campione con il numero di prove effettuate in un esperimento, quindi k = 600. Il risultato con 10 elementi è pari a 0.1361667.

```
# stima puntuale con 10 items
subdata <- data[0:10]
stima10 = mean(subdata)/600
stima10
```

```
> # stima puntuale con 10 items
>   subdata <- data[0:10]
>   stima10 = mean(subdata)/600
>   stima10
[1] 0.1361667
```

Vediamo, ora, per 30 elementi.

```
# stima puntuale con 30 items
subdata <- data[0:30]
stima30 = mean(subdata)/600
stima30
```

```
> # stima puntuale con 30 items
>   subdata <- data[0:30]
>   stima30 = mean(subdata)/600
>   stima30
[1] 0.1379444
```

Dal campione di 30 elementi otteniamo una stima leggermente più alta, pari a 0.1379444. Al momento sembra che al crescere dell'ampiezza del campione la stima sia sempre più accurata, quindi concludiamo adottando un campione composto da tutti e 50 gli elementi.

```
# stima puntuale con 50 items
subdata <- data[0:50]
stima50 = mean(subdata)/600
stima50
```

```
>   # stima puntuale con 50 items
>   subdata <- data[0:50]
>   stima50 = mean(subdata)/600
>   stima50
[1] 0.1382333
```

Notiamo come il campione da 50 elementi si avvicini di più rispetto al campione da 30 al reale valore della probabilità utilizzata.
Essendo una stima non si riuscirà a ricavare il preciso valore di p, ma considerando campioni ancora più grandi ci si avvicinerebbe ancora di più.
È già un ottimo risultato il fatto che con 50 elementi la stima puntuale si sia avvicinata moltissimo al reale valore di p, il quale ricordiamo essere 0.14.

Metodo della massima verosimiglianza

Il metodo della massima verosimiglianza è il più importante metodo per la stima dei parametri non noti di una popolazione ed è solitamente preferito rispetto al metodo dei momenti.

Dato un campione di ampiezza n estratto dalla popolazione, la funzione di verosimiglianza è la funzione di probabilità congiunta del campione casuale, ossia:

$$\begin{aligned} L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) &= L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k; x_1, x_2, \ldots, x_n) \\ &= f(x_1; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)\, f(x_2; \vartheta_1, \vartheta_2, \ldots, \vartheta_k) \cdots f(x_n; \vartheta_1, \vartheta_2, \ldots, \vartheta_k). \end{aligned}$$

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri non noti, quindi cercando di determinare i valori $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ che rendono massima la funzione di verosimiglianza. I valori che massimizzano la funzione di verosimiglianza vengono solitamente indicati con il cappelletto e costituiranno le stime di massima verosimiglianza dei parametri non noti della popolazione.

Ci si propone, quindi, di determinare lo stimatore di massima verosimiglianza del valore medio di una popolazione binomiale ricavata dal campione di elementi. Anche qui verificheremo se al crescere dell'ampiezza del campione, la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione, ergo più sarà grande il campione e più precisa sarà la stima. Ciò comporta al comprendere da quale campione proviene, quanto più plausibilmente, il parametro non noto, nel nostro caso p.

$$p_X(x) = \binom{k}{x} p^x (1-p)^{k-x}$$

Dove k è il numero di prove per il singolo esperimento, x è il numero di successi ottenuto.

Si ha:

$$L(p) = \binom{k}{x_1} p^{x_1} (1-p)^{k-x_1} \binom{k}{x_2} p^{x_2} (1-p)^{k-x_2} \cdots \binom{k}{x_n} p^{x_n} (1-p)^{k-x_n}$$

$$= \binom{k}{x_1}\binom{k}{x_2}\cdots\binom{k}{x_n} p^{x_1+x_2+\ldots+x_n} (1-p)^{nk-(x_1+x_2+\ldots+x_n)} \qquad (0<p<1),$$

In sostanza, si moltiplicano tra loro le funzioni di probabilità per ogni elemento del campione. Dopo svariati calcoli, arriveremo a dire che la stima di massima verosimiglianza del parametro kp è la seguente.

$$k\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{ossia} \qquad \widehat{p} = \frac{\overline{x}}{k}.$$

Quindi, per una popolazione binomiale lo stimatore di massima verosimiglianza e dei momenti del valore medio E(X) = kp è la media campionaria $\overline{X}$.

La formula ricavata per il calcolo del valore non noto p è identica a quella ricavata durante l'utilizzo del metodo dei momenti, ergo i valori di p ricavati sono praticamente identici.

Proprietà degli stimatori

Uno stimatore, è caratterizzato da diverse proprietà, e può essere:

- Corretto;
- Più efficiente di un altro;
- Corretto e con varianza uniformemente minima;
- Asintoticamente corretto;
- Consistente.

Correttezza: uno stimatore è detto corretto se il suo valore medio è pari al valore non noto calcolabile.

Efficienza: si misura tramite l'errore quadratico medio, il quale è il valore medio della differenza tra stimatore e valore stimato al quadrato. Tale errore permette di verificare quanto si discosta lo stimatore dal valore stimato. Il problema è che esistono tantissimi stimatori, ergo si misurano solo quelli corretti. Se lo stimatore è corretto, l'errore quadratico medio è pari alla sua varianza.

$$MSE(\widehat{\Theta}) = E\{[\widehat{\Theta} - E(\widehat{\Theta})]^2\} = \mathrm{Var}(\widehat{\Theta}).$$

Possiamo, quindi, dire che uno stimatore è migliore rispetto agli altri se la sua varianza MSE è più piccola degli altri. Verrebbe, quindi, da pensare che per trovare lo stimatore migliore sia necessario trovare la varianza più piccola, ma ciò non è possibile siccome per ogni stimatore deve valere la disuguaglianza di Cremèr-Rao:

$$\mathrm{Var}(\widehat{\Theta}) \geq \frac{1}{nE\left\{\left[\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right]^2\right\}}.$$

Se vale ciò, allora lo stimatore considerato è **corretto e con varianza minima**.

Asintoticamente corretto: si desidera verificare che uno stimatore è asintoticamente corretto della varianza di una popolazione. Nel caso di una popolazione di Bernoulli, sapendo che E(X) = p, si ricava:

$$\lim_{n\to+\infty} E(\widehat{\Theta}_n) = \lim_{n\to+\infty} E\Big(\frac{n\,\overline{X}+1}{n+2}\Big) = \lim_{n\to+\infty} \frac{n\,E(\overline{X})+1}{n+2} = p.$$

Ergo, lo stimatore è asintoticamente corretto del parametro p.

Consistenza: uno stimatore del parametro non noto della popolazione è detto consistente se e solo se, per ogni epsilon, si ha:

$$\lim_{n\to+\infty} P\big(|\widehat{\Theta}_n - \vartheta| < \varepsilon\big) = 1,$$

In sostanza, lo stimatore è detto consistente se e solo se converge in probabilità a ϑ.

Una condizione sufficiente affinché lo stimatore sia consistente e che sia asintoticamente corretto riguarda il fatto che la varianza tenderà a zero al crescere del campione.
Attenzione: uno stimatore può essere consistente senza essere asintoticamente corretto.

$$i)\ \lim_{n\to\infty} E(\widehat{\Theta}_n) = \vartheta,$$

$$ii)\ \lim_{n\to+\infty} \mathrm{Var}(\widehat{\Theta}_n) = 0.$$

Capitolo 4

Introduzione alle stime intervallari

Nel precedente capitolo abbiamo visto come è possibile stimare un valore non noto, quale è la probabilità, partendo da una popolazione raffigurante i casi in cui il totem del giocatore applica il suo effetto in svariate prove.

È preferibile, però, fornire per tale parametro non noto un determinato intervallo nel quale quest'ultimo possa ricadere, invece di un singolo valore ricavato dalla stima puntuale o dal metodo della massima verosimiglianza. Tale intervallo viene denominato come intervallo di confidenza ed è caratterizzato da due estremi C1 e C2 tra i quali può ricadere il parametro non noto, quindi la probabilità di successo. Il grado di fiducia della stima, deciso dal decisore, è dato dal coefficiente 1 – alpha, il quale, in sostanza, indica la probabilità entro la quale il valore non noto si trovi nell'intervallo specificato.

$$P(\underline{C}_n < \vartheta < \overline{C}_n) = 1 - \alpha$$

Difatti, come ben leggiamo dalla formula soprastante, la probabilità che il parametro non noto si trovi tra i due estremi è pari ad 1 – alpha.

Metodo pivotale

Il metodo pivotale è utile alla definizione degli intervalli di confidenza, quindi permette di stimare un parametro non noto di una popolazione tramite l'utilizzo di una variabile aleatoria detta "variabile di Pivot", la quale dipende dal campione casuale che si sta analizzando e dal parametro non noto (in tal caso la probabilità di successo). Infine, la sua

funzione di distribuzione non contiene il parametro non noto da stimare. Per la variabile di Pivot, vale la seguente legge:

$$P\big(\alpha_1 < \gamma(X_1, X_2, \ldots, X_n; \vartheta) < \alpha_2\big) = 1 - \alpha.$$

Quindi, la probabilità che la variabile di Pivot abbia valore compreso tra i coefficienti a1 e a2, i quali raffigurano un range, è proprio pari ad 1 – alpha, quindi al grado di fiducia della stima.

Se dalla disequazione interna si esterna il parametro non noto, si ottiene non altro che la relazione che lega la stima intervallare al grado di fiducia stabilito, dove le funzioni g(x) non fanno altro che calcolare i rispettivi limiti del range.

$$P\big(g_1(X_1, X_2, \ldots, X_n) < \vartheta < g_2(X_1, X_2, \ldots, X_n)\big) = 1 - \alpha.$$

$$P\big(\underline{C}_n < \vartheta < \overline{C}_n\big) = 1 - \alpha$$

Pertanto, il metodo pivotale è utilizzabile solo su popolazioni normali, ergo non è applicabile in tal caso siccome si sta trattando una popolazione di dati legata alla variabile binomiale. Pertanto, è possibile utilizzare un metodo approssimativo rispetto a quello pivotale, il quale prende il nome di metodo pivotale approssimato.

Stima intervallare tramite il metodo pivotale approssimato

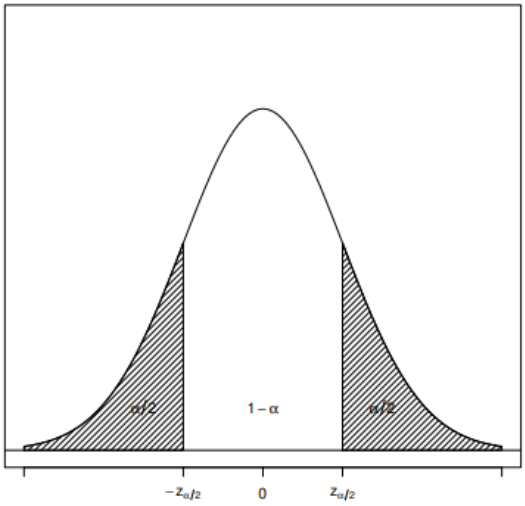
Consideriamo, quindi, l'utilizzo del metodo pivotale approssimato, il quale fa uso a sua volta del teorema centrale di convergenza per determinare un intervallo di confidenza con uno specifico grado di fiducia; inoltre, tale metodo dipende dalla dimensione del campione su cui si vuole effettuare la stima, il che deve essere statisticamente grande (cioè deve essere composto da almeno 30 elementi). Se X denota la variabile aleatoria con $E(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$, allora il teorema centrale di convergenza afferma che la variabile aleatoria converge in una distribuzione ad una variabile aleatoria normale standard.

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

A tal punto, la variabile aleatoria Zn può essere interpretata, quindi, come una variabile aleatoria di Pivot siccome dipende dal campione utilizzato, dipende dal parametro non noto tramite valore medio e varianza e, infine, per grandi campioni (grandezza del campione maggiore o uguale a 30) la sua funzione di distribuzione è approssimativamente normale standard siccome non contiene il parametro non noto da stimare.

In sostanza, quindi, tale variabile può essere vista come variabile di Pivot utilizzabile per il calcolo dell'intervallo [C1, C2] secondo il grado di fiducia stabilito. Vale, quindi, la forma approssimata:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \simeq 1 - \alpha.$$



Siccome, quindi, la dimensione del campione che analizzeremo è decisamente maggiore rispetto a 30 elementi (difatti è composta da ben 50 elementi), è possibile applicare il metodo pivotale in forma approssimata considerando come parametro non noto la probabilità di successo.

Ricapitoliamo, quindi, che nella nostra variabile binomiale Il valore medio vale E(X) = k p, mentre la varianza vale Var(X) = k p (1 - p), dove k è il numero di prove, quindi entrambi dipendono dal parametro non noto p.

Applicando il teorema centrale di convergenza, si ha:

$$\frac{mean(Xn) - E(X)}{\frac{\sqrt{Var(X)}}{\sqrt{n}}} = \sqrt{n} * \frac{mean(Xn) - E(X)}{\sqrt{Var(X)}} = \sqrt{n} * \frac{mean(Xn) - (k * p)}{\sqrt{(k * p(1-p))}}$$

La variabile aleatoria converge in distribuzione ad una variabile aleatoria normale standard. Per grandi campioni, l'intervallo di confidenza può essere determinato supponendo nel seguente modo:

$$P\left(-z_{\frac{alpha}{2}} < \frac{\sqrt{n} * \left(mean(Xn) - (k * p)\right)}{\sqrt{\left(k * p(1-p)\right)}} < z_{\frac{alpha}{2}} \right) \cong 1 - alpha$$

Prendendo la disuguaglianza, notiamo sia equivalente a…

$$-z_{\frac{alpha}{2}} \frac{\sqrt{n} * \left(mean(Xn) - (k * p)\right)}{\sqrt{\left(k * p(1-p)\right)}} < z_{\frac{alpha}{2}} \rightarrow \left[\frac{\sqrt{n} * \left(mean(Xn) - (k * p)\right)}{\sqrt{\left(k * p(1-p)\right)}} \right]^2 < z^2_{\frac{alpha}{2}}$$

$$\rightarrow \frac{n * \left(mean(Xn) - (k * p)\right)^2}{k * p(1-p)} < z^2_{\frac{alpha}{2}}$$

…il che è, a sua volta, uguale a…

$$k\left(nk + z^2_{\frac{alpha}{2}}\right)p^2 - k\left(2n\, mean(Xn) + z^2_{\frac{alpha}{2}}\right)p + n\, mean(Xn)^2 < 0$$

Volendo paragonare la disuguaglianza a quella di Bernoulli, notiamo che le disuguaglianze sono praticamente identiche se non per k, il quale ricordiamo raffiguri il numero di prove. Volendo essere esplicativi, la disuguaglianza binomiale è identica a quella di Bernoulli siccome quest'ultima possiede anch'essa k, ma viene omessa siccome viene effettuata un'unica prova (k = 1).

$$p^2\left(n + z^2_{\alpha/2}\right) - p\left(2\,n\,\overline{x}_n + z^2_{\alpha/2}\right) + n\,\overline{x}^2_n < 0.$$

Infine, le radici dell'equazione possono essere calcolate utilizzando la funzione che R mette a disposizione, polyroot. I coefficienti del polinomio, infine, sono i seguenti:

$$a_2 = k\left(nk + z^2_{\frac{alpha}{2}}\right); \quad a1 = -k\left(2n\,mean(Xn) + z^2_{\frac{alpha}{2}}\right); \quad a_0 = n\,mean(Xn)^2$$

È ora, quindi, di relazionare il nostro problema alla disequazione appena vista, per poi ricavare una stima dell'intervallo di confidenza.

Ricordiamo, quindi, che in un minuto la minigun spara ben 600 colpi e l'esecuzione di 50 esperimenti ci ha permesso di ricavare un dataset composto dal numero dei colpi andati a segno attivando l'effetto del totem con successo. Si voglia dare un grado di confidenza pari a 0.95, quindi alpha avrà valore 0.05, utilizzeremo R per ricavare gli estremi dell'intervallo di confidenza.

Con numberOfElements indicheremo la dimensione del campione, quindi 50; con kparam indicheremo il numero di colpi per ogni esperimento.

```
numberOfElements <- 50
kparam <- 600
alpha <- 1 - 0.95
zalpha <- qnorm(1-alpha/2, mean = 0, sd = 1)
medCamp <- sum(data)/numberOfElements

a2 <- kparam * (numberOfElements * kparam + zalpha^2)
a1 <- -kparam * (2 * numberOfElements * medCamp + zalpha^2)
a0 <- numberOfElements * medCamp^2
polyroot(c(a0, a1, a2))
```

Gli estremi dell'intervallo i seguenti rispettivi valori:

```
> a2 <- kparam * (numberOfElements * kparam + zalpha^2)
> a1 <- -kparam * (2 * numberOfElements * medCamp + zalpha^2)
> a0 <- numberOfElements * medCamp^2
> polyroot(c(a0, a1, a2))
[1] 0.1343740+0i 0.1421853+0i
```

Nel capitolo precedente abbiamo stimato tramite il metodo dei momenti il parametro non noto p, ottenendo con 50 elementi un valore pari a 0.1382333, ergo risulta essere compreso nell'intervallo. Tale risultato lo si è avuto con un grado di confidenza pari a 0.95, ergo p è incluso nell'intervallo con una probabilità del 95%.

Supponiamo, invece, di voler ridurre tale probabilità al 70%, considerando, quindi, un grado di confidenza pari a 0.70: l'intervallo prodotto ha rispettivi valori 0.1361810-0i e 0.1403116+0i, ergo la stima di p è ancora inclusa nell'intervallo.

Notiamo, quindi, che ad un valore basso del grado di confidenza corrisponde un restringimento dell'intervallo, comportando una probabilità più bassa nella quale p dovrebbe rientrare nell'intervallo. In sostanza, quindi, maggiore è il grado di confidenza e maggiore sarà la probabilità che il valore non noto sarà compreso in tale intervallo.

Capitolo 5

Verifica delle ipotesi

Le aree più importanti dell'inferenza statistica sono la stima dei parametri e la verifica delle ipotesi. La verifica viene utilizzata in molti ambiti reali, soprattutto quando vi sono da fare indagini di mercato o indagini sperimentali e industriali.

In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x; \vartheta)$, un'ipotesi su di un parametro non noto ϑ della popolazione ed un campione casuale X1,X2, . . . ,Xn estratto dalla popolazione. Occorre in primo luogo precisare il significato di ipotesi statistica.

Un'ipotesi statistica è un'affermazione o una congettura sul parametro non noto ϑ.
Se l'ipotesi statistica specifica completamente $f(x; \vartheta)$ è detta **ipotesi semplice**, altrimenti è chiamata **ipotesi composita**. Per denotare un'ipotesi statistica useremo il carattere H seguito dai due punti e successivamente dall'affermazione che specifica l'ipotesi.

L'ipotesi soggetta a verifica viene in genere denotata con H0 e viene chiamata **ipotesi nulla.** Si chiama test di ipotesi il procedimento o regola con cui si decide, sulla base dei dati del campione, se accettare o rifiutare H0. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa. Questa proposizione prende il nome di **ipotesi alternativa** ed è di solito indicata con H1. L'ipotesi nulla, cioè l'ipotesi soggetta a verifica, si ha quando $\vartheta \in \Theta 0$ e l'ipotesi alternativa si ha quando $\vartheta \in \Theta 1$ e si scrive

$$H0: \vartheta \in \Theta 0 \text{ e } H1: \vartheta \in \Theta 1$$

Avendo denotato con $\Theta 0$ e $\Theta 1$ due sottoinsiemi disgiunti dello spazio Θ dei parametri.
Il problema della verifica delle ipotesi consiste nel determinare un test ψ che permetta di suddividere, mediante opportuni criteri, l'insieme dei possibili campioni, ossia l'insieme delle n–ple (x1, x2, . . . , xn) assumibili dal vettore aleatorio X1,X2, . . . ,Xn, in due sottoinsiemi: una **regione di accettazione** A dell'ipotesi nulla ed una **regione di rifiuto** R dell'ipotesi nulla. Il test ψ può allora essere così formulato: accettare come valida l'ipotesi nulla se il campione osservato (x1, x2, . . . , xn) $\in$ A e rifiutare l'ipotesi nulla se (x1, x2, . . . , xn) $\in$ R.
Nel caso si verifichi che l'ipotesi nulla sia falsa, l'ipotesi alternativa sarà vera e viceversa.
Spesso si usa dire che l'ipotesi H0 va verificata in alternativa all'ipotesi H1.
Nel seguire questo tipo di ragionamento si può incorrere in due tipi di errori:

- **rifiutare l'ipotesi nulla H0** nel caso in cui essa risulti vera: viene commesso un **errore di tipo** I e la probabilità di commettere questo errore è denotata con α.
$$\alpha(\vartheta) = P(\text{rifiutare H0}|\vartheta), \vartheta \in \Theta 0;$$
- **accettare l'ipotesi nulla H0** nel caso in cui essa risulti falsa: viene commesso un **errore di tipo II** e la probabilità di commettere questo errore è denotata β.
$$\beta(\vartheta) = P(\text{accettare H0}|\vartheta), \vartheta \in \Theta 1.$$

Per campioni casuali di fissata ampiezza, se si diminuisce la probabilità di commettere un errore del tipo I aumenta quella di commettere un errore di tipo II e viceversa. Quindi, siccome non è possibile minimizzare entrambe le probabilità, quello che si fa è fissare la probabilità di commettere un errore di tipo I, con un valore piccolo, e cercare un test ψ che minimizzi la probabilità di commettere un errore di tipo II. Viene fissata l'errore di tipo I perché solitamente, quando vengono formulate le ipotesi, omettere questo tipo di errore risulta essere più grave della tipologia II in quanto corrisponde a rifiutare il vero. Solitamente la probabilità di commettere un errore di tipo I si sceglie uguale a 0.05, 0.01, 0.001 ed il test viene rispettivamente detto statisticamente significativo, statisticamente molto significativo e statisticamente estremamente significativo. Più piccolo è il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla.

I test statistici sono di due tipi:
- test bilaterali (detti anche test bidirezionali);

$$\mathbf{H}_0 : \vartheta = \vartheta_0$$
$$\mathbf{H}_1 : \vartheta \neq \vartheta_0,$$

- test unilaterali (detti anche test unidirezionali).
Test unilaterale sinistro

$$\mathbf{H}_0 : \vartheta \leq \vartheta_0$$
$$\mathbf{H}_1 : \vartheta > \vartheta_0$$

Test unilaterale destro

$$\mathbf{H}_0 : \vartheta \geq \vartheta_0$$
$$\mathbf{H}_1 : \vartheta < \vartheta_0,$$

I test unilaterale sinistro e test unilaterale destro sono testati avendo fissato a priori un livello di significatività α.
Passiamo alla verifica delle ipotesi sul campione da noi generato.

Test unilaterali e bilaterali

Consideriamo la nostra popolazione Binomiale, siamo interessati a costruire dei test unilaterali e bilaterali per il valore medio E(X)= μ.

Il test bilaterale può essere formulato come segue:

H0: μ= μ0 H1: μ ≠ μ0

Mentre per il test unilaterale sinistro:

H0: μ≤ μ0 H1: μ> μ0

E per il test unilaterale destro:

H0: μ ≥ μ0 H1: μ< μ0

Avendo fissato a priori un livello di significatività α. Essendo μ0= k*p0 e $\sigma^2$0= k p0(1 – p0), nei test unilaterali e bilaterali occorre considerare:

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - k*p0}{\sqrt{\frac{k*p0(1-p0)}{n}}}$$

Considerando il nostro campione, costituito da k prove con k=600 e composto da una lunghezza n con n=50. Nel capitolo 4 abbiamo mostrato che una stima dell'intervallo di confidenza di grado 1-α=0.95, con alpha=0.05 per p è (0.134,0.142).

Vogliamo verificare l'ipotesi H0: k*p ≥ k*p0 con p0=0.14, quindi H0: k*p≥84, e in alternativa assegniamo ad H1: k*p<k*p0, quindi H1: k*p< 84, con un livello di significatività α=0.05. Occorre considerare un test unilaterale destro. Utilizzando R si ha:

```
> # left unilateral test
> p0=0.14
> alpha<-0.05
> qnorm(1-alpha,mean=0,sd=1)
[1] 1.644854
> n
[1] 50
> (medCamp-kparam*p0)/sqrt((kparam*p0*(1-p0))/n)
[1] -0.8818648
```

Notiamo che zα=1.644854 e zos=-0.8818648 e cade nella regione di rifiuto.
Occorre quindi rifiutare l'ipotesi nulla che k*p≥ 84 con un livello di significatività del 5%.

Ora proviamo invece con p0=0.120. Ipotizziamo H0:k*p0 ≤ 72 e assegniamo a H1:k*p0>72 con un livello di significatività α=0.05.

Occorre considerare un test unilaterale sinistro. Utilizzando R si ha:

```
> #right unilateral test
> p0=0.12
> alpha<-0.05
> qnorm(alpha,mean=0,sd=1)
[1] -1.644854
> n
[1] 50
> (medCamp-kparam*p0)/sqrt((kparam*p0*(1-p0))/n)
[1] 9.718399
```

Notiamo che -zα=-1.644854 e zos=9.718399 e cade nella regione di accettazione.

Capitolo 6

Criterio del chi-quadrato

Con il criterio del chi-quadrato è possibile verificare che un certo campione, descritto da una variabile aleatoria X, sia caratterizzato da una funzione di distribuzione FX(x) con k parametri non noti da stimare. Denotiamo con H0 l'ipotesi soggetta a verifica (ipotesi nulla) e con H1 l'ipotesi alternativa. Il test chi-quadrato di misura α mira a verificare l'ipotesi nulla:

H0: X ha una funzione di distribuzione Fx(x) (avendo stimato k parametri non noti in base al campione)

In alternativa all'ipotesi

H1: X non ha una funzione di distribuzione Fx(x)

Dove α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera. Occorre determinare un test ψ di misura α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale (o a due code). Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi I1, I2, ..., Ir in modo che, a seconda della distribuzione ipotizzata, pi rappresenti la probabilità che la variabile aleatoria assuma un valore appartenente a Ii (i=1,2,…,r) .

Si calcola poi la quantità

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - n\,p_i}{\sqrt{n\,p_i}} \right)^2 .$$

Il criterio del chi-quadrato si basa sulla seguente statistica

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - n\,p_i}{\sqrt{n\,p_i}} \right)^2 ,$$

dove Ni è la variabile aleatoria che descrive il numero degli elementi del campione

casuale X1,X2, . . . ,Xn (costituito da n variabili aleatorie osservabili, indipendenti e identicamente distribuite con la stessa legge di probabilità FX(x) della popolazione) che cadono nell'intervallo Ii (i = 1, 2, . . . , r).

Se la variabile aleatoria X ha una funzione di distribuzione FX(x) con k parametri non noti, si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi–quadrato con r−k−1 gradi di libertà. Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta

$$\min(np_1, np_2, \ldots, np_r) \geq 5.$$

Per un campione sufficientemente numeroso di ampiezza n, il test chi–quadrato bilaterale di misura α è il seguente:

- *si accetti l'ipotesi* H_0 *se* $\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$,
- *si rifiuti l'ipotesi* H_0 *se* $\chi^2 < \chi^2_{1-\alpha/2,r-k-1}$ *oppure* $\chi^2 > \chi^2_{\alpha/2,r-k-1}$

dove $\chi^2_{\alpha/2,r-k-1}$ *e* $\chi^2_{1-\alpha/2,r-k-1}$ *sono soluzioni delle equazioni:*

$$P(Q < \chi^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}, \qquad P(Q < \chi^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}.$$

Considerando la nostra variabile aleatoria binomiale

$$px(x) = \binom{x}{k} p^x q^{x-k} \text{ con (x=0,1 , …).}$$

```
> #chi square
> data
 [1]  84  79  76  77  75  87  84  90  79  86  91  84  83  80  81  85  82  84  81  83  78 105  81
[24]  68  88  73  76  86  91  86  94  77  91  87  96  85  85  80  82  76  79  74  78  73  85  81
[47]  93  89  77  82
> n
[1] 50
> freq<-table(data)
> freq
data
 68  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  93  94  96 105
  1   2   1   1   3   3   2   3   2   4   3   2   4   4   3   2   1   1   1   3   1   1   1   1
> length(freq)
[1] 24
> p<-numeric(4)
> for(i in 1:4)
+   p[i]<-qbinom(0.2*i,size=600,prob=stima50)
> p
[1] 76 81 85 90
```

Supponiamo di considerare 5 sottoinsiemi I1, I2, . . . , I5, utilizziamo qbinom per ottenere gli intervalli con vettore di probabilità 0.2*i, con i(1,2,3,4), size=600 e la probabilità di successo prob=0.1382333. Otteniamo gli intervalli I1=(0,76) I2=[76,81) I3=[81,85) I4=[85,90) I5=[90,600). Ora andiamo a calcolare le probabilità associate agli intervalli:

```
temp_value<-pbinom(76,size=600,prob=stima50)
temp_value

temp_value1<-dbinom(77, size = 600, prob=stima50);
temp_value1<-temp_value1+dbinom(78, size = 600, prob=stima50);
temp_value1<-temp_value1+dbinom(79, size = 600, prob=stima50);
temp_value1<-temp_value1+dbinom(80, size = 600, prob=stima50);
temp_value1<-temp_value1+dbinom(81, size = 600, prob=stima50);
temp_value1

temp_value2<-dbinom(82, size = 600, prob=stima50);
temp_value2<-temp_value2+dbinom(83, size = 600, prob=stima50);
temp_value2<-temp_value2+dbinom(84, size = 600, prob=stima50);
temp_value2<-temp_value2+dbinom(85, size = 600, prob=stima50);
temp_value2

temp_value3<-dbinom(86, size = 600, prob=stima50);
temp_value3<-temp_value3+dbinom(87, size = 600, prob=stima50);
temp_value3<-temp_value3+dbinom(88, size = 600, prob=stima50);
temp_value3<-temp_value3+dbinom(89, size = 600, prob=stima50);
temp_value3<-temp_value3+dbinom(90, size = 600, prob=stima50);
temp_value3

temp_value4<-pbinom(90,size=600,prob=stima50,lower.tail = FALSE)
temp_value4
```

E otteniamo le seguenti probabilità associate agli intervalli:

```
> temp_value
[1] 0.2248963
> temp_value1
[1] 0.212963
> temp_value2
[1] 0.1860249
> temp_value3
[1] 0.1913217
> temp_value4
[1] 0.1847941
> temp_value+temp_value1+temp_value2+temp_value3+temp_value4
[1] 1
> min(primopb,temp_value,temp_value2,temp_value3,temp_value4)
[1] 0.1847941
> 50*temp_value4
[1] 9.239704
```

Come possiamo notare la somma degli intervalli equivale a 1 e, una volta trovato il minimo, abbiamo moltiplicato la probabilità minore per 50 avendo così ottenuto 9.24, valore maggiore di 5 quindi la condizione descritta precedentemente è stata soddisfatta.

Occorre ora determinare il numero di elementi del campione che cadono negli intervalli I1, I2, . . . , I5:

```
> r<-5
> nint<-numeric(r)
> nint[1]<-length(which(data<p[1]))
> nint[2]<-length(which((data>=p[1])&(data<p[2])))
> nint[3]<-length(which((data>=p[2])&(data<p[3])))
> nint[4]<-length(which((data>=p[3])&(data<p[4])))
> nint[5]<-length(which(data>=p[4]))
> nint
[1]  5 13 13 11  8
> sum(nint)
[1] 50
```

Calcoliamo ora x^2:

```
> chi2<-sum(((nint-n*totale)/sqrt(n*totale))^2)
> chi2
[1] 5.839637
```

Ossia x^2=5.84. In questo caso il numero di categorie è r = 5 e occorre porre k = 1 poiché la probabilità Binomiale contiene un parametro non noto, ossia p.

Pertanto, la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi–quadrato si ha r − k − 1 = 3 gradi di libertà e scegliendo α = 0.01 occorre calcolare $\chi^2_{\alpha/2,3}$ e $\chi^2_{1-\alpha/2,3}$ con α=0.05.

```
> r<-5
> k<-1
> alpha<-0.05
> qchisq(alpha/2,df=r-k-1)
[1] 0.2157953
> qchisq(1-alpha/2,df=r-k-1)
[1] 9.348404
```

Da cui segue che $\chi^2_{1-\alpha/2,r-k-1}$ =0.21 e $\chi^2_{\alpha/2,r-k-1}$ =9.34. Essendo 0.21 < x^2< 9.34, l'ipotesi H0 di popolazione Binomiale può essere accettata.