Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Statistica ed Analisi Dati

Variabile aleatorie Discrete

Distribuzione Geometrica

Docente

Prof.ssa

Nobile Amelia Giuseppina

Studente

De Rosa Gerardo

Matr. 0522500762

Anno Accademico 2019-2020

Sommario

Introduzione

L'**inferenza statistica** ha lo scopo di estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto. Uno dei problemi centrali dell'inferenza statistica è il seguente: si desidera studiare una popolazione **descritta** da una variabile aleatoria osservabile **X** la cui **funzione di distribuzione** ha una forma nota ma contiene un parametro $\vartheta \in \Theta$ **non noto**.
Il termine ***osservabile*** significa che si possono osservare i valori assunti dalla variabile aleatoria X e quindi il **parametro non noto** è presente soltanto nella **legge di probabilità** (funzione di distribuzione, funzione di probabilità, densità di probabilità). Ovviamente se ϑ è noto la legge di probabilità è completamente specificata. Per ottenere **informazioni sul parametro** non noto ϑ della popolazione, si può fare uso dell'inferenza statistica considerando un campione estratto dalla popolazione e effettuando su tale campione delle opportune misure.
L'inferenza statistica si basa su **due metodi** fondamentali di indagine: la **stima** dei parametri e la **verifica** delle ipotesi.

La **stima** dei parametri ha lo **scopo** di determinare i **valori non noti** dei parametri di una popolazione (come il **valore medio**, la **varianza**, …) per mezzo dei corrispondenti parametri derivati dal campione estratto dalla popolazione (come la **media** campionaria, la **varianza campionaria**, …). Si possono usare stime puntuali o stime per intervallo. Si parla di stima **puntuale** quando si stima un parametro non noto di una popolazione usando un **singolo** valore reale.
Alla stima puntuale di un parametro non noto di una popolazione spesso si preferisce sostituire un **intervallo di valori**, detto intervallo di **confidenza**, ossia si cerca di determinare in base al campione osservato due limiti entro i quali sia compreso il parametro non noto con un certo **grado di confidenza**, detto anche grado di fiducia.
La **verifica delle ipotesi** è un procedimento che consiste nel fare una congettura o un'ipotesi sul parametro non noto ϑ o sulla distribuzione di probabilità e nel **decidere**, sulla base del campione estratto se essa è **accettabile**.

Con l'impiego di R useremo variabili aleatorie per usare i due metodi d'indagine dell'inferenza statistica, ovvero la stima dei parametri e la verifica delle ipotesi.

Useremo per la nostra indagine una variabile aleatoria discreta con una funzione di ***distribuzione geometrica***.

1. Distribuzione Geometrica

Si consideri l'esperimento consistente in una successione di prove ripetute di **Bernoulli** di parametro p ∈ (0, 1). Si supponga di essere interessati all'evento

F_r = {il numero di fallimenti che precedono il primo successo `e r}

(r = 0, 1, . . .).

Dall'ipotesi di indipendenza delle prove si ricava che $P(F_r) = (1-p)^r p$. Sia Y la variabile aleatoria che descrive il numero di fallimenti che precedono il primo successo; è evidente che $P(Y = r) = P(F_r)$ per r = 0, 1, …

Una variabile aleatoria Y di funzione di probabilità

$$p_Y(y) = P(Y = x) = \begin{cases} (1-p)^y\, p, & y = 0, 1, \dots \\ 0, & altrimenti, \end{cases}$$

con 0 < p < 1 si dice avere distribuzione geometrica di parametro p. Da ciò segue immediatamente che $p_Y(r)$ è strettamente decrescente in r = 0, 1, ... Poiché

$$\sum_{r=0}^{k} p_Y(y) = p \sum_{r=0}^{k} (1-p)^r = 1 - (1-p)^{k+1},$$

la funzione di distribuzione della variabile aleatoria geometrica Y è la seguente:

$$F_Y(y) = P(Y \le y) = \begin{cases} 0, & y < 0 \\ 1 - (1-p)^{k+1}, & k \le y < k+1 \quad (k = 0, 1, \dots). \end{cases}$$

Per una variabile aleatoria geometrica Y si ha:

$$E(Y) = \frac{1-p}{p}, \qquad \mathrm{Var}(Y) = \frac{1-p}{p^2}.$$

Una proprietà della distribuzione geometrica è la seguente:

$$P(Y > r+n \,|\, Y > r) = P(Y > n),$$

con *r* e *n* interi non negativi. La formula di cui sopra esprime dunque la proprietà di assenza di memoria della distribuzione geometrica, ossia il numero di fallimenti fino al primo successo non dipende da *r*, ossia da quanto si è atteso, ma solo dal numero *n* di prove ancora da effettuarsi.

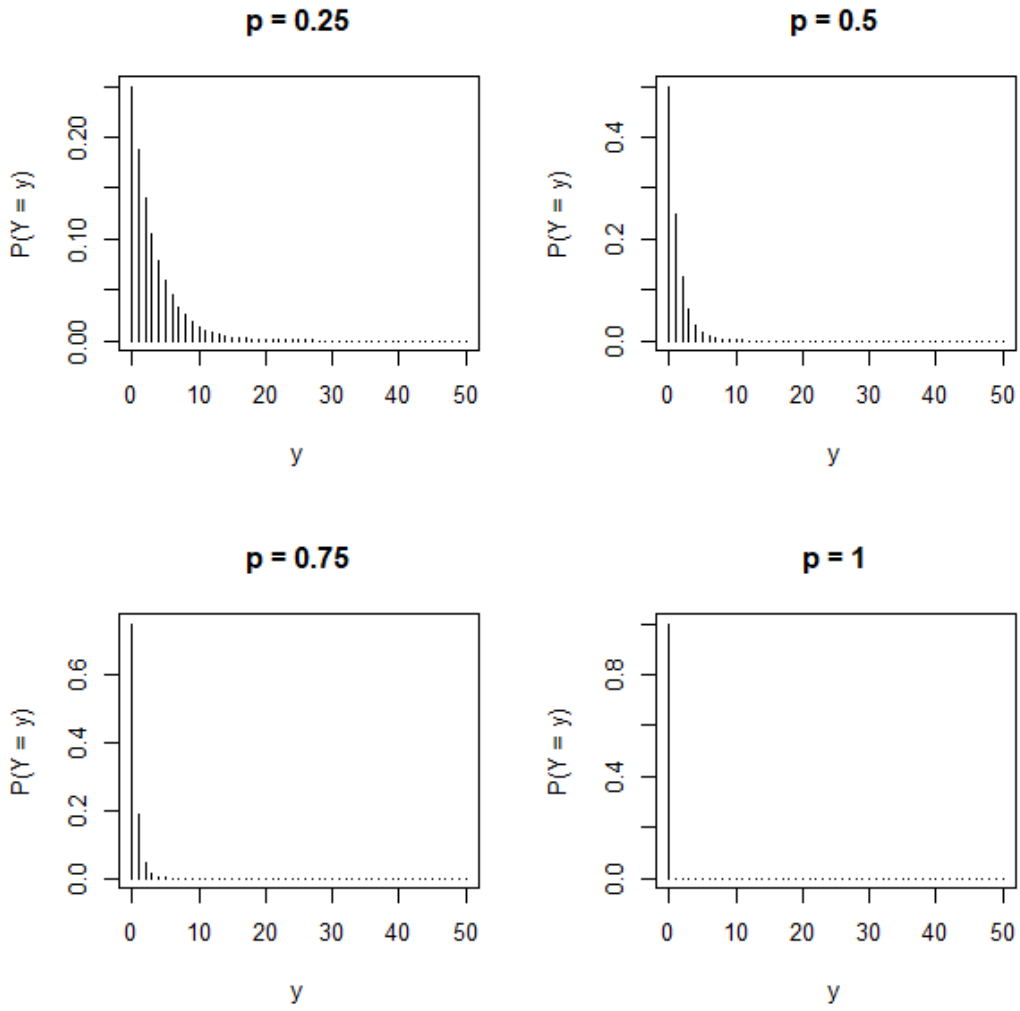
1.1 Studio della funzione di probabilità geometrica

Per il calcolo in R delle probabilità di una variabile aleatoria geometrica Y si utilizza la funzione: ***dgeom (x, prob)*** dove:

- *x* è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica considerata;
- *prob* è la probabilità di successo in ciascuna prova.

Usando i grafici abbiamo fatto un iniziale studio sulla probabilità geometrica, eseguendo l'esperimento con un numero di prove pari a **50** ed una probabilità **{0.25, 0.5, 0.75, 1}.**
Con le seguenti linee di codice mostriamo, usando la funzione ***par(),*** nella stessa finestra grafica, 4 grafici che permettono di visualizzare le funzioni di probabilità binomiale a 0.25, 0.5, 0.75 e 1.

```
y <- 0:50
par ( mfrow =c (2 ,2) )
plot (y , dgeom (y , prob =0.25) , xlab = "y" , ylab = "P(Y = y)" ,
      type ="h" , main = "p = 0.25")
plot (y , dgeom (y , prob =0.5) , xlab = "y" , ylab = "P(Y = y)" ,
      type ="h" , main = "p = 0.5")
plot (y , dgeom (y , prob =0.75) , xlab = "y" , ylab = "P(Y = y)" ,
      type ="h" , main = "p = 0.75")
plot (y , dgeom (y , prob =1) ,xlab = "y" , ylab = "P(Y = y)" ,
      type ="h" ,main = "p = 1")
```



1.2 Studio della funzione di distribuzione geometrica

Per il calcolo della funzione di distribuzione di una variabile aleatoria geometrica Y si utilizza la funzione: ***pgeom (x, prob, lower.tail = TRUE)***

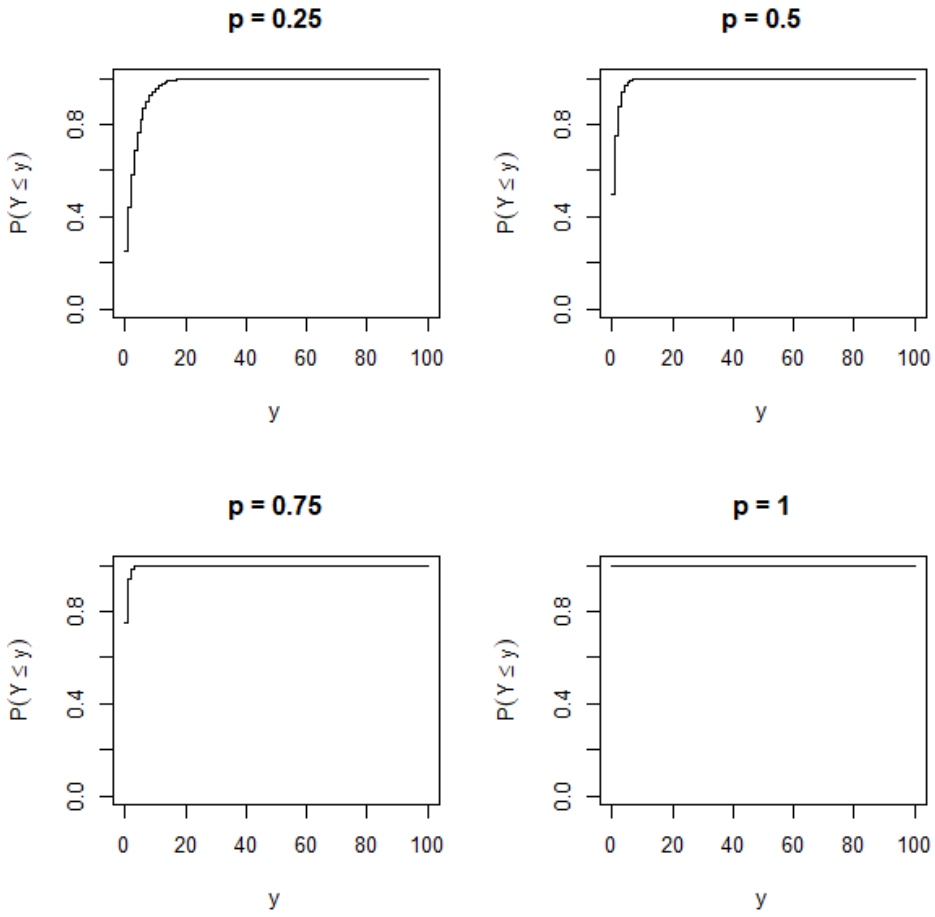
Dove:

- ***x*** è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica considerata;
- ***prob*** è la probabilità di successo in ciascuna prova;
- ***lower.tail*** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$.

La nostra analisi ha preso in esame una variabile X con **n = 100** e probabilità **p {0.25, 0.5, 0.75, 1}.**

Le seguenti linee di codice creano i 4 grafici mostrati:

```r
#FUNZIONE DI DISTRIBUZIONE GEOMETRICA
y <- 0:100
par ( mfrow =c (2 ,2) )
plot (y , pgeom (y , prob =0.25) ,xlab = "y" , ylab = expression (P(Y <= y)) , ylim = c(0 ,1) ,
      type="s" , main = "p = 0.25")
plot (y , pgeom (y , prob =0.5) ,xlab = "y" , ylab = expression (P(Y <= y)) , ylim = c(0 ,1) ,
      type="s" , main = "p = 0.5")
plot (y , pgeom (y , prob =0.75) ,xlab = "y" , ylab = expression (P(Y <= y)) , ylim = c(0 ,1) ,
      type="s" , main = "p = 0.75")
plot (y , pgeom (y , prob =1) ,xlab = "y" , ylab = expression (P(Y <= y)) , ylim = c(0 ,1) ,
      type="s" , main = "p = 1")
```



1.3 Valore medio, varianza, deviazione standard e coefficiente di variazione

È possibile calcolare il valore medio, la varianza, la deviazione standard e il coefficiente di variazione della distribuzione geometrica:

- *valore medio* = ***E(y) = (1-p)/p***
- *varianza* = ***Var(y) = (1-p)/p²***
- *deviazione standard* = ***D(y) =*** $\sqrt{Var(X)}$

- *coefficiente di variazione* = ***CV(y) = D(y)/E(y)***

In R è possibile valutare tali indici e vediamo come prendendo la probabilità di successo ***p*** pari a ***0.60:***

```
#MEDIA, VARIANZA, DEVIAZIONE STANDARD, COEFFICIENTE DI VARIAZIONE
p <- 0.60
E <- (1-p)/p
V <- (1-p)/(p*p)
D <- sqrt(V)
CV <- D/E

c(E,V,D,CV)
```

Grazie all'ultimo comando otteniamo i valori calcolati; in ordine Valore medio, Varianza, Deviazione Standard e Coefficiente di Variazione:

```
[1] 0.6666667 1.1111111 1.0540926 1.5811388
```

1.4 I quantili della distribuzione Geometrica

In R si possono calcolare anche i quantili (percentili) della distribuzione geometrica attraverso la funzione ***qgeom(z, prob)*** dove:

- ***z*** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z · 100–esimo;
- ***prob*** è la probabilità di successo in ciascuna prova. Per una distribuzione geometrica il percentile (quantile) z · 100–esimo è il più piccolo intero k tale che:

$$k \geq \frac{\log(1-z)}{\log(1-p)} - 1 \qquad (k = 0, 1, \ldots).$$

Creato un vettore **z** in cui indichiamo i **percentili**, in R si usa la funzione di cui sopra per calcolarli, per il nostro esperimento utilizzeremo una probabilità ***p*** di **0.20**.

```
#QUANTILI
z <-c (0 ,0.25 ,0.5 ,0.75 ,1)
qgeom (z , prob = 0.2)
```

Otteniamo i seguenti risultati:

```
[1]   0   1   3   6 Inf
```

I risultati mostrano che il **primo quartile** (25–esimo percentile) è ***Q1*** = 1, il secondo quartile o **mediana** (50–esimo percentile) è ***Q2*** = 3 e il terzo quartile (75–esimo percentile) è ***Q3*** = 6. Il minimo è ***Q0*** = 0 e il massimo è ***Q4*** = ∞.

1.5 Simulazione Variabile Geometrica

E' possibile simulare la variabile aleatoria geometrica in R generando una sequenza di numeri pseudocasuali mediante la funzione ***rgeom(N, prob)*** dove:

- N è la lunghezza della sequenza da generare;
- prob è la probabilità di successo in ciascuna prova.

Procediamo, quindi a generare una sequenza di ***40*** numeri pseudocasuali simulando una variabile aleatoria geometrica con ***p = 0.2*** si ha:

```
#SIMULAZIONE VARIABILE GEOMETRICA
sim <- rgeom (40 , prob =0.2)
sim
```

Otteniamo come risultato i seguenti numeri randomicamente generati:

```
 [1]  0  0 14  4  3  3 10  4  5  1  8  4  4  5  4  1  0  4  6
0  2  7 23  1  2  0  4  0  0 10  9 11  3  9  0  2  2  2 24  2
```

Utilizzando poi la funzione ***table(sim),*** possiamo calcolare il numero di occorennze per variabile:

```
> table ( sim )
sim
 0  1  2  3  4  5  6  7  8  9 10 11 14 23 24
 8  3  6  3  7  2  1  1  1  2  2  1  1  1  1
```

Ed infine utilizzando invece ***table(sim)/length(sim)*** otteniamo le frequenze relative:

```
> table(sim)/length ( sim )
sim
    0     1     2     3     4     5     6     7     8     9    10    11    14    23    24
0.200 0.075 0.150 0.075 0.175 0.050 0.025 0.025 0.025 0.050 0.050 0.025 0.025 0.025 0.025
```

2. Stima Puntuale

Uno dei problemi centrali **dell'inferenza statistica** è quello di voler studiare una **popolazione descritta** da una variabile aleatoria osservabile **X** la cui funzione di **distribuzione** ha una forma nota ma **contiene un parametro ϑ non noto**. Questo parametro è presente soltanto nella legge di probabilità (funzione di distribuzione, funzione di probabilità, densità di probabilità).

Per ottenere **informazioni** sul parametro non noto ϑ della popolazione, si può fare uso dell'inferenza statistica considerando *un* ***campione*** estratto dalla popolazione e effettuando su tale campione delle opportune misure. Affinché le conclusioni dell'inferenza statistica siano valide il campione deve essere scelto in modo tale da essere **rappresentativo della popolazione**.

Nei metodi di indagine dell'inferenza statistica si considera un campione casuale $\mathbf{X_1, X_2, \ldots, X_n}$ di **ampiezza n** estratto dalla popolazione e si cerca di ***ottenere informazioni*** sul parametro non noto facendo uso di alcune variabili aleatorie, che sono ***funzioni misurabili*** *del campione* casuale, dette **statistiche** e **stimatori**.

Una **statistica** $t(X_1, X_2, \ldots, X_n)$ è una **funzione** misurabile e osservabile **del campione** casuale $X_1, X_2, \ldots, X_n$. Essendo la statistica osservabile, i *valori da essa assunti* ***dipendono soltanto dal campione*** osservato $(x_1, x_2, \ldots, x_n)$ estratto dalla popolazione e i parametri non noti sono **presenti soltanto nella funzione** di distribuzione della statistica.

Un $\widehat{\Theta}$ ***stimatore*** $= t(X_1, X_2, \ldots, X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \ldots, X_n$ i cui **valori** possono essere *usati per* ***stimare*** *un* ***parametro non noto*** della popolazione. I valori assunti da tale stimatore sono detti ***stime*** *del parametro non noto*.

2.1 Metodi per la ricerca di stimatori

Sia $X_1, X_2, \ldots, X_k$ un campione casuale estratto da una popolazione con funzione di probabilità $f(x; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ dove $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ denotano i parametri non noti della popolazione. Lo scopo del decisore, dopo aver osservato i valori assunti dal campione casuale, è quello di stimare i parametri non noti della popolazione. I principali metodi di stima puntuale dei parametri sono ***il metodo dei momenti*** e ***il metodo della massima verosimiglianza***.

2.1.1 Metodo dei Momenti

Uno dei metodi più antichi di stima è il metodo dei momenti, per definirlo bisogna parlare prima dei **momenti campionari**.

Un momento campionario **r–esimo** relativo ai valori osservati $(x_1, x_2, \ldots, x_n)$ del campione casuale è il valore:

$$M_r(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i^r \qquad (r = 1, 2, \ldots)$$

Il **momento** campionario r–esimo è la media aritmetica delle **potenze r–esime** delle n osservazioni effettuate sulla popolazione. Se **r = 1** il momento campionario **M1**(x_1, x_2, . . ., x_n) coincide con il valore osservato della **media campionaria X**, ossia M1 = (x_1, x_2, . . ., x_n)/n.

Se esistono **k** parametri *da stimare*, il **metodo dei momenti** consiste **nell'uguagliare** i *primi **k** momenti della **popolazione*** in esame con i *corrispondenti **momenti** del **campione** casuale*. Quindi, se i primi k momenti esistono e sono finiti, tale metodo consiste nel risolvere il sistema di k equazioni:

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

Le **incognite del sistema** sono i parametri ϑ_1, ϑ_2, . . ., ϑ_k; affinché il metodo dei momenti sia utilizzabile occorre che il **sistema** ammetta un'unica soluzione.

Le **stime** dei parametri **ottenute** con tale metodo **dipendono** dal **campione** osservato (x_1, x_2, . . ., xn) e quindi al **variare** dei possibili **campioni** osservati si ottengono gli **stimatori** $\widehat{\Theta}$ **1,** $\widehat{\Theta}$ **2, …,** $\widehat{\Theta}$ **k** dei parametri non noti della popolazione, detti *stimatori del metodo dei momenti*. Alcune volte per ottenere tali stimatori è necessario utilizzare un numero maggiore di equazioni rispetto al numero dei parametri non noti da stimare.

Con il metodo dei momenti lo stimatore del parametro ***p*** di una popolazione geometrica descritta da una variabile aleatoria ***Y ∼ BN(1, p)*** con funzione di probabilità:

$$p_Y(y) = (1-p)^y\, p, \qquad y = 0, 1, \ldots \qquad (0 < p < 1).$$

Poiché ***E(Y) = (1 − p)/p***, ponendo ***ϑ = (1 − p)/p*** si ha:

$$\widehat{\vartheta} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \text{ossia} \quad \widehat{p} = \frac{1}{1 + \overline{x}}.$$

Il metodo dei momenti fornisce quindi come stimatore di ***ϑ = (1−p)/p*** la media campionaria X.

Consideriamo, ad esempio, un campione **campgeom** di ampiezza 30, generato randomicamente contenente come risultati il numero di fallimenti prima di ottenere il primo successo in lanci ripetuti di una moneta si ha:

```r
#METODO DEI MOMENTI
campgeom <-c (7 , 2, 4, 0, 1, 1, 0, 2, 5, 3, 1, 0, 3, 1, 0,
              10 , 11, 1, 6, 8, 0, 4, 3, 7 ,15 , 9, 1, 0, 14, 3)

stimap <-1/ (1+ mean ( campgeom ))
stimap
```

```
[1] 0.1910828
```

2.1.2 Metodo della massima verosimiglianza

Questo metodo è il più importante per la stima dei parametri non noti ed è di solito preferito al metodo dei momenti. Innanzitutto dobbiamo introdurre la **funzione di verosimiglianza**.

La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k; x_1, x_2, \ldots, x_n)$ del campione osservato $(x_1, x_2, \ldots, x_n)$ è la **funzione di probabilità congiunta** (nel caso di popolazione *discreta*)

del campione casuale $X_1, X_2, \ldots, X_n$, ossia:

$$L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k; x_1, x_2, \ldots, x_n)$$
$$= f(x_1; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)\, f(x_2; \vartheta_1, \vartheta_2, \ldots, \vartheta_k) \cdots f(x_n; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)$$

Il **metodo** della massima verosimiglianza **cerca** di determinare **da quale funzione** di *probabilità congiunta* è più **plausibile** che **provenga** il **campione** osservato ($x_1, x_2, \ldots, x_n$).

Pertanto si cercano di **determinare** i valori $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$ che **rendono massima** la funzione di verosimiglianza e che quindi offrano la *migliore spiegazione* del campione osservato (x1, x2, . . ., xn).

I *valori che massimizzano* la funzione di verosimiglianza vengono chiamati ***stime di massima verosimiglianza***. Le stime calcolate **dipenderanno** quindi dal campione osservato, quindi al variare del campione, **cambieranno** anche gli stimatori di questi ultimi. Questi **stimatori** $\hat{\Theta}$ **1,** $\hat{\Theta}$ **2, . . .,** $\hat{\Theta}$ **k** vengono chiamati *stimatori di massima verosimiglianza*.

Per la geometrica, lo stimatore è corretto con varianza minima e consistente per (1 − p)/p.

2.2 Proprietà degli stimatori

In generale esistono molti stimatori che possono essere utilizzati per stimare il parametro non noto di una popolazione. Occorre quindi definire delle **proprietà** di cui può più o meno godere uno stimatore; vediamone alcune:

- **Corretto** o non distorto;
- Più efficace di un altro;
- Corretto e con **varianza uniformemente minima**;

- Asintoticamente corretto;
- **Consistente**.

Uno stimatore del parametro non noto della popolazione è ***corretto*** se il valore **medio** dello **stimatore** è **uguale** al corrispondente **parametro non noto** della popolazione.

Gli stimatori che abbiamo visto per il **metodo dei momenti** che per quello della **massima verosimiglianza**, entrambi erano ***corretti*** perché il parametro non noto (**1-p)/p** era uguale alla **media campionaria.**

Dato che esistono molti stimatori per un parametro non noto di una popolazione, c'è bisogno di definire dei **criteri per confrontare** più stimatori dello **stesso** parametro.

Una misura molto importante è **l'errore quadratico medio**, che fornisce una misura di quanto si **discosta** lo stimatore dal **parametro** non noto.

L'errore quadratico medio è la quantità:

$$MSE(\widehat{\Theta}) = E\left[(\widehat{\Theta} - \vartheta)^2\right]$$

Per scegliere lo **stimatore migliore** del parametro non noto, bisogna scegliere lo stimatore con il **più piccolo** errore quadratico medio ***per ogni*** *valore ammissibile di* ***ϑ***.

Situazioni in cui esiste uno **stimatore migliore** di tutti gli altri si verificano **raramente** e spesso sono poco interessanti. La ricerca dello stimatore con errore quadratico uniformemente minimo deve essere quindi effettuata in opportune **classi** come, ad esempio, nella classe degli stimatori corretti.

Se $\widehat{\Theta}$ è uno stimatore corretto del parametro ϑ, allora:

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

Se si restringe quindi la ricerca alla classe degli stimatori corretti del parametro non noto ϑ, il problema del decisore consiste nel determinare in tale classe uno stimatore con varianza uniformemente minima.

3. Intervalli di confidenza

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore reale) spesso si preferisce sostituire un **intervallo di valori**, detto intervallo di confidenza (o intervallo di fiducia), ossia si cerca di determinare in base ai dati del campione, **due limiti** (uno inferiore ed uno superiore) entro i quali **sia compreso** il parametro non noto con un certo coefficiente di confidenza (**grado di fiducia**).

Sia $X_1, X_2, \ldots, Xn$ un campione casuale di ampiezza *n* estratto da una popolazione con **funzione di probabilità** f(x; ϑ), dove ϑ denota il parametro non noto della popolazione.

Denotiamo con $\underline{C}_n = g1(X_1, X_2, \ldots, X_n)$ e con $\bar{C}_n = g2(X_1, X_2, \ldots, X_n)$ **due statistiche** (funzioni osservabili del campione casuale) che soddisfino la condizione $\underline{C}_n < \bar{C}_n$, che godono della proprietà che ***per ogni*** *possibile fissato campione osservato* x = (x_1, x_2, . . ., x_n) risulti *g1(x) < g2(x)*.

Quindi fissato un coefficiente di confidenza **1 − α** (0 < α < 1), se è possibile scegliere le statistiche $\underline{C}_n$ e $\bar{C}_n$ in modo tale che:

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

allora si dice che $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza (intervallo di fiducia) di grado **1 − α** per ϑ. Inoltre, le statistiche $\underline{C}_n$ e $\bar{C}_n$ sono dette **limite** inferiore e superiore dell'intervallo di confidenza.

Se g1(x) e g2(x) sono i **valori assunti** dalle statistiche per il campione osservato x = (x_1, x_2, . . ., x_n), allora l'intervallo **(g1(x), g2(x))** è detto **stima dell'intervallo di confidenza** di grado 1 − α per ϑ ed i punti finali g1(x) e g2(x) di tale intervallo sono detti rispettivamente **stima del limite inferiore e stima del limite superiore** dell'intervallo di confidenza.

In generale esistono **molti intervalli** di confidenza dello **stesso grado** 1 − α per un parametro non noto ϑ della popolazione. La **scelta** dell'intervallo deve essere fatta sulla base di alcune **proprietà statistiche**. Alcune proprietà che si desiderano sono che la **lunghezza** dell'intervallo di confidenza:

$$L(X_1, X_2, \ldots, X_n; 1 - \alpha) = \overline{C}_n - \underline{C}_n$$

sia il più **piccolo** possibile oppure che la sua **lunghezza media** lo sia.

3.1 Metodo Pivotale

Un metodo per la **costruzione** degli intervalli di confidenza è il metodo pivotale. Tale metodo consiste essenzialmente nel **determinare** una **variabile aleatoria di pivot** $\gamma(X_1, X_2, \ldots, Xn : \vartheta)$ che dipende dal campione casuale e dal parametro non noto ϑ e la **cui funzione di distribuzione** non contiene il parametro da stimare.

Tale variabile aleatoria non è una statistica poiché dipende dal parametro non noto ϑ e quindi non è osservabile.

Per ogni fissato coefficiente **α** $(0 < \alpha < 1)$ siano α_1 e α_2 $(\alpha_1 < \alpha_2)$ due valori dipendenti soltanto dal coefficiente fissato α tali che per ogni $\vartheta \in \Theta$ si abbia:

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

Se per ogni possibile campione osservato $(x_1, x_2, \ldots, xn)$ e per ogni $\vartheta \in \Theta$, si riesce a dimostrare che:

$$\alpha_1 < \gamma(\mathbf{x}; \vartheta) < \alpha_2 \iff g_1(\mathbf{x}) < \vartheta < g_2(\mathbf{x})$$

Allora la precedente equazione equivale a:

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

Denotando con $\underline{C}_n = g_1(X_1, X_2, \ldots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \ldots, X_n)$, possiamo affermare che $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado 1- α per il parametro non noto ϑ della popolazione.

4. Intervalli di fiducia approssimati

Per campioni di grandi dimensioni (*n > 30)* **non** si possono effettuare delle stime **precise**, quindi quello che si fa è **costruire degli intervalli** di confidenza approssimati utilizzando il *teorema centrale di convergenza.* Infatti, se *X* denota la variabile aleatoria che descrive la popolazione con **E(X) =** μ e **Var(X) =** σ^2 (supposti entrambi finiti) e con (*X1,X2, . . . ,Xn*) il campione casuale, il **teorema** centrale di convergenza **afferma** che la variabile aleatoria:

$$Z_n = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

converge in distribuzione ad una variabile aleatoria **normale standard**. Pertanto per campioni di ampiezza elevata possiamo applicare il **metodo pivotale in forma approssimata** supponendo che:

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

Analizzeremo il seguente caso:

4.1 Intervallo di confidenza per il parametro p di una popolazione Geometrica

Prendiamo in esame una popolazione Geometrica descritta da variabile aleatoria X con funzione di probabilità: $P(X = x) = p^x (1-p)^{1-x} \qquad (x = 0, 1)$

Ricordiamo che il **valore medio** di una variabile aleatoria Geometrica **E(X) = (1-p)/p** e che la varianza è **Var(X) = *(1-p)/p²***, applicando il **teorema centrale di convergenza** abbiamo che la variabile aleatoria **converge** in distribuzione ad una **variabile aleatoria normale standard**.

Sia **X** la variabile aleatoria che descrive il numero associato alla prima vittoria riportata; la distribuzione di X è geometrica di parametro p, dove p, rappresenta la probabilità di vittoria in una singola estrazione. Se si effettuano **150** osservazioni di X, si nota che $\bar{x}_{150}$ = 47; sulla base di questi dati si vuole fare una **stima dell'intervallo** di confidenza di grado **1-α = 0 .90**.

Inoltre, essendo **α = 0.10**, si ha **α /2 = 0.05**. In R vediamo:

```
> #INTERVALLI DI CONFIDENZA
> alpha <- 1 - 0.9
> qnorm (1 - alpha /2, mean = 0, sd = 1)
[1] 1.644854
> zalpha <- qnorm (1 - alpha /2, mean = 0, sd = 1)
> n <- 150
> medcamp <- 47
>
> a2 <-n * medcamp ^2
> a1 <- -(2 *n * medcamp - zalpha ^2)
>
> a0 <-n - zalpha ^2
> polyroot (c(a0 ,a1 ,a2))
[1] 0.01844559-0i 0.02409944+0i
```

L'intervallo di confidenza per **p = (0.018, 0.024)**. Da notare che la stima puntuale della probabilità, ossia **p = 1/$\bar{x}_{150}$ = 0.0212** è compresa nell'intervallo.

Possiamo avere gli stessi risultati ma usando i comandi:

```
> (1 / medcamp ) -( zalpha / medcamp ^2) * sqrt ( medcamp *( medcamp -1) /n)
[1] 0.01844967
> (1 / medcamp ) +( zalpha / medcamp ^2) * sqrt ( medcamp *( medcamp -1) /n)
[1] 0.02410352
```

Da questi risultati possiamo dedurre che la ***stima** di probabilità è nella media*, non è né troppo bassa, né troppo alta.

5. Verifica delle Ipotesi

Come abbiamo già detto, le aree più importanti dell'**inferenza statistica** sono la **stima** dei parametri e la **verifica** delle ipotesi. Quest'ultima interviene spesso nelle ricerche di mercato, nelle indagini sperimentali e industriali, nei sondaggi di opinione, nelle indagini sulle condizioni sociali degli abitanti di una città o di una nazione.

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 f(x; ϑ), un'**ipotesi** su di un parametro **non noto** della popolazione ed un **campione** casuale $X_1, X_2, \ldots, X_n$ estratto dalla popolazione.

Diamo la definizione 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; ϑ)** è detta **ipotesi semplice**, altrimenti è chiamata ipotesi **composta**.

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 $\mathbf{H_0}$ e viene chiamata ipotesi **nulla**. Si chiama **test di ipotesi** il procedimento o regola con cui si decide se **accettare** o **rifiutare** $\mathbf{H_0}$.

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 $\mathbf{H_1}$.

Il **problema** della verifica delle ipotesi consiste nel determinare un **test** ν che permetta di **suddividere**, mediante opportuni criteri, l'insieme dei **possibili campioni**, ossia l'insieme delle ennuple ($x_1, x_2, \ldots, xn$) assumibili dal vettore aleatorio $X_1, X_2, \ldots, X_n$, in **due** sottoinsiemi: una regione di accettazione **A** dell'ipotesi **nulla** ed una regione di rifiuto **R** dell'ipotesi **nulla**. Il test ν può allora essere così formulato: **accettare** come valida l'ipotesi **nulla** se il campione osservato ($x_1, x_2, \ldots, xn$) $\in$ **A** e **rifiutare** l'ipotesi **nulla** se ($x_1, x_2, \ldots, xn$) $\in$ **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 $\mathbf{H_0}$ va verificata in alternativa all'ipotesi $\mathbf{H_1}$.

Si possono verificare due **errori** seguendo questo tipo di ragionamento; li riportiamo nel seguente schema:

	Rifiutare H_0	Accettare H_0
H_0 vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
H_0 falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

5.1 Verifica delle ipotesi per una variabile Geometrica

Un test di verifica delle ipotesi ha le seguenti **fasi**:

5.1.1 Test Bilaterale

L'ipotesi nulla afferma che $\mu = \mu_0$. In contrapposizione, l'ipotesi alternativa afferma che **μ è diverso da μ_0**:

$$\mathbf{H}_0 : \mu = \mu_0, \qquad \mathbf{H}_1 : \mu \neq \mu_0$$

Si **accetta** $\mathbf{H_0}$, se:

$$-z_{\alpha/2} < \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

Si **rifiuta** invece, se:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \qquad \text{oppure} \qquad \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

5.1.2 Test unilaterale sinistro

$$\mathbf{H}_0 : \mu \leq \mu_0, \qquad \mathbf{H}_1 : \mu > \mu_0$$

Si **accetta H_0**, se:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$$

Si **rifiuta H_0**, se:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

5.1.3 Test unilaterale destro

$$\mathbf{H}_0 : \mu \geq \mu_0, \qquad \mathbf{H}_1 : \mu < \mu_0$$

Si **accetta H_0**, se:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$$

Si **rifiuta H_0**, se:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

Vediamo ora i test con una geometrica, ricordando che **μ_0 = E(x) = $(1-p_0)/p_0$** e che **$\sigma = \sqrt{Var(X)} = \sqrt{(1-p)/p\text{^}2}$** .

5.1.4 Test Bilaterale (Geometrica)

Prendiamo un campione Geometrico con **p_0 = 0.80**, calcoliamo **μ_0 = E(x) = $(1-p_0)/p_0$ = *0.25***

- ***H_0:*** (1-p)/p = 0,25
- ***H_1:*** (1-p)/p ≠ 0,25

Con un grado di **fiducia** pari a **α = 0.20**, vediamo il seguente codice R:

```
> #VERIFICA DELLE IPOTESI
> sim <- rgeom (40 , prob =0.5)
> sim
 [1] 0 0 0 0 0 2 1 0 0 0 0 1 0 0 0 4 1 4 0 3 0 0 0 0 1 0 3 0 0 1 1 0 0 0 5 2 0 0 0 0
>
> p0<-0.8
>
> m<-mean(sim)
> m
[1] 0.725
>
> ((m-((1-p0)/p0))/(sqrt((1-p0)/(n*p*p))))
[1] 4.030509
>
> alpha<- 0.20
>
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.281552
>
> p0<-0.6
> ((m-((1-p0)/p0))/(sqrt((1-p0)/(n*p*p))))
[1] 0.35
```

In questo caso, utilizziamo un campione con ***n = 40*** e ***p= 0.5*** creato tramite la funzione ***rgeom()*** e possiamo notare come in effetti il risultato prodotto dal test con $p_0 = 0.8$, renda H_0 non accettabile poiché **4.03** è fuori dall'intervallo **(-1.28, 1.28);** scegliendo invece p_0 più vicino allo ***0.5***, in questo caso uguale a 0.6, otteniamo un valore compreso nell'intervallo e quindi H_0 diventa accettabile.

Possiamo provare anche a testare un campione creato non dalla funzione ***rgeom()*** e quindi non avente una probabilità di riferimento:

```
> #VERIFICA DELLE IPOTESI
> campgeom <-c (2 , 2, 4, 0, 1, 1, 0, 2, 5, 3, 1, 0, 8, 1, 0,
+               5 , 0, 1, 6, 8, 0, 4, 3, 7 ,4 , 9, 1, 0, 10, 3)
>
> p0<-0.25
>
> m<-mean(campgeom)
> m
[1] 3.033333
>
> ((m-((1-p0)/p0))/(sqrt((1-p0)/(n*p*p))))
[1] 0.1460593
>
> alpha<- 0.20
>
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.281552
```

In questo caso ad esempio, si nota come p_0 per essere valido debba aggirarsi vicino allo ***0.25,*** per ***alpha = 0.20***.

5.1.5 Test unilaterale sinistro (Geometrica)

Prendiamo il campione generato randomicamente in precedenza, con $\mathbf{p_0 = 0.80}$ e verifichiamo che

H_0: $p \leq 0{,}25$ o H_1: $p > 0{,}25$ sia valida:

```
> p0<-0.8
>
> m<-mean(sim)
>
> alpha<- 0.20
>
> qnorm(1-alpha,mean=0,sd=1)
[1] 0.8416212
>
> ((m-((1-p0)/p0))/(sqrt((1-p0)/(n*p*p))))
[1] 4.030509
```

In questo caso ***H_0*** non è valida, mentre è vera ***H_1***, questo poiché ***$4.03 > 0.84$*** (limite sinistro z_α), quando dovrebbe essere minore

5.1.6 Test unilaterale destro (Geometrica)

Prendiamo il campione generato randomicamente in precedenza, con $\mathbf{p_0 = 0.80}$ e verifichiamo che

H_0: $(1-p)/p \geq 0{,}25$ o H_1: $(1-p)/p < 0{,}25$ sia valida:

```
> p0<-0.8
>
> m<-mean(sim)
>
> alpha<- 0.20
>
> qnorm(alpha,mean=0,sd=1)
[1] -0.8416212
>
> ((m-((1-p0)/p0))/(sqrt((1-p0)/(n*p*p))))
[1] 4.030509
```

In questo caso otteniamo quindi come previsto (poiché il test unilaterale sinistro era negativo) esito positivo, difatti il valore ottenuto ***$4.03 \geq -0.84$ (z_α)*** è quindi maggiore uguale del limite unilaterale destro.