# DAL CAMPIONE ALLA POPOLAZIONE: LA STIMA DEI PARAMETRI

Andrea Onofri
Dipartimento di Scienze Agrarie ed Ambientali
Università degli Studi di Perugia
Versione on-line: http://www.unipg.it/ onofri/RTutorial/index.html

## Indice



## 1 Stima puntuale dei parametri

In genere, le unità sperimentali sottoposte ad un esperimento (o i risultati ottenuti da un qualunque procedimento di misura) sono solo un campione di quelle possibili, anche se il ricercatore è interessato a trarre conclusioni generiche, valide per l'intera popolazione da cui il campione è stato estratto (*stima dei parametri della popolazione*).

E' intuitivo pensare che, data una popolazione se da questa immaginiamo di estrarre a caso un campione di $n$ individui, è probabile che la media del campione sia pari alla media della popolazione da cui questo è stato estratto. Infatti gli individui intorno alla media nella popolazione di partenza sono i più frequenti e quindi sono quelli che hanno la massima probabilità di essere inclusi nel campione. E' ovvio che questo è vero se il campione è rappresentativo (cioè se è estratto a caso e sufficientemente numeroso). Questa osservazione intuitiva ci consente di affermare che dato un campione estratto casualmente da una popolazione normalmente distribuita, *la media e la deviazione standard del campione sono una stima non distorta della media e della deviazione standard della popolazione di origine*. Per la dimostrazione di questo assunto rimandiamo a pubblicazioni più specifiche, ma ricordiamo che solo la deviazione standard campionaria (cioè quella ottenuta come:

$$s = \sqrt{\frac{SS}{n-1}}$$

dove SS è la devianza del campione) è una stima corretta della deviazione standard della popolazione. Bisogna comunque notare che i reali valori dei parametri (media e deviazione standard) della popolazione di origine rimangono comunque ignoti, ma si può affermare che con la massima probabilità questi sono uguali a quelli del campione estratto.

Più in generale, dato un campione, le statistiche descrittive calcolate per questo campione (media, varianza, deviazione standard, parametri di regressione, correlazione ecc..) possono essere estrapolate alla popolazione che ha generato il campione stesso, senza che questo possa essere in qualche modo oggetto di critica. In fin dei conti è la migliore stima che abbiamo. Questo tipo di stima si definisce *stima puntuale*, perché ad ogni valore ignoto di un certo parametro della popolazione (ad es. la media) associamo una certa stima puntiforme, cioè costituita da un singolo valore.

**Esempio 1**

Da un terreno agrario è stato estratto casualmente un campione di 5 buste da 20 grammi ciascuna di terreno. Il terreno presente in ogni busta viene analizzato per conoscere il contenuto in fosforo assimilabile. I dato ottenuti sono 9, 10, 14, 16 e 13 ppm, rispettivamente per le cinque buste. Qual è il contenuto di fosforo nel terreno e qual è la sua deviazione standard (variabilità naturale del contenuto di fosforo nel terreno, errore di campionamento e di misura)?

Questo problema può essere risolto pensando che il campione da noi estratto (cinque buste) sia rappresentativo dell'intera popolazione e, di conseguenza, le statistiche descrittive del campione possono essere assunte come stime puntuali delle statistiche descrittive della popolazione.

La media delle cinque misure nel campione è pari a 12.4 ppm, mentre la deviazione standard è pari 2.88 ppm. Ne consegue che il coefficiente di variabilità è pari al 20.6%. Come abbiamo visto questi risultati possono essere estrapolati all'intera popolazione di tutte le misure possibili. Possiamo quindi concludere che il campione è estratto da un terreno il cui contenuto medio di fosforo è pari a 12.4 ppm con una deviazione standard pari a 2.88 ppm.

I reali valori di contenuto medio ed errore rimangono ignoti: le nostre conclusioni sono raggiunti solamente su base probabilistica; si tratta delle conclusione più probabili, ma non certe.

Quanto detto vale anche per la proporzione (la proporzione del campione p è una stima non distorta di $\pi$), mentre nel caso della varianza lo stimatore non distorto è la varianza campionaria (cioè quelle ottenuta dividendo la devianza per $n - 1$).

La stima puntuale è molto comoda, ma anche molto imprecisa: possibile che la popolazione intera abbia proprio la stessa media o la stessa deviazione standard del campione che noi abbiamo estratto?

Per esprimere questa incertezza è quindi necessario associare alla stima puntuale un intervallo, passando quindi alla cosiddetta stima per intervallo.

## 2 La precisione della stima e l'errore standard

Nel paragrafo precedente abbiamo affermato che la media incognita della popolazione ($\mu$) è stimata dalla media del campione. Tuttavia, il calcolo di probabilità illustrato nel capitolo precedente ci ha insegnato che, data una popolazione normale con media $\mu$ e deviazione standard $\sigma$;, se estraiamo infiniti campioni di $n$ elementi, le medie campionarie sono distribuite normalmente con media $\mu$ e deviazione standard pari alla quantità:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

detta *errore standard di una media*. Quest'ultima può essere guardato come la variabilità della media campionaria, cioè una misura dell'incertezza legata alla stima della media. In altre parole, l'errore standard è una misura dell'errore di stima.

In questo senso, l'errore standard è concettualmente molto diverso dalla deviazione standard della popolazione da cui il campione è estratto, che ne rappresenta la variabilità naturale ineliminabile. Infatti, se dalla popolazione di partenza avessimo estratto un campione di infinite misure (n = $\infty$) avremmo ottenuto una stima perfetta di $\mu$ (errore standard pari a 0), nonostante la variabilità naturale $\sigma$. In altre parole, la stima della media può essere perfetta anche se la misura è viziata da un errore (per esempio perché l'apparecchio non è perfettamente funzionante).

E' bene sottolineare ancora come l'errore standard (e quindi la precisione della stima di $\mu$) dipende sia dalla variabilità della misura, sia dal numero di repliche che effettuiamo; più precisamente, l'errore standard aumenta all'aumentare della deviazione standard e diminuisce all'aumentare del numero delle ripetizioni, annullandosi quando questo tende ad infinito.

**Esempio 2**

In un vigneto, si vuole conoscere la produzione d'uva per pianta. Non avendo tempo e risorse sufficienti per misurare tutte le pi-

ante del vigneto, si scelgono dieci piante a caso e si misura la loro produzione, che risulta pari rispettivamente a:

3.6, 4.2, 5.2, 3.4, 3.9, 4.1, 4.7, 4.2, 3.9, 3.8

La stima più probabile della produzione per pianta del vigneto è data dalla media delle misure effettuate, pari a 4.1

La variabilità della misura (che include, tra l'altro, la variabilità individuale delle viti, la variabilità della fertilità del terreno e l'errore di misura dell'operatore) può essere stimata dalla deviazione standard del campione, pari a 0.527

Come errore di stima della media possiamo prendere l'errore standard, che in R può essere calcolato con le consuete formule

```
x <- c(3.6, 4.2,  5.2,  3.4,  3.9,
       4.1,  4.7,  4.2,  3.9,  3.8)
mean(x)
[1] 4.1
se<-sqrt(var(x))/sqrt(length(x))
se
[1] 0.1666667
```

# 3 Intervalli di confidenza di una media

Nel capitolo precedente abbiamo già illustrato come il 95% delle medie campionarie sono comprese nell'intervallo $\pm$ 1.96 volte l'errore standard. Di conseguenza, se affermiamo che:

$$\mu = \bar{X} \pm 1.96 \times \sigma_{\bar{X}}$$

abbiamo una probabilità del 95% di essere nel giusto ed una probabilità d'errore del 5%.

Questo ragionamento, tuttavia, presuppone di conoscere la quantità $\sigma$ della popolazione. Più frequentemente, $\sigma$ viene stimato a partire da $s$, cioè dalla deviazione standard del campione. In questa situazione, abbiamo visto che le medie campionarie sono distribuite tra $\pm t_{\alpha,\nu}$, $\alpha$ è il grado di confidenza ricercato (ad esempio il 95%, equivalente al 5% di probabilità di errore del 5%) e $\nu$è il numero di gradi di libertà della deviazione standard del campione (n-1). Gli intervalli di confidenza della media, pertanto, possono essere costruiti in questo modo:

$$\mu = \bar{X} \pm t_{\alpha,\nu} \times s_{\bar{X}}$$

E'bene ribadire che se vogliamo usare R per il calcolo delle bande di confidenza, dobbiamo fare attenzione al valore $\alpha$, infatti se vogliamo la banda di inferenza del 95%, dobbiamo indicare un valore $\alpha$ = (1 - 0.95)/2 (distribuzione ad una coda).

In sostanza, dato un certo livello di probabilità d'errore (ad esempio il 5%), possiamo costruire un intervallo che molto probabilmente contiene il vero ed ignoto valore della media della popolazione da cui il campione è stato estratto. Più esattamente, questa affermazione è tanto probabile da lasciare solo il margine d'errore voluto.

**Caso studio 1**

Riprendendo i dati dell'Esercizio 9 abbiamo già osservato come, sulla base del campione esaminato, possiamo concludere che il valore più probabile della produzione media per pianta nel vigneto è pari a 4.1 kg. Questa stima ci lascia un po' insoddisfatti: come è possibile che la produzione per pianta di un intero vigneto sia proprio uguale a quella delle dieci piante misurate? Se ci calcoliamo allora l'intervallo di confidenza della media per un livello di probabilità $\alpha = 0.05$ otteniamo:

$$\mu = 4.1 \pm 2.262 \times 0.167 = 4.1 \pm 0.377$$

Questo ci permette di affermare che la produzione media per pianta del vigneto (quella vera, che rimane ignota) è compresa tra 4.447 e 3.723. Se il campione era effettivamente rappresentativo, possiamo avere fiducia che facendo questa affermazione non abbiamo più del 5% di probabilità d'errore.

Se volessimo essere ancora più tranquilli, potremmo calcolare l'intervallo di confidenza della media per un livello di probabilità $\alpha = 0.01$), ottenendo:

$$\mu = 4.1 \pm 3.250 \times 0.167 = 4.1 \pm 0.541$$

In questo caso possiamo affermare che la produzione media per pianta del vigneto è compresa tra 4.641 e 3.559, con una probabilità d'errore dell'1%. Come si vede, per diminuire la probabilità d'errore abbiamo dovuto allargare l'intervallo di confidenza.

In R, per il calcolo dei limiti di confidenza conviene ricorrere alla funzione anche`t.test()`, come nell'esempio seguente:

```
> t.test(x, conf.level=0.95)

        One Sample t-test

data:  x
```

```
t = 24.6, df = 9, p-value = 1.453e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.722974 4.477026
sample estimates:
mean of x
      4.1

> t.test(x, conf.level=0.99)

        One Sample t-test

data:  x
t = 24.6, df = 9, p-value = 1.453e-09
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 3.558361 4.641639
sample estimates:
mean of x
      4.1
```

# 4 L'errore standard e gli intervalli di confidenza nell'analisi di regressione

Come avrete intuito, il calcolo dell'errore standard e degli intervalli di confidenza ci consente di aggiungere alle nostre stime una banda d'incertezza; in questo modo possiamo comunque stare al riparo da errori macroscopici, anche se rimane il fatto che non potremo mai conoscere con assoluta precisione una certa caratteristica della nostra popolazione.

Lo stesso problema va affrontato nel caso dell'analisi di regressione. Come si ricorderà, eseguire una analisi di regressione in una popolazione di dati bivariata, consiste nel determinare due parametri: l'intercetta ($\beta_0$) e la pendenza ($\beta_1$) in modo da caratterizzare la retta che esprime la relazione funzionale tra le due variabili.

Anche in questo caso se non abbiamo a disposizione l'intera popolazione possiamo eseguire l'analisi di regressione su un campione rappresentativo che sia stato estratto da questa. In questo modo otterremo dei valori di intercetta ($b_0$)) e pendenza ($b_1$) che sono delle stime dei valori reali dell'intera popolazione. Anche queste stime, come nel caso della media, dovranno essere corredate dei relativi intervalli di confidenza.

Il calcolo degli intervalli di confidenza nell'analisi di regressione è un po' più complicato e verrà demandato ad R. Ricordiamo tuttavia che alla base

di questo calcolo vi è la determinazione degli errori standard delle stime, che hanno esattamente lo stesso significato dell'errore standard di una media.

**Caso studio 2**

Consideriamo il seguente campione (quattro individui):

| Num. Num. | Ricoprimento Infestanti | Produzione mais |
|---|---|---|
| 1 | 5.00 | 12.75 |
| 2 | 12.41 | 11.18 |
| 3 | 20.05 | 12.25 |
| 4 | 65.75 | 10.59 |

Eseguiamo su l'analisi di regressione con R:

```
> x<-c(5,12.41,20.05,65.75)
> y<-c(12.75,11.18,12.25,10.59)
> model<-lm(y~x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
        1         2         3         4
 0.478407 -0.885316  0.397364  0.009545

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 12.41078    0.56588  21.932  0.00207 **
x           -0.02784    0.01616  -1.722  0.22712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7651 on 2 degrees of freedom
Multiple R-Squared: 0.5973,     Adjusted R-squared: 0.396
F-statistic: 2.967 on 1 and 2 DF,  p-value: 0.2271

>
```

Per calcolare gli intervalli di confidenza possiamo utilizzare gli estrattori e le funzioni di probabilità:

```
>summary(model)$coefficients[4:4] *
        qt(0.975,df=model$df.residual)
0.06953662
>summary(model)$coefficients[3:3] *
        qt(0.975,df=model$df.residual)
2.434793
```

Il calcolo degli intervalli di confidenza è abbastanza importante, perché ci ha portato alla fine a fare un'affermazione di tipo probabilistico, che non è necessariamente vera, ma che invece è condizionata da una certa possibilità d'errore, che è comunque nota e fissata a priori, ancor prima di compiere la misurazione.

Questo modo di procedere è tipico della statistica inferenziale, nata appunto per le situazioni nelle quali non si possono avere certezze assolute, ma soltanto stime affidabile, fatto salvo un predefinito rischio d'errore.