**Immaginare un esperimento bernoulliano: cosa vuol dire estrarre un campione casuale e quali garanzie ci dà in probabilità? Qual è il ruolo della numerosità campionaria?**

Nicolò Finamore

Immaginiamo un esperimento focalizzato sullo studio della popolazione di bambini residenti in Italia in età campresa tra 6 e 10 anni dei quali siamo interessati a stimare la frequenza relativa F di bambini allergici nel periodo primaverile; è chiara l’impossibilità di effettuare l’analisi sulla totalità dei bambini in Italia. Per tale motivo, dovremo selezionare un campione che risulti eterogeneo e rappresentativo della popolazione, per consentirci di generalizzare il risultato dello studio a tutti i bambini. Il campionamento deve avvenire in maniera casuale, cioè tutti i bambini appartenenti alla popolazione devono avere la stessa probabilità di essere estratti. Disponendo di una lista numerata dei bambini appartenenti alla nostra popolazione, per la selezione può essere utile affidarci ad un software in grado di generare numero casuali.

Effettuato il campionamento possiamo descrivere il campione come vettore di variabili aleatorie (X1,… Xn) dove Xi è una variabile aleatoria dicotomica che può assumere valori: Xi=0 se l’i-esimo bambino reclutato bambino non è allergico, quindi non manifesta sintomi; Xi=1 quando è allergico e quindi manifesta sintomi. Definiamo la probabilità che il bambino estratto sia allergico, avremo quindi: P(Xi=1)= e P(Xi=0)=1-.

Il campionamento casuale ci garantisce l’uguaglianza fra , la probabilità che il bambino estratto sia allergico, ed F, ovvero la frequenza di bambini allergici nella popolazione. Il legame fra e F è vero poiché tutti i bambini sono equiprobabili, dunque è possibile applicare la definizione classica di probabilità: casi favorevoli (bambini allergici) su casi possibili (tutti i bambini). Quindi, la probabilità di estrarre un bambino allergico sarà uguale alla frequenza di bambini effettivamente allergici nella popolazione. Analogamente avremo che 1-=1-F. In altri termini la distribuzione di probabilità di Xi sarà uguale alla dstribuzione di frequenze nella popolazione. Poiché questo ragionamneto può essere ripetuto per ogni osservazione il nostro campione sarà un vettore di variabili aleatorie identicamente distribuite che assumeremo indipendenti.

Tuttavia quella stabilita è una garanzia in probabilità che dovremo traformare in una garanzia in frequenza. Ricordiamo infatti che il nostro obiettivo è che la frequenza f dei bambini allergici nel nostro campione sia simile a quella della popolazione. In questo passaggio è fondamentale il ruolo della numerosità del campione sancito dalla legge dei grandi numeri. Secondo tale legge, in presenza di n variabili aleatorie identicamente distribuite ed indipendenti, la loro media tenderà, al tendere di n ad infinito, al loro valore atteso. Abbiamo già chiarito che il nostro campione è un vettore di variabili aleatorie indipendenti ed identicamente distribuite secondo una legge Bernoulliana quindi la condizione iniziale è soddisfatta. Nel nostro caso la media delle osservazioni coincide con la frequenza relativa e il valore atteso è proprio . Avremo quindi

Quindi, se il campione è stato estratto casualmente e ha una numerosità sufficientemente elevata la frequenza dei bambini allergici nel campione tenderà alla probabilità che un bambino sia allergico, che sappiamo essere uguale alla F della popolazione, quindi la frequenza nel campione tenderà alla frequenza nella popolazione e potremmo finalmente generalizzare i risultati dell’esperimento. In simboli

**Che cosa vuol dire assumere un modello normale? Fare un esempio dove è ragionevole assumere un modello normale e chiarire perché è ragionevole.**

Federica Pitocco

Immaginiamo di voler stimare il peso alla nascita dei bambini italiani maschi nati al termine, con assenza di patologie. Si definisce una gravidanza a termine quella in cui il parto avviene tra le 37 e le 41 settimane.

Assumere un modello normale vuol dire immaginare che la distribuzione del peso nella popolazione d’interesse abbia le caratteristiche di una densità normale.

In altre parole immaginiamo di osservare il peso alla nascita di tutti i bambini appartenenti alla popolazione d’interesse e costruire un istogramma di frequenze sulla base dei dati osservati. Assumere un modello normale vuol dire ipotizzare che l’istogramma costruito abbia una forma approssimabile ad una densità normale dove μ è il peso medio e la varianza del peso nella nostra popolazione .

Ricordiamo che con il modello normale stiamo osservando una variabile continua e, poiché siamo nel continuo, assegniamo le probabilità ad intervalli di valori.

La normale è una famiglia di densità dalle caratteristiche in comune:

-sono unimodali: osserviamo un unico massimo in corrispondenza di μ

-sono simmetriche rispetto all’asse che passa per μ e ciò stà a significare che intervalli di uguale ampiezza a dx e a sx di μ hanno la stessa probabilità.   
Nel nostro caso è ragionevole assumere un modello normale perché, avendo escluso i bambini nati pretermine o con patologie, ci attendiamo che il processo di crescita fisiologico, che riguarda ciascun bambino, ci condurrà probabilmente ad osservare un peso alla nascita intorno ad un valore atteso. Inoltre pesi superiori o inferiori a quello atteso potranno essere attributi a fattori casuali.

**Immaginiamo di assumere un modello normale. Cosa vuol dire campionamento casuale? Quali garanzie ci offre in probabilità il campionamento casuale? Qual è il ruolo di *n* (della numerosità del campione)?**

Federica Pitocco

Immaginiamo di voler stimare il peso alla nascita dei bambini italiani maschi nati al termine, con assenza di patologie. Si definisce una gravidanza a termine quella in cui il parto avviene tra le 37 e le 41 settimane.

A questo punto, ipottiamo di assumere un modello normale (vedi ripsosta precedente).

Procediamo poi con l’osservazione di un campione per poter stimare μ (il peso medio nella popolazione) e σ2 (la variabilità del peso intorno al peso medio nella popolazione).

Estrarremo il nostro campione attraverso un campionamento casuale semplice, facendo sì che tutti i bambini appartenenti alla popolazione d’interesse abbiano la stessa probabilità di essere estratti.

Se iimaginiamo uno studio retrospettivo, per rispettare questa condizione di equi-probabilità dovremmo disporre di una lista di tutti i bambini nati ad esmpio nell’ultimo anno.

Descriviamo la nostra osservazione (il peso osservato nell’i-esimo bambino estratto) attraverso una variabile aleatoria (Yi) la quale spazia sui valori reali, poiché parliamo di una variabile aleatoria continua.   
Il nostro campione, perciò, diventa formalmente un vettore di variabili aleatorie continue (Y1,…,Yn), indipendenti e identicamente distribuite. Difatti, nel modello normale il campionamento casuale semplice ci garantisce che ciascuna variabile aleatoria del nostro campione abbia la stessa distribuzione che abbiamo ipotizzato nella popolazione

Questo perché: poiché tutti i bambini appartenenti alla popolazione d’interesse hanno la stessa probabilità di essere estratti, possiamo applicare la definizione classica di probabilità e dire che la probabilità che il peso osservato nell’i-esimo bambino estratto sia compreso nell’intervallo [a,b] è uguale alla frequenza relativa dei bambini che hanno alla nascita un peso compreso nell’intervallo [a,b].

Per definizione sappiamo che nel continuo la probabilità è l’area sottostante la funzione f(x), (nel nostro caso la densità di probabilità di Yi) compresa nell’intervallo [a,b] per cui:

Tale area corrisponderà all’area del rettangolo corrispondente all’intervallo [a,b] nell’istogramma che descrive la distribuzione del peso alla nascita nella nostar popolazione. Poiché questo è vero per qualsiasi rettangolo possiamo che affermare che la densità Xi sarà la stessa che abbiamo ipotizzato nella popolazione.

Dunque: Il nostro campione è formalmente un vettore di variabili aleatorie continue (X1,…,Xn), indipendenti e identicamente distribuite perché ognuna di esse eredita la stessa distribuzione che abbiamo ipotizzato nella popolazione. Stimiamo μ e σ2 sulla base dei dati campionari.   
Lo stimatore che andiamo ad utilizzare per μ sarà la media campionaria, variabile aleatoria continua perché somma di variabili aleatorie continue, il cui valore cambierà al cambiare dal campione che potremo osservare.

Poiché la media campionaria è somma di variabili aleatorie normali ( n che è una costante), la distribuzione della nostra media campionaria seguirà comunque una densità normale.

A questo punto, dovremmo capire quanto vale il valore atteso E[] e la varianza di questa variabile aleatoria normale.

Per cui:

Questa densità normale prende il nome di “distribuzione campionaria” della media; aree sotto questa densità descrivono la probabiità di osservare medie campionarie, cioè stime di μ, nell’intervallo corrispondente. In altri termini essa ci dice quali sono le stime più probabili e meno probabili al variare del campione che potremmo osservare. Essendo la densità concentrata proprio intorno a μ sarà molto probbaile osservare stime intorno al valore vero.

rappresenta la variabilità del nostro stimatore (la variabilità delle stime al variare del campione nello spazio dei campioni) e dipende da che rappresenta la varianza nella popolazione (la variabilità del fenomeno nella popolazione d’interesse).   
Questo implica che: aumentando la numerosità del campione (il numero delle osservazioni) possiamo concentrare la legge normale intorno a , il valore vero (in altre parole possiamo ridurre l’entità dell’errore nonostante la varianza resti la stessa).

Difatti, calcoliamo lo standard error (errore associato al nostro stimatore, la media campionaria) facendo la radice della varianza del nostro stimatore: .

**Qual è il percorso per la stima di μ in un modello normale? Definire lo stimatore, lo standard error e l’intervallo di confidenza (Daniele Antonucci)**

In un modello normale, μ rappresenta il valore medio effettivo della variabile di interesse nella popolazione. Il nostro problema è quello di poterlo stimare sulla base di un campione di osservazioni. Partiamo dal fatto che scegliamo il campione casualmente, questo ci consente di esprimere il campione come un vettore di variabili aleatorie (,…,) dove Yi indica il valore della variabile ogetto di studio che osserveremo sull’-iesmo individuo estratto. Il campionamento casuale ci garantisce che queste variabili aleatorie seguiranno la stessa densità normale della popolazione di riferimento. In queste condizioni lo stimatore più naturale per μ è la media campionaria che indicheremo con , calcolata appunto come la sommatoria delle osservazioni divisa per la numerosità campionaria. In generale uno stimatore è una funzione delle osservazioni campionarie cioè una variabile aleatoria che associa una stima del parametro incognito ad ogni possibile campione.

Una volta stimato il valore medio nel nostro campione dobbiamo però occuparci di valutare l’incertezza legata a questo risultato. L’errore viene valutato sulla base della variabilità delle stime che otterremmo al variare del campione nello spazio dei campioni, in altri termini, sulla base della varaibilità del nostro stimatore che si dimostra dipendere direttamente dalla variabilità della variabile di interesse nella popolazione e inversamente dalla numerosità campionaria (standard error della media campionaria = ). Ricordiamo che nell’ipotesi di un modello Normale anche la media distribuzione campionaria della media seguirà una densità Normale di valore atteso μ. Aumentando le osservazioni possiamo diminuire la variabilità e concentrare questa densità normale intorno al valore medio (μ).

E’ possibile combinare la stima di μ ed il suo standard error attraverso un intervallo di confidenza al 95% che rappresenta una stima per intervallo. La probabilità associata all’intervallo indica che in ipotetiche ripetizioni dell’esperimento, nel 95% dei casi l’intervallo stimato conterrà il valore reale. Si tratta di una garanzia sul metodo utilizzato e non sull’intervallo stimato. Per costruire l’intervallo di confidenza dobbiamo prima di tutto operare una standardizzazione della media campionaria

Sostituendo la varianza nella popolazione con il suo stimatore S2, avremo

che segue la legge nota come t di Student con n-1 gradi di libertà. Sulla base di questo risulttao potremo scrivere

dove è il percentile della densità t di Student che lascia alla sua destra un’area pari a 0.025.

**Qual è il percorso per la stima di in un modello Bernoulliano? Definire lo stimatore, lo standard error e l’intervallo di confidenza (Daniele Antonucci)**

Come prima cosa consideriamo un campionamento casuale che dunque ci consente di esprimere il campione come vettore di variabili aleatorie (,…,). In un modello Bernoulliano π rappresenta la probabilità di osservare la caratteristica di interesse nella singola osservazione. L’obiettivo dello studio è stimare la frequenza F con cui questa caratteristica compare nella popolazione. Ricordiamo che se il campionamento è casuale la probbailità π è uguale ad F.

Per stimare questa probabilità e quindi F, scegliamo come stimatore la frequenza relativa della caratteristica osservata nel nostro campione, calcolata come numero di osservazioni in cui compare diviso la numerosità campionaria.

Resta ora da valutare l’incertezza di questa stima. L’errore in statistica viene valutato sulla base della variabilità delle stime che otterremmo al variare del campione nello spazio dei campioni, in altri termini, sulla base della variabilità del nostro stimatore che si dimostra essere legata sia al numero di esperimenti effettuati, sia al valore di π secondo la legge: Var()= .

La radice della varianza, cioè la deviazione standard del nostro stimatore, prende il nome di standard error.

Una volta stimato π è possibile costruire un intervallo di confidenza al 95%. Per fare ciò sfruttiamo il teorema centrale del limite che afferma che la media standardizzata di variabili aleatorie indipendenti che seguono la stessa distribuzione tende alla distribuzione Normale standard al tendere ad infinito del numero delle variabili. Osserviamo che, essendo il nostro campione casuale le condizioni del terorema sono soddisfatte. Inoltre la frequenza camionaria è di fatto una media delle osservazioni. Per grandi campioni potremo pertanto approssimare la sua distribuzione ad una densità Normale. Avremo quindi

La probabilità associata all’intervallo indica che in ipotetiche ripetizioni dell’esperimento, nel 95% dei casi l’intervallo stimato conterrà il valore reale.

**Descrivere in un ipotetico studio sperimentale come si stima se assumiamo un modello normale. Definire lo stimatore, lo standard error e l’intervallo di confidenza. (Federica Pitocco e Ilaria Cirulli)**

Immaginiamo di voler stimare il peso alla nascita di bambini italiani maschi, nati a termine e con assenza di patologie. Si definisce una gravidanza a termine quella in cui il parto avviene tra le 37 e 41 settimane.

Assumiamo un modello normale.

Assumere un modello normale vuol dire immaginare che la distribuzione del peso nella popolazione abbia le caratteristiche di una densità normale. In altre parole: immaginiamo di osservare il peso alla nascita di tutti i bambini appartenenti alla popolazione d’interesse e costruire un istogramma di frequenze sulla base dei dati osservati. Assumere un modello normale vuol dire ipotizzare che l’istogramma costruito abbia una forma approssimabile ad una densità normale, dove μ è il peso medio nella popolazione e σ2 la variabilità del peso, rispetto al peso medio, sempre nella popolazione.

Poiché è chiara l’impossibilità di effettuare un’analisi sulla totalità dei bambini, estraiamo casualmente un campione eterogeneo e rappresentativo della popolazione per stimare μ.

Il campione deve essere estratto casualmente; in altri termini tutti i bambini appartenenti alla popolazione d’interesse devono avere la stessa probabilità di essere estratti (pari ad 1/N dove N è la numerosità della popolazione). La scelta del campionamento casuale ci consente di descrivere la nostra osservazione con una variabile aleatoria Yi che spazia su tutti i valori reali, proprio perché siamo nel continuo. Il nostro campione diventa formalmente un vettore di variabili aleatorie continue (Y1,…,Yn), indipendenti e identicamente distribuite. Difatti, nel modello normale il campionamento causale semplice ci garantisce che ciascuna variabile aleatoria del nostro campione erediterà la stessa distribuzione che abbiamo ipotizzato per la popolazione.

Stimiamo μ (il peso medio nella popolazione) e σ2 (la variabilità del peso nella popolazione) sulla base dei dati campionari. Lo stimatore che utilizziamo per μ è la media campionaria che indichiamo con . Uno stimatore è una funzione delle nostre osservazioni, ovvero una variabile aleatoria che associa una stima di μ ad ogni possibile campione.

Poiché la media campionaria è somma di variabili aleatorie normali, indipendenti e identicamente distribuite (i.d.d.) diviso n (una costante), la sua distribuzione seguirà un modello normale. A questo punto, dovremmo capire quanto vale il sul valore atteso E[] e la sua varianza V[].

Per cui:

Questa densità normale prende il nome di “distribuzione campionaria” della media; aree sotto questa densità descrivono la probabilità di osservare medie campionarie, ovvero stime di μ, nell’intervallo corrispondente. In altre parole: essa ci dice quali sono le stime più probabili e meno probabili al variare del campione che potremmo osservare.  
Essendo concentrata proprio intorno a μ sarà molto probabile osservare stime intorno al valore vero. Come abbiamo precedentemente detto, la media campionaria è una variabile aleatoria continua che associa una stima di μ ad ogni possibile campione. Calcoliamo l’errore come la variabilità delle stime che otterrei al variare del campione nello spazio dei campioni.

Poiché è chiara l’impossibilità di osservare tutti i campioni nello spazio dei campioni, sfruttiamo le leggi della probabilità e procediamo stimando l’errore attraverso la varianza del nostro stimatore.

Abbiamo dimostrato che questa varianza è e quindi dipende direttamente da (la variabilità del peso nella popolazione) e inversamente da n (la numerosità campionaria); per cui, possiamo concentrare la legge normale intorno a μ aumentando la numerosità campionaria.   
Calcoliamo lo standard error (l’errore associato al nostro stimatore) facendo la radice della varianza del nostro stimatore: .

A questo punto, possiamo combinare la stima di μ con lo standard error attraverso un intervallo di confidenza al 95% che rappresenta una stima per intervallo.   
In altre parole: l’intervallo aleatorio conterrà con una probabilità del 95%, al variare del campione nello spazio dei campioni, il valore vero μ.

Si tratta di una garanzia sul metodo utilizzato e non sull’intervallo stimato.

Per costruire l’intervallo di confidenza dobbiamo innanzitutto standardizzare la media campionaria.

Procediamo sostituendo σ2 con il suo stimatore S2 (la varianza campionaria) e otteniamo:

Tale variabile aleatoria seguirà una legge nota come t di Student con n-1 gradi di libertà (perché n-1 sono i gradi di libertà che abbiamo utilizzato per stimare σ2).

Sulla base di questo risultato possiamo scrivere:

Da cui:

Dove è quel percentile sotto la densità t di Student che lascia sulla destra un’area di 0,025.

Dunque, con una probabilità del 95%, al variare del campione nello spazio dei campioni, il campione osservato mi condurrà ad un intervallo che conterrà il peso medio nella popolazione.

**Descrivere in un ipotetico studio Bernoulliano come si stima .**

**Definire lo stimatore, lo standard error e l’intervallo di confidenza.**

**Federica Pitocco**

Immaginiamo un esperimento focalizzato sullo studio della popolazione di bambini residenti in Italia in età compresa tra 6 e 10 anni dei quali siamo interessati a stimare la frequenza relativa F di bambini allergici nel periodo primaverile.

Poiché è chiara l’impossibilità di effettuare l’analisi sulla totalità dei bambini, estraiamo casualmente un campione eterogeneo e rappresentativo della popolazione, per poi generalizzare il risultato dello studio a tutti i bambini.

Il campionamento deve essere casuale; ciò significa che tutti i bambini appartenenti alla popolazione devono avere la stessa probabilità di essere estratti (pari ad 1/N dove N è la numerosità della popolazione).

La scelta di un campionamento casuale ci consente di descrivere la singola osservazione come una variabile aleatoria Bernoulliana (Xi) che potrà assumere valori: Xi=0 se l’i-esimo bambino reclutato non è allergico, Xi=1 se è allergico.

Definiamo la probabilità che il bambino estratto sia allergico, avremo quindi: P(Xi=1)= e P(Xi=0)=1-.

Il nostro campione diventa formalmente un vettore di variabili aleatorie (X1,… Xn), indipendenti e identicamente distribuite (i.i.d.).

Ricordiamo che: il campionamento causale semplice garantisce che è uguale ad F, la frequenza dei bambini allergici nella popolazione.   
Se siamo di fronte ad una numerosità campionaria sufficientemente elevata, poiché il nostro campione è un vettore di variabili aleatorie i.i.d., ci possiamo affidare alla legge dei grandi numeri e affermare che la frequenza campionaria dei bambini allergici tenderà a e poiché è uguale ad F, tale frequenza sarà simile a quella che abbiamo nella popolazione.

In simboli

Per cui, stimeremo (e quindi F) sulla base della frequenza campionaria f che, essendo le nostre osservazioni dicotomiche, coincide con la media campionaria:

Calcoliamo l’errore associato alla nostra stima sulla base della variabilità delle stime che otterremmo al variare del campione nello spazio dei campioni. Poiché è chiara l’impossibilità di osservare tutti i campioni nello spazio dei campioni, sfruttiamo le leggi della probabilità e procediamo stimando l’errore attraverso la varianza del nostro stimatore

Si dimostra che: l’errore associato alle nostre stime dipende direttamente da e inversamente da n, per cui aumentando la numerosità campionaria possiamo ridurre l’entità del nostro errore.

Calcoliamo lo standard error (l’errore associato al nostro stimatore) facendo la radice della varianza del nostro stimatore. Otteniamo: .

A questo punto, possiamo combinare la stima di μ con lo standard error attraverso un intervallo di confidenza al 95% che rappresenta una stima per intervallo.   
In altre parole: l’intervallo aleatorio conterrà con una probabilità del 95% il valore vero di , al variare del campione nello spazio dei campioni.

Si tratta di una garanzia sul metodo utilizzato e non sull’intervallo stimato.

Per fare questo, sfruttiamo il teorema centrale del limite che afferma che la media standardizzata di variabili aleatorie indipendenti che seguono la stessa distribuzione tende alla distribuzione Normale Standard al tendere infinito del numero delle variabili.   
Poiché nel nostro campione sono soddisfatte queste condizionati, per grandi campioni possiamo approssimare la distribuzione della frequenza campionaria (o media campionaria) ad una densità normale.

Per cui:

Da cui:

Dunque: con una probabilità del 95%, l’intervallo aleatorio che ho costruito conterrà e quindi F, la frequenza relativa dei bambini allergici nel periodo primaverile. In termini intuitivi potremmo dire che, su cento intervalli di confidenza che potrei ottenere in ipotetiche ripetizioni dell’esperimento nelle medesime condizioni, 95 conterranno il valore vero di F. Si tratta di un garanzia sul metodo utilizzato e non sull’intervallo stimato sui nostri dati il quale potrebbe essere tra i 95 “buoni” oppure tra i 5 “cattivi”.

**Quali sono gli elementi che costituiscono un test statistico nel modello normale che in quello Bernoulliano? (Camilla Ciampa)**

Immaginiamo uno studio prospettico dove la nostra popolazione è costituita da individui in remissione dal tumore ai polmoni. Il nostro obiettivo è stimare la frequenza F con cui, in questa popolazione, comparirà una recidiva del tumore nell’arco di 5 anni. Estraggo casualmente un campione di 10 individui da questa popolazione. Dal momento che non dispongo di una lista della mia popolazione potrò campionare casualmente dalla lista dei centri specializzati per questa patologia. Per ogni centro estratto andrò ad estrarre un campione casuale dalla lista dei pazienti in remissione. Ipotizziamo di seguire questi individui per 5 anni per vedere se si presenta o meno una recidiva.

La scelta di un campionamento casuale ci consente di descrivere la singola osservazione come una variabile aleatoria Bernoulliana (Xi) che potrà assumere valore 0 se l’i-esimo paziente non presenta una recidiva ed 1 se compare una ricaduta. Definiamo la probabilità che l’i-esimo paziente presenti una recidiva. Avremo quindi: P(Xi=1)= e P(Xi=0)=1-. Il nostro campione diventa formalmente un vettore di variabili aleatorie (X1,…,X10), indipendenti ed identicamente distribuite.

Poiché il campione estratto è casuale, π sarà uguale a F.

Vogliamo verificare le seguenti ipotesi:

H0 -> ipotesi nulla (ipotesi che viene da lavori precedenti, quindi più conservativa)

H1 -> ipotesi alternativa (ipotesi sperimentale, che noi abbiamo in mente)

H0: π = 0,2

H1: π = 0,3

Per decidere a favore di una delle due ipotesi utilizzo i dati campionari sintetizzati in base alla statistica test S, funzione delle nostre osservazioni, che in questo caso è la somma delle osservazioni cioè il numero di pazienti nel campione che presenterà una recidiva. In simboli

Avendo preso un campione di 10 pazienti, S potrà assumere valori da 0 a 10. La statistica test semplifica lo spazio dei campioni originario Ω, infatti il numero di eventi elementari osservabili passa da 210 a 11.

Partendo da due ipotesi, la distribuzione campionaria di S ci dirà con che probabilità osserveremo i valori da 0 a 10 quando è vera H0 oppure H1. Dalla teoria della probabilità sappiamo che S seguirà una distribuzione Binomiale. In particolare avremo sotto H0

e ipotizzando vera H1

Definisco una regola di decisione che, sulla base dei valori di S, mi conduca a rifiutare o accettare H0. Formalmente faccio una partizione dell’Ω di S in due sottoinsiemi, la regione di accettazione e di rifiuto dell’ipotesi nulla.

Posso farlo utilizzando il principio di verosimiglianza: confronto le due distribuzioni campionarie e scelgo come regione di accettazione di H0 i valori di S in cui Pr{S = r|H0} > Pr{S = r|H1}, cioè quei valori per i quali è più verosimile H0. La regione di rifiuto di H0 comprenderà i valori di S in cui Pr{S = r|H0} ≤ Pr{S = r|H1}, cioè quei valori per i quali è più verosimile H1.

Quella che sto applicando è l’idea di verosimiglianza: prima calcolo le probabilità dei diversi valori che può assumere S ipotizzando vere le due ipotesi, poi immagino di aver osservato S=r (quindi S=r è adesso un evento noto) e utilizzo queste probabilità sotto le due ipotesi per valutare la verosimiglianza di H0 e di H1 alla luce del valore r che ho osservato.

Una volta individuata la regione di rifiuto, il test è completato formalmente.

Si possono però commettere degli errori. Commetto un errore di prima specie se rifiuto H0 quando invece è vera. La sua probabilità sarà la somma delle probabilità dato H0 dei valori di S che si trovano in regione di rifiuto. Commetto invece un errore di seconda specie se accetto H0 ma in realtà è vera H1. La sua probabilità sarà la somma delle probabilità dato H1 dei valori di S che si trovano in regione di accettazione.

Per migliorare il test, devo aumentare la numerosità del campione per riuscire a separare le distribuzioni di S sotto le due ipotesi e far in modo che non si sovrappongano.

Un’altra cosa che si potrebbe fare sarebbe distanziare ulteriormente le due ipotesi alternative (ma questo solitamente non è impossibile poiché le ipotesi sperimentali sono definite).