# QUANTI SUINI DEVO ESAMINARE? IL CALCOLO DELLA NUMEROSITÀ DEL CAMPIONE E L'INTERPRETAZIONE DEI RISULTATI

FABIO OSTANELLO

*Dipartimento di Scienze Mediche Veterinarie, Università di Bologna*

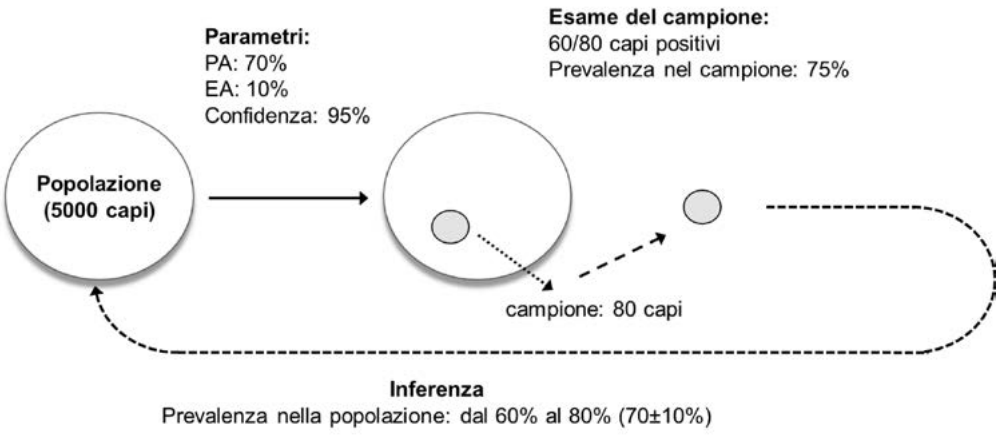
## INTRODUZIONE

Quando si devono prelevare dei campioni, ad esempio per diagnosticare una specifica malattia, raramente è possibile esaminare *ogni singolo* componente della popolazione: i fattori limitanti più frequenti sono in genere rappresentati dalle risorse disponibili (economiche, di tempo, di personale, di strutture diagnostiche, ecc.). L'esame di un campione, cioè di un numero ridotto di animali, consente di superare la maggior parte di questi problemi. Un campione è quindi un sottoinsieme di elementi estratti da una popolazione di interesse. Ne consegue che un campione è soltanto una parte, più o meno grande, dell'intera popolazione. Scegliere, cioè estrarre, un campione da una popolazione significa effettuare un campionamento. Il principale obiettivo di una indagine condotta su base campionaria è quello di raccogliere dati e informazioni che consentiranno di generalizzare, con un certo grado di incertezza, all'intera popolazione, le conclusioni ottenute dall'esame del campione. Questo processo di generalizzazione è detto inferenza (figura 1).

Quando si effettua una valutazione su base campionaria, è necessario tener presente che non si otterranno mai dei risultati del tutto affidabili, cioè applicabili con certezza all'intera popolazione da cui è stato estratto il campione. Ciò significa che il processo di inferenza è soggetto ad errore. Per valutare l'affidabilità di uno studio campionario è indispensabile tener conto di vari fattori, fra i quali i più importanti sono: i criteri di scelta della popolazione in esame, la numerosità ed il metodo con cui è stato selezionato il campione, il periodo di osservazione, i metodi adottati per identificare i casi di malattia, le tecniche di analisi, la precisione delle misure effettuate. In questa sede valuteremo solo gli aspetti relativi al calcolo della numerosità del campione, con particolare riferimento alla diagnosi delle malattie trasmissibili.

Se siete interessati agli aspetti matematici del problema, esistono molti buoni libri di statistica o di epidemiologia che vi possono fornire esaurienti spiegazioni relativamente alle formule utilizzate per il calcolo (vedi Bibliografia). Il mio scopo è invece quello consentire a chiunque di stabilire, in modo corretto, il numero di soggetti che devono essere estratti da una popolazione ed esaminati in modo da eseguire una stima di tipo campionario e valutare in maniera congrua i risultati. Per fare questo, esistono numerosi strumenti informatici gratuiti che, con un minimo di conoscenze, ci permettono di effettuare i calcoli necessari.

**Figura 1.** Rappresentazione schematica del campionamento e del processo di inferenza




## I LIMITI DEL CAMPIONAMENTO E IL CONCETTO DI RAPPRESENTATIVITÀ

Partiamo da una semplice osservazione: i dati che si ottengono da una indagine campionaria non saranno mai assolutamente precisi e non rispecchieranno mai la reale situazione sanitaria di un gruppo numeroso di soggetti. In altri termini, i dati che si ottengono sono una approssimazione (stima), più o meno valida, della realtà. E' chiaro però che questa approssimazione ha enormi vantaggi in termini economici, di tempo, ecc. Ne consegue che è possibile ottenere comunque delle informazioni anche in quelle condizioni in cui non è perseguibile la strada dell'esame di tutti gli animali. E' possibile inoltre stabilire a priori il livello di precisione desiderata tenendo però in considerazione il fatto che a maggiore precisione consegue la necessità di aumentare la numerosità del campione.

L'affidabilità delle conclusioni cui è possibile giungere attraverso indagini di tipo campionario dipende sostanzialmente da due elementi: il primo, quantitativo, è quello relativo al numero di soggetti che dovranno costituire il campione, cioè alla numerosità (dimensione) del campione e il secondo, qualitativo, è quello relativo alle caratteristiche che questi individui devono possedere per garantire la rappresentatività campionaria (es. composizione per sesso, età, peso, caratteristiche genetiche del campione).

Il numero di animali che compongono il campione è uno dei più importanti fattori che influiscono sulla precisione della stima: campioni di grandi dimensioni (rispetto al totale dei soggetti) permettono infatti stime più precise.

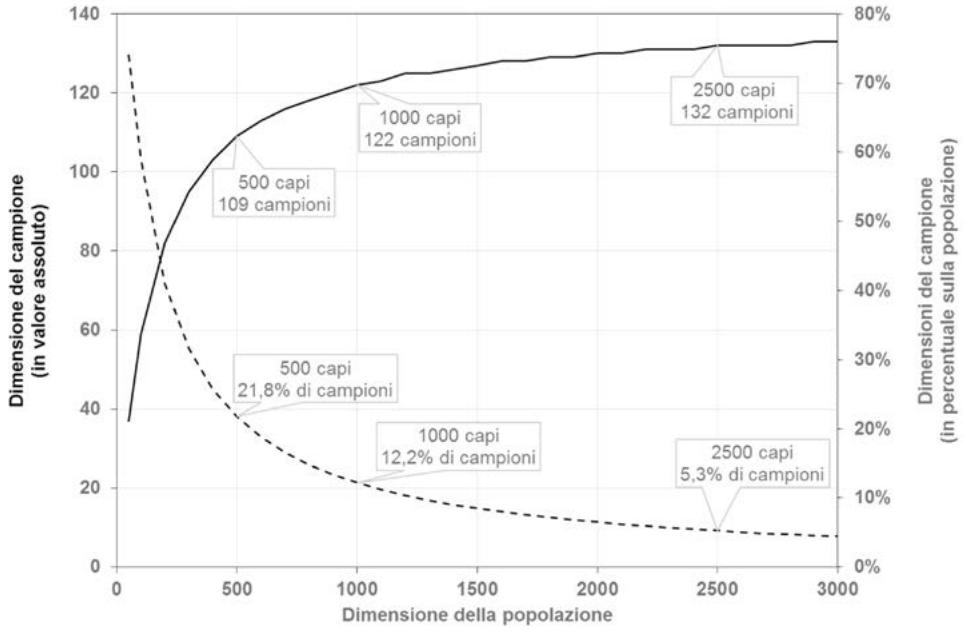
L'aspetto qualitativo viene garantito quando ciascun animale ha la stessa probabilità di entrare a far parte del campione. In tal caso il campione viene detto "randomizzato" o "casuale". Affidandosi al caso si ottiene, in una qualche misura, la garanzia che il campione sarà rappresentativo della popolazione stessa, ossia ne rifletterà le caratteristiche con una certa approssimazione. Tutto ciò vale però solo nell'ambito del campionamento di tipo probabilistico. Ci sono situazioni in cui può essere preferibile scegliere un'altra strada. Ad esempio, se la mia necessità è quella stabilire l'eventuale presenza di una malattia in allevamento, potrebbe essere opportuno selezionare solo gli animali con sintomi clinici. Questo tipo di campionamento, che viene definito "di convenienza", non permette però di generalizzare le osservazioni all'intera popolazione.

Ci sono altri 2 aspetto da tenere in considerazione: a) i vantaggi dell'esame campionario

crescono all'aumentare della dimensione di popolazione: per piccole popolazioni (50-100 capi) ha poco senso effettuare indagini campionarie; b) è concettualmente un grave errore stabilire un numero di campioni da esaminare sulla base di percentuali "fisse" che non tengono in considerazione il numero di animali che compongono la popolazione in esame.

Proviamo a spiegare questi ultimi aspetti con un esempio. In figura 2, la curva nera rappresenta il numero di campioni necessario a stimare la prevalenza di malattia (vedi dopo) in allevamento. E' evidente che all'aumentare del numero di animali presenti in allevamento aumenta (prima rapidamente, poi più lentamente) anche il valore assoluto dei campioni da prelevare (es. allevamento di 500 capi: 109 campioni; 2500 capi: 132 campioni). Se però, rapportiamo il numero dei campioni alla popolazione da cui saranno estratti, è possibile osservare un andamento inverso (curva tratteggiata). Quindi: a) il campionamento è tanto più conveniente quando più grande è il numero di animali presenti; b) stabilire, a priori, il numero di animali da esaminare sulla base di percentuali fisse è un grave errore. Se, ad esempio, avessi fissato arbitrariamente al 5% il numero di campioni da estrarre, in un allevamento di 500 capi avrei esaminato 40 animali (ottenendo così un campione numericamente non rappresentativo), mentre in un allevamento di 2500 capi avrei esaminato 200 suini (ottenendo così un campione numericamente rappresentativo ma inutilmente grande).

**Figura 2.** Variazione assoluta e percentuale della numerosità del campione al crescere delle dimensioni della popolazione (PA: 10%; EA: 5%; LC: 95%)



**AMBITI DI APPLICAZIONE DEL CAMPIONAMENTO**

Attraverso l'esame campionario, è possibile valutare 3 diverse condizioni:

1. stimare la presenza/assenza della caratteristica che stiamo cercando (es. malattia) all'interno di un gruppo di animali (a prescindere dal numero di soggetti interessati dalla malattia);
2. stimare la percentuale (cioè la prevalenza) di animali che hanno la caratteristica che stiamo cercando (es. anticorpi specifici nei confronti di PRRSV);
3. stimare il valore medio di un carattere qualitativo (misurabile) in un gruppo di soggetti (es. peso medio degli animali presenti).

Le prime 2 condizioni hanno un interesse prevalentemente diagnostico mentre la terza possibilità può essere utilizzata anche in altri ambiti. Vediamole in dettaglio.

1) *Stima della presenza/assenza della malattia.*

Questo tipo di campionamento può essere utilizzando quando lo scopo non è tanto quello di conoscere quale è la percentuale di animali malati all'interno del gruppo ma si vuole stabilire *se* una determinata malattia è presente, a prescindere dal numero di soggetti coinvolti. In questo caso, il risultato ultimo del campionamento è dicotomico: se individuo almeno 1 soggetto positivo, considero l'intero gruppo da cui proviene quel soggetto come positivo (malato, infetto, ecc.) viceversa, se non individuo nessun soggetto positivo, considero l'intero gruppo come negativo (sano, non infetto, ecc.),

Cosa devo tenere in considerazione per calcolare il numero di campioni necessari per valutare l'eventuale presenza di una malattia? Solo 3 valori:

a) il numero dei soggetti totali della popolazione. Questo è il parametro di più semplice impostazione: è sufficiente conoscere, anche in maniera approssimativa, quanti animali sono presenti in un determinato allevamento o in una certa categoria;
b) la prevalenza attesa (PA) cioè la percentuale di soggetti che mi aspetto possano essere positivi. La necessità di questo valore è apparentemente paradossale: secondo logica, se il mio scopo è quello di *escludere* la presenza di una certa condizione (es. malattia) il valore di PA dovrebbe essere zero. Purtroppo, per ragioni strettamente matematiche, la formula di calcolo richiede un valore di PA che deve essere maggiore di zero. Vediamo come è possibile risolvere, da veterinari, questo paradosso: se parliamo di malattie trasmissibili, quando l'agente eziologico entra per la prima volta in una popolazione, infetta, nella fase iniziale, solo una piccola frazione di soggetti (ma comunque sempre un numero di soggetti superiore a zero). Col passare del tempo, il numero di soggetti infetti tenderà, almeno inizialmente, ad aumentare, stabilizzandosi in una fase successiva. Se il mio scopo è quello di escludere la presenza di una malattia estremamente contagiosa (es. influenza), posso partire dal presupposto che, quando presente in allevamento, il virus influenzale interessi una percentuale di animali (prevalenza) molto alta, ad esempio superiore al 60%. E' chiaro che c'è la possibilità che, se l'infezione è appena entrata in allevamento, la prevalenza possa essere inferiore al 60% ma questa situazione può essere considerata relativamente poco probabile se riferita alla totalità degli allevamenti suinicoli infetti. Al contrario, se sono interessato ad escludere la presenza di una malattia poco contagiosa, posso scegliere un valore di prevalenza attesa più basso;
c) livello di confidenza (LC). Questo valore rappresenta il grado di "robustezza" della stima cioè l'affidabilità del risultato (vedi oltre). Tradizionalmente viene utilizzato un valore del 95% ma è possibile aumentarlo o diminuirlo. E' intuitivo però che la maggiore "affidabilità" della stima (es. LC 99%) si paga in termini di aumento del numero di soggetti da esaminare, a parità delle altre condizioni.

Questo tipo di campionamento può essere utilizzato anche per valutare se, in un determinato allevamento, la prevalenza di infezione è al di sopra o al di sotto di un valore fissato in maniera arbitraria (valore soglia). In questo modo posso classificare gli allevamenti in due categorie: allevamenti "problema" in cui la prevalenza è superiore ad una certa soglia ed allevamenti "virtuosi" in cui la prevalenza è al di sotto di una determinata soglia. E' il caso ad esempio del controllo sierologico per malattia di Aujeszky.

Facciamo qualche esempio e vediamo di interpretare i risultati:

- in un allevamento di 5000 capi, che non effettua la profilassi vaccinale nei confronti dei virus influenzali, voglio stabilire sierologicamente se c'è circolazione virale. Prudenzialmente, ipotizzo una prevalenza attesa del 20% (in realtà è noto che negli allevamenti infetti la prevalenza supera il 60%) e scelgo di utilizzare un livello di confidenza del 95%. Utilizzando uno dei software suggeriti alla fine di questo articolo, stabilisco che il numero di soggetti da esaminare (campione) è pari a 14. L'esito dell'esame sierologico mi informa che tutti i 14 campioni sono risultati sieronegativi. Posso concludere, con certezza, che l'infezione è assente? No, posso solo affermare che l'infezione, se presente, interessa meno del 20% degli animali, situazione che, come sappiamo, è poco probabile. Siamo quindi propensi a ritenere che la malattia non è presente in allevamento;
- per un allevamento da riproduzione di 1200 scrofe, i requisiti per l'ottenimento della qualifica di allevamento indenne da Malattia di Aujeszky prevedono il controllo di 59 capi (PA: 5%; LC: 95%). Questa numerosità campionaria garantisce una probabilità del 95% (livello di confidenza) di rilevare almeno 1 capo positivo alla ricerca di anticorpi anti gE quando la prevalenza aziendale supera la prevalenza del 5%. Per il mantenimento della qualifica, il valore di prevalenza attesa è del 10%: in questo caso il numero di campioni scende a 29. Se ritenessi necessario un risultato più affidabile, potrei calcolare il numero di soggetti da esaminare utilizzando un livello di confidenza più elevato (es. 99%). I soggetti da esaminare sarebbero 89 nel primo caso e 44 nel secondo: la maggiore affidabilità si paga in termini di aumento del numero di soggetti da esaminare.

*2) Stima della prevalenza*

In questo caso, lo scopo del campionamento è quello di stimare la percentuale di soggetti che presentano la caratteristica che sto cercando (es. anticorpi specifici verso un certo agente eziologico) all'interno di una popolazione (es. allevamento). Cosa devo tenere in considerazione per calcolare il numero di campioni da esaminare? solo 4 valori:

a) il numero (anche approssimativo) dei soggetti totali della popolazione (vedi paragrafo precedente);
b) la prevalenza attesa (PA) cioè la percentuale di soggetti che mi aspetto possano essere positivi. Anche in questo caso siamo difronte ad un paradosso ma la soluzione può essere trovata seguendo le indicazioni fornite in precedenza o riferendosi a valutazioni precedenti. Solo nel caso in cui non sia possibile ipotizzare nessun valore di prevalenza attesa (es. malattia mai segnalata in precedenza) è possibile utilizzare il valore 50%;
c) errore accettabile (EA). Rappresenta, in percentuale, il livello di precisione desiderata rispetto al valore della prevalenza attesa. Spieghiamolo con un esempio: gli exit pool elettorali. Il sondaggio elettorale effettuato all'uscita dal seggio permette di stimare in poche ore la percentuale (cioè la prevalenza) di elettori che hanno votato per il partito X. Questo consente di ottenere delle stime del risultato elettorale molti

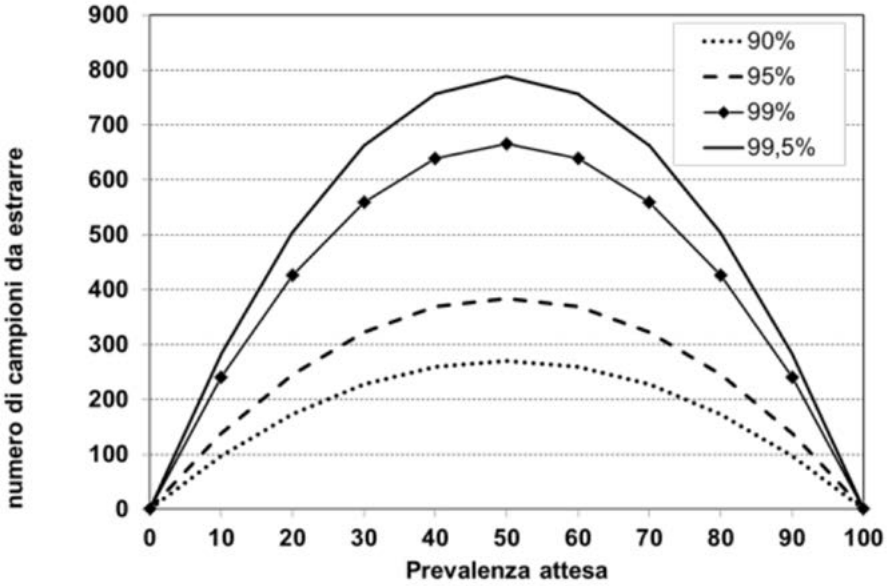
giorni prima della comunicazione dei risultati ufficiali. Il numero di soggetti da intervistare viene calcolato sulla base dei precedenti risultati elettorali del partito X (es. 22%) e di un certo valore di errore accettabile (es. 2%). Intervistiamo 1648 soggetti (0,006% di una popolazione di elettori pari a 30 milioni) e stabiliamo che il 21% degli intervistati ha votato per il partito X. Partendo dal presupposto che nessuno abbia mentito, è possibile concludere che, a livello nazionale, il partito X totalizzerà una percentuale di voti variabile da un minimo del 20% ad un massimo del 24% (22±2%). Se volete, potete verificare i calcoli utilizzando uno dei software suggeriti alla fine di questo articolo.

d) Livello di confidenza (LC, vedi paragrafo precedente).

Facciamo qualche esempio e vediamo di interpretare i risultati:

- in un allevamento da ingrasso di 5000 capi, voglio stabilire la sieroprevalenza per PRRSV. Ipotizzo (in base alla mia esperienza o da dati di letteratura) un valore di prevalenza attesa del 70%, un errore accettabile del 10% e un livello di confidenza del 95%. Utilizzando uno dei software suggeriti alla fine di questo articolo, stabilisco che il numero di soggetti da esaminare (campione) è pari a 80 (che rappresenta l'1,6% degli animali presenti). L'esito dell'esame sierologico mi informa che 60 campioni sono risultati sieropositivi. La prevalenza nel campione è del 75%. La prevalenza nell'intera popolazione potrà variare da un minimo del 60% ad un massimo dell'80% (70±10%), (figura 1);
- una malattia precedentemente mai segnalata si sta diffondendo nella popolazione. In questa situazione, non è possibile, realisticamente, stabilire un valore di prevalenza attesa. La soluzione da adottare è quella di utilizzare la prevalenza attesa che garantisca la massima numerosità del campione (figura 3). Questo *escamotage* da un lato mi mette al riparo da problemi legati alla rappresentatività numerica del campione ma dall'altro mi costringe ad aumentare le risorse da investire per la diagnosi.

**Figura 3.** Stima della prevalenza. Numero assoluto di campioni da estratte al crescere della prevalenza attesa e per diversi livelli di confidenza



*3) Stima del valore medio di un carattere qualitativo*

In questo caso, sono interessato a conoscere qual è il valore medio di una variabile di tipo quantitativo (es. peso) in un gruppo di soggetti. Cosa devo tenere in considerazione per calcolare il numero di campioni da esaminare? Solo 3 valori:

a) il numero (anche approssimativo) dei soggetti totali della popolazione (vedi paragrafo precedente).
b) la deviazione standard (DS). Proviamo a dare una spiegazione "grafica" del concetto di deviazione standard. La figura 4 rappresenta la curva di distribuzione Normale (o gaussiana). In questa distribuzione, il 95% delle osservazioni sono comprese nell'intervallo: media ± 2 DS. Immagina di aver pesato 2 gruppi diversi di suini (A e B) e di aver ottenuto, per entrambi i gruppi, lo stesso valore medio di peso vivo (100 Kg). Tuttavia: nel gruppo A il valore di Deviazione Standard è basso (2 Kg); la curva risulterà stretta e alta, in conseguenza dello scarso range di variabilità dei valori (96-103 Kg). Nel gruppo B il valore di Deviazione Standard è maggiore (4 Kg); la curva risulterà schiacciata, in conseguenza dell'ampio range di variabilità dei valori (92-108). In altri termini: nel gruppo A la popolazione è costituita da individui di peso relativamente omogeneo. In questa situazione, sarà sufficiente pesare pochi soggetti per ottenere una stima attendibile del peso vivo medio della popolazione. Nel gruppo B la popolazione è costituita da individui di peso poco omogeneo. In questa situazione, per ottenere una stima attendibile del peso vivo medio della popolazione sarà necessario pesare un numero maggiore di soggetti;
e) Errore accettabile (EA). Rappresenta, nella stessa unità di misura del parametro che sto valutando (es. peso: Kg) livello di precisione desiderato.
c) Livello di confidenza (LC, vedi paragrafo precedente).

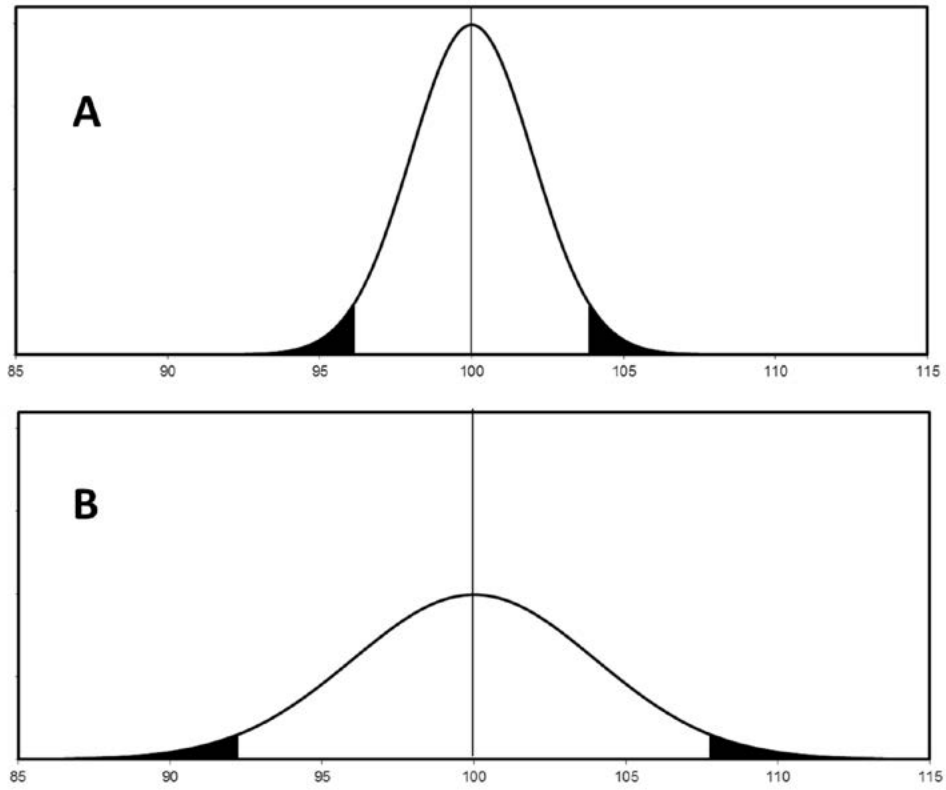
Facciamo qualche esempio e vediamo di interpretare i risultati:

- vogliamo determinare con un errore accettabile di 2 Kg, il peso vivo medio di due lotti di suini (A e B) composto ciascuno da 300 animali. Il lotto A è composto da animali di peso omogeneo; in base ai risultati ottenuti da una serie limitata di pesate preliminari, mi aspetto che la deviazione standard della popolazione sia di 3 Kg. Il lotto B è composto da animali di "scarto" con peso non omogeneo; in base ai risultati ottenuti da una serie limitata di pesate preliminari, mi aspetto che la deviazione standard della popolazione sia di 10 Kg. Utilizzando uno dei software suggeriti alla fine di questo articolo, stabilisco che il numero di soggetti da pesare (campione) è pari a 9 (3% dei suini) per il lotto A mentre, per il lotto B sarà pari a 73 (24% dei suini).

**Figura 4.** Distribuzione del peso vivo di 2 gruppi di suini:

A) peso vivo medio: 100 Kg; DS: 2 Kg; il 95% dei soggetti è compreso nell'intervallo di peso: 96-103 Kg

B) peso vivo medio: 100 Kg; DS: 4 Kg; il 95% dei soggetti è compreso nell'intervallo di peso: 92-108 Kg;




**BIBLIOGRAFIA**

1. Bottarelli E., Ostanello F. Epidemiologia. Teoria ed esempi di medicina veterinaria, Edagricole, Milano, 2011.
2. Thrusfield M. Veterinary epidemiology, 3rd ed. Blackwell Science Ltd, Oxford, 2007
3. Cannon R.M., Roe R.T. Livestock Disease Surveys: a Field Manual for Veterinarians. Australian Government Publishing Service, Canberra, 1982.

**Siti con software gratuiti per la stima della dimensione campionaria**

1. http://epitools.ausvet.com.au/content.php?page=SampleSize (in inglese)
2. http://www.winepi.net/uk/index.htm (in inglese)
3. http://www.quadernodiepidemiologia.it/epi/campion/dimens.xls (in italiano, foglio Excel per il calcolo della numerosità del campione necessaria stimare la prevalenza di una malattia)
4. http://www.quadernodiepidemiologia.it/epi/campion/cannon.xls (in italiano, foglio Excel per il calcolo della numerosità del campione necessaria e stimare la presenza di una malattia)