SECONDO APPUNTAMENTO CON LA SPERIMENTAZIONE IN AGRICOLTURA

Statistica descrittiva: prime informazioni dai dati sperimentali

La statistica descrittiva rappresenta la base di partenza per le applicazioni della statistica inferenziale che ha l'obiettivo di trarre conclusioni da delle ipotesi. Ha il compito di organizzare, riassumere, sintetizzare e presentare i dati in modo ordinato

di **Marco Delaiti, Mario Baldessari, Massimiliano Pasini**

In questo contributo viene descritto come organizzare e sintetizzare i dati in modo da poter evidenziare le loro caratteristiche importanti e soprattutto le informazioni da essi fornite. Per il momento non è importante se tali dati costituiscano l'intera popolazione o un campione più o meno rappresentativo estratto da essa. Dato che però l'analisi statistica si applica quasi sempre a dei campioni, possiamo assumere che i concetti che seguono siano riferiti a essi.

Quando si raccolgono dei dati, se il numero delle osservazioni fatte non è esiguo, ci si trova di fronte a un insieme disordinato di valori denominato dati grezzi o elementari, che probabilmente ci dicono poco o nulla, finché non sono stati messi in ordine in qualche modo. Di conseguenza, il primo problema che ci si trova ad affrontare è quello di **calcolare degli indicatori sintetici, utilizzando metodiche numeriche e/o grafiche, che siano in grado di sintetizzare ciò che abbiamo rilevato in campo senza alterarne il senso complessivo. Questa parte della statistica è nota con il nome di statisticaw descrittiva.**

Le variabili

Le variabili, in ambito statistico, si suddividono in qualitative e quantitative.

- Le **variabili qualitative esprimono una qualità dell'individuo** (ad esempio colore o forma delle foglie o dei frutti). Esse non comportano una misurazione, ma una classificazione in categorie sulla base delle modalità con cui si presentano (grappolo spargolo o compatto, sano o danneggiato; foglia verde o gialla o necrotica, ecc.).
- Le **variabili quantitative sono quelle caratterizzate da valori numerici, e possono essere continue** (ad esempio l'altezza delle piante, il peso della produzione) o discrete (ad esempio la produzione per ceppo, il numero di insetti per foglia, ecc.).

Le distribuzioni di frequenza

Molto spesso, avendo a che fare con un numero elevato di dati, una prima sintesi consiste nel considerare le frequenze, cioè la numerosità dei singoli valori raccolti.

- La **frequenza assoluta** (o **frequenza di classe**) **è il numero degli individui che presentano un certo valore numerico** (per un carattere quantitativo) **o una certa qualità** (per un carattere qualitativo).
- **Classi di frequenza.** Se abbiamo a che fare con variabili quantitative, prima di calcolare le frequenze è conveniente suddividere l'**intervallo delle misure in una serie di classi di frequenza aventi la stessa ampiezza**. Il modo di scegliere le classi non è univoco: potremmo scegliere un numero differente di classi a seconda dei casi, ma in ogni caso non devono sovrapporsi e devono contenere tutti i dati. Si arriva così a compilare quella che viene definita la tabella di frequenza (vedi *approfondimento* a pag. 34) Si tenga presente che troppe classi potrebbero rendere la tabella illeggibile; troppo poche la rende-

APPROFONDIMENTO

Come costruire le classi di frequenza

I dati riportati nella *tabella A* sono il risultato di 100 determinazioni del calibro di un campione di mele.

TABELLA A - Calibro delle mele campionate (100 osservazioni)

88	79	67	85	83	76	74	53
75	80	70	72	74	70	51	77
71	67	65	86	74	85	54	66
86	60	68	78	64	62	64	73
60	65	61	81	61	89	55	67
72	83	58	54	63	58	50	65
82	67	75	66	67	89	64	64
71	63	55	89	86	50	62	65
50	78	73	78	53	71	52	79
66	57	68	64	63	69	52	89
75	80	89	54	85	89	85	54
86	61	78	63	66	54	54	77
75	75	77	65				

Il calibro delle mele osservato è una variabile numerica continua, come classi vengono scelti degli intervalli di valori (< 55, 56 – 60, …, 81 – 85, > 85). Ogni osservazione viene assegnata alla classe di appartenenza. In questo modo si costruisce la tabella di frequenza (*tabella B*). In questa tabella la prima colonna indica la classe, la seconda la frequenza assoluta, detta semplicemente frequenza di classe, ossia il numero delle osservazioni che cadono in ciascuna classe; la terza colonna indica la frequenza relativa, ossia il rapporto tra la frequenza assoluta e il numero totale di osservazioni (in questo caso 100); la quarta è la frequenza percentuale, ossia la frequenza relativa moltiplicata per 100.

TABELLA B - Sintesi dei dati in classi di frequenza

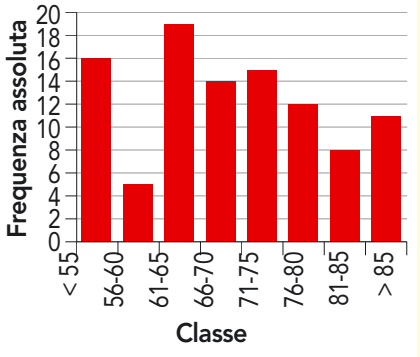
Classe	Frequenza		
	assoluta	relativa	percentuale
< 55	16	0,16	16
56-60	5	0,05	5
61-65	19	0,19	19
66-70	14	0,14	14
71-75	15	0,15	15
76-80	12	0,12	12
81-85	8	0,08	8
> 85	11	0,11	11

Molto usate risultano pure le rappresentazioni grafiche (*grafico A*).

L'osservazione dei dati in modalità grafica oltre a facilitarne la lettura può far notare irregolarità o comportamenti anomali non direttamente osservabili sui dati.

Le tipologie di grafici più utilizzati sono il diagramma a barre o il diagramma circolare (torta). ●

GRAFICO A - Rappresentazione grafica della frequenza



APPROFONDIMENTO

Media, moda e mediana

Come già osservato, uno dei limiti della media come misura della tendenza centrale è che essa è molto sensibile ai valori dei dati che cadono agli estremi dell'intervallo di variabilità; in questo senso può non rappresentare bene la collocazione dei dati. Nel nostro caso la mediana cade molto vicino alla media e questo ci fa capire che i dati si bipartiscono abbastanza bene attorno alla loro media. ●

TABELLA A - Esempio di media, moda e mediana (*)

Media	69,58
Mediana	68
Moda	89
Deviazione standard	11,20
Campo di variazione	39
Minimo	50
Massimo	89
Numero osservazioni	100

(*) Sono riportati nell'*approfondimento in alto* i valori descrittivi per il campione di 100 mele.

rebbero poco significativa. Una semplice regola pratica da adottare è quella di utilizzare un **numero di classi** pari alla radice quadrata del numero delle osservazioni (ad esempio: su 124 osservazioni le classi potrebbero essere 12).

Ricapitolando, la frequenza assoluta è il numero di osservazioni che cadono in una certa classe, la frequenza relativa è il rapporto tra frequenza assoluta e numero totale di osservazioni, la frequenza percentuale è la frequenza relativa moltiplicata per 100 (vedi *tabella B* dell'*approfondimento* in alto).

● **Valore centrale della classe.** Nel caso di variabili quantitative, una volta che i dati sono stati raggruppati come descritto sopra, ciascun valore esatto non viene più utilizzato, ma si rappresentano tutti i dati appartenenti a una certa classe utilizzando il suo punto medio, detto valore centrale della classe.

Media, moda e mediana

Nel caso di dati qualitativi, le frequenze delle classi costituiscono un'informazione sufficiente per un'analisi descrittiva dei dati. Nel caso di variabili quantitative, è possibile calcolare degli indici aggiuntivi, che rispecchino il più possibile le informazioni contenute nell'insieme dei dati. Uno di questi è la **media aritmetica**, che si calcola facendo la somma dei valori e dividendola per il numero delle osservazioni.

La **moda** è invece il valore o la classe a cui corrisponde la maggior frequenza.

La **mediana** è data dal valore che bipartisce la distribuzione di frequenza in modo da lasciare lo stesso numero di osservazioni a sinistra e a destra, una volta che si siano ordinati i dati in modo crescente o decrescente. Nel caso che il numero totale di osservazioni sia pari, la mediana è la media dei due valori centrali. La mediana è preferibile alla media quando si vogliono eliminare gli effetti

di valori estremi molto diversi dagli altri dati. Di solito nella descrizioni di risultati di sperimentazioni in campo agricolo sia la moda sia la mediana sono poco utilizzate (vedi *approfondimento* a pag. 34).

Indici dispersione

Come abbiamo visto, per sua stessa definizione la media non tiene conto della variabilità esistente fra i dati che la compongono: essa tende evidentemente ad appiattire tutte le differenze tra di loro. Vi sono infatti distribuzioni che, pur avendo la stessa media, sono molto diverse fra loro (vedi *approfondimento* a lato).

Quindi, quando si vuole descrivere un gruppo di valori è necessario utilizzare non solo un indice della tendenza centrale, ma anche un indice di variabilità, che ci consenta di stabilire come si collocano le singole osservazioni rispetto alla media.

Molto spesso viene usato il campo di variazione, che è la differenza tra la misura più bassa e la misura più alta. Non si tratta di un vero e proprio indice di variabilità, in quanto dipende solo dalle osservazioni estreme e non necessariamente cresce al crescere della variabilità degli individui.

Gli indici più importanti per la misura della variabilità di una distribuzione di frequenza sono la devianza, la varianza e la deviazione standard (vedi *glossario*).

La varianza e la deviazione standard sono detti indici di dispersione o indici di variabilità perché misurano la dispersione dei dati attorno alla media, in quanto il valore diventa tanto più grande quanto più i dati si discostano dalla media. In particolare la deviazione standard misura la dispersione delle osservazioni con la stessa unità di misura della media dei dati. Questa è la ragione principale per cui è più utilizzata rispetto alla varianza.

Il **coefficiente di variabilità** è un indice percentuale dato dal rapporto fra la deviazione standard e la media, moltiplicato per 100. Anch'esso è molto usato in statistica in quanto è indipendente dall'unità di misura.

Forma della distribuzione

Un'altra caratteristica dei dati che prendiamo in considerazione è la forma della loro distribuzione. Nonostante in agricoltura le popolazioni da campionare siano tantissime, si è notato che, come per la gran parte dei fenomeni biologici, possono essere ricondotte in ultima analisi a una distribuzione di frequenze denominata distribuzione normale, che ne rappresenta un po' il modello ideale (*grafico 1*).

APPROFONDIMENTO

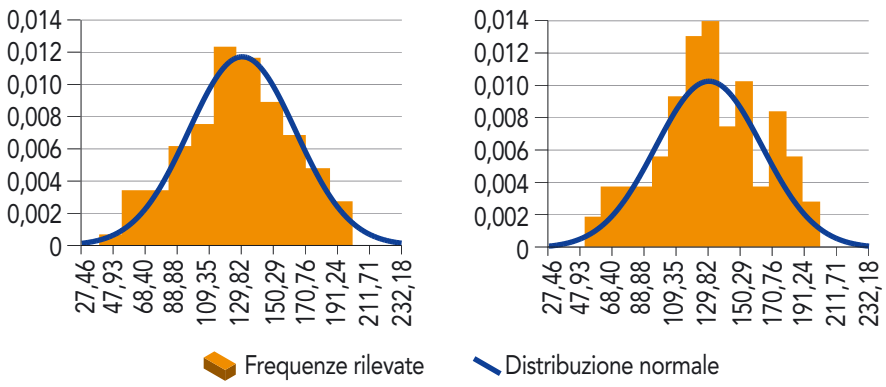
Capire la deviazione standard

Le tre aziende producono mediamente la stessa quantità di uva per ettaro. In realtà, però, all'interno delle tre situazioni esiste una diversa distribuzione delle produzioni. La deviazione standard ci aiuta nel capire questo; l'azienda 1 ha gli appezzamenti che producono costantemente dei valori vicini alla media dell'azienda, mentre l'azienda 2 ha grosse differenze nelle produzioni fra i diversi appezzamenti. ●

TABELLA A - Produzione di uva (q/ha) registrata in 3 aziende

Azienda	Rilievi (q/ha)				Media	Deviazione standard
	1°	2°	3°	4°		
1	90	110	100	100	100	7,07
2	80	83	100	137	100	22,68
3	80	90	110	120	100	15,81

GRAFICO 1 - Esempio di distribuzione dei dati e numero di classi



Nei due *grafici* viene rappresentata la distribuzione di frequenza degli stessi dati elementari. Nel primo caso vengono utilizzate 14 classi (a ***sinistra***) nel secondo 18 (a ***destra***). Un elevato numero di classi (nel secondo caso) porta a una eccessiva frammentazione e dispersione dei dati.

GLOSSARIO DEI TERMINI UTILIZZATI NEL TESTO

Asimmetria. In una distribuzione di dati di una popolazione o di un campione indica che i dati maggiori della media sono distribuiti in modo diverso da quelli minori della media.
Campo variazione. Differenza tra il valore massimo e minimo.
Classificazione. Operazione che si effettua sui dati elementari ogni volta che si ordinano a seconda di determinati criteri.
Coefficiente di variabilità. Rapporto percentuale tra la deviazione standard e la media.
Dati elementari o grezzi. Nelle prove sperimentali sono i dati che vengono raccolti in campo o in laboratorio a seguito di rilievi sulle variabili oggetto di studio.
Devianza o scarto quadratico medio. Sommatoria dei quadrati degli scarti di ciascun dato dalla media della popolazione o del campione.
Deviazione standard. Scostamento medio dei dati rispetto alla media.
Distribuzione normale. Distribuzione dei dati di un campione o di una popolazione secondo la curva a campana di Gauss (curva normale).
Frequenza. Numerosità di un determinato valore rilevato in una prova sperimentale o delle classi all'uopo formate.
Frequenza relativa. Rapporto percentuale tra la numerosità di una classe o di un valore e la numerosità totale del campione.
Gradi di libertà. In un insieme di osservazioni o in un campione: numero di dati veramente indipendenti, pari al numero totale di dati meno uno.
Indici di tendenza centrale. Indici che esprimono il centro attorno a cui ruotano i dati.
Indici di dispersione. Indici che esprimono il livello di variabilità dei dati.
Media. In una popolazione o in un campione, è il rapporto tra la somma di tutti i valori osservati e la loro numerosità.
Mediana. In una popolazione o in un campione ove i dati siano stati ordinati in forma crescente, è il valore centrale, oppure, se il numero di dati è pari, è la media dei due valori centrali.
Misura o misurazione. L'atto di rilevare il valore di una variabile attraverso l'uso di uno strumento (strumento di misura).
Misure di dispersione. Parametri di una popolazione o di un campione che descrivono la distribuzione dei dati intorno al valore centrale.
Misure di tendenza centrale. Parametri di una popolazione o di un campione che descrivono il centro attorno al quale si collocano i dati.
Moda. Valore più numeroso, cioè quello che ha la maggiore frequenza.
Popolazione. Insieme completo di individui o entità caratterizzate da determinate proprietà, alle quali è associato un valore (variabile).
Scala. Insieme di caratteristiche definito preventivamente allo scopo di migliorare la stima o rilievo di una variabile su un gruppo di oggetti o individui.
Variabile continua. Variabile che assume qualsiasi valore all'interno del campo di esistenza.
Variabile discreta. Variabile che assume solo determinati valori all'interno del campo di esistenza.
Variabile qualitativa. Variabile non numerica rappresentata da qualità o caratteristiche qualitative.
Variabile quantitativa. Variabile espressa da numeri o quantità.
Varianza. Rapporto tra la sommatoria dei quadrati degli scarti e la numerosità della popolazione. Nel caso di un campione, rapporto tra la sommatoria dei quadrati degli scarti e i gradi di libertà. ●

La curva normale ha una forma a campana e la distribuzione dei dati è simmetrica rispetto a una linea verticale che rappresenta la media. Questo indica che la frequenza dei valori superiori alla media è esattamente uguale alla frequenza dei valori inferiori alla media.

Partendo da una **popolazione** distribuita normalmente, con media «μ» e deviazione standard «σ», dopo opportuni passaggi matematici, si può dimostrare che la frequenza degli individui compresi tra μ + 1,96σ e μ – 1,96σ è pari al 95% e che la frequenza dei valori compresi tra μ + 2,575σ e μ – 2,575σ è pari al 99%.

Possiamo quindi calcolare per estensione quale è la frequenza di ogni possibile individuo che appartiene alla popolazione.

Nel caso dei **campioni**, la media si indica con «m» e la deviazione standard con «s» e quando si pubblicano i risultati delle prove sperimentali, molto spesso si utilizza la scrittura m ± s, che rappresenta l'intervallo di dati compreso tra m – s e m + s, al cui interno ricade il 68% dei dati. Ad esempio se il campione ha media 4 e deviazione standard 1,5 si scrive 4 ± 1,5. Ciò significa che il 68% dei dati è compreso tra il valore 2,5 (= 4 – 1,5) e il valore 5,5 (= 4 + 1,5).

Marco Delaiti
Mario Baldessari
Fondazione E. Mach
Istituto agrario S. Michele all'Adige (Trento)
Massimiliano Pasini
Agrea centro studi Verona

APPROFONDIMENTO

L'importanza della distribuzione normale

La distribuzione normale è importante in statistica per tre motivi fondamentali.

- Le variabili continue studiate in biologia e agricoltura seguono, almeno approssimativamente, una distribuzione normale.
- In virtù di quanto riportato nel testo e tenendo presente che la frequenza di una variabile ci dice anche la probabilità che abbiamo di estrarre quella variabile dalla popolazione, possiamo anche stimare la probabilità di estrarre una certa misura o un certo intervallo di misure da una popolazione distribuita secondo la curva.
- Proprio la distribuzione normale è alla base dell'inferenza statistica, che ha il compito principale di fare ipotesi probabilistiche sulla vera media (μ) delle popolazioni, spesso ignota dato che si lavora sui campioni estratti in maniera casuale. ●

Per commenti all'articolo, chiarimenti o suggerimenti scrivi a:
redazione@informatoreagrario.it

Per consultare gli approfondimenti e/o la bibliografia:
www.informatoreagrario.it/rdLia/12ia20_6358_web

ALTRI ARTICOLI SULL'ARGOMENTO

- *Capire facilmente l'analisi statistica in agricoltura*
  Pubblicato su *L'Informatore Agrario* n. 17/2012 a pag. 36.

www.informatoreagrario.it/bdo

ARTICOLO PUBBLICATO SU L'INFORMATORE AGRARIO N. 20/2012 A PAG. 33

L'INFORMATORE AGRARIO

Statistica descrittiva: prime informazioni dai dati sperimentali

TABELLA A - Esempio e formula degli indici di dispersione

In base alle definizioni date nel testo, consideriamo un campione estratto da una popolazione in cui X_i sia il valore generico del dato **i**-esimo e **n** il numero complessivo dei dati. Nel caso di una popolazione la media è indicata con la lettera X_i e il numero complessivo di dati *N*.

Indice	Popolazione	Campione
Media	$\mu = \frac{x_1 + x_2 + \ldots\ldots + x_n}{N}$	$m = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n}$
Devianza o scarto quadratico medio	$SQ = \sum(x_i - \mu)^2$	$sq = \sum(x_i - m)^2$
Varianza	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum(x_i - m)^2}{n - 1}$
Deviazione standard	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(x_i - m)^2}{n - 1}}$
Coefficiente di variabilità	$C.V. = \frac{\sigma}{\mu} \times 100$	$C.V. = \frac{s}{m} \times 100$

Esempio pratico

È stato misurato il peso di 5 frutti di anguria dopo un trattamento con un biostimolante. Il risultato è: 5,5 kg; 4,9 kg; 5,2 kg; 6 kg; 4 kg.

$m = 5{,}12$

$sq = (5{,}5 - 5{,}12)^2 + (4{,}9 - 5{,}12)^2 + (5{,}2 - 5{,}12)^2 + (6 - 5{,}12)^2 = 2{,}23$

$s^2 = \frac{(5{,}5 - 5{,}12)^2 + (4{,}9 - 5{,}12)^2 + (5{,}2 - 5{,}12)^2 + (6 - 5{,}12)^2}{5\text{-}1} = 0{,}56$

$s = \sqrt{0{,}56} = 0{,}75$

$C.V. = \frac{0{,}75}{5{,}12} \times 100 = 0{,}56$