

Alcune considerazioni sulla misurazione delle spore

di Daniele Uboldi

March 15, 2010

1 Premessa

La misurazione delle spore può essere convenientemente validata solo per spore da sporata; perchè sono le uniche mature.

E' inopportuno misurare spore da lamella (basidiocarpo fresco) o da exiccata.

Da fresco sono numerose le spore immature; da exsiccata possono risultare errate per via del rigonfiamento nella rigenerazione dei tessuti .

Sulla misurazione delle spore e degli altri elementi microscopici (basidi, cistidi.....) c'è molta confusione; prima ancora che sui metodi di rilevazione e gestione dei dati, sulla ragione stessa della misurazione.

La domanda non è retorica, perchè si misura?

La risposta potrebbe essere: “ per avere un riscontro oggettivo che consenta di definire i caratteri tipici del soggetto indagato”

Alcuni autori pensano che sia importante misurare per “distinguere” tra affini.

Dunque la conoscenza, per chi ha questa convinzione, non è importante “ in sè” ma per fini comparativi.

Ammesso che questo motivo possa essere valido, ciò non esclude che la misura debba essere anche autoreferenziale e “finita”; senza secondi scopi.

Dunque non ci possiamo accontentare di una misura che ci permetta di dire che il *typus* appartiene ad “A” piuttosto che a “B”: sarebbe un criterio approssimativo.

Quando si fa statistica bisogna avere chiaro un fatto basilare: **le misure campionarie devono potere essere generalizzate sull'insieme della popolazione da cui il campione è estratto.**

Tramite una porzione piuttosto piccola (campione) è possibile stimare qualche cosa di infinitamente grande (popolazione).

Stimare non è facile: richiede una corretta applicazione dei metodi statistici; cosa che, purtroppo, manca spesso nei libri e articoli di micologia, anche ad alto livello.

Su questi aspetti è necessario che si avvii una discussione che conduca a standardizzare le procedure e ad adottare comportamenti uniformi e validati a livello internazionale.

E' necessario che nei nostri convegni, a partire dai “comitati scientifici” si affronti e si definisca questo problema.

1.1 Nota:

Consiglio a tutti la lettura e la comprensione di quanto esposto di seguito. Tuttavia, per chi trovasse ostica la materia o fosse solo interessato agli aspetti pratici, consiglio l'adozione del software:

<http://www.freetiamo.altervista.org/index.php/ufficio/software-statistici/235-openstat-software-statistico-free.html>

Si tratta di un pacchetto statistico elaborato da Bill Miller, che si può scaricare gratuitamente da Internet; molto ben fatto e intuitivo (funziona col sistema a “tendina” tipico di Windows). E' inoltre possibile scaricare il corposo manuale, in lingua inglese.

Con esso si può facilmente ricavare media, scarto quadratico medio, moda, mediana, intervallo di confidenza, momenti di asimmetria e curtosi; inoltre grafici di probabilità normale, test di normalità ed altri numerosi test statistici assai utili alla comprensione del fenomeno indagato.

2 L'errore nelle misure

Il concetto di errore in statistica non è assimilato a “sbaglio” quanto, piuttosto ad “errare”: muoversi, discostarsi da un teorico “valore vero”.

Quando si ha a che fare con un campione, nel nostro caso di reperti micologici, dobbiamo avere ben presente che il campione “incorpora” almeno tre errori:

1) **L'errore dovuto alla variabilità intrinseca** del campione rispetto alla popolazione dal quale è estratto. Tale variabilità è relativa sia al fungo dal quale la sporata proviene, sia alla popolazione della specie a cui il carpoforo appartiene.

Può essere dovuta a fattori biochimici, all'altitudine di crescita del carpoforo indagato, dal versante su cui è cresciuto, dalla temperatura, umidità.....

Le cause della variabilità possono essere migliaia e per lo più ignote.

Se non possiamo conoscerne l'eziologia, quanto meno dobbiamo sapere che c'è e tenerne conto.

2) **L'errore dell'osservatore**, compiuto nel momento in cui misura.

Tali errori possono essere difetti della vista, distrazione, cattiva postura d'innanzi al microscopio...

3) **L'errore dello strumento** (microscopio) utilizzato per le rilevazioni: strumento tarato male o con difetti di assemblaggio, di costruzione, aberrazione nell'ottica.....

Per quanto l'osservatore si sforzi di mettere accuratezza nella propria ricerca (accuratezza dovuta ed auspicabile) questi tre tipi di errore sono **ineliminabili**.

Dunque, gli errori, se non eliminabili, devono essere attentamente “pesati”.

2.1 Caratteristiche del campione.

2.1.1 Gli outliers

Un errore che in molti commettono è quello di “scegliere” quali spore misurare.

Ad “occhio” taluni decidono che alcune delle spore che appaiono nell'oculare del microscopio siano o troppo grandi o troppo piccole che, in definitiva, si scostino dalla normalità; per cui le escludono.

Ebbene non si fa: il ricercatore non deve essere “arbitro” di ciò che vede ma, semplicemente, reporter, secondo il principio: “ vedo, misuro, trascrivo”.

E' indubbio che possiamo avere a che fare con degli outliers; però esistono metodi statistici per segregarli ed annullare il loro condizionamento sui parametri della distribuzione (media e varianza).

Se scegliessimo cosa misurare, correremmo il rischio, oltre agli outliers , di non misurare anche spore che outliers non sono ma che, semplicemente, si trovano agli estremi della curva di distribuzione.

In un capitolo successivo mostreremo come tenere conto delle spore abnormi senza doverle scegliere arbitrariamente

2.1.2 La natura del campione

Come dicevamo in premessa, il campione non è importante in “se” ma in quanto rappresenta la popolazione dalla quale è stato estratto. Infatti a noi non interessa sapere le misure di “quel”

campione ma, a partire da esso, quelle della popolazione tramite la stima; che deve essere la piu' accurata possibile.

Per fare un esempio comprensibilissimo, il medico preleva un campione di sangue al paziente, per valutare, tramite parametri, il suo stato di salute. Il campione di sangue non è importante "in se" ma per il fatto che le risultanze possono essere generalizzate a tutta la popolazione sanguinea appartenente a quel paziente e riferita al momento del prelievo.

Dunque alla base del campionamento vi deve essere la possibilità di **generalizzare i risultati**.

Questo processo logico-matematico in statistica prende il nome di "inferenza".

Perchè il campione sia rappresentativo deve avere le seguenti caratteristiche:

- 1) **essere estratto casualmente**
- 2) **essere rappresentativo**
- 3) **essere stocasticamente indipendente** (cioè non deve essere condizionato)

Sono norme intuitive ma giova metterle in rilievo.

Per esempio la misura delle spore da lamella non rispetta questi tre criteri, in quanto non è rappresentativa e non è dimostrata la sua indipendenza stocastica.

2.1.3 L'ampiezza del campione.

In termini teorici, piu' il campione è grande e piu' diventa accurata l'inferenza sulla popolazione.

Però ciò si scontra con degli aspetti pratici:

1) un campione eccessivamente grande non apporta grandi cambiamenti nei risultati. Attorno a $N=30$ osservazioni i parametri di media e varianza tendono a stabilizzarsi, con successive modifiche poco apprezzabili.

2) Un campionamento eccessivamente grande comporta dispendio di tempo, di energie.

E' sempre meglio privilegiare la qualità, l'accuratezza nell'indagine, piuttosto che la quantità.

In ogni caso è piu' opportuno l'esame di piu' raccolte, piuttosto che "accanirsi" su di uno o pochi campioni.

Non corrisponde a verità che i campioni debbano avere tutti uguale numero di unità statistiche.

Si può benissimo avere, per esempio, $C_1 = 30$; $C_2 = 18$ $C_k = 22$ $C_n = 32$. (1)

2.1.4 Le misure del campione e i parametri indagati

Non entreremo nel merito di cosa e come si misura, perchè questi aspetti sono stati trattati ampiamente nei testi di microscopia ed esistono precisi riferimenti in letteratura.

A noi preme rimanere legati agli aspetti teorici e pratici dei metodi statistici.

Premesso che la funzione prima della statistica è quella di generalizzare sulla popolazione gli elementi cognitivi campionari, si tratta di individuare quali siano i parametri da monitorare.

Ce ne sono molti, ma noi ne prenderemo in considerazione solo due, tipici della statistica parametrica: la media e la varianza.

La prima è una misura della "concentrazione" o, meglio " della tendenza centrale" la secondo è una misura di "dispersione" dei valori attorno alla media.

Dunque dal campione dobbiamo trarre almeno queste due importantissime informazioni: la media aritmetica , la varianza (e la sua radice quadrata, lo scarto quadratico medio).

I campioni, ovviamente appartenenti alla medesima popolazione ed estratti da essa con i criteri anzidetti, possono essere sommati tra loro.

Per ciascun campione calcoleremo la media:

$$C_1 = \frac{\sum_{i=1}^{30} x_i \cdot f_i}{30}$$

Procederemo analogamente per $C_2 \dots C_k \dots C_n$
 Poi calcoleremo lo scarto quadratico medio:

$$S_{C_1} = \sqrt{\frac{\sum_{i=1}^{30} (x_i - \bar{x})^2}{29}}$$

Procederemo analogamente per $S_{C_2} \dots S_{C_k} \dots S_{C_n}$
 Poi sommeremo le medie ponderate dei campioni per ottenere la media delle medie:

$$\bar{\bar{X}} = \frac{C_1 = \frac{\sum_{i=1}^{30} x_i \cdot f_i}{30} + C_2 = \frac{\sum_{i=1}^{18} x_i \cdot f_i}{18} + \dots + C_k = \frac{\sum_{i=1}^{22} x_i \cdot f_i}{22} + \dots + C_n = \frac{\sum_{i=1}^{32} x_i \cdot f_i}{32}}{102} *$$

Lo stesso faremo per trovare lo s.q.m. delle medie:

$$S_{c_1+c_2+\dots+c_k+\dots+c_n} = \frac{S_{C_1} = \sqrt{\frac{\sum_{i=1}^{30} (x_i - \bar{x})^2}{29}} + S_{C_2} = \sqrt{\frac{\sum_{i=1}^{18} (x_i - \bar{x})^2}{17}} + \dots + S_{C_k} = \sqrt{\frac{\sum_{i=1}^{22} (x_i - \bar{x})^2}{21}} + \dots + S_{C_n} = \sqrt{\frac{\sum_{i=1}^{32} (x_i - \bar{x})^2}{31}}}{98} *$$

Nota: il denominatore nel calcolo dello s.q.m. campionario è sempre dato da (n-1)

* (vedi esempio in appendice)

Quando abbiamo a che fare solo con due seriazioni, possiamo confrontare le medie di due campioni, tramite il t di Student, secondo la formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\frac{S_1}{\sqrt{n_1}} - \frac{S_2}{\sqrt{n_2}}}$$

dove:

$\bar{X}_1 - \bar{X}_2$ è la differenza tra due medie campionarie

$\frac{S_1}{\sqrt{n_1}} - \frac{S_2}{\sqrt{n_2}}$ è la differenza dei rispettivi errori standard

Le tavole dei valori critici di t a prestabiliti livelli di rischio (per esempio $\alpha = 0.05$) ci forniscono il rischio che corriamo di rifiutare l'Ipotesi Nulla quando è vera.

2.1.5 Dal campione alla popolazione: l'inferenza

Quasi tutti i testi di micologia, in riferimento alla misurazione delle spore, riportano una scrittura di questo tipo:

$$(8)9 \div 11, 5(13)x (3,5) 4,8 \div 6, 7(7,5)\mu m \quad (2)$$

Con tale notazione si evidenzia che è stata rinvenuta una lunghezza minima di (8); un intervallo in cui ricadono la maggior parte delle lunghezze, pari a $9 \div 11, 5$ e una misura massima pari a (13) μm . Quanto alla larghezza (3,5) è il valore minimo; $4,8 \div 6, 7$ l'intervallo di maggiore frequenza e (7,5) μm il valore massimo.

Questa rappresentazione, seppure universalmente accettata, non è scevra da critiche; per tre motivi:

a) La notazione in questione è relativa a una misurazione campionaria: descrive cioè semplicemente alcuni valori rilevati in un generico campione c_k ; perciò sono riferiti solo ad esso e non a tutta la popolazione.

Inoltre non permette inferenze sulla popolazione dalla quale è estratto.

b) Non fornisce nessun elemento indicativo circa la misura della tendenza centrale (media) ne sulla misura della dispersione (scarto quadratico medio). Infatti scrivendo

$$” 9 \div 11,5”; “4,8 \div 6,7”$$

non si può sapere se la media della lunghezza sia piu' vicina a “9” oppure a “11,5” e quella della larghezza a “4,8” oppure a “6,7”

c) E' di poco significato quantificare in modo separato “lunghezza” e “larghezza” delle spore; in quanto esse vanno rilevate (ed analizzate) in modo congiunto.

In pratica dobbiamo ricorrere alla statistica bivariata e rilevare coppie dei caratteri “lunghezza” (X) e “larghezza” (Y) ; vale a dire:

$$x_1y_1; x_2y_2\dots; x_ky_k\dots; x_ny_n$$

E' vero che, quando si misura, vengono lette contestualmente. per ciascuna spora, lunghezza e larghezza ma, non essendo considerate quali coppie di caratteri, le informazioni su ciascuna delle coppie viene messa nel “mucchio” rispettivamente, delle lunghezze e larghezze .Come già detto, il metodo piu' corretto è quello di ricorrere alla statistica bivariata, adottando metodi (regressione lineare, regressione multipla) di cui parleremo a parte.

Trascuriamo per un attimo la “larghezza” e supponiamo di volere rilevare solo le misure della “lunghezza”.

Immaginiamo di avere a disposizione la sporata di un certo carpoforo e che quattro campioni vengano da essa estratti e consegnati ad altrettanti osservatori.

Supponiamo anche che di ogni campione di numerosità n, a lavoro terminato avremo la seguente situazione:

$$Osservatore1 - C_1 \text{ con } \bar{X}_1 S_1$$

$$Osservatore2 - C_2 \text{ con } \bar{X}_2 S_2$$

$$Osservatore3 - C_3 \text{ con } \bar{X}_3 S_3$$

$$Osservatore4 - C_4 \text{ con } \bar{X}_4 S_4$$

Dove:

C indica il campione replicato 1,2..k..n volte

\bar{X} è la media campionaria

S è lo scarto quadratico medio (o deviazione standard) del campione

In ragione degli errori presenti nella rilevazione (con le caratteristiche evidenziate in premessa) molto probabilmente si avrà una situazione del tipo:

$$C_1 \neq C_2 \neq C_3 \neq C_4$$

Ovvero, per quanto siano accurate le nostre osservazioni, tarati gli strumenti, avremo sempre a che fare con medie e varianze diverse tra campione e campione.

Questo non significa che i campioni siano statisticamente diversi; significa solo che l'errore contenuto nella media e varianza dei campioni non ci fornisce, di per se, alcuna informazione circa la media vera della popolazione.

Senza gli errori (cioè una generica variabile casuale ξ in grado di riassumere tutte le accidentali fonti di variabilità) I campioni sarebbero uguali, in quanto estratti della medesima popolazione per cui

$$\bar{X}_1 - \bar{X}_2 = 0; \bar{X}_3 - \bar{X}_4 = 0 \text{ ecc.}$$

Analogamente:

$$\frac{S_1}{S_2} = 1; \frac{S_3}{S_4} = 1.. \text{ ecc.}$$

In pratica avremmo medie nulle e varianze sempre uguali a 1, in rapporto tra loro.

Cerchiamo di capire meglio questi concetti. Abbiamo già accennato agli “errori”: il primo errore, abbiamo detto, è costituito dalla naturale propensione delle spore (ma il discorso si può generalizzare alla gran parte dei fenomeni naturali e non solo ad essi) a variare. Tali variazioni possono essere di tipo **causale** (allora la causa va individuata e dimostrata) oppure **casuali** (in genere si tratta proprio di eventi di questa natura).

Una distribuzione che abbia media $\mu = 0$ e varianza $\sigma^2 = 1$ è *una distribuzione normale standardizzata* $N(0,1)$.

la cui densità di probabilità è data dalla formula:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

dove:

e è il numero di Neper = 2,71828

Z è una nuova variabile standardizzata, ricavata dal rapporto : $\frac{X_i - \bar{X}_1}{\sigma}$

A ben guardare a noi, atteso che la media, nella distribuzione normale standardizzata è sempre uguale a zero e la varianza sempre uguale a uno, quello che interessa è solo la quantificazione dell'errore; cioè di quella componente accidentale “ ξ ” di cui parlavamo, che riassume in se tutte le possibili fonti di variabilità; sia intrinseche alle unità statistiche indagate, sia alle fluttuazioni casuali dovute all'osservatore e allo strumento.

Il significato di tutto questo ragionamento sta nel fatto che noi **non siamo interessati ad ottenere dati campionari fini a se stessi; ma dati campionari in quanto rappresentativi dell'intera popolazione;**

per cui dobbiamo ricercare un nuovo modo di definire l'intervallo entro il quale si stima che sia contenuta la **media vera**. Questa è una delle ragioni per cui la (2) appare del tutto inefficace.

2.1.6 Medie campionarie e media vera

Una volta noti i dati campionari, è necessario compiere inferenze sulla popolazione costruendo l'intervallo della media vera.

Come detto poc'anzi a ciascuna delle n seriazioni statistiche ($C_1, C_2, \dots, C_k, \dots, C_n$) corrisponderanno altrettante medie e varianze diverse tra loro.

Dunque quale sarà la media “vera”, quella del primo campione, del secondo, dell'ultimo, la somma di tutti i campioni indagati o nessuna di queste?

Non è dato di saperlo perchè **la media vera è quasi sempre incognita.**

La possiamo **solo stimare.**

La media, per definizione è la tendenza centrale di una distribuzione; per cui siamo confidenti del fatto che essa si possa trovare compresa tra i valori minimo e massimo che abbiamo rilevato.

Ciò può essere vero ma può, in linea di principio, essere anche falso.

Dunque si può prospettare un quadro di questo tipo:

1) affermare che **la media vera sia contenuta** nell'intervallo tra minimo e massimo che abbiamo rilevati, **quando l'ipotesi è vera**

2) idem come sopra, però **quando l'ipotesi è falsa**

E' necessario allora **accettare un rischio (rischio α) cioè di respingere la Ipotesi Nulla (la media vera è contenuta nell'intervallo) quando essa è vera.** Questo errore è chiamato anche "errore di primo tipo" Il rischio α di solito è posto al 10%, 5%, 1 % ($\alpha = 0,1$; 0,05 ; 0,01) che corrisponde ad intervalli di confidenza del 90% ; 95 % ; 99 %.

Ovvio che piu' il rischio alfa è grande , piu' sarà piccolo l'intervallo; viceversa, piu' sarà piccolo il rischio, piu' sarà grande l'intervallo.

Piu' α è piccolo minore è il rischio di commettere un errore di "primo tipo"

Dunque porremo la disequazione:

$$\boxed{\bar{X}_k < \mu < \bar{X}_j}$$

Dove, ad una generica seriazione k viene attribuito un valore minimo e alla seriazione j il valore massimo della media aritmetica.

Alla disequazione dobbiamo aggiungere la formulazione di rischio, che poniamo $\alpha = 0,05$ (intervallo di confidenza 95%)

Essendo incognita la varianza σ^2 della popolazione , dovremo ricorrere al "**t di Student**" la cui formula risolutiva è:

$$\mu = \bar{X} \pm t_{\frac{\alpha}{2}(n-1)} \cdot \frac{S}{\sqrt{n}} \quad (3)$$

dove:

- μ è la media vera della popolazione
- n è il numero di dati
- S è la deviazione standard calcolata sui dati del campione
- $t_{\frac{\alpha}{2}(n-1)}$ è il valore tabulato di t ad un prestabilito rischio α con n-1 gradi di libertà.
- $\frac{S}{\sqrt{n}}$ è l'errore standard di ogni campione, dato dal rapporto tra lo s.q.m. e le unità statistiche rilevate.

Esempio:

sia rilevata da un osservatore la lunghezza di 14 spore di *Xerocomus chrysenteron* (Simonini, Pagine di Micologia, 1998) , come riportato in tabella:

Table 1: Spore di Xerocomus chrysenteron

	Unità stat.	Valore osserv.	Media	Scarto	Scarto ²
			$\bar{X} = \frac{\sum x_i}{n}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	1	14,25	13,18	1,07	1,15
	2	13,04	13,18	-0,14	0,02
	3	12,36	13,18	-0,82	0,67
	4	14,45	13,18	1,27	1,62
	5	14,23	13,18	1,05	1,11
	6	13,83	13,18	0,65	0,43
	7	11,96	13,18	-1,22	1,48
	8	11,94	13,18	-1,24	1,53
	9	13,25	13,18	0,07	0,01
	10	12,34	13,18	-0,84	0,7
	11	14,42	13,18	1,24	1,55
	12	11,90	13,18	-1,28	1,63
	13	13,09	13,18	-0,09	0,01
	14	13,40	13,18	1,22	0,05
Media	$\bar{X} = \frac{\sum x_i}{n}$	13,18			
Devianza	$\sum (x_i - \bar{X})^2$				11,94
Dev. Standard	$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$				0,96
Varianza	S^2				0,92

A noi interessa una “protezione bilaterale” cioè dividiamo a metà il rischio α sulle due “code” della distribuzione; perchè la media vera potrebbe trovarsi sia fuori dal valore minimo che da quello massimo della disequazione.

Sulle tavole statistiche possiamo leggere i valori critici $t_{0,025|13} \text{ g.d.l} = 2,160$ (nel nostro caso i gradi di libertà sono 13, perchè 14 le osservazioni) sostituendo alla **(3)** si avrà:

$$\mu = 13,18 \pm 2,160 \cdot \frac{0,96}{\sqrt{14}} = 12,62 \leq \mu \leq 13,73$$

Ciò significa che la **media vera** della lunghezza, con una probabilità non superiore al 5% di sbagliare (respingere l’ipotesi Nulla quando è vera), è compreso nell’intervallo tra

$$12,62 \div 13,73 \mu m$$

Dunque la scrittura completa e corretta sarà:

Table 2: Dati caratteristici della lunghezza delle spore di Xerocomus chrysenteron

Indicatori	Valori
Numero di osservazioni	14
Media campionaria	13,18
Dev. std. del campione	0,96
Rischio α	0,05 (95% prob.)
Intervallo di confidenza	$12,62 \leq \mu \leq 13,73$

Analoga procedura deve essere compiuta per la larghezza.

Successivamente, ma questo è un’altro aspetto del problema, sarà bene considerare come caratteri congiunti (xy) la lunghezza e la larghezza delle spore; valutare il loro grado di associazione e gli altri parametri della regressione, tramite gli strumenti della statistica bivariata.

3 Appendice

3.1 Come sommare piu' campioni della stessa popolazione.

Supponiamo di avere raccolto quattro campioni di *Crepidotus variabilis*, prelevati da altrettanti carpori, e di avere rilevato le seguenti caratteristiche, relative alle sole lunghezze (in μm):

Campione	n	media	s.q.m.
1	30	5,8	0,43
2	18	5,2	0,66
3	22	6,2	0,81
4	32	6,4	0,99

Il passo successivo è quello di estendere l'elaborazione dei dati e di ricercare l'errore standard, i valori critici del t di Student ai corrispondenti g.d.l. e l'intervallo di confidenza della media vera di ciascun campione:

Campione	n	\bar{X}	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$	$\frac{s}{\sqrt{n}}$	g.d.l.	$t_{\frac{\alpha}{2}}=0,025$	interv. di conf. 95%
1	30	5,8	0,43	0,08	29	2,36	$4,79 < \mu < 6,81$
2	18	5,2	0,66	0,16	17	2,44	$4,81 < \mu < 5,59$
3	22	6,2	0,81	0,17	21	2,41	$5,79 < \mu < 6,61$
4	32	6,4	0,99	0,18	31	2,35	$5,98 < \mu < 6,82$

Note:

\bar{X} : media campionaria

$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$: scarto quadratico medio (s)

$\frac{s}{\sqrt{n}}$: errore standard

$t_{\frac{\alpha}{2}}$: valore critico del t di Student nelle due code, ai rispettivi gradi di libertà (n-1)

$\mu = \bar{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$: formula dell'intervallo di confidenza

Tutti i passaggi nella tabella precedente non sono indispensabili: li abbiamo mostrati solo per mettere in evidenza come, al variare del numero delle unità statistiche, della media e dello s.q.m., cambi anche l'errore standard, il valore critico del t di Student e, conseguentemente, l'intervallo di confidenza.

Ora calcoliamo la media delle medie:

$$\bar{\bar{X}} = \frac{5,8 * 30 + 5,2 * 18 + 6,2 * 22 + 6,4 * 32}{102} = 5,97 \mu m$$

Lo scarto quadratico medio:

$$\bar{s} = \frac{0,43 * 30 + 0,66 * 18 + 0,81 * 22 + 0,9 * 32}{98} = 0,76$$

L'errore standard:

$$\frac{0,76}{\sqrt{102}} = 0,08$$

Il valore critico del t di Studente con 101 g.d.l. è: 2,27

per cui:

$$5,97 - 2,27 * 0,08 < \mu < 5,97 + 2,27 * 0,08$$

ovvero:

$$5,79 < \mu < 6,15 \mu m \quad (4)$$

che rappresenta l'**intervallo di confidenza della media vera** della lunghezza delle spore del *C. variabilis*.

Siamo dunque confidenti al 95% che, con formulazione di rischio $\alpha = 0.05$, protezione bilaterale (due code della distribuzione) la (4) sia l'intervallo entro cui si trovi la "media vera" della popolazione, col rischio di respingere l'Ipotesi Nulla, quando è vera, non più alto del 5%.

Naturalmente dovremo compiere lo stesso lavoro, con gli stessi metodi, anche per le larghezze sporali.

3.2 Gli outliers

In gergo statistico i dati anomali vengono denominati "outliers".

Non è raro che in una seriazione di dati appaiano delle unità statistiche i cui valori producano "genuina sorpresa" Cioè balzino all'occhio come valori abnormi rispetto a tutti gli altri.

Questa anomalia è rinvenuta, prima che da qualsiasi altro accorgimento, dal "buon senso" e dalla pratica corrente.

Immaginiamo la seguente successione numerica:

$$\boxed{10^{12}; 10^{25}; 11^{06}; 9^{89}; 9^{33}; 18^{36}; 10^{14}}$$

La nostra "genuina sorpresa" ci dice che quel "18³⁶" stona, come un valore estraneo alla seriazione.

In realtà ci può essere un "perchè" in questo valore "anomalo" e sarebbe un vero peccato se noi, un pò superficialmente, decidessimo di "liquidare" il caso semplicemente ignorando l'esistenza di tale valore.

E' fondata la preoccupazione che il dato anomalo (peggio ancora se sono più di uno!) alteri i parametri distributivi. Infatti, sia la media che lo scarto quadratico medio sono influenzati dall'unità statistica notevolmente diversa dalle altre.

Piuttosto che accantonarla, non prenderla in considerazione, considerarla come un aspetto indesiderato che disturba le nostre tranquille certezze, dobbiamo trovare il modo di "neutralizzarla".

A questo proposito possiamo ricorrere al metodo denominato: **Extreme Studentized Deviate** (ESD)

basato sul (t) di Student.

Inizialmente dobbiamo porci una domanda, semplice ma basilare: "quanto deve distare un valore dal resto delle altre unità statistiche per essere ritenuto un outlier"?

Se la "distanza" è notevole, la semplice "occhiata" può metterci sull'avviso che ci troviamo di fronte a un dato anomalo.

Le cose, invece, si complicano se la "distanza" è più ridotta ed entra nella regione critica, cioè in quella "zona grigia" dove accettare/rifiutare diventa difficile, se non impossibile, applicando il solo "buon senso".

Davanti a spore aberranti, decisamente riconosciute come tali, vi può essere la tentazione, per me ingiustificata, di non considerarle. Tutto però si complica se l'anomalia non è così evidente rispetto alla distribuzione normale.

Quindi è legittima la domanda: " quanto deve distare un valore dalle altre unità statistiche, per essere considerata una outlier"?

La risposta deve prendere in considerazione tre fattori:

- 1) la distanza del dato "anomalo" dalla media ($X_k - \bar{X}$)
- 2) la deviazione standard del campione (S)
- 3) La numerosità campionaria (N)

La ESD è data da:

$$ESD = \max_{j=1, \dots, N} \frac{|X_j - \bar{X}|}{S}$$

Tabella 3: Valori critici ESD

N	$\alpha=0.05$
30	2,91
35	2,98
40	3,04
45	3,09
50	3,13
60	3,20
70	3,26
80	3,31
90	3,35
100	3,38
150	3,52
200	3,61
300	3,72
400	3,80
500	3,86

In un campione di N dati, **nel quale non siano presenti outliers**, il valore massimo corrisponde, approssimativamente al percentile:

$$\frac{N}{N+1} \cdot 100$$

Esempio: supponiamo di avere rilevato le lunghezze di 60 spore di un certo basidiocarpo, senza outliers; il valore più alto dovrebbe essere:

$$\frac{60}{60+1} \cdot 100 = 98,36$$

non più distante dalla media di quanto lo sia il percentile 98,36.

C.P. Quesenberry e H.A. David, nel loro articolo “Some tests for outliers” Biometrika vol. 48 pagg. 379-399 ed. 1961 Propongono la seguente tabella di Valori Critici per l’ESD in outlier statistici

ricavati dalla relazione:

$$ESD_{n,p} = \frac{t_{N-2p} \cdot (N-1)}{\sqrt{N \cdot (N-2 + t_{N-2,p}^2)}}$$

Che utilizza il t di Student, dove P è la probabilità con

$$p = 1 - \frac{\alpha}{2N}$$

Ora dobbiamo individuare un numero credibile di valori anomali nella nostra seriazione. In genere tale numero non deve superare il rapporto

$$K = N/10$$

E mai per un numero superiore a 5 (a meno che il campione non sia molto numeroso: $N > 100$). Questo perchè la seriazione, con valori anomali maggiori, si allontanerebbe troppo dalla distribuzione normale e il t di Student perderebbe di significato.

Il test serve a verificare

Ipotesi Nulla H_0 :non è presente nessun outlier

Ipotesi Alternativa H_1 :sono presenti da 1 a K outlier

Ciò premesso, del campione indagato si calcola la media e la deviazione standard.

Poi si individuano i valori anomali (in numero non superiore a 5) e si calcola del valore più distante dalla media l’Extreme Studentized Deviate con

$$ESD^{(N)} = \frac{|\mathbf{X}^{(N)} - \bar{\mathbf{X}}^{(N)}|}{\mathbf{S}^{(N)}}$$

Dal campione complessivo di N dati si leva il primo valore sospettato di essere outlier; pertanto il campione diventa di dimensioni N-1

$$ESD^{(N-1)} = \frac{|\mathbf{X}^{(N-1)} - \bar{\mathbf{X}}^{(N-1)}|}{\mathbf{S}^{(N-1)}}$$

Si calcola nuovamente di questo campione la media e la deviazione standard e la sua ESD. Si procede allo stesso modo per tutti e 5 i valori ritenuti anomali.

Poi si confronta l'ESD calcolato con l'ESD critico (confronto con la tabella riportata dei valori critici) con formulazione di rischio $\alpha = 0,05$ oppure $\alpha = 0,01$

Verranno dichiarati significativi, e perciò rifiutata l'Ipotesi Nulla, quei valori che risulteranno superiori o uguali al valore critico tabulato.

Un esempio pratico servirà ad illustrare meglio il metodo.

Da un carpoforo di *Inocybe geophylla* var. *geophylla*, rinvenuto in prateria alpina a 2400 mt., è stata realizzata una sporata e , da questa, rilevato un campione N=60 osservazioni, i cui valori modali delle lunghezze rientrano negli intervalli riportati in letteratura ($8 \div 10^5 \mu m$); mentre i valori medi risultano alterati, per via della presenza di spore aberranti.

Pure rispettando il criterio della casualità nella misurazione delle spore, sono stati tuttavia inseriti 5 valori ritenuti "anomali".

Si vuole sapere se tali unità statistiche si possono considerare appartenenti (compatibili) con la popolazione da cui sono stati estratti quindi casuali, oppure se essi siano da considerare "outlier" e quindi in grado di fare rigettare l'Ipotesi Nulla.

Il test dell'Ipotesi sarà il seguente:

Ipotesi Nulla H_0 :nessun valore può essere considerato outlier

Ipotesi Alternativa H_1 :almeno un valore tra 1 e K può essere considerato outlier

Nella seguente tabella sono riportati i valori anomali,i valori medi, le deviazioni standard, l'ESD e la significatività

Tabella 4: Ricerca degli Outlier

N	X	\bar{X}	S	ESD	P
60	16,41	10,19	1,66	3,76	*
59	15,23	10,08	1,45	3,56	*
58	14,33	9,99	1,29	3,55	*
57	14,01	9,91	1,15	3,13	NS
56	13,08	9,77	1,02	3,18	NS

Come si può notare i primi tre valori sono significativi con formulazione di rischio $\alpha = 0,05$, essendo il valore critico tabulato di $ESD > 3,20$ (N=60 \rightarrow 3,20). Mentre invece gli altri due valori non sono significativi.

Stante i dati e il test dell'Ipotesi, si rifiuta l'Ipotesi Nulla e si ammette che nella seriazione indagata vi siano spore aberranti (outlier) con rischio di commettere un errore di I tipo $\leq 5\%$

3.3 Conclusioni

Con queste brevi considerazioni ci siamo sforzati di evidenziare che le spore "aberranti" presenti in alcuni carpofori di certe specie, in particolari condizioni, debbano essere prese in considerazione.

Questo per una ragione molto ovvia: se la natura ce le ha messe, a noi non è dato di toglierle, solo perchè le troviamo "scomode" e perchè potrebbero "inquinare" i nostri risultati.

Più che escluderle, è necessario metterle in condizione di non inficiare i parametri distributivi.

Abbiamo anche sottolineato come sia estremamente difficile “ a colpo d’occhio” decidere che cosa è aberrante e cosa non lo è. Infatti, non si può decidere, senza opportuni metodi, quando un’unità statistica è “anomala” oppure obbedisce alla Distribuzione Normale.

Il metodo dell’*Extreme Studentized Deviate* è utile allo scopo e non è l’unico che ci offre la statistica.

Possiamo infatti ricorrere alla “Median Absolute Deviation”, oppure alla tecnica del “Trimming data”

Tuttavia quella illustrata ci pare la migliore, perchè utilizza la statistica parametrica.

Non spetta allo statistico dire perchè ci siano le spore “aberranti” e se esse dipendano da particolari condizioni di stress termico o altro; al più può collaborare ad uno studio puntuale per accertarne le ragioni; tuttavia può fornire metodi di indagine che sarebbe aspicabile venissero applicati.

3.4 Bibliografia

V. Barnett e T. Lewis “**outliers in statistical data**” ed. Chichester Jhon Wiley & Sons 1994

J. Stevens “**Applied multivariate statistics for the social sciences**” ed. Mahwah, NJ lawrence Erlbaum Associates

C.P. Quesenberry H.A. David “**Some tests for outliers**” *Biometrika* vol. 48 1961

L. Soliani “**Manuale di Statistica per la Ricerca e la Professione**” ed. ott.2003 copleft

R. Heim “**Le Genre Inocybe**” *Paul Lechevalier & Fils editeurs -Paris- 1931*

Th. W. Kuyper “ **A Revision of the Genus Inocybe in Europe**” *Rijsherbarium, Leiden Netherlands - 1986*