È possibile trovare la popolazione di origine conoscendone un campione?

o meglio ….

È possibile conoscere σ e μ partendo dalla conoscenza di n, $\overline{x}$ e d.s.?

A partire da un campione estratto dalla popolazione è impossibile determinare esattamente la popolazione di origine, cioè conoscere μ e σ

A partire da un campione estratto dalla popolazione è però possibile stimare i valori più probabili della popolazione di origine, cioè stimare μ e σ

l'inferenza

Nella maggior parte dei casi noi non conosciamo in anticipo né la media né la deviazione standard della popolazione!
È proprio per avere questa informazione che scegliamo un campione e lo "misuriamo"!

cioè

Nella maggior parte dei casi il nostro **non è un caso di probabilità** ma **è un caso di inferenza** (si congetturano cioè le caratteristiche di una popolazione di origine sconosciuta a partire dalla descrizione* di un campione estratto casualmente dalla stessa).

* ricorda: i parametri statistici che descrivono una popolazione di dati sono: n, $\overline{x}$ e d.s.

La distribuzione del "t" di Student

Il matematico inglese Student (pseudonimo di W.S. Gosset) ha determinato il valore delle seguente quantità:

$$t = \frac{(\bar{x} - \mu)}{sm}$$

t di Student = $\frac{\text{media campione - media popolazione}}{\text{errore standard del campione}}$

ricorda sm = **d.s./√n**

$\bar{x}$ = Media del campione che oscilla intorno alla media vera secondo la distribuzione normale, quindi:
$(\bar{x} - \mu)$ oscilla, secondo la distribuzione normale, intorno allo Zero;

sm = Stima del vero errore standard della popolazione che, a sua volta oscillerà intorno al valore vero;

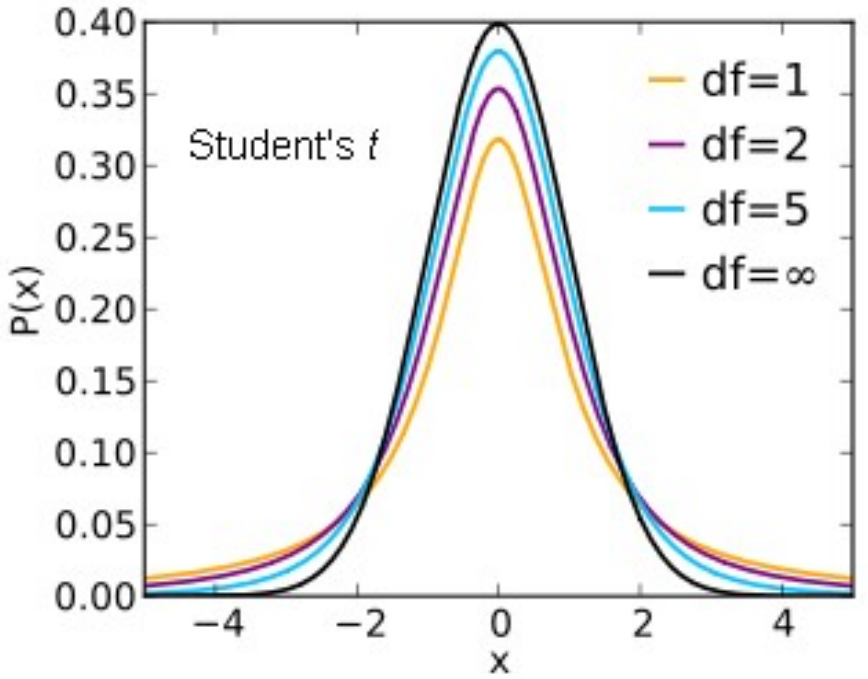
t = Varia (oscilla) più della distribuzione normale perché anche il denominatore varia (oscilla attorno a $\boldsymbol{\sigma}$)!

La distribuzione di **t** è **più complessa** della distribuzione **normale** in quanto si modifica anche in funzione della stima dell'errore standard (denominatore) che varia in funzione del numero dei g.l. utilizzati per il calcolo del *sm* (errore standard del campione)!

Dipende cioè dalla dimensione del campione

Non esiste pertanto una sola legge, rappresentata matematicamente, della distribuzione di t ma una famiglia di distribuzioni di t;
esiste cioè una distribuzione di t per ogni grado di libertà!

Student ha calcolato l'esatta distribuzione di t (risolvendo gli integrali definiti delle diverse equazioni) **ed ha redatto una tabella riassuntiva**



Ovviamente "aumentando i gradi di libertà" la distribuzione di t diviene "più stretta" (la stima di σ diviene più precisa).

Student's t-distribution has the probability density function

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

where ν is the number of *degrees of freedom* and Γ is the Gamma function.

Abbiamo quindi una tabella simile a quella di Z per ogni grado di libertà. Ciò è improponibile, abbiamo bisogno di semplificare.
Per ridurre il numero di tabelle ci si limita ad individuare per ciascuna distribuzione solo alcuni valori (generalmente 0,5 0,4 0,3 0,2 0,1 **0,05** 0,02 **0,01** 0,002 0,001) ma in pratica si esaminano solo due valori corrispondenti alle aree del **0,05 = 95%** e **0,01 = 99%**.

probabilità % di un valore più elevato di t trascurando il segno.										
due code	**0,5**	**0,4**	**0,3**	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
una coda	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,001	0,0005
g.l.										
1	1,000	1,376	1,963	3,078	6,314	12,710	31,820	63,660	318,310	636,620
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	3,195	3,416
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	3,174	3,390
1.000	0,675	0,842	1,037	1,282	1,646	1,962	2,330	2,581	3,098	3,300
infinito	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Tavola realizzata con la funzione invt del foglio di calcolo

Il significato della tabella è che se facciamo una estrazione a caso di un campione di n individui e calcoliamo il valore di ***t***, poi scegliamo una probabilità, nel punto di incrocio fra la colonna della probabilità scelta e la riga corrispondente ai gradi di libertà (g.l.= n-1) troviamo un valore di ***t*** che verrà superato dal valore che abbiamo calcolato solo un numero di volte inferiore a quello della probabilità scelta.

probabilità % di un valore più elevato di t trascurando il segno.										
due code	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,002	0,001
una coda	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,001	0,0005
g.l.										
1	1,000	1,376	1,963	3,078	6,314	12,710	31,820	63,660	318,310	636,620
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	5,893	6,869

Es. se facciamo una estrazione a caso di un campione di 6 individui e calcoliamo il valore di ***t***, c'è una probabilità di 0,05 cioè del 5% che il valore trovato superi il valore 2,571 (quinta colonna della tavola per g.l. = n-1 = 5).

La tabella Z riportava solo una metà questa riporta anche la somma dei due valori (due code)

Prendendo quindi: $t = \frac{(\bar{x} - \mu)}{sm}$ E considerando $t_{0,05}$

$$-t_{0,05} < \frac{(\bar{x} - \mu)}{sm} \quad \text{e} \quad \frac{(\bar{x} - \mu)}{sm} < +t_{0,05}$$

La probabilità che il valore calcolato per la differenza fra la media del campione e la media vera (della popolazione) fratto l'errore standard del campione sia compreso fra +***t*** e -***t*** è pari al valore scelto nella tavola in funzione dei gradi di libertà del nostro campione. Cioè:

$$-t_{0,05} * sm < (\bar{x} - \mu) \quad \text{e} \quad (\bar{x} - \mu) =< t_{0,05} * sm$$

o meglio..

$$\bar{x} - t_{0,05} * sm < (\mu) \qquad (\mu) < t_{0,05} * sm + \bar{x}$$

La media della popolazione è compresa fra la media del campione più o meno l'errore standard del campione.

Ovvero, se $\bar{x}$ e *sm* sono la media e l'errore standard di un campione estratto da una popolazione normale. Vi è una probabilità di 0,95 (o 95%) in favore dell'ipotesi che la **vera media** (quella della popolazione) sia compresa fra i valori:

$$\bar{x} - t_{0,05} * sm \quad \text{e} \quad t_{0,05} * sm + \bar{x}$$

Cioè abbiamo definito … … … … … … … ..

l'intervallo fiduciale di una media è

$$\bar{x} - t_{0,05} * sm < (\mu) \qquad (\mu) < t_{0,05} * sm + \bar{x}$$

attenzione:

~~La vera media (della popolazione) è situata entro questi limiti con una probabilità di …..~~

C'è una probabilità di x% di sbagliare affermando che la media sia situata entro questi limiti

esempio:

Calcola i limiti fiduciali per p=0,05

PESO ALLA	NASCITA DEI	BOVINI
matricola	PESO	SESSO
1	40	F
2	40	M
3	47	F
4	50	M
5	40	F
6	50	F
7	38	F
8	38	F
9	47	M
10	42	F

1 Calcolo i parametri del campione

Vedi lezioni precedenti

n	**10**
media	**43,2**
d.s.	**4,80**
e.s.=*sm.*	1,5178933

2 Individuo il valore di *t* per 10-1=9 gl.

probabilità % di un valore più elevato di t trascurando il segno.										
due code	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,002	0,001
g.l.										
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437

3 Calcolo t*sm: 2,262*1,5179 = 3,43

4 Sommo e sottraggo dalla media la quantità t*sm

43,2+3,43 e 43,2-3,43

limiti fiduciali allo 0,05

46,6 - 39,8

Tabella riassuntiva completa

serie PESO ALLA NASCITA DEI BOVINI			
		40	
		40	
		47	
		50	
		40	
		50	
		38	
		38	
		47	
		42	
n		**10**	
media		**43,2**	
d.s.		**4,80**	
e.s.		1,5178933	
t=		2,262	
t*sm =		3,4334746	
limiti fiduciali		**46,6**	
		39,8	

limiti fiduciali allo 0,05

Riepilogo: dati di misura continui

Una serie di misure deve essere sempre descritta da 3 parametri la **-media** la **-deviazione standard** e ed il **-numero** delle osservazioni: **n,** $\bar{x}$ e **d.s.**

Una serie di misure deve essere considerata come un campione estratto da una popolazione infinita che può essere descritta da 2 parametri la **-media** e la **-deviazione standard:** μ e σ

Dai parametri della **popolazione** posso individuare le dimensioni di un **campione** (**percentuale della popolazione o n se la popolazione di origine è finita**) da estrarre dalla popolazione tramite l'impiego della distribuzione normale standardizzata:

$$z = \frac{(X - \mu)}{\sigma}$$

Dai parametri della **popolazione** posso individuare probabilisticamente quale sarà l**a** media del **campione** perché la media del campione oscillerà intorno alla media della popolazione secondo una distribuzione normale la cui media è la media della popolazione e la cui deviazione standard è l'errore standard cioè:

$$\mu = \bar{x} \text{ e } \text{d.s.} = \sigma/\sqrt{n}$$

Dai parametri di un **campione** posso individuare probabilisticamente l**a** media della **popolazione** tramite la distribuzione di t perché: $sm = \text{d.s.}/\sqrt{n} = \text{e.s.}$

$$\bar{x} - t_{0,05} * \frac{d.s.}{\sqrt{n}} < (\mu) < \bar{x} + t_{0,05} * \frac{d.s.}{\sqrt{n}}$$

06-3

Dato come test il 2013-DIC-16

Calcola i limiti fiduciali al 95% delle medie relative alle 3 misure riportate in tabella:

		Altezza camera d'aria uovo	Peso uovo	Haugh Units
		mm	g	
uovo	n =12			
media		2,4	67	89,8
dev.st.		0,45	1,4	3,4

Misure riportate in tabella 0 misure di qualità delle uova:

- altezza camera d'aria uovo = dimensione della “bollicina d'aria” presente al polo ottuso delle uova = indice della freschezza;
- Peso uovo= peso in grammi dell'uovo = categoria di peso S, M, L, XL;
- Unità Haugh = misura della consistenza dell'albume = indice dell'altezza dell'albume diviso la supeficie che occupa l'albume dopo la rottura dell'uovo su una superficie piana = indice di freschezza e di carateristiche delle proteine dell'albume.

06-3

Sapendo che:

	camera d'aria mm	peso g	Haugh units
uovo n =12			
media	2,4	67	89,8
dev.st.	0,45	1,4	3,4
e.s.	0,130	0,404	0,981

$$\bar{x} - t_{0,05} * \frac{d.s.}{\sqrt{n}} < (\mu) \; (\mu) < \bar{x} + t_{0,05} * \frac{d.s.}{\sqrt{n}}$$

N=12;
g.l.=11
P=0,05

probabilità % di un valore più elevato di t trascurando il segno.										
due code	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,002	0,001
g.l.										
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,852	4,221

Radq di 12 = 3,464
e.s. = 0,45 diviso 3,464 = 0,130
0,130*2,201 = 0,286

2,400 -	2,400 +
0,286 =	0,286 =
2,114	2,686

e.s. = 1,4 diviso 3,464 = 0,404
0,404*2,201 = 0,890

67,00 -	67,00 +
0,890 =	0,890 =
66,110	67,890

e.s. = 3,4 diviso 3,464 = 0,981
0,981*2,201 = 2,160

89,800 -	89,800 +
2,160 =	2,160 =
87,640	91,960

n	12	12	12
media	2,4	67	89,8
d.s.	0,45	1,4	3,4
radq di 12	3,464	3,464	3,464
err.st.	0,130	0,404	0,981
$t_{0,05}$ =	2,201	2,201	2,201
intervallo=	0,286	0,890	2,160
min	2,114	66,110	87,640
MAX	2,686	67,890	91,960

Risposta

	media	limiti fiduciali a 0,05	
Altezza camera d'aria, mm =	**2,4**	**2,11**	**2,69**
Unità Haugh =	**89,8**	**87,64**	**91,96**
peso medio, g =	**67**	**66,11**	**67,89**