Alcuni parametri statistici di base

Misure di tendenza centrale:

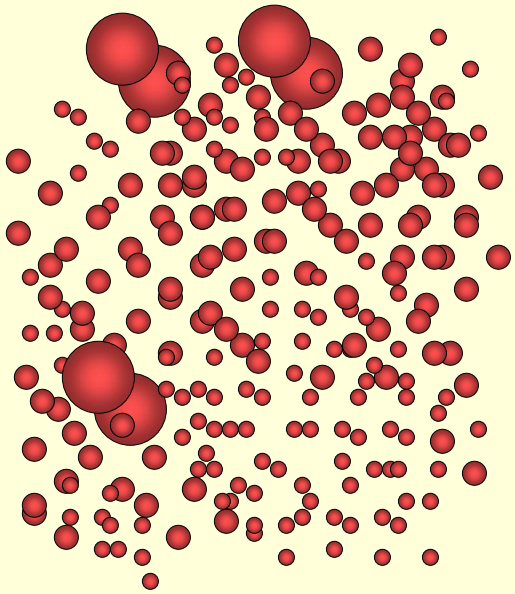
- **media**
- **mediana**
- **moda**

Misure di dispersione:

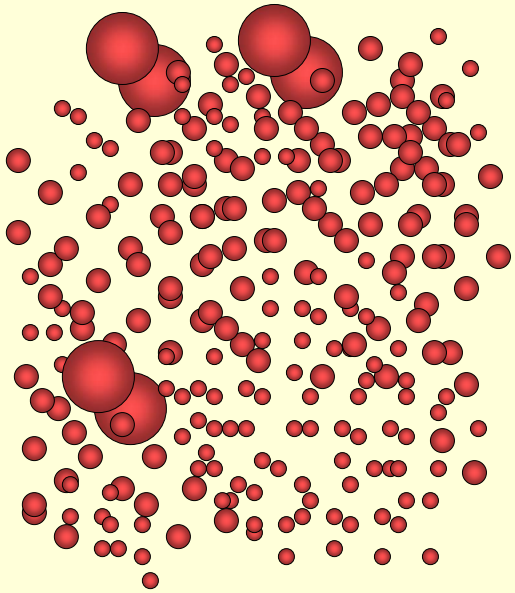
- **intervallo di variazione**
- **scarto medio**
- **varianza**
- **deviazione standard**
- **coefficiente di variazione**

Tendenza centrale - Media

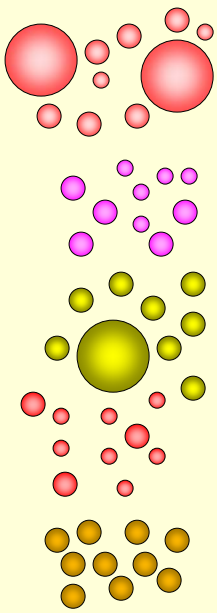
Popolazione di *Protopalla rotunda*



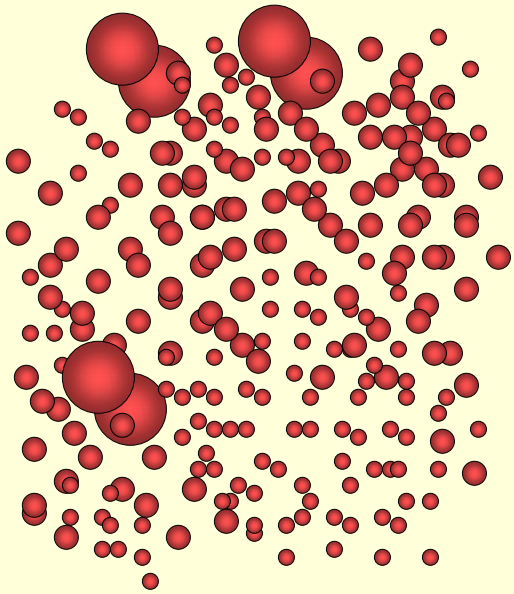
Popolazione di *Protopalla rotunda*



Campioni

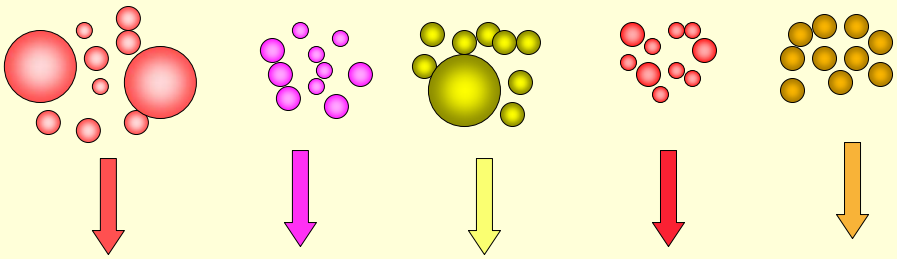


Popolazione di *Protopalla rotunda*



$$\mu = \frac{\Sigma X_i}{N}$$

Campioni



$$\bar{X} = \frac{\Sigma X_i}{n} \qquad \bar{X} = \frac{\Sigma X_i}{n} \qquad \bar{X} = \frac{\Sigma X_i}{n} \qquad \bar{X} = \frac{\Sigma X_i}{n} \qquad \bar{X} = \frac{\Sigma X_i}{n}$$

Se il campionamento è corretto…

$$\bar{X} = \frac{\Sigma X_i}{n}$$

$$\bar{X} = \frac{\Sigma X_i}{n}$$

$$\bar{X} = \frac{\Sigma X_i}{n}$$

$$\bar{X} = \frac{\Sigma X_i}{n}$$

$$\bar{X} = \frac{\Sigma X_i}{n}$$

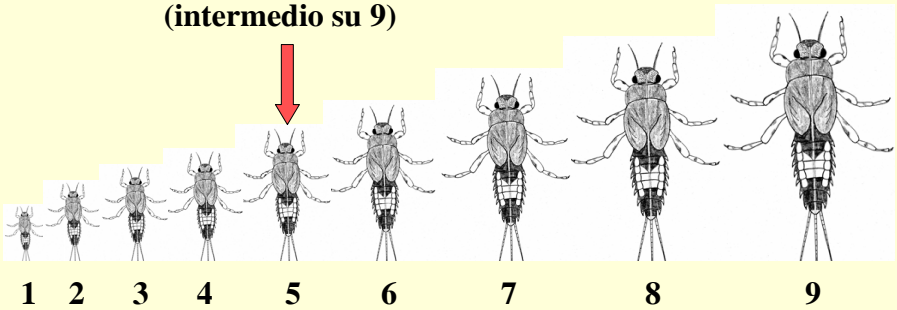
ogni stima approssima μ

Tendenza centrale - Mediana

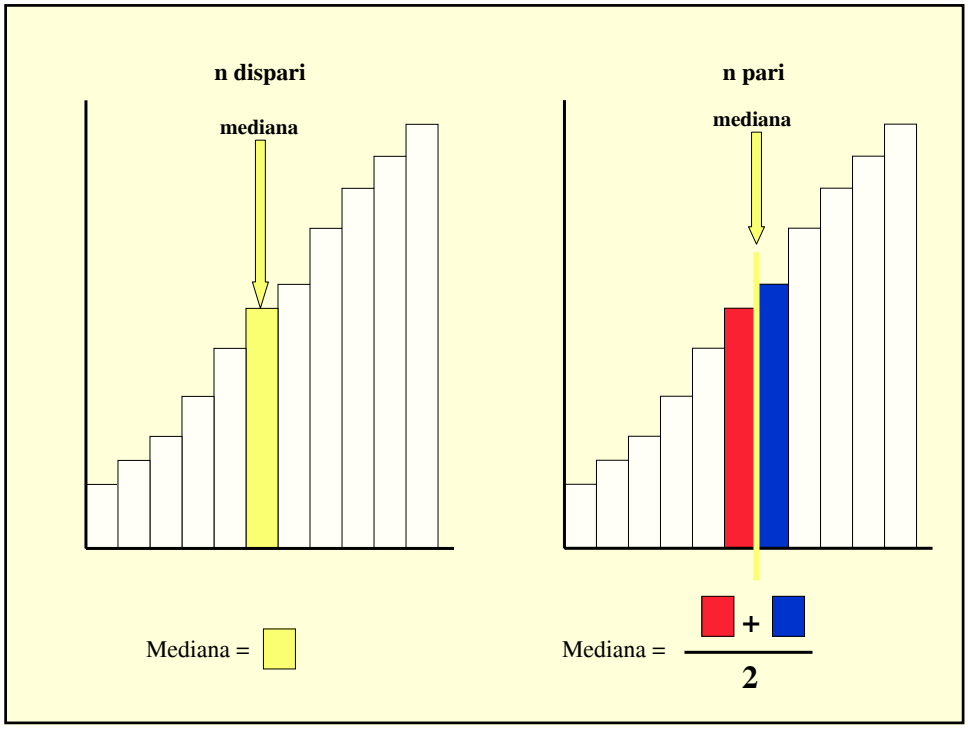
Mediana: valore intermedio
[(n-1)/2 valori maggiori, (n-1)/2 valori inferiori]

e.g. lunghezze di ninfe di Efemerotteri

5º valore
(intermedio su 9)

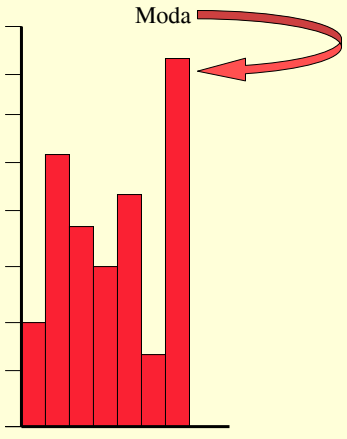


1 2 3 4 5 6 7 8 9



Tendenza centrale - Moda

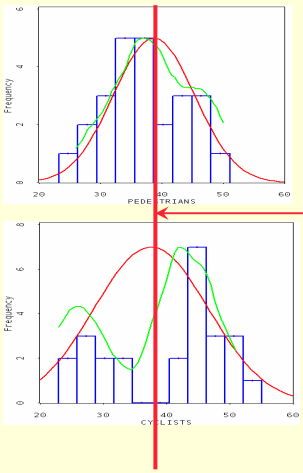
Moda: il valore più frequente



Misure di dispersione

Perchè sono importanti?

Perchè non tutte le popolazioni hanno le stesse caratteristiche

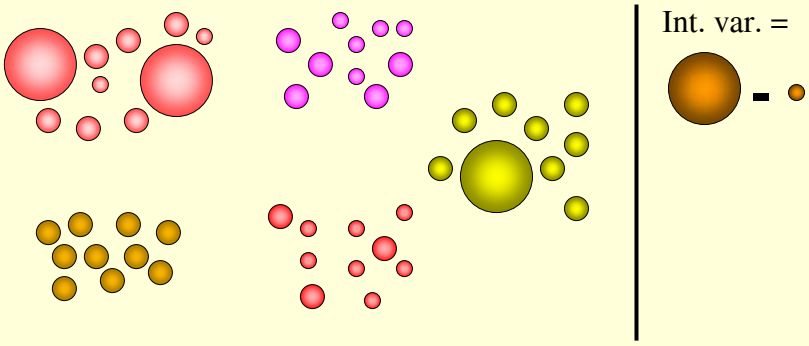


Distribuzioni diverse, ma
medie e mediane coincidenti!

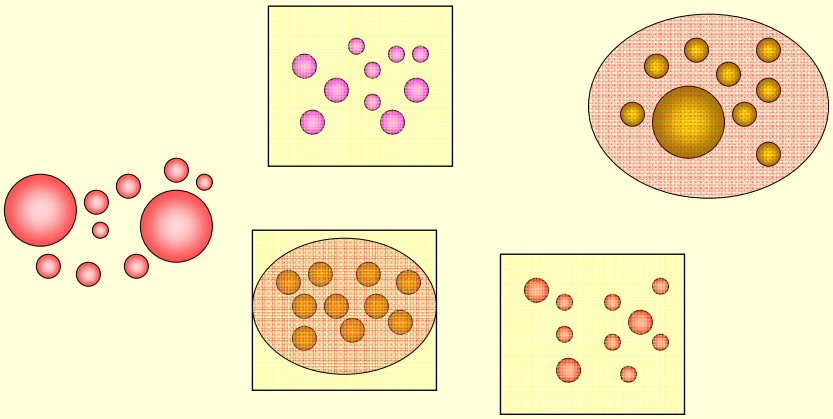
Media e mediana

Misure di dispersione - Intervallo di variazione

1. Intervallo di variazione: differenza fra min e max



Intervallo di varizione → semplice, ma poco informativo



NB – Tre campioni non includono un valore max

E due non includono un valore min

Misure di dispersione - Scarto medio

Si prende la differenza fra ogni valore e la media:

$$X_i - \overline{X}$$

$$\Sigma X_i - \overline{X} = 0$$

La somma di questi scarti è nulla, e quindi non serve…

Misure di dispersione - Scarto medio (segue)

Se però si prende in valore assoluto, allora è una misura di dispersione:

$$\Sigma \left| X_i - \overline{X} \right|$$

e

$$\frac{\Sigma \left| X_i - \overline{X} \right|}{n} = \text{scarto medio}$$

Misure di dispersione - Varianza

Per eliminare il segno dello scarto…

si prende il suo quadrato:

$$(X_i - \overline{X})^2$$

E, se si sommano le differenze quadratiche, si ha una "somma di quadrati":

$$\Sigma(X_i - \overline{X})^2$$

Misure di dispersione - Varianza (segue)

Una somma di quadrati può essere considerata a livello di popolazione o di campione:

Popolazione	Campione
$SS = \Sigma(X_i - \mu)^2$	$ss = \Sigma(X_i - \overline{X})^2$

Misure di dispersione - Varianza (segue)

Se si divide per la dimensione della popolazione o i *gradi di libertà* del campione, si ha lo scarto quadratico medio o *varianza*

Popolazione	Campione
$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$	$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$
Varianza della popolazione	Varianza del campione

Misure di dispersione - Deviazione standard

La radice quadrata della varianza

Popolazione

$$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$$

Campione

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$$

La deviazione standard è una misura molto utile:

Es. la maggioranza dei dati in una qualsiasi popolazione ha un valore che non si discosta dalla media di più di una deviazione standard

Misure di dispersione - Coefficiente di variazione

Lunghezza media: 2.4 m Varianza: 1.6 m

Dev. Std.: 1.26 m

Lunghezza media: 2.4 cm Varianza: 1.6 cm

Dev. Std.: 1.26 cm

Le orecchie degli elefanti sono 100 volte più variabili di quelle dei topi?

N.B. Varianza e dev. Std. hanno spesso ordini di grandezza dipendenti dalla scala dei dati

Misure di dispersione - Coefficiente di variazione (segue)

$$V = (s/\overline{X}) * 100\%$$

Elefanti: 1.26 m / 2.4 m *100% = 52.5%

Topi: 1.26 cm / 2.4 cm *100% = 52.5%

Campioni, confronti, ipotesi

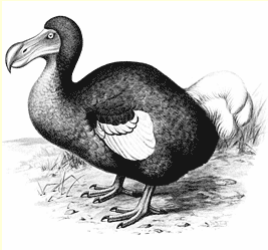
Due modi diversi di ragionare…

1. Inferenza deduttiva

2. Inferenza induttiva

Quanti campioni sono possibili?

Immaginiamo di essere tornati un po' indietro nel tempo e di aver potuto studiare la popolazione di Dodo prima della sua estinzione. Il nostro obiettivo era sapere quante uova deponeva in media ciascuna femmina (ne rimanevano solo 6!).



Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

$\mu = 4$

$\sigma^2 = 9.33$

$\sigma = 3.06$

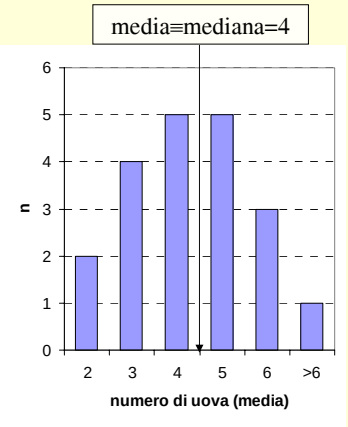
Quanti diversi campioni erano possibili per n=3?

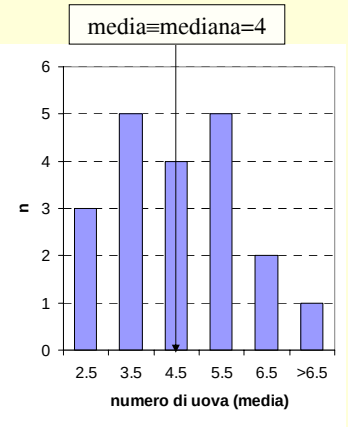
$$\frac{6!}{3!*3!} = 20$$

Medie stimate dai 20 campioni possibili

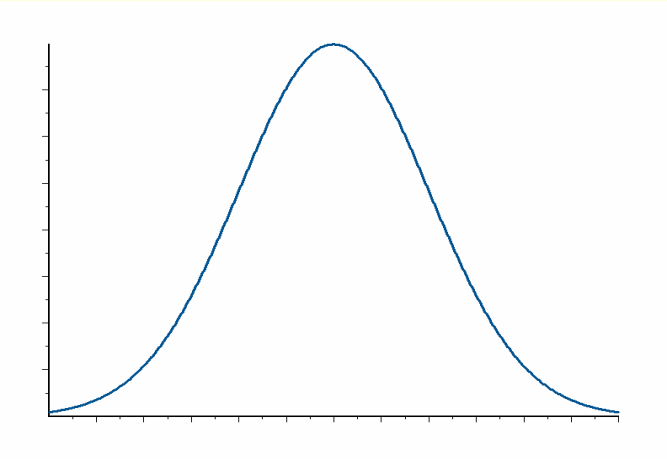
Dodo #1	Dodo #2	Dodo #3	Media del campione (m)
0	1	3	1.33
0	1	5	2.00
0	1	6	2.33
0	1	9	3.33
0	3	5	2.67
0	3	6	3.00
0	3	9	4.00
0	5	6	3.67
0	5	9	4.67
0	6	9	5.00
1	3	5	3.00
1	3	6	3.33
1	3	9	4.33
1	5	6	4.00
1	5	9	5.00
1	6	9	5.33
3	5	6	4.67
3	5	9	5.67
3	6	9	6.00
5	6	9	6.67

Distribuzione delle medie

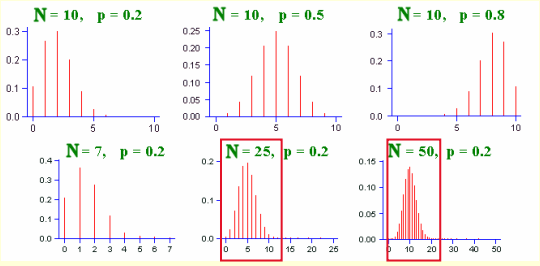




La distribuzione normale



Per grandi numeri, altre distribuzioni tendono a quella normale (teorema del limite centrale)

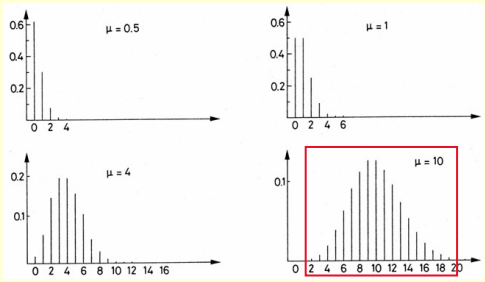


Distribuzione binomiale

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$P_x = e^{-\mu} \left[\frac{\mu^x}{x!} \right]$$

Distribuzione di Poisson

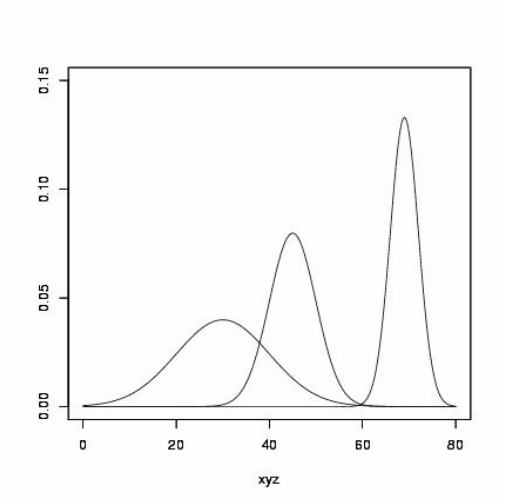


Una variabile casuale influenzata da numerosi fattori tende ad avere una distribuzione normale

Dati biometrici, tassi di vario tipo, misure fisiche in generale, etc.

Se i valori misurati sono influenzati da un numero elevato di eventi casuali, allora la distribuzione tenderà ad essere normale.

Le curve normali hanno forme variabili…



Quindi, per comparare più distribuzioni normali, dobbiamo standardizzarle in qualche modo…

Standardizzazione: la variabile Z

$$Z = \frac{\text{valore osservato var. casuale} - \text{media}}{\text{deviazione standard}}$$

ovvero

$$Z = \frac{x - \mu}{\sigma}$$

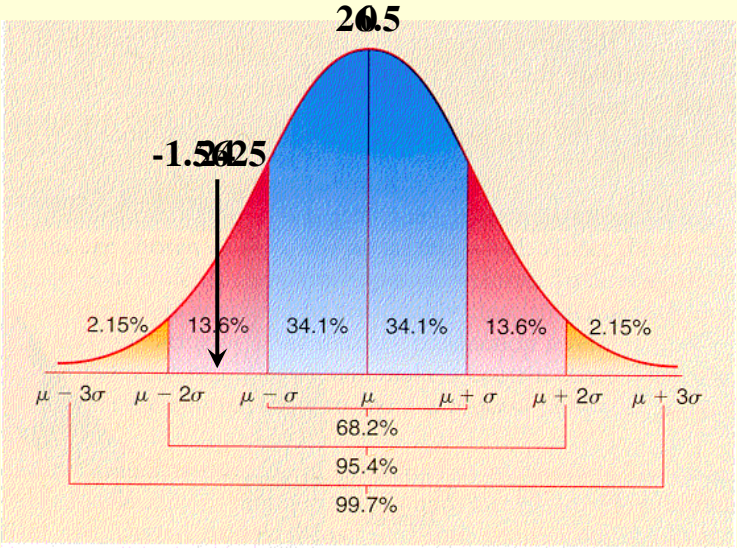
Esempio

Il voto medio di Metodologie Ecologiche è 26.5, mentre la deviazione standard è 1.6. Se hai avuto 24, qual'è stato il valore della variabile Z nel tuo caso?

$$Z = \frac{x - \mu}{\sigma}$$

$$Z = \frac{x - 26.5}{1.6} = -1.5625$$

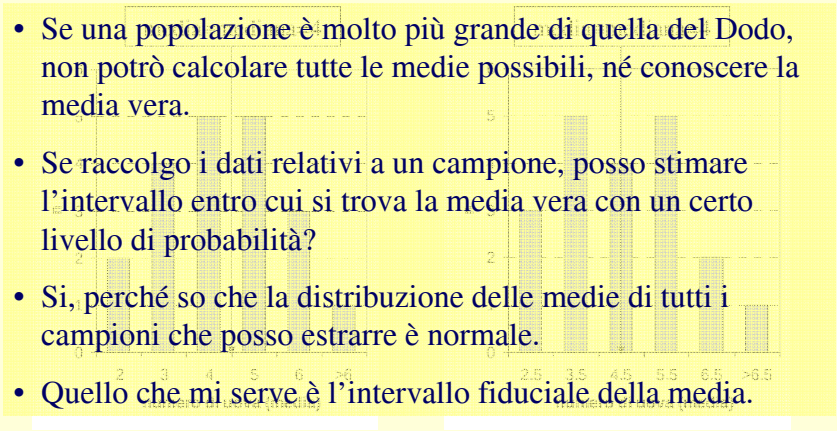
In pratica, Z ci dice di quante deviazioni standard un valore si scosta dalla media…



Ogni deviazione standard di scarto dalla media definisce un'area sotto la curva, che equivale a una certa percentuale di casi

Distribuzione delle medie

- Se una popolazione è molto più grande di quella del Dodo, non potrò calcolare tutte le medie possibili, né conoscere la media vera.
- Se raccolgo i dati relativi a un campione, posso stimare l'intervallo entro cui si trova la media vera con un certo livello di probabilità?
- Si, perché so che la distribuzione delle medie di tutti i campioni che posso estrarre è normale.
- Quello che mi serve è l'intervallo fiduciale della media.



Intervallo fiduciale della media

- Calcolo la media
- Calcolo la deviazione standard
- Calcolo l'errore standard della media: $s_e = \frac{\sigma}{\sqrt{n}}$
- La media μ della popolazione sarà compresa nell'intervallo fra la media campionaria *m* meno $t_{(n-1,p)} \cdot s_e$ e la media campionaria *m* più $t_{(n-1,p)} \cdot s_e$ dove $t_{(n-1,p)}$ è il valore del *t* di Student con *n-1* gradi di libertà per il livello di probabilità *p* desiderato

Intervallo fiduciale della media (in altre parole…)

$$m - t_{(n-1,p)} \cdot s_e \longleftarrow m \longrightarrow m + t_{(n-1,p)} \cdot s_e$$

μ

(con una probabilità *p*)

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

m=(9+3+1)/3=4.333

x	x-m	$(x-m)^2$
9	4.667	21.778
3	-1.333	1.778
1	-3.333	11.111

$s^2=(x-m)^2/(n-1) = 17.333$

$s=\sqrt{(x-m)^2/(n-1)} = 4.163$

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

m=4.333 s=4.163

$s_e = s/\sqrt{n} = 4.163/\sqrt{3} = 2.404$

$t_{(n-1,p)} = t_{(3-1,0.95)} = 4.303$

	Quantile (area to the left of t)									
df	0.600	0.700	0.800	0.900	0.950	0.975	0.980	0.990	0.995	0.9995
1	0.325	0.727	1.376	3.078	6.314	12.706	15.895	31.821	63.657	636.619
2	0.289	0.617	1.061	1.886	2.920	4.303	4.849	6.965	9.925	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	3.482	4.541	5.841	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	2.999	3.747	4.604	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	2.757	3.365	4.032	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	2.612	3.143	3.707	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.517	2.998	3.499	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.449	2.896	3.355	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.398	2.821	3.250	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.359	2.764	3.169	4.587

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

m=4.333 s=4.163

$s_e = s/\sqrt{n} = 4.163/\sqrt{3} = 2.404$

$t_{(n-1,p)} = t_{(3-1,0.95)} = 4.303$

$m - t_{(n-1,p)} \cdot s_e < \mu < m + t_{(n-1,p)} \cdot s_e$

$4.333 - 4.303 \cdot 2.404 < \mu < 4.333 + 4.303 \cdot 2.404$

$-6.011 < \mu < 14.677$ per p=0.95 (95%)

Test d'ipotesi

"There is one great difficulty with a good hypothesis. When it is completed and rounded, the corners smooth and the content cohesive and coherent, it is likely to become a thing in itself, a work of art... One hates to disturb it. Even if subsequent information should shoot a hole in it, one hates to tear it down because it once was beautiful and whole... ."

John Steinbeck/Ed Ricketts, 1941
Log from the Sea of Cortez

Confronti fra medie: il test *t* di Student

Il test t di Student

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_e} \qquad H_0 : \mu_1 = \mu_2$$

$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{gdl} = n_1 + n_2 - 2$$

Analisi della varianza

Alimentazione di pesci

Pesi degli animali al termine di una prova

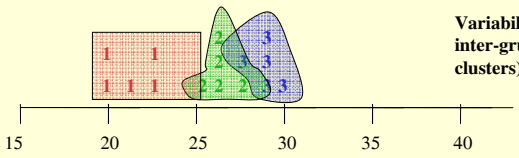
Prova 1			Prova 2		
Formula 1	Formula 2	Formula 3	Formula 1	Formula 2	Formula 3
20	25	28	18	27	17
22	27	28	24	20	37
21	26	27	17	29	29
22	26	29	22	31	21
20	26	28	24	23	36
21	**26**	**28**	**21**	**26**	**28**

Quale è la differenza di maggiore importanza nei dati?

La differenza sta nel modo in cui sono distribuiti

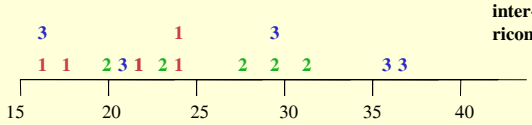
(I numeri colorati rappresentano le diverse formulazioni dell'alimento)

Prova 1



Variabilità intra-gruppo piccola, inter-gruppo grande (si formano clusters)

Prova 2



Variabilità intra-gruppo grande, inter-gruppo piccola (non si riconoscono clusters)

Per comparare le due prove
(c'è differenza fra di esse?),
formuliamo due ipotesi:

H_0: $\mu_1 = \mu_2 = \mu_3$ **Non c'è differenza fra le formulazioni**

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$ **C'è differenza fra le formulazioni**

Piano sperimentale ed ANOVA

Esperimento completamente randomizzato:

1) 15 pesci
2) assegnazione casuale alla dieta

Si usa un'ANOVA a una via
(o ad un fattore)

Ipotesi:

$\mathbf{H_0: \mu_1 = \mu_2 = \mu_3}$

$\mathbf{H_1: \mu_1 \neq \mu_2 \neq \mu_3}$

Perchè non analizzare invece:

$\mathbf{H_0: \mu_1 = \mu_2}$
$\mathbf{H_1: \mu_1 \neq \mu_2}$

$\mathbf{H_0: \mu_1 = \mu_3}$
$\mathbf{H_1: \mu_1 \neq \mu_3}$

$\mathbf{H_0: \mu_2 = \mu_3}$
$\mathbf{H_1: \mu_2 \neq \mu_3}$

Man mano che si aumenta il numero di confronti a coppie, aumenta la probailità di un errore di Tipo I (rigettare un'ipotesi vera)

Numero di confronti a coppie	Probabilità Errore Tipo I
1	.05
5	.23
10	.63
20	.92

La sola formula necessaria per un'ANOVA a una via:

$$\text{Varianza} = \frac{\text{Somma degli scarti quadratici dalla media}}{\text{Gradi di libertà}}$$

$$= \frac{\Sigma(Xi - X)^2}{n-1}$$

La varianza in un' ANOVA si chiama spesso "somma dei quadrati" o "SS"

- Ci sono più sorgenti di variazione (misurate dalla somma dei quadrati [SS]) nei dati.
- Lo scopo dell' ANOVA è di misurare queste variazioni e decidere da cosa dipendono.

La prima sorgente di variazione è la variabilità complessiva dei dati. Si misura con la Somma Totale dei Quadrati o SS_T

$$SS_T = \Sigma X^2 - \frac{\Sigma(X)^2}{N}$$

La variabilità complessiva dei dati può essere decomposta in due componenti:

Variabilità complessiva (SS_T)

Variabilità fra trattamenti
Deriva da:
1. Differenze fra soggetti
2. Errore sperimentale
3. Effetto dei trattamenti

Variabilità nei trattamenti
Deriva da:
1. Differenze fra soggetti
2. Errore sperimentale

Variabilità complessiva (SS_T)

Variabilità fra trattamenti
Deriva da:
1. Differenze fra soggetti
2. Errore sperimentale
3. Effetto dei trattamenti

Variabilità nei trattamenti
Deriva da:
1. Differenze fra soggetti
2. Errore sperimentale

Si comparano con una statistica F

$$F = \frac{\text{Variabilità fra trattamenti}}{\text{Variabilità nei trattamenti}}$$

$$= \frac{\text{Effetto dei trattamenti + differenze fra soggetti + errore sperimentale}}{\text{differenze fra soggetti + errore sperimentale}}$$

Considerando la prova 1…

Prova 1		
Formula 1	Formula 2	Formula 3
20	25	28
22	27	28
21	26	27
22	26	29
20	26	28
21	**26**	**28**

Media generale = 25

Come differiscono le diete?
Quale è la sorgente di variazione fra di esse?

Calcoli!

1) Variabilià complessiva
= Somma dei quadrati totale = SS_T

Prova 1		
Formula 1	Formula 2	Formula 3
20	25	28
22	27	28
21	26	27
22	26	29
20	26	28
21	26	28

Media generale = 25

$$SS_T = \Sigma X^2 - \frac{\Sigma(X)^2}{N}$$

$$= 9513 - \frac{140{,}625}{15}$$

$$= 138$$

Ma si può anche calcolare come:

$$SS_T = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 \ldots\ldots (X_n - \overline{X})^2$$

$$= (20 - 25)^2 + (22 - 25)^2 + (21 - 25)^2 + \ldots\ldots (28 - 25)^2$$

$$= 138$$

Calcoli!

2) Variabilità intra-gruppo
= $\mathbf{\Sigma SS_{intra}}$

Prova 1		
Formula 1	**Formula 2**	**Formula 3**
20	**25**	**28**
22	**27**	**28**
21	**26**	**27**
22	**26**	**29**
20	**26**	**28**
21	**26**	**28**

Media generale = 25

$SS_{intra\ F1} = \Sigma X^2 - \frac{\Sigma(X)^2}{N}$

$= 2209 - \frac{11025}{5}$

$= 4$

$SS_{intra\ F2} = 3382 - \frac{16900}{5}$

$= 2$

$SS_{intra\ F2} = 3922 - \frac{19600}{5}$

$= 2$

$\Rightarrow \Sigma SS = SS_{intra\ F1} + SS_{intra\ F2} + SS_{intra\ F3}$
$= 4 + 2 + 2 = 8$

Quindi $SS_{intra} = 8$

Calcoli!

3) Variabilità inter-gruppo,
si calcola in due modi

Prova 1		
Formula 1	**Formula 2**	**Formula 3**
20	**25**	**28**
22	**27**	**28**
21	**26**	**27**
22	**26**	**29**
20	**26**	**28**
21	**26**	**28**

Media generale = 25

$SS_{inter} = \frac{\Sigma X^2}{n} - \frac{\Sigma(X)^2}{N}$

oppure

Si sottrae SS_{intra} da SS_T

Poichè $SS_T = SS_{intra} + SS_{inter}$

Quindi $SS_{intra} = SS_T - Ss_{intra} = 138 - 8 = 130$

$SS_{inter} = 130$

Calcoli!

4) Si calcola il Quadrato Medio

Ricordare la formula della varianza (con una piccola modifica)

$s^2 = SS/gdl$

Nell'ANOVA si sostituisce s^2 col Quadrato Medio (MS):

$$MS_{inter} = \frac{SS_{inter}}{gdl_{inter}} = 130/2 = 65$$

$$MS_{intra} = \frac{SS_{intra}}{gdl_{intra}} = 8/12 = 0.66$$

E quindi F (la statistica dell'ANOVA) = $\frac{MS_{inter}}{MS_{intra}} = 65/.66 = 97.59$

Per una semplice ANOVA a una via

I gradi di libertà sono :

$Gdl_{intra} = N - K$

$Gdl_{inter} = K - 1$

$Gdl_T = N - 1$

Dove N = numero dei dati totali (15 pesci)
K = numero dei trattamenti (3 formulazioni)

I risultati dell'ANOVA si presentano in una tabella impostata come:

Sorgente di variazione	SS	gdl	MS (SS/gdl)
Inter-gruppo (trattamento)	130	2 (K-1)	66
Intra-gruppo (errore, residui)	8	12 (N-K)	0.65
Totale	138	14 (N-1)	

F = 97.59

Il passo finale!

Si cerca il valore critico per p = .05 e 2 e 12 gradi di livertà in una tavola di F

F(.05, 2, 12) = 5.10

Dato che il valore di F ottenuto (97.59) è molto maggiore di 5.10:

p <<< .05

Tornando alle ipotesi di partenza:

H_0: $\mu_1 = \mu_2 = \mu_3$ Non c'è differenza fra le formulazioni

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$ C'è differenza fra le formulazioni

Si rigetta H_0 e quindi si accetta H_1, cioè che esiste una differnza fra formulazioni

Sommario dell'ANOVA