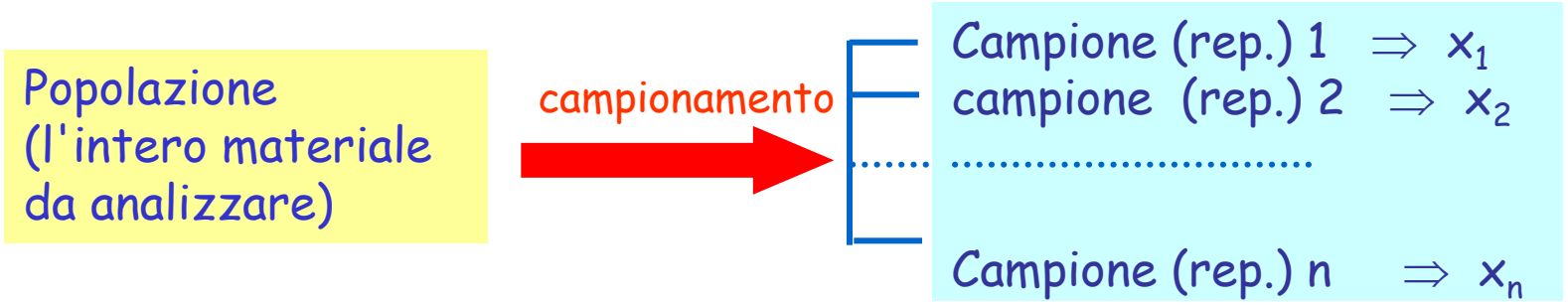
# Inferenza statistica ed intervallo di fiducia

Nel contesto chimico analitico la popolazione può rappresentare tutto il materiale da analizzare, ad esempio un terreno in cui occorre determinare la quantità di uno ione o di un pesticida, una cisterna contenente un prodotto industriale, ecc. In alternativa, può rappresentare l'ipotetico insieme di infinite misure replicate su uno stesso campione.

In entrambi i casi si può, di fatto, soltanto analizzare un insieme finito di campioni oppure ottenere un insieme finito di replicati:




I valori $x_1$, $x_2$,...$x_n$ rappresentano il set di dati analitici disponibili, dai quali ricavare i parametri campionari $\overline{X}$(media) e s (deviazione standard)

L'inferenza statistica è la procedura con cui i parametri di popolazione (ossia i valori desiderati) vengono stimati a partire da quelli campionari.

Poiché i parametri campionari hanno essi stessi una distribuzione, è possibile valutare per ciascuno di loro un intervallo di valori tale che la probabilità P che il parametro sia compreso in quell'intervallo sia numericamente definita (ad esempio il 90 o 95%).

Più precisamente, ciò significa che se si ricava quell'intervallo a partire da un certo numero di insiemi di dati, ciascuno costituito da n misure, il parametro di popolazione sarà contenuto nell'intervallo nel P% di casi.

- ✓ Tale intervallo viene definito intervallo di fiducia *(confidence interval)*
- ✓ i suoi estremi sono i limiti di fiducia
- ✓ il valore di P (di solito espresso come frazione dell'unità, ad esempio 0.95, o in percentuale, ad esempio 95%) è il coefficiente o livello di fiducia
- ✓ il complemento a 1 (o a 100) di P viene detto livello di significatività, indicato con $\alpha$

# Intervallo di fiducia per la media: calcolo nei vari casi

## Caso 1: popolazione distribuita normalmente e con varianza ($\sigma^2$) nota

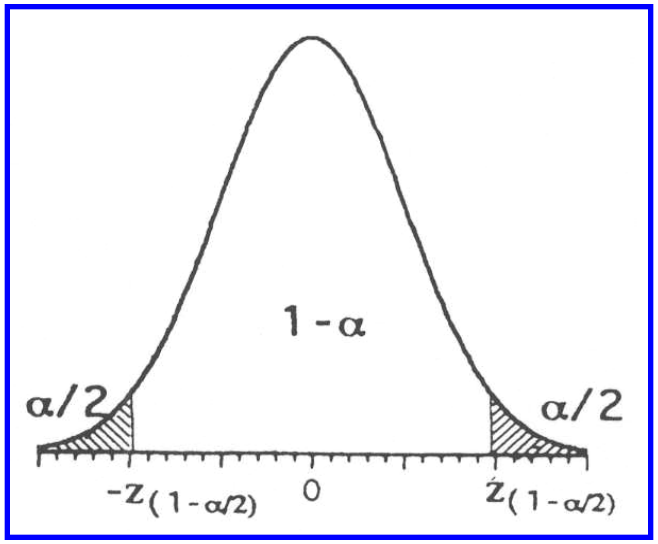
In questo caso la media campionaria, $\overline{X} = \Sigma_i x_i/n$, è anch'essa distribuita normalmente, con valore centrale pari alla media di popolazione $\mu$ e varianza $\sigma^2/n$, ossia:

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

Passando alla curva normale standardizzata, ossia introducendo la variabile:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$ si ha: $$z \sim N(0,1)$$

l'intervallo di fiducia corrispondente ad un livello di significatività $\alpha$ si può visualizzare graficamente sulla curva normale standard, tracciandone i limiti:




L'intervallo di fiducia è dunque espresso come:

$$-z_{(1-\alpha/2)} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{(1-\alpha/2)}$$

Trasformando l'equazione si ottiene:

$$\bar{X} - \frac{z_{(1-\alpha/2)}\,\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{(1-\alpha/2)}\,\sigma}{\sqrt{n}}$$

Il valore di $z_{(1-\alpha/2)}$ dipenderà da $\alpha$, ossia da P = 1- $\alpha$ :

| 1-α (%) | 99.9 | 99 | 95 | 90 | 50 |
|---|---|---|---|---|---|
| $z_{(1-\alpha/2)}$ | 3.29 | 2.58 | 1.96 | 1.64 | 0.67 |

## Caso 2: popolazione distribuita normalmente, con varianza ($\sigma^2$) ignota; numero di analisi sufficientemente elevato (n > 30)

In questo caso si può scrivere, in prima approssimazione:

$$z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

ossia considerare la varianza campionaria al posto della varianza di popolazione nel definire la variabile z ma impiegare comunque la distribuzione normale standard, invece che la t di Student, che sarebbe quella corretta, perché per $n \geq 30$ quest'ultima converge con la gaussiana standard.

L'intervallo di fiducia diventa dunque:

$$-z_{(1-a/2)} \leq \frac{\overline{X} - \mu}{s/\sqrt{n}} \leq z_{(1-a/2)}$$



$$\overline{X} - \frac{z_{(1-\alpha/2)}\, s}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{z_{(1-\alpha/2)}\, s}{\sqrt{n}}$$

## Caso 3: popolazione distribuita normalmente, con varianza ($\sigma^2$) ignota; numero di dati esiguo (n < 30)

E' il caso più frequente, legato alla difficoltà di effettuare un numero elevato di analisi o di replicati a causa del costo dell'analisi, della disponibilità limitata di materiale da analizzare, ecc..

Per le proprietà della distribuzione t di Student si può scrivere:

$$z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

L'intervallo di fiducia è dunque:

$$-t_{n-1\ (1-\alpha/2)} \leq \frac{\overline{X} - \mu}{s/\sqrt{n}} \leq t_{n-1\ (1-\alpha/2)}$$

$$\overline{X} - \frac{t_{n-1\ (1-\alpha/2)}\ s}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{t_{n-1\ (1-\alpha/2)}\ s}{\sqrt{n}}$$

| Gradi di libertà | $t_{0.95}$ | $t_{0.975}$ | $t_{0.995}$ |
|---|---|---|---|
| 1 | 6.31 | 12.71 | 63.66 |
| 2 | 2.92 | 4.30 | 9.92 |
| 3 | 2.35 | 3.18 | 5.84 |
| 4 | 2.13 | 4.78 | 4.60 |
| 5 | 2.01 | 2.57 | 4.03 |
| 6 | 1.94 | 2.45 | 3.71 |
| 7 | 1.89 | 2.37 | 3.50 |
| 8 | 1.86 | 2.31 | 3.55 |
| 9 | 1.83 | 2.26 | 3.25 |
| 10 | 1.81 | 2.23 | 3.17 |
| 20 | 1.72 | 2.09 | 2.85 |
| 30 | 1.70 | 2.04 | 2.75 |
| 60 | 1.67 | 2.00 | 2.66 |
| 120 | 1.66 | 1.98 | 2.62 |
| $\infty$ | 1.645 | 1.96 | 2.58 |

# Un esempio numerico: caso 3

Supponiamo di aver estratto cinque campioni di un polimero prodotto in un laboratorio industriale ed aver misurato la sua densità, ottenendo i valori:

| campione | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| densità ($g/cm^3$) | 1.2151 | 1.2153 | 1.2155 | 1.2145 | 1.2151 |

Facendo i calcoli opportuni si ottiene:

$\overline{X}$= 1.21510 s = 0.00037

Si noti che i due valori vengono considerati con una cifra in più rispetto a quelle significative, in vista dei calcoli successivi.

Nell'ipotesi che i cinque dati costituiscano un campione random estratto da una popolazione normale con media $\mu$ e varianza $\sigma^2$ (non nota) e poiché n < 30 l'intervallo per $\mu$ ad un livello di fiducia del 95% (P = 0.95 = 1-$\alpha \Rightarrow \alpha$ = 0.05) è dato da:

$$\overline{X} \pm t_{4\,(1-0.025)} \times s/\sqrt{n} = 1.21510 \pm 4.78 \times 0.00037/\sqrt{5} = 1.2151 \pm 0.0009$$

# Intervallo di fiducia per la differenza fra due medie

**Caso 1: popolazioni a cui si riferiscono le due medie distribuite normalmente e con varianze ($\sigma_1^2$ e $\sigma_2^2$) note**

Supponiamo di avere a disposizione due set di dati analitici ottenuti da campioni estratti dalle due popolazioni:

$x_1, x_2, \ldots, x_{n1}$ $\qquad$ $y_1, y_2, \ldots, y_{n2}$

l'ipotesi di partenza è: $x \sim N(\mu_1, \sigma_1^2)$ e $y \sim N(\mu_2, \sigma_2^2)$

Per le proprietà della funzione speranza matematica, la miglior stima per la differenza $\mu_1 - \mu_2$ è la differenza fra le due medie campionarie, a loro volta stimatori corretti di $\mu_1$ e $\mu_2$, rispettivamente:

$$\overline{X} = \sum_{i=1}^{n1} x_i / n_1 \quad \text{e} \quad \overline{Y} = \sum_{i=1}^{n2} y_i / n_2$$

Inoltre, la varianza relativa alla differenza $\overline{X} - \overline{Y}$ è data dalla somma delle varianze delle due medie campionarie: $\sigma_1^2/n_1$ e $\sigma_2^2/n_2$.

Si può scrivere quindi:

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

ovvero, passando alla normale standardizzata:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

In analogia con quanto visto per la singola media, l'intervallo di fiducia è dunque esprimibile così:

$$(\overline{X} - \overline{Y}) \pm \left( z_{1-\frac{\alpha}{2}} \right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Caso 2: popolazioni a cui si riferiscono le due medie distribuite normalmente, con varianze ($\sigma_1^2$ e $\sigma_2^2$) ignote; dimensioni dei campioni sufficientemente grandi ($n_1$ ed $n_2 > 30$)**

In questo caso si possono usare, con buona approssimazione, le varianze campionarie relative ai due set di dati al posto delle rispettive varianze di popolazione.

Si ha quindi:

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \overline{X})^2}{n_1 - 1} \quad e \quad s_2^2 = \frac{\sum_{i=1}^{n_2}(y_i - \overline{Y})^2}{n_2 - 1}$$

L'intervallo di fiducia è dato da:

$$(\overline{X} - \overline{Y}) \pm \left( z_{1-\frac{\alpha}{2}} \right) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Un esempio numerico: caso 2

Siano disponibili due set di dati relativi al $\Delta H$ di neutralizzazione dell'HCl con NaOH (dati espressi in kJ $mol^{-1}$):

1) $\overline{X} = 57.34$, $s_1^2 = 2.726$, $n_1 = 65$

2) $\overline{Y} = 56.99$, $s_2^2 = 1.130$, $n_2 = 32$

L'intervallo di fiducia per la differenza fra le due medie, al 95% di fiducia (P = 0.95, a = 0.05, 1-a/2 = 0.975) è dato da:

$$\mu_1 - \mu_2 = (57.34-56.99) \pm z_{(0.975)} \times \sqrt{2.726/65 + 1.130/32} = +0.4 \pm 0.5$$

La differenza include lo zero, perché l'intervallo si estende da -0.1 a + 0.9, il che significa che le due medie non sono significativamente diverse al 95 % di fiducia.

**Caso 3: popolazioni a cui si riferiscono le due medie distribuite normalmente, con varianze ($\sigma_1^2$ e $\sigma_2^2$) ignote ma uguali ($\sigma^2$); dimensioni dei campioni esigue ($n_1$ e/o $n_2 < 30$)**

In questo caso occorre applicare una delle proprietà fondamentali della distribuzione t di Student, ossia:

se $A \sim N(0,1)$ e $B \sim \chi^2_n$ allora $A/(B/\mathbf{n})^{1/2} \sim \mathbf{t_n}$

In particolare, come già visto in precedenza, risulta:

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1) \qquad (a)$$

d'altra parte valgono anche le relazioni:

$$(n_1\text{-}1)\, s_1^2 / \sigma^2 = \sum_{i=1}^{n_1}(x_i-\overline{X})^2/\sigma^2 \sim \chi^2_{n1-1}$$

$$(n_2\text{-}1)\, s_2^2 / \sigma^2 = \sum_{i=1}^{n_2}(y_i-\overline{Y})^2/\sigma^2 \sim \chi^2_{n2-1}$$

poiché la somma di due distribuzioni chi-quadro è essa stessa una distribuzione chi-quadro, avente un numero di gradi di libertà pari alla somma di quelli delle due distribuzioni, vale la relazione:

$$\frac{\sum_{i=1}^{n1}(x_i-\overline{X})^2+\sum_{i=1}^{n2}(y_i-\overline{Y})^2}{\sigma^2}=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}\sim\chi^2_{n1+n2-2} \qquad (b)$$

Applicando alle distribuzioni espresse dalle relazioni (a) e (b) la proprietà fondamentale della distribuzione t di Student vista in precedenza, si può scrivere:

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\Bigg/\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}\;\frac{1}{n_1+n_2-2}\right]^{1/2}\sim t_{n1+n2-2}$$

Ponendo:

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Si ottiene:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n1+n2-2}$$

L'intervallo di fiducia sulla differenza delle due medie si può esprimere dunque come:

$$(\overline{X} - \overline{Y}) \pm t_{n_1+n_2-2}(1 - \alpha/2) \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

# Un esempio numerico: caso 3

Supponiamo di avere a disposizione due set di dati relativi alla densità di un polimero prodotto industrialmente nelle stesse condizioni ma con monomeri prelevati da due fornitori diversi.

Occorre valutare se le densità medie dei polimeri ottenuti dai due monomeri siano significativamente diverse, ossia se la loro differenza sia significativamente diversa da zero.

Primo set di valori (x)

1.2151 1.2153 1.2155 1.2145 1.2151

Secondo set di valori (y)

1.2167 1.2176 1.2157 1.2158 1.2167

Facendo i calcoli opportuni si ottiene:

$\overline{X} = 1.21510$ $s^2(x) = 14.0 \times 10^{-8}$ $n_1 = 5$

$\overline{Y} = 1.21650$ $s^2(y) = 60.5 \times 10^{-8}$ $n_2 = 5$

In questo caso le dimensioni dei due campioni sono troppo esigue per applicare la formula del caso 2, dunque, ipotizzando che le varianze associate ai due set di dati siano le stesse (il che verrà dimostrato con un test apposito successivamente), calcoliamo:

$$s^2 = \frac{(5-1)14.0 + (5-1)60.5}{5+5-2} \times 10^{-8} = 37.25 \times 10^{-8} \Rightarrow s = 6.10 \times 10^{-4}$$

Nell'ipotesi di adottare un livello di fiducia del 95%, ossia un valore di P = 0.95 e quindi $\alpha$ = 0.05, calcoliamo:

$$t_{5+5-2}\,(0.975) = t_8\,(0.975) = 2.31$$

In definitiva l'intervallo di fiducia per la differenza fra le medie delle densità dei due polimeri è:

$$(1.21510\text{-}1.21650) \pm 2.31 \times 6.10 \times 10^{-4} \times \sqrt{2/5} = -0.0014 \pm 0.0009$$

L'intervallo di fiducia **NON** include lo zero, quindi la variazione del monomero ha provocato una variazione significativa di densità.

**Caso 4: popolazioni a cui si riferiscono le due medie distribuite normalmente, con varianze ($\sigma_1^2$ e $\sigma_2^2$) ignote e diverse; dimensioni dei campioni esigue ($n_1$ ed $n_2 < 30$) $\Rightarrow$ Problema di Fisher-Behrens**

In questa situazione si può solo adottare una soluzione approssimata che si basa sulla distribuzione:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

dove il numero n di gradi di libertà è espresso dalla relazione:

$$\frac{1}{\nu} = \frac{1}{n_1 - 1}\left(\frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2 / n_2}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2$$

L'intervallo di fiducia si può esprimere quindi con la relazione:

$$(\overline{X} - \overline{Y}) \pm \; t_{\nu}(1 - \alpha/2) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Un esempio numerico: caso 4

Un esempio tipico del problema di Fisher-Behrens è la verifica della presenza di un'eventuale differenza fra i valori forniti da due diversi metodi analitici sullo stesso campione.

Supponiamo di avere a disposizione i seguenti dati, ottenuti replicando 10 volte l'analisi della stessa soluzione con i due metodi:

| Dato | Metodo A | Metodo B |
|---|---|---|
| 1 | 5.11 | 5.18 |
| 2 | 5.14 | 5.13 |
| 3 | 5.13 | 5.27 |
| 4 | 5.17 | 5.12 |
| 5 | 5.12 | 5.27 |
| 6 | 5.08 | 5.29 |
| 7 | 5.15 | 5.17 |
| 8 | 5.20 | 5.28 |
| 9 | 5.16 | 5.14 |
| 10 | 5.14 | 5.18 |

Metodo A

$\overline{X} = 5.140$

$s_1^2 = 11.11 \times 10^{-4}$

$n_1 = 10$

Metodo B

$\overline{Y} = 5.203$

$s_2^2 = 45.34 \times 10^{-4}$

$n_2 = 10$

La differenza numerica (il rapporto è circa 1 a 4) suggerisce che le due varianze campionarie, $s_1^2$ e $s_2^2$, e quindi quelle di popolazione, $\sigma_1^2$ e $\sigma_2^2$, di cui esse sono, rispettivamente, stimatori corretti, siano significativamente diverse. Come si vedrà in seguito, questo verrà dimostrato da un test specifico per le varianze.

Ci si ritrova dunque nelle condizioni al contorno tipiche del Problema di Fisher-Behrens.

Il numero di gradi di libertà $\nu$ necessario in questo caso si ricava dall'equazione:

$$\frac{1}{\nu} = \frac{1}{9}\left(\frac{1.111}{5.645}\right)^2 + \frac{1}{9}\left(\frac{4.534}{5.645}\right)^2 \Rightarrow \quad \nu = 13.2 \approx 13$$

Scegliendo un livello di fiducia del 95% si ha $\alpha$ = 0.05 e quindi occorre valutare $t_{13}(0.975) = 2.16$.

L'intervallo di fiducia si ricava quindi dalla relazione:

$$(5.140\text{-}5.203) \pm 2.16 \times \sqrt{1.111 + 4.534} \times 10^{-2} = -0.06 \pm 0.05$$

L'intervallo al 90% di fiducia NON comprende lo zero, quindi esiste una differenza significativa fra i due metodi.