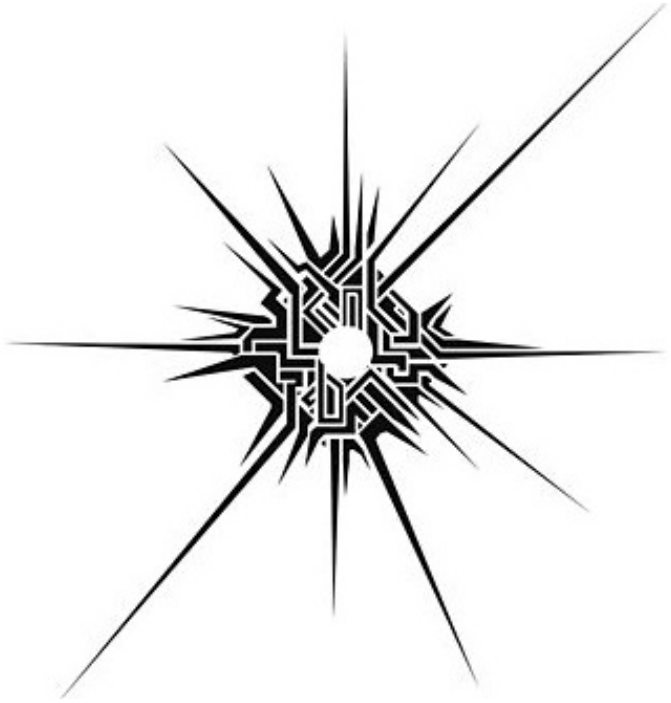


Chimica Analitica Magistrale

Docente
Andreas Stephan Lesch

Appunti di lezione

Redattore:

Alessandro Suprani

alessandro.suprani@studio.unibo.it

Laurea Magistrale in Chimica Industriale

Anno Accademico 2024/2025

Indice

Parte I

Statistica

1 Le basi della statistica

1.1 Popolazione e campione, normalità

1.1.1 Popolazione e campione

Nella statistica, i termini **popolazione** e **campione** fanno riferimento a concetti ben distinti:

- **Popolazione**: l'insieme di tutte le possibili osservazioni rilevabili.
- **Campione**: una porzione della popolazione selezionata per l'analisi.

1.1.2 Normalità

Dopo aver selezionato il campione, è necessario verificarne la normalità, ossia stabilire se i dati seguono una distribuzione normale (la "Campana di Gauss"). Se i dati sono normalizzati, è possibile applicare direttamente i test statistici. In caso contrario, si può applicare il teorema del limite centrale per trattare il campione come distribuito normalmente.

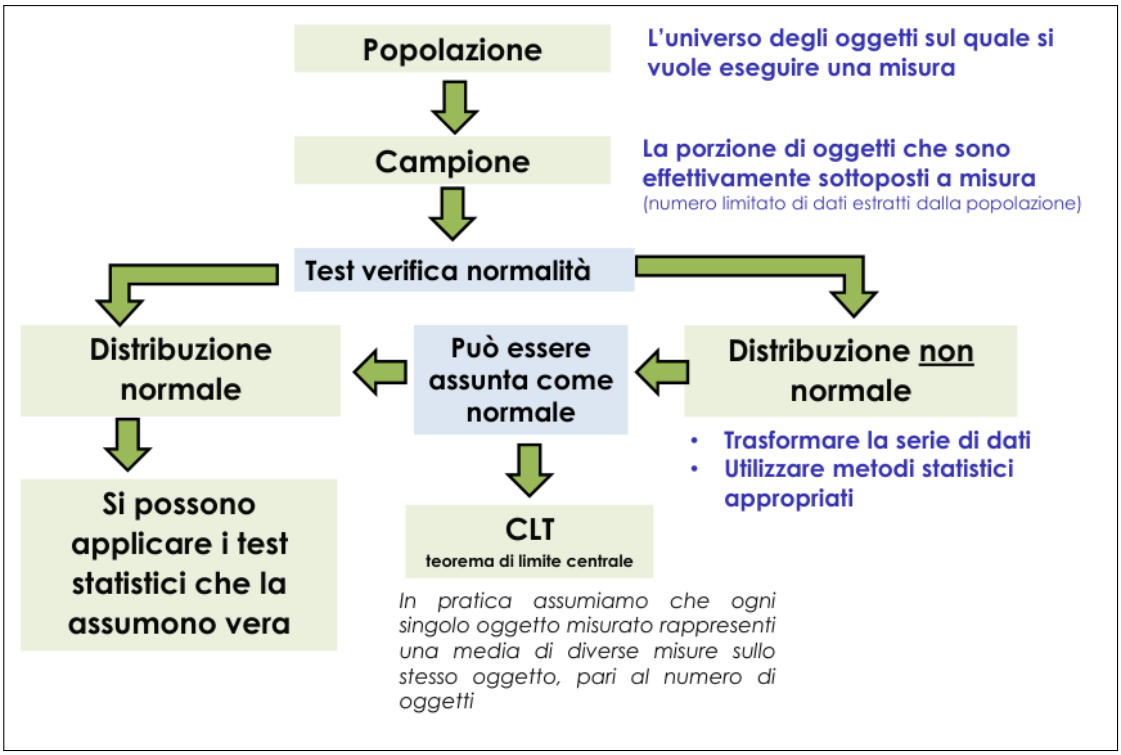


Figura 1: Diagramma di flusso per la gestione dei dati

1.2 Media e Deviazione Standard

1.2.1 Media

La media ($\overline{x}$ o μ) si utilizza per variabili a intervallo (con zero arbitrario, come la temperatura in °C) o a rapporto (zero assoluto, come la temperatura in K). Ed è calcolata:

$$\mu = \frac{\sum_{i=1} x_i}{n}$$

1.2.2 Normalità

Tuttavia, la media **non** è sufficiente a descrivere un campione in modo completo, perché campioni diversi possono avere la stessa media ma una dispersione di dati completamente diversa, come riportato in **Figura 2**.

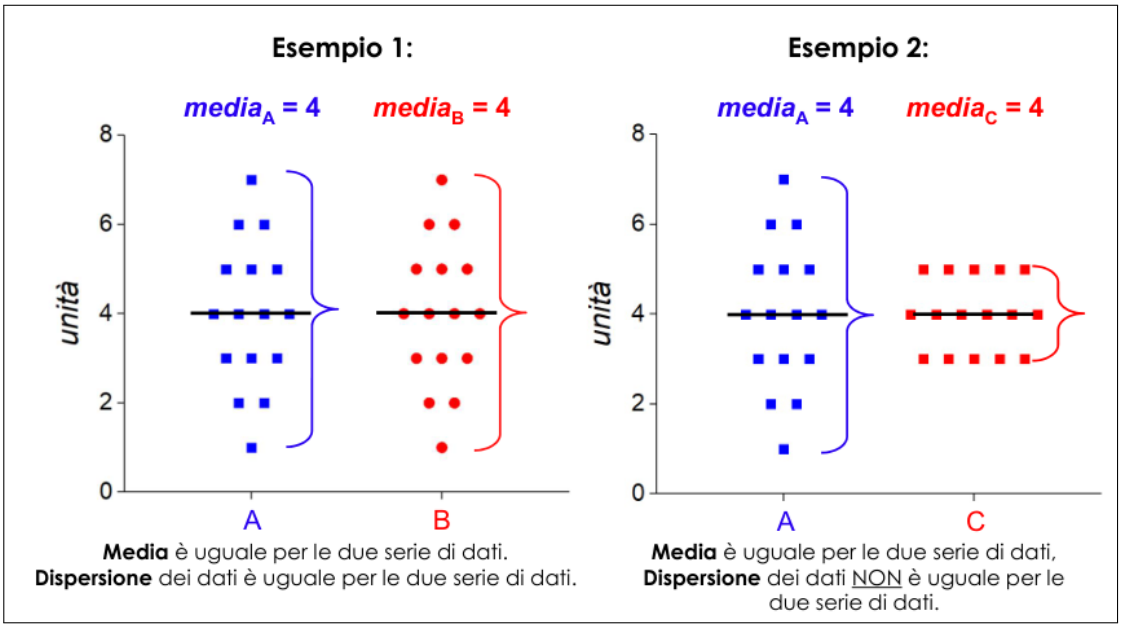


Figura 2: Esempi di come possono presentarsi i dati

Perciò, si calcola la **varianza della popolazione** (σ^2), che misura la dispersione dei dati rispetto alla media:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Per ripristinare le unità originali, si calcola la **deviazione standard della popolazione** (σ):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Varianza e deviazione standard del campione sono indicate rispettivamente con s^2 e s, ma seguono lo stesso principio.

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}$$

Varianza del Campione

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}}$$

Deviazione Standard del Campione

1.3 La distribuzione di Laplace-Gauss

Quando molteplici fattori indipendenti influenzano una misurazione, il risultato segue una distribuzione normale, detta anche **Campana di Gauss**. L'ampiezza della campana riflette la variabilità delle misure nel campione.

$$y = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Proprietà delle distribuzioni gaussiane è che l'area compresa nello stesso intervallo $\pm\sigma$ intorno alla media μ rappresenta la stessa percentuale dell'area totale.

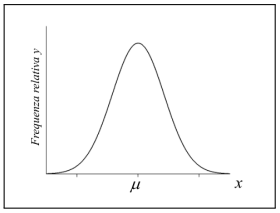


Figura 3: Distribuzione secondo la campana di Gauss

1.3.1 La variabile ridotta

Nasce la necessità di normalizzare il valore di $\mu \pm \sigma$ per poter confrontare le distribuzioni e quindi si introduce la **variabile ridotta z**:

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

La variabile ridotta presenta media pari a 0 e deviazione std uguale a 1.

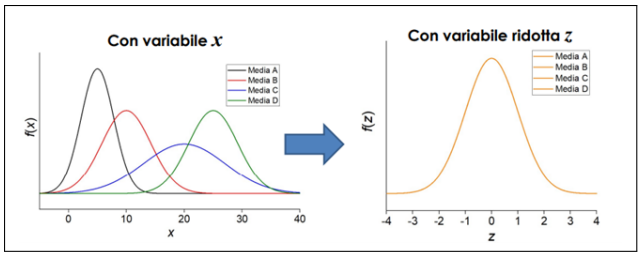


Figura 4: Effetto dell'introduzione della variabile ridotta z (notare il cambio di assi)

E' importante avere i dati distribuiti normalmente in quanto nell'intervallo di deviazione $\pm\sigma$ dal valore atteso è possibile trovare il 68,27% di tutti i valori, valore che cresce a 95,45% se l'intervallo di deviazione sale a $\pm 2\sigma$.

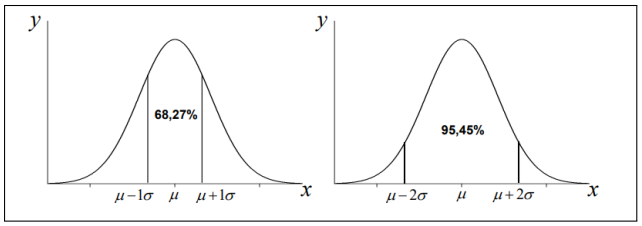


Figura 5: Rappresentazione grafica degli intervalli di deviazione

1.4 Teorema del limite centrale (CLT)

Il teorema del limite centrale afferma che la somma di n variabili indipendenti aventi identica distribuzione è una variabile che si distribuisce normalmente qualsiasi sia la tipologia di distribuzione di partenza Se si prendono tutti i possibili campioni, ognuno di dimensione n, da qualsiasi popolazione di media μ e deviazione standard σ, la distribuzione delle medie dei campioni avrà media uguale alla media della popolazione, varianza pari a $\frac{\sigma^2}{n}$ e la deviazione standard delle medie (definito **ERRORE STANDARD**) risulterà essere $\frac{\sigma}{\sqrt{n}}$ che sarà distribuita normalmente se lo era la distribuzione di origine oppure tenderà ad essere normale per numero grande di campioni come mostrato in **Figura 6**.

- La deviazione standard è una misura della **variabilità dei dati** e rimane costante all'aumentare di n
- L'errore standard misura la **precisione della media campionaria** e diminuisce all'aumentare di n

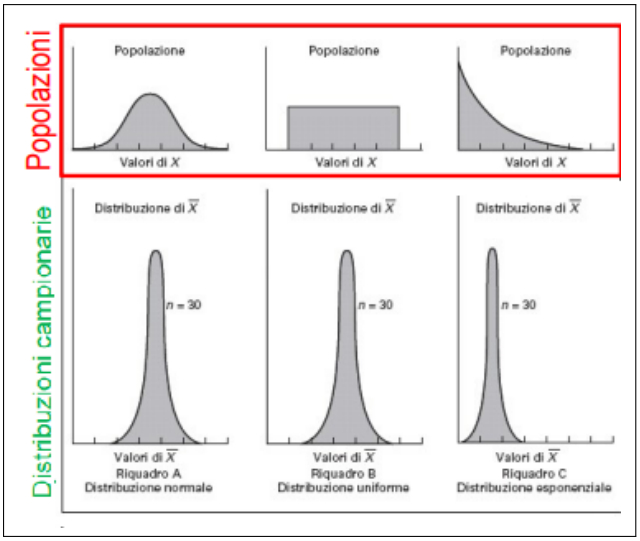
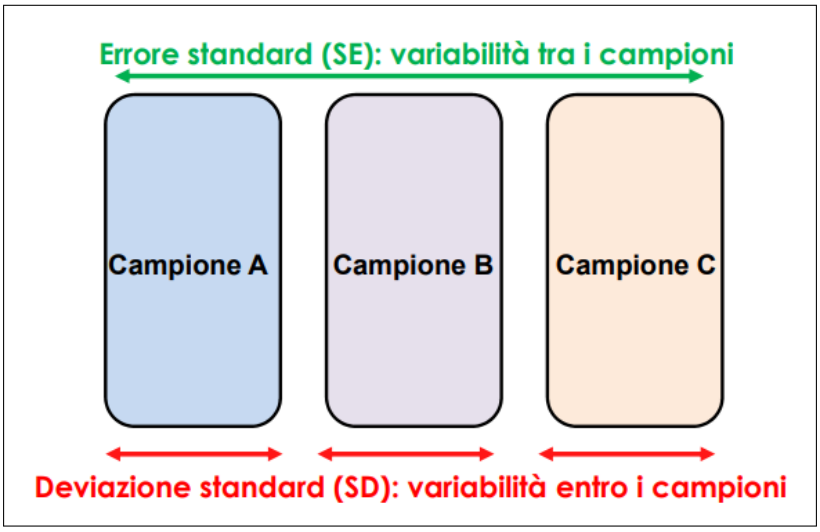


Figura 6: Distribuzione della media delle medie campionarie



Confidenza

Dato che il valore medio $\overline{x}$ è una stima del valore vero μ, abbiamo bisogno di definire un intervallo all'interno del quali si possa assumere che giaccia il valore vero: l'intervallo di confidenza.

$$\overline{x} - z\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{x} + z\left(\frac{\sigma}{\sqrt{n}}\right)$$

1.5 La legge di Student

$$\mu = \overline{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

Questa legge sostituisce la legge normale quando la deviazione standard è una stima campionaria.

Il valore t è un valore tabulato che dipende dal numero di gradi di libertà (df) e dall'intervallo di confidenza (ci) scelto. La distribuzione t viene utilizzata quando n è piccolo (<30) e σ è sconosciuto. Il termine t è utilizzato per applicare i test di significatività alle basi di dati ottenute.

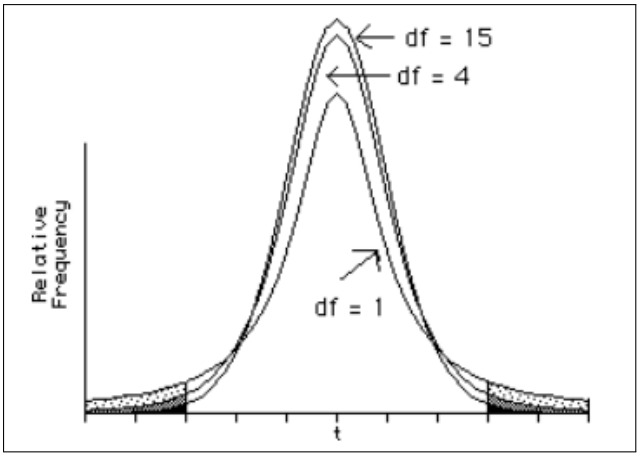


Figura 7: Differenza apportata dai gradi di liberà alla curva descritta.

1.6 Test di significatività

Nell'analisi delle misure sperimentali, è essenziale utilizzare test di significatività per valutare se vi siano differenze significative tra i risultati ottenuti. Questi test si fondano sulla formulazione di un'ipotesi nulla (H_0), la quale viene accettata o rifiutata in base ai dati raccolti. In alternativa, si prende in considerazione un'ipotesi alternativa (H_a o H_1), che rappresenta la negazione dell'ipotesi nulla.

Un metodo di stima è considerato "robusto" quando dimostra una bassa sensibilità a leggere deviazioni dalle ipotesi di base del modello, come piccoli scostamenti dalla distribuzione normale. Generalmente, un test respinge l'ipotesi nulla (assenza di differenza) quando la probabilità che tale differenza si verifichi per caso è inferiore al 5% (P < 0,05). In tal caso, la differenza viene considerata significativa, con una probabilità del 95% (P = 0,95), corrispondente a $1 - \alpha = 1 - 0{,}05$.

Tuttavia, se l'ipotesi nulla non viene rifiutata, ciò non dimostra necessariamente che sia vera (bensì che non è stata falsificata), poiché esistono due tipi di errori:

- Errore di prima specie (falso positivo): rifiutare H_0 quando essa è vera.
- Errore di seconda specie (falso negativo): accettare H_0 quando essa è falsa.

1.6.1 Test di accuratezza (Test t)

Confronto media - valore noto

Questo è un test a una coda che risponde alla domanda: "Il campione di dimensione n, con media $\overline{x}$ e varianza s^2, può considerarsi appartenente a una popolazione con media μ?"

H_0 : non ci sono differenze significative tra i due valori.

H_1 : ci sono differenze significative tra i due valori.

Si calcola il valore di t:

$$t_{\text{calc}} = \left| \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \right|$$

e si confronta con il valore critico t_{crit} (tabulato), corrispondente ai gradi di libertà $(n - 1)$.

Se $t_{\text{calc}} > t_{\text{crit}}$, allora H_0 viene rigettata.

Confronto media - media

Questo è un test a due code che risponde alla domanda: "I due campioni, di dimensioni n_1 e n_2, con medie $\overline{x}_1$ e $\overline{x}_2$ e varianze s_1^2 e s_2^2, possono considerarsi appartenenti a una popolazione con la stessa media?"

H_0 : non ci sono differenze significative tra le due medie.

H_1 : ci sono differenze significative tra le due medie.

Se s_1^2 e s_2^2 sono **omoschedastiche** (ovvero hanno la stessa varianza), possiamo calcolare la varianza combinata come:

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

e quindi calcolare:

$$t_{\text{calc}} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Se $t_{\text{calc}} > t_{\text{crit}}$, allora H_0 viene rigettata.

Test t per valori accoppiati

Quando lo stesso dato è ottenuto utilizzando due metodi diversi, si rende necessario adattare il test t. Si procede calcolando la differenza d tra ciascuna coppia di risultati:

$$d_{\text{coppia}} = y_{\text{Serie1}} - y_{\text{Serie2}}$$

e successivamente calcolando la media delle differenze:

$$d_{\text{med}} = \frac{\sum d_{\text{coppia}}}{n}$$

dove n è il numero di coppie. Da qui è possibile eseguire un test t, calcolando:

$$t_{\text{calc}} = \left| \frac{d_{\text{med}}}{\frac{s_d}{\sqrt{n}}} \right|$$

La H_0 afferma che non ci sono differenze significative tra le popolazioni dei due campioni, e viene rigettata se $t_{\text{calc}} > t_{\text{crit}}$.

1.6.2 Test di precisione (Test F)

Il test F confronta il rapporto tra due varianze per verificare se le due popolazioni siano normali e abbiano varianze identiche (H_0). Il valore di F è calcolato come:

$$F_{\text{calc}} = \frac{s_1^2}{s_2^2}$$

con $F > 1$.

Se $F_{\text{calc}} > F_{\text{crit}}$, allora H_0 viene rigettata.

2 ANalysis Of VAriance (ANOVA)

L'ANOVA è una tecnica statistica utilizzata per confrontare tra loro due o più medie o per valutare l'effetto di uno o più fattori di variazione sulle medie e stimarne gli effetti. È impiegata per confrontare le medie di tre o più gruppi e determinare se almeno uno di essi è significativamente diverso dagli altri. Questa tecnica risulta particolarmente utile quando si vogliono analizzare variabili categoriali (fattori) rispetto a una variabile continua (risposta).

2.1 ANOVA a una via

L'ANOVA a una via consente di valutare se vi sia una differenza significativa tra le medie di più di due campioni. Durante un'ANOVA, esistono due principali fonti di variazione:

- Errori casuali nella misurazione: intrinseci nell'esecuzione delle misure, dovuti alla casualità dei processi di misura.
- Fattori controllati: processi di produzione, metodi analitici, analisti coinvolti, indicati come gruppi.

L'ANOVA è in grado di separare la variazione dovuta ai fattori controllati da quella dovuta alla casualità. Per ottenere risultati affidabili, l'ANOVA richiede almeno tre medie da confrontare.

L'ipotesi nulla H_0 afferma che il fattore controllato non ha alcuna influenza sui risultati delle prove, il che implica che i gruppi appartengano alla stessa popolazione (stessa distribuzione stocastica). L'ipotesi alternativa, H_1, suggerisce che almeno uno dei gruppi abbia una media significativamente diversa dagli altri.

Per verificare H_0, si calcola la **devianza totale**, espressa come:

$$SS_{\text{totale}} = \sum_{i=1}^{n} (x_i - \mu)^2$$

dove μ è la media generale dei dati.

I gradi di libertà sono calcolati come:

$$df_{\text{tot}} = n_{\text{repliche totali}} - 1$$

Ora è necessario calcolare la devianza **entro** i gruppi e **tra** i gruppi.

Devianza entro i gruppi

La devianza entro i gruppi riflette la variazione tra le repliche di ciascun esperimento all'interno dello stesso gruppo. È calcolata come:

$$SS_{A,\text{mr}} = \sum_{i=1}^{n} (x_i - \mu_A)^2$$

$$SS_{B,\text{mr}} = \sum_{i=1}^{n} (x_i - \mu_B)^2$$

La devianza complessiva all'interno dei gruppi è data da:

$$SS_{\text{entro i gruppi}} = SS_{A,\text{mr}} + SS_{B,\text{mr}}$$

I gradi di libertà relativi alla devianza entro i gruppi sono:

$$df_{\text{entro}} = h(n-1)$$

dove h è il numero di gruppi e n è il numero di repliche.

Devianza tra i gruppi

La devianza tra i gruppi rappresenta la variazione tra le medie dei diversi gruppi. Per calcolarla, si separa la variazione tra le prove da quella interna ai gruppi, basandosi sull'assunzione che ogni risultato all'interno di un gruppo sia pari alla media dei risultati ottenuti nello stesso gruppo. Si calcola come segue:

$$SS_A = (\mu_A - \mu_{\text{gen}})^2 \times N_A$$

$$SS_B = (\mu_B - \mu_{\text{gen}})^2 \times N_B$$

dove μ_{gen} è la media generale, mentre N_A e N_B sono rispettivamente le dimensioni dei gruppi A e B.

La devianza tra i gruppi è quindi:

$$SS_{\text{tra i gruppi}} = SS_A + SS_B$$

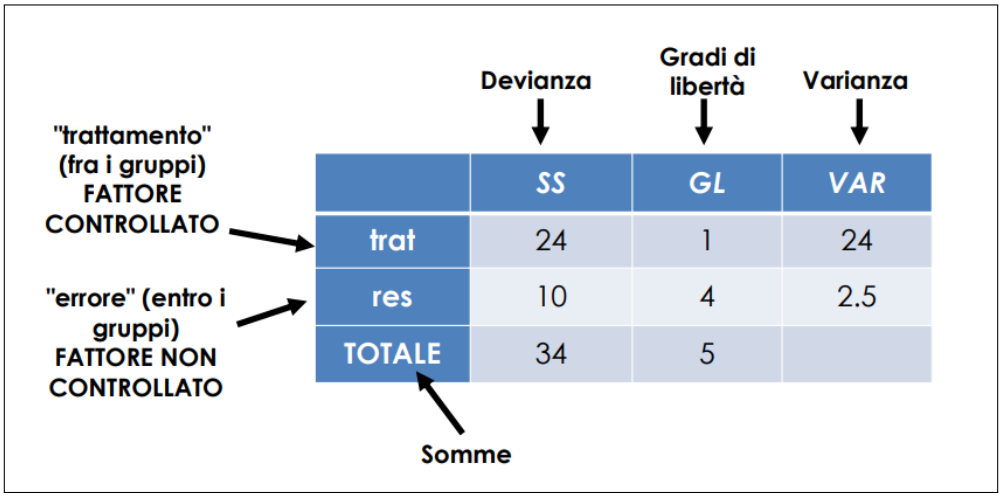
I gradi di libertà per la devianza tra i gruppi sono calcolati come:

$$df_{\text{tra}} = h - 1$$

Devianza totale

$$SS_{Totale} = SS_{entroigruppi} + SS_{traigruppi}$$

L'ANOVA può essere anche automatizzato



2.2 Test di Fisher

Il test di Fisher per l'ANOVA frappone le varianze precedentemente calcolate:

$$F_{calc} = \frac{VAR_{fraigruppi}}{VAR_{entroigruppi}}$$

Si sceglie il livello α di significatività del test e si individua il valore di F_{tab} ad una coda nelle tavole per probabilità P = 0,95 corrispondente ai gradi di libertà, considerando che **i df tra i gruppi sono in orizzontale mentre quelli entro i gruppi sono in verticale.**

Se $F_{calc>F_{tab}}$ l'ipotesi nulla è respinta (almeno un gruppo non appartiene all'origine degli altri)