**RELATIVE FREQUENCY** 🡪 è il rapporto tra il numero di volte in cui si verifica un evento di un'indagine statistica e il numero totale di eventi.

ANALISI DELLE COMPONENTI PRINCIPALI (**PCA**) = Tecnica per ridurre il numero più o meno elevato di variabili che descrivono un insieme di dati ad un numero minore di variabili latenti, limitando il più possibile la perdita di informazioni. Ciò avviene tramite una [trasformazione lineare](https://it.wikipedia.org/wiki/Trasformazione_lineare) delle variabili che proietta quelle originarie in un nuovo [sistema cartesiano](https://it.wikipedia.org/wiki/Sistema_cartesiano) nel quale la nuova variabile con la maggiore [varianza](https://it.wikipedia.org/wiki/Varianza) viene proiettata sul primo asse, la variabile nuova, seconda per dimensione della varianza, sul secondo asse e così via. Nell’ambito del **machine learning** viene utilizzata per generalizzare il modello riducendo la possibilità che possa essere soggetto ad overfitting, quando è presente un insieme molto ampio di features. E’ a tutti gli effetti un **algoritmo di machine learning non supervisionato** per ricavare le componenti primarie dei dati presenti nel dataset. Raramente la riduzione del set di features tramite PCA al minor numero possibile di PC (componenti primari) porta alle migliori performance, è necessario trovare il bilanciamento migliore tramite differenti test su modelli allenati con una differente livello di riduzione delle features.

**PRINCIPI PCA**: PCA is based on the **Pearson correlation coefficient framework** and inherits similar assumptions.

1. **Sample size:** Minimum of 150 observations and ideally a 5:1 ratio of observation to features (Pallant, 2010)
2. **Correlations:** The feature set is correlated, so the reduced feature set effectively represents the original data space.
3. **Linearity:** All variables exhibit a constant multivariate normal relationship, and principal components are a linear combination of the original features.
4. **Outliers:** No significant outliers in the data as these can have a disproportionate influence on the results.
5. **Large variance implies more structure:** high variance axes are treated as principal components, while low variance axes are treated as noise and discarded.