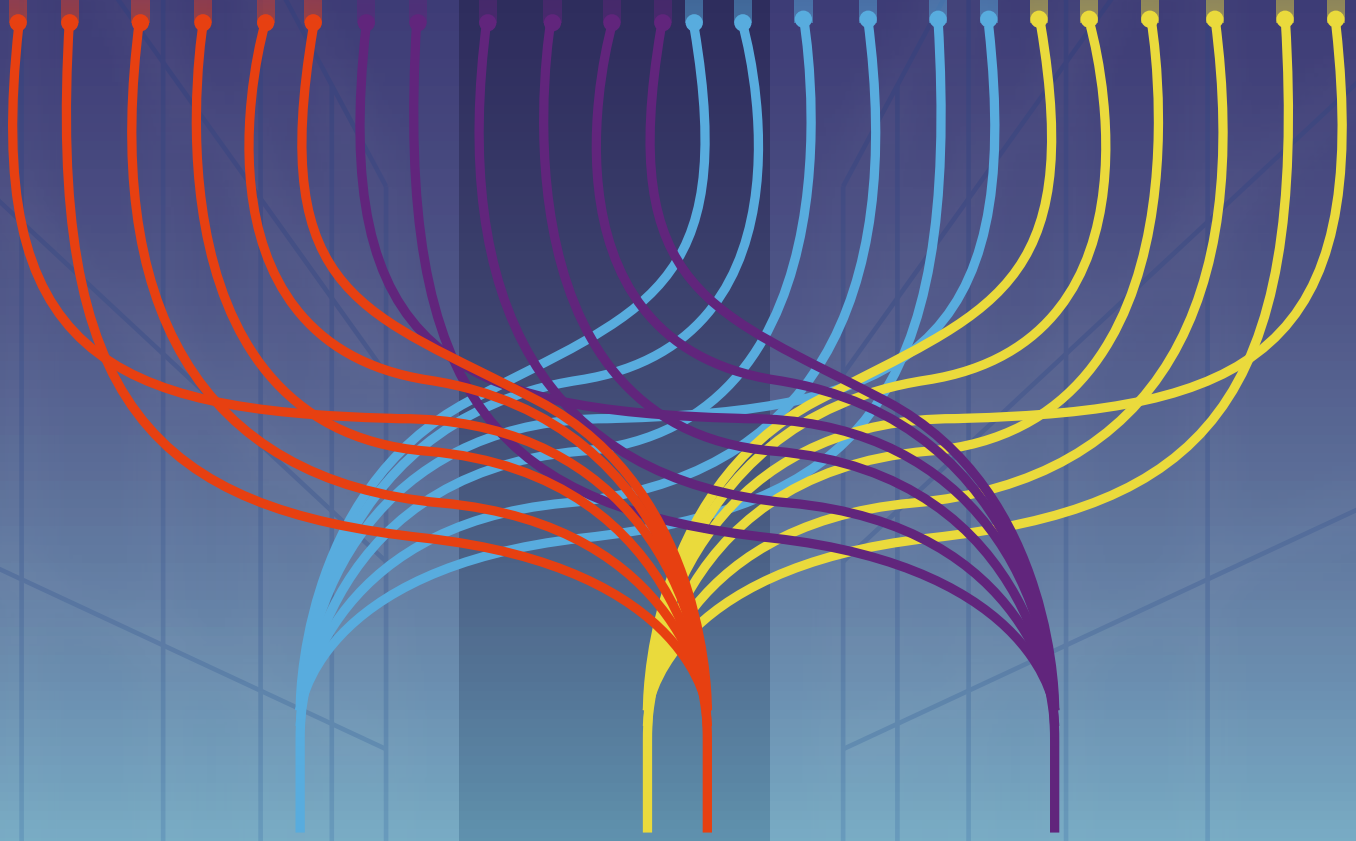
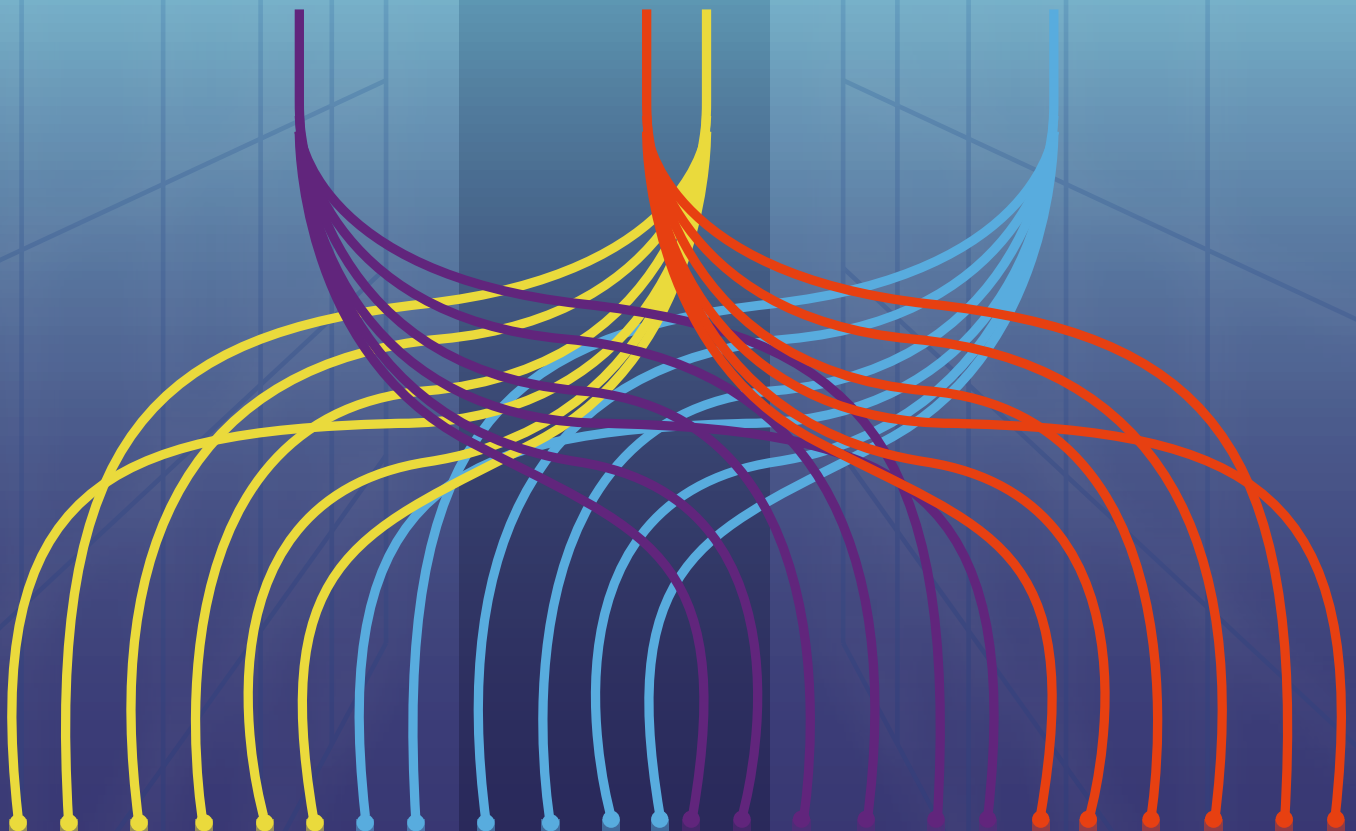
Maddalena Cavicchioli • Fabio Demaria • Maria Franco Villoria
Patrizio Frederic • Silvia Montagna • Isabella Morlini



Gimme (uni) MORE data:
workbook to the data challenge



Dipartimento di Economia
Marco Biagi

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Comunicazione
ed Economia

Indice


Ideas
15%
30%
30%
5%
20%
option
40%
40%
30%
30%
10%
35%
70%
Business

Prefazione

Cari studenti, care studentesse,

Benvenuti in un viaggio attraverso il mondo della statistica applicata. Questo volume è stato pensato per offrirvi un'introduzione pratica all'analisi statistica, utilizzando strumenti accessibili come Microsoft Excel e dati moderni che fanno parte della nostra vita quotidiana.

Per rendere l'esperienza il più concreta e interessante possibile, abbiamo scelto di lavorare con i dati forniti da Spotify, una piattaforma che molti di voi probabilmente utilizzano per ascoltare musica. **Spotify** dispone di API (Application Programming Interface), cioè un insieme di strumenti che consentono di accedere ed interagire con i dati della piattaforma. Ad esempio, le API permettono di cercare nel catalogo Spotify per trovare brani, artisti o playlist specifiche. Le API (con autorizzazioni specifiche) consentono agli sviluppatori Spotify di accedere alle playlist degli utenti, inserendo o rimuovendo brani. Quello che ci interessa in questo volume però è, in particolare, l'accesso ai dati sui brani e alle caratteristiche audio: Spotify, infatti, mette a disposizione informazioni dettagliate su ogni canzone quali, ad esempio, il titolo, l'album, l'artista, la popolarità del pezzo e le caratteristiche audio che descrivono il brano musicale. Tra le caratteristiche audio, l'API fornisce per ogni brano la sua durata, il tempo (BPM), la tonalità, il modo (maggiore o minore), la "ballabilità" (ovvero quanto è facile ballare un pezzo), ed altri attributi utili per l'analisi musicale.

Ora, immaginate di essere dei data analyst per una casa discografica. Quali sono le caratteristiche che rendono una canzone popolare? È il ritmo? La ballabilità? O forse la sua valenza, cioè la felicità che trasmette una canzone? Insomma, per avere un successo mondiale, il prossimo singolo dei Måneskin (se ci sarà!) dovrà essere struggente come *Coraline* o esplosivo come *Zitti e buoni*? Beh, a questa e ad altre domande possiamo provare a rispondere attraverso un'analisi statistica!
I dati che analizzeremo in questo volume possono essere liberamente scaricati utilizzando questo link:

https://www.economia.unimore.it/it/terza-missione/public-engagement

Questo dataset è veramente enorme: si tratta di oltre 230000 canzoni di tutti i generi musicali (Indie, Pop, Rock, R&B e molti altri ... incluse canzoni per film e colonne sonore) estratti da Spotify tra Agosto e Settembre 2018, per un totale di 33.71 MB di dati. Sicuramente, spulciando in questo dataset, troverete anche qualche artista di vostra conoscenza!
Una descrizione dettagliata delle informazioni disponibili per ogni brano è disponibile direttamente dalla Web API di Spotify:

https://developer.spotify.com/documentation/web-api/reference/get-audio-features

Utilizzando i dati Spotify, andremo a ripercorrere gli step principali di un'analisi statistica per cercare di estrarre informazioni utili dai dati stessi. Per l'analisi, utilizzeremo Microsoft Excel: grazie alle sue funzioni integrate, come tabelle pivot, formule statistiche e strumenti grafici, ci permetterà di organizzare, elaborare e visualizzare i dati in modo semplice e immediato.

Il volume si compone di tre Capitoli principali. Il primo Capitolo tratta la **statistica descrittiva**, il punto di partenza per qualsiasi analisi. Imparerete come riassumere e visualizzare i dati in modo semplice ed efficace, usando grafici e indicatori di posizione e dispersione. Questo vi permetterà di familiarizzare con i concetti fondamentali che vi guideranno nelle fasi successive.

Il secondo Capitolo è dedicato ad **inferenza e ai test statistici**. Scoprirete come verificare ipotesi e come determinare se esiste una relazione significativa tra le caratteristiche dei brani musicali, ad esempio tra la loro energia e la loro popolarità. Attraverso l'uso di Excel, esploreremo vari strumenti per condurre questi test in modo autonomo.

Infine, il terzo Capitolo introduce il concetto di **regressione lineare**, un potente metodo per modellare le relazioni tra variabili e fare previsioni. Imparerete come la regressione possa essere utilizzata per spiegare, ad esempio, come diverse caratteristiche di una canzone possono influire sul numero di ascolti o sulla sua presenza nelle playlist.

Il nostro obiettivo è quello di fornire non solo una base solida nella teoria statistica, ma anche gli strumenti pratici per applicarla. In un mondo sempre più orientato ai dati, saperli analizzare con competenza è una capacità fondamentale. Speriamo che questo volume vi aiuti a sviluppare queste competenze e a scoprire il potere che i numeri hanno nel descrivere il mondo che ci circonda, anche in contesti apparentemente lontani dalla matematica, come la musica.

https://

get-audio-features

API

Buon lavoro
e buon divertimento!

Capitolo[1]

Statistica descrittiva

I dati si presentano a noi attraverso la **matrice dei dati**, una tabella dove ogni riga corrisponde a una unità statistica, e ogni colonna ad un carattere misurato sulle unità statistiche. Ad esempio, nel nostro dataset le unità statistiche (e quindi le righe) sono canzoni su Spotify, e per ogni canzone abbiamo diverse informazioni (i.e., diverse colonne), come *l'artist_name*, la *popularity* o la *danceability*. Il numero di unità statistiche viene comunemente denotato con la lettera *n*. Il database Spotify è stato salvato in formato Excel e consta di $n = 232\,725$ brani (ogni riga è un brano) e di 20 variabili (o caratteri).
A seconda del valore che possono assumere, i caratteri si suddividono in due tipologie:

- **Caratteri quantitativi:** assumono valori numerici. Si possono suddividere in **discreti** (se assumono valori interi, ad esempio *popularity*) e **continui** (se assumono valori in un intervallo, ad esempio *loudness*)
- **Caratteri qualitativi:** le modalità non sono numeri, ma parole. Questi si classificano a loro volta in **ordinati**, se le diverse modalità/categorie hanno un ordine logico (ad esempio, il *Modo (maggiore/minore)* di un brano) o **sconnessi**, se non c'è un modo logico per ordinare le modalità (ad esempio, il *Genere musicale (R&B, Country, Hip-Hop, etc.)* di un brano). Attenzione: a volte le categorie di un carattere qualitativo vengono codificate numericamente!

È importante riconoscere il tipo di carattere perché determinerà il tipo di grafico/sintesi numerica che faremo. Possiamo considerare un carattere singolo (***statistica descrittiva univariata***) o più di uno. In particolare, vedremo come studiare le relazioni tra coppie di variabili (***statistica descrittiva bivariata***).

1.1 Statistica descrittiva univariata

Per un carattere **X** (ad esempio *Genere*), la collezione di tutte le modalità osservate per ogni singola unità statistica $\{x_1, \ldots, x_n\}$ viene chiamata **distribuzione unitaria semplice**. Nel nostro dataset, ogni colonna della tabella fornisce la distribuzione unitaria semplice del relativo carattere. L'obiettivo della statistica descrittiva è quello di capire l'andamento del collettivo come gruppo, non delle singole unità. Questo vuol dire che dobbiamo sintetizzare in qualche modo i dati.

La prima sintesi che possiamo fare è la costruzione della **distribuzione di frequenze**, cioè nell'identificare le diverse univoche K modalità che un carattere può assumere, e contare quante unità statistiche (nel nostro caso canzoni) presentano ciascuna modalità (**frequenze assolute** n_j, $j=1,\ldots,K$). A partire dalle frequenze assolute, possiamo costruire le **frequenze relative** ($f_j = n_j/n$) e **percentuali** ($p_j = f_j \times 100$), utili, ad esempio, per confrontare gruppi di diversa numerosità. Vediamo un esempio con la variabile *Genere* in Tabella 1. Notiamo, per esempio, che su 232 725 canzoni presenti nel database 51 451 sono canzoni pop e commerciali, ovvero il 22,11% del totale.

Genere	n_j	p_j
Pop e commerciali	51 451	22,11%
Rock e alternative	46 251	19,87%
Rap e urban	36 608	15,73%
Elettronica e moderni	36 299	15,60%
Classico e teatrale	34 988	15,03%
Jazz e tradizionali	27 128	11,66%
Totale complessivo	**232 725**	**100,00%**

Tabella 1: *Distribuzione del carattere Genere: la prima colonna indica le modalità del carattere, n_j il conteggio (frequenza assoluta), p_j la frequenza percentuale.*

HOW TO? **COME COSTRUIRE UNA DISTRIBUZIONE DI FREQUENZA IN EXCEL**

I fogli di calcolo (MS Excel, Calc di Open Office, i fogli di Google) sono strumenti molto comodi per presentare piccole basi di dati ma hanno grandi limitazioni per l'analisi. Ciò nonostante, alcune statistiche di base è possibile eseguirle. Useremo come esempio MS Excel, ma ogni procedura potrà essere implementata sugli altri fogli di calcolo.

Se volessimo costruire la distribuzione del macro-genere in colonna B, dovremmo procedere come in Figura 1: 1) selezionare la colonna B; 2) cliccare sulla scheda inserisci e inserire una tabella pivot; e 3 inserirla in un nuovo foglio. Una volta ottenuta la tabella dei conteggi procederemo a costruire la colonna delle frequenze percentuali come indicato in Figura 2: partiremo dalla cella C3 ottenuta dividendo B3 per il totale B9; i dollari $B$9 *bloccano* la riga 9 e la colonna B e potremo così trascinare la formula nelle celle sottostanti.

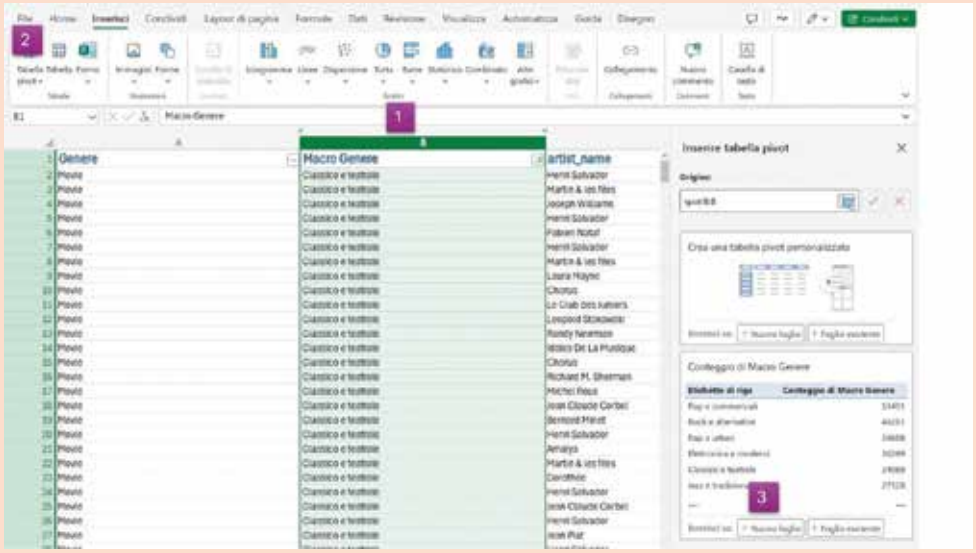


Figura 1: *Come creare una distribuzione di frequenza in Excel con le tabelle Pivot.*

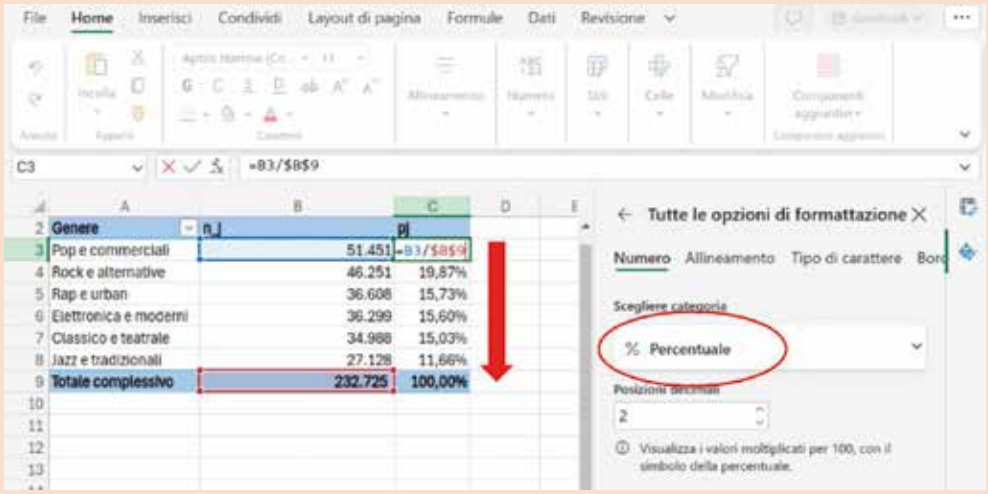


Figura 2: *Come creare la colonna delle frequenze relative.*

Spesso succede che le **variabili quantitative** assumono valori tutti diversi sulle *n* unità statistiche. In questo caso, la distribuzione di frequenze non è pratica, e si procede a una **suddivisione in classi**, ovvero, suddividere il range di possibili valori in intervalli tra loro disgiunti. Occorre decidere il numero di classi da utilizzare, e se considerare classi con la stessa ampiezza (differenza tra gli estremi dell'intervallo) o ampiezza variabile. Ad esempio, in Tabella 2 vediamo il *Tempo* delle canzoni suddiviso in classi; sotto ai 60 battiti al minuto (*bpm*), tra 60 *bpm* e 90 *bpm*, ecc. Per chiarezza espositiva abbiamo associato ad ogni classe un'etichetta descrittiva: molto lenti, lenti, moderati, veloci e molto veloci. Possiamo notare, per esempio, che il numero di brani molto lenti (1389) e quello di brani molto veloci (2584) rappresentano meno del 2% del totale (0,597% + 1,110% = 1,707%).

Classificazione	da	a	n_j	p_j
molto lento	0	60	1 389	0,597%
lento	60	90	4 7966	20,611%
moderato	90	140	128 338	55,146%
veloce	140	190	52 448	22,536%
molto veloce	maggiore di 190		2 584	1,110%
n			**232 725**	**100,000%**

Tabella 2: *Distribuzione del Tempo: la prima colonna indica le modalità del carattere, n_j il conteggio (frequenza assoluta), p_j la frequenza percentuale.*

HOW TO?

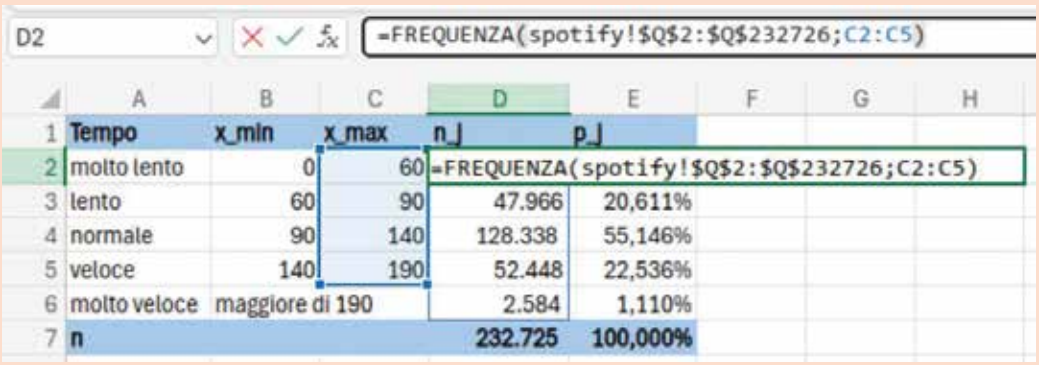
COME CREARE UNA DISTRIBUZIONE IN CLASSI SU EXCEL

Per creare la Tabella 2 abbiamo usato la funzione FREQUENZA, come esemplificato in Figura 3: abbiamo creato un nuovo foglio di lavoro, nelle colonne dalla C2 alla C5 abbiamo inserito gli estremi degli intervalli - abbiamo scelto 60, 90, 140, e 190 - che Excel userà per dividere e contare i dati, le celle D2:D6 saranno compilate automaticamente. Nella cella D2 scriveremo il comando

=FREQUENZA (spotify!$Q$2:$Q$232726;C2:C5)

Dove spotify!$Q$2:$Q$232726 indica le celle dalla Q2 alla 232 726, dove sono contenuti i dati sul tempo, nelle celle C2:C5 troveremo gli estremi degli intervalli decisi da noi.

Dato invio, in D2 troveremo il numero di brani con tempo inferiore a 60 *bpm*, in D3 il numero di brani con tempo compreso tra 60 *bpm* e 90 *bpm*, e così via. Aggiungeremo infine la colonna E per calcolare le frequenze percentuali come in precedenza.



	A	B	C	D	E	F	G	H
1	Tempo	x_min	x_max	n_j	p_j			
2	molto lento	0	60	=FREQUENZA(spotify!$Q$2:$Q$232726;C2:C5)				
3	lento	60	90	47.966	20,611%			
4	normale	90	140	128.338	55,146%			
5	veloce	140	190	52.448	22,536%			
6	molto veloce	maggiore di 190		2.584	1,110%			
7	n			232.725	100,000%			

Figura 3: *Come costruire un raggruppamento in classi in Excel.*

1.1.1 Rappresentazione grafica

La rappresentazione grafica è utile sia per sintetizzare i dati, che per comunicare gli eventuali risultati di un'analisi statistica. Consideriamo di seguito due tipi di grafici (ce ne sono altri) a seconda del tipo di carattere:

- **Grafico a barre** (per caratteri qualitativi): consiste nel disegnare un rettangolo per ogni modalità, dove l'altezza dei rettangoli (sull'asse delle ordinate) rappresenta la frequenza (assoluta, relativa o percentuale). I rettangoli sono staccati tra di loro ed hanno la stessa base, e sull'asse delle ascisse si riportano le modalità del carattere. Un grafico a barre rappresenta in forma grafica la tabella di frequenza e aiuta a cogliere in modo visuale i dati delle frequenze. In Figura 4, a titolo esemplificativo, mostriamo il grafico a barre della variabile *Metrica*: a sinistra la tabella di frequenza e a destra la sua rappresentazione grafica. Notiamo quanto spicchi la metrica in quattro quarti rispetto a tutte le altre: 86,26% risalta di più *guardato* sul grafico che non *letto* come numero
- **Istogramma** (per caratteri quantitativi suddivisi in classi): consiste nel disegnare un rettangolo per ogni classe, ma attaccati tra di loro. L'area di ogni rettangolo è proporzionale alla frequenza di ogni classe, mentre la base di ogni rettangolo corrisponde all'ampiezza della classe. In Figura 5 possiamo osservare l'istogramma dell'intensità sonora totale dei brani. La variabile *Loudness* (intensità sonora) misura il volume medio di una traccia in decibel (dB), con valori generalmente compresi tra –52.5 dB e 3.7 dB nel dataset. 0 dB rappresenta il massimo livello di volume senza distorsione, e i valori negativi indicano un volume inferiore a questo limite. Più il valore è negativo, più la traccia è silenziosa. Un valore positivo, come 3.7 dB, è insolito e potrebbe indicare un errore o distorsione nel brano. Notiamo una forma asimmetrica con coda lunga a sinistra e un picco intorno ai –6 dB

Per essere utili, i grafici devono essere chiari e semplici: ricorda di includere titolo ed etichette!

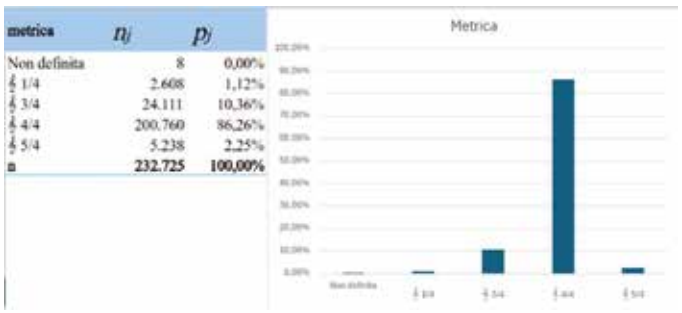


Figura 4: *Tabella di frequenza della variabile* Metrica *e relativo grafico a barre delle frequenze percentuali.*

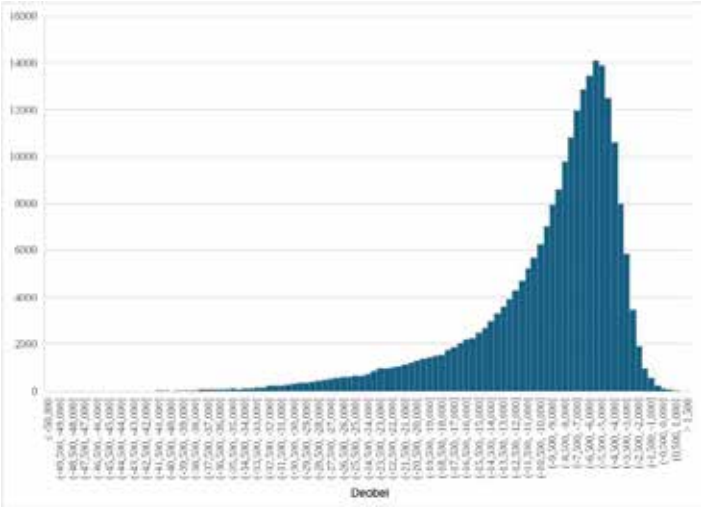


Figura 5: *Istogramma di densità della variabile* intensità sonora (Loudness).

HOW TO?

COSTRUIRE UN ISTOGRAMMA DI FREQUENZA IN EXCEL

Nelle versioni più recenti di Excel è stato aggiunto il grafico denominato istogramma statistico. Excel farà una divisione in classi di uguale ampiezza e rappresenterà l'istogramma delle frequenze assolute. Anche se l'opzione per disegnare l'istogramma su classi di ampiezza diverse, l'ampiezza delle classi può essere decisa dall'utente. In Figura 6 viene rappresentato il percorso.

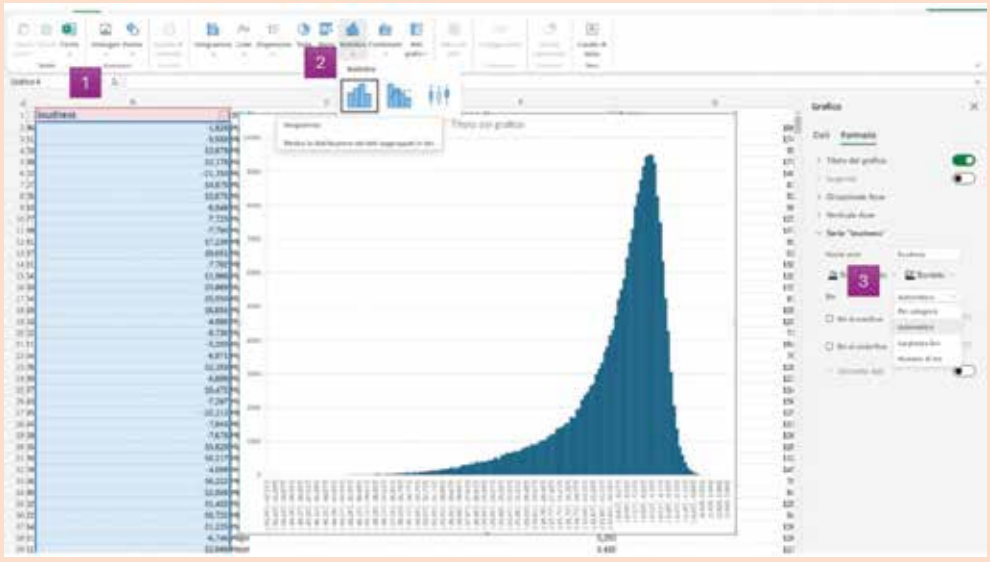


Figura 6: *Come inserire un istogramma in Excel.*

1.1.2 Sintesi numerica

Possiamo calcolare indici numerici che rappresentano alcuni aspetti essenziali della distribuzione di un carattere.

Indici di posizione: descrivono l'ordine di grandezza del carattere, i.e. i valori centrali attorno ai quali si concentrano i dati. I principali indici di posizione sono tre: **media aritmetica**, **mediana** e **moda**.

- **Media aritmetica** ($\bar{x}$): è la somma degli n valori divisa per il numero di osservazioni:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Nota che la media è calcolabile solo per caratteri quantitativi

- **Mediana:** è il valore centrale nella successione ordinata (in ordine crescente) delle n osservazioni individuali, i.e., bipartisce l'insieme delle osservazioni in due gruppi di uguale numerosità. Nota che la mediana presuppone che il carattere sia ordinabile, per tanto è calcolabile per caratteri quantitativi e qualitativi ordinati. Per calcolarla, basta trovare l'osservazione che si trova in posizione:

  - $\frac{n+1}{2}$ se n è dispari
  - $\frac{n}{2}$ e $\frac{n}{2}+1$ se n è pari (in caso di carattere quantitativo si calcola la media tra queste due osservazioni)

- **Moda:** è la modalità che presenta la frequenza maggiore. Può essere calcolata per qualsiasi tipo di carattere

Questi tre indici di posizione hanno la stessa unità di misura del carattere. La media è rappresentativa solo se la maggior parte dei valori sono vicini alla media; infatti, è fortemente influenzata dai valori estremi (mentre che la mediana è robusta rispetto ai valori estremi).
A titolo esemplificativo osserviamo come la variabile *Loudness* (l'intensità sonora), che abbiamo visto distribuita in Figura 5, ha media $\bar{x} = -9{,}570$, la mediana è pari a $-7{,}762$, mentre la moda è intorno a -6. La media è "abbassata" dai valori molto bassi evidenziati dalla coda lunga a sinistra, la mediana taglia la distribuzione esattamente a metà: metà brani più lenti di $-7{,}762$, e metà più veloci.

HOW TO? **CALCOLARE MEDIA, MEDIANA E VARIANZA SU EXCEL**

Le funzioni MEDIA e MEDIANA sono già implementate, Excel ignora in automatico le intestazioni e calcola queste statistiche anche su un'intera colonna. La funzione VAR.P funziona in maniera del tutto analoga e calcola la varianza di popolazione (da qui il suffisso.P) da non confondere con la varianza campionaria.

Per calcolare la media, la mediana e la varianza della colonna N *(Loudness)* del foglio coi dati Spotify, scriveremo:

=MEDIA(spotify!N:N) =MEDIANA(spotify!N:N) =VAR.P(spotify!N:N)

In tre celle libere.

Indici di dispersione: misurano la variabilità, ovvero la tendenza delle osservazioni ad assumere valori diversi. L'indice più comune per misurare la variabilità è la **varianza**, che può essere calcolata solamente per caratteri quantitativi.

- **Varianza** (σ^2): è la media dei quadrati degli scarti dalla media aritmetica

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

L'unità di misura della varianza è l'unità di misura del carattere al quadrato (e.g., se l'altezza di un individuo è misurata in *cm*, l'unità di misura della varianza sarà cm^2). Per questo si preferisce la deviazione standard σ, calcolata come la radice quadrata della varianza e con la stessa unità di misura del carattere. Attenzione: la varianza (e quindi anche la deviazione standard) può solo assumere valori positivi! La varianza è zero se tutte le osservazioni hanno lo stesso valore.

1.2 Statistica descrittiva bivariata

Per una coppia di caratteri (X,Y), la collezione delle coppie di modalità osservate per ogni singola unità statistica $\{(x_1,y_1),\ldots(x_n,y_n)\}$ viene chiamata **distribuzione doppia unitaria**. Per sintetizzare la distribuzione congiunta dei due caratteri (X,Y) possiamo costruire la **distribuzione doppia di frequenze** (o **tabella di frequenze a doppia entrata**). Questa consiste nell'identificare le diverse H modalità del carattere X e le K modalità del carattere Y, per poi contare quante unità presentano in contemporanea ogni combinazione possibile di modalità. Come nel caso univariato, per i caratteri quantitativi spesso viene usata la suddivisione in classi. Ad ogni incrocio della tabella abbiamo le **frequenze congiunte (assolute)** n_{ij}, $i=1,\ldots,H$, $j=1,\ldots,K$, che possiamo trasformare in frequenze congiunte relative ($f_{ij} = n_{ij}/n$) e percentuali ($p_{ij} = f_{ij} \times 100$). Ricorda che la somma di tutte le frequenze congiunte è uguale a n.

Osserviamo in Tabella 3, per esempio, la distribuzione bivariata dei caratteri *Genere* e *Metrica*. Leggiamo che i brani *Pop e Commerciali* in 4/4 sono $n_{11} = 43\,680$ (prima riga e prima colonna), quelli *Pop e Commerciali* in 3/4 $n_{12} = 5\,546$ (prima riga e seconda colonna), i brani *Rock e alternative*

in 4/4 $n_{21} = 42809$, i brani Jazz e tradizionali non definiti $n_{65} = 0$ (sesta riga e quinta colonna). L'ultima riga e l'ultima colonna rappresentato i totali di colonna e di riga, rispettivamente; ad esempio, $n_{.1} = 200760$, $n_{.2} = 24111$, mentre $n_{1.} = 51451$ e $n_{.2} = 46251$, e così via.

	Metrica					
Genere	4/4	3/4	5/4	1/4	Non def.	Totale complessivo
Pop e commerciali	43680	5546	1514	708	3	51451
Rock e alternative	42809	2825	396	221		46251
Rap e urban	34038	1736	672	161	1	36608
Elettronica e moderni	32951	2500	612	235	1	36299
Classico e teatrale	23075	8981	1816	1113	3	34988
Jazz e tradizionali	24207	2523	228	170		27128
Totale complessivo	**200760**	**24111**	**5238**	**2608**	**8**	**232725**

Tabella 3: *Distribuzione bivariata di* Genere *e* Metrica.

1.2.1 Studio dell'associazione tra due caratteri

A partire della tabella di frequenze a doppia entrata, possiamo calcolare l'indicatore simmetrico di associazione **Chi-quadrato di Pearson** X^2. Si chiama simmetrico perché misura la interdipendenza fra due caratteri, ovvero, assume che i due caratteri abbiano lo stesso ruolo (non implica l'esistenza di una causa ed un effetto). Per calcolarlo, abbiamo bisogno di due quantità:

- Le **frequenze teoriche** $E_{ij} = \frac{r_i \times c_j}{n}$, che sono le frequenze attese nel caso di indipendenza tra i due caratteri. I simboli r_i e c_j indicano la somma delle frequenze per ogni riga i (cioè il totale della riga i) ed ogni colonna j (totale colonna j) nella tabella a doppia entrata, rispettivamente
- Le **contingenze** $O_{ij} - E_{ij}$, ovvero, le differenze fra le frequenze osservate O_{ij} e quelle attese nel caso di indipendenza E_{ij}
- L'indice Chi-quadrato di Pearson X^2 si calcola come:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

L'indice $X^2 = 0$ se i due caratteri sono indipendenti (nota che, in questo caso, tutte le contingenze $O_{ij} - E_{ij} = 0$), $X^2 > 0$ e se i caratteri sono associati. Torneremo a parlare dell'indice Chi-quadrato nel prossimo capitolo, alla Sezione 2.5.

Come misura descrittiva dell'associazione può essere utilizzato l'indice V di Cramer, una versione normalizzata del X^2:

$$V = \sqrt{\frac{x^2}{n \cdot \min((H-1, K-1))}}$$

L'indice V varia tra zero ed uno: zero indica indipendenza tra i due caratteri, viceversa $V = 1$ indica perfetta dipendenza.

HOW TO? **COSTRUIRE DISTRIBUZIONI BIVARIATE E CALCOLARE IL CHI-QUADRO**

Excel permette di calcolare le distribuzioni bivariate con il comando *inserisci tabella pivot*. Per creare la distribuzione di *time_signature (Metrica)* e *Genere*, abbiamo copiato le due colonne su un foglio nuovo, le abbiamo selezionate (Figura 7a), abbiamo scelto la scheda inserisci, scelto *Tabella Pivot* e quindi *Da tabella/intervallo*, compariranno le opzioni di inserimento (Figura 7b).

Costruire la misura del Chi-quadro è un po' più brigosa e richiede alcuni passaggi aggiuntivi. Costruiremo dapprima la tabella delle frequenze teoriche n_{ij} (Figura 8): in cella H4 scriveremo

=B$10*$G4/$G$10

Dove B$10 è il totale della colonna *4/4*, $G4 è il totale della riga *Pop e commerciali*, e $G$10 è il totale dei brani. Trascineremo quindi la cella H4 verso destra e verso il basso per ricostruire tutti i 5 x 6 = 30 valori teorici; ricordiamo che il simbolo del $ serve per bloccare la riga o la colonna quando le formule vengono trascinate nelle celle adiacenti.

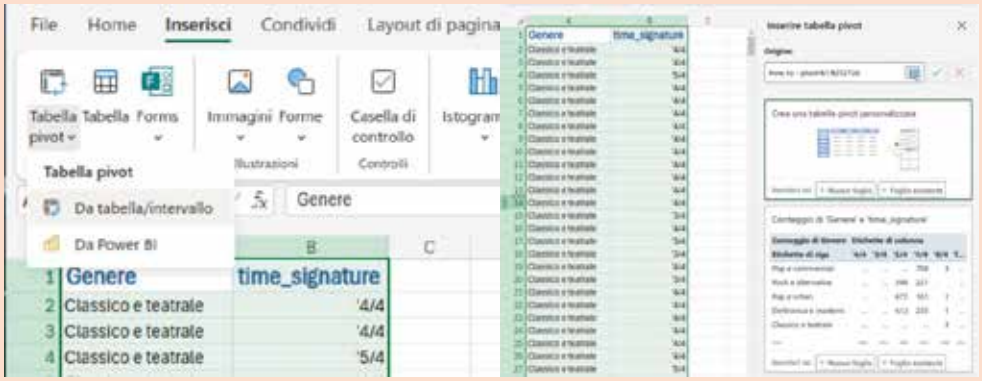


Figura 7a (sinistra): *Come inserire la tabella pivot.*
Figura 7b (destra): *Le opzioni di inserimento.*

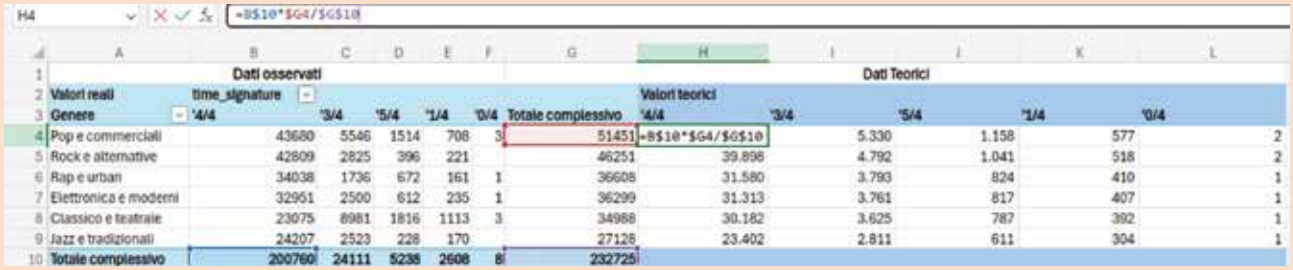


Figura 8: *Costruzione della tavola delle frequenze teoriche.*

Una volta costruita la tabella dei valori teorici procederemo a misurare la distanza, nella metrica del chi-quadro, dei valori osservati da quelli teorici. Per fare questo costruiremo un'altra tabella 6x5 che contiene tutte le distanze dai valori teorici a quelli osservati. In Figura 9 osserviamo la costruzione delle distanze: nella cella H4 inseriamo

=(B4-H4)^2/H4

dove B4 è il valore osservato n_{11} e H4 è il valore teorico n'_{11}. Trascineremo quindi la cella a destra per 5 colonne e in basso per 6 righe. Abbiamo colorato, a puro titolo illustrativo, le distanze: vanno dal verde quando sono zero al rosso quando tendono a numeri grandi. È interessante notare come la musica classica e teatrale siano molto più spesso in tre quarti degli altri generi, questo è colto dalla distanza del chi-quadro in cella N8. La somma di questi 30 valori fornisce il calcolo del chi-quadro (Figura 10); una volta noto il chi-quadro la V di Cramer è calcolata nella cella Q11

=RADQ(Q10/(G10*4))

La radice quadrata del chi-quadro (cella Q16) diviso n (G10) moltiplicato 4: il minimo tra 6–1 e 5–1.

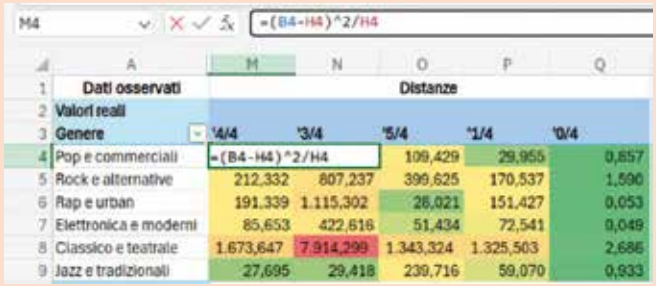


Figura 9: *Tabella delle distanze - ogni cella è ottenuta sottraendo il dato osservato dal dato teorico, elevato al quadrato e diviso per il dato teorico*

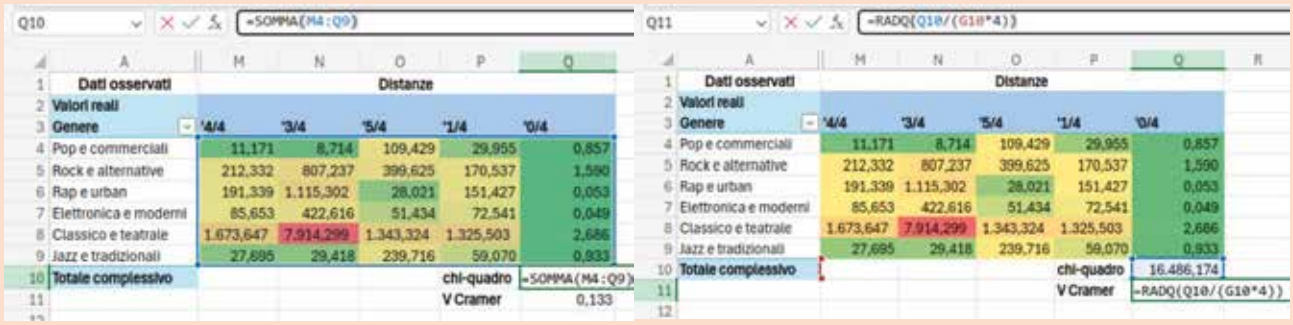
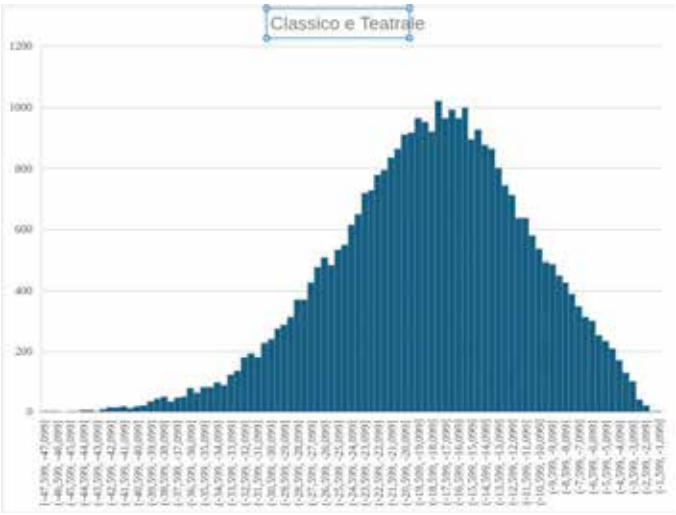


Figura 10 (sotto): *Calcolo del chi-quadro e della V di Cramer.*

1.2.2 Studio dell'associazione tra un carattere qualitativo ed uno quantitativo

I caratteri qualitativi possono creare gruppi all'interno dei quali una variabile quantitativa può assumere distribuzioni diverse. Per esempio, potremmo essere interessati a cogliere la relazione tra *Genere* e *Loudness*. Potremmo per esempio osservare gli istogrammi del *Loudness* al variare del *Genere*. In Figura 11 osserviamo per esempio che l'intensità sonora è molto più bassa nella musica classica e teatrale che nel Rock e alternative, dove tutta la distribuzione è spostata verso volumi più alti.

Possiamo anche sintetizzare le distribuzioni con indici sintetici al variare del gruppo: in Tabella 4 osserviamo come variano media e mediana al variare del gruppo.



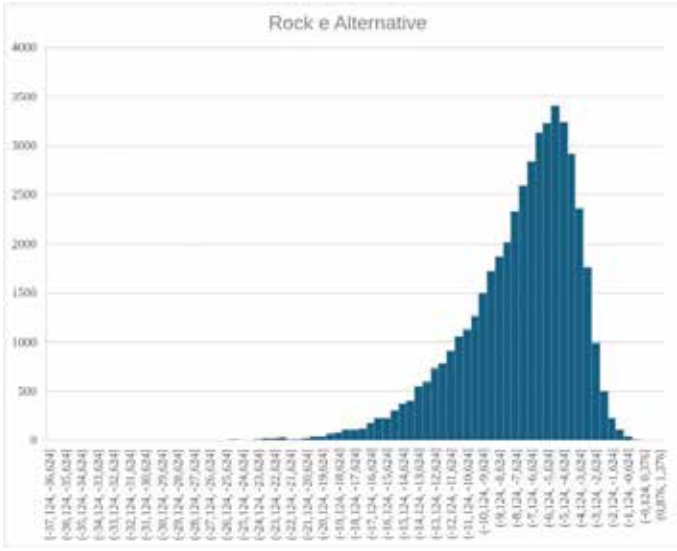


Figura 11: *Istogramma del* Loudness *per due generi diversi.*

Macro Genere	Intensità Sonora Media	Intensità Sonora Mediana
Classico e teatrale	-18,482	-18,067
Jazz e tradizionali	-9,258	-8,454
Elettronica e moderni	-8,311	-7,132
Pop e commerciali	-7,837	-6,680
Rock e alternative	-7,573	-6,838
Rap e urban	-7,491	-7,050
Tutti	**-9,570**	**-7,762**

Tabella 4: *Media e mediana dell'intensità sonora al variare del genere.*

HOW TO?

COSTRUIRE STATISTICHE AL VARIARE DI UNA VARIABILE QUALITATIVA

La funzione FILTRO consente di selezionare le righe del database al variare di un criterio, per esempio potremmo filtrare la colonna *Loudness* al variare delle modalità della variabile *Genere*. In Figura 12 possiamo osservare come procedere. Abbiamo copiato le colonne *Loudness* e *Genere* in un altro foglio e abbiamo copiato le sei etichette di genere nelle celle E1, F1, G1,... Nella cella E2 abbiamo scritto

=FILTRO($A:$A;$B:$B=E1)

Il resto della colonna conterrà solo i valori della colonna A a cui corrisponde la colonna B pari ad E1. Trascineremo quindi la cella E2 a destra fino alla cella J2. A questo punto avremo 6 diverse serie di dati su *Loudness*, una per ogni genere. Potremo operare grafici e statistiche per ogni colonna distintamente.

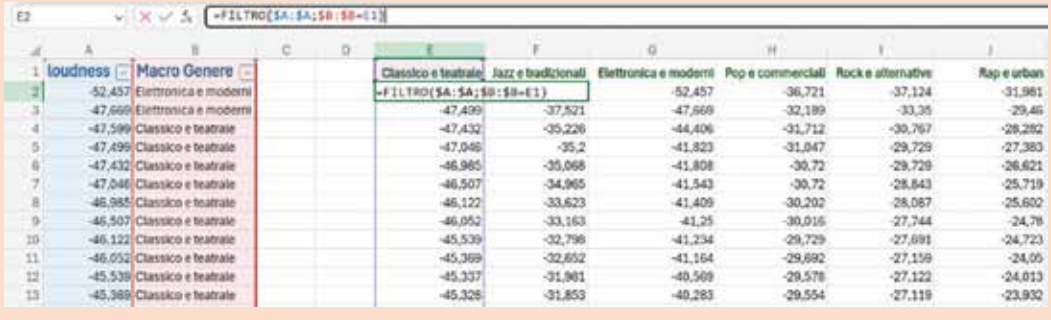


Figura 12: *Come usare la funzione filtro per selezionare serie di dati.*

1.2.3 Studio dell'associazione tra due caratteri quantitativi

La relazione tra due caratteri quantitativi X e Y si può rappresentare graficamente usando lo **scatterplot** o **grafico di dispersione**. In questo grafico, ogni coppia di valori (x_i, y_i) viene rappresentata da un punto nel piano cartesiano, i cui assi corrispondono ai due caratteri. Nel nostro caso, ci saranno tanti punti n come canzoni nel dataset. Dallo scatterplot possiamo capire se c'è una qualche relazione fra i caratteri, ed il tipo di relazione (lineare o no, positiva o negativa).

Osserviamo in Figura 13 lo scatterplot (in Excel lo scatterplot è chiamato grafico di dispersione) tra la variabile *Energy* (in ascissa) e la variabile *Loudness* (in ordinata). Possiamo osservare che all'aumentare dell'energia del brano cresce l'intensità acustica, ma la relazione non è lineare nelle estremità.

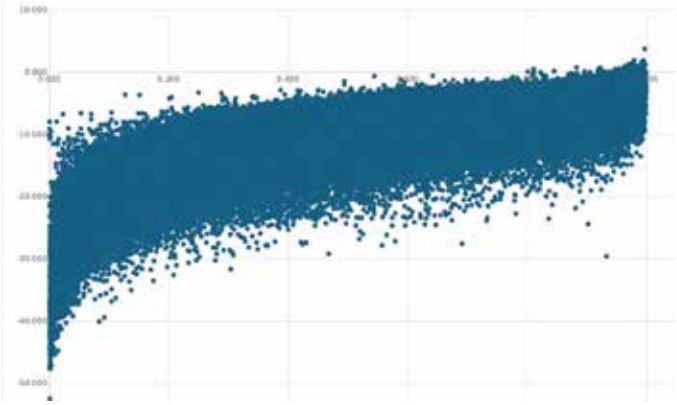


Figura 13: *Scatterplot dei caratteri* Energy *e* Loudness.

Un indice (simmetrico) per misurare la relazione **lineare** tra due caratteri è la **covarianza** σ_{XY}, calcolato come la media dei prodotti degli scostamenti delle variabili X e Y dalle rispettive medie:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Diversamente alla varianza, la covarianza può assumere sia valori positivi che negativi. In particolare:

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq \sigma_X \sigma_Y$$

Attenzione! Se i due caratteri sono indipendenti, allora $\sigma_{XY} = 0$, ma il contrario non è vero, i.e. se $\sigma_{XY} = 0$, potrebbe esserci una relazione di tipo non lineare.

Il valore della covarianza σ_{XY} è difficile da interpretare aldilà del fatto che, se è positiva, allora la relazione tra le due variabili è positiva (e negativa se la covarianza è negativa). Il coefficiente di correlazione lineare di Pearson ρ_{XY} non è altro che una versione normalizzata della covarianza:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Adesso abbiamo un indice che assume valori in un intervallo ben definito $-1 \leq \rho_{XY} \leq 1$, dove:

- $\rho_{XY} = -1$ indica correlazione negativa perfetta
- $\rho_{XY} = 0$ indica assenza di correlazione lineare (non necessariamente indipendenza)
- $\rho_{XY} = 1$ indica correlazione positiva perfetta

Torneremo nuovamente sul coefficiente di correlazione lineare nel Capitolo 3 (Sezione 3.1). Discuteremo infatti della sua relazione con il coefficiente di determinazione lineare quale indice per valutare la bontà di adattamento del modello di regressione ai dati.

HOW TO? DISEGNARE UN DIAGRAMMA DI DISPERSIONE E CALCOLARE LA CORRELAZIONE

Per ottenere il grafico in Figura 13 creeremo un nuovo foglio ed andremo ad incollare le colonne *Energy* e *Loudness*, le andremo a selezionare e quindi inseriremo un grafico a dispersione come indicato in Figura 14.

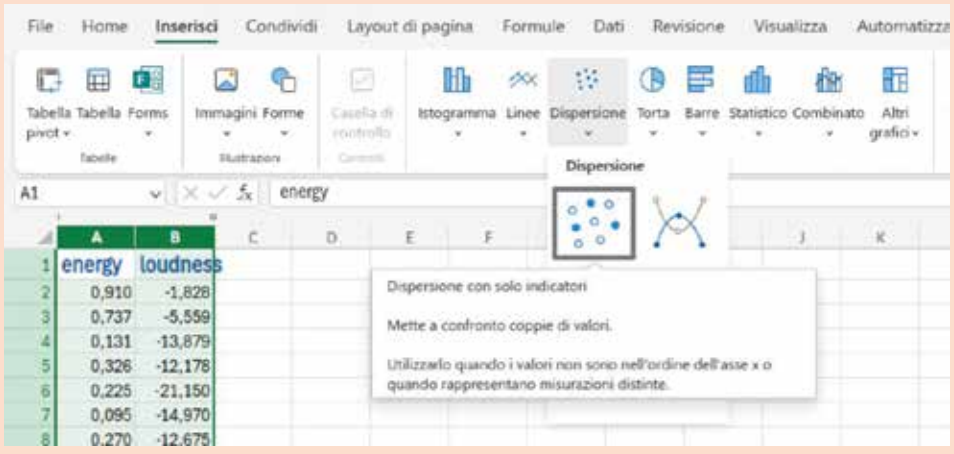


Figura 14: *Inserimento di uno scatterplot.*

Per calcolare la correlazione useremo la funzione CORRELAZIONE implementata in Excel

=CORRELAZIONE(A:A;B:B)

come indicato in Figura 15.

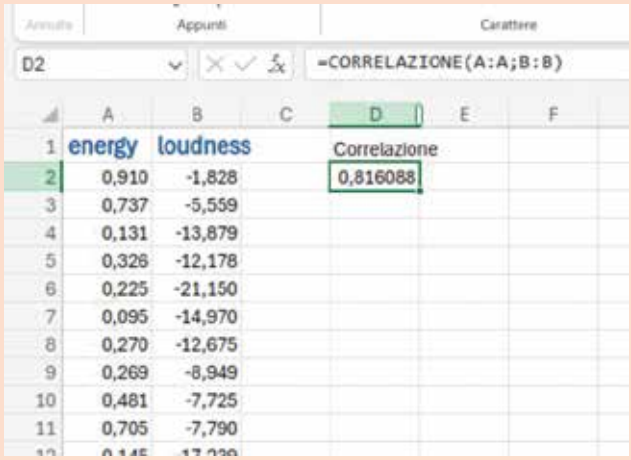
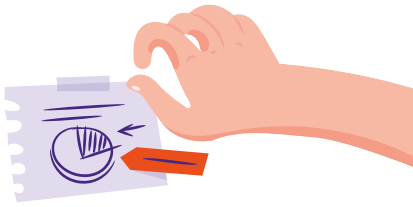


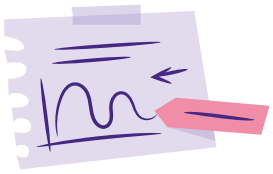
Figura 15: *Nella cella D2 inseriamo il risultato della correlazione tra* Energy *e* Loudness.

Nel prossimo capitolo analizzeremo i fondamenti teorici e pratici dell'inferenza statistica, concentrandoci sui test statistici più utilizzati nella ricerca empirica.

appunti



appunti



appunti

Capitolo[2]

Inferenza e test statistici

La statistica inferenziale è uno dei pilastri fondamentali dell'analisi quantitativa moderna, fornendo strumenti indispensabili per trarre conclusioni significative su intere popolazioni a partire dall'osservazione di campioni limitati. In un'epoca caratterizzata da una crescente disponibilità di dati, la capacità di effettuare inferenze accurate e verificare ipotesi statistiche è diventata una competenza essenziale in numerosi settori, dalla ricerca scientifica al processo decisionale aziendale.
Il processo inferenziale si basa su un principio tanto semplice quanto potente: attraverso l'analisi di un campione rappresentativo, possiamo estendere le nostre conclusioni all'intera popolazione di riferimento, quantificando al contempo il grado di incertezza associato a tali conclusioni. Questo approccio è particolarmente prezioso quando l'analisi dell'intera popolazione risulterebbe impraticabile o economicamente non sostenibile.
L'inferenza statistica si divide in due grandi aree metodologiche: la **stima dei parametri** e la **verifica delle ipotesi**. La stima dei parametri permette di determinare, con un certo grado di precisione, le caratteristiche della popolazione (come media, varianza, proporzioni) partendo dai dati campionari. La verifica delle ipotesi, d'altra parte, offre un framework rigoroso per valutare affermazioni sulla popolazione, permettendo di discriminare tra variazioni casuali e differenze statisticamente significative.

		Variabile dipendente	
		Categorica	**Continua**
Variabile indipendente	**Categorica**	Chi-quadrato & misure di associazione	2 Gruppi: T-test >2 Gruppi: ANOVA
	Continua	Regressione logistica	Correlazione Regressione

Tabella 5: *Tipologie di test statistici.*

Come illustrato nella Tabella 5, la scelta del test statistico dipende dalla natura delle variabili a disposizione (ad es., continue o categoriche). In questo Capitolo analizzeremo i fondamenti teorici e pratici dell'inferenza statistica, concentrandoci sui test statistici più utilizzati nella ricerca empirica: il T-test, l'ANOVA e il test del Chi-quadrato.

2.1 L'inferenza statistica: dal campione alla popolazione

L'inferenza statistica rappresenta uno strumento fondamentale per lo studio dei fenomeni sociali, economici e demografici, poiché permette ai ricercatori di trarre conclusioni valide su intere popolazioni attraverso l'analisi di campioni rappresentativi. Questa metodologia è particolarmente utile quando si affronta lo studio di popolazioni numerose, per le quali sarebbe impraticabile o impossibile raccogliere dati su tutti i soggetti.
Il principio fondamentale dell'inferenza statistica si basa sulla capacità di estendere (o "inferire") le conclusioni ottenute dall'analisi di un campione all'intera popolazione di riferimento, seguendo rigorosi criteri statistici che garantiscono l'affidabilità dei risultati. Questo processo consente di ottimizzare le risorse della ricerca, mantenendo al contempo un elevato livello di accuratezza nelle conclusioni.

2.2 La verifica delle ipotesi

Nell'ambito dell'inferenza statistica, la verifica delle ipotesi è il metodo che permette di trarre conclusioni sul fenomeno oggetto di studio. Questo processo si articola in quattro fasi fondamentali:

1. Formulazione delle ipotesi

Il processo inizia con la definizione di due ipotesi contrapposte:

- L'ipotesi nulla (H_0): rappresenta l'affermazione di nessun effetto o nessuna differenza. Nei test sulla media di un campione l'ipotesi nulla è definita come $H_0 : \mu = \mu_0$ (la media della popolazione μ è uguale alla media ipotizzata μ_0).
- L'ipotesi alternativa (H_1): rappresenta l'affermazione che si contrappone all'ipotesi nulla. A seconda della direzione dell'ipotesi alternativa, possiamo definire tre tipi di test:
  - Test bilaterale (a due code) → $H_1 : \mu \neq \mu_0$
  - Test unilaterale (a una coda) destro → $H_1 : \mu > \mu_0$
  - Test unilaterale (a una coda) sinistro → $H_1 : < \mu_0$

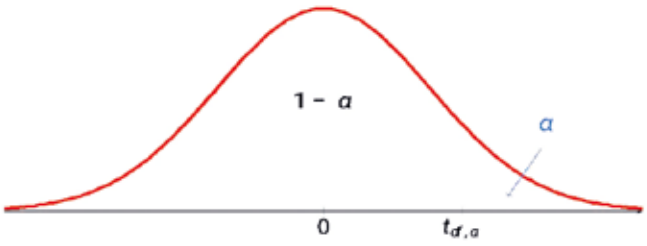


Figura 16: *Test unilaterale destro T di Student.*

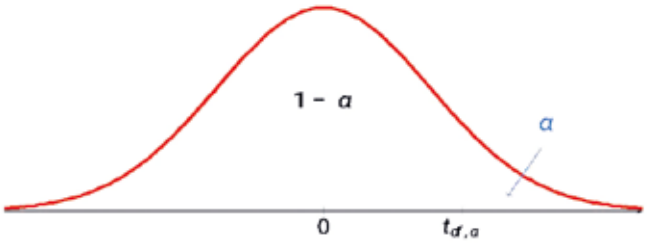


Figura 17: *Test bilaterale T di Student.*

2. Definizione del criterio decisionale

La scelta del livello di significatività (α) determina il rigore del test statistico.
Questo valore, tipicamente fissato al 5% o all'1%, rappresenta la probabilità di commettere un errore di Tipo I (falso positivo), ovvero di rifiutare un'ipotesi nulla quando vera. Il livello di confidenza, complementare al livello di significatività ($1 - \alpha$), indica l'affidabilità delle nostre conclusioni.

3. Calcolo della statistica test

La scelta della statistica test appropriata dipende dalle caratteristiche del campione e dalle informazioni disponibili sulla popolazione:

Z-test: utilizzato quando la dimensione del campione è grande ($n > 30$) e la varianza della popolazione è nota.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = media campionaria
μ_0 = media ipotizzata della popolazione
σ = deviazione standard della popolazione
n = dimensione campionaria
Il fattore $(\sigma / \sqrt{n})$ rappresenta l'*errore standard*

- **T-test:** utilizzato quando la dimensione del campione è piccola ($n < 30$) e la varianza della popolazione è ignota.

$$t_{df} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Gradi di libertà[1] (*degrees of freedom*): $df = n - 1$

4. Processo decisionale

Il confronto tra la statistica test calcolata e i valori critici corrispondenti al livello di significatività scelto porta alla decisione finale:

- Se |statistica test| > valore critico → Si rifiuta H_0
- Se |statistica test| < valore critico → Non si rifiuta H_0

Una coda	0.25	0.2	0.15	0.10	0.05	0.025	0.021	0.005	0.0005
Due code	1	0.50	0.40	0.40	0.20	0.10	0.05	0.02	0.01
df									
1	1.000	1. 376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.817	1. 061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.979	1.250	1.638	2.353	3.182	4.541	5.841	12.924
...	...	...	...	...	...	...	...	...	...
10	0.700	0.879	1.093	1.372	1.813	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.696	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
...	...	...	...	...	...	...	...	...	...

Tabella 6: *Valori critici distribuzione* T di Student.

Per individuare il valore critico sulle tavole statistiche occorre sapere i gradi di libertà ($df = n - 1$), il livello di significatività (α), ed il tipo di test (unilaterale o bilaterale). In caso di test bilaterale, bisognerà cercare sulle tavole il valore critico corrispondente ad (α/2). Ad esempio, considerando un test bilaterale con 10 gradi di libertà ($df = 10$) e significatività del 5% ($\alpha = 0.05$), il valore critico sarà pari a 2.228 (Tabella 6). L'ipotesi nulla (H_0) viene accettata quando la statistica test è minore di tale valore, poiché ricade nella regione di accettazione (Figura 16).

1 I gradi di libertà rappresentano il numero di osservazioni che possono variare liberamente in un'analisi statistica. Quando si calcola la statistica da un campione si perde un grado di libertà per ogni parametro stimato.

Il metodo *p-value*. Il *p-value* (o livello di significatività osservato) offre un metodo alternativo e più flessibile per il processo decisionale. Rappresenta la probabilità di osservare un valore della statistica test uguale o più estremo di quello calcolato, assumendo che l'ipotesi nulla sia vera. Trattandosi di una probabilità, il *p-value* è sempre compreso tra 0 e 1. Quanto più è piccolo, tanto più forte è l'evidenza contro l'ipotesi nulla. Viene tipicamente calcolato automaticamente dai software statistici.

- Se *p-value* < α → Si rifiuta H_0
- Se *p-value* > α → Non si rifiuta H_0

HOW TO? T-TEST SU EXCEL

Per condurre un T-test su un campione, abbiamo bisogno di alcune statistiche descrittive relative al campione che possiamo calcolare utilizzando il comando:
Dati - Analisi dei dati

Nella nuova finestra selezioniamo **"Statistica descrittiva"** per calcolare i principali indici statistici per la variabile che vogliamo testare (*Tempo*).

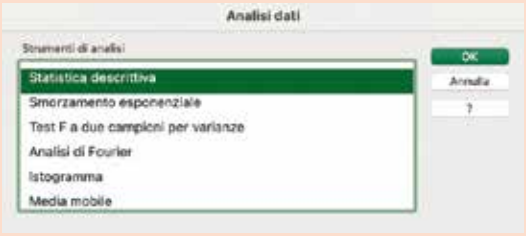


Figura 18: *Opzioni in "Analisi dei dati".*

Selezioniamo quindi l'intervallo della nostra variabile d'interesse e spuntiamo la casella "**Riepilogo statistiche".**

Figura 19: *Statistiche descrittive in "Analisi dei dati".*

tempo	
Media	117.67
Errore standard	0.06
Mediana	115.78
Moda	120.02
Deviazione standard	30.90
Varianza campionaria	954.74
Curtosi	-0.47
Asimmetria	0.40
Intervallo	212.52
Minimo	30.38
Massimo	242.90
Somma	27383955.97
Conteggio	232725.00

Una volta eseguito il comando otterremo una tabella riassuntiva delle statistiche descrittive per la variabile tempo.

Ora possiamo utilizzare la media ($\bar{x}$), l'errore standard (s) e il conteggio (n) per calcolare statistica test (t_{df}) e gradi di libertà (df).
Ipotizzando una media $\mu_0 = 115$, la statistica test sarà:

$$t_{232725} = \frac{117.67 - 115}{\frac{0.06}{\sqrt{232725}}} = 41.63 \text{ con } df = 232724$$

Figura 20: *Statistiche descrittive variabile tempo.*

In un test *bilaterale* le ipotesi da testare sono definite come:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Per un livello di significatività $\alpha = 0.05$, il valore critico può essere trovato attraverso la funzione:
=INV.T.2T(0.05; 232724)
In alternativa, possiamo calcolare il *p-value* associato alla statistica test utilizzando:
=DISTRIB.T.2T(stat.test; 232724)

tempo			
		Media ipotizzata	115.00
Media	117.67	Statistica T	41.63
Errore standard	0.06		
Mediana	115.78	**Verifica d'ipotesi**	
Moda	120.02	Valore critico	1.96
Deviazione standard	30.90	*p-value*	0.00
Varianza campionaria	954.74		
Curtosi	-0.47		
Asimmetria	0.40		
Intervallo	212.52		
Minimo	30.38		
Massimo	242.90		
Somma	27383955.97		
Conteggio	232725.00		

Figura 21. *Risultati T-test sulla media della variabile tempo.*

Siccome |statistica test| > valore critico, rifiutiamo H_0. Esiste quindi una differenza statisticamente significativa tra le due medie. Allo stesso modo, un *p-value* = 0.00 porta alla stessa conclusione.

2.3 Confronto tra due gruppi

Mentre i test su un campione singolo confrontano una statistica campionaria con un valore noto della popolazione, i test a due campioni permettono di valutare se esistono differenze statisticamente significative tra due gruppi distinti. Questa metodologia risulta particolarmente utile quando si vogliono confrontare due popolazioni o valutare l'efficacia di un trattamento rispetto a un gruppo di controllo. A tale proposito, nel confronto tra medie di due campioni le ipotesi saranno definite come:

- Ipotesi nulla → $H_0 : \mu_1 = \mu_2$ (le medie sono uguali)
- Ipotesi alternativa → $H_1 : \mu_1 \neq \mu_2$ (le medie sono statisticamente diverse)

2.3.1 Z-test a due campioni

Utilizzato quando:
- Le varianze delle popolazioni sono note
- I campioni sono sufficientemente grandi (n > 30) La statistica test è data da:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_1$ = media campione 1
$\bar{x}_2$ = media campione 2
μ_1 = media ipotizzata popolazione 1
μ_2 = media ipotizzata popolazione 2
σ_1^2 = varianza popolazione 1
σ_2^2 = varianza popolazione 2
n_1 = dimensione campione 1
n_2 = dimensione campione 2

2.3.2 T-test a due campioni

Impiegato quando le varianze delle popolazioni sono ignote e/o i campioni sono piccoli (n < 30). Si distinguono tre varianti:

a) T-test per campioni appaiati

Impiegato quando le osservazioni nei due gruppi sono naturalmente accoppiate o misurate sugli stessi soggetti in due momenti diversi.

$$t_{df} = \frac{\bar{x}_d}{\frac{s_d}{\sqrt{n}}}$$

dove $\bar{x}_d$ è la media delle differenze e s_d la loro deviazione standard.

b) T-test con varianze uguali

Il T-test a due campioni con varianza uguale è calcolato come segue:

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad df = n_1 + n_2 - 2$$

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

$\bar{x}_1$ = media campione 1
$\bar{x}_2$ = media campione 2
μ_1 = media ipotizzata popolazione 1
μ_2 = media ipotizzata popolazione 2
s_1^2 = varianza campione 1
s_2^2 = varianza campione 2
s_p^2 = varianza dei campioni accoppiati
n_1 = dimensione campione 1
n_2 = dimensione campione 2

c) T-test con varianze diverse (Welch's t-test)

Utilizzato quando non si può assumere l'omogeneità delle varianze.

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

I gradi di libertà sono dati da:

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

Per poter svolgere il T-test, occorre che la variabile dipendente sia continua e che i dati della variabile dipendente seguano una distribuzione normale all'interno di ciascun gruppo[22]. Inoltre, le osservazioni devono essere indipendenti l'una dall'altra.

2 Questa assunzione viene a meno per campioni con più di 30 unità per il *Teorema Centrale del Limite*. Secondo questo teorema, la distribuzione delle medie campionarie tende a seguire una distribuzione normale all'aumentare della dimensione del campione, tipicamente quando n > 30.

HOW TO?

T-TEST SU DUE CAMPIONI SU EXCEL

Supponiamo di voler testare se la media della variabile *Loudness* (volume) è statisticamente diversa a seconda della scala armonica (*Mode*) del brano.
Per prima cosa creiamo i nostri due campioni da testare: utilizzando la funzione **Ordina e filtra → Filtro**

Posizioniamoci ora sulla variabile *Mode* e filtriamo i valori in base ad una delle due categorie (es., "major"). Successivamente copieremo i valori filtrati in una nuova variabile *loudness_major*.

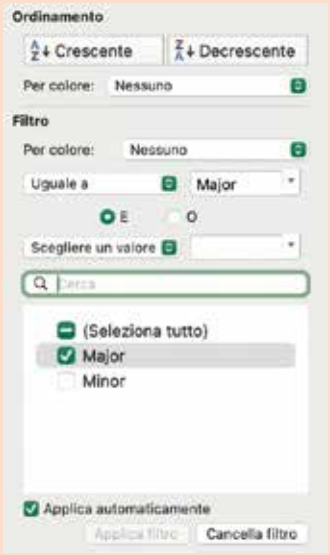


Figura 22: *Filtro applicato alla variabile mode.*

Ora che abbiamo definito i due campioni da testare, possiamo condurre il T-test su due campioni utilizzando il comando:
Dati → Analisi dei dati

A seconda delle assunzioni sulla varianza (uguale o diversa) possiamo scegliere il tipo di T-test. In questo caso sceglieremo il T-test su due campioni con **varianza uguale**.

Selezioniamo ora le due variabili appena create in "Intervallo variabile 1" e "Intervallo variabile 2", specificando una differenza ipotizzata per le medie pari a 0, e.

Figura 23: *T-test su due campioni in "Analisi dei dati".*

I risultati del T-test forniscono la statistica test, il valore critico ed il *p-value**.

Test t: due campioni assumendo uguale varianza		
	loudness_minor	*loudness_major*
Media	-9.35	-9.69
Varianza	35.69	36.09
Osservazioni	80981	151744.00
Varianza complessiva	35.95	
Differenza ipotizzata per le medie	0.00	
gdl	232723	
Stat t	13.08	
P(T<=t) una coda	0.00	
t critico una coda	1.64	
P(T<=t) due code	0.00	
t critico due code	1.96	

Figura 24: *Risultati T-test varianza uguale.*

La statistica test è stata calcolata come:

$$t_{232723} = \frac{-9.35 + 9.69}{35.95\left(\frac{1}{80981} + \frac{1}{151744}\right)} = 13.08 \text{ con } df = 232723$$

In un *test bilaterale* le ipotesi da testare sono definite come:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Con un valore critico di 1.96 (e un *p-value*=0.00), rifiutiamo H_0, concludendo che le due medie differiscono in maniera statisticamente significativa.
Il test produce gli stessi risultati anche assumendo **varianze diverse**:

Test t: due campioni assumendo varianze diverse		
	loudness_minor	*loudness_major*
Media	-9.35	-9.69
Varianza	35.69	36.09
Osservazioni	80981	151744.00
Differenza ipotizzata per le medie	0.00	
gdl	166157	
Stat t	13.11	
P(T<=t) una coda	0.00	
t critico una coda	1.64	
P(T<=t) due code	0.00	
t critico due code	1.96	

Figura 25: *Risultati T-test varianza diversa.*

* Il p-value può anche essere calcolato utilizzando la funzione:
=TESTT(variabile1, variabile2, coda, tipo)
In questo caso dovremo specificare le due variabili (gruppi) da testare, il numero di code (coda: 1 – Unilaterale, 2 – Bilaterale) ed il tipo di T-test.

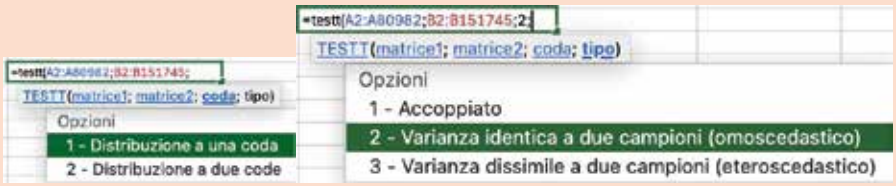


Figura 26: *Calcolo* **p-value** *con funzione* **TESTT().**

2.4 ANOVA: Analisi della varianza

L'ANOVA, acronimo di *Analysis of Variance*, è un test statistico che consente di confrontare simultaneamente le medie di tre o più gruppi, estendendo così il confronto oltre il semplice caso di due campioni. Questo metodo si basa sul confronto tra la variabilità all'interno dei gruppi e quella tra i gruppi, fornendo un approccio efficace per valutare se le differenze tra le medie siano statisticamente significative. In questo capitolo ci concentreremo sull'ANOVA ad una via (*One-way ANOVA*), ossia sul confronto delle medie di una variabile dipendente continua tra tre o più gruppi definiti da una sola variabile indipendente categorica.
Per eseguire l'ANOVA, è necessario rispettare alcuni requisiti fondamentali. Innanzitutto, la variabile dipendente deve essere continua, mentre la variabile indipendente deve essere categorica e dividere i dati in gruppi distinti da confrontare. Inoltre, le osservazioni devono essere indipendenti tra loro, cioè i dati di un gruppo non devono influenzare quelli di un altro. Un ulteriore requisito è l'omogeneità delle varianze: le varianze all'interno dei gruppi devono essere simili. Infine, si presuppone che la distribuzione dei dati all'interno dei gruppi sia normale e, idealmente, che i gruppi abbiano una numerosità simile per garantire un confronto equilibrato.

Statistica test F

L'ANOVA utilizza la statistica test F, che confronta la Varianza tra gruppi (*between*), ovvero la variabilità delle medie dei gruppi rispetto alla media generale, con la Varianza interna ai gruppi (*within*), che è la variabilità delle osservazioni all'interno di ciascun gruppo.

$$F = \frac{varianza\ tra\ gruppi}{varianza\ interna\ ai\ gruppi}$$

La statistica F utilizza due gradi di libertà (*df*), uno per la varianza tra i gruppi e uno per la varianza entro i gruppi. Questi si calcolano come segue:
- Gradi di libertà tra i gruppi: $df_1 = k - 1$, dove k è il numero di gruppi
- Gradi di libertà entro i gruppi: $df_2 = n - k$, dove n è il numero totale di osservazioni

La verifica delle ipotesi: ANOVA

Nell'ANOVA, le due ipotesi sono definite come:
- Ipotesi nulla → H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (tutte le medie sono uguali)
- Ipotesi alternativa → H_1: almeno una media è diversa dalle altre

Come per il T-test, la decisione finale sull'ipotesi nulla può essere presa confrontando la statistica test F con il valore critico sulle tavole della distribuzione F, oppure utilizzando il *p-value*. In caso di F-test statisticamente significativo, almeno una media di gruppo differisce dalle altre. Per determinare quali gruppi specifici differiscono tra loro, è necessario condurre analisi *post-hoc* appropriate, che non verranno trattate in questo contesto.

HOW TO? ANOVA SU EXCEL

Per indagare se l'energia (*Energy*) di un brano varia in maniera statisticamente significativa in base al genere, implementeremo il test ANOVA su Excel.
L'ANOVA richiede infatti una variabile dipendente continua (*Energy*) e una variabile indipendente categorica con più di due categorie (macro_genere).

Per prima cosa creiamo i nostri gruppi da testare: utilizzando la funzione **Ordina e filtra → Filtro** andremo a creare una variabile di energy per ciascuna categoria di macro_genere.

energy_classic	energy_electronic	energy_jazz	energy_pop	energy_rap	energy_rock
0.91	0.298	0.704	0.554	0.689	0.65
0.737	0.222	0.859	0.321	0.61	0.479
0.131	0.518	0.408	0.488	0.52	0.588
0.326	0.49	0.726	0.309	0.366	0.599
0.225	0.153	0.478	0.458	0.621	0.316
0.0948	0.0913	0.846	0.647	0.331	0.774
0.27	0.505	0.412	0.665	0.649	0.861
0.269	0.57	0.874	0.579	0.255	0.411
0.481	0.318	0.444	0.68	0.507	0.705
0.705	0.427	0.201	0.619	0.639	0.906
0.145	0.299	0.665	0.733	0.378	0.782
0.133	0.417	0.756	0.704	0.659	0.345
0.6	0.448	0.285	0.475	0.709	0.936
0.264	0.442	0.808	0.557	0.687	0.46
0.174	0.272	0.779	0.385	0.345	0.432
0.405	0.394	0.87	0.6	0.766	0.648
0.237	0.396	0.78	0.731	0.494	0.807
0.953	0.264	0.322	0.33	0.586	0.853
0.491	0.298	0.543	0.714	0.337	0.864

Figura 27: *Valori di* Energy *per ciascun macro_genere.*

Una volta costruita la nostra matrice di dati, possiamo implementare l'ANOVA utilizzando il comando:
Dati → Analisi dei dati → Analisi varianza: ad un fattore

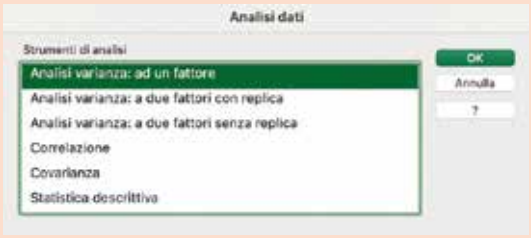


Figura 28a: *ANOVA ad un fattore in "Analisi dei dati".*

A questo punto dovremo definire l'intervallo dei dati selezionando tutta la matrice di dati appena creata.

Figura 28b: *ANOVA ad un fattore in "Analisi dei dati".*

Il risultato del test ANOVA riporta sia le statistiche descrittive di ciascun gruppo nella tabella di "RIEPILOGO", sia i risultati del test ANOVA.

Analisi varianza: ad un fattore

RIEPILOGO

Gruppi	*Conteggio*	*Somma*	*Media*	*Varianza*
energy_classic	34988	8071.44	0.23	0.04
energy_electronic	36299	23090.67	0.64	0.06
energy_jazz	27128	15446.02	0.57	0.06
energy_pop	51451	34085.09	0.66	0.04
energy_rap	36608	21898.50	0.60	0.03
energy_rock	46251	30284.41	0.65	0.06

ANALISI VARIANZA

Origine della variazione	*SQ*	*gdl*	*MQ*	*F*	*'o di significat*	*F crit*
Tra gruppi	4988.24	5	997.65	20794.87	0.00	2.21
In gruppi	11164.86	232719	0.05			
Totale	**16153.10**	**232724**				

Figura 29. *Risultati ANOVA*

In questo caso le ipotesi da verificare sono:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

H_1: almeno una media è diversa dalle altre

La statistica test F è stata calcolata come:

$$F = \frac{997.65}{0.05} = 20794.87 \text{ con } df_1 = 5 \text{ e } df_2 = 232719$$

In base al confronto tra statistica test F (20794.87) e valore critico (2.21), rifiuteremo l'ipotesi nulla (H_0). Questa conclusione è supportata anche dal *p-value* < 0.01.

2.5 Test del Chi-quadrato

Finora ci siamo concentrati su test statistici che richiedevano una variabile continua (numerica) e una categorica. Il test del Chi-quadrato è una tecnica statistica utilizzata per esaminare la relazione tra **due variabili categoriche**, organizzate in una *tabella di contingenza*.

		Variabile cat. 2			
		Categoria 1	**Categoria 2**	**Categoria 3**	**Totale riga**
Variabile cat. 1	**Gruppo A**	O_{11}	O_{12}	O_{13}	r_1
	Gruppo B	O_{21}	O_{22}	O_{23}	r_2
	Gruppo C	O_{31}	O_{32}	O_{33}	r_3
	Totale colonna	c_1	c_2	c_3	n

Tabella 7: *Tabella di contingenza delle frequenze osservate.*

Il test valuta se le frequenze osservate nelle varie categorie sono significativamente diverse dalle frequenze attese, assumendo che non vi sia alcuna associazione tra le variabili. Il test del Chi-quadrato è ampiamente utilizzato per verificare l'indipendenza tra due variabili categoriche o per testare la bontà dell'adattamento tra una distribuzione osservata e una distribuzione teorica.

HOW TO?

TABELLA DI CONTINGENZA SU EXCEL

Per costruire una tabella di contingenza su Excel possiamo usare il comando:
Inserisci → Tabella pivot

Figura 30: *Opzione tabelle pivot.*

Selezioniamo quindi l'intero dataset come input, in modo da poter selezionare successivamente le due variabili di interesse.

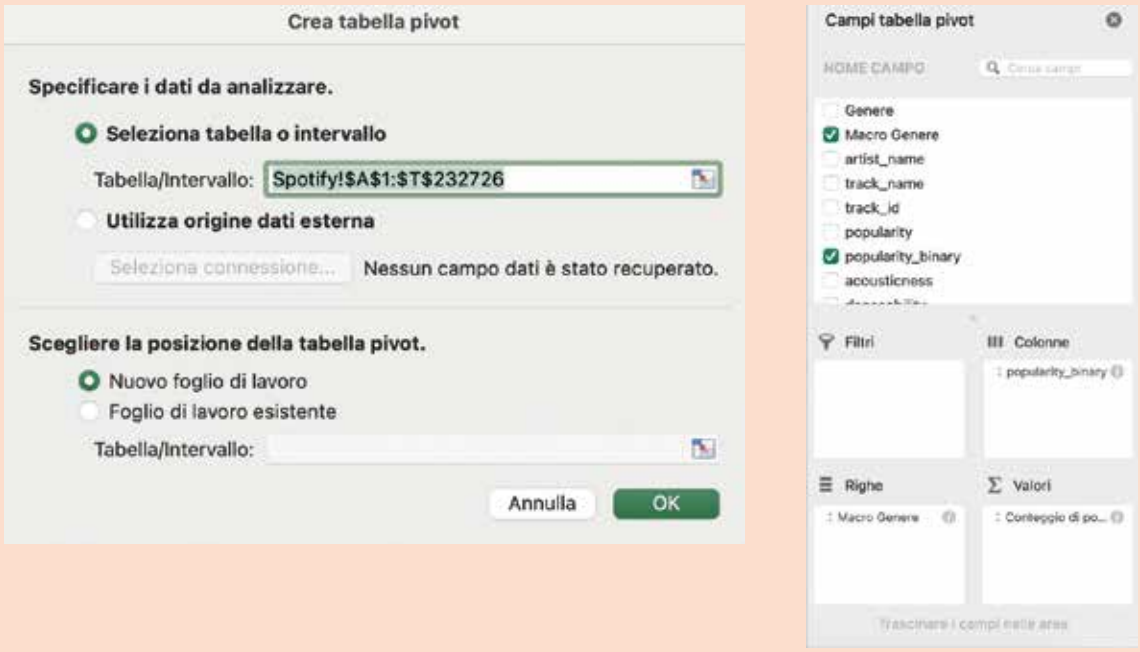


Figura 31: *Costruzione tabella pivot.*

Sulla destra apparirà la sezione **"Campi tabella pivot"**, nella quale possiamo selezionare le due variabili categoriche di interesse: macro_genere e popularity_binary.
Per costruire la tabella di contingenza trasciniamo la variabile continua nel riquadro **"Righe"** e la variabile categorica nel riquadro **"Colonne"**.
Successivamente, trasciniamo nel riquadro **"Valori"** la variabile categorica utilizzata per le colonne. All'interno dello stesso riquadro assicuriamoci di selezionare il conteggio dei valori facendo click destro sulla variabile → Impostazioni campo → Riepiloga per: **Conteggio**

Figura 32: *Impostazioni campo valori tabella pivot.*

Il risultato finale sarà la seguente tabella di contingenza:

Conteggio di popular	Etichette di col		
Etichette di riga	**0**	**1**	**Totale complessivo**
Classico e teatrale	34981	7	34988
Elettronica e moderni	36256	43	36299
Jazz e tradizionali	26986	142	27128
Pop e commerciali	47734	3717	51451
Rap e urban	34185	2423	36608
Rock e alternative	45024	1227	46251
Totale complessivo	**225166**	**7559**	**232725**

Figura 33: *Tabella pivot con macro_genere e popularity_binary.*

Statistica Chi-quadrato

Il test confronta le frequenze osservate nelle celle della tabella di contingenza con le frequenze attese, che si otterrebbero se le due variabili fossero indipendenti. Come anticipato al Capitolo 1 (Sezione 1.2), la statistica test del Chi-quadrato è calcolata utilizzando la seguente formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

dove:

- O_{ij} sono le frequenze osservate nella tabella di contingenza
- E_{ij} sono le frequenze attese, calcolate come $E_{ij}\ \frac{r_i \times c_j}{n}$, dove r_i è il totale della riga, c_j è il totale della colonna e n è il totale delle osservazioni.

Per poter svolgere il test del Chi-quadrato, occorre che entrambe le variabili siano categoriche, le osservazioni siano indipendenti, e che le frequenze attese in ciascuna cella della tabella di contingenza siano almeno 5.

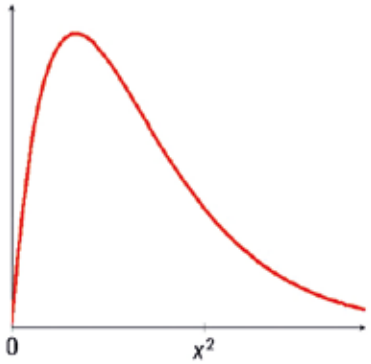


Figura 34:
Distribuzione Chi-quadrato.

Il valore della statistica χ^2 segue una distribuzione Chi-quadrato con gradi di libertà $df = (r-1)(c-1)$ dove r è il numero di righe e c è il numero di colonne della tabella di contingenza. Valori elevati di χ^2 indicano che le frequenze osservate si discostano significativamente da quelle attese, suggerendo una relazione tra le variabili.

La verifica delle ipotesi: Chi-quadrato

Le ipotesi per il test del Chi-quadrato di indipendenza sono:

- Ipotesi nulla → H_0: Non esiste relazione tra le variabili (le variabili sono indipendenti)
- Ipotesi alternativa → H_1: Esiste una relazione tra le variabili (le variabili non sono indipendenti)

Se il valore della statistica χ^2 supera il valore critico della distribuzione Chi-quadrato per un dato livello di significatività α, si rifiuta l'ipotesi nulla H_0, concludendo che esiste una relazione significativa tra le variabili. Come per i test precedenti, anche in questo caso la decisione può essere presa in base al *p-value* associato alla statistica test.

HOW TO? TEST CHI-QUADRATO SU EXCEL

Per calcolare la statistica Chi-quadrato occorre costruire la distribuzione delle frequenze attese. Costruiamo quindi una nuova tabella pivot utilizzando la stessa struttura di quella creata in precedenza.

Utilizzando la formula $E_{ij} = \frac{r_i \times c_j}{n}$ saremo in grado di calcolare le frequenze attese per ogni cella. Nell'esempio che segue è stata calcolata la frequenza attesa $E_{11} = \frac{34988 \times 225166}{232725} = 33851.58$

Frequenze osservate O_{ij}

Conteggio di popular	Etichette di colc		
Etichette di riga	0	1	Totale complessivo
Classico e teatrale	34981	7	34988
Elettronica e moderni	36256	43	36299
Jazz e tradizionali	26986	142	27128
Pop e commerciali	47734	3717	51451
Rap e urban	34185	2423	36608
Rock e alternative	45024	1227	46251
Totale complessivo	**225166**	**7559**	**232725**

Frequenze attese $E_{ij} = (r_i \times c_j)/n$

Genere	Popularity_0	Popularity_1
Classico e teatrale	=D4*B10/$D$10	1136.42
Elettronica e moderni	35119.99	1179.01
Jazz e tradizionali	26246.87	881.13
Pop e commerciali	49779.85	1671.15
Rap e urban	35418.96	1189.04
Rock e alternative	44748.75	1502.25

Figura 35:
Frequenze osservate e frequenze attese.

Una volta calcolate le frequenze attese, per poter calcolare la statistica test χ^2 costruiamo una terza tabella, dove ogni cella sarà calcolata come: $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

A questo punto abbiamo tutti gli elementi per poter calcolare χ^2. Ci basterà infatti sommare tutti gli elementi della tabella con la funzione **=SOMMA**() per ottenere $\chi^2 = 6896.63$.

	$(O_{ij} - E_{ij})^2 / E_{ij}$	
Genere	**Popularity_0**	**Popularity_1**
Classico e teatrale	37.68	1122.47
Elettronica e moderni	36.75	1094.57
Jazz e tradizionali	20.81	620.01
Pop e commerciali	84.08	2504.57
Rap e urban	42.99	1280.57
Rock e alternative	1.69	50.43
Chi-quadrato (χ^2)	6896.63	
p-value	0.00	

Figura 36. *Calcolo statistica χ^2 e p-value.*

Definiamo le ipotesi come:

H_0: Non esiste relazione tra le variabili
H_1: Esiste una relazione tra le variabili

Per calcolare la significatività statistica (*p-value*) utilizziamo la funzione:
=TEST.CHI.QUAD(freq. osservate; freq. attese)
Tramite la quale selezioneremo l'intervallo di frequenze osservate ed attese per generare il *p-value* associato alla statistica test.
Sulla base dei risultati, rifiutiamo H_0, concludendo che non esiste alcuna relazione tra il genere musicale e la popolarità dei brani.

2.6 Coefficiente di correlazione

Nel capitolo precedente abbiamo introdotto il coefficiente di correlazione di Pearson, che misura la forza e la direzione della relazione lineare tra due variabili quantitative. Per la verifica d'ipotesi sulla correlazione, la statistica test viene calcolata utilizzando il coefficiente di correlazione campionaria r:

$$t_r = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{n - 2}$$

dove n rappresenta la dimensione del campione. La statistica t_r segue una distribuzione T di Student con $n - 2$ gradi di libertà sotto l'ipotesi nulla.

Verifica delle ipotesi: correlazione

Nella verifica d'ipotesi per il coefficiente di correlazione, le ipotesi vengono formulate come segue:

- Ipotesi nulla → $H_0 : \rho = 0$ (non c'è correlazione tra le variabili nella popolazione)
- Ipotesi alternativa → $H_1 : \rho \neq 0$ (esiste una correlazione significativa tra le variabili)

Anche in questo caso, per decidere se accettare o rifiutare l'ipotesi nulla, si confronta il valore assoluto della statistica test $|t_r|$ con il valore critico $|t_\alpha|$ sulle tavole statistiche, per un dato α e con $n - 2$ gradi di libertà. Se $|t_r| \geq |t_\alpha|$, si rifiuta l'ipotesi nulla H_0, indicando che la correlazione è statisticamente significativa. La stessa conclusione può essere raggiunta analizzando il *p-value* associato alla statistica test.

HOW TO? **VERIFICA D'IPOTESI SUL COEFFICIENTE DI CORRELAZIONE SU EXCEL**

Calcoliamo il coefficiente di correlazione tra le due variabili energy e loudness, utilizzando la funzione **=CORRELAZIONE**(variabile1; variabile2)

Il risultato è $r = 0.816$, suggerendo una correlazione positiva (relazione diretta) e forte. Per testare la significatività statistica di questo coefficiente, svolgiamo la verifica d'ipotesi:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

E calcoliamo la statistica test:

$$t_r = \frac{0.816}{\sqrt{1 - 0.816^2}} = \sqrt{232725 - 2} = 681.21$$

Ora possiamo confrontare questo valore con il valore critico sulle tavole statistiche della distribuzione T di Student con $df = n - 2$ e $\alpha = 0.05$.

=INV.T.2T(0.05; 232723) $f_{0.05} = 1.96$

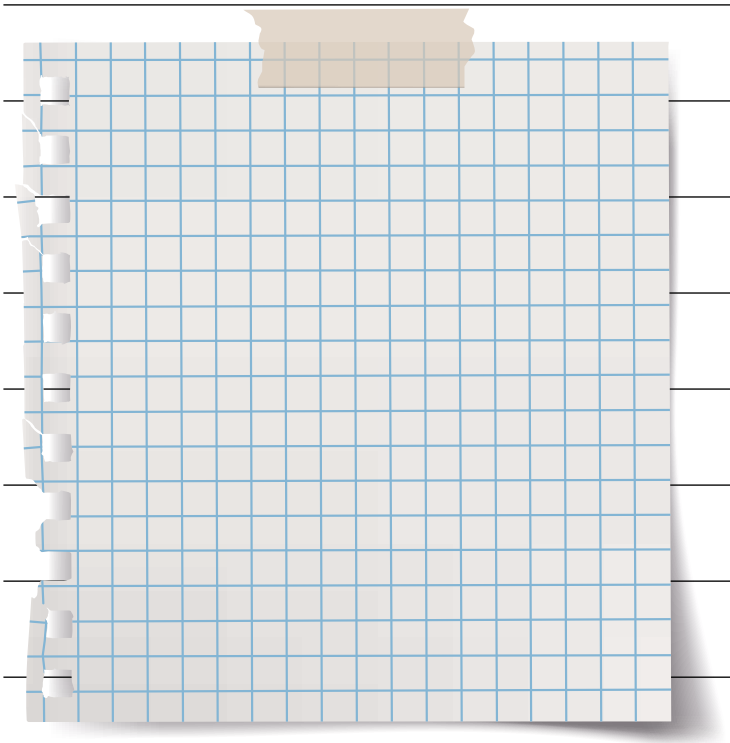
In alternativa, possiamo calcolare il *p-value* associato a questa statistica test, utilizzando la funzione:

=DISTRIB.T.2T(stat.test; 232724) $p.value = 0.00$

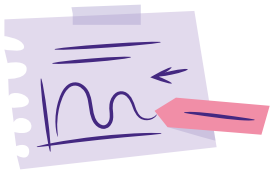
I risultati suggeriscono di rifiutare l'ipotesi nulla (H_0): il coefficiente di correlazione è statisticamente diverso da zero.

appunti

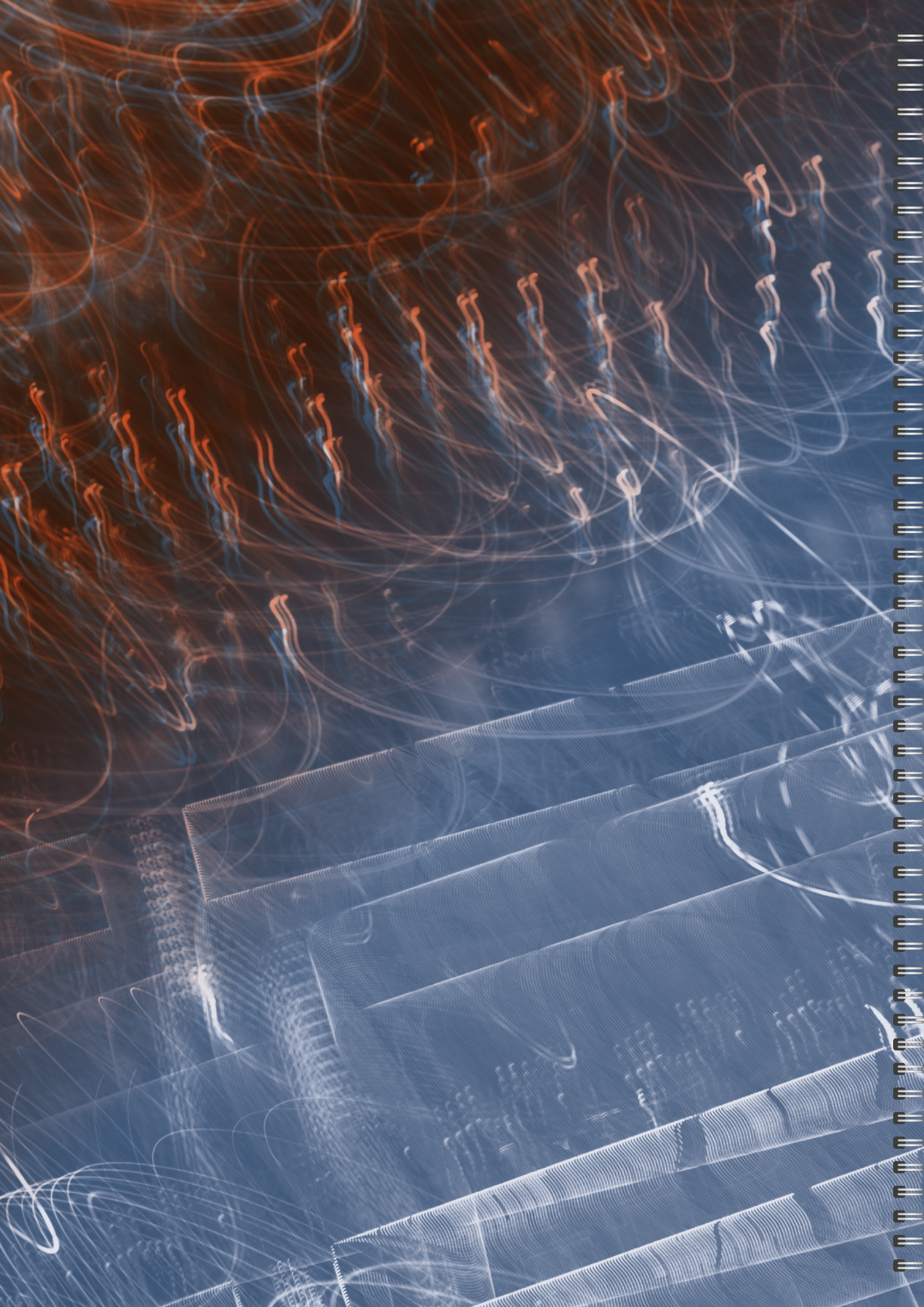
appunti



appunti



appunti



Capitolo[3]

Regressione lineare

Nello studio delle relazioni tra due (o più) variabili, oltre a misurare l'entità del legame esistente, si è anche interessati ad accertare come varia una di esse al variare dell'altra (o delle altre). Si vuole per questo individuare un'opportuna funzione che metta in relazione due o più variabili (di cui **una dipendente** e le altre **indipendenti o esplicative**).
Si studia una **relazione asimmetrica** attraverso la modellistica e la regressione lineare è uno dei possibili modelli statistici.

Obiettivi della modellistica sono:

- **Descrizione:** rappresentare con una funzione (semplice) l'andamento in media dei valori d'una variabile al variare dell'altra (o delle altre)
- **Interpretazione:** evidenziare la relazione tra variabili, per consentirne una spiegazione alla luce di precise teorie
- **Previsione o stima:** valutare il valore che assumerà la variabile dipendente in corrispondenza di un valore noto della variabile esplicativa

Se si utilizza un modello lineare, attraverso un **approccio esplorativo** dalla matrice dei coefficienti di correlazione si rilevano le coppie di variabili con correlazione abbastanza elevata per le quali è utile lo studio delle variazioni d'una di queste in funzione dell'altra. In questo caso sono i dati rilevati che indicano le relazioni da considerare.
Nell'**approccio modellistico** invece si definisce a priori una relazione lineare tra le variabili ed in base ad un campione di n unità statistiche, si stimano i parametri del modello e si accerta la validità dello stesso. La validità fornisce a posteriori una conferma (o smentita), di natura empirica, della relazione ipotizzata. In generale, per la costruzione di un modello statistico si utilizza il seguente diagramma di flusso (Figura 37):

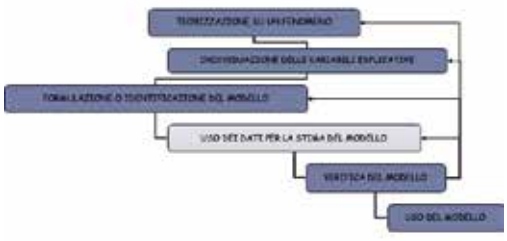


Figura 37: *Diagramma di flusso per la costruzione di un modello statistico.*

NotaBene > l'analisi statistica, volta ad accertare o a respingere un modello, è riferita ad un certo tempo, ad un dato luogo e ad un insieme di unità e le conclusioni raggiunte valgono limitatamente a tali circostanze; variando tali elementi cambiano i risultati.

Il modello lineare è quello di maggiore interesse e di più ampia applicazione, per i seguenti motivi:

- **Semplicità:** la retta è la più semplice funzione che lega due variabili ed è facile da interpretare
- **Effettiva linearità:** molte relazioni sono lineari o assai vicine alla linearità
- **Trasformazioni:** spesso è possibile ottenere una relazione lineare trasformando le variabili (es: passaggio ai logaritmi o ai quadrati)
- **Limitatezza dell'intervallo:** anche se la relazione tra due variabili non è lineare, considerando un intervallo limitato dei valori la retta può fornire una buona approssimazione

Se la variabile esplicativa è una sola, il modello si chiama *regressione lineare semplice* ed è così formulato:

$$y_i = a + bx_i + e_i \qquad i = 1, \ldots, n$$

y_i = Valore della variabile dipendente, rilevato nella i-esima unità
x_i = Valore della variabile esplicativa, rilevato nella i-esima unità
a = Intercetta (ordinata all'origine, costante)
b = Coefficiente angolare (coefficiente di regressione)
e_i = Il "residuo non spiegato" relativo all'osservazione i-esima
n = Numero di osservazioni (coppie di valori) rilevate

I residui

$$e_i = y_i - \hat{y}_i$$

Sono la differenza tra i valori stimati dal modello di regressione e i valori osservati. Hanno natura accidentale. Nel modello di regressione si assume che ciascun valore osservato della variabile dipendente sia esprimibile come funzione lineare del corrispondente valore della variabile esplicativa, più un termine residuo che traduce l'incapacità del modello di riprodurre con esattezza la realtà osservata.

Si considera la regressione come un problema di **interpolazione** cioè di adattamento di una funzione (in questo caso la retta) alla "nuvola" dei punti del diagramma di dispersione, in base a considerazioni di natura geometrica. La funzione lineare dei valori x_i rappresenta la retta di regressione:

$$\hat{y}_i = a + bx_i \qquad i = 1, \ldots, n$$

Consideriamo come esempio la variabile *Danceability* (la ballabilità del brano) per il genere musicale *A cappella* (canzoni senza strumenti). Dalla matrice di correlazione con le altre variabili numeriche vediamo che il maggior legame lineare è con la variabile *Valence*. Di seguito sono riportate le formule e i risultati per calcolare le correlazioni nel file Spotify:

CORRELAZIONI tra *deanceability* e le altre variabili numeriche		
	FORMULA	RISULTATO
popularity	=correlazione(G554:G672; E554:E672)	0.13
acousticness	=correlazione(G554:G672; F554:F672)	-0.59
duration_ms	=correlazione(G554:G672; H554:H672)	-0.28
energy	=correlazione(G554:G672; I554:I672)	0.64
instrumentalness	=correlazione(G554:G672; J554:J672)	-0.13
liveness	=correlazione(G554:G672; L554:L672)	0.08
loudness	=correlazione(G554:G672; M554:M672)	0.46
speechiness	=correlazione(G554:G672; O554:O672)	0.61
tempo	=correlazione(G554:G672; P554:P672)	0.06
valence	=correlazione(G554:G672; R554:R672)	0.80

Figura 38: *Formule per calcolare le correlazioni nel file Spotify, e relativi risultati.*

Possiamo quindi provare a formulare un modello di regressione lineare che ci dica come varia la *Danceability* in funzione di *Valence* in questo genere di canzoni. Il diagramma di dispersione di queste due variabili conferma un andamento lineare diretto (al crescere di una variabile cresce anche l'altra, come attestato dal coefficiente di correlazione positivo). La retta di regressione migliore è quella che più si avvicina all'insieme dei punti, ovvero alle coppie di valori (x_i, y_i). Per la stima dei parametri a e b si adotta il metodo dei minimi quadrati. Si cerca la retta che rende minima la somma dei quadrati dei residui:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = min$$

La retta di regressione i cui valori teorici soddisfano la condizione di accostamento è quella con il miglior adattamento. Essa si avvicina maggiormente all'insieme dei punti osservati, adottando come criterio il quadrato delle differenze. Gli errori, di natura accidentale, "spiegheranno" la diversa *Danceability* a parità di *Valence*.

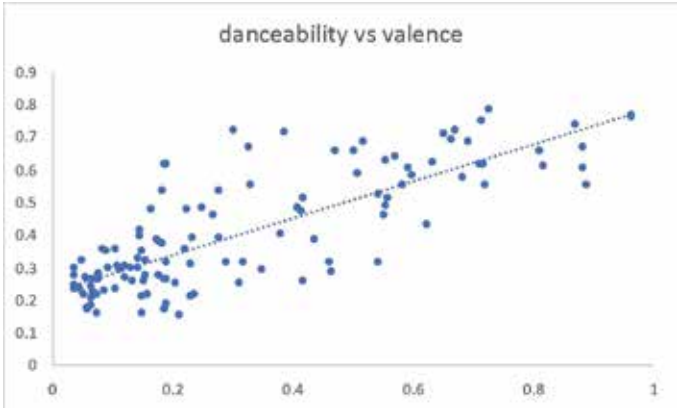


Figura 39: *Grafico di dispersione (scatterplot) tra i caratteri* Danceability *e* Valence. *La linea tratteggiata indica la retta stimata di regressione.*

Per la stima del modello:

Si pongono uguali a zero le derivate prime parziali rispetto alle incognite e (condizione di minimo):

$$2\sum_{i=1}^{n}(y_i - a - bx_i)(-1) = 0$$

e

$$2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i) = 0$$

Si moltiplicano primo e secondo membro di ciascuna equazione per (–1/2), ottenendo il sistema di equazioni lineari, detto **sistema di equazioni normali**:

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

e

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Isolando e si ottiene:

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

In Excel a e b sono calcolati con le funzioni intercetta e pendenza, che prendono come input i valori della variabile dipendente e quelli della variabile indipendente. Se nel file Spotify i valori della *Danceability* per il genere *A cappella* sono nelle celle dalla G554 al G672 e quelli della variabile *Valence* sono dalla R554 alla R672, le funzioni dovranno essere digitate nel modo seguente:

=intercetta (G554:G672; R554:R672)
=pendenze (G554:G672; R554:R672)

NotaBene > prima vanno inseriti i valori della variabile y e poi quelli della variabile x poiché le due funzioni non sono simmetriche (come ad esempio la correlazione, che dà risultati uguali scambiando l'ordine dei vettori inserito tra parentesi) ma sono asimmetriche.

I risultati numerici delle funzioni sono, rispettivamente, 0.23 e 0.56. Il modello di regressione stimato è quindi:

$$y_i = 0.23 + 0.56x_i + e_i$$

I parametri stimati hanno un significato interpretativo oltre a quello geometrico: all'aumentare di un punto nella *Valence*, la *Danceability* aumenta in media di 0.56. La *Danceability* stimata per una canzone di genere *A cappella* senza *Valence* è 0.23.

3.1 Bontà di adattamento

La verifica della validità o bontà di adattamento della retta di regressione è la fase conclusiva dell'analisi ed è diretta a controllare che la retta di regressione sia realmente in grado di spiegare l'andamento delle osservazioni. Si può adattare una retta con il metodo dei minimi quadrati anche nei casi in cui i punti non seguono una relazione lineare ed in queste circostanze la retta di regressione ha una capacità minima, o nulla, di riassumere la relazione tra le variabili. Questa circostanza deve essere evidenziata dalla verifica della bontà di adattamento.
L'indice che si calcola per la bontà di adattamento è chiamato indice di determinazione lineare e si basa sulla seguente scomposizione della devianza:

$$y_i = \hat{y}_i + e_i \qquad DEV(Y) = DEV(\hat{Y}) + DEV(E)$$

$$DEV(Y) = \sum_{i=1}^{n} (y_i - M_y)^2$$

Devianza della variabile dipendente (devianza totale)

$$DEV(\widehat{Y}) = \sum_{i=1}^{n} (\widehat{y}_i - M_y)^2$$

Devianza dei valori stimati (devianza di regressione)

$$DEV(E) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Devianza dei residui (devianza residua)

Le espressioni sono tutte realmente delle devianze: la media dei valori teorici è uguale alla media dei valori osservati e la media dei residui è nulla. La relazione fondamentale è la seguente:
La devianza totale dei valori osservati di Y è uguale alla somma della devianza di regressione e della devianza residua.

Ricordando che la devianza è una misura di variabilità, la variabilità delle osservazioni è il risultato di due componenti:

- La prima è la variabilità dei valori stimati (devianza di regressione)
- La seconda è la variabilità dei punti attorno alla retta adattata (devianza residua)
- La devianza di regressione è quella parte della devianza totale che è "spiegata" dalla relazione lineare con la variabile indipendente
- La retta di regressione fornisce una rappresentazione tanto più soddisfacente della relazione tra le variabili quanto più la *DEV(E)* risulta prossima a 0
- Il valore della devianza residua non è direttamente utilizzabile per misurare la bontà di adattamento, poiché è influenzato dall'ordine di grandezza e dall'unità di misura della variabile dipendente e dal numero di osservazioni

Una misura relativa (e normalizzata) è l'indice di determinazione lineare δ definito come rapporto tra la devianza di regressione e la devianza totale:

$$\delta = \frac{DEV(\widehat{Y})}{DEV(Y)} = 1 = \frac{DEV(E)}{DEV(Y)}$$

→ Assume valori in [0, 1]:

- $\delta = 0$ Adattamento pessimo
- $\delta = 0$ Adattamento perfetto

$$\delta = 0 \Rightarrow DEV(\widehat{Y}) = \sum_{i=1}^{n} (\widehat{y}_i - M_y)^2 = 0$$

→

$$\delta = 0 \Rightarrow \widehat{y}_i = M_y$$

I valori teorici sono tutti uguali alla media delle osservazioni di Y, qualunque sia il valore di X:

- La retta, che dovrebbe esprimere la dipendenza di Y da X, non spiega l'andamento dei punti più di quanto non lo spieghi la media di Y, senza tener conto di alcuna variabile esplicativa
- La retta adatta ha il coefficiente di regressione nullo ($b = 0$) e quindi pendenza nulla

$$\delta = 1 \Rightarrow DEV(E) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = 0$$

$$\delta = 1 \Rightarrow \widehat{y}_i = y_i$$

Tutti i punti si collocano esattamente sulla retta di regressione.

- Le situazioni $\delta = 0$ e $\delta = 1$ rappresentano dei casi limite. In pratica, si ha un indice di determinazione lineare interno all'intervallo [0, 1]
- Se l'adattamento della retta ai dati (misurato dall'indice di determinazione lineare) si rivela particolarmente buono (con un indice prossimo ad 1), si può dire che

la variabilità di Y è «spiegata» in misura notevole dalla retta di regressione
- Non si utilizza il termine «causata» poiché un valore di δ prossimo ad 1 non implica necessariamente un nesso causale tra le due variabili
- L'indice di determinazione lineare è uguale al quadrato del coefficiente di correlazione lineare:

$$\delta = r^2_{xy}$$

L'entità della relazione lineare tra due variabili può essere misurata in due modi:
- Attraverso l'indice di determinazione lineare
- Attraverso il coefficiente di correlazione lineare
  - Il giudizio non cambia (il primo è quadrato del secondo)
  - Cambia il significato interpretativo dei valori numerici

In Excel l'indice di determinazione lineare si calcola come quadrato del coefficiente di correlazione:

=(correlazione (G554:G672; R554:R672))^2

Il risultato è 0.64. Il modello stimato è quindi un buon modello poiché il 64% della variabilità della *Danceability* è spiegata dalla relazione lineare con la *Valence* e solo il rimanente 36% di variabilità è dovuto al caso o ad altri fattori.

Per calcolare congiuntamente a,b ed il coefficiente di determinazione si possono utilizzare due strategie alternative:

1) Cliccando su uno dei punti del grafico a dispersione, aprendo il menu con il tasto destro del mouse si seleziona "aggiungi linea di tendenza". Nelle opzioni che si aprono si seleziona "lineare", "visualizza l'equazione sul grafico" e "visualizza il valore di r quadrato sul grafico". In questo modo apparirà sul grafico l'equazione della retta con il valore del coefficiente di determinazione lineare:

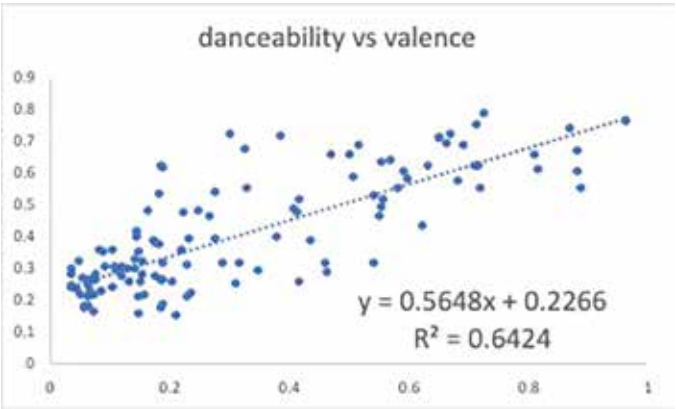


Figura 40: *Grafico di dispersione (scatterplot) tra i caratteri* Danceability *e* Valence. *La linea tratteggiata indica la retta stimata di regressione. Il grafico riporta inoltre l'equazione della retta stimata e il coefficiente di determinazione lineare.*

2) Si utilizza la funzione **regr.lin** che è una funzione in forma di matrice (restituisce più di un valore come output). Oltre a richiedere i valori della variabile dipendente e quelli della variabile indipendente, la funzione necessita di altri due valori di input:

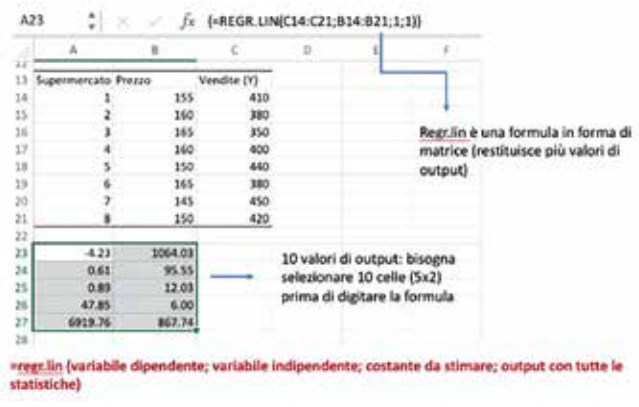
   - 1 o 0 a seconda che si voglia stimare una retta con qualsiasi intercetta (valore 1 da inserire) o una retta vincolata a passare per l'origine e quindi con solo la pendenza da stimare (valore 0 da inserire)
   - 1 o 0 a seconda che si vogliano come output tutti i dieci valori calcolati dalla funzione, fra cui il coefficiente di determinazione lineare, (valore 1 da inserire) o solo i parametri a e b (valore 0 da inserire)

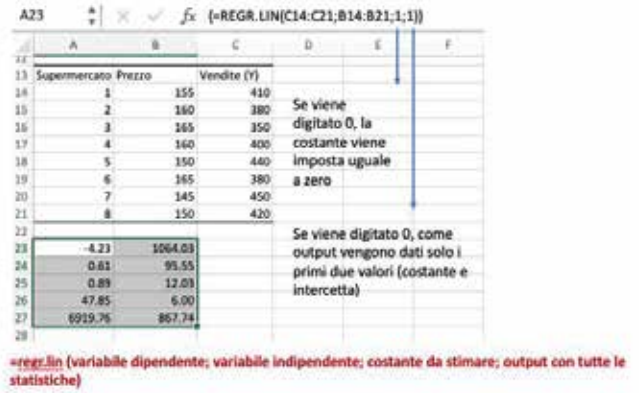
Nel nostro esempio, dopo aver selezionato 10 celle di output con il mouse nella forma 5x2, si dovrà digitare la seguente formula:

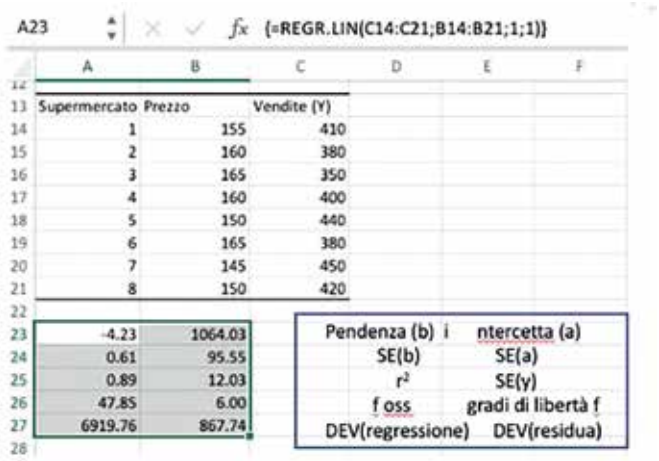
=regr.lin (G554:G672; R554:R672; 1; 1)

Stando attenti a digitare assieme al tasto di invio anche i tasti **control e alt** (in basso a sinistra nella tastiera) per inserire la formula in forma di matrice in maniera corretta. Se digitata correttamente, la formula apparirà scritta tra parentesi graffe.

Consideriamo come esempio un nuovo data set relativo a 8 supermercati ed il prezzo di vendita di un certo prodotto con le quantità vendute nell'ultimo mese. La Figura 42 riporta il data set con la funzione **regr.lin** ed i valori di output.







	A	B	C	D	E	F
13	Supermercato	Prezzo	Vendite (Y)			
14	1	155	410			
15	2	160	380			
16	3	165	350			
17	4	160	400			
18	5	150	440			
19	6	165	380			
20	7	145	450			
21	8	150	420			
22						
23	-4.23	1064.03	Pendenza (b) i	ntercetta (a)		
24	0.61	95.55	SE(b)	SE(a)		
25	0.89	12.03	r^2	SE(y)		
26	47.85	6.00	f oss	gradi di libertà f		
27	6919.76	867.74	DEV(regressione)	DEV(residua)		
28						

A23 {=REGR.LIN(C14:C21;B14:B21;1;1)}

Figura 41: *Esempio di Utilizzo della funzione regr.lin in Excel.*

Come evidenziato in Figura 41, il valore di b viene riportato nella prima colonna e prima riga, il valore di a nella seconda colonna e prima riga, il valore di r quadrato nella prima colonna e terza riga. Tralasciamo, in questa breve pubblicazione, la spiegazione degli altri valori calcolati dalla funzione.

La funzione stimata sarà quindi:

$$y_i = 1064 - 4.23x_i$$

Il modello indica una relazione inversa tra prezzo di vendita e quantità vendute: ad un incremento di un euro di prezzo corrisponde una diminuzione media dei pezzi venduti di 4.23 unità. Il valore (teorico) dei pezzi venduti con un prezzo di vendita di 0 euro è 1064.

Il modello è un buon modello perché l'89% della variabilità dei pezzi venduti è dato dalla relazione lineare con il prezzo e solo il rimanente 11% di variabilità è dovuto al caso o ad altri fattori.

Per questo data set, la Figura seguente riporta gli altri due metodi per calcolare i parametri e la bontà di adattamento del modello.

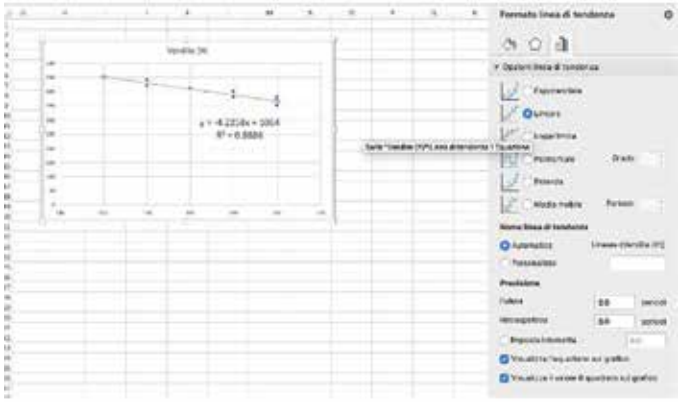
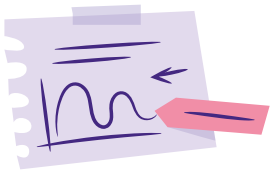


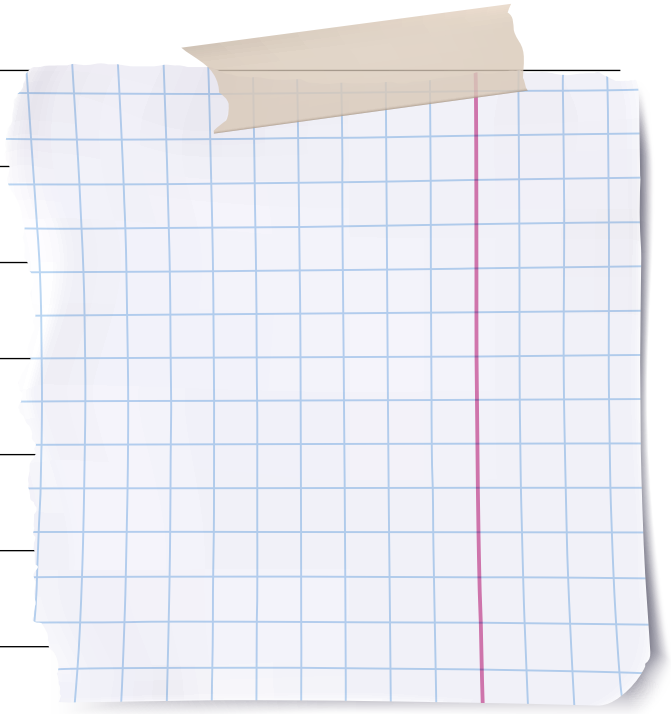
Figura 42: *Metodi alternativi in Excel per calcolare i parametri del modello di regressione lineare e la bontà di adattamento.*

	A	B	C	D	E
1					
2					
3	Supermercato	Dipendenti (X)	Fatturato (Y)	^2	y^2
4	1	10	1.9	100324	3.61
5	2	18	3.1	400	9.61
6	3	20	3.2		10.24
7	4	8	1.5		
8	5	30	6.2		
9	6	12	2.8		
10	7	14	2.3		
11	TOT	112	21	2128	77.28
12					
13	Supermercato	Prezzo	Vendite (Y)		
14	1	155	410		
15	2	160	380		
16	3	165	350		
17	4	160	400		
18	5	150	440		
19	6	165	380		
20	7	145	450		
21	8	150	420		
22					
23		-4.22581	=PENDENZA(C14:C21;B14:B21)		
24		1064.03226	=INTERCETTA(C14:C21;B14:B21)		
25		0.88857	=(CORRELAZIONE(C14:C21;B14:B21))^2		
26					

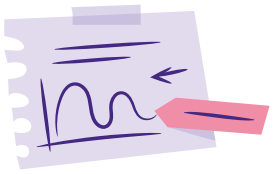
appunti



appunti



appunti



appunti

Conclusioni

In questo volume, abbiamo percorso insieme un viaggio nel mondo della statistica applicata, utilizzando dati reali per esplorare strumenti e tecniche fondamentali dell'analisi statistica. Il nostro compagno di avventure è stato il dataset Spotify, un enorme database di oltre 230.000 canzoni raccolte tra Agosto e Settembre 2018. Ogni canzone è stata per noi una "unità statistica" ricca di informazioni, come la popolarità, il genere musicale, la durata, la velocità (BPM), l'intensità sonora (*Loudness*), la ballabilità ed altre caratteristiche audio.

Riassumiamo i principali punti trattati e ciò che abbiamo scoperto in relazione al dataset Spotify.

Statistica descrittiva

Abbiamo iniziato dalle basi: la statistica descrittiva. Questa prima tappa ci ha permesso di "guardare" i dati per trarre alcune informazioni generali. L'obiettivo è sintetizzare i dati attraverso analisi grafiche (istogrammi, grafici a barre), tabelle di frequenza ed indici (media, mediana, moda, etc.) per acquisire una visione d'insieme dei dati ed individuare le caratteristiche principali delle canzoni.

Abbiamo scoperto che il dataset Spotify è dominato da generi come "Pop e commerciali" e "Rock e alternative", mentre altri generi, come il Jazz e la musica tradizionale, sono meno frequenti. Abbiamo anche visto come le canzoni con un tempo "moderato" siano le più comuni e che la maggior parte dei brani è in 4/4. Inoltre, abbiamo osservato una tendenza interessante: i brani con più "energia" tendono ad avere una intensità sonora più alta. Questa relazione non è perfettamente lineare per tutte le canzoni, ma abbiamo notato un trend generale che sembra rispecchiare l'intuizione.

Inferenza statistica e test delle ipotesi

Dopo aver compreso come descrivere i dati, siamo passati all'inferenza statistica. Questa fase ci permette di fare "salti di fiducia": partendo dai dati osservati (il campione), possiamo formulare ipotesi su un'intera popolazione. Abbiamo introdotto il concetto di ipotesi nulla e ipotesi alternativa per verificare se una determinata osservazione sui dati (ad esempio, la differenza di popolarità tra generi musicali) è significativa o solo casuale. Utilizzando strumenti come il t-test, l'ANOVA e il test del Chi-quadrato, abbiamo risposto a domande specifiche sui dati.

Ad esempio, il t-test ci ha permesso di confrontare il volume medio delle canzoni in tonalità maggiore e minore, evidenziando che c'è una differenza significativa tra le due: le canzoni in tonalità maggiore potrebbero avere una tendenza ad essere più "forti" o più "deboli" rispetto a quelle in tonalità minore. L'ANOVA ha mostrato che l'energia di un brano varia notevolmente a seconda del genere musicale, mentre il test del Chi-quadrato ci ha confermato che non c'è una relazione tra la popolarità di un brano e il suo genere. Il test d'ipotesi sul coefficiente di correlazione, infine, ci ha confermato quanto osservato qualitativamente nel Capitolo 1: esiste una relazione lineare significativa tra energia ed intensità sonora di un brano musicale.

Regressione lineare

Infine, abbiamo affrontato la regressione lineare: una tecnica per analizzare e prevedere relazioni tra due variabili.

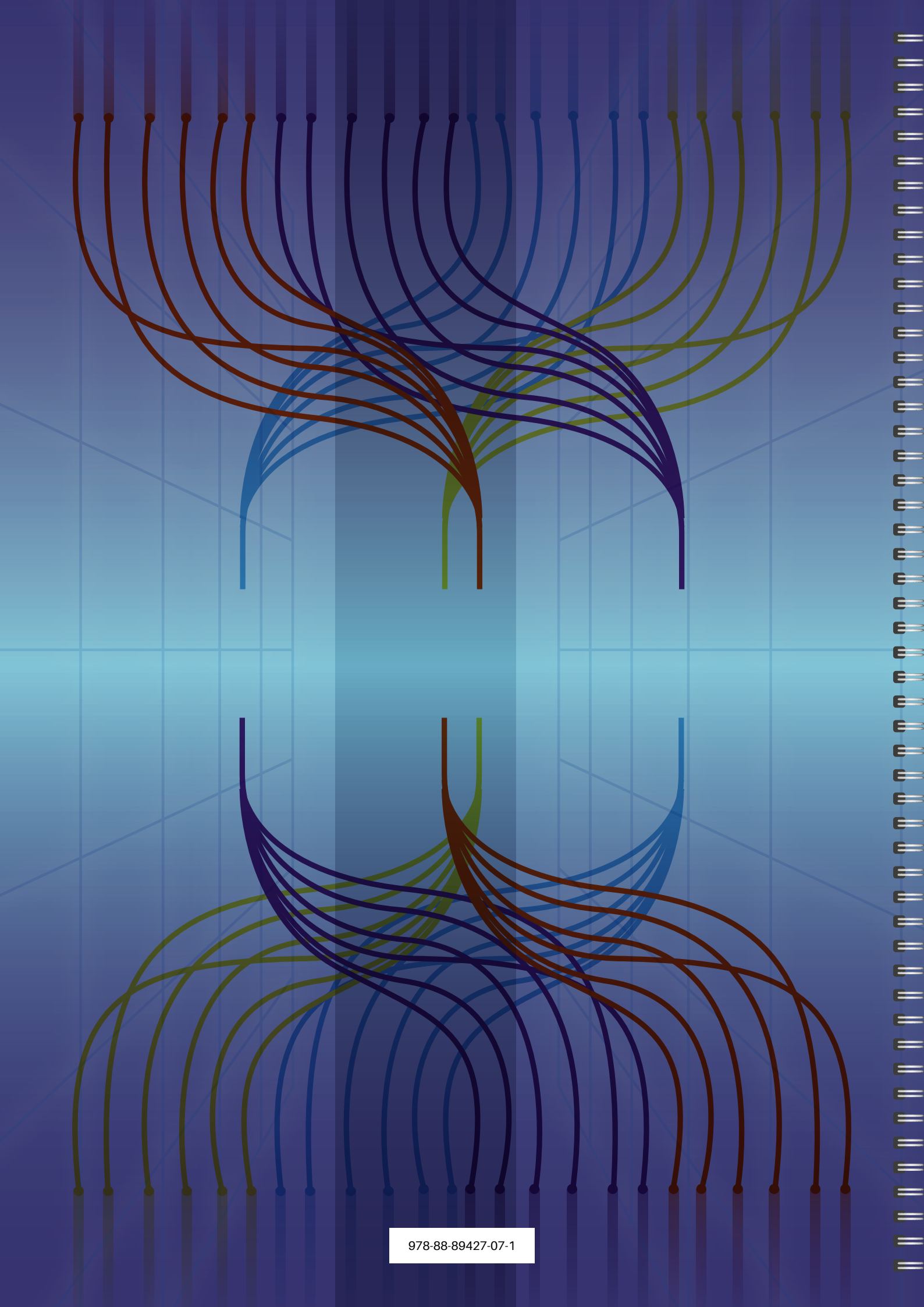
Abbiamo utilizzato il modello di regressione lineare (semplice) per studiare la relazione tra la ballabilità (*Danceability*) e la valenza emotiva (*Valence*) dei brani. Abbiamo osservato che nei brani *a cappella*, all'aumentare della positività emotiva, aumenta anche la ballabilità: le canzoni più "positive" sembrano essere anche più facili da ballare. Questo legame è stato descritto tramite una retta di regressione, che ci ha mostrato come la ballabilità tenda ad aumentare di 0.56 punti per ogni punto di valenza. Grazie all'indice di determinazione, abbiamo potuto anche giudicare la bontà del modello, che si è dimostrato affidabile ($\delta = 0.64$) per questo tipo di analisi.

Applicazione pratica con Excel

Durante tutto il percorso, abbiamo utilizzato Excel come strumento pratico per le analisi statistiche. Ci siamo esercitati con funzioni, tabelle pivot e grafici, imparando ad usare un programma accessibile per gestire e analizzare dati. Questi esercizi non solo aiutano a consolidare le competenze statistiche, ma forniscono anche una base pratica che sarà utile sia in ambito accademico che lavorativo.

In sintesi, questo volume ha fornito una panoramica delle principali tecniche statistiche, cercando di renderle accessibili e concrete. Speriamo che queste conoscenze vi abbiano aiutato a sviluppare una mentalità orientata ai dati e vi abbiano dato gli strumenti per rispondere a domande complesse in modo critico e informato. La statistica può sembrare una materia difficile, ma, con pazienza e pratica, diventa uno strumento potente per comprendere meglio la realtà, anche nei contesti meno prevedibili, come il mondo della musica.

E ora tocca a voi: riuscireste a replicare queste analisi o a scoprire qualcosa di nuovo? È arrivato il momento di mettervi alla prova!



978-88-89427-07-1