GUIDA OPERATIVA: COME COSTRUIRE UN DATASET PER UNA TESI SPERIMENTALE

Introduzione

Queste indicazioni devono essere considerate come base per la strutturazione e lo sviluppo di un dataset nell'ambito di una tesi sperimentale. L'obiettivo è fornire un percorso chiaro e guidato che permetta allo studente di raccogliere dati in modo coerente, organizzato e metodologicamente corretto.
La costruzione del dataset rappresenta una fase centrale nello sviluppo di una tesi sperimentale. Non si tratta di un passaggio puramente tecnico, ma di un momento in cui la domanda di ricerca viene tradotta in forma operativa. Un dataset ben progettato consente di svolgere analisi coerenti, riduce il rischio di errori e facilita l'interpretazione dei risultati.

Definizione degli obiettivi

La prima fase consiste nella definizione dell'obiettivo della ricerca. È fondamentale chiarire cosa si vuole indagare e quali sono le domande di ricerca. Un dataset efficace nasce sempre da un problema ben definito: ad esempio, analizzare un fenomeno, confrontare gruppi diversi oppure osservare cambiamenti nel tempo.

Ogni dataset nasce da una domanda di ricerca. Prima di iniziare a costruire la tabella dei dati, è necessario avere chiaro cosa si vuole studiare e quali relazioni si intendono analizzare.

Si consideri il seguente esempio:

"L'utilizzo dei social media influisce sul benessere psicologico e sul rendimento accademico degli studenti e delle studentesse?"

A partire da questa domanda, è possibile individuare gli elementi principali della ricerca. In particolare, è necessario distinguere tra ciò che si vuole spiegare e i fattori che potrebbero influenzarlo. Nel caso in esame, il benessere psicologico e il rendimento accademico rappresentano le variabili che si intendono spiegare (variabile dipendente o *outcome*), mentre l'utilizzo dei social media rappresenta uno dei possibili fattori esplicativi (variabile indipendente). Questo passaggio è fondamentale, poiché consente di identificare in modo chiaro le variabili che dovranno essere incluse nel dataset e analizzate nello studio.

Identificazione e definizione delle variabili

Le variabili costituiscono le colonne del dataset e devono essere definite in modo chiaro e coerente con gli obiettivi della ricerca. Ogni variabile deve avere un significato preciso e deve essere misurabile.

Nel nostro esempio, le variabili principali possono essere definite come segue:

- il tempo trascorso sui social media, espresso in ore giornaliere;
- il livello di benessere psicologico, rilevato attraverso una scala;

- il rendimento accademico, misurato attraverso il voto medio.

A queste si aggiungono variabili che descrivono le caratteristiche degli studenti, come l'età, il genere e il corso di studi. Queste variabili sono importanti perché permettono di controllare eventuali effetti o differenze tra gruppi.

È importante sottolineare che ogni variabile deve essere definita prima di iniziare la raccolta dei dati. Modificare le variabili in corso d'opera può generare incoerenze difficili da correggere.

Dalla variabile alla sua operazionalizzazione

Una variabile concettuale, come il "benessere psicologico", non è direttamente osservabile e deve essere trasformata in una misura concreta. Questo processo prende il nome di operazionalizzazione. Nel caso del benessere, ad esempio, si può utilizzare una scala composta da più item, il cui punteggio complessivo rappresenta la variabile finale. Allo stesso modo, il tempo trascorso sui social può essere rilevato chiedendo allo studente di indicare il numero medio di ore giornaliere. Questa fase richiede attenzione, perché la qualità dei dati dipende direttamente da come le variabili vengono misurate.
Per quanto riguarda le misure si raccomanda l'utilizzo di strumenti validati o già utilizzati in letteratura. È sempre utile fare una ricerca in letteratura degli strumenti utilizzati per valutare quello specifico costrutto.

Strutturazione del dataset

Una volta definite le variabili, è possibile costruire la struttura del dataset. Il formato standard prevede una tabella in cui ogni **riga** rappresenta un'unità di analisi (es., ogni soggetto), mentre ogni **colonna** rappresenta una variabile.

Nel caso in esame, ogni riga corrisponde a uno studente e contiene tutte le informazioni raccolte su di lui.

Una possibile struttura è la seguente:

ID	Età	Genere	Nazionalità	Ore_social	Benessere_1	Benessere_2	Benessere_3	Benessere_4	Benessere_5	Voto_medio

È buona pratica inserire una prima colonna contenente un codice identificativo univoco (ID), che consente di distinguere ogni osservazione.

I nomi delle variabili devono essere brevi, chiari e privi di spazi. Questo facilita l'utilizzo del dataset nei software di analisi.

Nella tabella potete trovare:

- *Informazioni demografiche:* Età, Genere, Nazionalità
- *Ore passate sui social*
- *Scala del benessere con tutti gli item di riferimento*
- *Voto medio esami*

Codifica delle informazioni

Non tutte le informazioni possono essere inserite direttamente in forma testuale. Le variabili qualitative, come il genere o il corso di studi, devono essere trasformate in valori codificati. Ad esempio, il genere può essere rappresentato assegnando il valore 0 ai maschi e il valore 1 alle femmine. Allo stesso modo, i corsi di studio possono essere identificati con numeri distinti. La codifica deve essere definita in anticipo e mantenuta costante in tutto il dataset. È fondamentale evitare cambiamenti nella codifica, perché renderebbero i dati incoerenti e difficili da analizzare.

Inserimento dei dati

Raccolta dei dati manuali

Una volta definita la struttura, è possibile procedere con l'inserimento dei dati. Questa fase richiede precisione, perché errori anche minimi possono influenzare i risultati.

È importante evitare:

- celle vuote non gestite;
- valori inseriti in formati diversi;
- errori di digitazione.

Raccolta dei dati tramite strumenti digitali

Nel caso in cui i dati vengano raccolti attraverso strumenti digitali, come questionari online, è importante considerare che molte piattaforme consentono di esportare direttamente le risposte in formato strutturato.
Strumenti come Google Moduli o LimeSurvey permettono infatti di scaricare automaticamente i dati raccolti in formato Excel o CSV. Questo consente di ottenere un dataset già organizzato in forma tabellare, in cui ogni riga corrisponde a un rispondente e ogni colonna a una domanda del questionario.

Tuttavia, anche in questi casi, è necessario verificare la qualità e la struttura dei dati esportati. In particolare, è importante controllare che:

- i nomi delle variabili siano chiari e coerenti;
- le modalità di risposta siano correttamente codificate;
- eventuali valori mancanti siano gestiti in modo appropriato.

L'utilizzo di questi strumenti facilita la fase di raccolta dati, ma non sostituisce la necessità di una corretta organizzazione e preparazione del dataset per l'analisi.

Controllo e pulizia del dataset

Prima di procedere all'analisi, è necessario verificare la qualità dei dati. Questo passaggio, spesso sottovalutato, è essenziale per garantire l'affidabilità dei risultati.

Il controllo deve riguardare:

- la presenza di valori mancanti;

- la presenza di valori non plausibili;
- la coerenza tra le variabili;
- eventuali duplicazioni.

Ad esempio, un voto superiore al massimo previsto o un'età negativa indica chiaramente un errore che deve essere controllato ed eventualmente corretto.

Documentazione del dataset

Un dataset non è completo senza una documentazione che ne descriva il contenuto. Questa documentazione, spesso chiamata *"codebook",* permette di comprendere il significato delle variabili e la loro codifica.

Per ogni variabile è necessario indicare:

- il nome;
- la descrizione;
- il tipo (numerica o categoriale);
- la codifica utilizzata.

Questa fase è fondamentale per rendere il dataset comprensibile anche a distanza di tempo o da parte di altri ricercatori, tramite la creazione di un codebook.

Conclusione

La costruzione di un dataset rappresenta un passaggio cruciale nello sviluppo di una tesi sperimentale. Richiede attenzione, pianificazione e coerenza metodologica. Un dataset ben costruito non solo facilita l'analisi, ma contribuisce in modo determinante alla qualità complessiva del lavoro.

Per vedere un esempio consultare il file nominato *"Template dataset"*