PERCORSO AGENZIA DELLE ENTRATE

Statistica descrittiva: Concetti introduttivi

Introduzione

Benvenuti!

In questa lezione tratteremo le principali definizioni della statistica descrittiva, introducendo le distribuzioni di frequenza e alcuni dei principali indici di posizione e medie analitiche.

In particolare, andremo ad approfondire che cosa si intende per:

- carattere o variabile statistica
- popolazione
- unità statistica
- modalità

Successivamente andremo a definire le varie tipologie dei caratteri o variabili che potremmo incontrare nelle diverse analisi statistiche. Infine, una volta introdotte le distribuzioni di frequenza, passeremo a definire i principali indici di posizione come:

- la moda
- la mediana
- i quartili
- per poi introdurre medie analitiche come ad esempio la media aritmetica

Bene, non ci resta che cominciare…

Per iniziare...

Iniziamo definendo che cosa sia la statistica. Ebbene la statistica è uno strumento conoscitivo atto ad analizzare in termini quantitativi un fenomeno collettivo. In pratica la statistica fornisce gli strumenti metodologici quantitativi per sintetizzare numericamente le informazioni derivanti da un elevato numero di osservazioni di fatti che noi riusciamo a percepire singolarmente.

La statistica può essere suddivisa in due grandi branche:

- la statistica descrittiva
- la statistica inferenziale

In questa prima parte, noi parleremo di statistica descrittiva, cioè dell'insieme delle metodologie atte a descrivere e riassumere le caratteristiche della distribuzione di una o più variabili statistiche, mentre nella seconda lezione si parlerà di statistica inferenziale, che ha gli stessi obiettivi della statistica descrittiva, ma li persegue rilevando solo una parte della popolazione che è chiamata campione. Verranno quindi introdotte tecniche che consentiranno di utilizzare le informazioni misurate sul campione e renderle valide per l'intera popolazione.

Classificazione dei caratteri

Passiamo ora alla terminologia essenziale di base. Partiamo con il definire che cosa si intende per carattere o variabile statistica.

Un *carattere* è il fenomeno che siamo interessati ad analizzare come ad esempio l'età di una popolazione, il titolo di studio o il reddito. Ognuno di questi fenomeni viene osservato su un *collettivo o popolazione*. Un collettivo è l'insieme di unità statistiche. Viene detta *unità statistica* l'oggetto elementare al quale facciamo riferimento come osservazione. Su ciascuna unità statistica, osserveremo il carattere che si presenta attraverso quelle che vengono chiamate *modalità*, cioè il modo con cui si presenta un carattere.

Ogni carattere ha delle caratteristiche distintive. In particolare, possiamo classificare le diverse variabili statistiche in due grandi famiglie:

- caratteri qualitativi
- caratteri quantitativi

Un carattere è classificato qualitativo se le modalità assunte dalle diverse unità statistiche esprimono un attributo o una qualità. I caratteri quantitativi sono invece quei caratteri le cui modalità sono numeriche ed esprimono una misura, una quantità.

I caratteri qualitativi a loro volta possono essere suddivisi in due categorie:

- qualitativi sconnessi
- qualitativi ordinati

Facendo riferimento alle modalità assunte dai diversi caratteri, un carattere qualitativo sconnesso è un carattere le cui modalità non sono ordinate. Non esiste, quindi, un ordinamento implicito tra le modalità. Per i caratteri qualitativi ordinati invece è possibile osservare un ordine naturale tra le diverse modalità.

Per i caratteri quantitativi invece possiamo considerare due ulteriori classificazioni in:

- caratteri quantitativi discreti
- caratteri quantitativi continui

I caratteri quantitativi discreti vengono definiti discreti laddove le modalità assunte dal carattere sulle diverse unità statistiche siano numerabili, generalmente in ordine finito. Un carattere viene detto quantitativo continuo se le sue modalità invece possono essere qualsiasi sull'intera retta dei numeri reali.

È importante distinguere in modo appropriato la tipologia dei diversi caratteri. Vedremo che alcune operazioni tra modalità si possono o non possono fare a seconda delle diverse tipologie di caratteri. Andando avanti durante il corso vedremo che alcuni indici potranno essere calcolati solamente per alcune tipologie di carattere.

In particolare, possiamo notare che:

- la relazione di uguaglianza/disuguaglianza tra due modalità può essere definita sempre per qualsiasi tipologia di carattere

- l'ordinamento tra modalità non è possibile definirlo per i caratteri qualitativi sconnessi, mentre lo è per caratteri qualitativi ordinati e ovviamente per i caratteri quantitativi, in cui le modalità numeriche sono intrinsecamente ordinati
- le relazioni ed operazioni matematiche tra le modalità, come ad esempio l'addizione o la sottrazione, sono possibili solo laddove le modalità siano numeriche e, di conseguenza, solo in presenza di un carattere quantitativo

ESEMPI: Vediamo insieme ora alcuni esempi di caratteri sia qualitativi, che quantitativi. Cominciamo con il carattere *sesso*. Per definire se il carattere *sesso* sia un carattere qualitativo o quantitativo dobbiamo vedere quali sono le modalità, il modo con cui si presenta il carattere. Quali sono le modalità assunte dal carattere *sesso*? Sappiamo tutti quanti che maschio e femmina sono le modalità con cui osserviamo il carattere sesso sulle nostre unità statistiche. Maschio e femmina sono ovviamente degli attributi, delle qualità, quindi il carattere *sesso* sicuramente è un carattere qualitativo. Dobbiamo ora decidere se il carattere *sesso* sia un carattere qualitativo sconnesso o ordinato. È possibile ordinare le modalità? Viene prima maschio o viene prima femmina? Non esiste un ordinamento delle modalità, di conseguenza il carattere sesso verrà classificato come qualitativo sconnesso.

Lo stesso può essere fatto per il carattere *religione* in cui le modalità, ad esempio cattolica, musulmana, induista, sono delle qualità, ma non contengono un ordine intrinseco. Passiamo ora a classificare il *titolo di studio*. Le possibili modalità del titolo di studio sono analfabeta, elementare, licenza media, licenza superiore, laurea, dottorato. Come vedete sono delle qualità e, quindi, anche il titolo di studio è un carattere qualitativo. Però, in questo caso, è un carattere qualitativo ordinato perché le modalità osservate sulle diverse unità statistiche sono ordinate, cioè per ottenere la licenza superiore è necessario aver ottenuto la licenza media, e prima ancora la licenza elementare. Quindi come vediamo c'è un ordinamento tra le modalità. Lo stesso per il grado di soddisfazione con modalità da per nulla soddisfatto, poco soddisfatto, molto soddisfatto del tutto soddisfatto. Sono delle qualità che però hanno un ordinamento. Quindi anche il grado di soddisfazione farà parte dei qualitativi ordinati.

Continuiamo considerando il numero di figli o il numero di pezzi prodotti da un macchinario. Quali sono le modalità che possiamo osservare sulle diverse unità statistiche? Le possibili modalità sono 0,1,2,3, i numeri interi. Quindi, le modalità sono finite e numerabili. Lo stesso vale per il voto ad un esame, che può essere 18, 19, 20, fino a 30. Anche in questo caso, le modalità assunte dal carattere voto ad un esame sono solo determinati valori numerici. Quindi questi tre caratteri vengono classificati come quantitativi discreti. Infine, guardiamo al peso, al reddito, all'altezza. Sono caratteri quantitativi continui perché il reddito di una persona può essere un qualsiasi numero positivo, lo stesso per peso e altezza.

Distribuzioni di frequenza

Una volta classificati correttamente i caratteri, possiamo passare a *sintetizzare* le informazioni contenute in ciascun carattere. Questo avviene attraverso l'utilizzo di *distribuzioni di frequenza*. Le distribuzioni di frequenza sono lo strumento più utilizzato per sintetizzare le informazioni contenute da un carattere. La sintesi avviene tramite la costruzione di tabelle riassuntive delle informazioni contenute in un carattere.

- La distribuzione di frequenze semplice associa a ciascuna modalità del carattere che indicheremo con la lettera X maiuscola (indifferentemente se sia esso qualitativo o quantitativo) le corrispondenti frequenze

Quindi, detto X maiuscolo il carattere, indicheremo con x1, x2, la prima e la seconda modalità del carattere X a cui associamo *le frequenze assolute* n1 e n2. In altre parole, n1 è il numero di volte che si presenta la modalità x1, n2 è il

numero di volte che si presenta la modalità x2. La somma delle frequenze assolute deve restituire la numerosità del collettivo, che indicheremo con la lettera n. Le frequenze assolute sono sicuramente il primo passaggio per sintetizzare le informazioni a disposizione, ma non consentono il confronto tra collettivi di numerosità diversa. Per poter effettuare dei confronti tra collettivi di numerosità diversa dovremo far ricorso alle *frequenze relative* o alle *frequenze percentuali*.

- La frequenza relativa è data dal rapporto tra la frequenza assoluta e la numerosità del collettivo.
- La frequenza percentuale è data dalla frequenza relativa moltiplicata per 100

Le frequenze relative percentuali non sono soltanto utili per eseguire confronti tra distribuzioni in collettivi diversi, ma ci consentono anche di capire l'importanza di una determinata modalità all'interno del collettivo.

- La somma delle frequenze relative è sempre pari a uno
- La somma delle frequenze percentuali è sempre pari a 100

Distribuzioni di frequenze cumulate

Accanto alle frequenze assolute, relative e percentuali esistono anche le *frequenze cumulate*: le frequenze assolute cumulate, le frequenze relative cumulate, le frequenze percentuali cumulate. Dato un qualsiasi carattere con K grande numero di modalità ordinate in senso crescente, la frequenza cumulata, che sia essa assoluta, relativa, percentuale è data dalla somma delle frequenze. Le distribuzioni di frequenze cumulate sono utili per capire la frequenza con cui si presentano modalità di ordine inferiore o uguale ad una certa modalità.

- L'ultima delle frequenze relative cumulate è pari a 1
- L'ultima delle frequenze percentuali cumulate pari a 100
- L'ultima delle frequenze assolute cumulate è pari alla numerosità del collettivo

Le frequenze cumulate possono essere calcolate e hanno senso solamente laddove le modalità siano almeno *ordinate*. Questo fa sì che le frequenze cumulate non abbiano alcun senso per i caratteri qualitativi sconnessi, perché ricordiamo che per i caratteri qualitativi sconnessi non è possibile determinare l'ordinamento tra le modalità.

Distribuzioni di caratteri suddivisi in classi

Un ulteriore aspetto che merita attenzione riguarda i *caratteri quantitativi continui*. Nei caratteri quantitativi continui spesso il numero di modalità osservate è elevato e si possono quindi avere notevoli difficoltà nella comprensione dei dati osservati o nella loro sintesi. Per rendere la sintesi più fruibile all'utilizzatore finale, le modalità vengono suddivise in intervalli tra loro disgiunti che vengono chiamate *classi di modalità*. Come è possibile definire le classi?

- Le classi devono essere in numero abbastanza piccolo da fornire una sintesi adeguata, ma sufficientemente grande da mantenere l'informazione contenuta dal carattere con un livello accettabile di dettaglio
- Le classi devono comprendere tutte le possibili modalità del carattere e se possibile avere la stessa ampiezza, tuttavia nelle applicazioni reali spesso le classi non hanno la stessa ampiezza e di questo noi dovremo tenere conto nelle varie analisi

Statistica descrittiva: Le medie

La moda

Introduciamo ora il primo indice di posizione. Si tratta della *moda*. La moda è l'indice più semplice da calcolare per una qualsiasi distribuzione di frequenza.

- La moda di un collettivo non è nient'altro che la *modalità prevalente del carattere,* ossia quella cui è associata la massima frequenza che sia essa assoluta, relativa o percentuale

Consideriamo il seguente esempio in cui abbiamo un collettivo di 220 donne su cui abbiamo misurato il carattere numero di figli e abbiamo osservato che il numero di figli su queste 220 donne può essere 0, 1, 2, 3, 4. Qual è la modalità a cui è associata la frequenza, in questo caso assoluta, più alta? La frequenza assoluta più alta è 100, di conseguenza la modalità a cui è associata la frequenza più alta è 2. La moda può essere calcolata per qualsiasi tipologia di carattere, in quanto per il calcolo della moda è necessario solamente conoscere le frequenze.

Quando abbiamo a che fare con un carattere quantitativo suddiviso in classi dobbiamo porre più attenzione nel calcolo della moda. Nel caso in cui le classi avessero ampiezza diversa, la classe modale, cioè la moda per carattere suddivisi in classi, è individuata come la classe alla quale corrisponde *la densità* più alta. Che cos'è la densità?

- La densità è il rapporto tra frequenza ed ampiezza della classe

Attraverso la densità pesiamo le frequenze per l'ampiezza della classe stessa.

La moda ha delle proprietà:

- la moda è interna, è sempre compresa tra la modalità più piccola e la modalità più grande
- può essere calcolata per caratteri di qualsiasi tipo, sia qualitativi che quantitativi
- la moda non è necessariamente unica, possiamo avere distribuzioni con due mode dette bimodali, con tre mode dette trimodali e così via

Mediana e quantili

- La *mediana* è la modalità che divide in due parti di uguale numerosità il collettivo delle unità ordinate in senso non decrescente

Come faccio a identificare l'unità che si trova al centro della distribuzione ordinata? Nel caso in cui la numerosità del collettivo fosse *dispari*, l'unità centrale si trova nella posizione (n+1)/2, dove n è la numerosità del collettivo. Nel caso in cui la numerosità del collettivo fosse *pari*, avremo due unità centrali, che si trovano alla posizione n/2 e n/2+1. Quindi avremo che la mediana sarà data dalle modalità in corrispondenza di queste due unità centrali. Nel caso di distribuzioni di frequenza, l'identificazione della mediana, ed in generale di qualsiasi indice di posizione, avviene attraverso l'utilizzo delle frequenze relative cumulate.

- La mediana, in distribuzioni di frequenza, è la modalità in corrispondenza della prima frequenza relativa cumulata maggiore uguale di 0,5

Nel caso di distribuzioni di caratteri quantitativi suddivisi in classi c'è un passaggio aggiuntivo da fare per individuare la mediana. La prima cosa da individuare è la classe mediana pari anche in questo caso alla modalità in corrispondenza della prima frequenza relativa cumulata maggiore di 0,5. Successivamente, all'interno della classe mediana andremo a identificare il valore specifico della mediana attraverso una formula specifica.

La mediana ha delle proprietà:

- è interna, è sempre compresa tra modalità minima e massima
- il numero di scarti positivi è uguale al numero di scarti negativi
- è possibile calcolarla per tutte le tipologie di caratteri tranne i qualitativi sconnessi, perché per il calcolo della mediana le modalità devono essere ordinate in senso non decrescente e per i caratteri qualitativi sconnessi non è possibile ordinare le modalità

Oltre alla mediana è possibile calcolare diversi indici di posizione che prendono il nome di *quantili*. Tra i quantili, i *quartili* sono gli indici di posizione più utilizzati.

- Il primo quartile è la modalità che separa il 25% dei valori più piccoli dal 75% dei valori più grandi
- La mediana viene anche detta secondo quartile
- Il terzo quartile è il valore che separa il 75% dei valori più piccoli dal 25 dei valori più grandi

La media aritmetica

La *media aritmetica* è la media analitica più conosciuta.

- La media aritmetica della distribuzione di un carattere è data dalla somma delle modalità assunte dalle unità del collettivo divisa per la numerosità del collettivo stesso.
- La media aritmetica, siccome implica operazioni matematiche tra le modalità, può essere calcolata solamente per carattere di tipo quantitativo.
- La media aritmetica ha delle proprietà molto importanti:
    - la somma degli scarti dalla media aritmetica è nulla
    - la somma dei quadrati degli scarti da un qualsiasi valore è minima quando questo valore è la media
    - la media aritmetica è interna, cioè sempre compresa tra il minimo e il massimo
    - la media aritmetica è invariante per trasformazioni lineari del tipo y = a + bx
    - la media aritmetica gode della proprietà associativa

Conclusioni

Bene, con questo siamo giunti alla fine di questa prima video lezione.

Ti ricordo che abbiamo introdotto:

- la classificazione dei caratteri
- le distribuzioni di frequenza e le distribuzioni di frequenza cumulate
- infine, abbiamo approfondito gli indici di posizione (media, mediana e quantili) con le rispettive proprietà

Grazie per l'attenzione!