1 Presentazione

Una caratterristica del corso sarà l'esposizione di teoria (elementi di calcolo delle probabilità) e pratica (analisi statistiche di dati tramite programma R) presentate simultaneamente o comunque in modo alternato, non sequenziale.

Partiamo da una tabella di dati che vogliamo "capire", "leggere", "esaminare". La tabella (file *indicatori_benessere* di tipo *Excel,* foglio 1) riporta, relativamente all'anno 2002, relativamente ad ogni regione italiana, i valori di certi indicatori di benessere sociale: PLIC = posti letto in istituti di cura, SC = spese mensili complessive, SA/SC = rapporto tra spese alimentari e spese complessive, TD = tasso di disoccupazione, TMI = tasso di mortalità infantile. I dati sono stati presi dal sito dell'ISTAT:

$$http://www.istat.it/dati/$$

dove abbiamo aperto *Sistema di indicatori territoriali*:

$$http://sitis.istat.it/sitis/html/index.htm$$

dove poi si apre *Consulta i dati.* Provate ad esempio a rintracciare i dati sul tasso di disoccupazione.

La mole di dati è non indifferente. Pur potendo estrarre alcune informazioni tramite lettura accurata dei dati, è quasi impossibile avere una visione d'insieme.

DOMANDA 1: è possibile avere una visione d'insieme di questi dati, possibilmente grafica?

Altre domande rilevanti emergeranno strada facendo. Vediamo una soluzione non banale a questa prima domanda. Vedremo che non è semplice capire i principi dietro questa soluzione. Questa sarà la motivazione di una serie di approfondimenti teorici che effettueremo subito dopo la visualizzazione.

Rispondiamo alla domanda 1 tramite software. I dati sono riportati su un normale foglio Excel. Tale programma è però troppo elementare per realizzare quello che vedremo. Usiamo il programma R, scaricabile in rete dal sito:

$$http://www.r-project.org/$$

Scaricamento ed installazione sono abbastanza semplici. L'uso di R però non è immediato. Eseguiamo quindi alla cieca i comandi che ora vedremo, rinviando il momento in cui li capiremo a fondo.

Le fasi sono:

- Conviene riportare i dati ad una stessa unità di misura, altrimenti la grandezza dei valori SC rispetto agli altri può falsare le indagini e le rappresentazioni grafiche. Pertanto, sul foglio 1 del file calcoliamo media e deviazione di ogni indicatore, poi costruiamo a fianco una tabella simile con i valori standardizzati (se x è un valore di un certo indicatore e $\overline{x}$, s sono media e deviazione standard campionarie, $\frac{x-\overline{x}}{s}$ è il valore standardizzato).

- Conviene avere nomi brevi per le regioni (così come per gli indicatori). Abbreviamoli. Attenzione: ogni nome non deve contenere spazi, altrimenti questi verranno equivocati in formato testo (interpretati come tabulazioni).

- Dobbiamo salvare i dati Excel in formato testo per poterli acquisire con R. Copiamoli su Foglio 2. Poi salviamo come testo (delimitato da tabulazione). Teniamo nota della cartella in cui sono salvati. Chiudiamo Excel.

- Possiamo verificare l'aspetto di questi dati aprendoli con Blocco note.

- Osservazione: i dati standardizzati hanno un segno, a differenza di quelli originari. Il segno fornisce già un'indicazione interpretativa.

- Apriamo R. Apriamo un file di testo in cui avremo elencato alcuni comandi frequenti di R, noiosi da ricordare o ritrovare ogni volta con l'help. In questo file ci sarà il comando:

$$\begin{aligned} &IB <- read.table(file = \\ &\text{“}C:/...percorso.../indicatori_benessere.txt\text{”}, \\ &header = TRUE) \end{aligned}$$

  che carica i dati dal file .txt al foglio di R. I dati sono ora in una tabella di nome IB. Se ad esempio scriviamo IB compare la tabella.

- scriviamo

$$pca <- princomp(IB)$$

  poi

$$biplot(pca)$$

- Compare un grafico con i nomi delle regioni, due assi cartesiani denominati Comp.1 e Comp.2, e tanti vettori rossi quanti gli indicatori. L'interpretazione completa di questo grafico è complessa e costituirà uno dei nostri primi scopi del corso. Tuttavia, a livello molto superficiale, possiamo vedere in questo grafico vari elementi di immediata lettura. La direzione SC è opposta alle direzioni TD ed SA/SC, segno che questi sono indicatori *correlati negativamente* (l'aumentare dell'uno corrisponde al diminuire degli altri). Interessante è la correlazione negativa tra le spese complessive e la percentuale dedicata a quelle alimentari.

- Gli altri due indicatori sono abbastanza perpendicolari ai precedenti, segno che sono abbastanza scorrelati dai precedenti, però con una lieve correlazione positiva nelle direzioni che ci si aspetta (maggiori posti letto ove ci sono maggiori spese, maggior mortalità infantile ove c'è più disoccupazione e prevalenza di spese alimentari). Invece, PLIC e TM sono negativamente correlati, come è naturale.

- La posizione delle regioni va letta rispetto al verso crescente delle frecce. Per esempio, Veneto ed EmiliaRomagna sono le due regioni con maggiori spese, Calabria e Campania quelle con meno.

- Le regioni compaiono raggruppate o comunque posizionate in modo "istruttivo", che suggerisce delle considerazioni (per quanto naturali e più o meno note, visto che sappiamo abbastanza bene la situazione nazionale del benessere). Indicativamente, nel grafico, la regione di maggior benessere è quella nella direzione positiva di SC, con un po' di spostamento verso PLIC. In tale regione di alto benessere si trovano fortemente raggruppate varie regioni, che pertanto risultano molto simili. Alcune regioni come Lazio, Friuli e Liguria non eccellono in SC quanto quelle del gruppo più forte, ma eccellono in PLIC, indicazione della buona cura sanitaria, nonostante un tenore di vita medio più modesto di Lombardia ecc. Altre regioni invece sono negative per certi versi, con accentuazioni per alcune. Tra le note negative, rilevante è il raggruppamento particolarmente negativo di Puglia, Campania, Sicilia e Basilicata, rispetto ad altre regioni del sud o insulari, che nell'immaginario collettivo potremmo credere ad esse simili (indicazione forse di uno sforzo di miglioramento). Per inciso, questa indagine è un rifacimento alla data del 2002 di un'indagine più datata

proposta nel bel testo di Giuseppe D'Angelo, *Elementi di Analisi Statistica Spaziale*, Monduzzi Editore (in particolare il Capitolo 4.2), il cui spirito da studioso dei problemi del Sud d'Italia era quello di evidenziare quantitativamente gli squilibri regionali in Italia ed il divario tra Nord e Sud.

- Infine, esiste un modo automatico ed oggettivo di individuare delle classi di regioni simili dal punto di vista del benessere (misurato attraverso questi indicatori)? Eseguiamo

$$clust3 < -pam(IB, 3)$$
$$plot(clust3)$$

e vediamo lo stesso grafico suddiviso in tre *cluster*. Siccome la Valle d'Aosta risulta un po' forzata, proviamo con 4 cluster:

$$clust4 < -pam(IB, 4)$$
$$plot(clust4)$$

Ora la suddivisione è molto naturale.

DOMANDA 2: quali principi matematici sono alla base della raffigurazione dei dati vista ora?

Per rispondere, partiremo un po' dall'inizio, ricordando alcuni concetti di base di calcolo delle probabilità, alcune distribuzioni importanti (ampliando un po' rispetto allo stretto necessario rispetto all'esempio visto ora), arrivando a descrivere il metodo delle *componenti principali.*

Exercise 1 (obbligatorio) *Rintracciare sul sito ISTAT alcuni indicatori (regionali) relativi a ricerca e sviluppo, sia di tipo positivo sia negativo e ripetere l'analisi precedente.*

Exercise 2 (suggerito) *Rintracciare, se esistono, degli indicatori che favoriscano regioni diverse dalle solite regioni del centro-nord (Lombardia, Emilia-Romagna, ecc.).*

Prima di passare alle basi teoriche, è però ragionevole tentare di capire intuitivamente cosa ha fatto il software per darci questa raffigurazione.

Ogni regione italiana viene vista come un punto in uno spazio a 5 dimensioni (5 è il numero di indicatori di benessere adottati). Ad esempio, Piemonte è il punto

$$(0.088, 0.471, -0.707, -0.607, -0.395)\,.$$

Le diverse regioni vengono viste quindi come una *nuvoletta* di punti in uno spazio a 5 dimensioni (non raffigurabile su un foglio bidimensionale). Immaginiamo che questa nuvoletta abbia una forma approssimativamente *ellissoidale* - se questo non è vero, il software esegue compunque meccanicamente delle operazioni che sono suggerite dal caso ellissoidale. Il SW cerca gli assi dell'ellissoide, che sono 5 assi ortogonali nello spazio a 5 dimensioni, diversi però dagli assi coordinati. Poi, il SW, esegue una *rotazione* e legge i dati nella nuova base formata dagli assi dell'ellissoide. Se fossimo in due dimensioni invece che 5, il SW ci mostrerebbe un grafico in cui l'asse delle x è l'asse principale dell'ellissoide, l'asse delle y è il secondo asse dell'ellissoide, ed i punti sono le regioni viste in questo nuovo sistema di assi coordinati; inoltre, le frecce rosse sarebbero i vecchi assi (quelli del sistema di coordinate originario) visti nel nuovo sistema di assi. Ad esempio, se il primo asse del sistema originario corrispondeva ai valori PLIC, vedremo nella raffigurazione una freccia rossa PLIC che è quell'asse visto nel nuovo sistema degli assi dell'ellisoide.

In due dimensioni non c'è un gran vantaggio grafico nell'operare questa rotazione, questa trasformazione di punto di vista: infatti, i dati raffigurati nel sistema di assi originario già apparirebbero come una nuvoletta ellissoidale opportunemente inclinata e potremmo fare varie considerazioni intuitive su quel grafico. Se però siano in più di 2 dimensioni, dove è impossibile raffigurare i dati originari, il SW raffigura solo i primi due assi dell'ellissoide, quelli di maggior lunghezza, che chiameremo *assi principali.* Poi raffigura la proiezione dei punti (le regioni) e degli assi originari (gli indicatori) sul piano di tali assi principali. Così facendo, *il SW ha individuato il piano che meglio distingue i dati*, essendo il piano in cui i dati hanno la *dispersione maggiore.* Questo è il potere di visualizzazione del metodo delle **Componenti Principali**: trovare un sistema di coordinate in cui i dati siano più separati possibile e raffigurarli rispetto ai due assi di maggior separazione. Questo aumenta la nostra capacità visiva di giudizio e permette ad esempio una divisione in cluster più efficace. Gli stessi dati, raffigurati rispetto ad un piano qualsiasi, potrebero apparire forzatamente concentrati tutti insieme e non "intellegibili".

Si potrebbero aggiungere varie altre considerazioni, come ad esempio il fatto che il primo asse principale, essendo quello in cui lo sparpagliamento dei dati è maggiore, è quello che *spiega in modo più netto la diversità* tra le regioni. E' quindi quello che meglio ci aiuta nel giudizio, nella classificazione. Per questo, a volte viene interpretato come il *miglior criterio di giudizio del benessere.* Se ad esempio il primo asse principale e_1 ha coordinate 0.2 rispetto a PLIC e 0.8 rispetto a SC, diremo che il miglior indicatore di benessere non è SC oppure PLIC, ma una grandezza di misurazione meno diretta data da

$$0.8 \cdot SC + 0.2 \cdot PLIC.$$

Per essere poi più precisi, dovemmo aggiungere il peso dato dalle coordinate, negative, rispetto agli altri indicatori.

PROBLEMA: come fa il SW a calcolare gli assi principali a partire dai dati sperimentali?

Potremmo arrangiare una spegazione senza basi di probabilità, basata solo su intuizioni geometriche. Però questo è un primo problema in cui senza dubbio l'inquadramento probabilistico rende ovvia la risposta.