Descrizione Progetto Python:

Implementazione del Test del Chi-Quadro e/o del Test ANOVA

Michela Rossi

IN400 - Modulo A

Introduzione

Nel mio progetto in Python vorrei implementare uno tra due test statistici fondamentali (o, eventualmente, entrambi):

- **Test del Chi-Quadro:** utilizzato per verificare l'indipendenza tra due variabili categoriali.

  Esempio. Supponiamo di voler verificare se c'è una relazione tra il genere (Maschio/Femmina) e la preferenza per tre diversi tipi di snack (Cioccolato, Patatine, Biscotti).

- **Test ANOVA:** impiegato per determinare se le medie di due o più popolazioni distribuite normalmente, con uguale varianza non nota, sono uguali.

  Esempio. Un insegnante vuole confrontare il rendimento scolastico medio di studenti che hanno seguito tre diversi metodi di studio: individuale, di gruppo, e con tutor; quindi suddivide i risultati del test che ha loro sottoposto in tre gruppi e si vuole confrontare la media di queste ultime.

Si potrebbero raccogliere i dati in un file excel e importare questo file all'interno del codice oppure dare gli input da console.

Analisi Specifica

Test del Chi-Quadro

Nel test del χ^2 si costruisce la statistica test, che avrà appunto la distribuzione di una χ^2. In particolare, si dimostra che se si hanno due variabili da analizzare (nell'esempio di sopra genere e preferenze di snack), supponendo che la prima variabile sia suddivisa in ℓ categorie (2 nell'esempio di sopra) e la seconda in m categorie (3 nell'esempio di sopra), la statistica sarà una $\chi^2_{(\ell-1)(m-1)}$ ed è definita come segue:

$$\chi^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{m} \frac{(N_{ij} - \frac{r_i c_j}{n})^2}{\frac{r_i c_j}{n}}$$

dove

- N_{ij}= numero di osservazioni della $i-esima$ categoria per la prima variabile e $j-esima$ categoria per la seconda variabile.
- n= numero totale di osservazioni
- r_i= numero di elementi nella $i-esima$ riga
- c_j= numero di elementi nella $j-esima$ colonna

Per chiarire, facciamo un esempio pratico:

	Cioccolato	Patatine	Biscotti	Totale
Maschi	20	30	10	60
Femmine	40	50	60	150
Totale	60	80	70	210

in questo esempio si avrà:

- $\ell = 2$ $m = 3$
- $n = 210$
- $N_{2,1} = 40$

- $r_1 = 60$

- $c_3 = 70$

Dopodiché si sceglie un livello di significatività α (tipicamente 0.05) cioè una soglia numerica che definisce il margine di errore che si è disposti a "tollerare" prendendo una decisione basata sui dati. Assumiamo che le due variabili siano indipendenti, indichiamo con H_0 tale assunzione e con H_1 l'assunzione contraria. La significatività di un test (in generale) rappresenta la probabilità di assumere H_1 come vero quando H_1 è falso.
Nel test del χ^2 si valuta $\chi^2 = \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{(N_{ij} - \frac{r_i c_j}{n})^2}{\frac{r_i c_j}{n}}$; se tale quantità è maggiore del quantile α della $\chi^2_{(\ell-1)(m-1)}$ (indicato con $\chi^2_{(\ell-1)(m-1),\alpha}$) allora si concluderà che le variabili non sono indipendenti.

Il programma Python seguirà i seguenti passi:

1. Acquisizione delle osservazioni.

2. Calcolo della statistica χ^2.

3. Decisione sull'indipendenza in base al livello di significatività scelto.

Test ANOVA

Il test di ANOVA si basa anch'esso sul calcolo di una statistica che sarà distribuita come una Fisher. Supponiamo di avere m gruppi (nel caso dell'esempio 3) e ogni gruppo ha n osservazioni. La distribuzione di Fisher da calcolare è:

$$F = \frac{\frac{SSb}{m-1}}{\frac{SSw}{mn-m}}$$

Dove definiamo

- X_{ij} l'osservazione $j-esima$ del gruppo $i-esimo$

- $\bar{X}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n}$

- $\bar{X} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} X_{ij}}{nm}$

- "Sum of Squares Between" $SSb = n\sum_{i=1}^{m}(X_i - \bar{X})^2$
- "Sum of Squares Within" $SSw = \sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ij} - \bar{X}_i)^2$

Anche in questo caso, si sceglie un livello di significatività α, si valuta la quantità $\frac{\frac{SSb}{m-1}}{\frac{SSw}{mn-m}}$ e se tale quantità è maggiore del quantile α della $F_{(m-1,m(n-1))}$ (indicato con $F_{(m-1,m(n-1)),\alpha}$) allora si concluderà che le variabili non hanno la stessa media.

Il programma Python includerà:

1. Lettura dei dati per più gruppi.
2. Calcolo della statistica F.
3. Interpretazione del risultato per determinare se le medie dei gruppi sono diverse.