Gestione e Analisi dei Dati

Lezione 7

Introduzione all'analisi dei dati

Popolazioni, campioni, unità statistica, variabili

Studio e rappresentazione della frequenza

Obiettivi del modulo

- Scopo del corso è la presentazione di metodi per la estrazione di informazioni da insiemi di dati
  - Le informazioni possono essere
    - di sintesi (p.e. la media di un insieme di valori numerici)
    - di andamento (p.e. il grafico dei valori di una variabile nel tempo)
- La base concettuale è la *statistica (descrittiva* e *qualche elemento di inferenziale)*
- Lo strumento operativo è il foglio elettronico (spreadsheet), in particolare OpenOffice Calc

Tipiche domande

- Che incassi sono stati registrati l'anno passato per ciascuna regione e ciascuna categoria di prodotto?
- Che correlazione esiste tra l'andamento dei titoli azionari dei produttori di PC e i profitti trimestrali lungo gli ultimi 5 anni?
- Quali sono gli ordini che massimizzano gli incassi?
- Quale di due nuove terapie risulterà in una diminuzione della durata media di un ricovero?
- Che rapporto c'è tra i profitti realizzati con spedizioni di meno di dieci elementi e quelli realizzati con spedizioni di piu` di dieci elementi?

Outline

- Elementi di statistica descrittiva e uso di OpenOffice Calc per i calcoli:
  - Variabili e tabelle
  - Manipolazione di tabelle (sort ecc.)
  - Distribuzioni di frequenze e grafici
  - Statistica descrittiva per una singola variabile (media, varianza, ...)
  - Statistica descrittiva: misure di associazione.
- Data warehouses e OLAP (On Line Analitycal Processing)
- OLAP in oCalc: pivot tables
- Elementi di data cleaning

Concetti di base

- Una *popolazione* (population)include tutti gli oggetti di interesse
- Esempi di popolazione:
  - tutti i potenziali votanti per l'elezione del rettore
  - tutti gli abbonati alla RAI
  - tutte le fatture ricevute dal Dipartimento di Informatica nel 2003
- Un *campione* (sample) è un sottoinsieme di una popolazione, spesso scelto in modo casuale e preferibilmente rappresentativo dell'intera popolazione.

Variabili e Osservazioni

- Una *unità statistica* {*osservazione, caso*} è una ennupla di valori che caratterizza un elemento di una popolazione o di un campione
- Una *variabile* {*carattere, attributo, campo*} è l'identificatore (nome) di uno dei valori dell'osservazione
- Se una popolazione (campione) è rappresentata in forma di *tabella* le righe della tabella sono le osservazioni, i nomi delle colonne sono le variabili e il contenuto di ciascuna riga è la lista dei valori delle variabili

Un esempio

Name	Gender	DomesticGross	ForeignGross	Salary
Angela Bassett	F	32	17	2,5
Jessica Lange	F	21	27	2,5
Winona Ryder	F	36	30	4
Michelle Pfeiffer	F	66	31	10
Whoopi Goldberg	F	32	33	10
Emma Thompson	F	26	44	3
Julia Roberts	F	57	47	12
Sharon Stone	F	32	47	6
Meryl Streep	F	34	47	4,5
Susan Sarandon	F	38	49	3
Nicole Kidman	F	55	51	4
Holly Hunter	F	51	53	2,5
Meg Ryan	F	43	55	8,5
Andie Macdowell	F	26	75	2
Jodie Foster	F	62	85	9

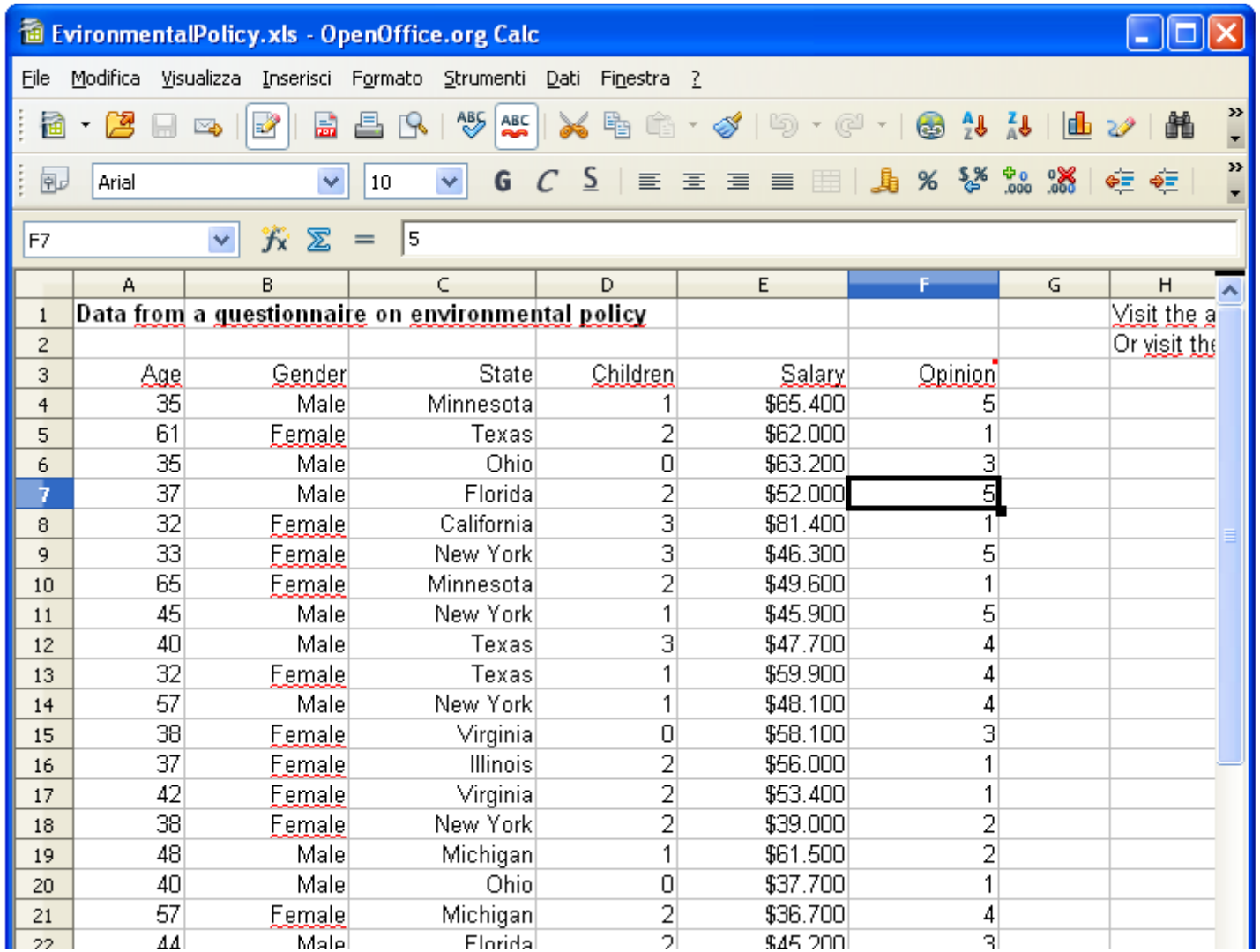
Tipi di Valori (1)

- Possiamo classificare le variabili in base alla tipologia dei valori che possono assumere;
- Distinguiamo tra:
  - Variabili numeriche *(quantitative)*,se sui valori e` possibile compiere un insieme significativo di operazioni aritmetiche;
  - Variabili categoriche *(qualitative)*, altrimenti

Tipi di Valori (2)

- Le variabili numeriche possono essere:
  - discrete, se i valori possono essere contati
  - continue, se sono il risultato di una misura continua
- Le variabili categoriche possono essere:
  - ordinali, se esiste un ordine naturale sui possibili valori (es. giudizi scolastici: insufficiente, sufficiente ecc.)
  - nominali, altrimenti (es. colori)

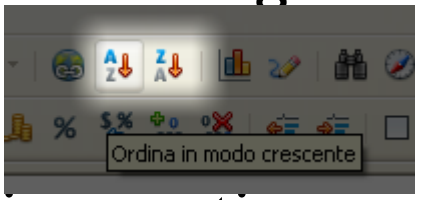
Esempio



	A	B	C	D	E	F	G	H
1	Data from a questionnaire on environmental policy							Visit the a
2								Or visit the
3	Age	Gender	State	Children	Salary	Opinion		
4	35	Male	Minnesota	1	$65.400	5		
5	61	Female	Texas	2	$62.000	1		
6	35	Male	Ohio	0	$63.200	3		
7	37	Male	Florida	2	$52.000	5		
8	32	Female	California	3	$81.400	1		
9	33	Female	New York	3	$46.300	5		
10	65	Female	Minnesota	2	$49.600	1		
11	45	Male	New York	1	$45.900	5		
12	40	Male	Texas	3	$47.700	4		
13	32	Female	Texas	1	$59.900	4		
14	57	Male	New York	1	$48.100	4		
15	38	Female	Virginia	0	$58.100	3		
16	37	Female	Illinois	2	$56.000	1		
17	42	Female	Virginia	2	$53.400	1		
18	38	Female	New York	2	$39.000	2		
19	48	Male	Michigan	1	$61.500	2		
20	40	Male	Ohio	0	$37.700	1		
21	57	Female	Michigan	2	$36.700	4		
22	44	Male	Florida	2	$45.200	3		

Interpretatione dei dati: ordinamento

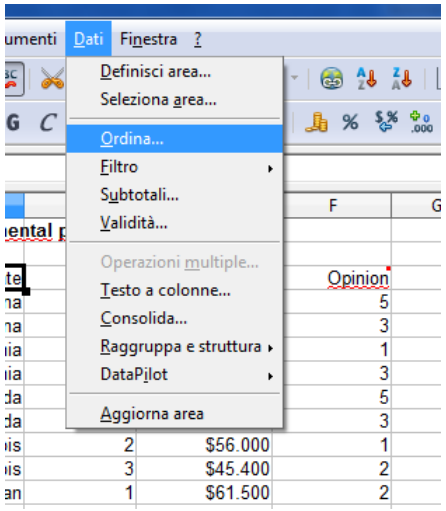
- Il più potente e il più semplice strumento di elaborazione di una tabella è l'*ordinamento*
- In oCalc è possibile ordinare la tabella scegliendo:
  - la colonna (l'attributo)
  - la modalità: crescente, decrescente

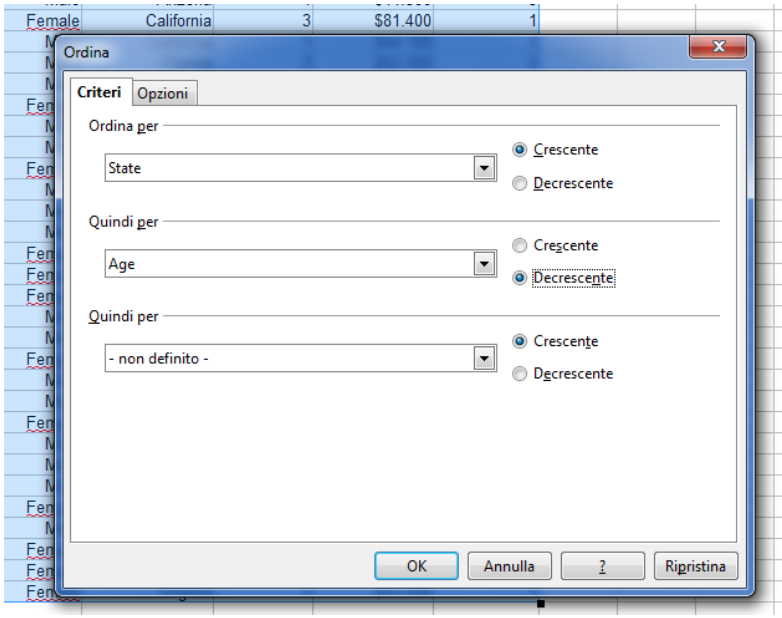


- E` inoltre possibile effettuare ordinamenti annidati, ovvero ordinare rispetto ad un attributo e, tra gli elementi che hanno lo stesso valore dell'attributo, effettuare l'ordinamento rispetto a un secondo attributo, e così via;
- Esempio su environmental policy (slide precedente);

Ordinamento

- E' anche possibile effettuare l'ordinamento di una tabella di dati tramite il menu *Dati→Ordina...*
- Dalla finestra di dialogo è possibile impostare i criteri per l'ordinamento anche su più attributi





Interpretazione mediante distribuzione dei valori degli attributi

- Nello studio di un campione di osservazioni in cui alcune variabili sono di tipo categorico o categorizzabile, può essere molto informativo vedere come le osservazioni si distribuiscono sulle categorie;
- Una *tabella di frequenze* riporta il numero di osservazioni che ricadono in ciascuna delle categorie stabilite;
- Un *istogramma* è una tecnica di visualizzazione di una tabella di frequenza tramite un diagramma a barre.

Creazione di una tabella di frequenze (1)

- oCalc consente di creare una tabella di frequenze a partire da una tabella di dati mediante la funzione:
  - **FREQUENZA(dati,classi)**
- dove:
  - **dati** è una tabella monodimensionale (array) che contiene l'insieme di valori di cui vogliamo calcolare le frequenze.
  - **classi** è una tabella monodimensionale che contiene gli intervalli in cui vogliamo raggruppare i valori in **dati**.

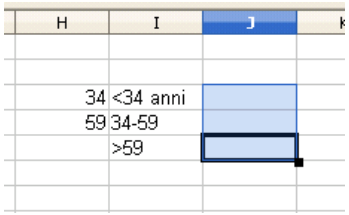
Creazione di una tabella di frequenze (2)

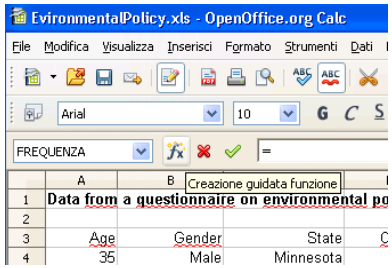
- *passo 1:* in un'area dello stesso foglio che contiene la tabella dei dati o su un altro foglio costruire una tabella monodimensionale (un array) contenente il valore superiore per ciascuna categoria
- *passo 2:* per rendere la tabella delle frequenze leggibile, creare una colonna affiancata a quella dei limiti superiori contenente una descrizione della categoria.

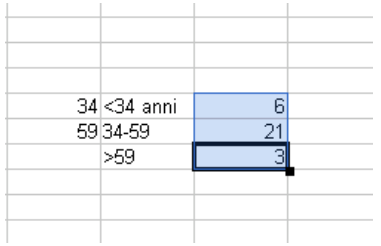
Female	Texas	2	$62.000	1			
Male	Ohio	0	$63.200	3			
Male	Florida	2	$52.000	5			
Female	California	3	$81.400	1			
Female	New York	3	$46.300	5		34	<34 anni
Female	Minnesota	2	$49.600	1		59	34-59
Male	New York	1	$45.900	5			>59
Male	Texas	3	$47.700	4			
Female	Texas	1	$59.900	4			
Male	New York	1	$48.100	4			
Female	Virginia	0	$58.100	3			
Female	Illinois	2	$56.000	1			
Female	Virginia	2	$53.400	1			

Creazione di una tabella di frequenze (3)

- Selezionare una colonna contenente un numero di celle pari alle categorie create al passo precedente
- digitare la formula utilizzando lo strumento "Creazione guidata funzione":
  - Per il campo "Dati" selezionare la colonna delle età; per il campo "Classi" selezionare i valori nella tabella di riferimento creata prima {34, 59, }
- Completare i passi della creazione guidata
- *Esercizio:* generare le distribuzioni di Salary e ForeignGross a partire dal file ACTORS

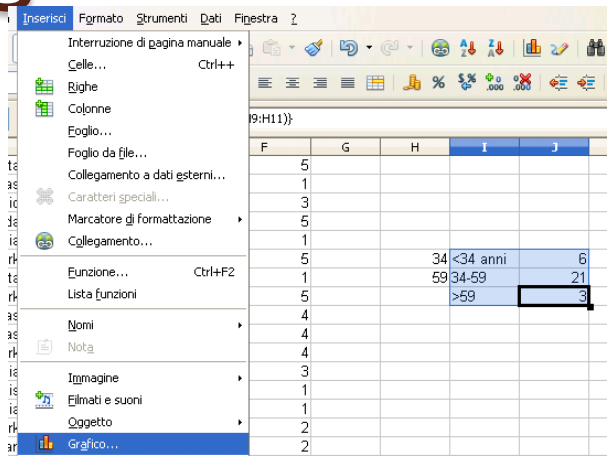


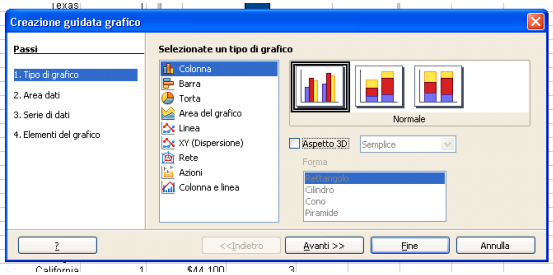


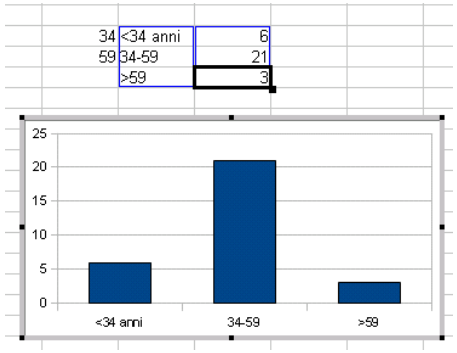


Creazione di un Istogramma

- Utilizzare il menu Inserisci→Grafico a partire dalle colonne:
  - delle spiegazioni delle categorie
  - delle frequenze (calcolate al passo precedente)
- e scegliendo come output il diagramma a barre
- *Esercizio:* costruire gli istogrammi relativi alle tabelle di frequenze ottenute per il file ACTORS

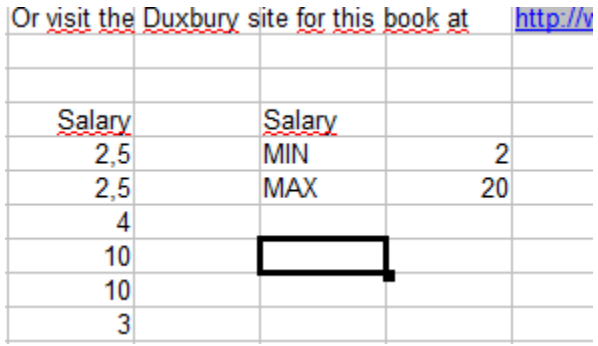






Esercitazione (1)

- Generare le distribuzioni di Salary a partire dal file ACTORS
- Procediamo a definire le soglie per i valori di Salary, in modo tale da avere dei range uniformi per ogni soglia
  - Selezioniamo il MIN e il MAX per i valori utilizzando le funzioni del foglio di calcolo
    - =MIN(E5:E71)
    - =MAX(E5:E71)



Esercitazione (2)

- I valori di Salary sono compresi nell'intervallo [2,20]
- La dimensione dell'intervallo è 20-2=18
- Fissiamo tre soglie: Low_salary, medium_salary e high_salary
- La dimensione di ogni sottointervallo sarà: dim_intervallo/num_intervalli, ovvero 18/3=6
- I valori soglia per ogni intervallo saranno
  - Min+dim_intervallo = 2+6 = 8
  - soglia_precedente+dim_intervallo = 8+6=14
  - soglia_precedente+dim_intervallo = 14+6=20

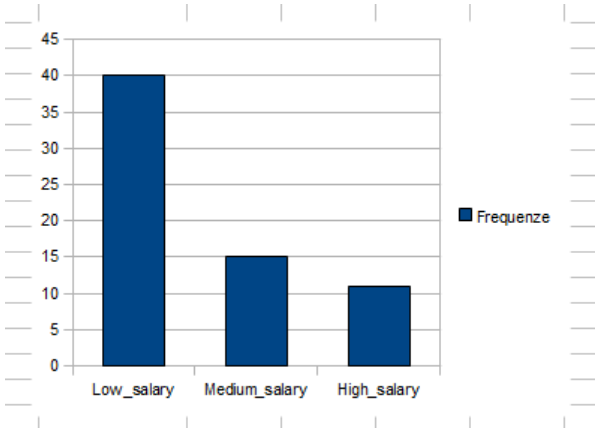
Salary	
MIN	2
MAX	20
Soglie	
Low_salary	8
Medium_salary	14
High_salary	20

Esercitazione (3)

- La distribuzione delle frequenze si ottiene dalle tre soglie definite prima utilizzando la funzione:
  - =FREQUENZA(E6:E71;H10:H12)
- Dai valori di frequenza ottenuti si deriva l'istogramma tramite il diagramma a barre

I10 {=FREQUENZA(E6:E71;H10:H12)}

	E	F	G	H	I
4					
5	Salary		Salary		
6	2,5		MIN	2	
7	2,5		MAX	20	
8	4				
9	10		Soglie		Frequenze
10	10		Low_salary	8	40
11	3		Medium_salary	14	15
12	12		High_salary	20	11
13	6				



Esercitazione (1)

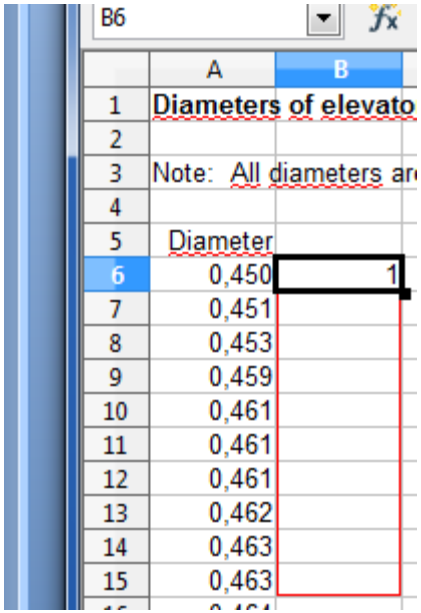
- La fabbrica di ascensori OTIS ha misurato il diametro dei cavi da ascensore prodotti (file OTIS1). Generare la distribuzione (scegliendo opportunamente le categorie) e l'istogramma relativo e discuterne la forma
- I valori contenuti nel file sono 399
  - In questo caso, sceglieremo tre soglie in modo che ogni sottointervallo contenga lo stesso numero di valori

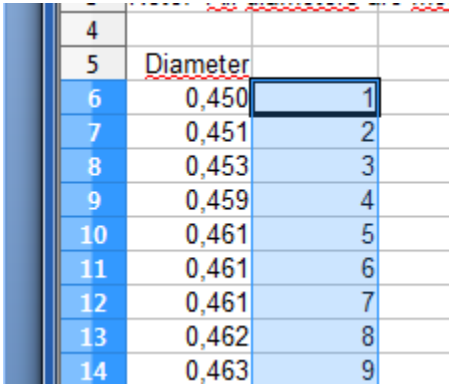
Esercitazione (2)

- Iniziamo ordinando i valori in senso crescente
- Dato che vogliamo tre intervalli della stessa dimensione e che i valori totali sono 399, la prima soglia sarà il valore corrispondente alla 133esima riga, la seconda soglia sarà la 266esima riga e l'ultimo valore sarà l'ultima riga

Esercitazione (3)

- Possiamo evitare di contare le righe sullo schermo utilizzando una serie numerica da affiancare ai valori ottenuti
- Inseriamo un 1 accanto alla prima riga e poi trasciniamo l'angolo in basso a destra della cella appena inserita, fino all'ultima cella della prima colonna





Esercitazione (4)

- Cerchiamo le righe dove vogliamo effettuare il taglio e annotiamo i valori corrispondenti come soglie
  - 0.492
  - 0.510
  - 0.548
- La distribuzione della frequenza si ottiene come prima