Salvatore Ingrassia
Carmela Senatore

Laboratorio di Statistica I

Guida alle Attività

Facoltà di Economia, Università della Calabria
Corso di Laurea in Statistica
Anno Accademico 2002-2003

Indice

Capitolo 1

Statistiche Univariate

1.1 Importazione di un file .data

Nell'analisi statistica è fondamentale l'uso di un foglio elettronico come *Excel*[1] soprattutto in considerazione del gran numero di dati che viene gestito durante una elaborazione. Tuttavia, spesso, i dati a disposizione non sono formattati per poter essere consultati in modo immediato con *excel*, è quindi necessario procedere con un'imporatazione guidata del testo. In genere i file di dati hanno estensione *.data* o *.dat* o a volte *.txt*. Prima di procedere con l'importazione è essenziale accertare che le impostazioni relative ai numeri prevedano lo stesso separatore previsto nel file dati da importare. Nella finestra opzioni internazionali, del pannello di controllo, è possibile fare queste verifiche e modificare le modalità di visualizzazione della data, dell'ora, della valuta, dei numeri interi e decimali (vedi Figure 1.1e 1.2).

Figura 1.1: *Opzioni internazionali*

[1] Il programma Excel fa parte del software Microsoft Offi ce.

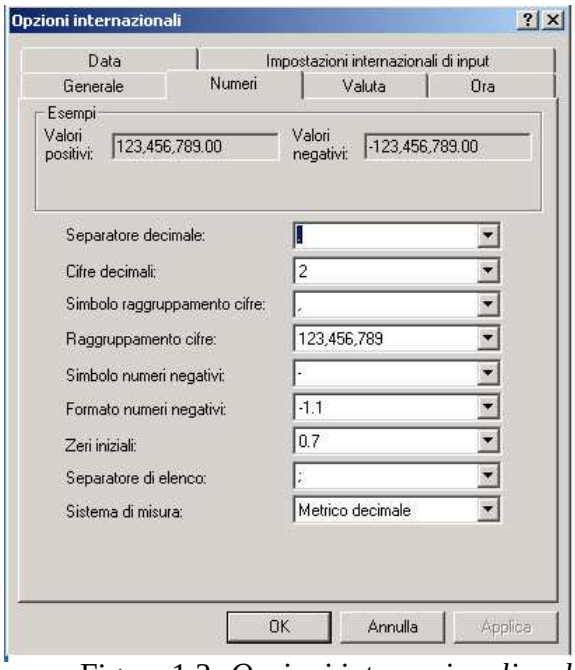

Figura 1.2: *Opzioni internazionali: scheda numeri*

La procedura d'importazione guidata consente di leggere i file di testo formattati in vari modi:

1. file delimitati da,
    - tabulatori, elementi separati da un carattere di tabulazione;
    - spazi, elementi separati da spazi;
    - virgole, elementi separati da virgola;
    - un punto e virgola, elementi separati da un punto e virgola;
    - un altro carattere, elementi separati da un altro carattere;
2. file in formato fisso, tutti gli elementi del file di testo sono della stessa lunghezza.

In questa applicazione, supportata da Excel, si utilizzerà il set di dati "wine recognition data", relativo a particolari caratteristiche del vino che permettono d'indentificarne diverse tipologie di vini. Questi dati sono il risultato di un'analisi chimica di vini prodotti nella stessa regione d'Italia, ma in diverse coltivazioni. Il dataset in esame consta di $N = 178$ osservazioni relative a $p = 13$ variabili quantitative concernenti le caratteristiche rilevate. Per gli scopi di questa prima parte del lavoro considereremo un sottoinsieme costituito da 15 casi e 6 variabili. Le informazioni descritte dalle variabili sono le seguenti:

1. alcol;
2. acido malico;[2]
3. magnesio;

[2]Ossiacido bicorbassilico presente in diversi frutti di piante e nel vino.

4. totale fenoli;[3]
5. intensità del colore;
6. colore.

Il file *wine.data* (relativo al set di dati sopra descritto) è reperibile sul sito internet www.economia.unical.it/STATistica/Laboratori/dati/wine6.dat; dopo averlo salvato in un'apposita cartella si può procede con l'esportazione del file.

In Excel, la procedura di importazione del file è la seguente:

- dal menù file scegliere apri e quindi selezionare dalla cartella precedentemente creata il file *wine.data*, si accede così alla finestra di dialogo "Importazione Guidata del testo", se il file è delimitato, come in questo caso, si selezionerà il pulsante di opzione "delimitati", se il file è in formato fisso si selezionerà il pulsante di opzione "larghezza fissa"(vedi Figura 1.3). Continuare cliccando sul pulsante avanti;

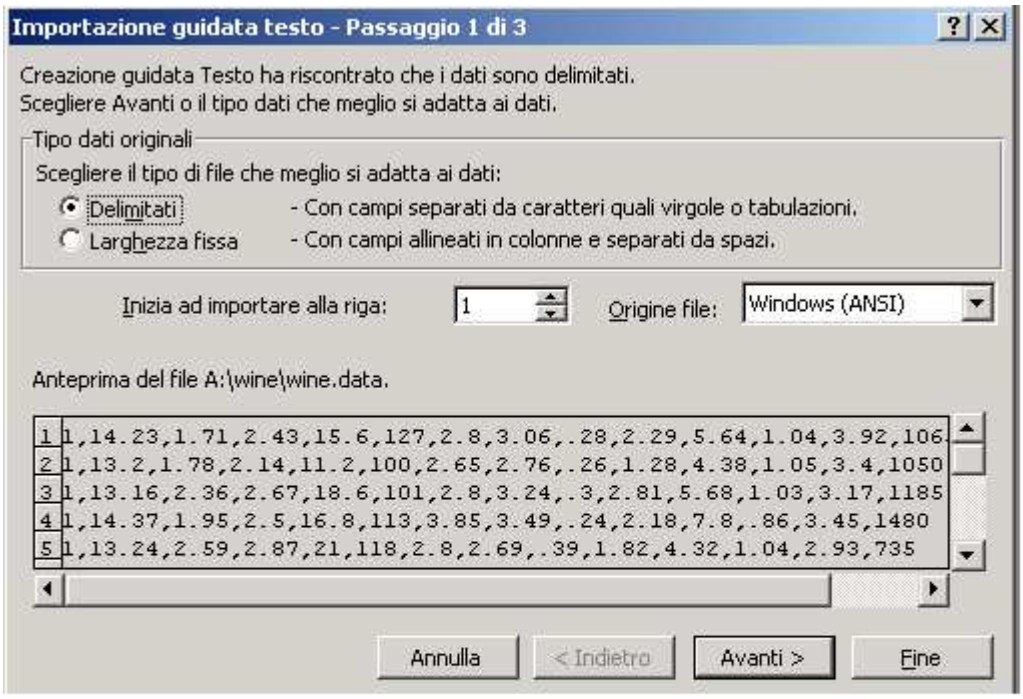

Figura 1.3: Importazione guidata testo

- a questo punto si accede alla finestra successiva dove si deve selezionare il tipo di delimitatore se il file è delimitato e creare, eliminare, spostare interruzioni di colonna se il file è a larghezza fissa (vedi Figura 1.4e1.5);

[3]Composto chimico utilizzato nelle industrie chimiche come disinfettante.

Figura 1.4: *Importazione guidata testo: file delimitato*

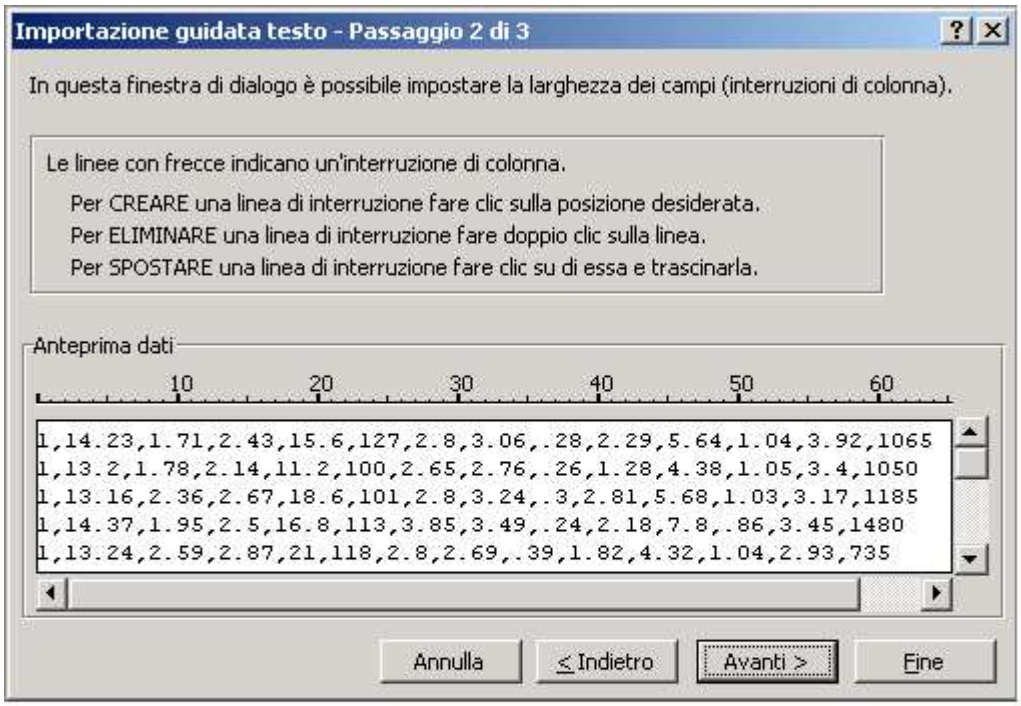

Figura 1.5: *Importazione guuidata testo: file a larghezza fissa*

- l'ultima finestra di dialogo permette di specificare il tipo di separatore decimale o delle migliaia da visualizzare in un file di testo con il pulsante *avanzate* [4] e di selezionare la colonna di testo

[4] All'apertura del file con Microsoft Excel i separatori visualizzati corrisponderanno a quelli specificati per il paese

convertito desiderata nella casella anteprima [5] dati, quindi selezionare il formato di dati che si desidera applicare alla colonna nella casella di gruppo Formato dati per colonna. Confermare tutte le operazioni con il pulsante *fine* (vedi Figura 1.6).

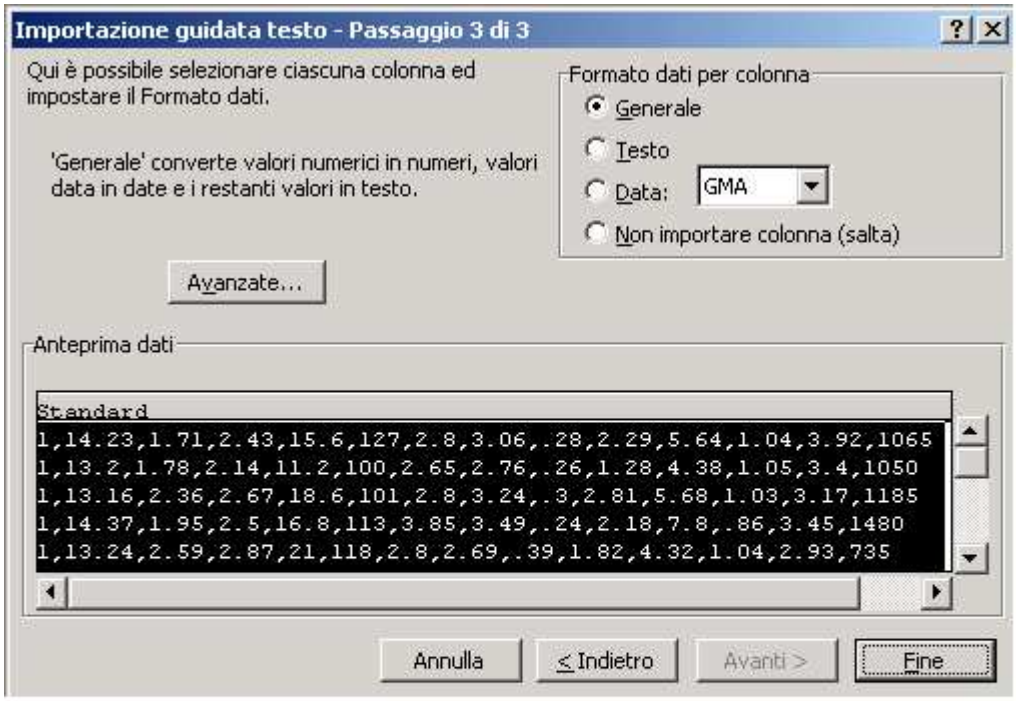

Figura 1.6: *Importazione guuidata testo: file a larghezza fissa*

Completate le operazioni di esportazione il file di dati si presenta come in Figura 1.7 si salva, quindi, il file in formato excel (.xls).

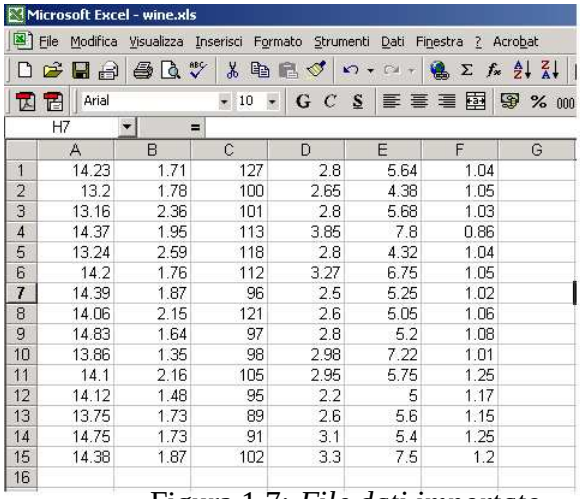

Figura 1.7: *File dati importato*

scelto in Impostazioni internazionali (Pannello di controllo).

[5] In tutte le finestre di dialogo Importazione dati il file di testo viene visualizzato in una finestra di anteprima.

L'operazione successiva prevede l'organizzazione dei dati, in modo da poter essere utilizzati durante l'analisi statistica.

I passi da eseguire sono i seguenti:

- inserimento di una colonna iniziale;
- inserimento di una riga iniziale;
- immissione della descrizione della statistica che s'intende calcolare. Le statistiche oggetto di questa esercitazione sono: media aritmetica, armonica e geometrica, mediana, quartili, varianza (varianza pop.), deviazione standard (deviazione standard pop); per comodità, in questo caso, le descrizioni saranno inserite a partire dalla cella **A17** fino alla **A29**
- inserimento dei nomi delle variabili: in questo caso verranno inseriti dalla cella **B1** alla cella **G1**

A questo punto il data set si presenta come in Figura 1.8.

	A	B	C	D	E	F	G	H
1		Alcol	Acido Malico	Magnesio	Tot.fenoli	Intesità colore	Colore	
2		14.23	1.71	127	2.8	5.64	1.04	
3		13.2	1.78	100	2.65	4.38	1.05	
4		13.16	2.36	101	2.8	5.68	1.03	
5		14.37	1.95	113	3.85	7.8	0.86	
6		13.24	2.59	118	2.8	4.32	1.04	
7		14.2	1.76	112	3.27	6.75	1.05	
8		14.39	1.87	96	2.5	5.25	1.02	
9		14.06	2.15	121	2.6	5.05	1.06	
10		14.83	1.64	97	2.8	5.2	1.08	
11		13.86	1.35	98	2.98	7.22	1.01	
12		14.1	2.16	105	2.95	5.75	1.25	
13		14.12	1.48	95	2.2	5	1.17	
14		13.75	1.73	89	2.6	5.6	1.15	
15		14.75	1.73	91	3.1	5.4	1.25	
16		14.38	1.87	102	3.3	7.5	1.2	
17	Media Aritmetica							
18	Media Geometrica							
19	Media Armonica							
20	Mediana							
21	Valore minimo							
22	Primo quartile							
23	Secondo Quartile							
24	Terzo Quartile							
25	Valore massimo							
26	Deviazione Standard							
27	Dev.St. Pop.							
28	Varianza							
29	Var. Pop.							
30								

Figura 1.8: *Organizzazione dei dati*

1.2 Medie e variabilità

Le procedure per il calcolo delle funzioni statistiche in Excel sono qui di seguito elencate. Dalla barra dei menù standard cliccare sull'icona f_x (funzioni)[6] e quindi scegliere tra l'elenco delle *categorie* delle funzioni disponibili *statistiche* e dal *tipo di funzione* quella che interessa; a questo punto si aprirà una finestra nella quale si dovranno inserire o l'intervallo di dati o la matrice di cui si vuole calcolare il valore della statistica (vedi Figura 1.9).

[6]Lo stesso risultato si ottiene selezionando dalla barra dei men`u, il men`u inserisci quindi funzioni.

Figura 1.9: *Funzioni statistiche*

I nomi delle funzioni che si intendono calcolare sono i seguenti:

- media;
- media.geometrica;
- media.armonica;
- quartile;
- dev.st.;
- dev.st.pop (calcola la deviazione standard in base all'intera popolazione);
- var;
- var.pop (calcola la varianza in base all'intera popolazione);

Per le specifiche della sintassi si veda l'appendice.

Nella Figura 1.10 si osserva il risultato dell'applicazione delle procedure di calcolo delle suddette funzioni, per la variabile *alcol*[7].

[7]La variabile alcol sar`a oggetto di tutti gli esempi relativi all'applicazione

16		14.38	1.87
17	**Media Aritmetica**	**14.04**	
18	**Media Geometrica**	**14.03**	
19	**Media Armonica**	**14.02**	
20	**Mediana**	**14.12**	
21	**Valore minimo**	**13.16**	
22	**Primo quartile**	**13.81**	
23	**Secondo Quartile**	**14.12**	
24	**Terzo Quartile**	**14.38**	
25	**Valore massimo**	**14.83**	
26	**Deviazione Standard**	**0.52**	
27	**Dev.St. Pop.**	**0.50**	
28	**Varianza**	**0.27**	
29	**Var. Pop.**	**0.25**	
30			

Figura 1.10: *Valori medi e variabilità*

1.3 Distribuzioni di frequenze

Prima di argomentare la procedura della costruzione delle distribuzioni di frequenza è necessario puntualizzare che le variabili considerate sono tutte quantitative; è, quindi, ragionevole costruire innanzitutto delle classi di frequenza. Si consideri, ancora una volta, la variabile alcol e si costruiscano per essa, a partire dal valore più piccolo, quattro classi di frequenza di ampiezza 0.5. A questo punto a partire dalla cella **A34** fino alla cella **H34** si digitino le seguenti voci:

- etichette classi;
- estremi superiori;
- frequenza assoluta (numero di volte che ciascuna modalità viene osservata);
- frequenza relativa (rapporto tra la frequenza j-esima e il totale delle unità);
- frequenza percentuale;
- frequenza cumulata (somma delle frequenze della modalità j e tutte le frequenze delle modalità precedenti);
- frequenza cumulata relativa[8];
- frequenza cumulata percentuale.

[8]Con un carattere quantitativo continuo suddiviso in classi, le frequenze cumulate si possono rappresentare tramite la funzione di ripartizione.

Sotto la cella *"etichette classi"* si riportino gli intervalli delle classi più la voce totale; la successiva colonna, in cui si inseriscono gli estremi superiori, viene costruita per poter poi calcolare le frequenze assolute mediante la funzione di Excel. La procedura da seguire per il calcolo delle **frequenze assolute** è quella relativa al calcolo di ogni funzione statistica (vedi paragrafo 1.2). Prima di iniziare la procedura è però necessario selezionare le celle che dovranno contenere le frequenze, in questo caso dalla **C35** alla **C39**. Durante la procedura di calcolo si selezionerà come tipo di funzione *frequenza* e si aprirà ,quindi, una finestra in cui si dovranno inserire nel campo *matrice dati* i dati relativi alla variabile alcol, e nel campo *matrice classi* i dati relativi alla colonna *estremi superiori*. Dopo aver confermato l'operazione attraverso la selezione del pulsante ok e necessario per poter calcolare tutte le frequenze digitare CTRL+MAIUSC+INVIO. Essendo le frequenze successive ricavabili manipolando le frequenze assolute Excel non prevede funzioni di calcolo. Nella Figura 1.11 si riportano i valori delle frequenze della variabile alcol ottenute seguendo le procedure appena descritte.

32								
33	Tabella delle frequenze relative alla variabile Alcol							
34	*classi di frequenza*	*estremo superiore*	*Frequenza assoluta*	*Frequenza relativa*	*Frequenza percentuale*	*Frequenza cumulata*	*Frequenza cumulata relativa*	*Frequenza cumulata percentuale*
35	13.16-13.66	13.66	3	0.20	20.00	3	0.20	20.00
36	13.66-14.16	14.16	5	0.33	33.33	8	0.53	53.33
37	14.16-14.66	14.66	5	0.33	33.33	13	0.87	86.67
38	14.66-15.16	15.16	2	0.13	13.33	**15**	**1.00**	**100.00**
39	TOTALE		**15**	**1.00**	**100.00**			
40								

Figura 1.11: *Tabella di frequenze*

1.4 Rappresentazioni Grafiche

Per ottenere un grafico con il programma Excel occorre selezionare l'area della tabella dei dati, che sarà stata preventivamente digitata nel *foglio di lavoro Excel*, includendo eventualmente le etichette dei dati che si vogliono rappresentare graficamente.

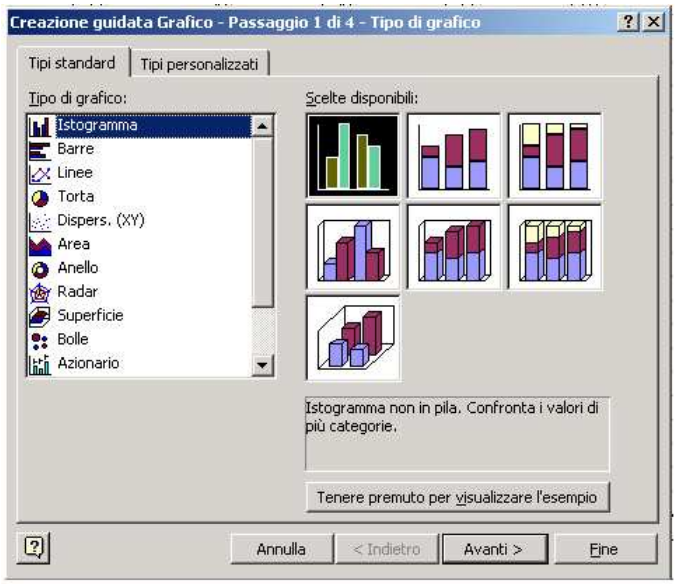

Figura 1.12: *Creazione guidata del grafico*

Una volta selezionate le caselle delle tabelle dei dati sulla barra dei menù standard, occorre cliccare sull'icona *"Creazione Guidata grafico"*(vedi Figura 1.12): i bordi dell'area delle celle selezionate della tabella dovranno allora apparire evidenziati, ossia lampeggianti, e ciò indica che si può procedere per scegliere il tipo di grafico adatto a rappresentare i dati selezionati. Appariranno diversi tipi di grafici e una volta cliccato su quello scelto occorrerà fornire via via le informazioni richieste per aggiungere le legende, i titoli e altre opzioni possibili, disponendo ogni volta di un'anteprima del grafico elaborato e anche della possibilità di tornare indietro per cambiare (vedi Figura 1.13).

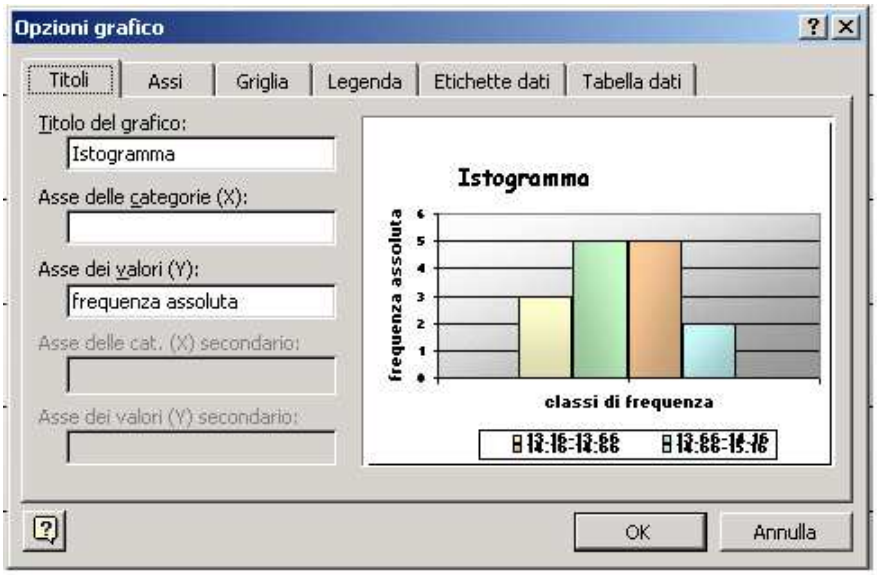

Figura 1.13: *Opzioni grafico*

Si fa osservare che occorre selezionare correttamente le serie dei dati che si vogliono andare a rappresentare fornendo esattamente le opzioni di scelta relative a i dati specificando se ci si riferisca alle righe o alle colonne. Una volta completato il grafico può essere aperto per essere modificato facendo doppio clic su una qualsiasi parte interna al quadrato contenente il grafico: appariranno quindi le diverse opzioni relative alle varie modifiche attuabili.

Quando la distribuzione considerata è relativa alle frequenze di una variabile divisa in classi, le rappresentazioni grafiche adatte sono:

- l'**istogramma**;
- il **poligono di frequenza**.

Negli istogrammi e nei poligoni di frequenza le frequenze sono proporzionali all'area (delimitata dalla spezzata che li costituisce e inclusa tra due valori reali sull'asse orizzontale), e non all'altezza della Figura. Ovviamente quando le classi hanno tutte la stessa ampiezza, l'area è proporzionale anche all'altezza[9]. Quindi, in pratica la rappresentazione grafica per istogrammi consiste nel riportare tanti rettangoli contigui, ciascuno avente base unaguale all'ampiezza della classe e altezza uguale o proporzionale alla frequenza assunta dalle classi dei valori considerati. Per l'istogramma in excel il

[9]Si ricorda che quando le classi non sono della stessa ampiezza `e necessario calcolare la densit`a media di frequenza affi nch`e le altezze di ciasun rettangolo corrispondenti alle frequenze delle classi siano confrontabili.

tipo di grafico è *istogramma non in pila*, per quanto riguarda il poligono delle frequenze può essere costruito mediante il tipo *linee* o con il tipo *dispersione (x,y)*, in quest'ultimo caso le coordinate dei punti hanno come ascissa i valori centrali delle classi e come ordinata le frequenze corrispondenti della classe.

Nelle Figure 1.14 e 1.15 si osservano l'istogramma e il poligono delle frequenze per la variabile alcol.

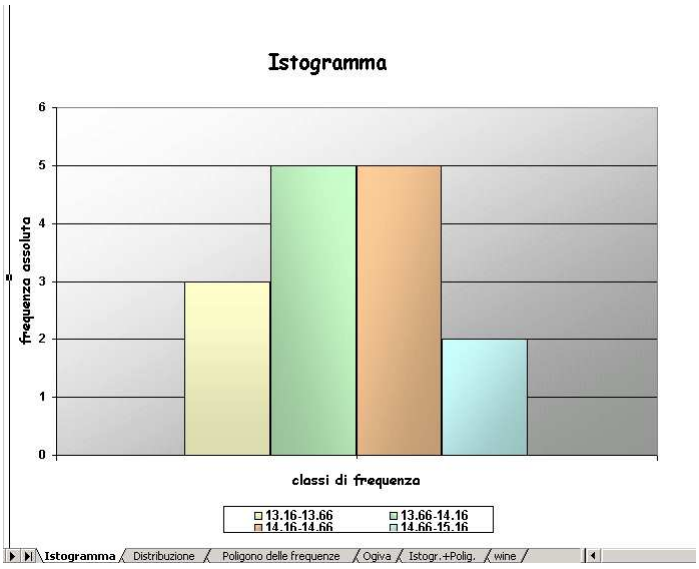

Figura 1.14: *Istogramma*

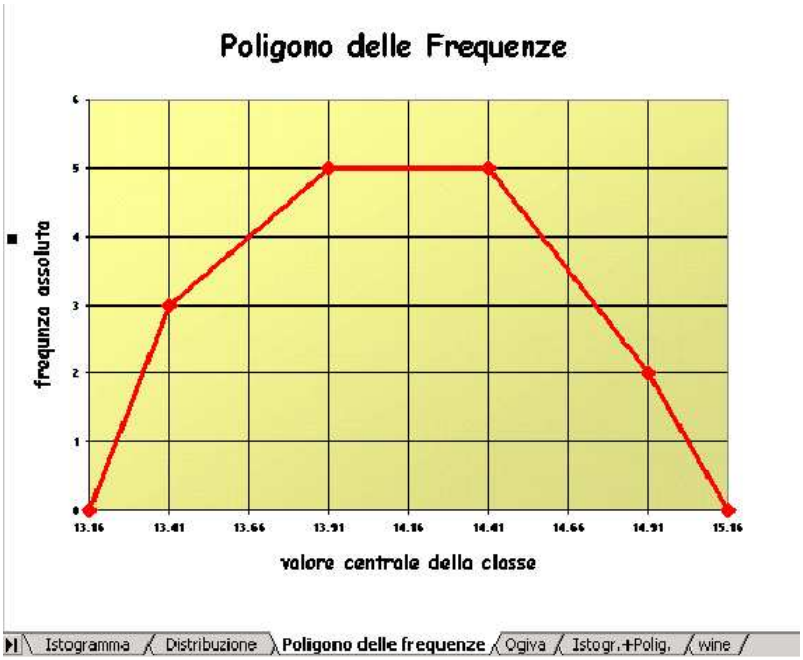

Figura 1.15: *Poligono di frequenza*

Quando si è in presenza di variabili quantitative può essere interessante, per evidenziare maggiormente l'andamento di una distribuzione, proiettare contemporaneamente l'istogramma e il poligono di frequenza, in Figura 1.16 si osserva tale grafico per la variabile alcol.

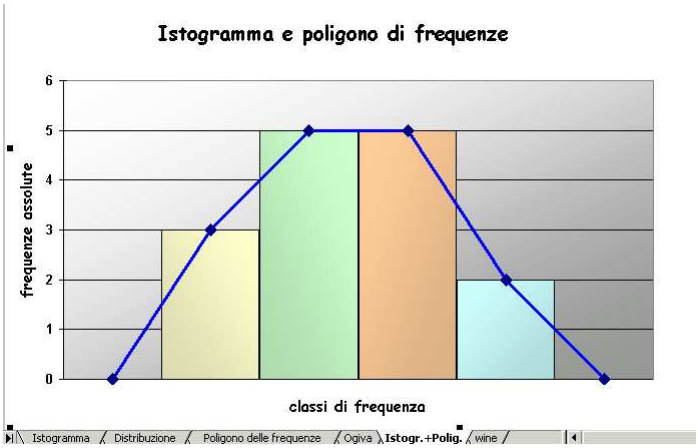

Figura 1.16: *Proiezione simultanea di istogramma e poligono delle frequenze*

In riferimento alle distribuzioni di frequenza cumulate,per le variabili continue, la rappresentazione grafica adatta è l'**ogiva** o poligono della frequenza cumulata [10] e si costruisce impiegando un diagramma in coordinate cartesiane ortogonali. In Excel il tipo di grafico utilizzato è *linee con indicatori di livello*. L'ogiva di frequenza per la variabile alcol è rappresentata in Figura 1.17.

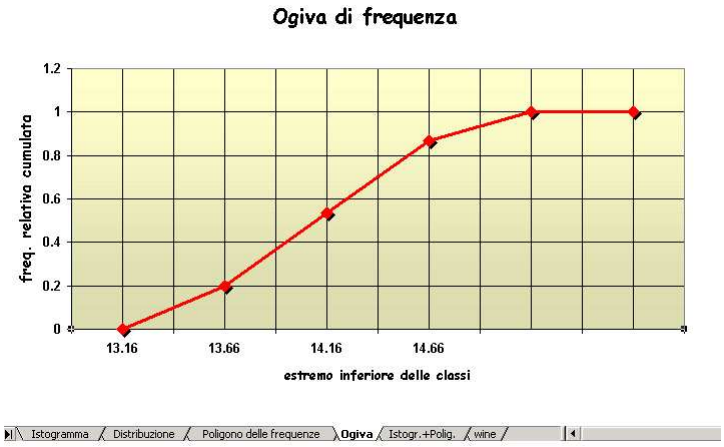

Figura 1.17: *Ogiva di frequenza*

Un'altra rappresentazione grafica, interessante per una variabile continua potrebbe essere il sua andamento intorno al valore medio. In questo caso il grafico excel utilizzato è del tipo *dispersione (x,y)*.

[10] Funzione di ripartizione

In Figura 1.18 si osserva tale rappresentazione per la variabile alcol.

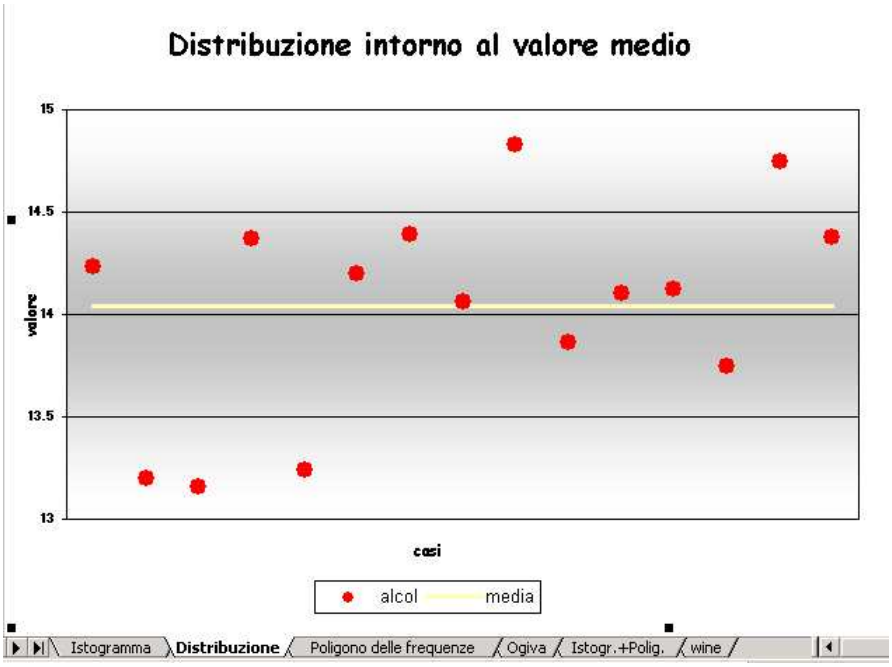

Figura 1.18: *Valori della variabile alcol intorno alla media*

Infine, un'ultima interessante rappresentazione grafica per un set di dati è il *box-plot*. Questo tipo di grafico al contrario dell'istogramma, che offre una visione generale delle caratteristiche dei dati, descrive simultaneamente diverse importanti caratteristiche di un data set.

Il *il box-plot* è un grafico caratterizzato da tre elementi:

1. un punto che indica la posizione centrale (di solito la mediana);
2. un rettangolo (box) di altezza legato alla variabilità dei valori "prossimi alla media" (scarto interquartile);
3. 2 segmenti che partono dai lati del rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione.

Il procedimento per creare dei *box-plot* in *Excel* è il seguente[11].

Si considerino i dati relativi alle sei variabili d'interesse, si costruisca una tabella contenente i valori del primo e terzo quartile, della mediana e dei valori minimi e massimi di ciascuna distribuzione come mostrato in Figura 1.19

[11]Il metodo qui illustrato `e una semplice rielaborazione del procedimento suggerito da Naville Hunt (*Conventry University*).

	Alcol	Acido Malico	Magnesio	Tot.fenoli	Intesità colore	Colore
q1	13.805	1.72	96.5	2.625	5.125	1.035
min	13.16	1.35	89	2.2	4.32	0.86
med	14.12	1.78	101	2.8	5.6	1.05
max	14.83	2.59	127	3.85	7.8	1.25
q3	14.375	2.05	112.5	3.04	6.25	1.16

Figura 1.19: *Tabella per la costruzione dei box-plot.*

I dati mostrati nella tabella possono essere facilmente calcolati con le funzioni disponibili in Excel, rispettando lordine qui indicato; selezionare quindi la colonna relativa alla prima variabile, escludendo solo liintestazione, attivare la procedura per la Creazione guidata grafico scegliendo il tipo di grafico linee e quindi, procedendo con il tasto AVANTI, selezionare serie in righe (si noti che Excel è predisposto per considerare i dati per colonna) e terminare la procedura con il tasto FINE.

Nel grafico così organizzato, i valori sono connessi mediante linee, che però non hanno alcun interesse ai fini dei box-plot. Sarà necessario quindi rimuovere queste linee: a tal fine è necessario selezionare la linea e quindi scegliere Serie dei dati selezionati dal menu Formato e, nel quadro Motivo, attivare lopzione Linea Assente; infine nel quadro Opzioni, selezionare le due voci Linee di Min-Max e Barre cresc.-decresc.

Per rendere maggiormente leggibile il grafico si può aggiungere una legenda (seguire la sequenza dei menù Grafico - Opzioni grafico - Legenda) e si possono modificare i simboli e i colori dei vari punti (in particolare quelli che rappresentano la mediana, affinch risultino ben evidenziati).

Il risultato finale è visibile in Figura 1.20.

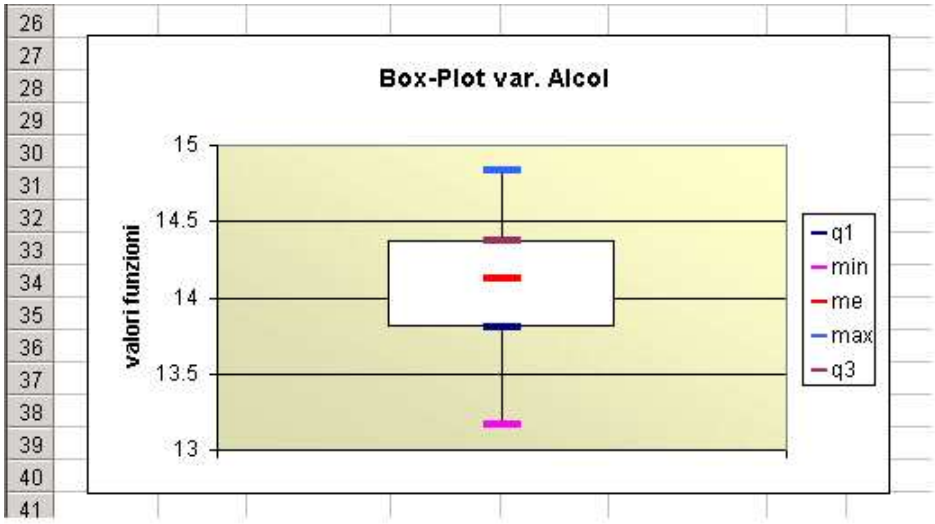

Figura 1.20: *Box-plot*

1.5 Analisi dati

In Excel è possibile generare rapporti di statistiche univariate anche con lo strumento di analisi "Statistica Descrittiva". In particolare esso genera informazioni sulla tendenza centrale e la variabilità dei dati analizzati. Selezionando l'opzione riepilogo statistiche per ogni variabile coinvolta nell'analisi sono restituiti i valori di: media, errore standard (della media), mediana, moda, deviazione standard, varianza, curtosi, asimmetria, intervallo, minimo, massimo somma , conteggio, più grande (numero), più piccolo (numero) e livello di confidenza.

Le procedure per il calcolo delle suddette statistiche sono qui di seguito riportate. Dal menù strumenti selezionare analisi dati [12], si accede quindi alla finestra di dialogo "Analisi dati" (vedi Figura 1.21), in essa sono elencati i diversi strumenti di analisi disponibili tra essi selezionare statistica descrittiva;

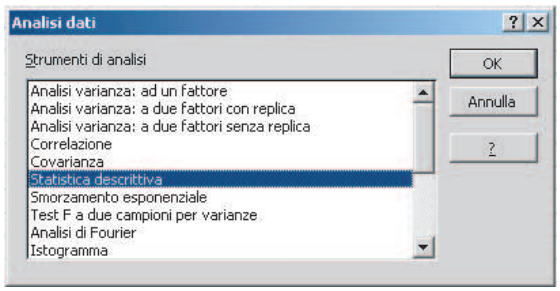

Figura 1.21: *Strumenti di analisi.*

a questo punto si aprirà la relativa finestra di dialogo nella quale inserire l'intervallo di dati da analizzare e selezionare l'output che s'intende calcolare, in questo caso **riepilogo statistiche**. Tali statistiche dovranno essere visualizzate in un nuovo foglio di lavoro nominato riepilogo statistiche (vedi Figura 1.22).

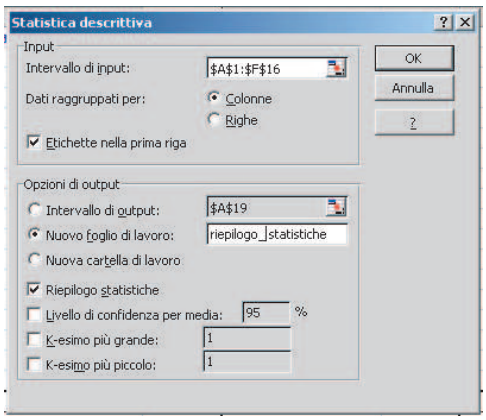

Figura 1.22: *Opzioni statistiche descrittive.*

[12] Se nel men`u strumenti la voce analisi dati non `e disponibile, scegliere dall'elenco componenti aggiuntive quindi selezionare strumenti di analisi, a questo punto lo strumento analisi dati sar`a disponibile nel men`u strumenti

In figura 1.22 si osserva il risultato dell'applicazione appena descritta.

18						
19	***Alcol***		***Acido Malico***		***Magnesio***	
20						
21	Media	14.04266667	Media	1.875333	Media	104.3333
22	Errore standard	0.134229607	Errore standard	0.084159	Errore standard	2.936903
23	Mediana	14.12	Mediana	1.78	Mediana	101
24	Moda	#N/D	Moda	1.87	Moda	#N/D
25	Deviazione standard	0.519869031	Deviazione standard	0.325946	Deviazione standard	11.37457
26	Varianza campionaria	0.27026381	Varianza campionaria	0.106241	Varianza campionaria	129.381
27	Curtosi	-0.444021557	Curtosi	0.455031	Curtosi	-0.556376
28	Asimmetria	-0.507937995	Asimmetria	0.690712	Asimmetria	0.659894
29	Intervallo	1.67	Intervallo	1.24	Intervallo	38
30	Minimo	13.16	Minimo	1.35	Minimo	89
31	Massimo	14.83	Massimo	2.59	Massimo	127
32	Somma	210.64	Somma	28.13	Somma	1565
33	Conteggio	15	Conteggio	15	Conteggio	15
34						

Figura 1.23: *Statistiche descrittive.*

1.6 Esercizi

Esercizio 1.1
Importare il file wine.data reperibile dal sito
www.economia.unical.it/STATistica/Laboratori/dati/wine.dat e per ogni variabile effettuare le seguenti operazioni:

1. calcolare:
    - media, media geometrica, media armonica, quartili, deviazione standard, varianza, distribuzioni di frequenza;
2. costruire i seguenti grafici:
    - istogramma, poligono di frequenza, ogiva di frequenza, box-plot;
3. calcolare le statistiche di riepilogo utilizzando lo strumento di analisi “Statistiche descrittive”.