Organizzazione dei Dati

Raw Data

L’informazione raccolta per descrivere la caratteristica di una popolazione o di un campione è una sequenza di dati casuali non ordinati.

|  |
| --- |
| **Raw Data o dati grezzi.** Insieme di dati registrati secondo la sequenza di raccolta che **non** sono stati sottoposti ad alcuna analisi o manipolazione preliminare. |

La Tabella 2.1 riporta l’età in anni di 50 studenti iscritti alla Università di Ferrara. I valori raccolti in successione: 20, 19, 27, ecc. sono un esempio di un **raw data-set di tipo quantitativo**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tabella 2.1. Età degli studenti | | | | | | | | | |
| 20 | 19 | 27 | 20 | 20 | 27 | 24 | 27 | 31 | 33 |
| 24 | 20 | 25 | 19 | 18 | 19 | 20 | 19 | 27 | 18 |
| 20 | 29 | 22 | 28 | 26 | 18 | 20 | 19 | 20 | 21 |
| 23 | 18 | 21 | 29 | 20 | 20 | 24 | 26 | 23 | 28 |
| 24 | 28 | 18 | 26 | 20 | 21 | 24 | 25 | 21 | 29 |

Agli stessi studenti è stato chiesto di indicare il Corso di Laurea che frequentano: Medicina e Chirurgia (MC), Odontoiatria (O), Scienze Motorie (SM), Infermiere (I), Tecnico di Radiologia (TR), Fisioterapia (F), Dietologo Alimentarista (DA), Riabilitazione Psichiatrica (RP). Le risposte riportate in Tabella 2.2 sono un esempio **raw data-set di tipo qualitativo (categoriali)**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tabella 2.2. Corso di laurea | | | | | | | | | |
| MC | MC | F | MC | MC | F | I | F | DA | I |
| I | SM | I | I | MC | MC | MC | MC | F | MC |
| MC | DA | SM | DA | TR | SM | TR | I | MC | SM |
| I | MC | O | RP | MC | MC | I | TR | I | DA |
| I | F | SM | I | SM | MC | I | MC | MC | RP |

I dati in Tabella 2.1 e Tabella 2.2 sono detti anche **dati non raggruppati**; ciascun valore presente nel data set contiene informazione sugli membri (individui) del campione o della popolazione.

Organizzazione e Rappresentazione di Dati Qualitativi

La ***variabile*** del data set in Tabella 2.2 che indica il Corso di Laurea al quale sono iscritti gli studenti è di tipo ***qualitativo***ed un variabile ***nominale***[[1]](#footnote-1),[[2]](#footnote-2)*.*

La Tabella 2.3 riporta i dati di Tabella 2.2 in modo organizzato; cioè, elenca i corsi di laurea ai quali sono iscritti gli studenti ed il numero di studenti che frequentano ciascun corso: ad ex. 17 studenti frequentano Medicina, 6 frequentano Scienze Motorie, ecc.

|  |
| --- |
| Categorie  frequenza |

Nella I colonna sono elencati i corsi di laurea cioè le **categorie** del campione esaminato. Le categorie sono mutuamente esclusive cioè lo studente appartiene ad una sola categoria. Nella II colonna è indicata la **frequenza** di ciascuna categoria cioèil numero di studenti iscritti a ciascun corso di laurea.

L’insieme di tutti gli **elementi** del campione (popolazione) cioè le categorie ed i valori della **variabile** “No. Studenti” la **distribuzione delle frequenze** del campione cioè il numero di studenti per ciascuna categoria. La **tabella delle frequenze** 2.3 è la distribuzione delle frequenze della variabile nominale di Tabella 2.1.

|  |
| --- |
| La **distribuzione delle frequenze** per un set di dati qualitativi è un elenco di tutte le categorie e degli elementi del data set che appartengono a ciascuna categoria |

**Esempio 2.1.** Ad un campione di 30 studenti neodiplomati che intendono proseguire gli studi è stato chiesto a quale corso di laurea triennale intendono iscriversi, gli indirizzi proposti dal questionario sono: Medico Sanitario (MS), Economico (Ec), Industriale (In), Scienze di Base (SB), Altri (Al).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| In | Al | SB | Al | Al | MS | Al | Ec | In | Ec |
| Al | Al | Al | Ec | In | Ec | Ec | In | Al | Al |
| MS | In | Al | MS | MS | Ec | SB | Al | Al | MS |

La distribuzione delle frequenze è riportata in Tabella 2.4. Le variabili del data set sono le 5 categorie elencate nel questionario, a lato sono riportati la frequenza di ciascuna risposta[[3]](#footnote-3). La somma di tutti i valori nella colonna delle frequenze deve dare come risultato le dimensioni del campione o *frequenza totale*

**Tabella 2.4**. Distribuzione delle frequenza delle risposte al questionario

|  |  |
| --- | --- |
| Indirizzo di Studio | Frequenza |
| Altri | 12 |
| Economia | 6 |
| Ingegneria | 5 |
| Medico Sanitario | 5 |
| Scienze di Base | 2 |
| Totale | 30 |

Distribuzione delle Frequenza Relativa e della Percentuale

La **frequenza relativa** di una categoria si calcola dividendo la frequenza della categoria per la frequenza totale. La *distribuzione delle frequenze relative* è l’elenco delle frequenze relative di tutte le categorie del data set.



La **percentuale** di una categoria si ottiene moltiplicando la frequenza relativa della categoria per 100. La *distribuzione delle percentuali* è l’elenco delle percentuali di tutte le categorie presenti nel data set.



La somma delle frequenze relative è sempre uguale a 1.0 e la somma delle percentuali è uguale a 100.

In Tabella 2.5 sono indicate la distribuzione delle frequenze relative e della percentuale del data set qualitativo.

**Tabella 2.5.** Distribuzione della Frequenza Relativa e della Percentuale per la Tabella 2.4.

|  |  |  |
| --- | --- | --- |
| Indirizzo di Studio | Frequenza  Relativa | Percentuale |
| Altri | 12/30 = 0.400 | 0.400\*100 = 40.0 |
| Economia | 6/30 = 0.200 | 0.200\*100 = 20.0 |
| Ingegneria | 5/30 = 0.167 | 0.167\*100 = 16.7 |
| Medico Sanitario | 5/30 = 0.167 | 0.167\*100 = 16.7 |
| Scienze di Base | 2/30 = 0.067 | 0.067\*100 = 6.7 |
| Totale[[4]](#footnote-4) | 1.001 | 100.1 |

Rappresentazione grafica di dati qualitativi

I valori della frequenza, della frequenza relativa e della percentuale di un set di dati qualitativi possono essere rappresentati in modo grafico: il *grafico a barre* è utilizzato per rappresentare i valori della frequenza e della frequenza relativa, il *grafico a torta* per rappresentare i valori della probabilità.

Grafico a Barre

Il grafico a barre è un diagramma cartesiano dove sull’asse orizzontale (asse x) sono indicate in intervalli di larghezza costante le categorie presenti nel data set, e nell’asse verticale (asse y) la frequenza, la frequenza relativa o la probabilità[[5]](#footnote-5). Nel grafico la frequenza di ciascuna categoria viene indicata con una *barra* la cui altezza è pari alla frequenza della rispettiva classe.

La Figura 2.1 riporta il grafico a barre per la Tabella 2.4.

****

**Figura 2.1.** Grafico a barre per la distribuzione delle frequenze di Tabella 2.4

|  |
| --- |
| **Grafico a Barre**. Grafico costituito da barre la cui altezza rappresenta la frequenza, la frequenza relativa o la probabilità delle categorie. |

Grafico a Torta (Pie Chart)

La **pie chart** o Grafico a Torta è il tipo di grafico utilizzato per rappresentare i valori di probabilità[[6]](#footnote-6) di una serie di dati qualitativi. La torta intera (una circonferenza di 360 gradi) rappresenta l’intero campione o la popolazione. Per indicare la percentuale di ciascuna categoria la torta viene divisa in porzioni (settori di circonferenza) proporzionali al valore percentuale corrispondente. L’ampiezza in gradi di ciascun settore si ottiene moltiplicando la corrispondente frequenza relativa di ciascuna categoria per 360 gradi[[7]](#footnote-7).

**Tabella 2.6.** Calcolo degli angoli della Pie Chart di Tabella 2.4.

|  |  |  |
| --- | --- | --- |
| Indirizzo di Studio | Frequenza  Relativa | Angolo in gradi |
| Altri | 0.400 | 0.400\*360 = 144 |
| Economia | 0.200 | 0.200\*360 = 72 |
| Ingegneria | 0.167 | 0.167\*360 = 60 |
| Medico Sanitario | 0.167 | 0.167\*360 = 60 |
| Scienze di Base | 0.067 | 0.067\*360 = 24 |
| Totale[[8]](#footnote-8) | 1.001 | 360 |

In Tabella 2.6 sono indicate le misure in gradi delle categorie del questionario di Tabella 2.5; nella pie chart di Figura 2.2 sono indicati i valori della percentuale delle risposte al questionario sull’indirizzo di studio scelto dagli studenti.

**Figura 2.1.** Pie Chart per la distribuzione delle frequenze di Tabella 2.4

Il foglio elettronico <EserciziCap2.xls> riporta alcuni esercizi sviluppati in ambiente Excel dove è stata calcolata la distribuzione delle frequenze, il grafico a barre e la pie-chart di alcuni data set qualitativi. Gli esercizi sono stati tratti dal Cap-2 del libro di testo.

Il foglio elettronico <HomeRun.xls> ed il documento [HomeRun.doc](Baseball_HomeRun.doc) illustrano un esempio di calcolo della distribuzione delle frequenze e del relativo istogramma per un serie di data set nel quale sono riportati i “fuori campo” di due giocatori della “Baseball legue USA” nel campionato 1998.

Organizzazione e Rappresentazione di Dati Quantitativi

Distribuzione delle Frequenze

La Tabella 2.7 elenca lo stipendio settimanale di 100 dipendenti di una azienda: nella prima colonna sono indicate le *classi* che indicano la variabile quantitativa *stipendio settimanale*. Per una variabile quantitativa la *classe* è un intervallo che comprende tutti i valori che sono compresi fra due valori noti come *limite inferiore* e *superiore*. Le classi non si sovrappongono ed i valori della variabile quantitativa possono appartenere ad una sola classe. La seconda colonna indica il numero di impiegati il cui stipendio appartiene ad una data classe, i valori nella seconda colonna sono le **frequenze** *f* di ciascuna classe.

**Tabella 2.7.** Stipendio settimanale di 100 impiegati di una azienda

|  |  |  |
| --- | --- | --- |
| **Stipendio settimanale**  **(Euro)** | **Numero di impegati**  **( frequenza** *f* **)** | |
| 301 a 400 | 9 |  |
| 401 a 500 | 16 |  |
| |  | | --- | | 501 a 600 | | |  | | --- | | 33 | |  |
| 601 a 700 | 20 |  |
| 701 a 800 | 14 |  |
| 801 a 900 | 8 |  |

I valori riportati in Tabella 2.7 sono la rappresentazione tipica della **tabella della distribuzione** **delle frequenze** per dati quantitativi. I valori in evidenza indicano la terza e la settima classe della tabella delle frequenze e in queste i limiti inferiori (501 e 801) e superiori (600 e 900).

Tabella delle frequenze per Dati Quantitativi:

* La **Distribuzione delle frequenze** è un elenco delle **classi** nelle quali è indicata la variabile quantitativa con associato il numero di elementi (**la frequenza**) appartenenti a ciascuna classe.
* I dati rappresentati in una tabella delle frequenze suddivisi in **Classi** sono detti **dati raggruppati**.
* Il **Confine fra le classi** è Punto intermedio fra l’estremo superiore di una classe e l’estremo inferiore della classe successiva. Il confine fra la prima e la seconda classe è pari a

[(400+401)/2 = 400.5].

* La **Larghezza della classe** è la Differenza fra gli estremi superiori ed inferiori di ciascuna classe.

Larghezza = [Estremo superiore - Estremo inferiore]

* Il **Punto centrale o marker della classe.**  Differenza fra estremo superiore ed inferiore di ciascuna classe.

Marker = [Estremo superiore - Estremo inferiore]/2

Costruzione della Tabella di distribuzione delle Frequenze

* Il **Numero delle classi** varia da 5 a 20 all’aumentare del numero di osservazioni nel data set.
* La **Larghezza delle classi** si calcola a partire dai valori massimo e minimo presenti nel data set. La larghezza approssimativa di ciascuna classe è data dalla formula[[9]](#footnote-9)

Larghezza = [Valore massimo - Valore minimo]/Numero delle classi

la larghezza approssimata viene successivamente arrotondata ad un valore conveniente che viene assunto come larghezza delle classi.

* Il **Limite inferiore della prima classe** è un valore numerico pari o inferiore al valore minimo.
* Il **Limite superiore dell’ultima classe** è un valore numerico pari o superiore al valore minimo.

**Esempio 2-3**. La tabella 2-9 indica gli ingaggi ([ingaggi.xls](Ingaggi.xls)) pagati ai propri giocatori dalle 30 squadre più importanti della lega di basket. Vogliamo costruire la tabella delle frequenze del data set degli ingaggi.

**Tabella 2-9.** Ingaggi delle principali squadre di basket

|  |  |  |  |
| --- | --- | --- | --- |
| **Squadra** | **Ingaggi (M$)** | **Squadra** | **Ingaggi (M$)** |
| Anaheim | 51 | Milwaukee | 43 |
| Arizona | 70 | Minnesota | 16 |
| Atlanta | 79 | Montreal | 15 |
| Baltimora | 75 | New York Meets | 72 |
| Boston | 72 | New York Yankees | 92 |
| Chicago Cubs | 55 | Oakland | 25 |
| Chicago White Sox | 25 | Philadelphia | 30 |
| Cincinnati | 38 | Pittsburgh | 24 |
| Cleveland | 74 | St. Luis | 46 |
| Colorado | 54 | San Diego | 47 |
| Detroit | 37 | San Francisco | 46 |
| Florida | 15 | Seattle | 45 |
| Houston | 56 | Tampa Bay | 38 |
| Kansas City | 17 | Texas | 81 |
| Los Angeles | 77 | Toronto | 49 |

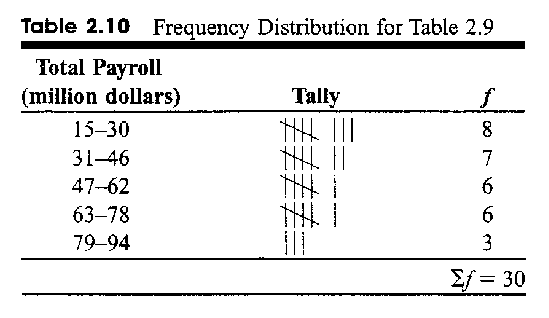
1. Scegliamo quale **Numero di classi** per descrivere la distribuzione delle frequenze 5
2. Dati *Max* = 92 e *Min* = 15 calcoliamo la *larghezza approssimativa* delle classi:



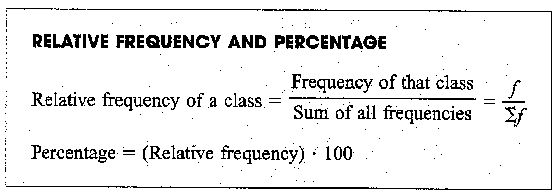
1. Scegliamo quale **larghezza di ciascuna** classe un valore intero maggiore di 15.4, ad esempio 16, e quale **limite inferiore** il valore minimo 15.
2. Le **classi** di eguale ampiezza per il calcolo della distribuzione delle frequenze sono rispettivamente

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 15-30 | 31-46 | 47-62 | 63-78 | 79-94 |

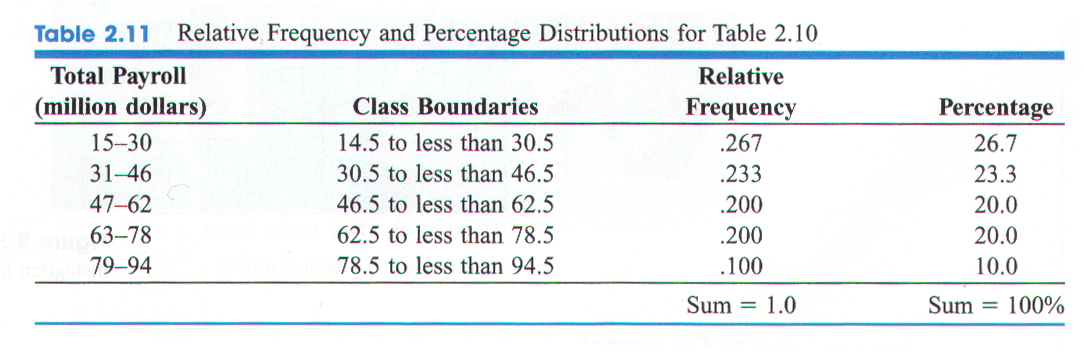
Costruiamo ora la tabella delle frequenze elencando le classi nella prima colonna (Payroll) e nella seconda le frequenze *f* di ciascuna classe[[10]](#footnote-10).



La tabella 2.9 riporta tutte le squadre della lega Basket cioè tutta la popolazione e per questa ragione la somma dei valori di frequenza *f*i di tutte le classi è pari a S *f*i = 30 ed è pari a tutti i dati presenti nella popolazione.

**Importante:** La tabella di distribuzione delle frequenze non conserva alcuna informazione sulle osservazioni individuali presenti in un campione di dati raggruppati.

Distribuzione della Frequenza Relativa e della Percentuale

Calcolo della Frequenza relativa e delle percentuale dalla Tab.210

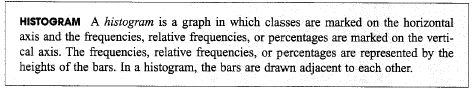
La distribuzione dei valori percentuali è l’informazione usata più di frequentemente per riassumere rapidamente i valori nel data set quantitativo, ad esempio

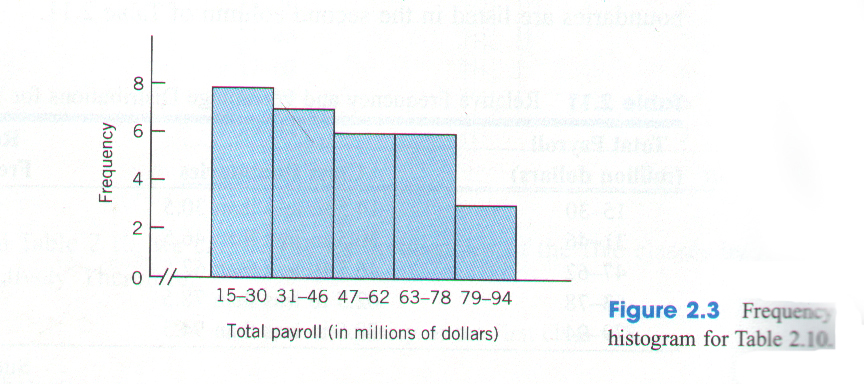
* Circa il 27% dei team spende una cifra inferiore a 30M$ per gli ingaggi
* Il 50% di tutti i Team della lega Basket spende al più 46M$ per gli ingaggi
* Solo il 10% dei Team spende più 78M$ in ingaggi.

Rappresentazione grafica di dati Raggruppati

I dati di tipo quantitativo raggruppati in classi posso essere rappresentati tramite un *istogramma* o una *poligonale*.

Istogramma

L’**istogramma** delle frequenze è un diagramma cartesiano dove in ascissa sono riportate le classi in cui sono stati raggruppati i dati, ed in corrispondenza di ciascuna classe è disegnata una barra verticale la cui altezza corrisponde alla frequenza ad essa corrispondente. L’istogramma è utilizzato per rappresentare la distribuzione delle frequenze, delle frequenze relative e anche della probabilità.

 La Figure 2.3 riportano l’istogramma con la distribuzione delle frequenze dei dati raggruppati di tabella 2.10[[11]](#footnote-11).

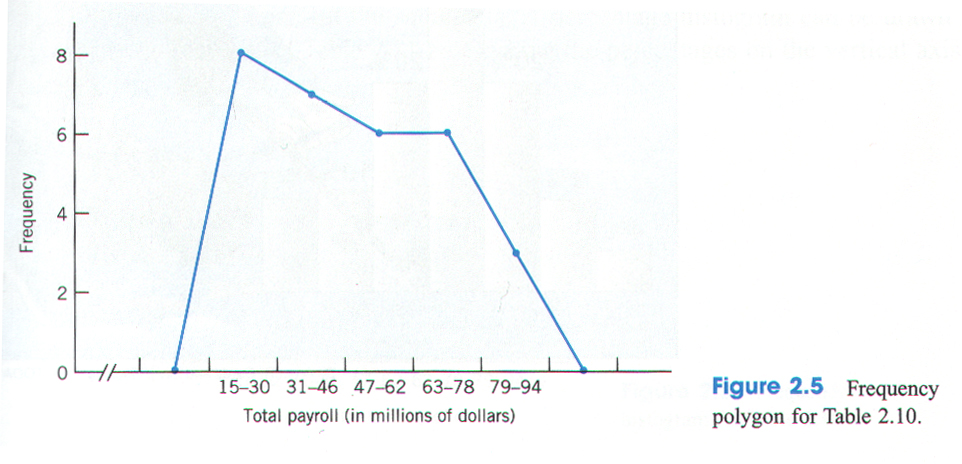
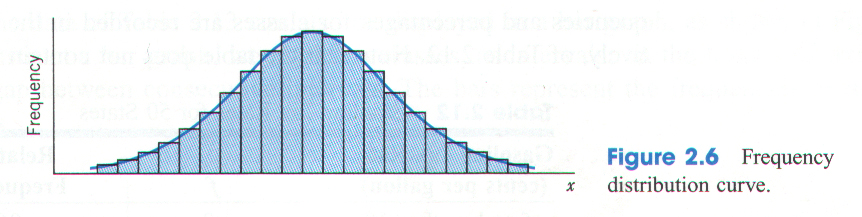
Gli istogrammi della distribuzione delle frequenze, delle frequenze relative e della probabilità hanno la medesima ascissa le classi e ordinate diverse. Poiché le barre hanno base uguale a 1 l’area di ciascuna barra è esattamente uguale alla frequenza della classe corrispondente, e l’area dell’istogramma è uguale alla somma di tutte le frequenze: *f*i=*n* ([ingaggi.xls](Ingaggi.xls))*.*

Poligoni

I **poligoni** o le **poligonali** si ottengono congiungendo con una spezzata il punto centrale (class mark) di ciascuna della barre che indicano la frequenza di ciascuna classe. Il grafico a poligoni richiede la presenza di due ulteriori classi con frequenza nulla che precedono e succedono le classi in cui sono stati raggruppati i dati.

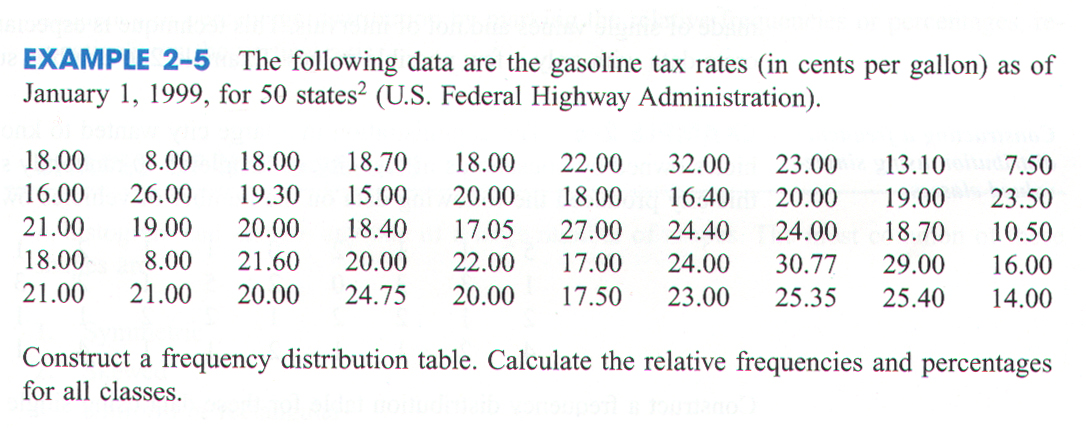


Analogamente alla rappresentazione tramite istogramma è possibile costruire il poligono delle frequenze relative e delle probabilità.

**Importante.** Nel caso in cui il data set esaminato abbia grandi dimensioni un al suo interno grande numero di dati (valori di *n* grandi se un campione o *N* grandi se una popolazione) il numero di classi in cui raggruppare i dati aumenta, se contemporaneamente l’ampiezza delle classi diminuisce la poligonale si trasforma in una curva continua. Questa curva assume un profilo simile a quello indicato nella Fig.2.6 e prende il nome **di curva di distribuzione delle frequenze** o **curva delle frequenze**.

Esempi di Dati raggruppati di tipo diverso

Distribuzione delle frequenze per Classi delimitate dall’estremo superiore.

I valori del data set in cui sono indicati gli ingaggi dei giocatori della lega NBA sono stati ottenuti approssimando l’ingaggio reale al valore interi in M$ più vicino. Nel caso in cui il data set sia composto da valori frazionari è opportuno raggruppare i dati calcolando la frequenze con cui appaiono *i valori che sono minori o uguali ad un valore di riferimento* rappresentato dall’estremo superiore di ciascuna classe. Illustriamo la tecnica con un esempio nel quale si chiede di calcolare la distribuzione delle frequenze e delle frequenze relative di un insieme di dati nel quale sono riportati i valori delle imposte ([imposte.xls](Benzina.xls)) sulla benzina (in centesimi di $) in ciascuno dei 50 stati USA.

Soluzione.

* Valori Max = 32, Min = 7.5, numero delle classi = 6
* Ampiezza approssimativa di ciascuna classe ; arrotondata =5.

|  |
| --- |
| Estremo superiore di ciascuna classe  **Tabella 2.12**. Imposte sul consumo di benzina negli stati USA |

La tecnica di suddivisione dei dati basata sull’estremo superiore della classe e utilizzata dalle funzione predefinite di Excel; ed è utilizzata per tutti esercizi svolti negli appunti.

Classi corrispondenti a valori discreti (Sigle-Valued Classes)

Se le osservazioni presenti nel data set sono contraddistinte da un insieme di **valori discreti** può essere appropriato costruire la distribuzione delle frequenze utilizzando quali valori delle classi i singoli valori discreti assunti dalla variabile esaminata.

**Esempio 2-6**. Gli amministratori di una grande città sono interessati ad avere informazioni dettagliate sul numero dei veicoli posseduti da ciascuna famiglia residente. Da un campione di 40 famiglie scelte a caso si è ricavato il seguente data set dove sono riportati i raw data ([automobili.xls](Automobili.xls)).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 1 | 1 | 2 | 0 | 1 | 1 | 2 | 1 | 1 |
| 1 | 3 | 3 | 0 | 2 | 5 | 1 | 2 | 3 | 4 |
| 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 |
| 4 | 2 | 1 | 1 | 2 | 1 | 1 | 4 | 1 | 3 |

Il data set *ha solo valori discreti* X={0,1,2,3,4,5} e poiché il valori minimo e massimo sono facilmente individuabili da una osservazione diretta del data-set, la tabella delle frequenze si ricava facilmente elencando a sinistra il numero di auto possedute per famiglia e a destra il numero di famiglie che possiedono 0, 1, 2, … automobili.



**Tab. 2-13**. Distribuzione delle frequenze dei veicoli posseduti dalle famiglie.

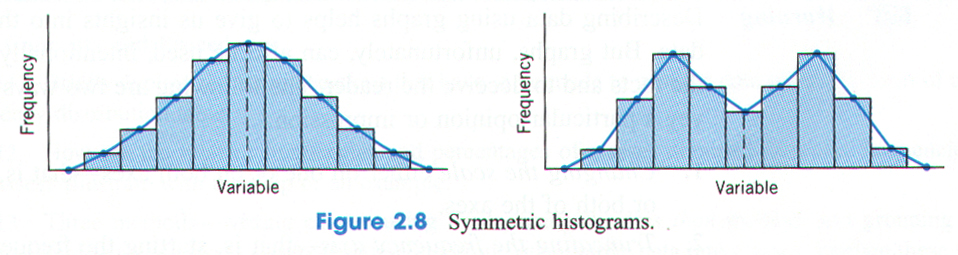
Dalla Tab. 2-13 si ricava un grafico a barre (istogramma) dove sull’asse orizzontale sono riportati come **indicatori di classe** il numero di autoveicoli posseduti e sull’asse verticale la frequenza (frequenza relativa o percentuale) di ciascuna classe.

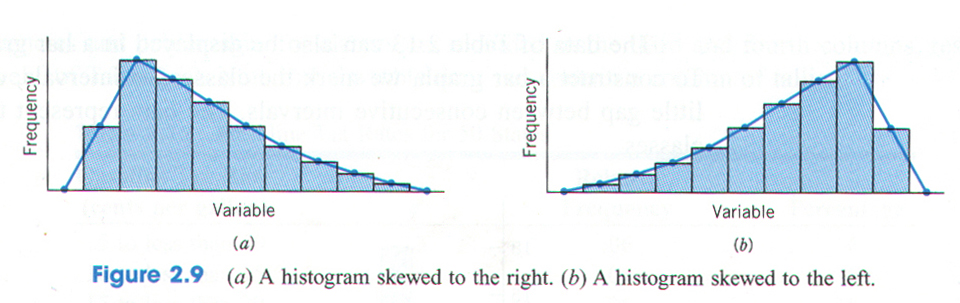
Istogrammi con forme diverse

L’istogramma può assumere forme diverse ma alcune fra queste si ripetono frequentemente ed è importante darne indicazione (forme.xls). Fra forme più comuni si individuano istogrammi:

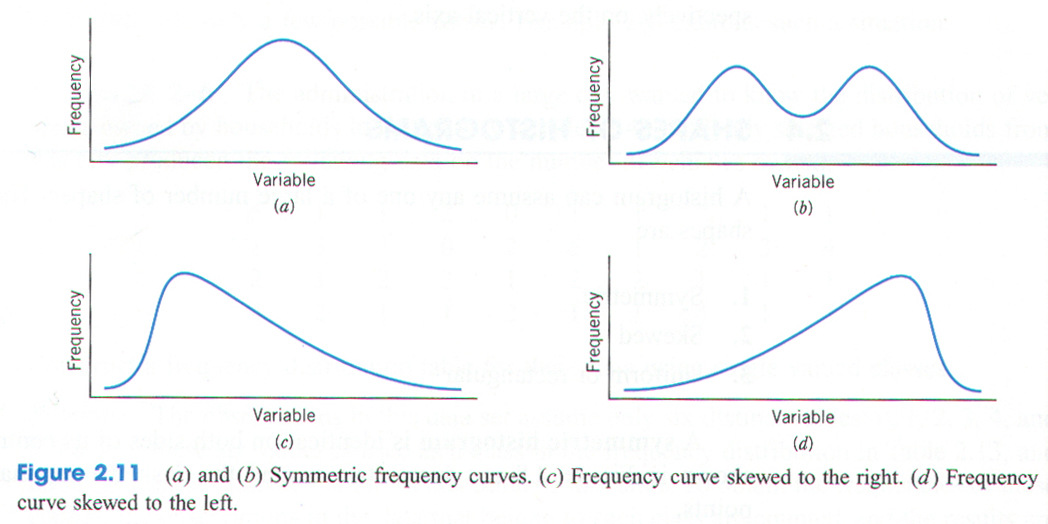
1. Simmetrici
2. Asimmetrici (Skewed)
3. Uniformi o rettangolari

Un istogramma è detto **simmetrico** quando ha forma speculare rispetto un punto centrale scelto come riferimento. Entrambi gli istogrammi di figura 2.4 hanno forma simmetrica, l’istogramma di sinistra ha una forma detta a campana con code simmetriche rispetto al valore centrale.

Un istogramma **non simmetrico** (Skewed) mostra una coda molto più ha estesa dell’altra se la coda estesa è a destra (sinistra) del valore di frequenza massima è detto **simmetrico verso destra** (sinistra) nel gergo statistico skewed to the right o skewed to the left. Esempi di istogrammi skewed rispettivamente verso destra e verso sinistra sono indicati in Fig.2.9.

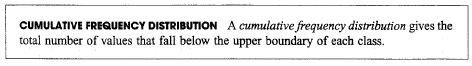
Un istogramma che descrive una distribuzione di frequenza uniforme (istogramma rettangolare) ha la forma caratteristica di Figura 2.10.

|  |
| --- |
| Par-2 |

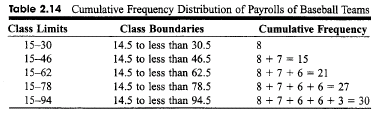
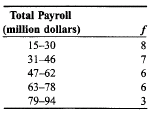
Le medesimi indicazioni di distribuzione simmetrica, asimmetrica e uniforme possono essere utilizzate per la curva delle distribuzioni delle frequenza; esempi di distribuzioni simmetrica asimmetrica e uniforme sono riportati in Fig.2.11.

Distribuzione Cumulativa della Frequenza

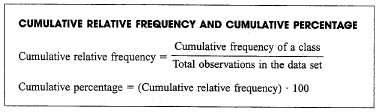
Dalla tabella della distribuzione delle frequenze è possibile ricavare informazioni *puntuali* quali la frequenza, la frequenza relativa o la probabilità di ciascuna classe nella quale sono stati raggruppati i dati. Ad esempio dalla Tab.2-10 è facile ricavare il numero di società della NBA che spendono una cifra compresa fra i 15-30M$ per gli ingaggi, non è però immediato individuare quante società spendono *al più* 46M$. Per ottenere la risposta a una domanda che relativa ad informazioni *non puntuali* è opportuno costruire la **distribuzione delle frequenze cumulative**.

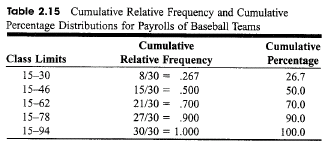


Esempio 2-7. Distribuzione delle frequenze cumulative della Tab.2-10.



Dalla Tab. 2-14 si ricava che 15 squadre della NBA spendono e *al più* 46M$ in ingaggi[[12]](#footnote-12).

Le distribuzioni della frequenza relativa cumulativa e della percentuale cumulativa si ottengono tramite le relazioni:

 Dal calcolo si ricavano i valori che seguono

Ogiva

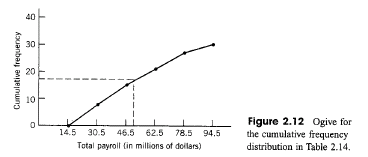
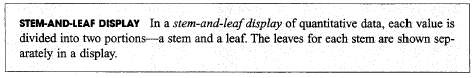
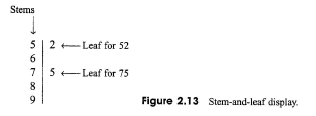
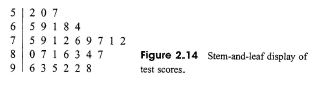
I valori della distribuzione della frequenza cumulativa riporatati in un grafico danno luogo ad una curva continua detta **ogiva** che collega i valori della frequenza cumulativa. In ascissa il grafico ogiva riporta i valori di confine fra le classi, o l’estremo supereiore di ciascuna classe, in ordinata i valori della frequenza cumulativa.

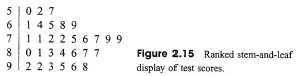
Grafico Tronco-Foglia (Steam-and-Leaf Display)

Lo Steam-and-Leaf Display è una tecnica usata per la rappresentazione di dati quantitativi in forma condensata. Contrariamente alla distribuzione delle frequenze lo Steam-and-Leaf display conserva le informazioni (il valore) sulle singole osservazioni. Lo Steam-and-Leaf display non può essere utilizzato per dati qualitativi.

**Esempio 2-8.** Punteggio conseguito al test di Statistica da 30 studenti.

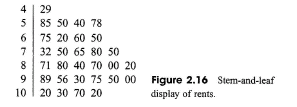
Costruiamo lo Steam-and-Leaf display. Separiamo ciascun valore nelle corrispondenti decine ed unità: ad esempio il valore minimo 52 ha cinque decine, il **tronco** e 2 unità che rappresentano le **foglie**. Elenchiamo il valore del tronco per ciascun valore a sinistra di una linea di riferimento verticale; riportiamo a destra della linea il numero di foglie corrispondenti. La Fig.2-13 mostra la costruzione dello Steam-and-Leaf per i valori 52 e 75.

Ripetiamo il procedimento per tutti i valori presenti nel data set

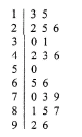
Dalla Fig.2-14 si ricavano facilmente informazioni simili alla tabella delle distribuzioni delle frequenze: il “tronco 5” ed il “tronco 7” hanno rispettivamente la frequenza più elevata e meno elevata. Se ordiniamo lo Steam-and-Leaf display secondo un ordine crescente otteniamo lo schema di Fig.2.15

che dimostra che lo Steam-and-Leaf descrive il data set esaminato conservando i valori delle singole osservazioni. Infatti, dalla osservazione diretta dello Steam-and-Leaf è possibile dedurre che un solo studente ha ottenuto un punteggio pari a 98, che due studenti hanno superato la prova con un punteggio pari a 72, ecc. Si osservi che la distribuzione delle frequenze non consente di conservare i valori delle singole osservazioni.

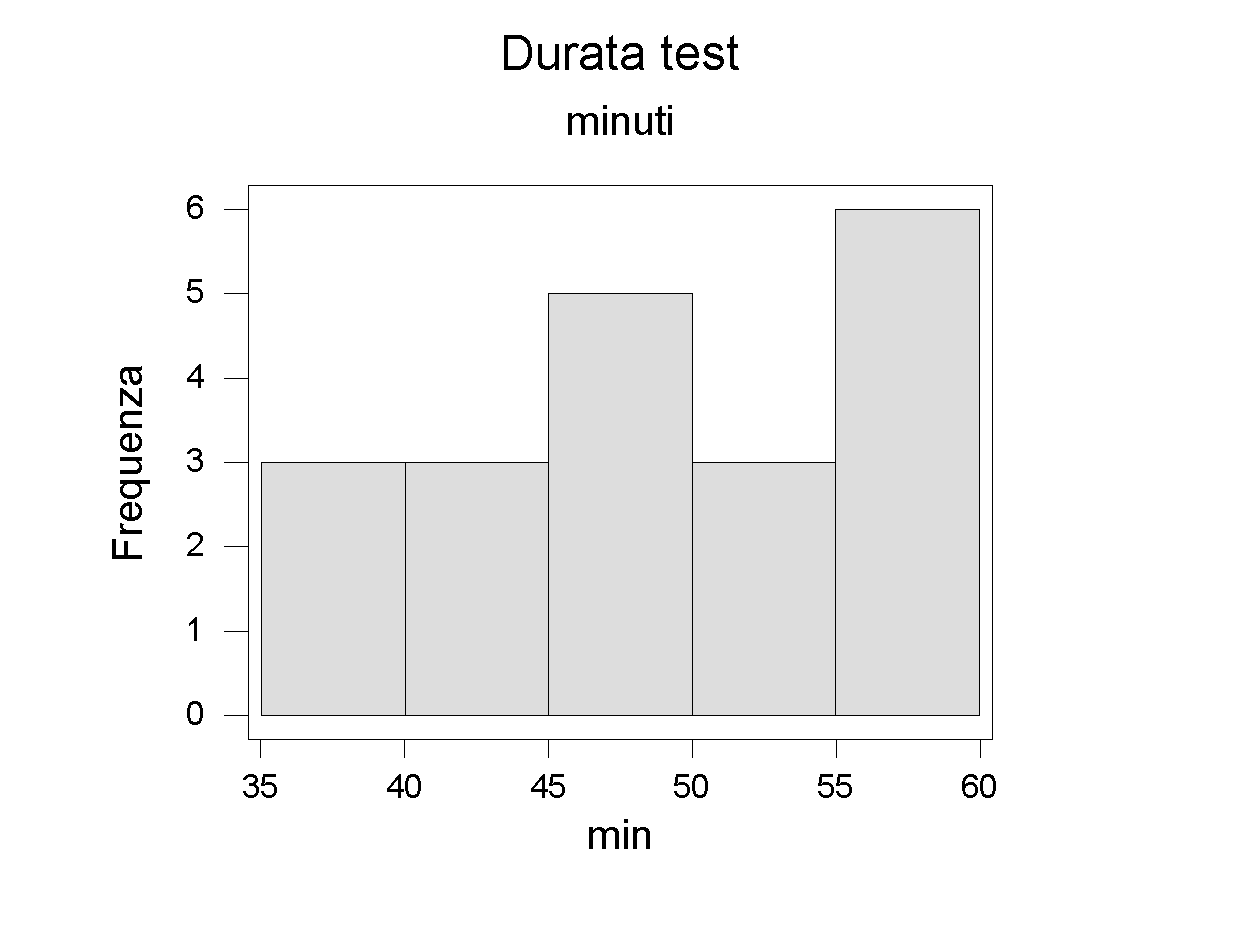
**Esempio 2-9**. Affitto mensile pagato da 30 inquilini scelti a caso in diversi quartieri della città. I valori sonno in euro.

La Fig.2.16 riporta lo Steam-and-Leaf display dove le centinaia sono scelte come tronco.

**Esempio 2-10**. Lo Steam-and-Leaf display che segue è stato ricavato da un data-set dove sono riportati i prezzi in euro relativi alle vendite a stock di 21 grossisti di cancelleria.

Lo Steam-and-Leaf display può essere condensato raggruppando i valori di steam 1-3, 4-6, 7-9. Il simbolo \* è stato inserito per separare il valore di osservazione (leaf) di ciascun gruppo di tronchi.

**Esercizio 2-48**. La tabella riporta il tempo in minuti impiegato dai 20 studenti che hanno eseguito il test di Statistica. Si vuole costruire lo Steam-and-Leaf display per questo data-set organizzando le foglie di ciascuno steam in ordine crescente.



Cumulativa

Leaff

Cumulativa

Mediana

Steam

Steam-and-leaf display

Istogramma delle frequenze

1. Una **variabile nominale** consiste in una serie di categorie con nomi distinti, tra le quali non è possibile stabilire un criterio di ordinamento logico. [↑](#footnote-ref-1)
2. Fra le variabili qualitative vanno annoverate anche le **variabili ordinali** che rappresentano una serie ordinata di categorie dove la differenza fra le diverse categorie non può essere considerata costante. Un esempio di variabile qualitativa ordinale è lo schema di valutazione utilizzato nella scuola media di I grado, in base al quale viene assegnato un punteggio ordinale del tipo insufficiente, buono, distinto, ottimo, eccellente. [↑](#footnote-ref-2)
3. La distribuzione delle frequenze si ottiene contando il valori della variabile per ciascuna categoria. In ambiente Excel questo si ottiene con la funzione predefinita di Excel “**CONTA.SE**(**data set**; **categorie**)”. [↑](#footnote-ref-3)
4. Il valore totale della frequenza relativa e della percentuale sono diversi dai rispettivi valori di 1.0 e 100 causa le approssimazioni del calcoli. [↑](#footnote-ref-4)
5. Talvolta il grafico a barre riporta le categorie in ordinata e la frequenza in ascissa. [↑](#footnote-ref-5)
6. La pie chart viene talvolta utilizzata anche per rappresentare la distribuzione delle frequenze relative. [↑](#footnote-ref-6)
7. La misura in gradi delle frequenze relative di ciascuna categoria viene eseguita automaticamente da tutti i programmi di analisi statistica, ivi compreso l’Excel. [↑](#footnote-ref-7)
8. Il valore totale della frequenza relativa e della percentuale sono diversi dai rispettivi valori di 1.0 e 100 causa le approssimazioni del calcoli. [↑](#footnote-ref-8)
9. Questa relazione è valida solo nel caso in cui tutte le classi abbiano la medesima larghezza [↑](#footnote-ref-9)
10. Operativamente ciascun valore viene confrontato con gli estreme di ciascuna classe ed assegnato alla classe di appartenenza. Questa procedura è eseguita comunemente da tutti i software di analisi statistica oggi disponibile incluso il workbook per l’Analisi dei Dati presente in Excel. [↑](#footnote-ref-10)
11. Il simbolo di interruzione - //- presente sull’asse delle classi indica un troncamento dell’asse utilizzato per rappresentare una scala molto più grande dei valori indicati sull’asse. [↑](#footnote-ref-11)
12. Dalla distribuzione delle frequenze cumulative è possibile ricavare risposte anche a domande del tipo: quanti Team pendono *almeno* 46M$ in ingaggi. Infatti, poiché il numero totale delle squadre è pari a 30 si ha:

    **Team che spendono almeno 46M$ = No. Totale dei Team - Team che spendono al più 46M$**

    Per esercizio calcolare: a) il numero di Team che spendono almeno 62M$ e b) una cifra compresa fra i 46M$ e 72M$. [a) 9 Team; b) 12 Team]. [↑](#footnote-ref-12)