Corso Integrato di Statistica Informatica e Analisi dei Dati Sperimentali

Note A.A. 2009-10

1 Dati: organizzazione, sintesi e rappresentazione

1.1 Dati e variabili

Molti fenomeni fisici sono di tipo aleatorio o stocastico, significa che l'esito di un certo evento non é prevedibile con esattezza. Ovvero la realizzazione di un esperimento ripetuto in condizioni del tutto uguali potrá avere risultati anche molto diversi. Esempi di fenomeni aleatori sono: il risultato di un lancio di dati, un risultato elettorale, l'esito di un test diagnostico, un carattere ereditario.

L'aleatorietá dei fenomeni fisici dipende da diversi fattori concorrenti: alcuni insiti nel fenomeno stesso, altri dovuti ad una scarsa o incompleta conoscenza di tutti i fattori che concorrono alla realizzazione di un evento.

Le principali cause di incertezza nella realizzazione di un evento sono:

- l'aleatorietá intrinseca di alcuni fenomeni quali, ad esempio, il lancio di dadi, i decadimenti radioattivi, le estrazioni del lotto, la trasmissione di caratteri ereditari.
- La definizione incerta della grandezza che si sta osservando o che si vuole misurare. Supponiamo ad esempio di voler misurare l'accelerazione di gravitá al livello del mare. Sappiamo che

  $$g = G\frac{M_T}{R_T^2}$$

  dove R_T é la distanza, al livello del mare, dal baricentro terrestre, M_T é la massa della terra e G la costante di gravitazione universale.

  Anche assumendo di conoscere con precisione G, sicuramente R_T non é ben definito sia perché il livello del mare non é ben definito sia percé il baricentro della terra non é ben definito. Poi c'é l'effetto della rotazione della terra che produce una forza centrifuga diversa in funzione della latitudine: maggiore all'equatore e nulla ai poli. Se vogliamo essere molto precisi c'é anche l'effetto dell'attrazione della luna, quindi il valore esatto di g dipende anche dal giorno e dall'ora in cui viene misurata.
- La realizzazione imprecisa delle condizioni di osservazione. Spesso le osservazioni sono fatte cercando di riprodurre nella pratica esperimenti trattati in modo ideale. I corsi di fisica sono pieni di esempi quali il pendolo inestensibile, piano liscio privo di attrito, gas perfetto, caduta libera di un grave. Nella pratica la realizzazione di un esperimento non é mai ideale e, pur cercando di mantenere un certo controllo sui parametri, alcuni di questi non sono noti con esattezza, ció genera incertezza.
- Le misure effettuate su un campione piuttosto che sull'intera popolazione. quale é la densitá di globuli rossi nel sangue? Quale é la vita media di una specie di farfalle? Quale é il risultato previsto alle prossime elezioni? Nella maggior parte dei casi non é possibile accedere ai dati dell'intera popolazione ma bisogna accontentarsi di un campione. Ovviamente i risultati sono influenzati dalle caratteristiche del campione scelto.
- Gli effetti delle condizioni ambientali possono far si che le diverse realizzazioni dello stesso fenomeno non siano rigorosamente eguali: variazioni di temperatura possono avere effetto sulle misure di densitá di un materiale.

Tutti questi fattori fanno si che la realizzazione di un evento dia risultati non prevedibili con esattezza.

L'analisi statistica ci consente trattare in modo rigoroso e quantitativo i fenomeni aleatori assegnando un valore di probabilitá ai possibili risultati in modo da tenere sotto controllo l'incertezza. In questo modo é possibile effettuare valutazioni quantitative e confronti che sono alla base del metodo scientifico.

Tipicamente lo studio di un fenomeno avviene tramite l'osservazione (**misura**) di una o piú caratteristiche (**caratteri**). Ad esempio volendo studiare la maturazione di un frutto (fenomeno) ne osserviamo (misura) il colore, il peso il sapore (caratteri) ottenendo una seria di valori (**dati**) come in tabella:

Fenomeno	Caratteri	Dati
Maturazione	peso	50 (g)
	colore	giallo
	sapore	aspro

Dal punto di vista *matematico* i caratteri del fenomeno in esame si definiscono mediante *variabili*: una variabile é una legge che associa un *valore* ad un carattere dell'oggetto o del fenomeno in esame. Le variabili possono essere di tipo:

variabile		
Nominale	Ordinale	
	discreta	continua

- Variabili nominali: non c'é una relazione che permette di ordinare a priori i possibili valori in successione: ad esempio il colore degli occhi, il sesso, la religione, la nazionalitá.

- variabili ordinali: i diversi valori si possono mettere in relazione con valori numerici interi (variabili discrete: numero di pagine di un libro, numero di studenti in aula) o con numeri reali (variabili continue: altezza, peso, densitá)

Lo scienziato ha una certa confidenza con i numeri quindi spesso é preferisce ad associare numeri (interi) anche a grandezze nominali, quindi si possono presentare tabelle del tipo:

colore degli occhi			
Nero	Marrone	Verde	Celeste
1	2	3	4

senza che che ció indichi un'effettivo ordine quantitativo. Infatti la tabella potrebbe essere anche riscritta cosí:

colore degli occhi			
Celeste	Marrone	Nero	Verdi
1	2	3	4

senza che ció ne alteri il significato.

Nel seguito indichiamo con lettere maiuscole le variabili e con lettere minuscole corrispondenti i valori (dati) numerici o nominali. Quindi x é un valore che si riferisce alla variabile X. Se si vuole indicare in modo specifico un dato in una serie di dati si indica con un pedice: x_i é l'i-esimo elemento della serie di valori della variabile X.

Proprio per la natura stocastica dei fenomeni fisici spesso, per ridurre l'incertezza, si effettuano piú osservazioni dello stessa variabile in modo da avere una maggior confidenza riguardo al valore osservato. Quindi, come risultato di un'osservazione, abbiamo generalmente una serie di valori numerici o nominali che corrispondono alle osservazioni effettuate.

Ad esempio supponiamo di voler sapere **quanti studenti frequentano** il corso. Posso contare gli studenti presenti oggi in aula, tuttavia potrebbe darsi che proprio oggi, vuoi per sciopero dei mezzi, vuoi per la pioggia o altro molti siano assenti. Quindi per avere un idea piú precisa non mi accontento di una sola misura ma li riconto in diverse occasioni e ottengo una tabella di dati come quella che segue:

Numero di studenti

#	X	#	X	#	X
1	96	11	92	21	91
2	90	12	98	22	73
3	101	13	93	23	90
4	85	14	95	24	87
5	65	15	84	25	96
6	101	16	69	26	82
7	102	17	87	27	79
8	95	18	98	28	80
9	98	19	101	29	89
10	99	20	103	30	86

Le colonne # indicano il numero progressivo della misura (ma potrebbe essere la data), X é la variabile che sto studiando: numero di studenti presenti. Il valore $x_{10} = 99$ é il numero di studenti presenti in aula alla decima osservazione. Questi non sono tutti i possibili valori degli studenti in aula: ho notato che alcuni arrivano tardi, altri vanno via prima. Quindi i dati della tabella rappresentano un "campione" di tutte i possibili valori della variabile **numero di studenti**. Spesso in un osservazione ci si limita ad un campione, si pensi alla misura della densitá di globuli rossi nel sangue di un paziente: sarebbe poco pratico dissanguarlo per avere un conteggio esatto di tutti i globuli rossi.

Una tabella come quella vista sopra ci fornisce un'informazione completa riguardo al campione in esame, tuttavia é poco intuitiva, poco maneggevole (a volte si devono trattare centinaia di valori, se non migliaia) quindi é necessario sintetizzare le informazioni in modo da renderle piú chiare e intuitive, pur mantenendo il piú possibile intatta l'informazione.

Per far questo si usano gli strumenti della statistica descrittiva: tabelle di frequenza, grafici, indici di posizione e variabilitá.

Attenzione: il processo di sintesi dei dati provoca sempre un degrado dell'informazione, deve quindi essere fatto con cura e, a volte, deve essere rivisto alla luce delle conclusioni o di nuove osservazioni. Quindi é bene conservare i dati originali in caso sia necessaria una revisione o integrazione degli stessi.

Esistono diversi strumenti che ci permettono di sintetizzare le informazioni contenute in una serie di dati che possono essere suddivisi in:

- **Qualitativi**: tabelle di frequenza: assolute, relative, relative integrate
- **Grafici**: istogrammi, diagrammi a barre, diagrammi a torta, etc...
- **Quantitativi**: indici statistici di posizione, dispersione e asimmetria

1.2 Tabelle di frequenza

Abbiamo visto come una lunga serie di dati sia difficile da interpretare e come le informazioni siano poco intuitive se presentate in questa forma. La prima operazione di sintesi consiste nel dividere i dati in *classi* e contare il numero di osservazioni in ciascuna classe.

La divisione in classi dipende dall'informazione che si vuole rilevare e evidenziare (anche per questo é sempre bene conservare i dati originati in vista di una eventuale diversa suddivisione). Ad esempio i risultati di un lancio di un dado possono essere raggruppati in base ai valori letti sulla faccia in alto: 1, 2, 3,...6, oppure divisi in pari e dispari (P, D, ovvero 0=pari, 1=dispari). I dati su

un infezione possono essere suddivisi per gravitá: nulla, bassa, media, alta; oppure in base agli effetti: arrossamento, bolle, piaghe.

Pur con una certa libertá di scelta, nel definire le classi bisogna che le seguenti condizioni siano sempre verificate:

- ogni dato deve sempre essere associato ad una e una sola classe;
- le classi devono considerare tutte le possibilitá, ovvero che non ci devono essere dati esclusi.

Una volta definite le classi si costruisce una *tabella di frequenze assolute* ovvero si conta il numero di osservazioni (dati) in ciascuna delle classi definite. Nel seguito indicheremo con N_i il numero di osservazioni (frequenza assoluta) nella classe i-esima. Indicheremo con N_T il numero totale di osservazioni (numero di dati misurati) e indichiamo con K il numero delle classi.

Queste tabelle descrivono come si distribuiscono i dati nelle varie classi, rappresentano cioé la *distribuzione di frequenze assolute* dei dati. Una distribuzione di frequenze assolute fornisce in modo relativamente intuitivo alcune informazioni importanti sul fenomeno in esame come l'intervallo dei valori, i valori piú frequenti o quelli piú rari.

Attenzione a come si definiscono le classi! Consideriamo una suddivisione come questa:

X = N. studenti presenti

X	N	X	N	X	N	X	N
65	1	75	0	85	1	95	2
66	0	76	0	86	1	96	2
67	0	77	0	87	2	97	0
68	0	78	0	88	0	98	3
69	1	79	1	89	1	99	1
70	0	80	1	90	2	100	0
71	0	81	0	91	1	101	3
72	0	82	1	92	1	102	1
73	1	83	0	93	1	103	1
74	0	84	1	94	0	104	0

Notare che la variabile *numero di studenti* é associata alla colonna X mentre N conta il numero di volte che si é osservato un dato valore della variabile. Una tabella come la precedente ci dice poco di piú della tabella originale e rimane ancora abbastanza confusa. Se peró raggruppiamo insieme piú valori allargando le classi e riducendone il numero il vantaggio in termini di chiarezza si vede immediatamente:

X = N. studenti presenti

X	N
61 - 65	1
66 - 70	1
71 - 75	1
76 - 80	2
81 - 85	3
86 - 90	6
91 - 95	5
96 - 100	6
101 - 105	5
106 - 110	0

Con una tabella del genere le informazioni sono piú intuitive e chiare: posso vedere che é abbastanza frequente osservare studenti tra 90 e 100, raramente osservo meno di 80 o piú di 105 studenti.

Le frequenze assolute non sono il modo piú efficace per presentare le osservazioni: se ad esempio se devo confrontare serie di dati in cui il numero di osservazioni é diverso, é piú comodo utilizzare le *frequenze relative* definite come il rapporto tra frequenza assoluta (N_i) e il numero totale di osservazioni (N_T):

$$f_i = \frac{N_i}{N_T}$$

Le seguenti proprietá delle f_i seguono dalla definizione di frequenza relativa:

$$i) \qquad 0 \leq f_i \leq 1$$

$$ii) \qquad \sum_{i=1}^{m} f_i = 1$$

Usando le frequenze relative la tabella precedente diventa:

X = N. studenti presenti

X	f
61 - 65	0.033
66 - 70	0.033
71 - 75	0.033
76 - 80	0.067
81 - 85	0.100
86 - 90	0.200
91 - 95	0.167
96 - 100	0.200
101 - 105	0.167
106 - 110	0.000
$N_T = 30$	

Nota: nel riportare le frequenze relative é bene riportare anche la numerositá del campione.

Per scegliere la ripartizione in classi non c'é una regola precisa: devono essere abbastanza numerose da conservare l'informazione sulla distribuzione dei dati, ma non tante da renderla confusa. Come regola empirica il numero di classi K dovrebbe essere:

$$k \sim \sqrt{N_T}$$

oppure

$$K \sim 1 + \log_2 N_T$$

con N_T numero di dati e K intero. In base al valore di K scelto si calcola l'ampiezza delle classi (intervalli eguali):

$$Delta \sim \frac{x_{max} - x_{min}}{K}$$

Ovviamente si cerca con un po' di buon senso di scegliere gli intervalli in modo ragionevole evitando valori strani o numeri complicati e con lunghe parti decimali. Ad esempio nel caso precedente avremmo:

N_{Tot}	K	x_{min}	x_{max}	Δ
30	5-6	65	103	6-7

Per comoditá abbiamo scelto gli intervalli delle classi a passi di 5 piuttosto che 6 o 7. In questo modo il numero di classi é piú grande di 6 ma spesso scegliere valori di Δ come 2, 5, 10 é piú chiaro che valori tipo 3, 7, 13.

Per indicare le frequenze (relative o assolute) relative ad una data classe posso scrivere:

$$N(91-95) = 5; \quad f(91-95) = 0.167$$

indicando in modo esplicito che per 5 volte ci sono stati tra 91 e 95 studenti presenti ovvero nel 16.7% dei casi ho osservato un numero di studenti tra 91 e 95.

Un modo alternativo per descrivere la distribuzione dei risultati é usare le *frequenze cumulative* o *frequenze integrate*. La frequenza integrata assoluta T di una classe rappresenta il numero di volte che si osserva un valore minore o eguale al limite destro della classe, la frequenza integrata relativa F rappresenta la frazione di valori minori o eguali al valore del limite destro della classe:

$$F_i = \frac{T_i}{N_T}$$

Le frequenze integrate per i dati precedenti sono:

X = N. studenti presenti		
X	T	F
61	0	0.000
65	1	0.033
70	2	0.067
75	3	0.100
80	5	0.167
85	8	0.267
90	14	0.467
95	19	0.633
100	25	0.833
105	30	1.000
110	30	1.000

In tabella abbiamo presentato sia T che F insieme per brevitá, nella pratica basta solo una di esse. Nel caso si riportino i valori relativi é sempre bene indicare la numerositá del campione in esame. In questo caso i valori di X si riferiscono all'estremo destro delle classi.

Anche le frequenze integrate forniscono in modo relativamente intuitivo informazioni sulla distribuzione dei dati: ad esempio $F(65) = 0.033$ indica che solo il 3.3% delle volte ho osservato un numero di studenti minore o eguale a 65; $F(90) = 0.467$ indica che il circa il 50% delle volte ho osservato un numero di studenti minore o eguale a 90. Circa il 73% dei dati é compreso tra 76 e 100. In alcuni casi usare le frequenze integrate (F, T) da informazioni piú intuitive rispetto alle frequenze (N, f).

La determinazione delle frequenze assolute é relativamente piú semplice dal punto di vista del calcolo rispetto al calcolo delle frequenze: una volta ordinati i dati per x crescenti é facile stabilire le frequenze integrate semplicemente contando il numero progressivo Ad esempio dalla tabella con i dati originali ordinati diventa:

X=Numero di studenti presenti					
X	T	X	T	X	T
65	1	87	11	96	21
69	2	89	12	98	22
73	3	90	13	98	23
79	4	90	14	98	24
80	5	91	15	99	25
82	6	92	16	101	26
84	7	93	17	101	27
85	8	95	18	101	28
86	9	95	19	102	29
87	10	96	20	103	30

é facile vedere che la frequenza assoluta della classe 95 é 19, mentre per la classe 75 é 3: nel caso di una serie di dati ordinati il numero progressivo ci dice anche la frequenza assoluta integrata associata a quel valore. Attenzione: se piú dati coincidono la T associata a quel valore é quella dell'ultimo: ad esempio dalla tabella precedente $T(98) = 24$.

Variabili continue

Se abbiamo a che fare con variabili continue le cose si complicano: dal momento che il risultato di un osservazione é a priori un valore reale é impossibile (e sarebbe poco significativo) definire una classe per ogni valore osservato. Quindi, nel caso di una variabile continua una classe si definisce *sempre* in un intervallo di valori della variabile X.

Per evitare sovrapposizioni o esclusioni, gli intervalli devono essere contigui e semiaperti. Ad esempio, volendo riportare in una tabella la distribuzione in frequenza del peso di un campione di frutti, l'i-esima classe é definita dall'intervallo di valori $(x_{i-1}, x_i]$ e comprende tutti i dati che sono *maggiori* di x_{i-1} e *minori o uguali* a x_i. La frequenza della classe N_i o $N(x_i)$ si calcola

contando il numero di osservazioni tali che:

$$x_{i-1} < x \leq x_i$$

Qualora gli intervalli delle classi non siano costanti é bene specificarli nella definizione si frequenza: $N(x_{i-1}, x_i)$, $f(x_{i-1}, x_i)$.

Per variabili continue, dal punto di vista del calcolo, é piú comodo calcolare le frequenze integrate. Vediamo ad esempio il caso di una misura del peso di alcuni frutti. I dati originali sono:

X = Peso frutti

#	X[g]	#	X[g]	#	X[g]
1	14.1	11	12.9	21	14.0
2	11.2	12	9.9	22	13.9
3	15.7	13	9.5	23	19.0
4	18.8	14	12.1	24	14.7
5	18.6	15	12.7	25	14.4
6	20.2	16	8.6	26	13.5
7	8.4	17	13.3	27	20.9
8	14.3	18	13.8	28	17.6
9	18.3	19	15.4	29	22.1
10	11.7	20	13.9	30	13.0

se ordinati per valori crescenti di peso abbiamo:

X = Peso frutti

X[g]	T	X[g]	T	X[g]	T
8.4	1	13.3	11	15.4	21
8.6	2	13.5	12	15.7	22
9.5	3	13.8	13	17.6	23
9.9	4	13.9	14	18.3	24
11.2	5	13.9	15	18.6	25
11.7	6	14.0	16	18.8	26
12.1	7	14.1	17	19.0	27
12.7	8	14.3	18	20.2	28
12.9	9	14.4	19	20.9	29
13.0	10	14.7	20	22.1	30

in cui il numero progressivo é direttamente la frequenza assoluta integrata per il valore i-esimo (a parte per valori ripetuti) ed é quindi immediato calcolare la funzione $T(x)$ per qualunque valore di x: es. $T(13.3) = 11$, $T(20.2) = 28$. Dal momento che la variabile X é una funzione continua, anche la $T(x)$ é una funzione a priori continua, definita anche per valori non presenti in tabella, ad esempio $T(20) = 27$.

La $T(x)$ é una funzione a gradini dal momento che é costante tra due valori contigui. All'aumentare del numero di dati la $T(x)$ diventa sempre piú *liscia* con gradini via via piú ravvicinati.

Dalla definizione della $T(x)$ si calcolano le frequenze relative integrate:

$$F(x_i) = \frac{T(x)}{N_T}$$

dove N_T é il numero totale di osservazioni.

X = Peso frutti [g]

X[g]	F	X[g]	F	X[g]	F
8.4	0.033	13.3	0.367	15.4	0.700
8.6	0.067	13.5	0.400	15.7	0.733
9.5	0.100	13.8	0.433	17.6	0.767
9.9	0.133	13.9	0.467	18.3	0.800
11.2	0.167	13.9	0.500	18.6	0.833
11.7	0.200	14.0	0.533	18.8	0.867
12.1	0.233	14.1	0.567	19.0	0.900
12.7	0.267	14.3	0.600	20.2	0.933
12.9	0.300	14.4	0.633	20.9	0.967
13.0	0.333	14.7	0.667	22.1	1.000

La tabella delle frequenze integrate relative ci permette di ottenere in modo veloce alcune informazioni sulla distribuzione dei dati: il 50% dei frutti pesa meno di 13.9 [g], circa 80% dei frutti ha un peso tra 9.5 e 19 [g], é raro trovare frutti che pesano meno di 8.5 g o piú 21 g.

Una volte definiti gli estremi delle classi é relativamente semplice costruire una tabella di frequenze (assolute o relative). As esempio, nel caso precedente $N_T = 30$, $x_{min} = 8.4$, $x_{min} = 22.4$, quindi K $5 - 6$ e possiamo definire $\Delta = 2$. Abbiamo quindi la tabella seguente:

X = Peso frutti [g]

X[g]	N	f
≤8	0	0.000
(8-10]	4	0.133
(10-12]	2	0.067
(12-14]	9	0.300
(14-16]	7	0.233
(16-18]	1	0.033
(18-20]	4	0.133
(20-22]	2	0.067
(22-24]	1	0.033
>24	0	0.000

dove le frequenze assolute sono:

$$N(x_{i-1}, x_i) = T(x_i) - T(x_{i-1})$$

e le frequenze relative:

$$f(x_i) = \frac{T(x_i) - T(x_{i-1})}{N_T} = F(x_i) - F(x_{i-1})$$

Utilizzando un foglio elettronico (EXCEL o OpenOffice) si puó usare la funzione:

`Frequenza`(`matrice_dati`, x)

che, data una serie di dati (`matrice_dati`), calcola il numero di valori minori o eguali al valore x ovvero la frequenza assoluta integrata $T(x)$. L'analoga funzione della distribuzione OpenOffice é la funzione `Frequency` con lo stesso formato.

In Excel la funzione `Frequenza` puó essere utilizzata in "forma di matrice" per fornire direttamente le frequenze assolute, piuttosto che i valori integrati. Per far questo:

- **Selezionare** le celle in cui inserire i valori delle frequenze, chiaramente rispettando il numero di classi definito.
- **inserire**: `=Frequenza(xxxx,yyy)` , attenzione usare "," o ";" per separare gli argomenti a seconda delle impostazioni di sistema.
- **xxx**: matrice dei dati, **yyy**: matrice delle classi.
- inviare il comando premendo `Ctrl + shift + Enter`.

Riassumendo, per preparare una tabella di frequenze bisogna:

- calcolare i valori estremi dei dei valori ottenuti: x_{min}, x_{max} e il numero totale di osservazione N_T;
- definire il passo Δ delle classi;
- calcolare gli estremi delle classi;
- calcolare le frequenze assolute integrate T;
- calcolare N o f e preparare la tabella;
- indicare chiaramente: il tipo di dato con l'unitá di misura, se si usano F o f e, in tal caso, indicare il numero totale di osservazioni N_T.

1.3 Istogrammi di frequenza e diagrammi a barre

La presentazione sotto forma di tabelle é sicuramente utile quando il numero delle classi é limitato. Quando, come spesso accade, il numero delle classi é elevato, le tabelle non riescono sintetizzare in modo abbastanza intuitivo e chiaro le informazioni e diventa piú difficile il confronto tra diversi campioni. In questi casi é meglio ricorrere a metodi grafici per sintetizzare le informazioni contenute nei dati poiché questi consentono una visione piú immediata e globale delle informazioni e anche il confronto tra diverse osservazioni é piú semplice e immediato.

Esistono diversi metodi per presentare graficamente una distribuzione di frequenza, i piú usati sogno gli *istogrammi di frequenza* e i *diagrammi a barre*. Diagrammi a torta o circolari e poligoni di frequenza sono altri tipi di strumenti grafici spesso utilizzati per presentare le distribuzioni dei dati di una variabile.

Un diagramma a barre é utilizzato per rappresentare la frequenza (assoluta, relativa o integrata) di classi nominali o discrete. Consiste in una serie di colonne, tante quante sono le classi (modalitá del carattere), la cui altezza é proporzionale alla frequenza della classe. Il diagramma a barre si rappresenta in un sistema di assi cartesiani: sull'asse delle ascisse (x, orizzontale) si riportano le etichette delle classi, sull'asse delle ordinate (y, verticale) la scala delle frequenze (assolute, relative o integrate).

L'istogramma é un metodo grafico per presentare la distribuzione di frequenze di una variabile continua. L'istogramma é simile al diagramma a barre, tanto che spesso sono considerati sinonimi: é costituito da rettangoli adiacenti la cui base é eguale all'ampiezza della classe e l'altezza é proporzionale al rapporto tra la frequenza e l'ampiezza della classe. Ovvero indicando con Δ_i l'ampiezza della classe i-esima e con

$$h_i = \alpha\ N_i/\Delta_i$$

l'altezza rettangolo corrispondente, l'area del rettangolo

$$A_i = \alpha N_i$$

é proporzionale alla frequenza della classe cui si riferisce. L'altezza dei rettangoli é quindi proporzionale alla *densitá di frequenza* della classe.

Nota: nel paragrafo precedente abbiamo definito le frequenze relative:

$$f(x_{i-1}, x_i) = F(x_i) - F(x_{i-1})$$

Dove x_i indica l'estremo di una classe.

Abbiamo anche visto che la $F(x)$ é una funzione continua definita per ogni valore di x. Supponendo di avere un numero molto grande di dati possiamo immaginare di scegliere intervalli molto piccoli, al limite infinitesimi dx, ma mantenendo finito il rapporto:

$$\frac{F(x) - F(x - dx)}{dx}$$

In questo caso posso definire:

$$f(x) = \lim_{dx \to 0} \frac{F(x) - F(x - dx)}{dx} = \frac{dF(x)}{\delta x}$$

ovvero la $f(x)$ é la derivata della $F(x)$ e rappresenta la densitá di frequenza relativa nel punto x. Quindi $f(x)dx$ é la frequenza relativa di osservazioni nell'intervallo infinitesimo tra x e $x+dx$. Con questa definizione si ha:

$$F(x) = \int_{x_{min}}^{x} f(x)dx \quad T(x) = N_T \int_{x_{min}}^{x} f(x)dx$$

Anche se al momento non é essenziale, é bene ricordarsi di questa definizione delle $f(x)$ come "densitá di frequenza" che é l'analogo della densitá di probabilitá per una distribuzione continua, che vedremo piú avanti.

Per evitare confusione é sempre preferibile utilizzare classi di eguale ampiezza, in questo modo la densitá di frequenza é costante e l'altezza dei rettangoli é proporzionale alla frequenza delle classi. In questo modo istogramma e diagramma a barre sono effettivamente sinonimi.

Come per una tabella, anche per un istogramma (diagramma a barre) la scelta delle classi non é univoca, in

particolare nel caso di caratteri continui, ma va fatta con una certa cura: se si riduce troppo Δ si hanno frequenze basse e molti intervalli vuoti, l'istogramma diviene rumoroso e poco intuitivo. Al contrario se si scelgono classi troppo ampie si mascherano dettagli che possono essere importanti. Il numero delle classi K e l'ampiezza delle stesse (Δ) si sceglie quindi in base al numero totale di osservazioni (N_T) e dall'intervallo di variazione del carattere R usando regole empiriche e buon senso. Come abbiamo visto per le tabelle si puó seguire la regola empirica per determinare il numero delle classi K:

$$K \sim \sqrt{N}$$

oppure

$$K \sim 1 + \log_2 N$$

e da questo calcolare l'ampiezza delle classi:

$$\Delta \sim \frac{x_{max} - x_{min}}{K}$$

scegliendo poi valori in modo da rendere i grafici semplici e leggibili, in particolare evitando valori con troppi decimali o troppe cifre.

Nel presentare un istogramma bisogna ricordarsi di specificare correttamente cosa si sta graficando in modo da rendere semplice la lettura e l'interpretazione. In particolare bisogna:

- Indicare il tipo di dato e l'unitá di misura (asse X)
- indicare cosa si sta graficando (T, F, N, f)
- indicare con opportune etichette dati appartenenti a set diversi per facilitare i confronti.

Quando si usa Excel le frequenze sono calcolate utilizzando l'estremo destro delle classi. Questo puó dar luogo a qualche confusione se si utilizza l'indice utilizzato da Excel come etichetta delle classi dal momento che sul grafico sembra che questi valori siano associati al centro delle classi. Conviene definire una nuova colonna in cui inserire i valori corretti da usare come matrice delle "etichette".

1.4 Sintesi numerica

Grafici e tabelle sono strumenti molto utili per sintesi e confronto di dati. Tuttavia le informazioni che si possono ricavare in questo modo sono spesso solo qualitative o semi quantitative. La statistica descrittiva utilizza una serie di indicatori numerici per sintetizzare e caratterizzare in modo quantitativo una distribuzione. Questi indicatori si dividono in indici di posizione, indici di dispersione e indici di asimmetria. E' evidente che questi indici si possono applicare solamente a grandezze ordinali, siano essi continui o discreti, mentre non si possono definire per caratteri nominali.

Supponiamo di avere un campione di N dati per una grandezza ordinale: x_1, x_2, ... x_N. La **media** dei dati si indica con $\bar{x}$ ed é definita come:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

La **mediana** é quel valore che divide a metá la distribuzione dei dati, cioé si hanno 50% dei valori minori della mediana e 50% dei valori maggiori. Se ordiniamo i dati in ordine crescente la mediana é quel valore che divide in due la distribuzione ordinata dei dati. Nel caso di distribuzioni simmetriche media e mediana coincidono. Se i dati presentano una coda a sinistra la media e a sinistra della mediana, se i dati presentano uno scodamento a destra la media é a destra della mediana.

La **moda** é quel valore che si presenta con maggior frequenza nella distribuzione dei valori. Nel caso che la distribuzione dei dati presenti piú massimi allora la moda non é ben definita, si chiamano distribuzioni plurimodali.
`MEDIA(matrice)`
`MODA(matrice)`
`MEDIANA(matrice)`
sono funzioni predefinite in Excel e Open Office, dove `matrice` indica l'insieme delle celle contenenti i dati

I principali indici di dispersione sono: **intervallo** di variazione (o range), **distanza interquartile**, **varianza**, **deviazione standard**, **coefficiente di variazione**.

L'*intervallo di variazione* dei dati é la differenza tra valore massimo e valore minimo

$$R = x_{max} - x_{min}$$

`MAX(matrice)` e `MIN(matrice)` sono funzioni predefinite di Excel e OpenOffice che calcolano rispettivamente il valore massimo e il valore minimo di un insieme di dati.

Il primo quartile (Q_1) é il valore per cui il 25% dei dati é minore o uguale a Q_1, il secondo quartile (Q_2) é il valore per cui il 50% dei dati é minore o uguale a Q_2 (Q_2 = Mediana), il terzo quartile (Q_3) é il valore per cui il 75% dei dati é minore o uguale a Q_3. La distanza interquartile é definita come:

$$\Delta Q = Q_3 - Q_1$$

Segue dalla definizione che il 50% dei dati cade tra il primo e il terzo quartile.

La funzione Excel `QUARTILE(matrice,quarto)` calcola il valore del quartile `quarto` (=1, 2, 3) per una matrice di dati.

La **varianza** campionaria é definita come il valor medio del quadrato degli scarti, ovvero:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{2}$$

La **deviazione standard** é la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Le funzioni Excel `VAR(matrice)` e `DEV.ST(matrice)` calcolano la varianza e la deviazione standard di un insieme di dati (`matrice`).

Gli indici statistici che abbiamo visto fin qui si calcolano utilizzando la serie dei dati cosi come sono stati acquisiti, prima di eventuali raggruppamenti in classi. A volte abbiamo a che fare non con i dati di origine, che come abbiamo detto occupano lunghi elenchi, ma con dati sintetizzati in tabelle di frequenza o istogrammi.

Supponiamo di avere un campione di N dati per una grandezza ordinale discreta e supponiamo di avere osservato k valori x_1, x_2, ... x_k con frequenze N_1, N_2, ...N_k. Per calcolare la media usiamo la definizione precedente (1) ma invece di sommare i valori ripetuti possiamo moltiplicare ognuno dei k valori per la sua frequenza assoluta:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{k} N_j x_j \tag{3}$$

con $N = \sum_j N_j$ che puó essere riscritta utilizzando le frequenze relative:

$$\bar{x} = \sum_{j=1}^{k} f_j x_j \tag{4}$$

In questo caso si parla di media "pesata", ovvero é la media dei valori delle diverse classi osservate, ogni valore é pesato per la frequenza della classe.

Per il calcolo della varianza si segue un ragionamento analogo a partire dalla definizione di varianza:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \tag{5}$$

$$= \frac{N}{N-1}\sum_{j=1}^{k} f_j (x_j - \bar{x})^2 \tag{6}$$

dove il termine $\frac{N}{N-1} \sim 1$ se N é abbastanza grande.

Quando abbiamo a che fare con classi che raggruppano piú valori si usa, come riferimento per il calcolo di media e varianza, il valore medio dell'intervallo della classe. Consideriamo dati che si riferiscono ad una grandezza continua. Nella tabella di frequenze N_j conta gli eventi che hanno valore compreso nell'intervallo assegnato alla classe j-esima: $x_{j-1} \leq x_i < x_j$. Se ora abbiamo la tabella con la distribuzione di frequenze e non i dati originali possiamo ancora calcolare la media e la varianza utilizzando le formule appena viste per la media:

$$\bar{x} = \sum_{j=1}^{k} f_j \bar{x}_j \tag{7}$$

dove $\bar{x}_j$ é il punto medio dell'intervallo che definisce la classe j-esima: $\bar{x}_j = x_{j+1} - x_j$. Per la varianza:

$$\sigma^2 = \frac{N}{N-1}\sum_{j=1}^{k} f_j (\bar{x}_j - \bar{x})^2 \tag{8}$$

dove il rapporto $(N/N-1)$ puó essere considerato 1 se N é abbastanza grande.

Come esempio riprendiamo una delle tabella viste sopra, per calcolare la media dobbiamo utilizzare i valori medi $\bar{x}$ degli intervalli delle classi:

X = Peso frutti [g]

X[g]	$\bar{x}$[g]	N	f
≤8		0	0.000
(8-10]	9	4	0.133
(10-12]	11	2	0.067
(12-14]	13	9	0.300
(14-16]	15	7	0.233
(16-18]	17	1	0.033
(18-20]	19	4	0.133
(20-22]	21	2	0.067
(22-24]	23	1	0.033
>24		0	0.000

Attenzione perché Excel e Oo utilizzano l'estremo destro per il calcolo delle frequenze e, se non sene tiene conto, si rischia di sovrastimare il valore medio di $\Delta/2$

E' intuitivo che data una serie di dati, diverse suddivisioni in classi possono dar luogo a valori diversi della media e varianza calcolate sui dati ragggruppati. Questo effetto ci fa vedere come nel passare da dati grezzi alle tabelle di frequenze l'informazione non si conservi perfettamente ma sia parzialmente degradata.