2 Statistica descrittiva

2.1 La raccolta dei dati

Per descrivere un esperimento è necessario saper distinguere i vari tipi di dati e i diversi possibili livelli di misurazione dei loro valori. I dati, cioè i risultati osservati di fenomeni o caratteristiche, possono essere:
Qualitativi: danno luogo a risposte non numeriche (sì/no, maschio/femmina, ...)
Quantitativi: danno luogo a risposte numeriche (numero abbonamenti, altezza, peso,...) che si suddividono in:

- discreti: forniscono risposte che derivano da un processo di conteggio
- continui: forniscono risposte che derivano da un processo di misurazione

2.2 I campioni

Abbiamo definito campione la parte di una popolazione che si seleziona per l'analisi. Esistono fonda-

mentalmente due tipi di campioni: quelli non probabilistici, in cui gli oggetti o individui sono inclusi senza tener conto della loro probabilità di appartenere al campione e quelli probabilistici, in cui gli oggetti sono scelti sulla base di probabilità note. Il campionamento probabilistico risulta essere il solo metodo che consente di ottenere inferenze corrette sulla base di un campione.

I tipi di campionamento probabilistico maggiormente usati sono:

- il campionamento casuale semplice (ogni oggetto o individuo della popolazione ha la stessa probabilità, nota e costante, di essere selezionato) (con o senza ripetizione)

- il campionamento sistematico (si divide la numerosità N della popolazione per la dimensione desiderata n del campione e $k = N/n$ è chiamato passo del campionamento. Si sceglie un oggetto ad ogni k, iniziando con l'estrazione a caso di un oggetto tra 1 e k)

- il campionamento stratificato (gli oggetti della popolazione vengono suddivisi in sottopopolazioni distinte ed omogenee, chiamate *strati*, sulla

base di una caratteristica comune; si conduce un campionamento casuale semplice in ogni strato e poi i risultati dei singoli campionamenti vengono accorpati in un campione complessivo)

- il campionamento a grappolo (gli oggetti della popolazione sono suddivisi in molti gruppi, detti *grappoli*, in modo che ogni grappolo sia rappresentativo dell'intera popolazione. Si estrae poi un campione casuale di grappoli e gli oggetti di ogni grappolo sono inclusi nel campione complessivo)

2.3 Organizzazione dei dati quantitativi

Quando il numero delle osservazioni è elevato, diventa utile sintetizzare i dati ricorrendo a tabelle e grafici in grado di fornire informazioni sul fenomeno che si sta analizzando.
L'ordinamento: utile per individuare i valori estremi, quelli più frequenti, quelli attorno ai quali c'è una maggior concentrazione di un carattere (p.es. ordine crescente);
Il diagramma ramo-foglia: utile per individuare intorno a quali valori si concentrano le osservazio-

ni; si costruisce dividendo ogni osservazione nella sua parte principale (il ramo dell'albero) e in quella secondaria (le foglie dell'albero).
Esempio: salario mensile in euro di un campione di otto impiegati:

$$555, 490, 648, 832, 710, 590, 576, 627$$

Come rami si considera la colonna delle centinaia e quella delle decine come le foglie dell'albero, trascurando la colonna delle unità:

4|9
5|597
6|42
7|1
8|3

Poichè questi metodi si rivelano inadeguati all'aumentare del numero di osservazioni è opportuno utilizzare:
La distribuzione delle frequenze assolute: una tabella in cui i dati sono organizzati in opportune classi di raggruppamento.

È importante la scelta del numero delle classi, cioè il numero opportuno di classi in cui i dati devono essere raggruppati (solitamente min 5, max 15) e l'ampiezza di ogni classe. In caso di classi di uguale ampiezza la lunghezza di ogni intervallo si ottiene:

$$\text{Ampiezza intervallare} = \text{range}/\text{numero delle classi}$$

dove il range è l'intervallo dei valori che la variabile considerata assume. Bisogna poi stabilire quali sono gli estremi delle classi per evitare sovrapposizioni e scegliere il punto medio di ogni classe come rappresentativo dei valori che cadono all'interno dell'intervallo considerato.
Per migliorare ulteriormente la comprensione dei dati si ricorre a due varianti:
La distribuzione delle frequenze relative: una tabella in cui si rapportano le frequenze assolute al numero di osservazioni.
La distribuzione delle percentuali: una tabella in cui si moltiplica per 100 ogni frequenza relativa.
Usare una base pari a 1 (frequenze relative) o pari a 100 (percentuali) diventa essenziale quando bisogna confrontare più insiemi di dati, ognuno con un numero differente di osservazioni.

Un altro metodo di presentare i dati è:
La distribuzione cumulativa: una tabella che si ottiene a partire sia dalle frequenze assolute, che relative, che percentuali. La frequenza cumulativa, nel caso si utilizzino per esempio i dati di una tabella delle frequenze assolute, è la somma totale della frequenza assoluta dell'elemento della serie preso in esame e di tutte le frequenze assolute dei valori che lo precedono.

I grafici adatti alla rappresentazione di dati numerici, sintetizzati in distribuzioni di frequenza o percentuali sono:
L'istogramma: diagramma a barre verticali rettangolari aventi come base gli intervalli in cui sono state raggruppate le osservazioni e come altezza il numero, o la proporzione o la percentuale di osservazioni che cadono in ogni classe.

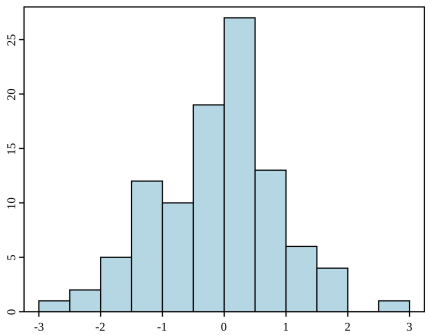

Figura 2: Istogramma

Il poligono: si costruisce scegliendo il punto medio di ogni classe per rappresentare tutte le osservazioni che cadono in quella classe, si associa ad ogni punto medio, in ordinata, il numero, o la proporzione o la percentuale di osservazioni per ogni intervallo di raggruppamento e si congiungono tali punti con linee rette.

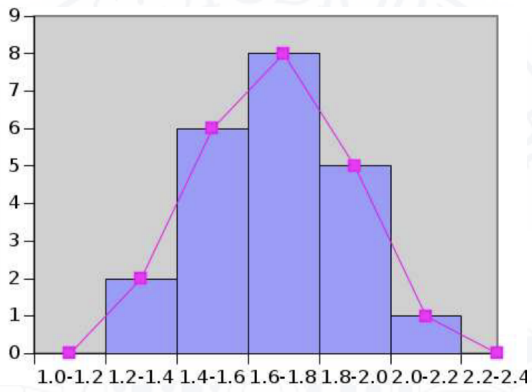

Figura 3: Poligono

È possibile rappresentare graficamente anche la tabella delle frequenze cumulate attraverso un grafico chiamato *il poligono cumulativo o ogiva*.

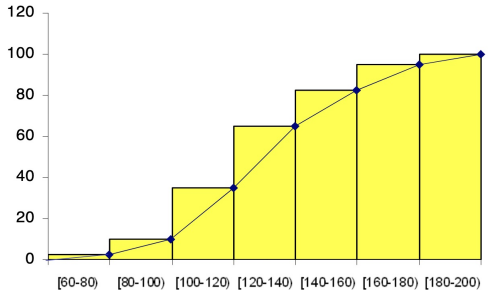

Figura 4: Ogiva

2.4 Organizzazione dei dati qualitativi

Anche i dati qualitativi possono essere sintetizzati in forma tabellare con le stesse caratteristiche delle tabelle delle frequenze viste in relazione ai dati quantitativi. Da un punto di vista grafico è possibile rappresentare i caratteri qualitativi in vari modi, per esempio attraverso:
Il diagramma a barre: ciascuna barra del diagramma rappresenta una modalità del dato e la sua lunghezza è proporzionale alla frequenza o alla percentuale di osservazioni.

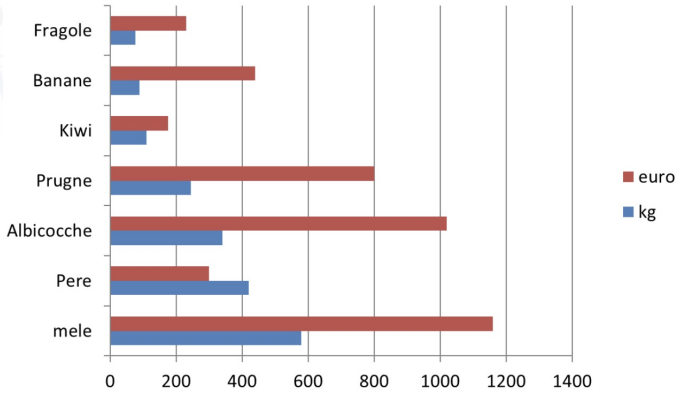

Figura 5: Barre

Il diagramma a torta: è uno strumento grafico ampiamente utilizzato suddividendo l'angolo a 360 gradi in fette la cui dimensione è proporzionale alla percentuale di osservazioni che cadono in ogni categoria.

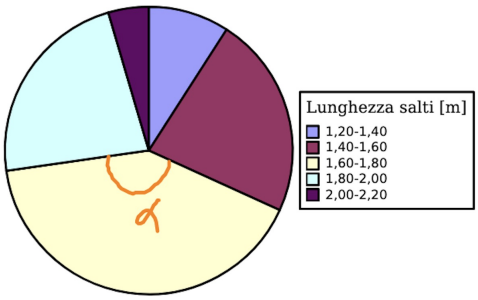

Figura 6: Torta

Entrambi i diagrammi hanno lo stesso scopo: rendere immediatamente leggibile la distribuzione delle frequenze di tali dati.

2.5 Sintesi e descrizione dei dati quantitativi

La rappresentazione dei dati, sebbene sia una componente essenziale della statistica descrittiva, non è di per sè esaustiva. Quando si considerano dati quantitativi, non è sufficiente presentare adeguatamente i dati e trarre indicazioni su questi a partire dall'osservazione di tali rappresentazioni. Una buona analisi dei dati richiede anche che le caratteristiche principali delle osservazioni siano sintetizzate con misure opportune e che tali misure siano adeguatamente analizzate o interpretate.

Molte sono le misure che rappresentano le caratteristiche di posizione, di variabilità e di forma. Queste

misure di sintesi sono chiamate *statistiche* se calcolate sulla base di un campione, *parametri* se calcolate a partire dall'intera popolazione.

2.6 Misure di posizione

Nella maggior parte degli insiemi di dati, le osservazioni mostrano una tendenza a raggrupparsi attorno ad un valore centrale e in generale risulta possibile selezionare un valore tipico per descrivere l'intero insieme di dati. Questo valore è una misura di **posizione** o di **tendenza centrale**. Le misure di posizione più comuni sono la *media (aritmetica)*, la *mediana*, la *moda*.

La media: si calcola dividendo la somma dei valori osservati per il numero totale di osservazioni:

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

La media si presenta come un punto di equilibrio, ma poichè si basa su tutte le osservazioni, tale misura risulta influenzata dai valori estremi, che possono fornire una rappresentazione distorta dei dati.

Si può ricorrere ad un'altra misura che non sia influenzata dalle osservazioni estreme di un insieme di dati.

La mediana: è il valore centrale in una successione ordinata di dati, che lascia alla destra il 50% delle osservazioni e a sinistra il 50% delle osservazioni.

Se l'ampiezza del campione è un numero n dispari, la mediana coincide col valore centrale e occupa la posizione $(n+1)/2$ nella serie ordinata di dati. Se l'ampiezza del campione è un numero n pari, il valore centrale si trova a metà tra le due osservazioni centrali della serie ordinata e la mediana coincide con la media dei valori corrispondenti a queste due osservazioni centrali.

Quando si calcola la mediana non si tiene conto delle eventuali ripetizioni dei dati. Il suo calcolo è influenzato dal numero delle osservazioni, non dalla grandezza dei valori estremi.

La moda: è il valore più frequente in un insieme di dati.

Non è influenzata dai valori estremi, ma viene usata per soli scopi descrittivi. Un insieme di dati può non aver moda, oppure avere più mode.

Le misure di posizione "non centrale" maggiormente usate sono i *quartili* (caso particolare di una categoria di misure dette *quantili*), per descrivere o sintetizzare le caratteristiche di insiemi di dati quantitativi molto ampi.
I quartili: sono misure descrittive che dividono i dati ordinati in quattro parti.
Il *primo quartile* Q_1 è il valore tale che il 25% delle osservazioni è più piccolo di Q_1 e il 75% è più grande di Q_1 e corrisponde alla $(n+1)/4$ osservazione della serie ordinata di dati.
Il *terzo quartile* Q_3 è il valore tale che il 75% delle osservazioni è più piccolo di Q_3 e il 25% è più grande di Q_3 e corrisponde alla $3(n+1)/4$ osservazione della serie ordinata di dati.

Varie sono regole usate per calcolarli. Altri quantili spesso usati sono i decili (dati ordinati divisi in dieci parti) e i percentili (dati ordinati divisi in cento parti).

2.7 Misure di variabilità

Una seconda caratteristica importante di un insieme di dati è la variabilità, cioè la quantità di *disper-*

sione presente in essi. Tra le varie misure di variabilità, le più importanti sono il *range*, la *varianza*, lo *scarto quadratico medio*.
Il range o intervallo di variazione: è la differenza tra l'osservazione più grande e quella più piccola in un insieme di dati. Rappresenta una misura della dispersione totale nell'insieme di dati e il suo limite consiste nel fatto che non tiene conto di come si distribuiscano i dati tra il valore più piccolo e quello più grande.
La varianza: è la somma dei quadrati degli scarti di ogni osservazione dalla media divisa per $n-1$:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

La varianza sintetizza la dispersione dei dati osservati attorno alla loro media.
Lo scarto quadratico medio o *deviazione standard*: è la radice quadrata della varianza.

Esso ci aiuta a stabilire se e quali dati sono concentrati o dispersi intorno alla loro media. Per quasi tutti gli insiemi di dati, la maggior parte dei valori osservati si trova nell'intervallo centrato sulla me-

dia e i cui estremi distano dalla media per 1 scarto quadratico medio. Pertanto la conoscenza della media e dello scarto quadratico medio in genere aiuta a definire in quale intervallo si concentra almeno la maggior parte dei dati osservati.

L'unità di misura della varianza è il quadrato dell'unità di misura dei dati e per tale motivo si preferisce usare come misura di variabilità lo scarto quadratico medio. Nè la varianza nè lo scarto quadratico medio possono essere negativi. Possono essere entrambi nulli nel caso in cui non ci sia variabilità nei dati, cioè quando le osservazioni sono uguali tra loro. In tal caso anche il range è uguale a zero.

Osservazione: È importante sottolineare che nella definizione di varianza non si può usare come numeratore la quantità:

$$\sum_{i=1}^{n}(X_i - \bar{X}_n) = \sum_{i=1}^{n} X_i - n\bar{X}_n \equiv 0$$

poichè la somma degli scarti dalla media è sempre nulla, essendo la media il punto di equilibrio tra le osservazioni.

2.8 La forma

La terza caratteristica dei dati è la *forma* della loro distribuzione. Essa può essere *simmetrica* o meno. In tal caso si dice *asimmetrica* o *obliqua*.

Per descrivere la forma di una distribuzione di dati è sufficiente confrontare la media con la mediana. Se sono uguali allora la distribuzione è simmetrica; se la media è maggiore della mediana allora la distribuzione si dice obliqua a destra; se la media è più piccola della mediana la distribuzione si dice obliqua a sinistra.

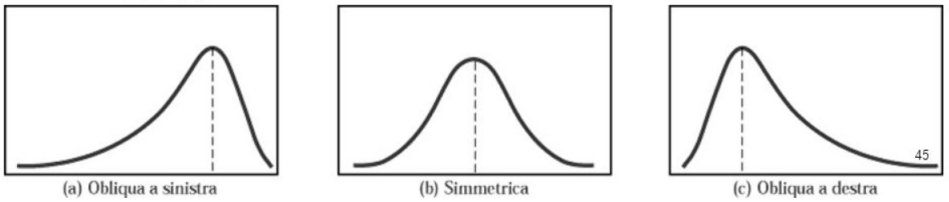

Figura 7: Forma della distribuzione

Se i dati hanno distribuzione obliqua a destra (c), essi si raggruppano alla sinistra, cioè al di sotto della mediana: la lunga coda a destra è dovuta alla presenza di valori estremamente grandi che attirano la media verso l'alto rendendola più grande della mediana.

Per dati con distribuzione obliqua a sinistra (a) le osservazioni tendono a concentrarsi a destra, cioè al di sopra della mediana: la lunga coda a sinistra è dovuta alla presenza di valori estremamente piccoli che attirano la media verso il basso rendendola più piccola della mediana.

Per dati distribuiti in modo simmetrico (b) media e mediana coincidono e le osservazioni tendono a concentrarsi attorno a queste misure di posizione. In tale situazione possiamo usare la cosiddetta regola empirica per esaminare la variabilità dei dati e per analizzare il significato dello scarto quadratico medio.

La regola empirica

La regola empirica afferma che, nella maggior parte degli insiemi di dati, circa il 67% delle osservazioni si trova ad una distanza dalla media pari ad 1 volta lo scarto quadratico medio, circa il 95% delle osservazioni si trova ad una distanza dalla media pari a 2 volte lo scarto quadratico medio, circa il 99% delle osservazioni si trova ad una distanza dalla media pari a 3 volte lo scarto quadratico medio.