QUADERNI DEL DIPARTIMENTO DI SCIENZE DELL'UOMO

DSU 03/2005

Giovanni Delli Zotti (*dellizottig@sp.units.it*)

ANALISI E SINTESI DI UNA VARIABILE

Maggio 2005

Università degli Studi di Trieste
www.dsu.units.it

Quaderni del Dipartimento di Scienze dell'Uomo
Università degli Studi di Trieste
DSU: 03/2005
(www.dsu.units.it)

Analisi e sintesi di una variabile

di Giovanni Delli Zotti

Sommario

Premessa

La matrice dei dati è una "macro tabella" usata per registrare informazioni che sono state raccolte con tecniche a volte assai diverse. Tali informazioni diventano dati nel momento in cui vengono trascritte nelle celle della matrice, dopo essere state eventualmente codificate[1]. I dati contenuti in una singola colonna della matrice costituiscono l'insieme delle rilevazioni effettuate per tutti i casi su una singola proprietà. I dati sono a volte decine, qualche volta centinaia o migliaia, dunque si pone il problema di sintetizzare questa mole imponente di dati analiticamente registrati. L'esame più semplice che possiamo effettuare sui dati è denominato analisi monovariata, per effettuare la quale sono disponibili numerose tecniche che verranno illustrate in questa sede.

Nell'esempio che proponiamo è stata scelta una variabile cardinale perché le caratteristiche di questo tipo di variabili rendono virtualmente possibile l'illustrazione di tutte le tecniche di analisi disponibili. Le variabili nominali sono semplici elenchi di categorie (ad esempio le professioni, oppure i comuni di residenza) ed i codici numerici che si usano per registrare in matrice gli stati su queste variabili non godono in realtà di alcuna proprietà dei numeri. Se le categorie sono ordinate (variabili ordinali, come ad esempio il livello d'istruzione), si possono applicare solamente le proprietà ordinali dei numeri ("maggiore di" o "minore di"), ma non le operazioni aritmetiche di somma, sottrazione, moltiplicazione e divisione. È legittimo invece usare tutti gli strumenti statistici con le variabili cardinali, che derivano dalla rilevazione mediante conteggio o misurazione di proprietà discrete o continue come, ad esempio, il reddito, l'età o il numero dei figli)[2].

Solo in casi molto particolari è possibile trasformare una variabile nominale in una variabile cardinale o quasi-cardinale e, come vedremo, è invece possibile ridurre ad una serie di ca-

[1] Si veda l'articolo *Tipologia delle matrici utilizzate nella ricerca sociale* (Delli Zotti 1985), disponibile online all'indirizzo: www.uniud.it/dest/docenti/dellizotti/matrix.pdf .

[2] Sui tipi di proprietà, di variabili, e sulle diverse definizioni operative, si veda *Introduzione alla ricerca sociale. Problemi e qualche soluzione* (Delli Zotti 2004) e, per maggiori approfondimenti, *Il linguaggio delle variabili* (Cardano e Miceli, 1991).

tegorie (ordinate) una variabile cardinale. Perciò, si possono usare tecniche adatte all'analisi di una varabile categoriale partendo da una variabile cardinale, ma non viceversa.

Il proposito di questo scritto è mostrare la varietà di soluzioni disponibili anche con gli strumenti di analisi più semplici, quelli dell'analisi monovariata, ma sarebbe riduttivo non mostrare altre tecniche di analisi, peraltro disponibili solo tra quelle solitamente classificate all'interno dell'analisi bivariata. Per questo motivo spesso "sconfineremo" in quella direzione. Ciò verrà fatto, tra l'altro, nella consapevolezza che la distinzione tra analisi monovariata e bivariata non è poi così nitida come appare. Si può sostenere, infatti, che si tratta di una classificazione che ha un valore "locale", e cioè vale solo con riferimento ad una specifica matrice dei dati. Infatti, se replichiamo su un campione rappresentativo di popolazione residente in Friuli-Venezia Giulia un'indagine riguardante la fiducia verso i partiti svolta in precedenza nel Lazio, e confrontiamo le risposte fornite nelle due rilevazioni mediante un grafico a barre, possiamo dire che stiamo rappresentando due distribuzioni monovariate. Lo stesso tipo di dati può però essere rilevato in un'indagine nazionale, nella quale Lazio e Friuli-Venezia Giulia sono due categorie della variabile "regione di residenza". Per mettere a confronto le distribuzioni, dobbiamo, in questo caso, costruire una tabella di contingenza (strumento statistico dell'analisi bivariata), incrociando la regione di residenza con le domande relative alla fiducia.

Un altro criterio che adotteremo nella trattazione è quello del passaggio da tecniche che consentono l'analisi con la minima perdita dell'informazione contenuta nella matrice dei dati, a tecniche che sintetizzano "pesantemente" i dati originari, fino a ridurre l'intera distribuzione di frequenza ad un unico valore riassuntivo. Le due situazioni sono identificabili nella distribuzione di frequenza, ad un estremo, e nella media aritmetica, dall'altro. In realtà, come si vedrà nel prossimo paragrafo, c'è una tecnica di analisi che, rispetto alla distribuzione di frequenza, è ancor più conservativa dei dati originali.

Uno dei miei intenti è quello di mostrare come le tecniche di analisi, tutto sommato, siano in qualche misura equivalenti tra loro, anche se è certamente corretto, e spesso più efficace, usare quella di volta in volta più adatta per il tipo di variabile sul quale si sta lavorando (i tipi di variabili, se stiamo effettuando un'analisi bivariata). Ma, tutto sommato, esiste una tale varietà di tecniche che appare un po' stucchevole il sussiego con il quale vengono trattati coloro che non praticano con assiduità tutto l'armamentario statistico, dimenticando che ciò che davvero conta è formulare un interessante quesito scientifico e raccogliere dati validi ed attendibili da cui cercare di ottenere risposte non equivoche.

C'è, peraltro, un motivo tecnico che induce a consigliare versatilità nell'uso delle diverse tecniche di analisi e nella capacità di eventualmente operare adeguate trasformazioni delle variabili originarie. Le tecniche di analisi multivariata, infatti, richiedono in genere che le variabili siano dello stesso tipo ed è perciò importante saper trasformare un tipo di variabile in un altro, quando ciò sia possibile. Le tecniche di regressione, ad esempio, esigono che le variabili siano del tipo cardinale (o quasi-cardinale) ed è perciò opportuno che alcune variabili strategiche, come l'appartenenza religiosa in alcuni modelli causali, vengano ridotte ad una serie di variabili dicotomiche (dette variabili *dummy*), in modo da poterle includere nel modello stesso[3]. Se, invece, si usano come strumento di analisi le tabelle di contingenza (oppure i più complessi modelli di analisi loglinare), si può decidere di ridurre a categoriale una variabile cardinale (magari procedendo anche in questo caso ad una drastica dicotomizzazione)[4].

[3] Si può trasformare la modalità "cattolico" in una nuova variabile con la modalità "sì" per coloro che così si sono definiti e la modalità "no" per tutti gli altri.

[4] Si veda *Il problema più importante per noi* (Delli Zotti 1992a), nel quale sono illustrate una serie di opzioni disponibili nel formulare una determinata domanda in un questionario (cioè, diverse definizioni operative per rilevare la stessa proprietà). Il lavoro è pubblicato on-line in un'edizione (Delli Zotti 1992b) nella quale vengono mostrate alcune possibilità di trasformazione delle variabili generate da ognuna delle definizioni operative ed i passaggi tecnici necessari a realizzare tali trasformazioni mediante il programma Spss®.

1. Variabili cardinali *vs.* categoriali e casi noti *vs.* casi anonimi

Una colonna in una matrice dei dati si presenta come una serie di numeri che necessitano, per essere interpretati, di alcune informazioni riguardanti la loro natura perché, altrimenti, essi non rivestono pressoché alcun significato per colui che li esamina. Si è detto "numeri", in quanto solitamente, come già accennato, per ragioni di praticità anche gli stati su proprietà rilevate mediante procedura di classificazione vengono riportati in matrice dopo che ad ogni categoria che compone la classificazione è stato attribuito un codice numerico che rende più veloce la registrazione e la "manipolazione" dei dati da parte dei programmi elettronici di analisi statistica.

Già a questo livello va comunque segnalata la rimarchevole differenza tra le variabili di tipo cardinale e le classificazioni (ordinate o non ordinate). I numeri trascritti in matrice acquistano immediatamente significato nel caso delle variabili cardinali, una volta che venga precisata la proprietà rilevata ed eventualmente indicata l'unità di misura usata nella definizione operativa. A volte non è nemmeno necessario precisare l'unità di misura, in quanto essa è presumibile quasi con certezza: se la proprietà è il numero dei figli, essi saranno riportati in matrice mediante la serie dei numeri naturali e siamo in grado di interpretare il valore 1, oppure 2, come la presenza di un figlio (oppure due). Nel caso i valori presentino cifre decimali, anche se non viene precisata possiamo immaginare che l'unità di analisi sia di tipo aggregato (in genere si tratta di file "ecologici", ovvero su base territoriale), dunque i valori devono essere intesi come numero medio di figli per famiglia o coppia. In altri casi invece è necessario conoscere l'unità di misura: se, ad esempio, la proprietà è il reddito, sapere se è stato registrato in matrice in euro oppure in migliaia di euro, può aiutare a capire se ci troviamo di fronte a "comuni mortali", a milionari (in euro), oppure a fatturati di imprese, o alla ricchezza prodotta da aggregati territoriali (comuni, regioni, ecc.). Nel caso del reddito sarà anche necessario sapere se si tratta di reddito mensile o annuale e, nel caso di aggregati, se si tratta di reddito complessivo o pro capite.

Se invece la proprietà è rilevata mediante classificazione, i codici numerici non sono di norma interpretabili, a meno che non si possieda il c.d. libro codice (*codebook*) e, appreso che che, per registrare il genere dell'intervistato, è stato usato il codice 1 per i maschi ed il codice 2 per le femmine, sarà possibile svolgere una, ancorché molto approssimativa, analisi della colonna dei dati: scorrendola, si potrà infatti giungere alla valutazione che nel campione prevalgono i maschi, oppure le femmine, o che, invece, i due generi grosso modo si equivalgono[5].

L'interesse riguardo al singolo dato presente in una cella, in questo specifico esempio, è ovviamente prossimo allo zero, in quanto sapere che una specifica persona intervistata è un maschio, oppure una femmina, è di per sé è privo di qualsiasi utilità. Ciò accade perché, nell'ipotesi di una rilevazione da un campione di popolazione, i casi sono anonimi; dunque, i dati acquistano significato solo "complessivamente", dopo averne analizzato, ad esempio, la distribuzione di frequenza. Quando i casi (ad esempio, aziende, distretti scolastici, regioni, etc.) sono noti, la matrice dei dati non è solo uno strumento comodo per la registrazione di dati che, per poter essere adeguatamente "apprezzati", necessiteranno di analisi statistica, ma è utile anche per ricavarne singole informazioni riguardanti uno specifico caso.

La tabella 1 riporta (su quattro colonne per ragioni di impaginazione) il contenuto della colonna dei dati che si riferiscono al tasso medio di occupazione (occupati ogni 100 abitanti) nelle diverse province italiane. Proprio perché i casi non sono anonimi, tale informazione è

[5] Tale analisi, a vista, è facilitata da programmi come Spss® dalla possibilità di passare dalla visualizzazione della matrice dei dati codificata alla versione "in chiaro", nella quale i codici numerici sono sostituiti dalle eventuali etichette descrittive.

stata affiancata da un'altra variabile presente nella matrice dalla quale sono stati estratti i dati. Si tratta della variabile identificatrice, riportata "in chiaro" e cioè non usando il codice numerico, che si sarebbe potuto attribuire ad ognuna delle province.

Tab. 1 – Tasso di occupazione: occupati per 100 abitanti nelle province italiane (1999)

Provincia	Tasso	Provincia	Tasso	Provincia	Tasso	Provincia	Tasso
Torino	46,5	Vicenza	54,9	Livorno	40,6	Brindisi	37,2
Vercelli	46,5	Belluno	50,1	Pisa	45,6	Lecce	33,1
Novara	49,4	Treviso	50,6	Arezzo	46,3	Potenza	34,0
Cuneo	48,3	Venezia	46,9	Siena	50,6	Matera	38,9
Asti	46,4	Padova	46,8	Grosseto	42,7	Cosenza	33,4
Alessandria	40,2	Rovigo	47,9	Prato	51,8	Catanzaro	30,5
Biella	47,8	Udine	46,0	Perugia	44,7	Reggio Calabria	31,1
Verbano-Cusio-Ossola	44,6	Gorizia	45,7	Terni	39,7	Crotone	28,2
Aosta	50,0	Trieste	41,5	Viterbo	37,2	Vibo Valentia	31,2
Imperia	42,2	Pordenone	48,5	Rieti	40,6	Trapani	33,2
Savona	41,1	Piacenza	46,0	Roma	43,4	Palermo	28,7
Genova	40,2	Parma	49,1	Latina	41,3	Messina	33,1
La Spezia	39,6	Reggio nell'Emilia	51,2	Frosinone	35,5	Agrigento	29,5
Varese	49,6	Modena	53,7	Caserta	31,4	Caltanissetta	34,4
Como	48,9	Bologna	51,5	Benevento	40,0	Enna	26,9
Sondrio	48,3	Ferrara	46,3	Napoli	31,2	Catania	34,0
Milano	49,1	Ravenna	48,1	Avellino	37,0	Ragusa	38,0
Bergamo	52,3	Forlì	50,2	Salerno	38,6	Siracusa	34,8
Brescia	49,6	Rimini	47,1	L'Aquila	37,8	Sassari	38,8
Pavia	45,5	Pesaro e Urbino	47,6	Teramo	42,9	Nuoro	37,0
Cremona	46,9	Ancona	44,8	Pescara	41,0	Cagliari	36,0
Mantova	49,8	Macerata	45,2	Chieti	39,8	Oristano	35,7
Lecco	50,1	Ascoli Piceno	48,1	Campobasso	37,6		
Lodi	49,2	Massa-Carrara	38,8	Isernia	39,0	Delimitazione di regione	
Bolzano	56,0	Lucca	43,6	Foggia	34,4	Delimitazione di ripartizione	
Trento	49,9	Pistoia	49,4	Bari	36,6		
Verona	47,4	Firenze	44,4	Taranto	33,7		

Scorrendo la tabella è possibile togliersi qualche curiosità, come leggere il tasso di occupazione nella provincia in cui si risiede e, visto che le province sono ordinate secondo l'appartenenza alle diverse regioni[6], è possibile confrontarlo con quello delle province circostanti, oppure effettuare confronti con altre province, anche molto lontane territorialmente dalla nostra. I casi non sono anonimi e perciò ogni singolo dato è associato ad un caso del quale abbiamo anche una conoscenza implicita, che ci permette di effettuare comparazioni "intenzionali" con altri casi presenti in matrice e anche di formulare ipotesi sulle ragioni di un valore particolarmente alto (o basso).

Se invece la colonna di dati si fosse riferita al tipo di occupazione di una campione di intervistati, riportare nella tabella il codice loro attribuito non avrebbe rivestito alcuna utilità perché, anche qualora servisse ad identificare il singolo intervistato, sarebbe utile solo al fine del reperimento del questionario cartaceo, ad esempio, per eliminare eventuali incongruenze nei dati. Identificare un particolare individuo come "questionario 145", oppure come "Mario Rossi", non riveste alcuna differenza, perché il caso rimane pur sempre anonimo dal punto di vista della conoscenza implicita. Non sappiamo di lui nulla di più di quanto non sia riportato in matrice e ciò, a nostro avviso, provoca una serie di conseguenze che si estende dalla "lettura" della matrice dei dati ad altre tecniche statistiche, ed in particolare a quelle di analisi e rappresentazione grafica, delle quali vedremo alcuni esempi nel seguito ed ulteriori approfondimenti in un volume in corso di pubblicazione (Delli Zotti 2005a).

[6] Le regioni, in linea di massima, sono disposte in tabella secondo la scansione territoriale nord-sud (e, per quanto riguarda il nord, tendenzialmente da ovest verso est).

2. Ordinamento dei dati e distribuzione di frequenza

A parte l'uso della matrice dei dati come repertorio, e cioè come fonte diretta di informazione sui singoli casi, l'analisi viene solitamente effettuata sintetizzando i dati con qualche tecnica statistica. Uno strumento molto semplice è costituito dalla tabella della distribuzione di frequenza nella quale i casi, anche fossero in origine noti, diventano anonimi, perché tutti quelli che condividono lo stesso valore sulla variabile analizzata vengono raggruppati all'interno della corrispondente categoria. Ci troviamo nella situazione raffigurata nella tabella 3, dove nella prima colonna, è riportato il valore rilevato e nella seconda la sua frequenza. Apparentemente non c'è molta differenza rispetto alla tabella 1: visto che si tratta di una variabile cardinale rilevata con una certa precisione, non ci sono quasi le "categorie". Pur presente, è infatti poco frequente la situazione in cui due casi condividono lo stesso valore e, dunque, volendo, per molti valori è possibile risalire al rispettivo caso, ritornando ad osservare la matrice dei dati.

Il lettore attento si accorgerà che, rispetto alle informazioni riportare nella matrice, è praticamente intervenuto un solo mutamento davvero sostanziale: i valori, che in matrice sono disposti in modo "disordinato"[7], nella distribuzione di frequenza sono invece ordinati, solitamente dal valore più basso a quello più elevato. Ho detto "solitamente", in quanto alcuni programmi consentono tipi diversi di ordinamento: oltre che secondo valori decrescenti, secondo l'ordine crescente (o decrescente) delle eventuali etichette descrittive assegnate ai valori (in caso di variabili categoriali), oppure secondo valori crescenti (o decrescenti) delle frequenze (riportate nella seconda colonna della tabella 3).

Costruendo la tabella della distribuzione di frequenza, l'operazione fondamentale consiste nell'ordinamento dei valori. Ciò ci suggerisce uno strumento di analisi per certi aspetti ancor più utile della tabella di distribuzione di frequenza, in particolare quando i casi non sono anonimi. Si tratta di uno strumento, tra l'altro, a basso costo, per il quale non è necessario ricorrere a più o meno sofisticati programmi specifici di analisi statistica. La procedura di ordinamento è presente infatti in tutti i programmi di archiviazione delle informazioni (come Excel® e Access®), ma anche in programmi di videoscrittura come Word®. Nulla impedisce, dopo avere costruito una qualsiasi tabella di casi per variabili, che si chieda di effettuare un ordinamento simile a quello che ha prodotto la tabella 2 a partire dai dati della tabella 1. Nel caso la tabella sia stata prodotta mediante un foglio elettronico, si può scegliere quale colonna debba essere usata per effettuare l'ordinamento; se invece la tabella è costruita mediante un programma di videoscrittura, si dovrà avere l'accortezza di trasferire all'inizio dei paragrafi i valori su cui si deve basare l'ordinamento, perché il programma lo realizza sulla base dei caratteri alfanumerici che trova all'inizio dei paragrafi stessi (è in genere possibile scegliere l'ordinamento crescente o quello decrescente).

Nella tabella 2 si è perso l'ordinamento territoriale-amministrativo dei casi, ed è perciò più difficile cercare una singola provincia al fine riscontrare quale sia il suo tasso di occupazione, ma si guadagnano alcune opportunità di analisi dei dati che non vanno trascurate e che rendono questa procedura simile alla realizzazione della tabella di distribuzione di frequenza, ma con il vantaggio che non si perde la possibilità di identificare i singoli casi.

Ad esempio, dalla tabella si vede che il più basso tasso di occupazione è pari a 26,9 e che il più alto è pari a 56,0; ma sappiamo anche che il primo tasso si registra nella provincia di Enna

[7] La loro disposizione dipende dalla sequenza con cui si sono registrati i casi in matrice è, specialmente nel caso di sondaggi su campioni anonimi, l'inserimento dei dati dipende banalmente dalla successione con la quale i questionari arrivano alla scrivania di chi cura l'inserimento di dati, molto spesso anche più persone che lavorano in luoghi diversi. Se i casi sono noti, come nel nostro esempio, la matrice presenta un qualche ordinamento, alfabetico o territoriale, ma i valori che si leggono in qualsiasi altra colonna, che non sia quella denominabile come variabile "identificativa", sono pur sempre "disordinati".

e il secondo invece a Bolzano. È possibile anche vedere che la provincia di residenza di chi scrive (Udine) registra un tasso di occupazione che la colloca nella metà più alta della distribuzione ordinata dei valori. Nella tabella sono state evidenziate anche le altre province della regione Friuli-Venezia Giulia; esse erano collocate l'una accanto all'altra nella tabella 1 e ciò consentiva di leggere facilmente i diversi tassi di occupazione, ma non era altrettanto facile apprezzare le posizioni relative delle province. Nella tabella 2 invece si nota chiaramente che i valori assai simili di Udine e Gorizia collocano queste due province una accanto all'altra, mentre Trieste si trova 13 posizioni più in basso e Pordenone 19 posizioni più in alto.

Tab. 2 – Elenco ordinato delle province secondo valori crescenti del tasso di occupazione

Provincia	Tasso	Provincia	Tasso	Provincia	Tasso	Provincia	Tasso
Enna	26,9	Brindisi	37,2	Firenze	44,4	Sondrio	48,3
Crotone	28,2	Campobasso	37,6	Verbano-Cusio-Ossola	44,6	Pordenone	48,5
Palermo	28,7	L'Aquila	37,8	Perugia	44,7	Como	48,9
Agrigento	29,5	Ragusa	38,0	Ancona	44,8	Milano	49,1
Catanzaro	30,5	Salerno	38,6	Macerata	45,2	Parma	49,1
Reggio Calabria	31,1	Massa-Carrara	38,8	Pavia	45,5	Lodi	49,2
Napoli	31,2	Sassari	38,8	Pisa	45,6	Novara	49,4
Vibo Valentia	31,2	Matera	38,9	Gorizia	45,7	Pistoia	49,4
Caserta	31,4	Isernia	39,0	Udine	46,0	Varese	49,6
Lecce	33,1	La Spezia	39,6	Piacenza	46,0	Brescia	49,6
Messina	33,1	Terni	39,7	Ferrara	46,3	Mantova	49,8
Trapani	33,2	Chieti	39,8	Arezzo	46,3	Trento	49,9
Cosenza	33,4	Benevento	40,0	Asti	46,4	Aosta	50,0
Taranto	33,7	Alessandria	40,2	Torino	46,5	Lecco	50,1
Potenza	34,0	Genova	40,2	Vercelli	46,5	Belluno	50,1
Catania	34,0	Livorno	40,6	Padova	46,8	Forlì	50,2
Foggia	34,4	Rieti	40,6	Cremona	46,9	Treviso	50,6
Caltanissetta	34,4	Pescara	41,0	Venezia	46,9	Siena	50,6
Siracusa	34,8	Savona	41,1	Rimini	47,1	Reggio nell'Emilia	51,2
Frosinone	35,5	Latina	41,3	Verona	47,4	Bologna	51,5
Oristano	35,7	Trieste	41,5	Pesaro e Urbino	47,6	Prato	51,8
Cagliari	36,0	Imperia	42,2	Biella	47,8	Bergamo	52,3
Bari	36,6	Grosseto	42,7	Rovigo	47,9	Modena	53,7
Avellino	37,0	Teramo	42,9	Ravenna	48,1	Vicenza	54,9
Nuoro	37,0	Roma	43,4	Ascoli Piceno	48,1	Bolzano	56,0
Viterbo	37,2	Lucca	43,6	Cuneo	48,3		

Isole
Sud
Friuli-Venezia Giulia

Siccome la tabella è disposta su quattro colonne per occupare meno spazio nella pagina, ci troviamo di fronte ad un'altra possibilità di analisi interessante: quella di individuare alcuni valori posizionali, ben noti a chi ha una conoscenza anche elementare della statistica. Se si divide una distribuzione di frequenza in quattro parti è possibile individuare i quartili, cioè i valori che suddividono la distribuzione in quattro gruppi di casi: un primo gruppo che registra i valori più bassi (nel nostro esempio, le province con tassi di occupazione dal minimo al confine del primo quartile, rappresentato dal valore 37,2), poi un secondo gruppo di casi che raggiunge la metà della distribuzione ordinata dei casi (cioè la mediana, il cui valore è pari a 43,6). Nella parte alta della distribuzione identifichiamo il terzo quartile, che si arresta al valore 48,3, al di sopra del quale abbiamo il quarto quartile, che identifica le province con i più alti tassi di occupazione (fino al 56,0%).

L'individuazione dei quartili è facilitata nella tabella 2 dalla disposizione dei dati su quattro colonne di lunghezza approssimativamente uguale (con la sola eccezione dell'ultima colonna, dovuta al fatto che il numero dei casi è dispari), ma anche se i dati fossero inseriti in un foglio elettronico non è difficile effettuare la stessa operazione. Basterà dividere per quattro il numero dei casi ed il valore ottenuto individua la riga in corrispondenza della quale possiamo leggere il tasso di occupazione che delimita il primo quartile; lo stesso valore, moltiplicato per due e per tre, indica la riga (e cioè il caso) corrispondente alla mediana e al terzo quartile. Se,

ad esempio, i casi sono 275, l'incremento è pari a 68,75 e perciò dovremmo individuare la riga corrispondente al caso n. 69 (arrotondato), al n. 138 (137,5) e al n. 206 (206,25).

Ovviamente, è possibile individuare questi valori soglia anche nella tabella 3, ma ribadiamo che la tabella 2 possiede un valore aggiunto: la semplice disposizione ordinata dei casi, senza il loro accorpamento in categorie, permette l'identificazione e ciò è molto interessante quando i casi non sono anonimi. Scorrendo la lista, infatti, è facile usare la conoscenza implicita delle province italiane, che ci permette di notare che nel quartile inferiore vi sono esclusivamente province del Sud, mentre nel quartile superiore sono elencate solamente province del Nord, in particolare del Nord-est.

Per segnalare l'appartenenza alle aree territoriali del Paese, si può aggiungere ad ogni provincia una sigla, oppure, analogamente a quanto abbiamo fatto per contrassegnare le province del Friuli-Venezia Giulia, si può evidenziare l'appartenenza alle aree con colori diversi (noi ci siamo limitati a farlo per le province del Sud e delle due Isole). Questa codifica mediante i colori ci permette di individuare immediatamente due province delle Isole (Ragusa e Sassari) che si trovano in una posizione nettamente più avanzata rispetto alle altre che appartengono alla stessa ripartizione territoriale. Anche Pescara e Teramo sono in una posizione distante dalle altre province del Sud, ma la nostra conoscenza extra-matrice ci permette di osservare che probabilmente è convenzionale, e non reale, la loro attribuzione al Sud (sono le province del Sud collocate geograficamente più a Nord) e questo giustifica la loro posizione rispetto al dato economico esaminato; al contrario, Frosinone e Viterbo sono due province del Centro che si collocano più in basso rispetto alle altre della stessa ripartizione e, siccome, la loro posizione centrale sul territorio italiano non è in discussione, si tratta probabilmente di province che presentano difficoltà occupazionali maggiori rispetto a quelle delle province circostanti.

Tab. 3 – Distribuzione di frequenza del tasso di occupazione

Valori	N	%	% cum.	Valori	N	%	% cum.	Valori	N	%	% cum.
26,9	1	1,0	1,0	39,6	1	1,0	35,0	46,9	2	1,9	68,0
28,2	1	1,0	1,9	39,7	1	1,0	35,9	47,1	1	1,0	68,9
28,7	1	1,0	2,9	39,8	1	1,0	36,9	47,4	1	1,0	69,9
29,5	1	1,0	3,9	40,0	1	1,0	37,9	47,6	1	1,0	70,9
30,5	1	1,0	4,9	40,2	2	1,9	39,8	47,8	1	1,0	71,8
31,1	1	1,0	5,8	40,6	2	1,9	41,7	47,9	1	1,0	72,8
31,2	2	1,9	7,8	41,0	1	1,0	42,7	48,1	2	1,9	74,8
31,4	1	1,0	8,7	41,1	1	1,0	43,7	**48,3**	2	1,9	**76,7**
33,1	2	1,9	10,7	41,3	1	1,0	44,7	48,5	1	1,0	77,7
33,2	1	1,0	11,7	41,5	1	1,0	45,6	48,9	1	1,0	78,6
33,4	1	1,0	12,6	42,2	1	1,0	46,6	49,1	2	1,9	80,6
33,7	1	1,0	13,6	42,7	1	1,0	47,6	49,2	1	1,0	81,6
34,0	2	1,9	15,5	42,9	1	1,0	48,5	49,4	2	1,9	83,5
34,4	2	1,9	17,5	43,4	1	1,0	49,5	49,6	2	1,9	85,4
34,8	1	1,0	18,4	**43,6**	1	1,0	**50,5**	49,8	1	1,0	86,4
35,5	1	1,0	19,4	44,4	1	1,0	51,5	49,9	1	1,0	87,4
35,7	1	1,0	20,4	44,6	1	1,0	52,4	50,0	1	1,0	88,3
36,0	1	1,0	21,4	44,7	1	1,0	53,4	50,1	2	1,9	90,3
36,6	1	1,0	22,3	44,8	1	1,0	54,4	50,2	1	1,0	91,3
37,0	2	1,9	24,3	45,2	1	1,0	55,3	50,6	2	1,9	93,2
37,2	2	1,9	**26,2**	45,5	1	1,0	56,3	51,2	1	1,0	94,2
37,6	1	1,0	27,2	45,6	1	1,0	57,3	51,5	1	1,0	95,1
37,8	1	1,0	28,2	45,7	1	1,0	58,3	51,8	1	1,0	96,1
38,0	1	1,0	29,1	46,0	2	1,9	60,2	52,3	1	1,0	97,1
38,6	1	1,0	30,1	46,3	2	1,9	62,1	53,7	1	1,0	98,1
38,8	2	1,9	32,0	46,4	1	1,0	63,1	54,9	1	1,0	99,0
38,9	1	1,0	33,0	46,5	2	1,9	65,0	56,0	1	1,0	100,0
39,0	1	1,0	34,0	46,8	1	1,0	66,0	**Totale**	**103**	**100,0**	

La disposizione dei casi per valori crescenti permette di individuare esattamente la loro posizione in graduatoria, ma non altrettanto bene di apprezzare le reali distanze tra di loro sulla dimensione che si sta studiando. Ogni caso occupa uno specifico posto nella lista ma, come si vede dalla distribuzione di frequenza, alcuni casi adiacenti si trovano in realtà in una situazione di pari merito e perciò la loro distanza sul continuo è nulla. Quando non è nulla, la distanza

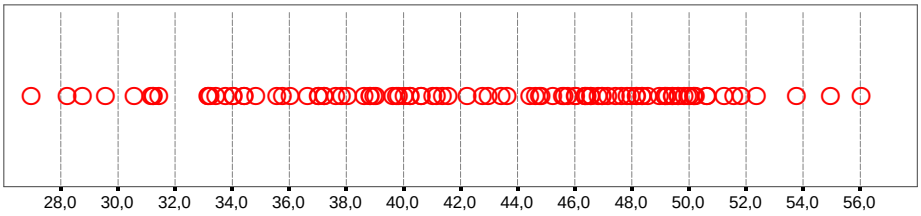
tra casi contigui può essere molto variabile; nel nostro esempio si limita in genere a pochi decimali, ma tra Caserta e Lecce c'è una distanza di 1,7 punti. Altre tecniche, come vedremo, permettono di posizionare esattamente i casi sulla dimensione studiata e consentono perciò di apprezzare meglio le reali distanze[8].

Nella distribuzione di frequenza della tabella 3 sono evidenziati i valori che corrispondono ai quartili e si vede chiaramente come solo alcune (poche) volte vi siano casi con lo stesso valore. Quest'ultima è una prima forma usata per sintetizzare l'informazione contenuta nella colonna dei dati; una forma poco efficace quando la variabile è cardinale e la definizione operativa prevede una misurazione o un conteggio preciso della proprietà. Non va però dimenticato che la sintesi mediante la distribuzione di frequenza può essere alquanto drastica quando, per esempio, la proprietà rilevata è il sesso: in questo caso le decine, centinaia o anche migliaia di casi si riducono a due righe nella tabella della distribuzione di frequenza che riportano la numerosità (assoluta e relativa) delle due categorie di intervistati (maschi e femmine).

3. Strumenti grafici di visualizzazione della distribuzione

Prima di passare a tecniche che permettono una sintesi ancor più drastica dell'informazione contenuta in una variabile analiticamente rilevata come il tasso di occupazione, vediamo alcune altre tecniche che, come il semplice ordinamento, usano i valori presenti in matrice senza sottoporli ad alcuna sintesi. Alcune tecniche grafiche consentono di posizionare i casi sul continuo che varia dal minimo al massimo dei valori empirici rilevati, in modo da visualizzare non solo la gerarchia, ma anche l'esatta distanza tra i casi.

Fig. 1 – Grafica a dispersione del tasso di occupazione

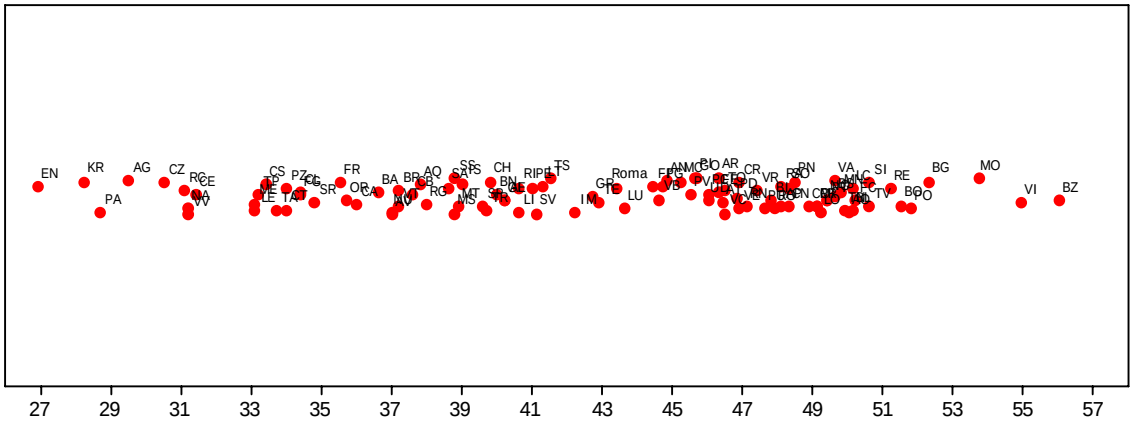


Nella figura 1 si è usata la tecnica del grafico a dispersione (*scatterplot*), solitamente impiegata per mostrare, mediante una "nuvola di punti", l'andamento congiunto di due variabili cardinali ed identificare eventuali "regolarità" come, ad esempio, la crescita tendenziale del reddito all'aumentare dell'età o, più appropriatamente, dell'anzianità di servizio. Se, come nella figura 1, non si specifica la seconda variabile, il grafico mostra la collocazione dei casi su un'unica dimensione e la figura suggerisce in modo immediato la complessiva dispersione dei dati, l'eventuale concentrazione in certe zone del continuo e anche la presenza di casi anomali o estremi. I casi possono essere contrassegnati in modo da poterli identificare e, trattandosi di province, si sarebbe potuta utilizzare la sigla automobilistica, ma abbiamo rinunciato a causa della loro eccessiva numerosità. Come vedremo in seguito, è opportuno farlo quando la maggior parte dei casi vengono riassunti in una figura come il box-plot, al di fuori della quale rimangono solo i casi anomali ed estremi che, per definizione, sono poco numerosi.

[8] Ne abbiamo discusso ampiamente anche in altra sede: si veda il quaderno del Dipartimento di Scienze dell'Uomo dal titolo *Come 'fare a fette' una distribuzione di frequenza* (Delli Zotti 2005b).

Il grafico della figura 1, costretto nelle dimensioni ridotte del foglio, non consente di valutare bene quanti siano i casi presenti nelle zone del continuo in cui i simboli si sovrappongono. Si può ovviare a questo inconveniente introducendo uno scostamento casuale (*jitter*) della posizione dei casi sull'asse delle categorie, che non ha alcun significato sostanziale e consente solamente di creare un'"artificiale" nuvola di punti. L'opzione è disponibile anche in Spss, ma ho realizzato il grafico a dispersione della figura 2 mediante il programma MiniTab® con il quale sono riuscito ad ottenere una figura più leggibile. In genere è opportuno che lo scostamento sia appena sufficiente a consentire di distinguere i punti che altrimenti si sovrapporrebbero, ma in questo caso ho preferito accentuarlo allo scopo di mostrare che lo scostamento casuale consente anche di inserire le sigle automobilistiche che permettono di identificare i casi (per lo meno nelle zone del grafico non troppo affollate)[9].

Fig. 2 – Grafico a dispersione del tasso di occupazione con la posizione dei casi sfalsata e le sigle identificative dei casi



Possiamo usare la stessa tecnica per visualizzare la posizione dei casi appartenenti ad ognuna delle categorie di una variabile categoriale: nella figura 3 ho usato la ripartizione geografica secondo le macro-aree geografiche che, tra l'altro, identificano i collegi elettorali per le elezioni europee. Abbiamo così i casi disposti su cinque linee, che permettono di effettuare un'interessante analisi bivariata dell'andamento della variabile che stiamo studiando. I dati all'interno delle ripartizioni geografiche sono molto meno dispersi (in ognuna c'è un campo di variazione meno ampio) ed occupano porzioni diverse del continuo complessivo. Le province delle Isole sono tutte collocate su valori inferiori al tasso di 40 occupati per 100 abitanti e, al contrario, tutte le province del Nord-est si collocano al di sopra di questo valore soglia.

Nella figura 3 ho rinunciato all'identificazione dei casi mediante la sigla automobilistica e, ciononostante, anche se non possiamo identificarle come nella tabella 2, possiamo notare anche con questo strumento alcune province "devianti". Si vede, ad esempio, che il Nord-est sarebbe ancor più compatto se non fosse per una provincia dal tasso di occupazione notevolmente più basso delle altre (sappiamo dalla tabella 2 che si tratta di Trieste, una provincia che conta una proporzione molto elevata di popolazione anziana). Nel Nord-ovest, in complesso abbastanza ben posizionato, si notano quattro province dai valori sensibilmente più bassi e anche in questo caso è possibile identificarle mediante la tabella 2.

In pratica, si può individuare, area per area, la collocazione sul continuo di tutte le province "fuori posto" rispetto alla collocazione prevalente di ciascuna area e ciò è reso più agevole dalla seguente figura 4, realizzata sempre mediante MiniTab, nella quale, anche se vi è certa-

[9] Siccome lo scostamento viene attribuito sulla base del generatore di numeri casuali, se lo stesso grafico viene tracciato una seconda volta, i casi avranno scostamenti diversi e si potrà riprodurre una seconda volta lo stesso identico grafico solo fissando il valore di base per la generazione dei numeri casuali.

mente una maggiore confusione, si leggono piuttosto distintamente le sigle della province devianti e, del resto, quelle raggruppate al centro delle singole distribuzioni sono meno interessanti e, dunque, è meno importante identificarle. Ovviamente, in questa sede ci premeva mostrare le alternative disponibili, mentre in pratica si tratterà di decidere di volta in volta quale delle due soluzioni sia più idonea.

Fig. 3 – Grafico a dispersione del tasso di occupazione per ripartizione

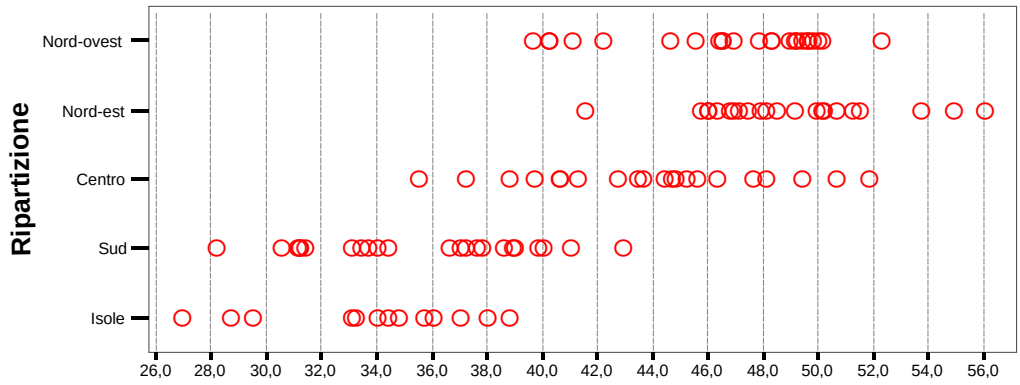
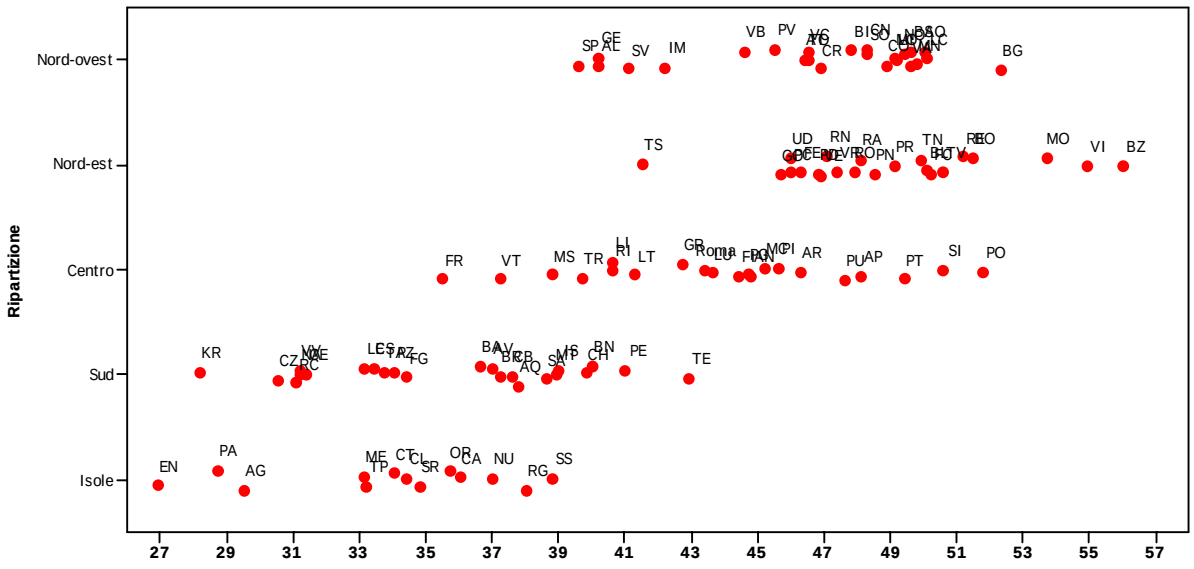


Fig. 4 – Grafico a dispersione del tasso di occupazione con la posizione dei casi sfalsata e le sigle identificative dei casi per ripartizione



Il grafico appena esaminato usa e visualizza tutti i valori della variabile che stiamo esaminando e, da questo punto di vista, appartiene alla categoria delle tecniche di analisi che non riducono la quantità di informazione presente nella variabile originaria. In pratica non è così, perché il grafico non permette di leggere i valori originari della variabile, se non molto approssimativamente. Anche se chiediamo al programma di tracciare nel grafico delle linee guida in corrispondenza dei valori indicati nella scala (come nella figure 2 e 4), otteniamo una lettura grossolana e non certamente precisa al decimale, come nella tabella delle distribuzioni di frequenza. Non ce ne rammarichiamo, naturalmente, perché lo scopo della rappresentazione è quello di mostrare l'andamento complessivo dei tassi nelle ripartizioni geografiche e di

permetterci di individuare eventuali casi devianti o estremi all'interno delle ripartizioni stesse. Si deve perciò ammettere che quella che abbiamo appena visto, in realtà è una tecnica che per certi aspetti sintetizza l'informazione.

Nel 1976 Tuckey ha proposto una tecnica, che si può definire davvero geniale, la quale allo stesso tempo consente una raffigurazione grafica complessiva della distribuzione dei valori e una lettura relativamente precisa dei dati originari. Come vedremo nel seguito, la genialità di Tuckey si è rivelata anche proponendo una seconda tecnica, che ha riscosso ancor maggiore successo (i c.d. box-plot o grafici a scatola), la quale permette di mostrare in una figura riassuntiva la posizione dei valori (che per l'appunto si chiamano posizionali) che abbiamo già incontrato (mediana e quartili, oltre a eventuali casi devianti ed estremi).

Il grafico "ramo-foglia" (*stem & leaf*) viene costruito usando molto ingegnosamente gli stessi valori che costituiscono la distribuzione di frequenza al fine di realizzare una raffigurazione grafica che assomiglia molto ad un istogramma (che vedremo nel seguito), con le barre disposte in orizzontale, anziché in verticale. Come nell'istogramma, vengono costruite automaticamente classi di valori di uguale ampiezza, ma con il vantaggio che i valori possono essere direttamente letti sul grafico.

Fig. 5 – Grafico "ramo-foglia" del tasso di occupazione

```
Frequenza    Ramo &  foglia
    1,00        2 .  6
    3,00        2 .  889
    5,00        3 .  01111
    5,00        3 .  33333
    7,00        3 .  4444455
    8,00        3 .  66777777
    9,00        3 .  888889999
    9,00        4 .  000001111
    5,00        4 .  22233
    8,00        4 .  44445555
   15,00        4 .  666666666677777
   15,00        4 .  888888999999999
    9,00        5 .  000000111
    2,00        5 .  23
    1,00        5 .  4
    1,00        5 .  6

Larghezza ramo:  10,0
Ciascuna foglia: 1 caso(i)
```

Nel nostro esempio (fig. 5) il computer ha costruito un ramo per ogni incremento di due unità e perciò cinque rami per ogni incremento di 10 punti nel tasso di occupazione. Ogni ramo ha una "larghezza" di 10 punti, il che significa che i valori indicati sui rami vanno moltiplicati per 10 al fine di ricostruire i valori originari (se i valori sono ancor più elevati, il ramo può rappresentare centinaia, migliaia o altri valori). Ogni foglia nella figura 5 corrisponde ad un caso ma, se essi sono più numerosi, le foglie ne possono raffigurare una quantità più elevata (indicata in legenda). Siccome la raffigurazione permette di rappresentare solo due cifre, i valori vengono troncati infatti alle prime due cifre significative e, nel nostro esempio, viene sacrificato il decimale (i tassi di occupazione sono troncati all'unità). Il troncamento è una semplificazione un po' "brutale" e sarebbe stato certamente meglio se la riduzione alle prime due cifre significative fosse attuata mediante arrotondamento. Non possiamo escludere che altri programmi effettuino questo tipo di operazione e, ovviamente, se i valori originari fossero composti di due sole cifre, il grafico ramo e foglia avrebbe consentito la lettura diretta e precisa di tutti i valori, senza troncamenti.

Si può così leggere direttamente dal grafico che il valore (troncato) più basso è 26% (il valore 2 si legge sulla colonna "ramo" e il 6 sulla colonna "foglia"); non c'è nessuna provincia che registra un tasso di occupazione del 27%; ve ne sono due con il 28%, una con il 29%, una con il 30%, cinque con il 31%, e così via. All'altro estremo della distribuzione vi sono due

province che registrano valori "isolati", pari al 54% e 56% di occupati per 100 abitanti. Nella prima colonna della figura è riportata la distribuzione di frequenza per ognuno dei rami e la figura complessivamente mostra una certa concentrazione sui valori da 46% a 50% ed una notevole asimmetria, cioè pochi casi che registrano valori più elevati rispetto a questo "picco" ed una coda molto più lunga verso i valori più bassi della distribuzione.

Possiamo riproporre (figura 6) la raffigurazione ramo e foglia per ognuna delle ripartizioni geografiche e da essa notiamo innanzitutto che, quando presenti, i casi estremi vengono indicati separatamente. Ciò serve a segnalare che, per una corretta lettura della figura, si deve tenere conto che essi avrebbero dovuto essere collocati in una posizione più scostata rispetto agli altri casi, in pratica lasciando una o più righe vuote.

Si può anche notare che la lettura della situazione interna alle diverse ripartizioni è relativamente facile, ma non è per nulla agevole confrontare le posizione delle diverse province nel continuo complessivo dei valori. Si è costretti a leggere i valori, infatti, per notare che la forma della distribuzione nel Nord-est non è molto dissimile da quella del Sud, ma i tassi d'occupazione del Nord-est variano dal 41 al 56%, mentre quelli del Sud passano da un minimo del 28% al massimo del 42%.

Fig. 6 – Grafico "ramo-foglia" del tasso di occupazione per ripartizione

```
      Frequenza    Ramo &  foglia
Nord-ovest 1,00       3 .  9
           5,00       4 .  00124
          15,00       4 .  566667888999999
           3,00       5 .  002

Nord-est   1,00       4 .  1
          13,00       4 .  5666667778899
           7,00       5 .  0001134
           1,00       5 .  6

Centro     4,00       3 .  5789
           9,00       4 .  001233444
           6,00       4 .  556789
           2,00       5 .  01

Sud        1,00       2 .  8
          10,00       3 .  0111133344
           9,00       3 .  677778899
           3,00       4 .  012

Isole      2,00  Estremi  (=<29)
           1,00       2 .  9
           5,00       3 .  33444
           5,00       3 .  56788

Larghezza ramo:  10,0
Ciascuna foglia: 1 caso(i)
```

Il diverso posizionamento del campo di variazione delle due ripartizioni sul continuo complessivo era invece molto facilmente visibile nel grafico delle figure 2 e 4. Ciò sia detto a riprova che non c'è una tecnica ideale, ma solo tecniche diverse, che di volta in volta ottimizzano una o l'altra delle esigenze conoscitive o "raffigurative" del ricercatore e perciò il presente lavoro, che può apparire un po' pignolo, forse non è inutile.

Anche se altri strumenti permettono di risolvere il problema (oltre ai grafici a dispersione "adattati", sono davvero preziosi da questo punto di vista i box-plot), possiamo provare a piegare anche i grafici ramo e foglia all'esigenza del confronto tra ripartizioni, auspicando di non essere accusato di lesa maestà nei confronti di Tuckey. Possiamo infatti provare a disporre i rami e foglia affiancati, su un unico continuo, come abbiamo fatto nella figura 5, rinunciando, per facilitare la lettura, alla colonna delle frequenze e tracciando una riga in corrispondenza del passaggio del valore del tasso di occupazione alla decina successiva. Non è una figura particolarmente efficace e, visto che vi sono soluzioni alternative, vi si può rinunciare senza ec-

cessivo rimpianto, però ha il merito di mostrare abbastanza chiaramente quale sia il problema cui si allude. Disposti così, i grafici ramo e foglia fanno per lo meno vedere che le singole distribuzioni di frequenza si collocano in regioni diverse del continuo.

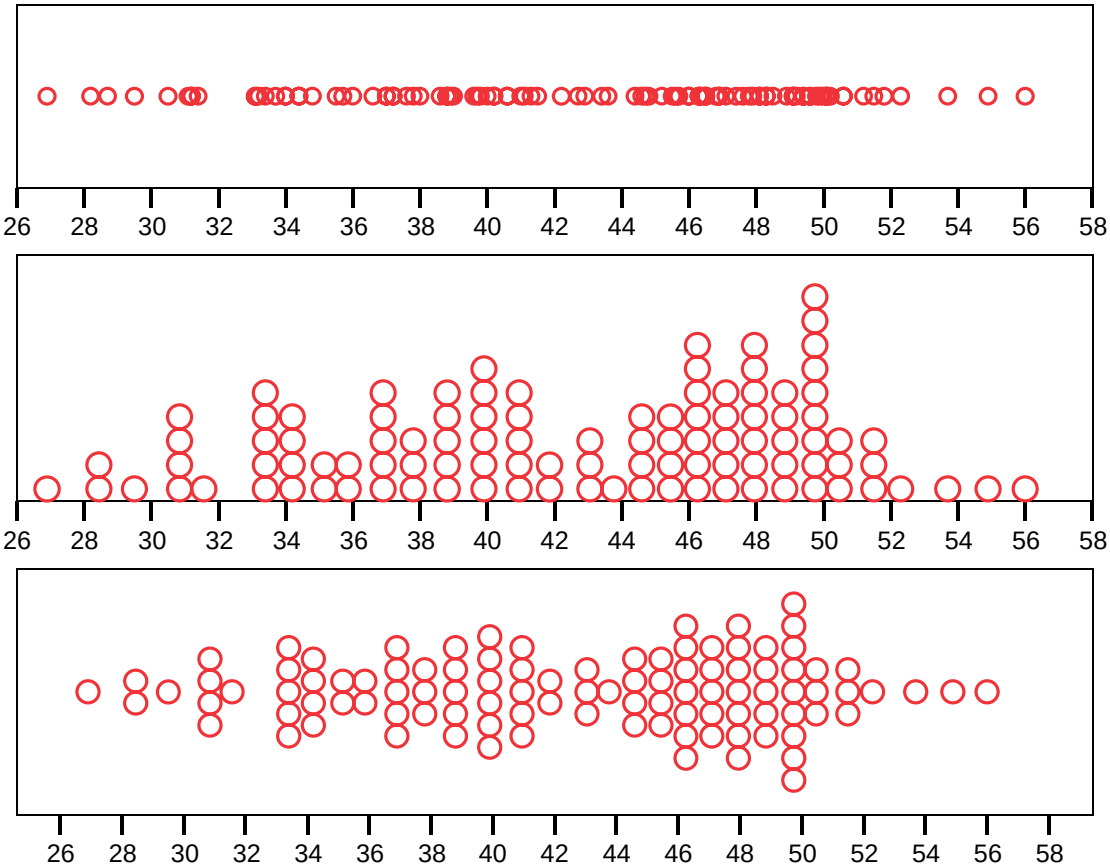
Fig. 7 – Grafico "ramo-foglia" del tasso di occupazione per ripartizione disposto orizzontalmente

Nord-ovest	Nord-est	Centro	Sud	Isole (2 casi <29)
			2. 8	2. 9
			3. 0111133344	3. 33444
3. 9		3. 5789	3. 677778899	3. 56788
4. 00124	4. 1	4. 001233444	4. 012	
4. 566667888999999	4. 5666667778899	4. 556789		
5. 002	5. 0001134	5. 01		
	5. 6			

Larghezza ramo: 10,0
Ciascuna foglia: 1 caso(i)

Un ultimo tipo di grafico permette la rappresentazione della posizione dei singoli casi sul continuo: si tratta del grafico a punti (*dot chart*). La prima versione, definita "piatta" (*flat*) il grafico è praticamente del tutto simile a quello della figura 1. È però possibile risolvere il problema della sovrapposizione di più casi in corrispondenza di determinati valori della variabile in due modi diversi. In entrambi i casi il problema viene risolto posizionando i simboli che rappresentano i casi uno sopra l'altro. Nella versione asimmetrica ciò viene fatto a partire dalla linea di base su cui è tracciata la scala dei valori; nella versione asimmetrica, invece, i pallini sono ripartiti al di sopra di una linea ideale "tracciata" al centro del grafico.

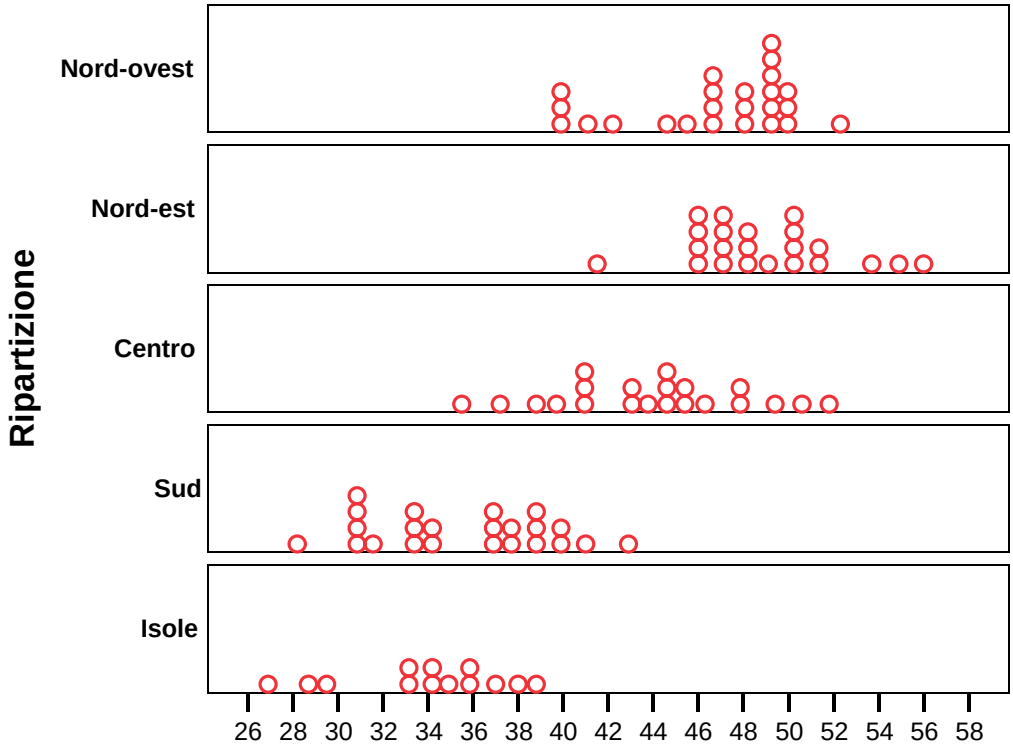
Fig. 8 – Grafico a punti del tasso di occupazione nelle versioni piatta, asimmetrica e simmetrica



Proprio la possibilità di scegliere tra queste varie versioni del grafico impedisce di tracciare la scala verticale che permetterebbe di leggere la numerosità dei casi disposti in ogni colonna di "pallini" che, in effetti, avrebbe senso solo nella versione asimmetrica del grafico.

Nella figura 9, dove il grafico viene proposto solo nella versione asimmetrica, si vede il risultato ottenibile mediante la solita ripartizione della distribuzione della variabile secondo le aree territoriali.

Fig. 9 – Grafico a punti del tasso di occupazione per ripartizione



4. L'istogramma e la curva normale

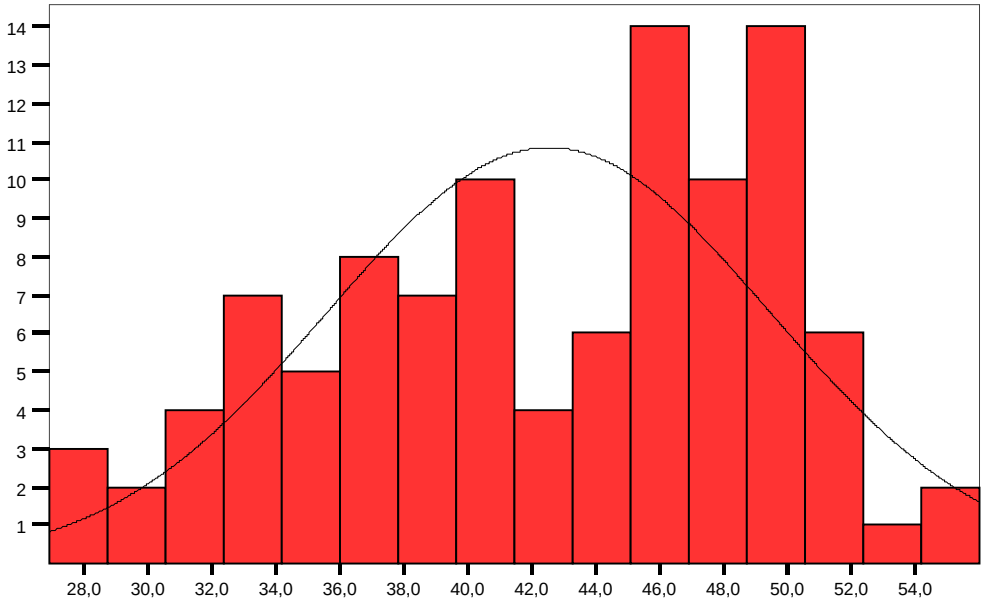
Come tecnica di raffigurazione grafica della distribuzione di frequenza è certamente più conosciuto l'istogramma, il quale, pagando anche in questo caso qualche prezzo, permette di apprezzare diverse caratteristiche della distribuzione dei dati. L'istogramma mostra la distribuzione dei valori di una variabile cardinale dividendo il campo di variazione (differenza tra il valore massimo e il minimo nella distribuzione) in intervalli di uguale ampiezza e tracciando barre la cui altezza rappresenta il conteggio dei casi in ciascun intervallo.

Come si vede dalla figura 10, pur essendo il disegno di gran lunga più accattivante, la forma complessiva è molto simile al grafico ramo e foglia della figura 5 e si può notare, infatti, che anche in esso è ben visibile la concentrazione dei casi intorno ai valori dal 45% al 50% circa del tasso di occupazione. È ben chiara anche l'asimmetria della distribuzione, con la coda verso i valori più elevati che scende repentinamente e molto di meno invece quella verso i valori più bassi. Si nota forse con maggior chiarezza che nella figura 5 che la distribuzione dei dati è in realtà bimodale, con una seconda concentrazione di casi intorno a valori del tasso di occupazione grosso modo del 35-40%. La forma bimodale della distribuzione è suggerita anche dal fatto che la sommità della curva normale non corrisponde alle barre più alte (valori di poco inferiori al 50%), ma molto più a destra (poco sopra al 40%), proprio a causa dell'influenza del secondo picco di valori (c'è una seconda moda sotto al 40%). Abbiamo volutamente

usato i termini “intorno” e “grosso modo”, per sottolineare lo svantaggio insito in questa tecnica, che presenta qualche grado di astrazione che la rende scarsamente utile a scopi diversi da quello di apprezzare la forma complessiva e alcune specificità della distribuzione.

I programmi di elaborazione dei dati creano automaticamente gli intervalli all’interno dei quali viene contata la frequenza dei casi e molto spesso i valori calcolati contengono decimali e si sovrappongono a quelle “cifre tonde” che ci è più facile usare ed apprezzare. Conviene perciò guidare il processo, invece di lasciar fare al computer, e ciò può essere fatto in due modi: ricodificando “manualmente” i valori della variabile originale all’interno di intervalli prefissati ed ancorati a numeri interi, oppure, dopo che il grafico è stato realizzato, agire sui parametri che definiscono il massimo ed il minimo della scala e gli intervalli da costruire, in modo da ottenere lo stesso risultato.

Fig. 10 – Istogramma del tasso di occupazione con curva normale



Ciò è stato fatto con la figura 11, per realizzare la quale si è intervenuti su quella di figura 6, imponendo un valore minimo pari a 26 e poi intervalli di 2 punti percentuali. La forma della figura risultante è ora più simile a quella del grafico ramo e foglia della figura 5 ed è più facile leggere ed apprezzare a cosa si riferiscono le barre che compongono l’istogramma. Si può rilevare facilmente, ad esempio, che vi sono 15 province con un tasso di occupazione tra il 48 e il 49% e solo 5 province con tassi di occupazione tra il 42 e il 43%.

Si nota anche che è sparita la barra della figura 10 che corrispondeva all’incirca ad un tasso di occupazione del 48% ed era un po’ più bassa delle due circostanti: una “depressione” proprio al centro della “zona modale” della distribuzione. Osservando attentamente i dati contenuti nel grafico ramo e foglia ci accorgiamo però che tra i 15 casi contenuti nell’intervallo 46-47% ve ne sono 10 con il valore 46 (arrotondato rispetto ai valori originali) e solo 5 con il valore 47; analogamente tra i 15 casi contenuti nel successivo intervallo 48-49% ve ne sono solo 5 con il valore 48 e 10 con il valore 49. La depressione nella distribuzione di frequenza in corrispondenza di tassi di occupazione del 47-48% è reale e perciò più realistico l’istogramma della figura 10. Con la costruzione manuale degli intervalli abbiamo ottenuto una migliore leggibilità dei dati, ma abbiamo perso un’importante caratteristica della distribuzione e, una volta di più, si può concludere che è bene conoscere i diversi modi di operare e le diverse opzioni disponibili per integrare l’informazione ottenibile con i diversi strumenti.

Fig. 11 – Grafico a barre del tasso di occupazione aggregato in classi

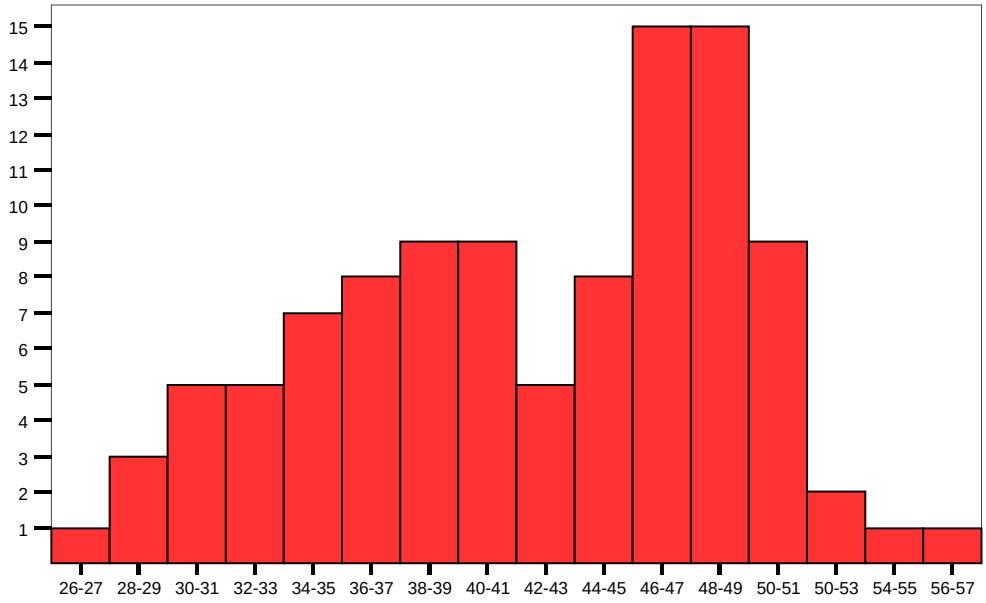
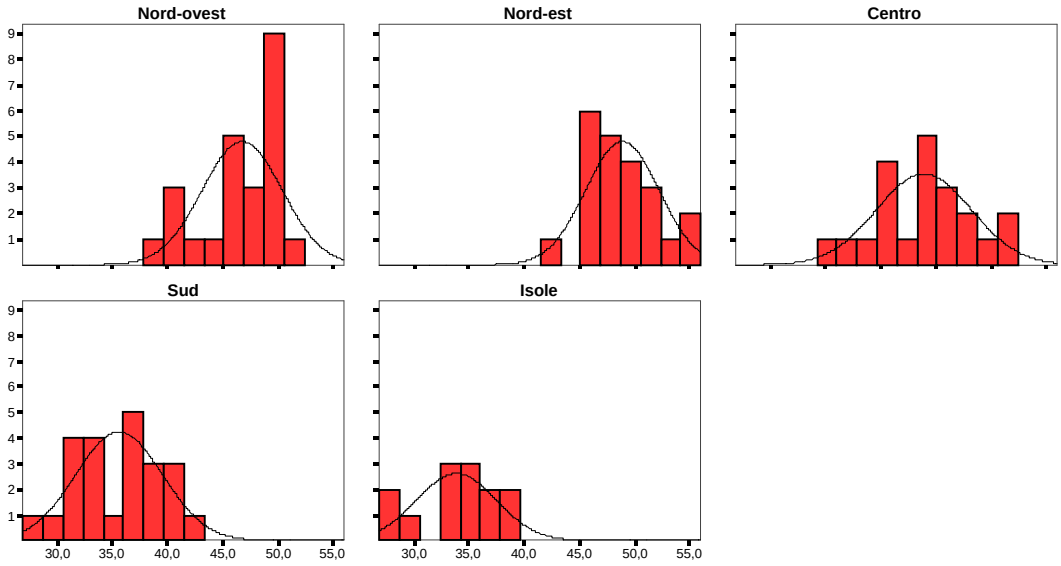


Fig. 12 – *Panel* di istogrammi del tasso di occupazione per ripartizione



Notiamo anche che la figura 11 è un grafico a barre, nel quale è stata annullata la distanza tra le barre stesse in modo da simulare la figura dell'istogramma, nel quale le barre sono automaticamente accostate. Normalmente le barre non sono accostate perché i grafici a barre sono adatti a rappresentare la distribuzione di frequenza di variabili categoriali: l'accostamento delle barre suggerisce una contiguità tra le categorie che è corretta solo nel caso in cui la variabile sia ordinale o cardinale. In ogni caso, rappresentando variabili categoriali non è possibile (nemmeno quando le categorie sono ordinate) tracciare la curva normale.

Per simmetria con quanto effettuato con le altre tecniche fin qui proposte, anche l'istogramma lo riproponiamo per ognuna delle ripartizioni territoriali del paese (fig. 12). A differenza di quanto visto con la serie di grafici ramo e foglia, nella consapevolezza che i grafici multipli (*panel*) vengono costruiti proprio al fine di comparare le diverse situazioni, la scala dei valori per ognuno degli istogrammi che compongono il panel è la stessa e perciò si può constatare abbastanza agevolmente che i valori del tasso di occupazione delle province del

Sud (e delle Isole) si collocano all'interno di un campo di variazione ben diverso da quello che si registra invece nel Nord-est.

5. La riduzione in classi

Potremmo definire "analitiche" le tecniche che abbiamo visto finora, visto che usano "analiticamente" i valori presenti nella matrice dei dati costruendo tabelle o figure che permettono di cogliere l'andamento complessivo delle distribuzione e/o alcune sue specifiche caratteristiche. In alternativa, i valori possono essere "trasformati" e poi essere esaminato il risultato di queste operazioni, che non fanno altro che operare una preliminare sintesi, o semplificazione, dei dati di partenza.

In realtà, come abbiamo visto, una semplificazione preliminare già avviene con la procedura di costruzione dell'istogramma, che prevede la riduzione automatica del campo di variazione in una serie di intervalli uguali. Abbiamo anche visto che per alcuni fini è opportuno che l'individuazione degli intervalli sia guidata consapevolmente, decidendo quale debba essere l'ampiezza delle classi. I programmi offrono anche l'opportunità di decidere il numero delle classi e, di conseguenza, l'intervallo viene calcolato automaticamente. Insomma, dai casi si passa alle categorie, trasformando la variabile cardinale in una variabile ordinale.

Quest'ultima affermazione va meglio precisata perché, quando gli intervalli sono uguali, si può ritenere che la variabile mantenga le caratteristiche della cardinalità, in quanto è cambiata semplicemente l'unità di misura. Il tasso di occupazione che stiamo esaminando era originariamente misurato con la precisione di un decimale e, se gli intervalli non fanno altro che raggruppare i valori arrotondandoli all'unità o alle decine, non cambia la natura della variabile, che rimane cardinale anche se la misurazione è meno precisa. Non cambia la natura della variabile nemmeno se i valori, arrotondati all'unità, sono ulteriormente raggruppati a due a due, come abbiamo fatto per costruire il grafico a barre della figura 11 e ha fatto automaticamente il programma di elaborazione per costruire il grafico ramo e foglia della figura 5. La variabile diventa invece ordinale quando si opera diversamente, scegliendo intervalli di ampiezza diversa (come nel caso della suddivisione in quartili che vedremo di seguito), oppure sezionando la variabile in prossimità di valori soglia o di discontinuità riscontrate nella distribuzione, come illustrato nel citato quaderno dedicato alla suddivisione di una distribuzione di frequenza (Delli Zotti 2005b).

Tab. 4 – Distribuzione di frequenza del tasso di occupazione aggregato in classi di 10 punti percentuali

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	21-30	4	3,9	3,9	3,9
	31-40	35	34,0	34,0	37,9
	41-50	52	50,5	50,5	88,3
	51-60	12	11,7	11,7	100,0
	Totale	103	100,0	100,0	

Per costruire la distribuzione visualizzata nella tabella 4 si è proceduto nel primo modo (si sono costruite le classi per incrementi di 10 unità nella percentuale di occupati) ed è del tutto naturale che le classi abbiano consistenza diversa (nel nostro caso anche perché la prima e l'ultima classe in realtà iniziano e terminano molto dopo e molto prima che sia raggiunto il confine di classe). Questa tecnica di segmentazione permette l'ancoramento delle classi a valori facilmente memorizzabili e, in genere, le "cifre tonde" sono più facilmente apprezzabili.

La riduzione in classi è particolarmente utile se si vuole usare la tabella di contingenza per analizzare il tasso di occupazione in ognuna delle ripartizioni. Infatti, se è difficile apprezzare la distribuzione della variabile originaria usando la tabella di frequenza, si immagini di ripetere l'operazione con una tabella di contingenza, composta nel nostro caso di circa 500 celle. La riduzione in classi ha permesso invece di costruire la tabella 5, nella quale è molto più facilmente leggere il fenomeno, che può essere descritto in vari modi, anche con l'ausilio delle percentuali che sono state ovviamente calcolate per annullare la diversa numerosità delle province nelle 5 ripartizioni.

Tab. 5 – Tabella di contingenza del tasso di occupazione in classi di 10 punti per ripartizione

		Tasso di occupazione: occupati per 100 abitanti (1999)								Totale	
		21-30		31-40		41-50		51-60			
		N	%	N	%	N	%	N	%	N	%
Ripartizione	Nord-ovest			1	4,2	21	87,5	2	8,3	24	100,0
	Nord-est					14	63,6	8	36,4	22	100,0
	Centro			4	19,0	15	71,4	2	9,5	21	100,0
	Sud	1	4,3	20	87,0	2	8,7			23	100,0
	Isole	3	23,1	10	76,9					13	100,0
Totale		4	3,9	35	34,0	52	50,5	12	11,7	103	100,0

Si vede così che le 4 province che non superano il tasso di occupazione del 30% sono collocate tutte nel Meridione (3 nelle Isole e 1 al Sud), mentre le 12 province con tassi superiori al 50% sono tutte collocate al di fuori di quest'area, con una netta prevalenza nel Nord-est. Si può anche leggere la tabella per riga, e sottolineare che le 22 province del Nord-est hanno tutte tassi di occupazione superiori al 40%. Analogamente, avvalendoci delle percentuali, possiamo sottolineare che quasi un quarto delle province delle Isole appartengono al gruppo che registra tassi di occupazione non superiori al 30% e il resto (oltre i tre quarti) comunque non supera il 40% di occupati. Nel Nord-est, invece, oltre un terzo delle province è collocato nella classe con tassi superiori al 50% ed il resto (meno di due terzi), nella fascia contigua con tassi tra il 41% e il 50%.

Con l'aggregazione in classi decennali c'è spesso, come già segnalato, l'inconveniente che alcune possono avere una numerosità non molto elevata o essere addirittura vuote. Al di là del fatto che questa può essere un'informazione importante, la presenza di classi con queste caratteristiche rende la nuova variabile poco usabile nell'analisi mediante tabelle di contingenza. Nel nostro caso non abbiamo dato molta importanza al fatto che fossero solo 4 le province appartenenti alla fascia con tassi non superiori al 30%, perché questo è pur sempre un dato di fatto e le classi che abbiamo costruito ci convincono, in quanto la progressione di 10 punti percentuali è abbastanza ben figurabile nella mente. Ma se i nostri dati riguardassero un campione casuale di individui, sarebbe assai problematico ritenere fondata qualsiasi generalizzazione perentoria basata sul comportamento o sulle affermazioni di sole 4 persone. Immaginando di sostituire il tasso di occupazione con il reddito, ci troveremmo a cercare di trarre conclusioni sull'effetto che la residenza ha sul reddito, basandoci sulla situazione di solo 4 intervistati poveri.

È meglio allora rinunciare alle cifre tonde e suddividere la distribuzione in modo da generare classi di consistenza pressoché uguale: se, ad esempio, si vogliono creare 4 classi, risulta ovviamente molto agevole usare i quartili (tab. 10). Le categorie, come si vede nella tabella 6, essendo state costruite senza badare ai valori (spesso ben lontani dalle cifre tonde che siamo abituati a maneggiare con facilità), possono essere denominate usando espressioni che evochino la posizione della categoria sul continuo, visto che la variabile è diventata una classifi-

cazione a categorie ordinate. La categoria "basso", ad esempio, identifica il quarto delle province italiane che occupa il quartile inferiore della disposizione ordinata dei casi.

Tab. 6 – Distribuzione di frequenza del tasso di occupazione aggregato in fasce

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	Basso	27	26,2	26,2	26,2
	Medio-basso	25	24,3	24,3	50,5
	Medio-alto	27	26,2	26,2	76,7
	Alto	24	23,3	23,3	100,0
	Totale	103	100,0	100,0	

La variabile così costruita può essere usata per creare la tabella di contingenza 7, del tutto simile alla tabella 5, ma nella quale per valutare le percentuali c'è un ausilio in più. Visto che la distribuzione di frequenza relativa al complesso dei casi è stata volutamente suddivisa in quattro porzioni di uguale entità, corrispondente ognuna a circa il 25% dei casi, abbiamo sempre presente, senza bisogno di controllare i marginali, che qualsiasi valore si discosti significativamente da questo indica la sovra o sottorappresentazione di quella specifica categoria nella ripartizione territoriale esaminata.

Per quanto le categorie basate su intervalli di 10 punti percentuali potessero sembrare più convincenti, perché ancorate a cifre tonde, a nostro avviso è invece più realistico e convincente sottolineare dalla tabella 7 che quasi l'85% delle province delle Isole presentano valori "bassi" e che tutte le province del Sud si collocano nella metà inferiore della distribuzione dei tassi occupazionali. Nel Nord-est, invece, la metà delle province presentano valori "alti e quasi tutte le altre valori "medio-alti" nei tassi di occupazione.

Tab. 7 – Tabella di contingenza del tasso di occupazione in fasce per ripartizione

		Tasso di occupazione: occupati per 100 abitanti (1999)								Totale	
		Basso		Medio-basso		Medio-alto		Alto			
		N	%	N	%	N	%	N	%	N	%
Ripartizione	Nord-ovest			5	20,8	9	37,5	10	41,7	24	100,0
	Nord-est			1	4,5	10	45,5	11	50,0	22	100,0
	Centro	2	9,5	8	38,1	8	38,1	3	14,3	21	100,0
	Sud	14	60,9	9	39,1					23	100,0
	Isole	11	84,6	2	15,4					13	100,0
Totale		27	26,2	25	24,3	27	26,2	24	23,3	103	100,0

Non a caso abbiamo commentato la tabella anche con riferimento all'importante discrimine costituto dalla mediana, che divide in due parti uguali la distribuzione: spesso è proprio la dicotomizzazione della variabile lo strumento usato per semplificare l'analisi. Essa costituisce il massimo della sintesi ottenibile accorpando i valori originali (creazione di categorie). Questa sintesi è molto elevata in particolare quando si parte da una variabile cardinale con i valori misurati con una certa precisione (ogni caso, o quasi, presenta un valore unico all'interno della distribuzione).

Il quadro con la variabile dicotomizzata è alquanto semplificato, ma allo stesso tempo indubitabilmente chiaro (tab.8): tutte le province del Sud e delle Isole hanno tassi di occupazione che appartengono alla metà inferiore della distribuzione (valori pari o inferiori alla mediana); le province del Centro sono pressoché equamente ripartite e, invece, le province del Nord

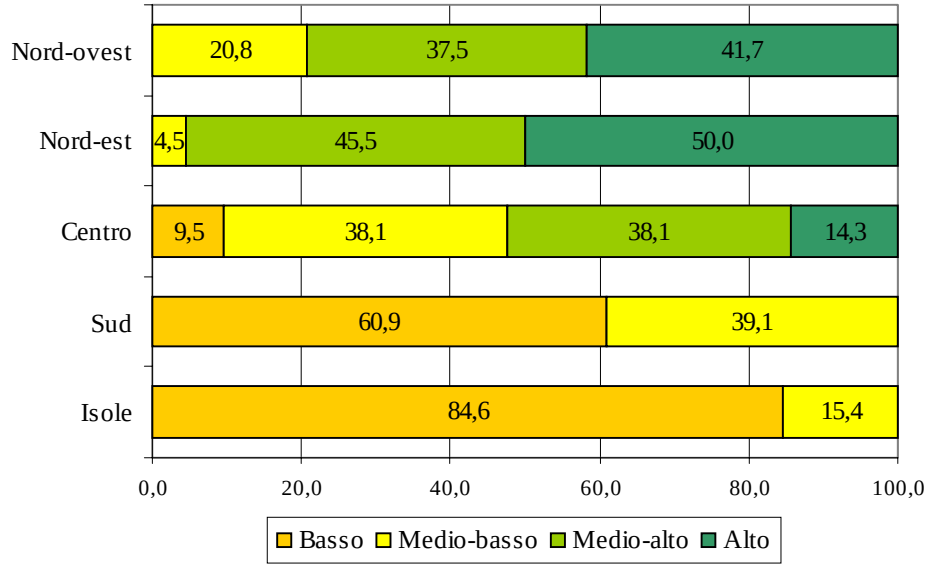
(in particolare del Nord-est, dove si registra la sola eccezione di Trieste) presentano prevalentemente tassi di occupazione superiori al tasso mediano.

Tab. 8 – Tabella di contingenza del tasso di occupazione dicotomizzato per ripartizione

		Tasso di occupazione				Totale	
		Medio-basso		Medio-alto			
		N	%	N	%	N	%
Ripartizione	Nord-ovest	5	20,8	19	79,2	24	100,0
	Nord-est	1	4,5	21	95,5	22	100,0
	Centro	10	47,6	11	52,4	21	100,0
	Sud	23	100,0			23	100,0
	Isole	13	100,0			13	100,0
Totale		52	50,5	51	49,5	103	100,0

Uno strumento senz'altro utile, e intuitivamente corretto, per la corretta rappresentazione grafica di quanto visto in questo paragrafo, è costituito dalle barre percentuali suddivise, che nella figura 13 mostrano, per ognuna delle ripartizioni, come sono ripartite le province nelle categorie di livelli occupazionali. Analizzando con attenzione la figura si riesce a constatare che nel Sud e nella Isole non vi sono province che presentano tassi alti o medio-alti, però, paradossalmente, sembrano "graficamente" più convincenti le tabella 7 e 8. Ciò avveniva perché nelle tabelle erano immediatamente evidenti le celle vuote del Sud e delle Isole in corrispondenza delle categorie "alto" e "medio-alto".

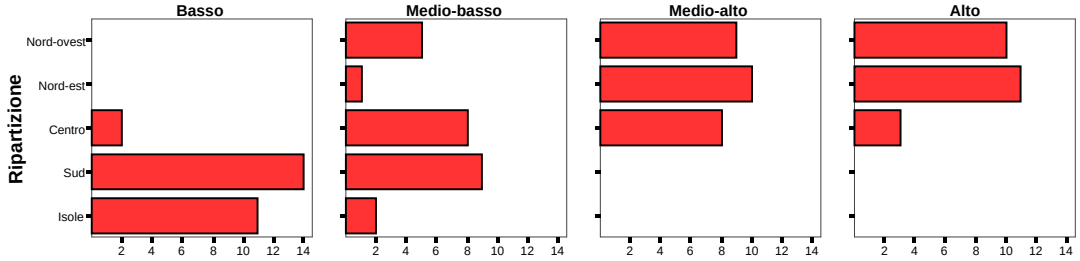
Fig. 13 – Grafico a barre suddivise del tasso di occupazione in fasce per ripartizione



Con il grafico della figura 9, certamente corretto, bisogna invece lavorare di immaginazione e decodificare i colori con l'ausilio della legenda, per accorgersi che alcune categorie sono mancanti in certe ripartizioni. Ogni barra, dovendo essere posta uguale a 100, occupa tutto lo spazio disponibile sulla sua riga e ciò avviene perché su una delle dimensioni del grafico abbiamo una scala di valori (percentuali) comune; invece, nella tabella entrambe le dimensioni sono assi di categorie e la percentuale registrata da ognuna delle categorie in ognuna delle ripartizioni è indicata dal valore numerico registrato nella cella.

Con Spss nella versione 13 è possibile usare un tipo di grafico (nuovo nella galleria di grafici offerti in questo programma, ma ben conosciuto in statistica), che può essere adattato al nostro scopo di rappresentare con maggiore efficacia quanto riportato nella tabella 7. Si tratta della piramide di età che, come abbiamo sostenuto anche nel quaderno dedicato alla segmentazione delle distribuzioni di frequenza (Delli Zotti 2005b), si presta ad un uso molto più generalizzato di quanto non suggerisca il suo nome. Le piramidi di età si realizzano solitamente per rappresentare con barre "a specchio" le classi di età di una popolazione, suddivise nella dicotomia maschi/femmine. Spss consente di rinunciare alla disposizione a specchio e, inoltre, di realizzare le piramidi anche con variabili con più di due categorie (politomiche). La figura 14 è per molti aspetti simile alla tabella 7: i valori assoluti, che nella tabella indicano il numero di province che ricade in ogni categoria congiunta, sono sostituiti nella figura da barre di lunghezza ad essi proporzionale[10].

Fig. 14 – Grafico a piramide con barre non a specchio del tasso di occupazione in fasce per ripartizione



Benché abbastanza soddisfatti dalla capacità descrittiva della figura 10, riteniamo utile proporre un grafico alternativo e riteniamo che anch'esso possa avere un ambito applicativo più generale rispetto all'esempio qui proposto. Il ruolo particolare della mediana, al centro della distribuzione, può essere enfatizzato usandola come una spartiacque da cui far partire l'asse della rappresentazione grafica. In pratica, si può cercare di facilitare il confronto tra il numero di casi che stanno al di sopra e al di sotto della mediana, visualizzando la dicotomizzazione presentata nella tabella 8. Nel nostro caso abbiamo conservato nella rappresentazione grafica le 4 categorie originarie, ma abbiamo suggerito la sostanziale dicotomizzazione (valori sopra e sotto la mediana) usando due gradazioni dello stesso colore per ognuno dei due versanti della "piramide".

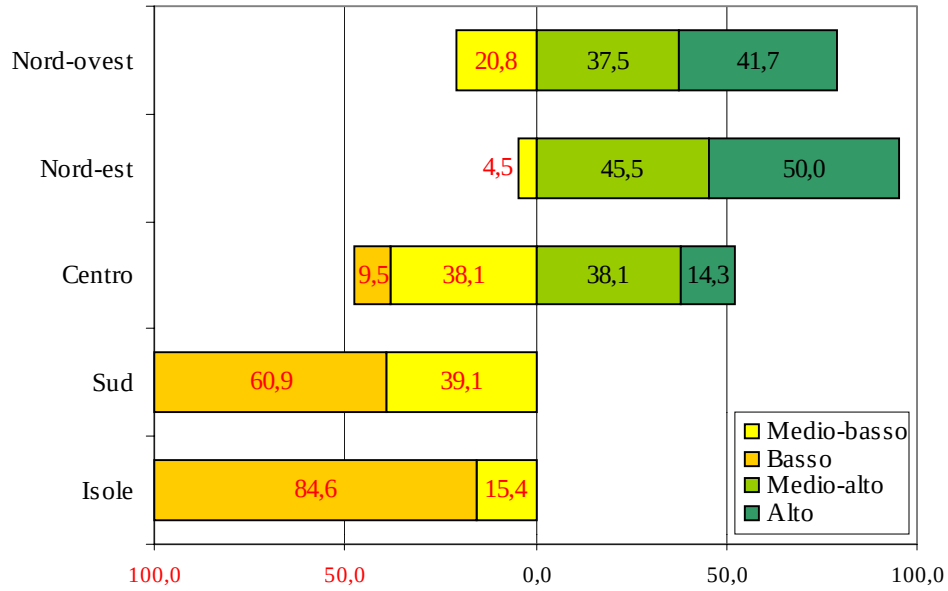
Per realizzare il grafico con Excel è necessario ricorrere a qualche accorgimento: i numeri che saranno visualizzati a sinistra nel grafico devono essere scritti con il segno negativo e le due categorie devono essere scritte nel foglio elettronico in ordine inverso, rispetto a quello in cui appaiono nel grafico (operando in questo modo vanno al loro posto le barre, ma rimane comunque invertita la legenda per i valori a sinistra della mediana). Per eliminare l'antiestetico e controintuitivo segno meno nella scala e nei valori visualizzati sulle barre, è necessario scegliere il formato che consente di rappresentare con il carattere rosso i numeri negativi.

Questo tipo di grafico a barre "fluttuanti" (fig. 15) si presta bene, a mio avviso, per rappresentare le percentuali di risposta a scale di Likert, suddividendo, ad esempio, le risposte positive (o che segnalano accordo, oppure fiducia nelle istituzioni, ecc.) da quelle che indicano un atteggiamento contrario. Lo stesso si può fare con le scale autoancoranti, rappresentando a sinistra e destra rispettivamente la metà inferiore e superiore dei valori della scala. Se il numero dei valori (punteggi) è dispari, rimane da risolvere il problema dell'aggregazione della catego-

[10] Il programma consente di rappresentare solo i conteggi della variabile, ma riteniamo che nella prossime versioni (analogamente a quanto accade per altri tipi di grafici offerti dal programma) sarà possibile scegliere di rappresentare anche le frequenze percentuali. La figura, in ogni modo, è stata ottenuta con una procedura leggermente diversa, a partire dai grafici a barre interattivi che consentono l'utilizzo dei valori percentuali.

ria intermedia, che può essere aggregata ad uno dei due versanti, visualizzandola però con un colore leggermente diverso, a significare che l'aggregazione è convenzionale, visto che essa, a rigore, non appartiene a nessuno dei due insiemi.

Fig. 15 – Grafico a "piramide" con barre sovrapposte del tasso di occupazione in fasce per ripartizione



6. Valori posizionali e parametri sintetici

Con la dicotomizzazione della variabile siamo arrivati al massimo della sintesi per accorpamento dei valori, ma vi sono altri modi per riassumere, in modo più o meno drastico, l'informazione contenuta in una distribuzione di frequenza. Si tratta dei valori posizionali e dei parametri sintetici, che a loro volta possono essere suddivisi tra quelli che servono ad indicare la tendenza centrale ovvero la dispersione dei valori.

Queste caratteristiche, se disposte in una tabella a doppia entrata (tab.9), individuano quattro tipi di strumenti statistici usabili nell'analisi monovariata di variabili cardinali:

1. *Valori posizionali di tendenza centrale*: la **mediana**, che abbiamo già incontrato, è il valore sopra e sotto il quale ricade la metà dei casi. Quando il numero delle osservazioni è pari essa corrisponde alla media delle due osservazioni centrali.
2. *Valori posizionali di dispersione*: abbiamo già incontrato anche il massimo ed il minimo della distribuzione che, sottratti l'uno dall'altro, forniscono il **campo di variazione** (*range*, o intervallo della variabile). Si può dire che questi valori posizionali mostrano quale porzione del continuo teorico individuato dalla definizione operativa della variabile sia effettivamente occupata dalla distribuzione empirica dei valori.
3. *Parametri statistici di tendenza centrale*: si tratta evidentemente della **media**, assai conosciuta e utile a volte in alternativa alla mediana. Quest'affermazione può apparire paradossale e un po' provocatoria, perché generalmente è la mediana ad essere considerata un utile sostituto della media, in determinate circostanze. Noi invece non vediamo ragioni per un uso più frequente della mediana: tra le sue virtù c'è il fatto che può essere usata anche con variabili ordinali e, inoltre, è quasi indispensabile quando nella distribuzione vi sono valori devianti o estremi. La media, infatti, è calcolata sommando tutti i valori (eventualmente moltiplicati per la loro frequenza) ed è perciò influenzata da eventuali valori anomali, men-

tre la mediana, essendo semplicemente il valore che si colloca al centro della distribuzione, non è influenzata da "ciò che accade" agli estremi[11].

4. *Parametri statistici di dispersione*: il più noto è la **deviazione standard** (detta anche scarto-tipo), calcolata a partire della media e definibile come la media degli scarti dei valori dalla media aritmetica della distribuzione. La deviazione standard, espressa nella stessa unità di misura della media, se aggiunta e sottratta alla media stessa, individua il campo di variazione all'interno del quale si concentrano i due terzi delle osservazioni[12].

Tab. 9: Strumenti statistici sintetici nell'analisi monovariata

	Tendenza centrale	**Dispersione**
Valori posizionali	Mediana	Campo di variazione
Parametri statistici	Media	Deviazione standard

I parametri statistici or ora ricordati sono elencati nella tabella 10 che commentiamo rapidamente. I casi sono 103 e per tutti è stata rilevata l'informazione (mancati = 0). Il tasso di occupazione medio (arrotondato al primo decimale) è pari a 42,5 occupati ogni 100 abitanti e quello mediano è molti simile (43,6), in quanto non vi sono casi veramente anomali nella distribuzione. Comunque, se si esaminano con attenzione la tabella 3 e la figura 1, si nota che sui valori più bassi c'è un gruppo di 9 province leggermente distaccate dal "corpo centrale" della distribuzione, mentre tra i valori alti le province con valori leggermente distanti sono solo 3: ciò fa sì che la media si abbassi di circa un punto rispetto alla mediana.

Tab. 10 – Parametri statistici di tendenza centrale e dispersione del tasso di occupazione

Tasso di occupazione: occupati per 100 abitanti (1999)		
N	Validi	103
	Mancanti	0
Media		42,51
Mediana		43,60
Deviazione std.		6,89
Intervallo		29,1
Minimo		26,9
Massimo		56,0
Percentili	25	37,20
	50	43,60
	75	48,30

La media è "accompagnata" dalla deviazione standard (pari a circa 7 punti), la quale indica (aggiungendo e sottraendo 6,9 punti alla media) che i due terzi delle province presentano tassi di occupazione che variano da circa 35,5 (42,51 - 6,89) a circa 49,5% (42,51 + 6,89). Anche la mediana si accompagna a valori (il massimo ed il minimo della distribuzione) che indicano

[11] Si può neutralizzare l'effetto dei valori anomali anche calcolando la media *trim* (troncata): Spss offerte questa opzione, con la quale la media viene calcolata dopo avere eliminato il 5% delle osservazioni collocate ai lati opposti della distribuzione.

[12] Tra gli altri parametri statistici c'è l'asimmetria, meno nota e sostituibile dall'esame dell'istogramma e della curva normale: una curva normale è simmetrica e presenta perciò un valore di asimmetria pari a zero; un valore negativo indica una coda della distribuzione più accentuata a sinistra; un valore positivo che la coda è più lunga a destra. Inoltre, c'è la curtosi: se è positiva, la distribuzione ha più concentrazione nelle code rispetto ad una distribuzione normale; viceversa, se la curtosi è negativa c'è più concentrazione al centro della distribuzione.

quanto i dati sono dispersi intorno ad essa: nel nostro caso il valore minimo è di poco inferiore ad un tasso del 27% e quello massimo è esattamente pari al 56%, con un campo di variazione (intervallo) pari a quasi 30 punti.

Come abbiamo visto in precedenza, è utile individuare altri due valori posizionali che, collocandosi a metà strada tra il minimo e la mediana, e tra la mediana ed il massimo, permettono di frazionare la distribuzione in quattro gruppi contenenti un pari numero di casi. L'efficacia dei parametri sintetici e dei valori posizionali appare assai evidente quando vengono usati per confrontare la distribuzione di frequenza di variabili diverse, oppure, come è il nostro caso, le distribuzioni condizionate: cioè distribuzioni relative a gruppi di casi individuati dai valori di una variabile categoriale. È ovviamente difficile comparare le distribuzioni esaminando analiticamente i dati originali, e finora abbiamo visto come invece si può fare confrontando la forma degli istogrammi (fig. 12), oppure riducendo le variabili in classi e usando le frequenze percentuali (tabelle 7 e 8).

Un'altra strada praticabile è proprio quella di impiegare i parametri statistici e i valori posizionali: esaminando la tabella 11, essi ci permettono di constatare assai rapidamente che il tasso di occupazione medio nel Nord-ovest è di circa il 47%, sale ulteriormente nel Nord-est (a quasi il 49%), mentre è più basso nel Centro (44%) e scende "drammaticamente" nel Sud (36%) e in particolare nelle Isole (34%). Come abbiamo visto nella figura 2, la media del Nord-ovest è un po' penalizzata, rispetto a quella del Nord-est, dalle province della Liguria, che all'interno della ripartizione costituiscono casi un po' devianti: la mediane invece (meno influenzate da essi) sono invece molto simili nelle due ripartizioni.

Tab. 11 – Parametri statistici del tasso di occupazione per ripartizione

Tasso di occupazione: occupati per 100 abitanti (1999)

		Ripartizione				
		Nord-ovest	Nord-est	Centro	Sud	Isole
N	Validi	24	22	21	23	13
	Mancanti	0	0	0	0	0
Media		46,75	48,88	43,90	35,59	33,85
Mediana		48,05	48,30	44,40	36,60	34,40
Deviazione std.		3,63	3,33	4,30	3,94	3,60
Intervallo		12,7	14,5	16,3	14,7	11,9
Minimo		39,6	41,5	35,5	28,2	26,9
Massimo		52,3	56,0	51,8	42,9	38,8
Percentili	25	44,83	46,68	40,60	31,40	31,30
	50	48,05	48,30	44,40	36,60	34,40
	75	49,55	50,75	46,95	38,90	36,50

Il Centro (ma anche il Sud) presenta valori più dispersi intorno alla media di quanto non si riscontri nelle altre ripartizioni, ed in particolare nel Nord-est che, a parte il caso di Trieste, è assai compatto nella sua situazione di relativo "benessere" occupazionale. Il massimo, il minimo e l'intervallo complessivo forniscono ulteriori spunti di commento delle diverse situazioni. Per fare un solo esempio, ci permettono di notare che, pur mostrando valori medi assai più bassi, le province dove c'è maggiore occupazione al Sud e nelle Isole presentano valori che quasi raggiungono il valore medio registrato al Centro e sono molto più elevati di quelli delle province del Centro con tassi di occupazione più bassi. Questo è un ulteriore segno che, pur potendosi evidenziare sostanziali differenze complessive tra le ripartizioni, esse anche sono piuttosto diversificate al loro interno.

Nonostante la consistente variabilità interna, l'appartenenza di una provincia ad una determinata ripartizione è un buon predittore del tasso di occupazione: infatti, le due variabili sono significativamente associate, come si evince dal test di significatività evidenziato nella tabella 12. A rigore, il test di significatività non ha un particolare significato nel nostro caso in quanto non c'è un problema di inferenza: la rilevazione non è campionaria e dunque non si pone il problema di stimare quale sia il margine di errore che si rischia di commettere nell'inferire che le differenze delle medie campionarie siano presenti anche nella popolazione di riferimento. Lo abbiamo qui usato a fini didattici: un valore della significatività così alto mostra pur sempre che le differenze tra le medie sono tali che, se si trattasse di un campione, non sarebbero certo dovute al caso. Più appropriatamente, possiamo usare la misura di associazione "eta", pari a .84, la quale, elevata al quadrato, indica in oltre il 70% la percentuale di varianza nella distribuzione attribuibile all'appartenenza alle ripartizioni, mentre il resto (meno del 30%) è attribuibile alle specificità provinciali.

Tab. 12 – Analisi della varianza del tasso di occupazione per ripartizione

Tabella ANOVA

			Somma dei quadrati	df	Media dei quadrati	F	Sig.
Tasso di occupazione: occupati per 100 abitanti (1999) * Ripartizione	Fra gruppi	(Combinati)	3441,134	4	860,283	60,075	,000
	Entro gruppi		1403,383	98	14,320		
	Totale		4844,517	102			

Misure di associazione

	Eta	Eta quadrato
Tasso di occupazione: occupati per 100 abitanti (1999) * Ripartizione	,843	,710

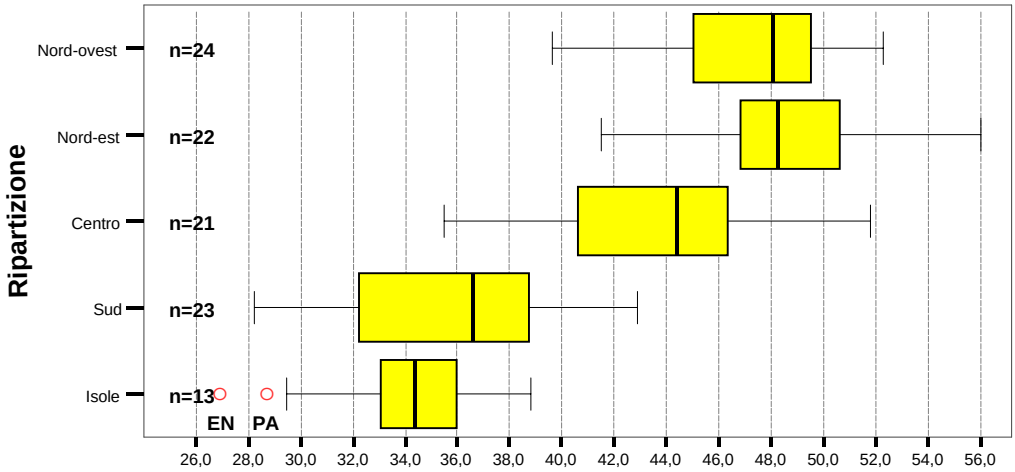
7. La rappresentazione grafica dei valori posizionali e dei parametri sintetici

I test statistici che abbiamo visto sono interpretabili abbastanza facilmente con un po' di esperienza ma, quando diventano numerose le distribuzioni da confrontare, può diventare difficile figurarsi la situazione. Il già ricordato Tuckey (1977) ha proposto una figura (il box-plot o grafico a scatole) utile a mostrare in modo compatto la posizione dei dati attorno alla mediana, con l'ulteriore possibilità di marcare singolarmente quei valori che, collocandosi ad una distanza dalla mediana superiore a determinate soglie, vengono definiti devianti o estremi.

Nella figura 16 sono state tracciate linee di riferimento che consentono una lettura, ancorché approssimativa, dei valori usati per costruire le scatole (box) e i "baffi" (*whiskers*), che rappresentano l'estensione dei quattro quartili in cui è divisa la distribuzione (i valori sono leggibili con precisione nella tabella 11). Nella figura si nota, ad esempio, che il tasso di occupazione mediano nel Nord-ovest è pari al 48% e che le province collocate a ridosso di questa posizione centrale (secondo e terzo quartile) presentano valori tra 45% e meno del 50%.

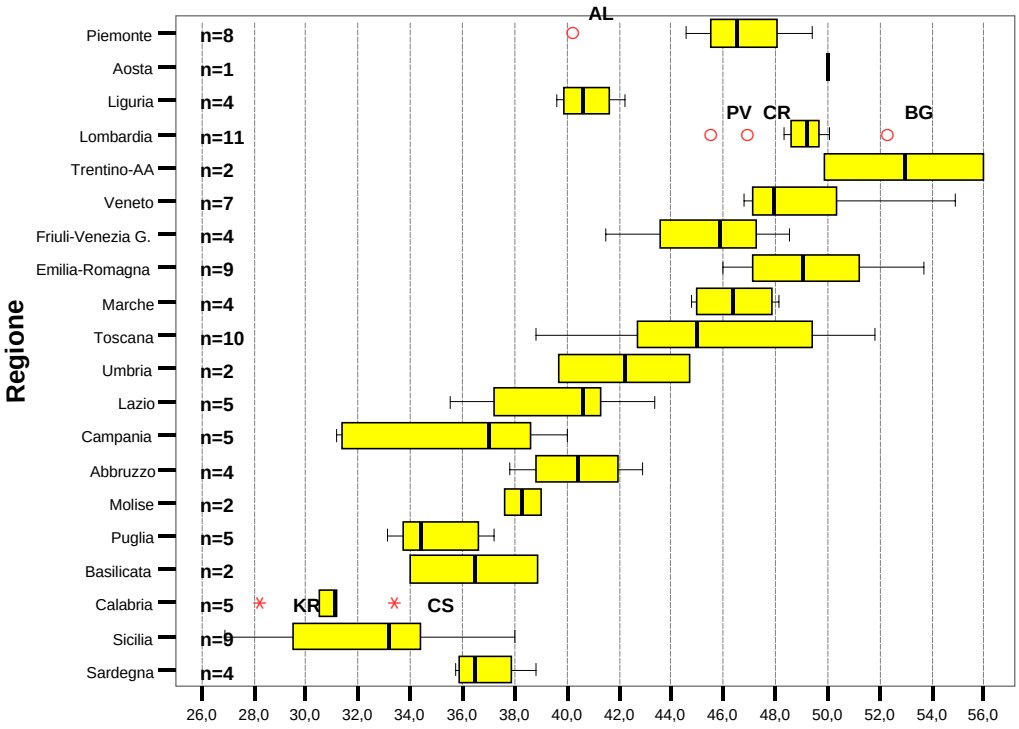
È immediatamente evidente che la distribuzione è asimmetrica, in quanto i 12 casi a sinistra della mediana sono dispersi all'interno di un (sotto)campo di variazione molto più ampio (da meno di 40 a 48%) di quello nel quale sono collocate le 12 province che hanno tassi di occupazione più elevati (da 48 a poco più di 52%). Nelle Isole si registrano due casi devianti (le province di Enna e Palermo) e anche con questo strumento si vede con chiarezza che Sud e Isole occupano uno spazio che poco si sovrappone con quello delle altre ripartizioni.

Fig. 16 – Box-plot del tasso di occupazione per ripartizione



La visualizzazione è in parte simile a quella della figura 2, dove veniva proiettata la posizione precisa di ogni caso, ma con i box plot c'è la possibilità di identificare i valori anomali (compresi tra 1,5 e 3 lunghezze di scatola a partire dalla sua estremità) ed estremi (collocati a più di 3 lunghezze) e di individuare facilmente i principali punti di suddivisione della variabile. La sola controindicazione rispetto a queste figure, per molti aspetti "geniali" (come il grafico ramo e foglia), riguarda il fatto che un occhio non esercitato può non essere in grado di valutare correttamente il significato della maggiore o minore estensione dei box e dei baffi. Si può essere indotti erroneamente a ritenere che la maggiore estensione di un box (o di un baffo) rappresenti una maggiore numerosità di casi all'interno, quando invece è esattamente il contrario: la maggiore estensione significa maggiore dispersione perché, per definizione, in ognuna delle quattro parti della distribuzione è compreso lo stesso numero di casi.

Fig. 17 – Box-plot del tasso di occupazione per regione

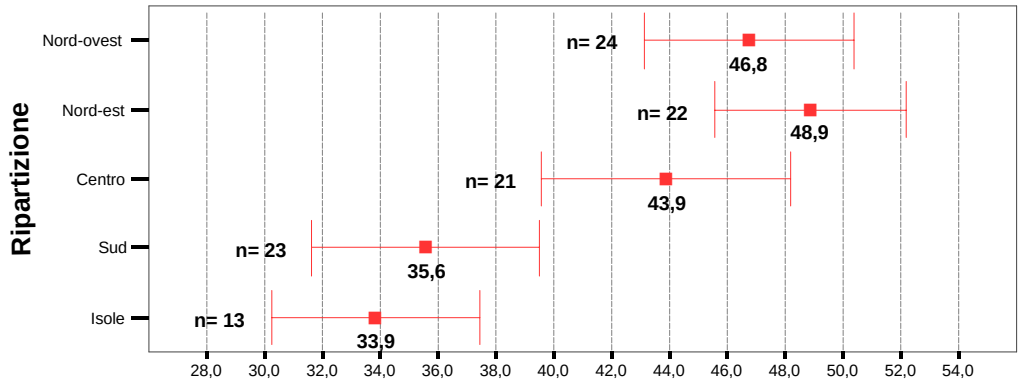


I box plot sono tanto più utili quanto più numerose le categorie da porre a confronto e a tal fine abbiamo realizzato la figura 17 dove sono rappresentate tutte le 20 regioni. Ovviamente, si tenga conto che alcune ripartizioni non hanno un numero di province sufficiente a tracciare correttamente i box e/o i baffi: la Valle d'Aosta comprende una sola provincia e, ovviamente non si applica nemmeno il concetto di dispersione dei dati; nelle regioni che comprendono due province la mediana è pari alla media aritmetica dei valori registrati nelle due province; nelle regioni con un numero pari di casi (province) la mediana è pari alla media dei due valori centrali della distribuzione.

La scarsa numerosità di casi in alcune regioni non riveste alcuna importanza per quanto riguarda il problema della significatività statistica dei dati che, come abbiamo ricordato, non trattandosi di un'indagine campionaria non si pone nemmeno. Invece, la possibilità di rappresentare molte distribuzioni in una spazio relativamente ridotto ci consente di notare nuovi elementi molto interessanti, uno dei quali, ad esempio, è il fatto che la ripartizione Isole è in realtà un'"astrazione". Dalla figura 17 si vede infatti in modo estremamente chiaro che le distribuzione dei tassi di occupazione delle province della Sicilia e della Sardegna si collocano in due sezioni piuttosto distanti del continuo.

Come per i valori posizionali, anche per la media e la deviazione standard si è trovata una tecnica grafica di rappresentazione, denominata "barre degli errori" (*error graphs*) che consente facili confronti ed è abbastanza evidente la genesi della sua denominazione. La variabilità di comportamenti, opinioni, e caratteristiche socio-economiche sono la ricchezza della società (il mondo è bello perché è vario), mentre in molte altre scienze la variabilità che si vuole studiare è spesso dovuta ad "errori", dovuti all'imperfezione degli strumenti o della procedura di misurazione, oppure ai materiali o alle tecniche di produzione di qualche prodotto.

Fig. 18 – Grafico degli "errori" del tasso di occupazione per ripartizione



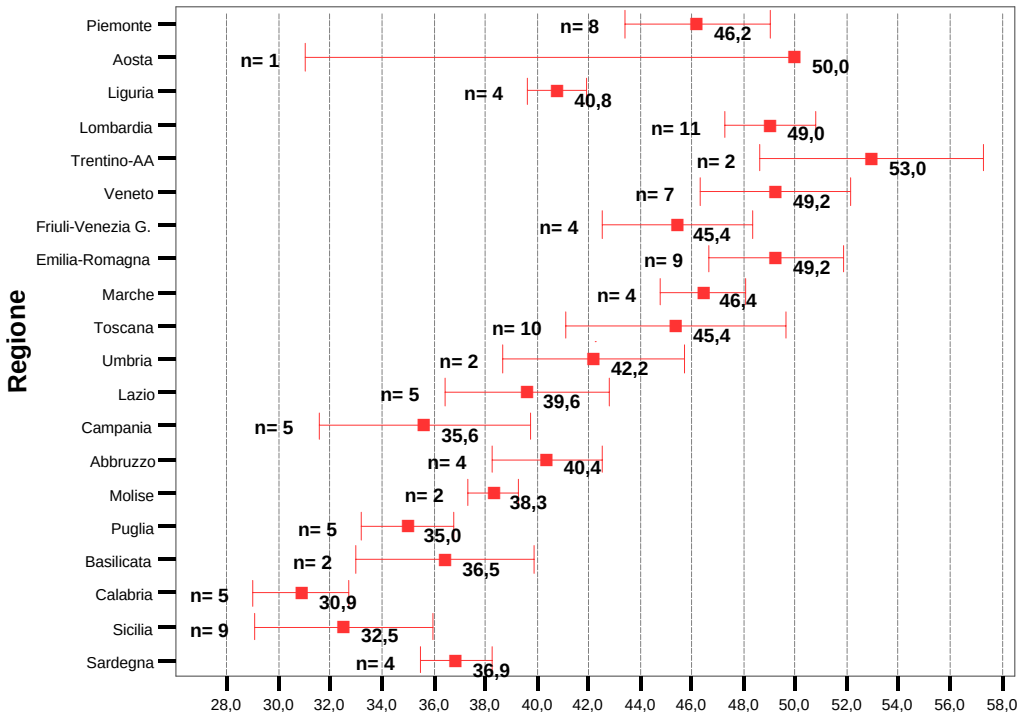
La figura 18 mostra diverse analogie con la figura 16, ma anche alcune differenze, a partire dal fatto che l'informazione incorporata in quest'ultima è molto più ricca. Si nota più di tutto che manca completamente nella figura 18 la possibilità di apprezzare eventuali asimmetrie nella distribuzione dei valori. Ciò non è costituisce un problema per la rappresentazione dei veri e propri errori, in quanto si può assumere che essi siano dovuti alla mera casualità e perciò presumibilmente simmetrici (si può sbagliare per eccesso o per difetto). Molto fenomeni sociali invece sono asimmetrici: si pensi al reddito, un tipico fenomeno la cui distribuzione è caratterizzata da una notevole asimmetria. I poveri sono poveri grosso modo alla stessa maniera, perché sotto ad una determinata soglia non si può scendere, in quanto è messa in pericolo la sopravvivenza e/o intervengono correttivi come le politiche di welfare, che fissano soglie quali la pensione minima; all'opposto, non c'è quasi limite alla ricchezza che un numero esiguo di persone/famiglie può accumulare.

Il fatto che i grafici degli errori siano assai semplici consente di tracciarne un numero assai elevato e questo può essere un motivo per preferirli in alcune circostanze ai grafici a scatola. Si potrebbe rilevare che i valori visualizzati nella figura 19 sono facilmente interpretabili anche leggendoli nella tabella 13, ma è decisamente più facile valutare nella figura la diversità tra le medie da una regione all'altra che, nella metafora spaziale, si traduce in distanza. La dispersione dei dati intorno alle medie è invece tutto sommato abbastanza simile e dipende anche dal fatto che, a parità di altre condizioni, il numero delle province fa di per sé aumentare la variabilità nelle regioni più "affollate". Correttamente non viene calcolata la deviazione standard nel caso della Valle d'Aosta (c'è una sola provincia e dunque non ha senso parlare di deviazione da una media che è convenzionale), ma nel grafico essa viene tracciata (si tratta probabilmente di un errore del programma), anche se su un solo lato della "media".

Tab. 13 – Media e deviazione standard del tasso di occupazione per regione

Regione	Media	N	Deviazione std.
Piemonte	46,21	8	2,83
Aosta	50,00	1	.
Liguria	40,78	4	1,13
Lombardia	49,03	11	1,75
Trentino-AA	52,95	2	4,31
Veneto	49,23	7	2,92
Friuli-Venezia G.	45,43	4	2,90
Emilia-Romagna	49,24	9	2,62
Marche	46,43	4	1,67
Toscana	45,38	10	4,26
Umbria	42,20	2	3,54
Lazio	39,60	5	3,20
Campania	35,64	5	4,10
Abbruzzo	40,38	4	2,14
Molise	38,30	2	,99
Puglia	35,00	5	1,81
Basilicata	36,45	2	3,46
Calabria	30,88	5	1,86
Sicilia	32,51	9	3,48
Sardegna	36,88	4	1,40
Totale	**42,51**	**103**	**6,89**

Fig. 19 – Grafico degli "errori" del tasso di occupazione per regione



La tabella 13 e la figura 19, insomma, mostrano chiaramente la differenza tra il dato inserito in tabella, preciso ma non facilmente apprezzabile comparativamente, e la rappresentazione mediante le barre degli errori che peraltro è altrettanto precisa, in quanto il grafico consente, oltre al posizionamento spaziale della media, l'indicazione del suo preciso valore numerico.

È opportuno sottolineare una differenza tra i grafici degli errori e i grafici a scatola che sfugge nella nostra esemplificazione, nella quale è stata prevalentemente usata una variabile cardinale con una distribuzione di valori assai diversificati. A volte le variabili cardinali presentano pochi valori costituiti da numeri interi: si pensi ad una indagine in cui l'unità di analisi è la famiglia e la proprietà rilevata il numero dei figli. È anche il caso delle variabili quasi-cardinali generate dalle scale auto-ancoranti: ad esempio, si può chiedere di indicare il proprio atteggiamento verso una serie di istituzioni, indicandolo su una scala di sette valori, dove 1 significa "minima fiducia" e 7 "massima fiducia".

Usando i grafici a scatola per rappresentare la distribuzione della variabile nella quale abbiamo raggruppato in classi i tassi di occupazione otteniamo la figura 20, nella quale risulta evidente che le scatole presentano qualche anomalia. Nel caso delle Isole la scatola non viene nemmeno costruita e la spiegazione è semplice: questi grafici raffigurano valori posizionali (come la mediana) che non sono il prodotto di un calcolo statistico, ma vengono semplicemente "individuati" tra individuati tra i valori che fanno parte della distribuzione. I parametri statistici (come la media) si ottengono invece effettuando un calcolo e sono dunque "astrazioni" che spesso non fanno parte dei valori riscontrabili all'interno della distribuzione. Se in una famiglia c'è un figlio e in due altre famiglie ve ne sono due, si può affermare che nelle tre famiglie c'è in media un figlio e due terzi, anche se il valore 1,67 non appartiene alle distribuzione. Non esistono "terzi di figlio", ma non ci preoccupiamo di questo esito del calcolo, perché sappiamo benissimo che la media è un'astrazione statistica; la mediana, invece, viene individuata tra i valori della distribuzione ordinata del numero dei figli, e nell'esempio è pari a 2 (non ci si può fermare al valore 1, perché si raggiunge solo un terzo della distribuzione).

Tab. 14 – Parametri statistici del tasso di occupazione raggruppato in fasce per ripartizione

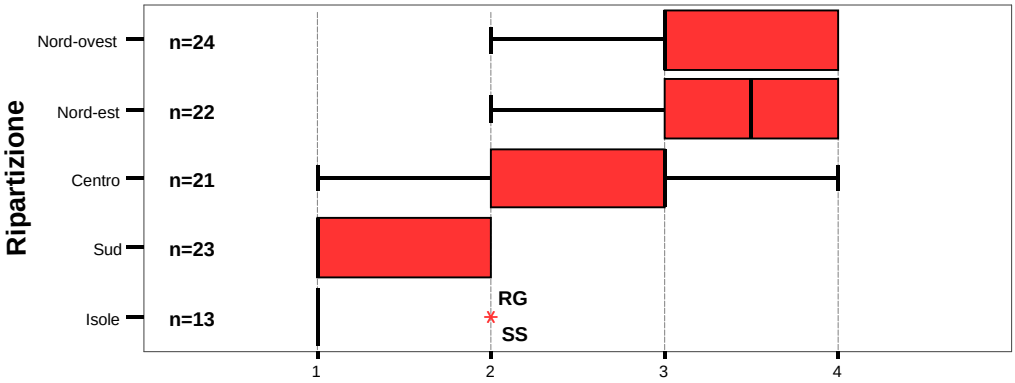
Tasso di occupazione: occupati per 100 abitanti (1999)

		Ripartizione				
		Nord-ovest	Nord-est	Centro	Sud	Isole
N	Validi	24	22	21	23	13
	Mancanti	0	0	0	0	0
Media		3,21	3,45	2,57	1,39	1,15
Mediana		3,00	3,50	3,00	1,00	1,00
Deviazione std.		,779	,596	,870	,499	,376
Minimo		2	2	1	1	1
Massimo		4	4	4	2	2
Percentili	25	3,00	3,00	2,00	1,00	1,00
	50	3,00	3,50	3,00	1,00	1,00
	75	4,00	4,00	3,00	2,00	1,00

Quando i valori presenti nella distribuzione sono molto pochi (anche se i casi fossero molto numerosi), può accadere, come si vede chiaramente dalla tabella 20, che ad un determinato valore corrispondano più valori posizionali. Infatti, la scatola relativa alle Isole nella figura 20 non viene tracciata, perché quasi tutti i casi registrano il valore 1, che perciò contiene tutti i valori posizionali della distribuzione, ad eccezione dei casi estremi (Ragusa e Sassari), che vengono indicati singolarmente. Nel Sud si riesce a costruite la scatola ma, visto che per quasi il 70% delle province si registra il valore 1, ad esso corrisponde la mediana, il minimo e il primo quartile della distribuzione; al valore 2 corrisponde il terzo quartile e, allo stesso tempo, il massimo della distribuzione. Interessante è il caso del Nord-est nel quale (piccola eccezione alla regola), siccome i casi sono in numero pari, vi sono due province al centro della distribu-

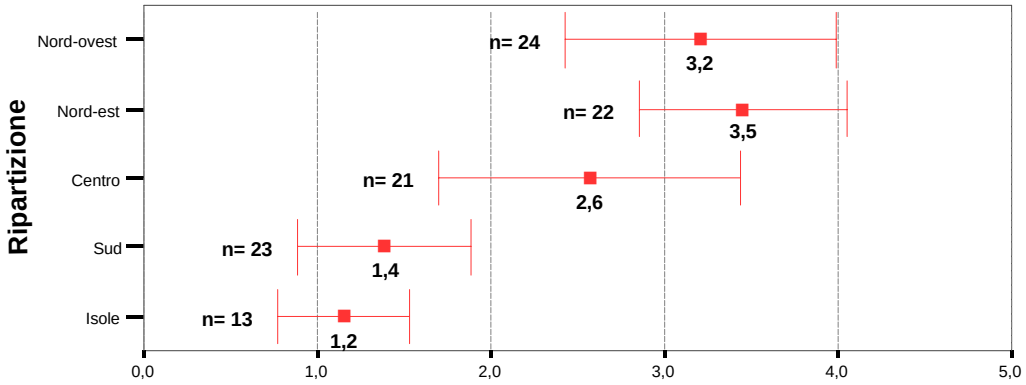
zione e la mediana (3,5) viene calcolata come media tra i valori 3 e 4 registrati in queste due province. Infine, le scatole relative al Nord-ovest e nel Nord-est sono del tutto simili: l'unica eccezione è costituita del diverso posizionamento della mediana, e peraltro sarebbe anch'essa stata pari a 3, se vi fosse stata una sola provincia in meno con il valore 4[13].

Fig. 20 – Grafico a scatole del tasso di occupazione in fasce per ripartizione



La media e la distribuzione standard vengono invece calcolate, ed ecco perché nel grafico degli errori della figura 21 vi sono valori che presentano decimali ed è possibile distinguere meglio le diverse situazioni. Si tenga conto che stiamo analizzando una variabile rispetto alla quale la ripartizione territoriale fa emergere differenze molto grandi, ma spesso le differenze non sono così nette e perciò può accadere che i grafici a scatola facciano apparire le distribuzioni molto simili tra loro. *In primis* ciò può avvenire perché, se la tendenza centrale della distribuzione è rappresentata dalle mediane, molte di esse corrisponderanno allo stesso valore che, invece, si differenzia se come misura di tendenza centrale usiamo la media.

Fig. 21 – Grafico a scatole del tasso di occupazione in fasce per ripartizione



Quando una variabile cardinale è ridotta in classi è possibile calcolare la media, usando il valore centrale di ognuna di esse, se hanno tutte un'uguale ampiezza. Nel nostro esempio abbiamo invece costruito le classi raggruppando un uguale numero di casi dispersi su un intervallo variabile di valori e perciò calcolando la media abbiamo operato una forzatura.

Confrontando le figure 20 e 21 possiamo desumere che, se si usa la mediana la rappresentazione è poco efficace, se si usa la media l'efficacia aumenta, ma si agisce in modo statisticamente non corretto. Di fronte a questa insoddisfacente alternativa, a mio avviso si può pro-

[13] Si noti anche che nella figura non sono state scritte le etichette descrittive accanto ai valori sulla scala (si poterebbero aggiungere manualmente), in quanto la variabile deve essere dichiarata "scala" (in Spss così si chiamano le variabili cardinali) per poter essere usata nella costruzione del grafico a scatole.

porre una soluzione di compromesso, molto utile quando ci si trovi ad operare con le scale di Likert. Queste scale sono molto diffuse e non a torto; a differenza delle scale ancoranti, che producono variabili quasi-cardinali ma sono piuttosto "astratte", le scale di Likert usano aggettivi (es. molto, abbastanza, poco, per nulla) o avverbi (sempre, spesso, raramente, mai) che consentono un ancoramento semantico più "convincente". È infatti usuale nel linguaggio quotidiano usare questi termini per riferire le nostre opinioni o i nostri comportamenti, molto di più che individuare su una scala il numero o la posizione spaziale corrispondente.

Ma le scale Likert producono variabili ordinali per le quali, a rigore, non è consentito il calcolo della media e, d'altro canto, usando per il confronto delle distribuzioni la mediana, si constata che molto spesso essa corrisponde alle due categorie intermedie e avremo perciò molte distribuzioni a "pari merito". La media, accoppiata alla mediana, può egregiamente servire proprio a disambiguare il significato delle mediane che hanno lo stesso valore. In una scala Likert con quattro categorie, due mediane pari al valore 3 (abbastanza) potranno essere apprezzate diversamente, tenendo sotto controllo anche la media. Riguardando i valori posizionali esposti nella tabella 14 e visualizzati nella figura 20, si vede che le mediane hanno il valore 3 sia nel Nord-ovest che nel Centro, ma le due medie (rispettivamente pari a 3,2 e 2,6) ci dicono che le due mediane "guardano" in direzioni opposte, e in particolare quella del Centro verso il valore 2.

Conclusioni

Abbiamo visto nel corso della trattazione in quanto modi la variabile che abbiamo deciso di scegliere per la nostra esercitazione può essere analizzata: dal semplice ordinamento dei valori, che consente un uso "idiografico" della variabile, all'estrema sintesi costituita dalla constatazione che nel 1999 in Italia era occupato il 42,5% della popolazione.

Abbiamo anche provato a mostrare come le varie tecniche si applichino all'analisi bivariata, nell'intento di provare quanto sia diversa la situazione occupazionale nelle grandi ripartizioni territoriali (che sono decisamente anche zone socio-economiche) in cui può essere divisa l'Italia. Ricordiamo che l'intento dello scritto era anche quello di mostrare che le tecniche statistiche e grafiche sono largamente equivalenti e che la sofisticazione dell'analisi statistica non è essenziale quanto operare con dati validi ed attendibili.

Si è visto, ad esempio, che:

- ordinando i valori dei tassi di occupazione, le province del Sud e delle Isole vanno ad occupare in larga prevalenza i gradini più bassi della graduatoria;
- il campo di variazione del tasso di occupazione è molto diverso dal Nord al Sud. Nelle province del Sud è incentrato su valori bassi (da 28 al 43% circa) e in particolare nelle Isole sono così bassi (dal 27 al 39%), che non c'è alcuna sovrapposizione con le due aree del Nord, che complessivamente hanno tassi di occupazione dal 40 al 56%;
- dividendo in classi la distribuzione, le province del Nord si collocano quasi esclusivamente ai livelli alti o medio-alti, mentre quelle del Sud e delle Isole si collocano esclusivamente ai livelli bassi o medio-bassi dei tassi;
- anche i valori mediani e medi sono molto più bassi al Sud e, in generale, scendono dal 48% circa che si registra al Nord con entrambi gli strumenti statistici a valori tra il 34 e il 36% che si registrano invece nel Sud e nelle Isole.

Tutto ciò è stato anche visualizzato mediante alcune tecniche grafiche di analisi e rappresentazione dei dati: grafici a barre, istogrammi, grafici ramo-foglia, a scatole, etc. che rendono evidenti le differenze in vario modo e con diversa efficacia: ciò che comunque non cambia è il problema occupazionale del Sud e delle Isole, che risalta in tutta la sua drammaticità, qualunque sia lo strumento usato.

Riferimenti bibliografici

Cardano M. e Miceli R. (a cura di) (1991), *Il linguaggio delle variabili*, Rosenberg & Sellier, Torino.

Delli Zotti G. (2005a), *Tecniche grafiche di analisi e rappresentazione dei dati*, Angeli, Milano, *in corso di pubblicazione*.

Delli Zotti, G. (2005b), *Come fare "a fette" una distribuzione di frequenza*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 1-2005, Trieste (www.dsu.units.it/quaderni/dsu0105.pdf).

Delli Zotti G. (2004), *Introduzione alla ricerca sociale. Problemi e qualche soluzione. Nuova edizione aggiornata e rivista*, Angeli, Milano.

Delli Zotti G. (1992a), *Il problema più importante per noi*, in A. Marradi, G. Gasperoni (cur.), *Costruire il dato 2. Vizi e virtù di alcune tecniche di raccolta delle informazioni*, Angeli, Milano, pp. 130-141.

Delli Zotti G. (1992b), *Il problema più importante per noi ... Opzioni nella formulazione, codifica ed elaborazione di domande di atteggiamento*, Quaderni dell'Isig. Programma "Metodologia", 92-1, Isig, Gorizia (www.uniud.it/dest/docenti/dellizotti/probimp.pdf).

Delli Zotti G. (1985), "Tipologia delle matrici utilizzate nella ricerca sociale", *Rassegna Italiana di Sociologia*, XXVI, 2, pp. 141- 168 (www.uniud.it/dest/docenti/dellizotti/matrix.pdf).

Tukey J.W. (1977), *Exploratory Data Analysis*, Reading, Addison-Wesley.

Altri testi di interesse metodologico disponibili online:

Delli Zotti, G. (2005), *Come creare un indice o una tipologia*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 2-2005, Trieste (www.dsu.units.it/quaderni/ dsu0205.pdf).

Delli Zotti G. (1999), *L'analisi esplorativa delle tabelle di contingenza. Nuova edizione - esempi realizzati con Spss per Windows 7.5*, Quaderni del Dipartimento Est, 99-15, Dest, Udine (www.uniud.it/dest/docenti/dellizotti/tabmulti.pdf). On-line anche nella Rassegna Italiana di Valutazione, n.1, 2000, (www.valutazioneitaliana.it/riv/rivista2000/tabmulti.doc).