Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013
Corso di laurea in *Teorie, culture e tecniche per il Servizio sociale*
Università degli Studi di Macerata

STATISTICA PER LE SCIENZE SOCIALI

MACERATENSIS · SIGILLVM · VNIVERSITATIS

il Mulino

GEORGE W. BOHRNSTEDT
DAVID KNOKE

L'analisi della varianza

In questo capitolo presenteremo una versione particolare del modello lineare generale, in cui la variabile indipendente è formata da un insieme di categorie discrete e la variabile dipendente è di tipo continuo. Questa tecnica è chiamata **analisi della varianza** o, in breve, ANOVA. In sostanza, mostreremo come sottoporre a verifica l'ipotesi secondo la quale le medie campionarie di J gruppi diversi provengono dalla stessa popolazione, anziché da popolazioni differenti. Presenteremo anche un caso speciale di ANOVA in cui vengono analizzate le differenze fra le medie di due soli gruppi ($J = 2$). In questo capitolo prenderemo in considerazione solo modelli in cui le osservazioni sono classificate secondo un'unica variabile discreta e che, per questo motivo, sono noti come analisi della varianza *a una dimensione* (o *a una via*). Modelli ANOVA più complessi nei quali le osservazioni sono classificate simultaneamente secondo due o più variabili sono troppo complessi per essere discussi in questa sede.

1. La logica dell'ANOVA

1.1. Un esempio

Per illustrare la logica dell'analisi della varianza faremo ricorso a un esempio basato su dati relativi agli Stati Uniti. Oggetto della nostra analisi è una questione dibattuta da tempo: è possibile affermare che stati appartenenti a regioni diverse sono caratterizzati da tassi di criminalità differenti? Specificamente, ci si chiede se sia vero che la criminalità è più diffusa negli stati del Sud e dell'Ovest, in virtù della presunta «cultura della violenza» che si ritiene caratterizzi queste due regioni. Per sottoporre a verifica questa ipotesi utilizzeremo dati del

1985 relativi ai cinquanta stati americani, classificati secondo le quattro grandi regioni definite dall'Istituto centrale di statistica statunitense. La variabile dipendente sarà il tasso di crimini gravi noti alla polizia, inteso come numero di crimini per 1.000 abitanti.

La nostra ipotesi di ricerca iniziale è che i tassi di crimini gravi siano più alti negli stati del Sud (*S*), seguiti – in ordine decrescente – dagli stati dell'Ovest (*O*), del Nordest (*N*) e del cosiddetto Midwest (*M*). L'ipotesi nulla, invece, sostiene che le medie di queste quattro «popolazioni» – indicate con i simboli μ_S, μ_O, μ_N e μ_M – sono uguali una all'altra. Dunque, secondo questa ipotesi le quattro medie regionali sono anche uguali alla **media generale** dell'intera popolazione di 50 stati, indicata con μ. Tuttavia, se l'ipotesi alternativa è vera e l'ipotesi nulla viene rifiutata, ci aspettiamo di trovare che $\mu_S > \mu_O > \mu_N > \mu_M$.

I modelli ANOVA permettono di verificare l'ipotesi nulla secondo la quale tutte le J medie campionarie provengono dalla stessa popolazione e, quindi, sono uguali una all'altra. Formalmente, l'ipotesi nulla implicita in un modello ANOVA è la seguente:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_J$$

L'ipotesi alternativa, invece, afferma che almeno una delle medie campionarie proviene da una popolazione la cui media differisce dalle medie delle altre popolazioni. Il rifiuto dell'ipotesi nulla implica una fra diverse possibilità alternative.

1. La media di ogni popolazione differisce dalla media di tutte le altre popolazioni, cioè

$$H_1: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_J$$

2. Le medie di alcuni sottoinsiemi di popolazioni differiscono l'una dalle altre (ad esempio, μ_1 è diversa da μ_2 ma è uguale a μ_3 e μ_4).
3. Alcune combinazioni di medie differiscono da una singola media o da un'altra combinazione di medie (ad esempio, μ_1 differisce dalla media di μ_3 e μ_4).

L'analisi della varianza sottopone a verifica l'ipotesi nulla secondo la quale le medie delle varie popolazioni sono uguali. Se i risultati del test portano al rifiuto dell'ipotesi nulla, rimane da stabilire in che modo le medie differiscono fra loro.

1.2. Gli effetti delle variabili

Per esaminare l'effetto esercitato da una variabile discreta su una variabile dipendente continua si consideri un'unica popolazione all'interno della quale la media della variabile dipendente è uguale a μ. Se, come afferma l'ipotesi nulla H_0, le medie J dei vari gruppi sono uguali

Page 3 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

una all'altra, allora esse sono anche uguali alla media generale della popolazione μ. Questa semplice relazione rappresenta la base per misurare gli **effetti** esercitati dalla variabile indipendente discreta sulla variabile dipendente continua. L'effetto esercitato dal gruppo j, indicato con il simbolo α_j, viene definito come la differenza fra la media di quel gruppo e la media generale:

$$\alpha_j = \mu_j - \mu$$

Se il gruppo j non esercita alcun effetto sulla variabile dipendente, allora $\alpha_j = 0$, cioè $\mu_j = \mu$. Se, invece, il gruppo j esercita un effetto, allora α_j assumerà un valore positivo e negativo, a seconda che la media del gruppo sia maggiore o minore della media generale μ.

Riprendendo il nostro esempio sui tassi di criminalità degli stati americani, i dati mostrano che, nel 1985, il tasso medio di crimini gravi (per 1.000 abitanti) era pari a 4,77. I tassi regionali sono i seguenti: Sud, $\mu_S = 4{,}59$; Ovest, $\mu_O = 5{,}8$; Nordest, $\mu_N = 4{,}36$; e Midwest, $\mu_M = 4{,}21$. Pertanto, gli effetti esercitati dalle singole regioni sono: $\alpha_S = -0{,}18$; $\alpha_O = +1{,}03$; $\alpha_N = -0{,}41$; e $\alpha_M = -0{,}56$. Possiamo già vedere che il Sud e l'Ovest sono caratterizzati da tassi di criminalità maggiori di quelli osservati nelle altre due regioni, anche se l'effetto associato all'Ovest è inaspettatamente molto più intenso di quello associato al Sud. In effetti, la media del Sud è minore della media generale. Ci rimane ancora da stabilire se queste differenze osservate sono statisticamente significative.

1.3. Il modello ANOVA

L'analisi della varianza è finalizzata a determinare quale proporzione della variazione totale nella variabile dipendente Y è attribuibile alle singole osservazioni i appartenenti al gruppo j. Il modello ANOVA generale con una singola variabile indipendente scompone ogni valore osservato in tre componenti:

$$Y_{ij} = \mu + \alpha_j + e_{ij}$$

dove:

Y_{ij} = valore dell'osservazione i del gruppo j
μ = media generale, comune a tutte le osservazioni della popolazione
α_j = effetto esercitato dal gruppo j, comune a tutte le osservazioni di quel gruppo
e_{ij} = errore, unico all'osservazione i del gruppo j.

Riordinando questa formula, possiamo vedere che l'**errore** – o residuo – è quella parte di un valore osservato che non può essere attribuita né alla componente comune né alla componente di gruppo:

$$e_{ij} = Y_{ij} - \mu - \alpha_j$$

Nell'analisi della varianza l'errore può essere visto come la discrepanza fra i valori osservati e quelli predetti dall'appartenenza a un dato gruppo. In sostanza, tale errore tiene conto del fatto che i vari casi i appartenenti a un gruppo j non hanno tutti lo stesso valore osservato Y_j.

2. Le tavole ANOVA

2.1. La somma dei quadrati

Per determinare la proporzione della varianza di Y_{ij} attribuibile agli effetti di gruppo (α_j) e la proporzione che, invece, è attribuibile all'errore (e_{ij}), bisogna partire dal numeratore della varianza campionaria:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

Ognuna delle N osservazioni presenti in un campione appartiene a uno dei J gruppi precedentemente identificati. Se n_j indica il numero di osservazioni appartenenti al gruppo j, allora $n_1 + n_2 + \ldots + n_J = N$. In altri termini, la somma delle osservazioni appartenenti a ciascuno dei J gruppi è uguale alla dimensione complessiva del campione N. Pertanto, se a ciascuna osservazione associamo sia un indice individuale (i) che un indice di gruppo (j), possiamo riscrivere il numeratore della varianza come segue:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2$$

Il termine destro dell'equazione rappresenta la **somma dei quadrati totale**, spesso indicata con il simbolo SQ_{TOTALE}. In sostanza, essa equivale alla somma delle deviazioni al quadrato di ciascun valore dalla media generale di tutti i gruppi ($\overline{Y}$).

Per fare un esempio, supponiamo di avere un campione con $N = 5$ osservazioni, ognuna delle quali appartiene al gruppo 1 o al gruppo 2 (cioè, $J = 2$); supponiamo anche che $n_1 = 3$ e $n_2 = 2$. Espandendo il termine destro dell'espressione riportata sopra, si ottiene:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2 = [(Y_{11} - \overline{Y})^2 + (Y_{21} - \overline{Y})^2 + (Y_{31} - \overline{Y})^2] +$$

$$+[(Y_{12} - \overline{Y})^2 + (Y_{22} - \overline{Y})^2]$$

Page 5 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

La prima coppia di parentesi quadre contiene la somma relativa al gruppo 1, mentre la seconda coppia di parentesi quadre contiene la somma relativa al gruppo 2.

L'analisi della varianza a una via suddivide la somma dei quadrati totale in due componenti: 1) la **somma dei quadrati fra i gruppi** ($SQ_{INTERGRUPPO}$), pari alla somma delle deviazioni al quadrato di ciascuna media di gruppo dalla media generale dei gruppi; 2) la **somma dei quadrati nei gruppi** ($SQ_{INTRAGRUPPO}$), pari alla somma delle deviazioni al quadrato delle osservazioni di ciascun gruppo dalla media di quel gruppo. Nel formare questa suddivisione è possibile aggiungere e sottrarre simultaneamente un dato valore da qualsiasi espressione senza alterarne il valore. Pertanto, si sottragga e si aggiunga $\overline{Y}_j$ alla deviazione di ciascun caso dalla media generale:

$$Y_{ij} - \overline{Y} = Y_{ij} + (\overline{Y}_j - \overline{Y}_j) - \overline{Y}$$

A questo punto si devono riordinare i quattro termini a destra del segno uguale in modo tale da ottenere due deviazioni:

$$Y_{ij} - \overline{Y} = (Y_{ij} - \overline{Y}_j) + (\overline{Y}_j - \overline{Y})$$

Il primo termine a destra del segno uguale, cioè $(Y_{ij} - \overline{Y}_j)$, rappresenta la deviazione dell'osservazione i del gruppo j dalla media di quel gruppo; in sostanza, si tratta della stima campionaria dell'errore e_{ij}. Il secondo termine a destra del segno uguale, cioè $(\overline{Y}_j - \overline{Y})$, rappresenta invece la deviazione della media del gruppo j dalla media generale di tutti i gruppi; in sostanza, si tratta della stima campionaria dell'effetto esercitato dal gruppo j, cioè $\alpha_j = \mu_j - \mu$. Elevando al quadrato entrambi i lati dell'equazione e sommando i diversi valori si ottiene:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2 + \sum_{j=1}^{J} n_j(\overline{Y}_j - \overline{Y})^2$$

Questa uguaglianza dimostra che la somma dei quadrati totale può essere sempre suddivisa in due componenti:

$$SQ_{INTRAGRUPPO} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

e

$$SQ_{INTERGRUPPO} = \sum_{j=1}^{J} n_j(\overline{Y}_j - \overline{Y})^2$$

Pertanto:

$$SQ_{TOTALE} = SQ_{INTRAGRUPPO} + SQ_{INTERGRUPPO}$$

Se si eccettuano i casi in cui tutte le osservazioni di un dato campione assumono lo stesso valore, la loro varianza è sempre maggiore di zero. Una parte di questa varianza può essere attribuita agli effetti dei gruppi ai quali le osservazioni appartengono. In altri termini, la somma dei quadrati *fra i gruppi* esprime l'effetto complessivo esercitato sulla variabile dipendente continua dalla variabile indipendente discreta. Tuttavia, all'interno di ogni singolo gruppo le osservazioni possono essere diverse fra loro a causa di fattori casuali come la variazione campionaria o di effetti esercitati da variabili indipendenti non osservate. Dunque, la somma dei quadrati *nei gruppi* esprime l'influenza di fattori non misurati. Trattare le osservazioni appartenenti a un dato gruppo come se fossero tutte uguali sarebbe scorretto.

2.2. La somma dei quadrati nell'esempio dei tassi di criminalità

Se l'ipotesi nulla relativa alle differenze regionali fra stati in termini di tassi di crimini gravi fosse vera, allora dovremmo osservare la stessa media in tutti e quattro i gruppi considerati, cioè: $\mu_S = \mu_O = \mu_N = \mu_M$. In questo caso la somma dei quadrati *fra i gruppi* ($SQ_{INTERGRUPPO}$) sarebbe uguale a zero e, quindi, $SQ_{TOTALE} = SQ_{INTRAGRUPPO}$. In altri termini, tutta la variazione osservata nei tassi di criminalità statali sarebbe dovuta a «errore», cioè a fattori casuali. Nei casi in cui la variabile indipendente non esercita alcun effetto, il modello ANOVA generale per i dati campionari si riduce a:

$$Y_{ij} = \overline{Y} + e_{ij}$$

Supponiamo, invece, che la regione influisca sul tasso di criminalità, e cioè che $\mu_S > \mu_O > \mu_N > \mu_M$. Inoltre, supponiamo che tutti gli stati appartenenti a ogni data regione siano caratterizzati dallo stesso tasso di crimini gravi, cioè che le deviazioni delle singole osservazioni dalle medie dei gruppi corrispondenti siano tutte uguali a zero. In questo caso, i tassi di criminalità osservati negli stati appartenenti alle quattro regioni sarebbero uguali a:

$$Y_{i,S} = \overline{Y} + a_S$$

$$Y_{i,O} = \overline{Y} + a_O$$

$$Y_{i,N} = \overline{Y} + a_N$$

$$Y_{i,M} = \overline{Y} + a_M$$

dove i simboli a_j indicano le stime campionarie degli effetti α_j all'opera nella popolazione.

In questo caso la somma dei quadrati *nei gruppi* ($SQ_{INTRAGRUPPO}$) è uguale a zero e, quindi, $SQ_{TOTALE} = SQ_{INTERGRUPPO}$. Ciò significa che *tutta* la variazione osservata nei tassi di criminalità statali è attribuibile alla regione alla quale ogni stato appartiene. Naturalmente, in realtà nessuna singola variabile indipendente può essere in grado di spiegare *tutta* la variazione osservata in una variabile dipendente. In primo luogo, quest'ultima potrebbe essere influenzata in modo *sistematico* anche da altre variabili indipendenti *non misurate*. In secondo luogo, la variabile dipendente potrebbe essere influenzata anche da fattori *casuali* di vario tipo. Poiché la variazione osservata nella variabile dipendente può derivare da fattori non misurati di tipo sia sistematico che casuale, le equazioni che rappresentano i dati campionari devono includere anche l'errore:

$$Y_{i,S} = \bar{Y} + a_S + e_{i,S}$$

$$Y_{i,O} = \bar{Y} + a_O + e_{i,O}$$

$$Y_{i,N} = \bar{Y} + a_N + e_{i,N}$$

$$Y_{i,M} = \bar{Y} + a_M + e_{i,M}$$

La tabella 4.1 mostra che in tutte e quattro le regioni considerate esiste una sostanziale variazione *intragruppo* (errore) fra i diversi stati. Inoltre, il fatto che le quattro medie regionali siano chiaramente diverse suggerisce la presenza anche di un certo grado di variazione *fra i gruppi*. Pertanto, la variazione complessiva osservata nei tassi di criminalità statali (SQ_{TOTALE}) sembra essere attribuibile sia a $SQ_{INTRAGRUPPO}$ che a $SQ_{INTERGRUPPO}$. Per determinare se queste differenze apparenti sono statisticamente significative bisogna calcolare diverse misure, inclusi le varie somme dei quadrati, i quadrati medi e il quoziente F.

Per calcolare la somma dei quadrati totale si deve sottrarre il valore di ciascuna osservazione dalla media generale, elevare al quadrato il risultato ottenuto e sommare tutti i quadrati:

$$SQ_{TOTALE} = (7{,}56 - 4{,}77)^2 + (6{,}57 - 4{,}77)^2 + \ldots + (2{,}69 - 4{,}77)^2 = 80{,}13$$

La somma dei quadrati fra i gruppi viene calcolata direttamente sottraendo la media di ciascun gruppo dalla media generale, elevando al quadrato il risultato ottenuto, moltiplicandolo per la numerosità del gruppo e sommando tutti i valori ottenuti. Nel caso del nostro esempio si ha:

$$SQ_{INTERGRUPPO} = (16)(4{,}59 - 4{,}77)^2 + (13)(5{,}80 - 4{,}77)^2 + (9)(4{,}36 - 4{,}77)^2 + (12)(4{,}21 - 4{,}77)^2 = 19{,}38$$

TAB. 4.1. *Tassi di criminalità registrati negli Stati americani nel 1985*

Stato	Y_{ij}	Stato	Y_{ij}
Sud		Ovest	
Florida	7,56	Colorado	6,92
Texas	6,57	Arizona	6,90
Louisiana	5,71	Oregon	6,73
Oklahoma	5,42	Nevada	6,68
Maryland	5,38	New Mexico	6,54
Delaware	4,96	Washington	6,53
Georgia	4,89	California	6,52
South Carolina	4,83	Alaska	5,82
Tennessee	4,32	Utah	5,31
North Carolina	4,11	Hawaii	5,20
Alabama	3,92	Montana	4,33
Virginia	3,78	Wyoming	4,01
Arkansas	3,58	Idaho	3,86
Mississippi	3,27		
Kentucky	2,95		
West Virginia	2,23		
$\Sigma Y_{ij} = 73{,}49$		$\Sigma Y_{ij} = 75{,}34$	
$\overline{Y}_S = 4{,}59$		$\overline{Y}_O = 5{,}80$	
$s_S = 1{,}37$		$s_O = 1{,}13$	
Nord-Est		Midwest	
New York	5,56	Michigan	6,36
Connecticut	5,42	Illinois	5,27
New Jersey	5,10	Missouri	4,65
Massachusetts	4,72	Kansas	4,35
Rhode Island	4,71	Ohio	4,26
Vermont	3,87	Indiana	4,16
Maine	3,67	Minnesota	4,13
New Hampshire	3,23	Wisconsin	3,98
Pennsylvania	2,99	Iowa	3,94
		Nebraska	3,78
		South Dakota	2,98
		North Dakota	2,69
$\Sigma Y_{ij} = 39{,}26$		$\Sigma Y_{ij} = 50{,}56$	
$\overline{Y}_N = 4{,}36$		$\overline{Y}_M = 4{,}21$	
$s_N = 0{,}95$		$s_M = 0{,}96$	
	$\overline{Y} = 4{,}77$		
	$s_Y = 1{,}28$		

Infine, la somma dei quadrati nei gruppi può essere calcolata sottraendo il valore di ciascuna osservazione dalla media del gruppo corrispondente, elevando al quadrato il risultato ottenuto e sommando tutti i quadrati:

$$SQ_{INTRAGRUPPO} = (7{,}56 - 4{,}59)^2 + (6{,}57 - 4{,}59)^2 + \ldots + (2{,}69 - 4{,}21)^2 = \\ = 60{,}75$$

In alternativa, la somma dei quadrati nei gruppi può essere calcolata sottraendo la somma dei quadrati fra i gruppi dalla somma dei quadrati totale: 80,13 – 19,38 = 60,75.

Page 9 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

Page 10 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

2.3. I quadrati medi

Il passo successivo, nell'analisi della varianza, consiste nel calcolare i **quadrati medi** corrispondenti a $SQ_{INTERGRUPPO}$ e a $SQ_{INTRAGRUPPO}$; ognuno di essi rappresenta la stima di una varianza: la prima attribuibile agli effetti di gruppo, la seconda attribuibile all'errore. Se non esiste alcun effetto di gruppo, allora le due stime saranno identiche. Se, al contrario, esiste un effetto di gruppo significativo, allora il **quadrato medio fra i gruppi** ($SQ_{INTERGRUPPO}$) sarà maggiore del **quadrato medio nei gruppi** ($SQ_{INTRAGRUPPO}$).

I quadrati medi vengono calcolati dividendo ciascuna somma di quadrati per l'appropriato numero di gradi di libertà. I gradi di libertà associati alla varianza fra i gruppi sono semplicemente pari a $J-1$, perché una volta che la media generale e la media dei primi $J-1$ gruppi sono note, la media del gruppo J è determinata automaticamente. Pertanto, per calcolare il quadrato medio fra i gruppi è sufficiente dividere la somma dei quadrati fra i gruppi per $J-1$:

$$QM_{INTERGRUPPO} = \frac{\sum_{j=1}^{J} n_j (\overline{Y}_j - \overline{Y})^2}{J-1} = \frac{SQ_{INTERGRUPPO}}{J-1}$$

Nel nostro esempio, quindi, si ha:

$$QM_{INTERGRUPPO} = 19{,}38/(4-1) = 6{,}46$$

I gradi di libertà associati alla varianza nei gruppi sono uguali a $N-J$. Ciascun gruppo, infatti, ha $n_j - 1$ gradi di libertà; sommando i gradi di libertà di tutti i gruppi si ottiene:

$$\begin{aligned}(n_1 - 1) + (n_2 - 1) + \ldots + (n_J - 1) &= \\ = (n_1 + n_2 + \ldots + n_J) - (1 + 1 + \ldots + 1) &= \\ = N - J\end{aligned}$$

Dunque, per calcolare il quadrato medio nei gruppi bisogna dividere la somma dei quadrati nei gruppi per $N-J$:

$$QM_{INTRAGRUPPO} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2}{N-J} = \frac{SQ_{INTRAGRUPPO}}{N-J}$$

Nel nostro esempio, quindi, si ha:

$$QM_{INTRAGRUPPO} = 60{,}75/(50 - 4) = 1{,}32$$

Come possiamo vedere, nel nostro esempio la varianza stimata attribuibile agli effetti di gruppo è maggiore di quella attribuibile all'errore (6,46 contro 1,32). Una differenza di questa entità si verificherebbe solo se la regione esercitasse un effetto significativo sul tasso di crimini gravi. L'operazione finale, nell'analisi della varianza, consiste nel determinare quanto maggiore $QM_{INTERGRUPPO}$ deve essere rispetto a $QM_{INTRAGRUPPO}$ affinché l'ipotesi nulla possa essere rifiutata.

2.4. Il quoziente *F*

Nell'analisi della varianza la statistica F viene calcolata semplicemente come rapporto fra i due quadrati medi:

$$F_{J-1,N-J} = \frac{QM_{INTERGRUPPO}}{QM_{INTRAGRUPPO}}$$

La distribuzione campionaria di F può essere utilizzata per verificare l'ipotesi nulla secondo la quale nessuna parte della varianza osservata nella variabile dipendente è dovuta agli effetti di gruppo. Due assunti devono essere soddisfatti:

1. i J gruppi devono essere estratti indipendentemente da una popolazione normalmente distribuita;
2. la varianza della popolazione deve essere identica alle varianze dei J gruppi.

Quest'ultimo assunto è noto come **omoschedasticità**. Quando le J varianze della popolazione differiscono, si è in presenza di *eteroschedasticità*.

Se entrambi gli assunti sono soddisfatti, la statistica F risulta distribuita secondo una distribuzione F teorica con $J - 1$ gradi di libertà al numeratore e $N - J$ gradi di libertà al denominatore. Poiché nell'analisi della varianza l'ipotesi alternativa tipicamente si attende che la stima della varianza fra i gruppi sia maggiore della stima della varianza nei gruppi, il test di significatività appropriato è a una coda. Se il quoziente F ottenuto è maggiore del valore critico associato a un dato livello α, rifiuteremo l'ipotesi nulla e concluderemo che la variabile indipendente influenza la variabile dipendente in modo significativo.

Per $\alpha = 0{,}05$ e 0,01 la tavola B4 in appendice B fornisce 2 valori critici corrispondenti a numerose distribuzioni F. I gradi di libertà del numeratore (ν_1) sono riportati nelle colonne, mentre quelli del denominatore (ν_2) sono riportati nelle righe. Nella cella che si trova all'in-

TAB. 4.2. *Tavola ANOVA relativa all'esempio sulla criminalità*

Fonte	*SQ*	*gl*	*QM*	*F*
Fra i gruppi	19,38	3	6,46	4,89[a]
Nei gruppi	60,75	46	1,32	
Totale	80,13	49		

[a] Significativo al livello $\alpha = 0{,}01$.

crocio della colonna e della riga appropriate si trova il valore critico di F necessario per rifiutare l'ipotesi nulla, dato il livello α prescelto.

Tornando al nostro esempio, il quoziente F assume il seguente valore:

$$F_{3,46} = \frac{6,46}{1,32} = 4,89$$

Ponendo $\alpha = 0{,}01$, consultiamo la tavola B4 dell'appendice B, cercando la cella corrispondente a $\nu_1 = 3$ e $\nu_2 = 46$. Come possiamo vedere il valore critico è uguale a 4,24. Poiché il quoziente F da noi calcolato equivale a 4,89, possiamo rifiutare l'ipotesi nulla con una probabilità di commettere un errore di I tipo pari a 0,01; concluderemo, dunque, che nel 1985 la regione influiva in modo significativo sui tassi di criminalità osservati nei 50 stati americani.

I risultati dell'analisi della varianza a una via sono generalmente presentati sotto forma di una **tavola ANOVA riassuntiva**, di cui troviamo un esempio nella tabella 4.2. La tavola ANOVA offre tutte le informazioni rilevanti di cui si ha bisogno per interpretare il test di significatività: le somme dei quadrati, i gradi di libertà, i quadrati medi, il quoziente F e il livello di probabilità.

3. Il test delle differenze fra due medie

Generalmente l'analisi della varianza viene applicata quando la variabile indipendente è formata da tre o più categorie. Tuttavia, l'approccio ANOVA è appropriato anche quando le medie da confrontare sono solo due. In tali casi è possibile fare ricorso anche alle distribuzioni Z e t, discusse nel capitolo precedente. In questo paragrafo mostreremo che i test delle differenze fra due medie basati su queste distribuzioni sono casi speciali della più generale analisi della varianza e danno luogo agli stessi risultati. A titolo di esempio utilizzeremo i dati della General Social Survey del 1991 e sottoporremo a verifica l'ipotesi nulla secondo la quale l'orientamento politico degli uomini e quello delle donne non sono significativamente diversi. La

Page 12 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

variabile «orientamento politico» è stata operativizzata chiedendo agli intervistati di autocollocarsi su una scala a sette posizioni che va da «fortemente progressista» (punteggio 1) a «fortemente conservatore» (punteggio 7), con la categoria «moderato» in posizione intermedia (punteggio 4). La nostra ipotesi alternativa è che gli uomini e le donne sono caratterizzati da orientamenti politici differenti; poiché non abbiamo nessun motivo plausibile per formulare aspettative su quale dei due sessi sia più conservatore, l'ipotesi alternativa assume un carattere bidirezionale:

$$H_0: \mu_F = \mu_M$$

$$H_1: \mu_F \neq \mu_M$$

3.1. Il test *Z*

Come abbiamo già visto (cfr. cap. 3, par. 5), il teorema del limite centrale assicura che, per campioni sufficientemente numerosi, la distribuzione di tutte le medie campionarie (cioè la distribuzione campionaria) sarà normale, e la sua media sarà una stima corretta della media della popolazione dalla quale il campione è stato estratto. Quando si tratta di verificare se esiste una differenza fra le medie di due popolazioni, assume un ruolo rilevante il seguente corollario del teorema del limite centrale:

La distribuzione della differenza fra due medie campionarie, calcolate su campioni casuali di numerosità N_1 e N_2 tratti da popolazioni con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2, segue una distribuzione normale con media pari a $\mu_1 - \mu_2$ e deviazione standard (errore standard) pari a $\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}$. La forma delle distribuzioni delle popolazioni può essere di qualunque tipo.

In sostanza, questo corollario ci dice che:

$$\mu_{(\bar{Y}_1 - \bar{Y}_2)} = \mu_1 - \mu_2$$

e

$$\sigma_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}$$

La tavola delle probabilità associate alla distribuzione normale ci permette di effettuare una **verifica di ipotesi sulla differenza fra medie** avente per oggetto le medie di due popolazioni. I requisiti da soddisfare sono solo due: N_1 e N_2 devono essere sufficientemente grandi (in pratica, $N_1 + N_2$ dovrebbe essere *almeno* pari a 60, preferibilmente maggiore di 100), e le varianze di entrambe le popolazioni (σ_1^2 e σ_2^2) devono essere note. Poiché i parametri della popolazione sono rara-

Page 13 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

mente noti, generalmente vengono stimati utilizzando le statistiche campionarie. Se $N_1 + N_2$ è sufficientemente grande, le varianze campionarie s_1^2 e s_2^2 possono essere usate per sostituire le varianze delle popolazioni σ_1^2 e σ_2^2 nella stima dell'errore standard:

$$\hat{\sigma}_{(\bar{Y}_1-\bar{Y}_2)} = \sqrt{s_1^2/N_1 + s_2^2/N_2}$$

Per sottoporre a verifica l'ipotesi nulla secondo la quale $\mu_1 = \mu_2$ (che equivale ad affermare che $\mu_1 - \mu_2 = 0$) si può fare riferimento alla procedura descritta nel quadro 3.4. Per il nostro esempio sull'orientamento politico delle donne e degli uomini americani, poniamo $\alpha = 0{,}01$. Dato l'ampio numero di casi ($N_1 = 621$ uomini e $N_2 = 838$ donne), la distribuzione Z rappresenta la distribuzione teorica più appropriata per effettuare il test desiderato. La tavola B1 in appendice B mostra che i valori critici per un test a due code con $\alpha = 0{,}01$ equivalgono a $-Z_{\alpha/2} = -2{,}58$ e $+Z_{\alpha/2} = 2{,}58$.

Secondo l'ipotesi nulla H_0: $\mu_1 - \mu_2 = 0$, la media della prima popolazione è uguale alla media della seconda popolazione. Il nostro test consiste nel determinare se la differenza osservata fra le due medie campionarie $(\bar{Y}_1 - \bar{Y}_2)$ si sarebbe potuta verificare nel caso in cui la vera differenza fra le medie delle due popolazioni fosse uguale a zero (cioè se $\mu_1 - \mu_2 = 0$). Formalmente:

$$Z_{(\bar{Y}_1-\bar{Y}_2)} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma_{(\bar{Y}_1-\bar{Y}_2)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}}$$

Utilizzando invece la stima campionaria di $\sigma_{(\bar{Y}_1-\bar{Y}_2)}$, la formula del test diventa:

$$Z_{(\bar{Y}_1-\bar{Y}_2)} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{(\bar{Y}_1-\bar{Y}_2)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

La figura 4.1 illustra due esempi di distribuzione campionaria in cui l'ipotesi nulla H_0: $\mu_1 - \mu_2 = 0$ è vera. Nel riquadro superiore la differenza osservata fra le due medie campionarie è pari a un piccolo valore positivo prossimo alla differenza nulla ipotizzata. In un caso come questo non rifiuteremmo l'ipotesi nulla in quanto la piccola dif-

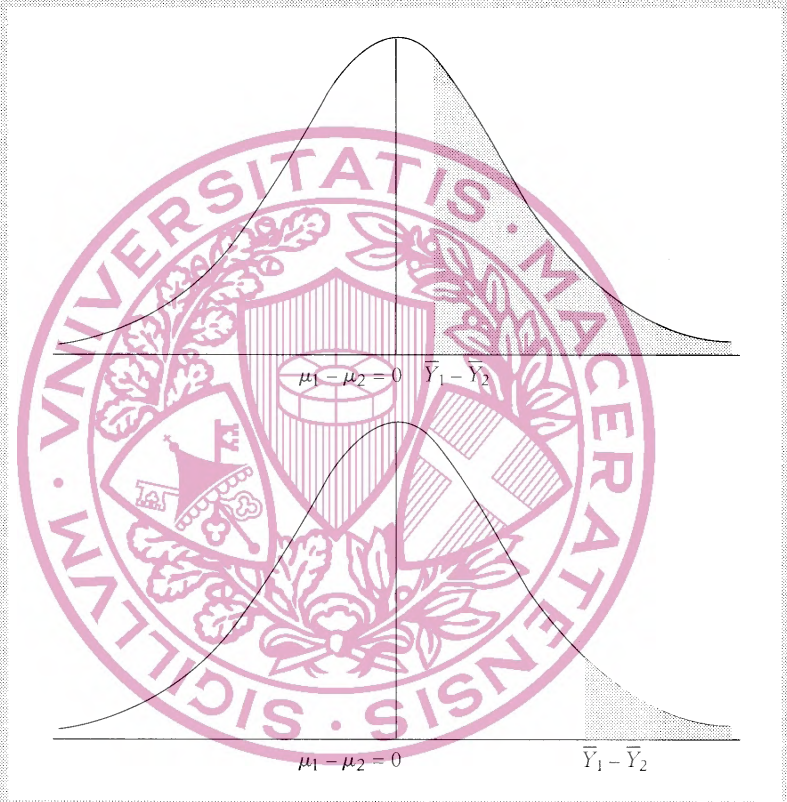

FIG. 4.1. Esempi di risultati ottenuti sottoponendo a verifica ipotesi sulle differenze fra due medie (ipotesi nulla vera).

Fonte: Adattato da D. Knoke e G.W. Bohrnstedt, *Basic Social Statistics*, Itasca, Ill., Peacock Publishers, Inc., 1991, p. 202.

ferenza osservata fra le due medie campionarie è molto probabile in una popolazione in cui $\mu_1 - \mu_2 = 0$. Passando al riquadro inferiore, in questo caso la differenza campionaria osservata appare sostanziale. Con ogni probabilità, dunque, rifiuteremo la nostra ipotesi nulla, in quanto l'ampia differenza osservata fra le due medie campionarie si verificherebbe solo raramente se $\mu_1 - \mu_2 = 0$. Piuttosto, è probabile che le due medie campionarie provengano da due popolazioni nelle quali $\mu_1 > \mu_2$.

Tornando al nostro esempio, l'orientamento politico medio degli uomini risulta pari a 4,14, mentre quello delle donne è uguale a 4,04. Le due varianze campionarie sono pari, rispettivamente, a 1,90 e 1,66. Pertanto, la statistica del test può essere calcolata come segue:

Page 15 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

$$Z_{(\bar{Y}_1-\bar{Y}_2)} = \frac{4,14-4,04}{\sqrt{\frac{1,90}{621}+\frac{1,66}{838}}} =$$

$$= \frac{0,100}{0,071} =$$

$$= 1,41$$

Poiché il valore di Z è inferiore al valore critico del test (2,58), non possiamo rifiutare l'ipotesi nulla a favore di quella alternativa secondo la quale uomini e donne sono caratterizzati da orientamenti politici differenti. Dunque, concluderemo che, probabilmente, uomini e donne hanno orientamenti politici molto simili.

3.2. La verifica delle ipotesi sulle proporzioni

Nel secondo capitolo abbiamo visto che la media di una variabile dicotomica è uguale alla proporzione di casi che assumono valore 1, cioè equivale a p_1. Per esempio, nella General Social Survey del 1991, 1.078 dei 1.414 soggetti intervistati hanno dichiarato di essere a favore della pena di morte nel caso di condanna per omicidio. In questo caso, $p_1 = 1.078/1.414 = 0,762$, cioè 76,2%.

Le stesse formule utilizzate nel paragrafo precedente per sottoporre a verifica le ipotesi sulle differenze fra le medie possono essere applicate alle variabili dipendenti dicotomiche per effettuare **test di significatività delle differenze fra proporzioni**. La varianza di una variabile dicotomica è uguale a pq, dove $q = 1 - p$. Pertanto, riprendendo l'esempio del giudizio sulla pena di morte, possiamo dire che in questo caso la varianza campionaria equivale a $s_Y^2 = (0,762)(1 - 0,762) = (0,762)(0,238) = 0,181$. A sua volta, l'errore standard di una distribuzione campionaria di proporzioni è uguale a:

$$s_p = \sqrt{pq/N}$$

In questo esempio,

$$s_p = \sqrt{(0,762)(0,238)/1.414} = 0,011.$$

Supponiamo di ipotizzare che i repubblicani, convenzionalmente considerati più conservatori, siano più a favore della pena di morte dei democratici. Formalmente:

$$H_0: p_R = p_D$$

$$H_1: p_R > p_D$$

dove: p_R è la proporzione di repubblicani che si dichiarano a favore della pena di morte, mentre p_D è l'equivalente proporzione di democratici. Nella General Social Survey del 1991, su 611 repubblicani che hanno espresso un'opinione l'86,1% hanno dichiarato di essere a favore della pena di morte, contro il 67% dei 622 democratici. Dunque, $p_R = 0{,}681$ e $p_D = 0{,}670$. Se poniamo $\alpha = 0{,}05$, la statistica del test della differenza fra le due proporzioni può essere calcolata come segue:

$$Z_{(p_R - p_D)} = \frac{0{,}861 - 0{,}670}{\sqrt{\frac{(0{,}861)(0{,}139)}{611} + \frac{(0{,}670)(0{,}330)}{622}}} = \frac{0{,}191}{0{,}023} = 8{,}3$$

Poiché il valore critico, in un test a una coda, è pari a 1,645, rifiutiamo l'ipotesi nulla e concludiamo che la propensione dei repubblicani a sostenere la pena di morte è maggiore di quella manifestata dai democratici.

La formula dell'errore standard utilizzata nel calcolo della statistica del test mostra chiaramente che all'aumentare di N_1 e N_2 l'errore standard diminuisce. Se N_1 e N_2 sono sufficientemente grandi, quasi *ogni* differenza fra $\overline{Y}_1$ e $\overline{Y}_2$ diventerà significativa. Per questa ragione, si dovrebbero adottare strategie di verifica delle ipotesi più prudenti. Ad esempio, ogni differenza fra medie inferiore a un quarto della deviazione standard dovrebbe essere considerata irrilevante, indipendentemente dalla sua significatività statistica. Un approccio ancora più preferibile consiste nello stimare la forza della relazione fra variabili. In ogni caso, i semplici test di significatività statistica non rivelano mai l'intera storia.

3.3. Il test *t*

Quando i dati consistono in due campioni di piccole dimensioni – cioè quando $N_1 + N_2 < 100$ – gli assunti che sottendono il test Z della differenza fra due medie sono violati. In questi casi si può ricorrere alla famiglia delle distribuzioni t, purché siano rispettati due assunti: 1) entrambi i campioni sono estratti casualmente da due popolazioni

Page 17 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

indipendenti e normalmente distribuite; 2) le varianze delle due popolazioni sono omoschedastiche, cioè $\sigma_1^2 = \sigma_2^2 = \sigma^2$. A differenza del test Z, in questo caso la forma delle distribuzioni delle due popolazioni è importante. In pratica, tuttavia, la violazione di questo assunto produrrà effetti limitati sui risultati del test.

Per ottenere un'unica stima della varianza della popolazione, le due varianze campionarie vengono aggregate, utilizzando la seguente formula:

$$s^2 = \frac{(N_1 - 1)(s_1^2) + (N_2 - 1)(s_2^2)}{N_1 + N_2 - 2}$$

dove:
$N_1 + N_2 - 2$ = gradi di libertà associati alla stima aggregata.

I gradi di libertà del test corrispondono alla somma dei gradi di libertà associati a s_1^2 e s_2^2. Precisamente:

$$gl_1 + gl_2 = (N_1 - 1) + (N_2 - 1) = N_1 + N_2 - 2.$$

Dunque, per piccoli campioni la statistica del test della differenza fra due medie è uguale a:

$$t_{(N_1+N_2-2)} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{s_{(\overline{Y}_1 - \overline{Y}_2)}} =$$

$$= \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s^2/N_1 + s^2/N_2}} =$$

$$= \frac{\overline{Y}_1 - \overline{Y}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

dove:
$s_{(\overline{Y}_1 - \overline{Y}_2)}$ = stima dell'errore standard della differenza fra le medie.

Per illustrare l'applicazione del test t prenderemo in esame i tassi di sindacalizzazione statunitensi e ci chiederemo se negli stati che possiedono una legislazione per il «diritto al lavoro» la partecipazione sindacale è più bassa che negli stati privi di tale legislazione. La tabella 4.3 riporta i dati rilevanti. Poiché i due gruppi di stati hanno una numerosità limitata (rispettivamente 21 e 30 casi), in questo caso bisogna

TAB. 4.3. *Tassi di sindacalizzazione della forza lavoro statunitense, secondo la presenza/assenza di una legislazione per il diritto al lavoro*

Stati con legislazione per il diritto al lavoro		Stati senza legislazione per il diritto al lavoro	
Stato	Tasso di sindacalizzazione	Stato	Tasso di sindacalizzazione
Iowa	20,5	New York	35,6
Alabama	18,2	Michigan	33,7
Alaska	18,2	Distretto di Columbia	33,4
Tennessee	17,3	Washington	32,9
Utah	16,8	Hawaii	31,5
Nevada	16,3	West Virginia	28,9
Nebraska	16,3	Illinois	27,5
Wyoming	15,6	Oregon	27,5
North Dakota	14,2	Ohio	27,4
Louisiana	13,8	Pennsylvania	27,0
Arkansas	13,2	Missouri	26,6
Arizona	12,8	California	25,4
Georgia	12,7	Indiana	25,1
Texas	12,5	Wisconsin	24,5
Kansas	12,0	Minnesota	24,5
Virginia	10,9	Montana	21,7
South Dakota	10,3	Kentucky	20,4
Florida	9,6	Delaware	20,3
Mississippi	9,3	New Jersey	19,9
North Carolina	8,9	Massachusetts	19,7
South Carolina	5,8	Rhode Island	19,4
		Connecticut	18,9
		Maryland	18,6
		Maine	18,5
		Colorado	18,0
		Idaho	16,1
		Oklahoma	12,9
		New Mexico	12,8
		New Hampshire	12,3
		Vermont	11,9
$N_1 = 21$		$N_2 = 30$	
$\overline{Y}_1 = 13{,}58\%$		$\overline{Y}_2 = 23{,}10\%$	
$s_1 = 3{,}70$		$s_2 = 6{,}73$	

ricorrere al test t, assumendo che i dati costituiscano campioni indipendente estratti casualmente da due distribuzioni normali.

L'ipotesi nulla e quella alternativa – dove il gruppo 1 comprende gli stati che possiedono la legislazione per il diritto al lavoro e il gruppo 2 comprende gli stati che ne sono privi – possono essere espresse come segue:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Ponendo $\alpha = 0{,}01$, la tavola B3 in appendice B mostra che, per

Page 19 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

21 + 30 – 2 = 49 gradi di libertà, il valore critico di un test a una coda è uguale a –2,407. Calcolando s^2 e estraendone la radice quadrata si ottiene:

$$s = \sqrt{\frac{(21-1)(3,70)^2 + (30-1)(6,73)^2}{21+30-2}} = 5,69$$

Quindi, t equivale a:

$$t_{49} = \frac{13,58 - 23,10}{(5,69)\sqrt{1/21 + 1/30}} = \frac{-9,52}{1,62} = -5,88$$

Come si può vedere, i risultati del test corroborano l'ipotesi secondo la quale gli stati che possiedono una legislazione per il diritto al lavoro sono caratterizzati da tassi di sindacalizzazione inferiori a quelli osservati negli altri stati.

Generalmente i ricercatori utilizzano il test t, anziché il test Z, per analizzare le differenze fra due medie. All'aumentare di N, tuttavia, le distribuzioni t e Z convergono verso i valori critici della seconda. Inoltre, all'aumentare di N (oltre i 100 casi) il teorema del limite centrale rende sempre meno rilevante l'assunto secondo il quale i campioni devono provenire da popolazioni normalmente distribuite. Infine, poiché i ricercatori non conoscono quasi mai la deviazione standard della popolazione, il test t è preferibile.

3.4. Confronto fra due distribuzioni mediante diagramma rami e foglie e «boxplot»

Nel paragrafo 8 del secondo capitolo abbiamo visto come si costruiscono i diagrammi rami e foglie e i *boxplot* per rappresentare graficamente singole distribuzioni. In questo paragrafo, invece, mostreremo come gli stessi strumenti possano essere utilizzati per confrontare due distribuzioni.

Impiegando i dati riportati nella tabella 4.3, la figura 4.2 illustra due diagrammi rami e foglie che rappresentano la distribuzione dei tassi percentuali di sindacalizzazione rispettivamente negli stati americani che possiedono una legislazione per il diritto al lavoro e negli stati che, invece, ne sono privi. La figura 4.3, invece, rappresenta le stesse distribuzioni mediante due *boxplot*. La mediana del primo gruppo di stati corrisponde all'11^{ima} osservazione, che assume valore 13,2; il cardine inferiore (H_1) corrisponde alla sesta osservazione, che ha valore pari a 10,9, mentre il cardine superiore (H_2) corrisponde alla 16^{ima} osservazione, con valore uguale a 16,3. Lo spazio fra i cardini, dun-

que, è uguale a $HS = 16{,}3 - 10{,}9 = 5{,}4$. Per verificare l'eventuale presenza di valori estremi dobbiamo calcolare il margine interno inferiore e il margine interno superiore:

$$MII = 10{,}9 - (1{,}5)(5{,}4) = 2{,}8$$

e

$$MIS = 16{,}3 + (1{,}5)(5{,}4) = 24{,}4$$

Poiché in questo primo gruppo di stati non esistono osservazioni con valori inferiori a 2,8 o superiori a 24,4, possiamo concludere che in esso non sono presenti valori estremi. Il *boxplot* per questi dati è disegnato nella parte sinistra del grafico riportato nella figura 4.3.

Poiché il gruppo degli stati privi di legislazione per il diritto al lavoro è formato da 30 casi, in questo caso la mediana della distribuzione si colloca fra la 15^{ima} e la 16^{ima} osservazione e, quindi, è uguale a $(21{,}7 + 24{,}5)/2 = 23{,}1$. I valori assunti dagli altri parametri sono i seguenti: $H_1 = 18{,}6$, $H_2 = 27{,}5$, $HS = 8{,}9$, $MII = 5{,}25$ e $MIS = 40{,}85$. Poiché nessuna osservazione assume valori inferiori a 5,25 o superiori a 40,85, possiamo concludere che anche in questo gruppo di stati non sono presenti valori estremi.

La figura 4.3 offre una conferma visiva del fatto che gli stati che possiedono una legislazione per il diritto al lavoro risultano caratterizzati da tassi di sindacalizzazione più bassi. Inoltre, la figura conferma che nel gruppo degli stati privi di legislazione per il diritto al lavoro la variazione dei tassi di sindacalizzazione è più ampia.

Rami	Foglie	(*N*)
Stati con legislazione per il diritto al lavoro		
0	699	(3)
1	00123333446667788	(17)
2	1	(1)
Stati senza legislazione per il diritto al lavoro		
0		
1	223368999	(9)
2	000025555777889	(15)
3	23346	(5)

FIG. 4.2. Diagrammi rami e foglie raffiguranti le distribuzioni dei tassi di sindacalizzazione della forza lavoro negli stati americani con legislazione per il diritto al lavoro e negli stati americani senza legislazione per il diritto al lavoro.

Page 21 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

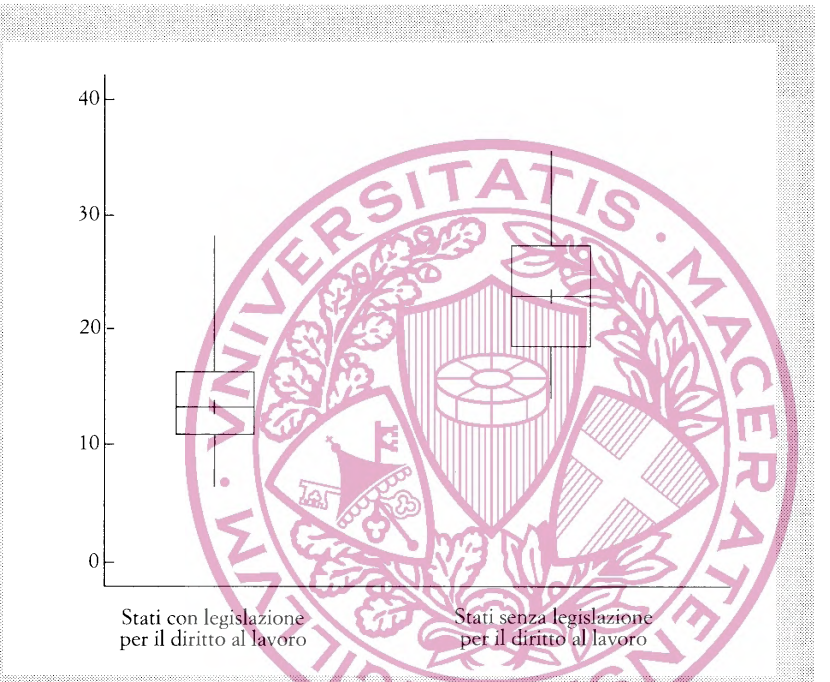

FIG. 4.3. *Boxplot* raffiguranti le distribuzioni dei tassi di sindacalizzazione della forza lavoro negli stati americani con legislazione per il diritto al lavoro e negli stati americani senza legislazione per il diritto al lavoro.

3.5. Intervalli di confidenza e stime puntuali

Seguendo la procedura illustrata nel paragrafo 6 del terzo capitolo, intorno alla stima puntuale della differenza fra due medie è possibile costruire un intervallo di confidenza. Per ogni dato livello di confidenza $1 - \alpha$, la formula dell'intervallo di confidenza è la seguente:

$$(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2} s_{(\overline{Y}_1 - \overline{Y}_2)}$$

Ad esempio, per calcolare un intervallo di confidenza del 95% intorno alla stima campionaria della differenza fra i due gruppi di stati discussi sopra in termini di tasso medio di sindacalizzazione, bisogna partire dalla stima puntuale di tale differenza, cioè $\overline{Y}_1 - \overline{Y}_2 = -9{,}52\%$. L'errore standard stimato è pari a $s_{(\overline{Y}_1 - \overline{Y}_2)} = 1{,}62$. La tavola B3 in appendice B ci mostra che per un intervallo di confidenza del 95% con $N_1 + N_2 - 2 = 49$ gradi di libertà, il valore critico $t_{\alpha/2}$ è circa pari a 2. Pertanto, il limite superiore dell'intervallo sarà uguale a – 9,52% + (2) (1,62) = –6,28%, mentre il limite inferiore sarà uguale a –9,52% – (2) (1,62) = –12,76%. In altri termini, possiamo essere «fiduciosi» al 95% che l'intervallo compreso fra –12,76% e –6,28% contiene la vera dif-

Page 22 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

ferenza fra i due gruppi di stati in termini di tasso medio di sindacalizzazione.

La formula per calcolare un intervallo di confidenza intorno alla differenza fra due proporzioni è la seguente:

$$(p_1 - p_2) \pm t_{\alpha/2} s_{(p_1 - p_2)}$$

Riprendendo l'esempio sull'atteggiamento degli americani nei confronti della pena di morte, la stima puntuale della differenza fra la proporzione dei Repubblicani e quella dei Democratici a favore è uguale a $p_R - p_D = 0{,}191$, con errore standard uguale a $s_{(p_R - p_D)} = 0{,}023$ e gradi di libertà pari a $611 + 622 - 2 = 1.231$. La tavola B3 in appendice B ci mostra che a un intervallo di confidenza del 99% corrisponde un valore critico pari a 2,58. Pertanto, il limite inferiore dell'intervallo sarà uguale a $0{,}191 - (2{,}58)(0{,}023) = 0{,}132$, mentre il limite superiore sarà uguale a $0{,}191 + (2{,}58)(0{,}023) = 0{,}250$. Possiamo allora essere «fiduciosi» al 99% che la vera differenza fra la proporzione di Repubblicani e quella di Democratici a favore della pena di morte è compresa fra 0,132 e 0,250. Indubbiamente, questa differenza è significativamente e sostanzialmente maggiore di zero.

3.6. La relazione fra *t* e *F*

Il test t e il test F per la verifica delle ipotesi sulle differenze fra le medie sono strettamente correlati. Quando il confronto riguarda solo due medie, cioè quando $J = 2$, tanto l'analisi della varianza quanto il test t producono lo stesso risultato. Infatti, quando applicati agli stessi dati, la radice quadrata di un quoziente F con 1 e ν_2 gradi di libertà equivale al valore t con ν_2 gradi di libertà. Formalmente:

$$t_{\nu_2} = \sqrt{F_{1,\nu_2}}$$

Generalmente, quando $J = 2$ vengono riportati solo i risultati del test t, mentre quando $J > 2$ si ricorre all'analisi della varianza.

4. Il quoziente di correlazione: eta-quadrato

Il test Z, il test t e l'analisi della varianza permettono al ricercatore di rifiutare, se i dati lo consentono, l'ipotesi nulla. Il passo successivo dell'analisi consiste nel chiedersi: qual è la forza della relazione fra le variabili? Come abbiamo già osservato, poiché un campione sufficientemente numeroso rende significativa quasi ogni differenza fra medie osservata, i risultati dei test di significatività non ci dicono nulla sulla

Page 23 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

forza – e, quindi, sulla rilevanza sostanziale – della relazione di interesse. Pertanto, dopo avere rifiutato un'ipotesi nulla è necessario stimare la forza della relazione fra le due variabili oggetto di analisi calcolando il **quoziente di correlazione** o **eta-quadrato** o η^2.

Nel paragrafo 2.1 abbiamo visto che:

$$SQ_{TOTALE} = SQ_{INTERGRUPPO} + SQ_{INTRAGRUPPO}$$

Dividendo entrambi i lati dell'equazione per la somma dei quadrati totale si ottiene:

$$1 = \frac{SQ_{INTERGRUPPO}}{SQ_{TOTALE}} + \frac{SQ_{INTRAGRUPPO}}{SQ_{TOTALE}} =$$

= «Somma dei quadrati spiegata» + «Somma dei quadrati non spiegata»

Il quoziente del rapporto fra la somma dei quadrati fra i gruppi e la somma dei quadrati totale rappresenta la proporzione di quest'ultima che è «spiegata» (in senso statistico) dalla variabile indipendente. Analogamente, il quoziente del rapporto fra la somma dei quadrati nei gruppi e la somma dei quadrati totale può essere visto come la proporzione di quest'ultima rimasta senza «spiegazione». La somma di queste due componenti è uguale a 1. Pertanto, la proporzione di varianza della variabile dipendente attribuibile alla variabile indipendente è definita come segue:

$$\eta^2 = \frac{SQ_{INTERGRUPPO}}{SQ_{TOTALE}}$$

Poiché il simbolo η^2 denota un parametro della popolazione, la sua stima campionaria è indicata con lo stesso simbolo sormontato da un accento circonflesso: $\hat{\eta}^2$. Eta-quadrato è sempre compreso fra 0 e 1. Maggiore è la differenza complessiva fra le medie e minori sono le varianze campionarie, maggiore è la somma dei quadrati fra i gruppi e, quindi, più vicino a 1 è il valore assunto da $\hat{\eta}^2$.

Utilizzando le statistiche ANOVA riportate nella tabella 4.2, si ha:

$$\hat{\eta}^2 = \frac{19,38}{80,13} = 0,242$$

In altri termini, il 24,2% della variazione osservata nei tassi statali di criminalità è «spiegato» dalla regione alla quale i vari stati appartengono. È difficile stabilire quanto grande deve essere il valore assunto da $\hat{\eta}^2$ affinché si possa parlare di una relazione «sostanziale». Tipica-

mente, nelle scienze sociali una singola variabile indipendente «spiega» non più del 25-30% della varianza di una variabile dipendente, e spesso il suo potere di «spiegazione» non supera il 2-5%.

5. La verifica delle ipotesi sulle differenze fra le medie («post hoc»)

Come abbiamo già osservato (cfr. par. 1), nell'ambito dell'analisi della varianza l'ipotesi alternativa può assumere diverse forme. Specificamente, si danno tre possibilità:

1. la media di ogni popolazione differisce dalla media di tutte le altre popolazioni;
2. le medie di alcuni sottoinsiemi di popolazioni differiscono l'una dalle altre;
3. alcune combinazioni di medie differiscono da una singola media o da un'altra combinazione di medie.

Il valore della statistica F, in sé, non offre alcuna informazione che possa aiutarci a stabilire quale di queste tre alternative è vera in una data situazione. Utilizzare una serie di test t per esaminare tutte le possibili coppie di medie è scorretto, in quanto non tutti i $J^2 - J$ confronti fra le medie dei J gruppi sono indipendenti fra loro. Dunque, bisogna fare ricorso a qualche metodo che sia guidato dalle idee del ricercatore sulle probabili differenze esistenti.

Per confrontare le medie sono disponibili due approcci di base. Il primo, chiamato **confronto pianificato** o *a priori*, sfortunatamente richiede un trattamento matematico più avanzato di quello che possiamo utilizzare in questo libro. Il secondo approccio, chiamato **confronto *post hoc*** o *a posteriori*, è statisticamente meno potente ma comunque utile nell'ambito delle scienze sociali. In questo paragrafo offriremo una breve presentazione di un particolare tipo di confronto post hoc, il **test di Scheffé**.

Per effettuare confronti multipli fra J medie è possibile creare un contrasto. Un **contrasto** fra le medie di J popolazioni, indicato con il simbolo Ψ (che rappresenta la lettera greca *psi*), è definito come segue:

$$\Psi = c_1\mu_1 + c_2\mu_2 + \ldots + c_J\mu_J$$

dove:
c_j = pesi con vincolo $c_1 + c_2 + \ldots + c_J = 0$.

Facendo riferimento ancora una volta all'esempio sui tassi di criminalità, possiamo ipotizzare che il Sud e l'Ovest esibiscano tassi maggiori di quelli osservati nel Nordest e nel Midwest. In base a questa ipotesi, dobbiamo formare contrasti che oppongano le prime due regioni alle altre due nel seguente modo: $c_S = c_O = -1/2$ e $c_N = c_M = 1/2$. Utilizzando pesi di uguale valore ma di segno opposto, si garantisce

Page 25 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

che la somma dei pesi sia uguale a zero. Il contrasto per questa specifica ipotesi, indicato con il simbolo Ψ_1, è formulato come segue:

$$\Psi_1 = (-1/2)\mu_S + (-1/2)\mu_O + (1/2)\mu_N + (1/2)\mu_M =$$

$$= \frac{-\mu_S - \mu_O}{2} + \frac{\mu_N + \mu_M}{2}$$

In sostanza, la media fra il Sud e l'Ovest è opposta alla media fra il Nordest e il Midwest. Sono possibili altri contrasti compatibili con la nostra ipotesi. Ad esempio, avremmo potuto fissare $c_S = 1$, $c_O = 1$, $c_N = -1$ e $c_M = -1$, rispettando il vincolo della somma zero. Una volta che abbiamo formato un contrasto utilizzando le medie campionarie, dobbiamo confrontare il suo valore con il corrispondente errore standard, esattamente come si fa in un test t. Se il quoziente del rapporto è sufficientemente grande, possiamo concludere che il confronto fra le medie è significativo. In caso contrario, non possiamo rifiutare l'ipotesi nulla secondo la quale la vera differenza fra le medie è nulla. Una stima corretta di un contrasto basato sui dati campionari è data da:

$$\hat{\Psi} = c_1\bar{Y}_1 + c_2\bar{Y}_2 + \ldots + c_J\bar{Y}_J$$

A sua volta, la stima della varianza di un contrasto è data da:

$$\hat{\sigma}^2_{\hat{\Psi}} = QM_{INTRAGRUPPO}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \ldots + \frac{c_J^2}{n_J}\right)$$

dove:
n_j = numero di osservazioni del gruppo j.

La statistica del test per un confronto *post hoc* corrisponde al quoziente del rapporto fra il valore assoluto di $\hat{\Psi}$ e il suo errore standard:

$$t = \frac{|\hat{\Psi}|}{\hat{\sigma}_{\hat{\Psi}}}$$

Il valore critico rispetto al quale bisogna valutare questa statistica del test è dato da:

$$\sqrt{(J-1)(F_{J-1,N-J})}$$

dove:
$F_{J-1,\,N-J}$ = valore critico corrispondente al livello α prescelto per verificare l'ipotesi nulla nell'analisi della varianza.

Dunque, ogni volta che $|\hat{\Psi}|/\hat{\sigma}_{\hat{\Psi}} \geq \sqrt{(J-1)(F_{J-1,N-J})}$ possiamo rifiutare l'ipotesi nulla. A questo test t sono associati $N-J$ gradi di libertà.

Per illustrare il modo in cui si effettua praticamente un confronto multiplo fra medie, torniamo all'esempio dei tassi di criminalità. Il contrasto precedentemente ipotizzato è il seguente:

$$\hat{\Psi}_1 = \frac{(4,36+4,21)}{2} + \frac{(-4,59+-5,80)}{2} = -0,91$$

con

$$\hat{\sigma}^2_{\hat{\Psi}_1} = (6,46)\left[\frac{(1/2)^2}{16} + \frac{(1/2)^2}{13} + \frac{(-1/2)^2}{9} + \frac{(-1/2)^2}{12}\right] = 0,539$$

e

$$\hat{\sigma}_{\hat{\Psi}_1} = \sqrt{0,539} = 0,73$$

Partendo da questi valori, la statistica t risulta uguale a $t = |-0,91|/0,73 = 1,25$. Il valore critico del test, per $\alpha = 0,01$, è uguale a $\sqrt{(4-1)(4,22)} = 3,56$. Dato che 3,56 è maggiore di 1,25, possiamo concludere che le differenze regionali ipotizzate devono essere rifiutate. Poiché, dall'analisi della varianza, sappiamo che comunque le quattro medie regionali non sono uguali, il passo successivo dell'analisi dovrebbe consistere nell'ipotizzare e sottoporre a verifica un nuovo tipo contrasto.

Problemi

1. Uno studioso del comportamento elettorale ipotizza che nel 1992 il sostegno alla coppia Clinton-Gore sia stato maggiore fra i democratici, seguiti dagli indipendenti e, quindi, dai repubblicani. Scrivi l'ipotesi nulla e quella alternativa in forma simbolica.

2. Utilizzando una scala di capitale culturale un sociologo della famiglia scopre che la media generale in un campione casuale di famiglie è pari a 26,8. Se il punteggio medio è uguale a 18,3 fra le famiglie che vivono nei centri urbani, a 22,6 fra le famiglie che vivono nei centri rurali e a 37,4 fra le famiglie che vivono nei centri suburbani, quali sono gli effetti α_j esercitati dai tre tipi di famiglia?

3. Uno psicologo sociale somministra un test di flessibilità intellettuale a un campione di 75 studenti universitari. Il punteggio medio del

Page 27 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013

test risulta pari a 8,65 fra i 30 studenti di scienze dell'educazione, a 11,42 fra i 20 studenti di scienze politiche e a 9,26 fra i 25 studenti di ingegneria. Quali sono gli effetti α_j esercitati dalle tre facoltà universitarie, utilizzando la media campionaria ponderata come stima della media della popolazione?

4. Trova i gradi di libertà e i valori critici di F per le seguenti combinazioni:

a) $\alpha = 0{,}05$, 2 gruppi, 30 osservazioni
b) $\alpha = 0{,}01$, 2 gruppi, 300 osservazioni
c) $\alpha = 0{,}01$, 6 gruppi, 60 osservazioni
d) $\alpha = 0{,}01$, 20 gruppi, 600 osservazioni.

5. Trova i gradi di libertà e i valori critici di F per le seguenti combinazioni:

a) $\alpha = 0{,}05$, $n_1 = 8$, $n_2 = 9$
b) $\alpha = 0{,}05$, $n_1 = 12$, $n_2 = 10$, $n_3 = 11$
c) $\alpha = 0{,}01$, $n_1 = 22$, $n_2 = 18$, $n_3 = 20$
d) $\alpha = 0{,}01$, $n_1 = 15$, $n_2 = 15$, $n_3 = 15$, $n_4 = 15$.

6. Uno sperimentatore divide 200 soggetti in tre gruppi, due dei quali hanno la stessa numerosità, mentre il terzo è grande quanto la somma dei primi due. Se la media dei tempi impiegati per risolvere una certa serie di problemi è uguale a 8 minuti nel primo gruppo e a 10 minuti nel secondo gruppo, qual è il punteggio medio del terzo gruppo, tenendo conto che la media generale è uguale a 11 minuti?

7. Un insieme di 45 soggetti, suddivisi in tre gruppi di uguale dimensione, dà luogo a una somma di quadrati totale pari a 284,25 e a una somma di quadrati nei gruppi uguale a 239,75. A quanto ammonta il quadrato medio fra i gruppi?

8. Un gruppo di sociologi del lavoro seleziona quattro campioni casuali di lavoratori da altrettante fabbriche di automobili, con dimensioni campionarie pari a 13, 15, 12 e 11. Misurando le intenzioni di mettersi in proprio manifestate da questi lavoratori utilizzando una scala a 10 posizioni, i ricercatori hanno ottenuto $SQ_{TOTALE} = 186{,}33$ e $SQ_{INTERGRUPPO} = 24{,}22$. Qual è il valore assunto da eta-quadrato? La probabilità che le medie reali delle quattro fabbriche siano significativamente differenti fra loro è minore di 0,05? Calcola il quoziente F e il valore critico appropriati.

9. Qui di seguito sono riportati i punteggi ottenuti in una prova di storia dai 15 studenti di una classe, precedentemente assegnati a tre diversi metodi di studio: 1) nessuna preparazione; 2) studio individuale; 3) studio in coppia con un compagno di classe:

Nessuna preparazione	Studio individuale	Studio in coppia
4	4	5
5	6	6
6	7	8
6	7	9
8	8	9

Calcola gli effetti α_j esercitati dai tre metodi di studio. Quindi calcola la somma dei quadrati totale, la somma dei quadrati fra i gruppi e la somma dei quadrati nei gruppi. Poi, determina i quadrati medi e il quoziente F, e valuta quest'ultimo rispetto all'ipotesi nulla secondo la quale le medie delle tre popolazioni sono uguali. Riporta questi risultati in una tavola ANOVA riassuntiva. Infine, calcola $\hat{\eta}^2$ e interpreta i risultati.

10. Utilizzando i dati del problema precedente e ponendo $\alpha = 0{,}05$, verifica l'ipotesi *post hoc* secondo la quale studiare in coppia con un compagno di classe da risultati superiori a quelli prodotti dagli altri due metodi. Riporta i valori di $\hat{\Psi}$ e $\hat{\sigma}^2_{\hat{\Psi}}$.

Page 29 of 29 Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013