Università degli Studi di Padova
DIPARTIMENTO DI SCIENZE STATISTICHE
Corso di Laurea Triennale in
Statistica e Gestione delle imprese

Relazione finale
L'algoritmo reticolare: una nuova tecnica per la diagnosi del tumore adrenocorticale

Relatore:
Prof.ssa Ventura Laura
Dipartimento di Scienze Statistiche

Laureanda:
Crepaldi Marica
Matricola: 1011195

Anno Accademico
2012/2013

Indice

Introduzione

In questo studio vengono analizzati dati relativi ad un campione di n=245 pazienti affetti da un raro tumore alle ghiandole surrenali. I dati provengono da 5 diversi centri, ossia Firenze, Milano, Padova, Treviso e Torino.

Mediante questi dati si vuole valutare l'accuratezza di una nuova tecnica per la diagnosi del tumore adrenocorticale: l'*algoritmo reticolare*. Questo nuovo metodo è importante per classificare il tumore al surrene come benigno o maligno. Il vantaggio di questa nuova tecnica è che risulta più veloce, economica e di facile interpretazione rispetto alle tecniche usate precedentemente.

Il dataset contiene i giudizi di otto patologi, con diversa esperienza in campo surrenale. Usando il giudizio di questi patologi si vuole valutare se effettivamente questa tecnica diagnostica è accurata, indipendentemente dall'esperienza dal patologo.

Questa relazione è divisa in tre capitoli.

Nel primo capitolo viene brevemente spiegato che cosa sono la ghiandole surrenali, che cos'è il tumore adrenocorticale e si descrivono i dati a disposizione. Nel secondo capitolo, dopo un'analisi esplorativa, si ipotizza un modello di regressione logistica per valutare quale relazione intercorre tra la variabile di interesse, che indica se il tumore è benigno o maligno, e la altre variabili presenti nel dataset. Nel terzo capitolo, infine, si verifica la concordanza tra i giudizi dei diversi patologi, attraverso degli indici specifici, ossia l'α di Cronbach, l'indice di correlazione tetracorico, il k di Cohen e il k di Fleiss.

Capitolo 1

Descrizione della patologia

In questo capitolo si descrive brevemente che cosa sono le ghiandole surrenali, che cos'è il carcinoma adrenocorticale, cos'è e in cosa consiste l'algoritmo reticolare. Vengono poi definiti alcuni termini chiave usati per spiegare il problema di questo studio. Infine, vengono presentate la variabili presenti nel dataset.

1.1 Le ghiandole surrenali

Le ghiandole surrenali, chiamate anche surreni, sono due ghiandole endocrine di colore brunogiallastro e di forma per lo più triangolare, situate sopra i reni (da ciò deriva il loro nome, sur-rene appunto) e che misurano all'incirca 5 cm in lunghezza e 2.5 cm in larghezza.

Ogni surrene è diviso in due parti distinte.

La parte centrale è chiamata midollare del surrene, e produce alcune sostanze chimiche che svolgono importanti funzioni nel sistema nervoso centrale, come l'adrenalina e la noradrenalina. Queste due sostanze, tra varie altre funzioni, stimolano l'attività del cuore, aumentano il tasso glicemico nel sangue facilitando così l'assorbimento degli zuccheri da parte dei tessuti e stimolano la coagulazione del sangue.

La parte esterna è detta corticale. Essa produce gli ormoni steroidi, fra cui l'aldosterone, che contribuisce a regolare la pressione arteriosa, il cortisolo, alcuni ormoni sessuali maschili e alcuni ormoni sessuali femminili.

La presenza di tali ghiandole è indispensabile per la vita, ma è possibile vivere con una ghiandola sola.

1.2 Il carcinoma adrenocorticale

Il carcinoma adrenocorticale (AAC) è un tumore alle ghiandole surrenali.

È più comune nei bambini di età inferiore ai 5 anni e negli adulti tra i 30 e i 50 anni.

I fattori di rischio non sono noti. Non è stata documentata alcuna relazione con il fumo e con la familiarità. È stata tuttavia rilevata un'associazione con alcune mutazioni genetiche.

Il tipo più comune di tumore alle ghiandole surrenali è un tumore benigno, detto *adenoma surrenalico*. Fra i tumori maligni il più frequente è la metastasi che origina da tumori situati in altri organi, e più raramente un tumore maligno insorge primitivamente nella ghiandola surrenalica.

Molti adenomi surrenalici vengono riscontrati casualmente durante l'esecuzione di ecografie, TAC o risonanze magnetiche per altri motivi, e spesso non è necessario asportarli, ma solo controllarli mediante la ripetizione di periodiche analisi. Se, invece, provocano sintomi a causa dell'alterata produzione ormonale, essi vanno asportati chirurgicamente.

La diagnosi certa sulla natura del tumore avviene mediante biopsia, cioè attraverso il prelievo di una piccola parte di tessuto, ed esame istologico del campione raccolto. Alcuni tumori corticali pongono notevoli difficoltà diagnostiche anche ad un patologo esperto nella categorizzazione in lesioni benigne e maligne.

Nel livello di diagnosi di carcinoma del cortico-surrene viene utilizzato come criterio di malignità quello proposto da Weiss (Weiss, 1984). Inolte, la morfologia, assieme alla colorazione istochimica, è la combinazione migliore nei casi di dubbia interpretazione tra tumore benigno o maligno.

1.3 L'algoritmo reticolare

La diagnosi patologica del carcinoma adrenocorticale è basata sul riconoscimento di molti parametri morfologici, che si combinano in un sistema di punteggio, che comprende fino a 12 criteri micro e macroscopici. Questa procedura è dispendiosa, di difficile riproduzione ed è poco usata anche da patologi specializzati in materia.

Attualmente, il sistema diagnostico per i tumori maggiormente usato è il sistema di Weiss. Tuttavia questa tecnica è scarsamente riproducibile nel tipo di tumore studiato in questa relazione. Pertanto, è stato proposto recentemente un diverso approccio: l'*algoritmo reticolare*. Questa tecnica definisce il tumore adrenocorticale maligno basandosi su un processo che avviene in due fasi: nella prima si analizza la struttura del reticolo, attraverso una colorazione a base d'argento; quindi, se è stata trovata una rottura, la malignità del tumore è ulteriormente definita attraverso l'identificazione di almeno tre parametri di malignità (necrosi, alto tasso micotico e invasione venosa). Questo algoritmo di classificazione è simile al sistema di Weiss, ma è più facile e veloce da applicare. La seconda parte di questo algoritmo, essendo simile al sistema proposto da Weiss, è già stata validata in uno studio di riproducibilità studiato da un gruppo di francesi (si veda

Duregon *et al.*, 2013 e i riferimenti qui citati). La parte che necessita di validazione è la prima, ossia quella riguardante la colorazione reticolare.

A tale scopo è stato progettato uno studio multicentrico, volto a valutare la riproducibilità dell'interpretazione della colorazione reticolare in 245 casi di tumore ai surreni raccolti in 5 centri. Questi 245 casi includono classici tumori adrenocorticali, una speciale variante di essi, oltre ad un consistente numero di tumori benigni.

Per definire lo stato della struttura reticolare, ogni istituzione ha eseguito una colorazione istochimica del reticolo, usando un kit di base disponibile in commercio (Bio Optica, Milano). Quindi i 245 vetrini di reticoli colorati sono stati rivisti da un patologo locale per verificare che sia stato selezionato, per ogni caso, un blocco rappresentativo e che questo sia stato riclassificato in accordo con la tecnica dell'algoritmo reticolare. Tutti i vetrini che differiscono dalla struttura normale della ghiandola surrenale sono stati registrati come "alterati".

Per valutare la riproducibilità dell'interpretazione della colorazione reticolare, i vetrini sono stati distribuiti tra 8 patologi, con differente esperienza nelle patologie surrenali.

Nella prima fase dello studio, è stato chiesto ai patologi, all'oscuro della diagnosi iniziale, di vedere e classificare separatamente i 245 vetrini, scegliendo tra due possibili opzioni (normale o alterato) e basandosi soltanto sulla descrizione del modello reticolare fornito dai loro precedenti studi.

Nella seconda fase, invece, tutti i casi discordanti della prima fase, sono stati rivalutati dopo una formazione specifica dei patologi.

Oltre allo studio delle concordanze nei giudizi, in questo studio si evidenzia che la colorazione reticolare è una tecnica più veloce, economica e di facile interpretazione dei metodi usati finora, dato che considera sia i cambiamenti quantitativi che qualitativi della struttura reticolare.

Questa tecnica ha un'alta riproducibilità, che giustifica un uso esteso dell'approccio in due fasi dell'algoritmo reticolare per la diagnosi del tumore alle ghiandole surrenali.

Di seguito vengono descritte alcune parole chiave usate in questo studio (da treccani.it - L'enciclopedia italiana).

endòcrino agg. – In fisiologia, si riferisce a ghiandola o a cellula che concorre a una secrezione interna; sistema e., l'insieme delle ghiandole endocrine (ipofisi, epifisi, tiroide, ecc.) che versano il loro prodotto di secrezione nei capillari sanguigni o linfatici, con le pareti dei quali si trovano a intimo contatto.

istochìmica s. f. [comp. di isto- e chimica]. – Ramo dell'istologia che ha per oggetto l'individuazione e la misura quantitativa dei costituenti chimici delle cellule e dei tessuti con metodi diversi: coloranti differen-

ziali, reazioni enzimatiche e immunologiche, uso di anticorpi marcati o fluorescenti, autoradiografia

necròsi s. f. In patologia, complesso di alterazioni strutturali irreversibili, dovute a cause di diversissima natura (fisiche, chimiche, microbiche, ecc.), che comportano la perdita di ogni vitalità, ossia la morte, di gruppi cellulari, zone di tessuto, porzioni di organo in un organismo vivente [...].

mitòtico agg. [der. di mitosi] (pl. m. -ci). – In biologia, che si riferisce alla mitosi [...].

mitosi Processo di divisione cellulare che costituisce il tipico modo di riproduzione cellulare negli organismi.

1.4 Il dataset

Il dataset contiene 245 casi di tumore adrenocorticale, raccolti dagli archivi dei reparti di patologia di cinque istituzioni.

Le variabili presenti nel dataset sono:

- Cod: è una variabile numerica che indica il codice del paziente all'interno del relativo ospedale.
- Proven: è una variabile che identifica la provenienza del vetrino che assume 5 livelli:
  - FI: vetrini raccolti dal 1993 al 2011 presso l'Università di Firenze (provenienti dall'Ospedale Careggi);
  - MI: vetrini raccolti tra il 1994 e il 2007 presso il Dipartimento dell'Ospedale Niguarda Ca' Granda di Milano;
  - PD: vetrini recuperati presso l'Università di Padova tra il 2000 e il 2008;
  - TO: vetrini raccolti tra il 2009 e il 2012 presso l'Università di Torino (Ospedale San Luigi);
  - TV: vetrini raccolti tra il 1998 e il 2012 presso l'Ospedale di Treviso.
- WS: è un punteggio ricavato col sistema di Weiss. Varia tra 0 e 9 ed indica la gravità della malattia. Un WS minore di tre indica che il tumore è da considerarsi adenoma, e quindi benigno, mentre un WS maggiore di 2 indica carcinoma, ossia tumore maligno.
- Stato: è una variabile dicotomica che vale 1 se WS è maggiore o uguale a 3, e vale 0 se WS è minore di tre. Indica, quindi, se il tumore è benigno o maligno.

- Sesso: è una variabile che indica il sesso del paziente. Assume due livelli: M nel caso di paziente di sesso maschile, F nel caso di paziente di sesso femminile.

- Eta: è una variabile che indica l'età del paziente (in anni).

- Med2-Med8: è una variabile dicotomica che vale 0 o 1 a seconda che il vetrino sia considerato, rispettivamente, normale o alterato. Vale quindi 0 se il tumore è classificato come benigno, e vale 1 se il tumore è classificato come maligno. Ogni colonna si riferisce ad un patologo con differente esperienza. I vetrini sono stati valutati da due medici interni all'ospedale, due medici giovani e quattro consulenti istopatologici.

- Somma: è una variabile che varia da 0 a 8 che indica la somma delle 8 colonne precedenti, ossia le colonne riferite al giudizio degli otto patologi. Il valore 0 indica che tutti i patologi hanno classificato il tumore come benigno, mentre il valore 8 indica che tutti i patologi hanno classificato il tumore come maligno. Questa variabile è da considerarsi una misura della gravità della malattia del paziente.

- Conc: è una variabile che varia da 4 a 8 che indica il numero di patologi concordi con lo stesso giudizio.

Nel prossimo capitolo si analizzeranno le variabili a disposizione allo scopo di verificare eventuali relazioni tra esse. Si cercherà poi di ipotizzare un modello di regressione logistica per la variabile dicotomica Stato.

Capitolo 2

Analisi dei dati

In questo capitolo si svolge una prima analisi delle variabili presenti del dataset. Inoltre, dopo aver definito alcune parole chiavi di questo studio, viene ipotizzato un adeguato modello per la variabile Stato.

Alcuni testi di riferimento per le tecniche usate in questo capitolo sono Piccolo (1998), Azzalini (2001) e Pace e Salvan (2010).

Il software utilizzato per le analisi è R (www.R-project.org).

2.1 Analisi univariata

Nel dataset sono presenti 245 pazienti.

Gli adenomi, ovvero i tumori benigni, sono 61 (24.9 %), mentre i restanti 184 (75.1 %) casi sono carcinomi, ovvero tumori maligni.

La distribuzione del punteggio di Weiss è riportata in Tabella 2.1 e in Figura 2.1.

WS	0	1	2	3	4	5	6	7	8	9	Media (sd)
Fr. ass.	36	19	6	10	12	21	39	43	38	21	5.07 (3.00)

Tabella 2.1: Frequenze assolute, media (e deviazione standard) del punteggio di Weiss.

Le donne sono 149 (60.8%), mentre gli uomini sono 96 (39.2%); il rapporto F/M risulta 1.55.

Nel dataset mancano 3 valori nella variabile Età. A questi tre valori mancanti si attribuisce il valore della media (ossia 50) e si ottengono le statistiche descrittive riportate nella Tabella 2.2 e il grafico in Figura 2.2.

La distribuzione della variabile Età risulta simmetrica. Nel dataset sono presenti 4 bambini di età inferiore ai 10 anni e un anziano di 97 anni.

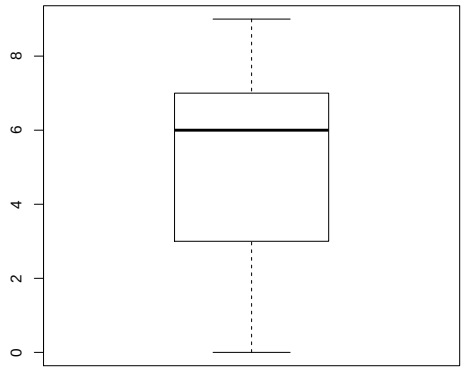

Figura 2.1: Boxplot relativo al punteggio di Weiss.

Min	1° Qt	M.na	Med (st. err)	3° Qt	Max
0	38.00	51.00	49.64 (16,44)	61.00	97.00

Tabella 2.2: Riassunto della variabile Età.

Il valore del test di normalità di Shapiro-Wilk per la variabile Età risulta pari a 0.99 che, con un *p-value* pari a 0.28, porta all'accettazione dell'ipotesi nulla di normalità a tutti i livelli di α usuali.

I vetrini analizzati sono stati raccolti in 5 centri, con le frequenze (assolute e relative) indicate nella Tabella 2.3.

Città	Firenze	Milano	Padova	Treviso	Torino
Freq. assolute	61	15	42	117	10
Freq. Relative	0.25	0.06	0.17	0.48	0.04

Tabella 2.3: Frequenze assolute e relative per la Provenienza.

La Tabella 2.4 riporta la distribuzione dei valori della variabile Concordanza. Si nota che la maggior parte dei valori (75%) sono posizionati nel valore 8, e solo il 5% dei dati è posizionato nel valore 5. La media della variabilie Concordanza risulta pari a 7.6 ($\pm$0.81).

La Tabella 2.5 e il boxplot in Figura 2.3 riportano la distribuzione della variabile Somma. La media di tale variabile risulta pari a 6.62 ($\pm$2.60).

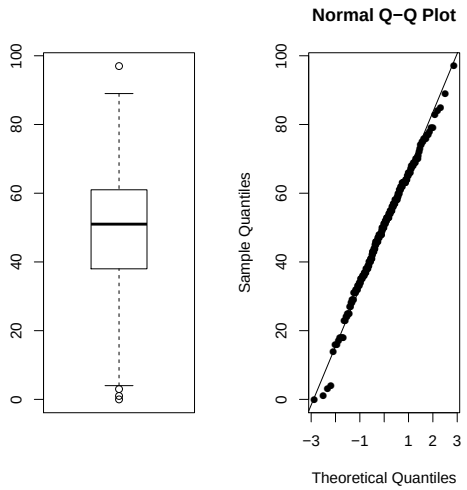

Figura 2.2: Boxplot e qqplot normale relativi alla variabile Età.

Medici concordi	5 su 8	6 su 8	7 su 8	8 su 8
Proporzione	0.05	0.061	0.135	0.755

Tabella 2.4: Distribuzione della variabile Concordanza.

Non c'è nessun caso dove i patologi si dividono esattamente a metà tra l'affermare che il tumore sia benigno o maligno, ossia per cui Conc=4 e Somma=4.

2.2 Analisi bivariata

Nella Figura 2.4 sono riportati i boxplot relativi alle variabili Età, Concordanza e Somma divisi nei due tipi di tumore (Stato).

Il t-test per verificare l'ipotesi nulla $H_0 : \mu_0 = \mu_1$, dove μ_0 è l'età media del gruppo affetto da tumore benigno e μ_1 è l'età media del gruppo affetto da tumore maligno, risulta pari a 3.13 (p-*value*=0.0022) che porta al rifiuto, a livello α=0.05, dell'ipotesi di uguaglianza tra le medie dei due gruppi. L'età media del gruppo di persone affette da tumore maligno è significativamente più alta dell'età media nell'altro gruppo.

Il test non parametrico di Mann-Whitney per verificare l'ipotesi nulla $H_0 : me_0 = me_1$, dove me_0 è la mediana della variabile Concordanza nel gruppo dei pazienti affetti da tumore benigno e me_1 è la mediana della variabile Concordanza nel gruppo dei pazienti affetti da tumore maligno, risulta pari a MW=2600 e, poichè p-*value* <0.001, si rifiuta l'ipotesi di uguaglianza delle due mediane a tutti i livelli di α usuali.

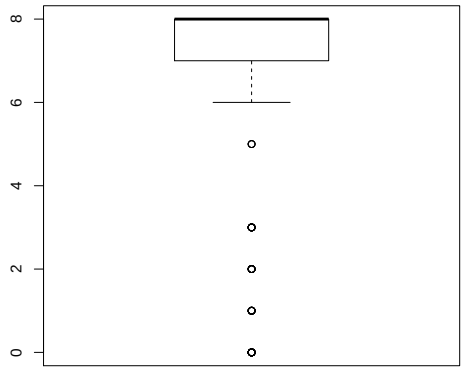

Figura 2.3: Boxplot relativo alla variabile Somma.

Somma	0	1	2	3	4	5	6	7	8
Fr.ass	17	9	8	9	0	3	7	23	169
Fr.Rel	0.07	0.04	0.03	0.04	0	0.01	0.03	0.09	0.69

Tabella 2.5: Distribuzione della variabile Somma.

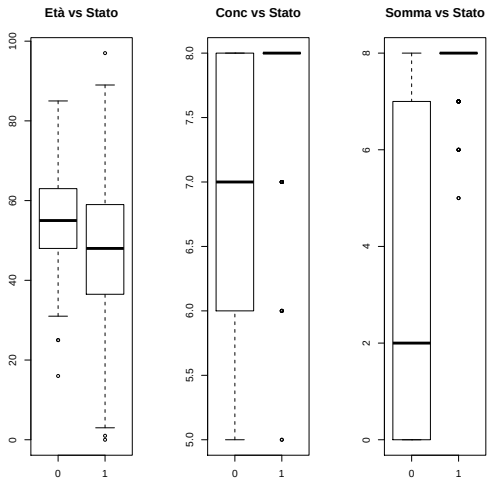

Figura 2.4: Boxplot relativi alle variabili Età, Concordanza e Somma divise per i due tipi di tumore.

Anche le due mediane relative ai due gruppi della variabile Somma risultano significativamente diverse a tutti i livelli di α usuali (MW=776, *p-value* <0.001)

Viene riportata in Tabella 2.6 la divisione per Sesso della variabile Stato.

	Sesso		
Stato	Femmina	Maschio	Tot
0	40	21	61
1	109	75	184
Tot	149	96	245

Tabella 2.6: Divisione della variabile Stato per Sesso del paziente.

Il test χ^2 di indipendenza risulta pari a 0.77 che, poichè *p-value*=0.38, porta all'accettazione dell'ipotesi di indipendenza tra le due variabili a tutti i livelli di α usuali. Non c'è quindi dipendenza tra il Sesso e lo Stato del paziente.

La correlazione tra la variabile Età e il punteggio di Weiss risulta pari a -0.12. La correlazione di Spearman tra il punteggio di Weiss e la variabile Concordanza risulta pari a 0.522, mentre la correlazione tra il punteggio di Weiss e la variabile Somma risulta pari a 0.688. Quest'ultime due correlazioni sono positive e risultano significative.

In Figura 2.5 viene riportato il boxplot della variabile WS divisa per il Sesso del paziente.

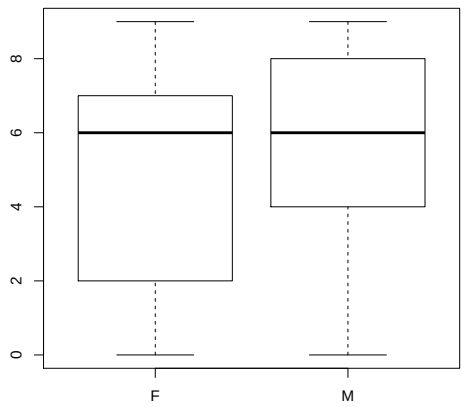

Figura 2.5: Boxplot relativo alle variabile WS divisa nei due gruppi della variabile Sesso.

Il test non parametrico di Mann-Whitney per verificare l'uguaglianza delle mediane dei punteggi di Weiss rispetto al Sesso del paziente, risulta pari a MW=6727 e, poichè *p-value*=0.4283, si accetta l'ipotesi nulla di uguaglianza delle mediane a tutti i livelli di α usuali.

2.3 Un modello per la variabile Stato

Nel dataset è presente la variabile dicotomica Stato, che fornisce informazioni sul tipo di tumore del paziente. È quindi interessante capire se le variabili presenti nel dataset sono in relazione con tale variabile.

Si ipotizza a tale scopo un modello di regressione logistica.

La variabile risposta è la variabile Stato, ossia una variabile dicotomica che vale 1 se il tumore è maligno e 0 se il tumore è di tipo benigno. Le variabili di cui è interessante studiare la relazione con Stato sono l'Età del paziente, il Sesso del paziente, la Concordanza e la Somma dei giudizi degli otto patologi.

Dato che la variabile risposta è una variabile dicotomica ed è di interesse modellare la probabilità che un paziente abbia un tumore di tipo maligno, il modello considerato è un modello di regressione logistica.

Siano

$$Y_i \sim Ber(\pi_i), \quad \pi_i \in [0,1],$$

con

$$E(Y_i) = \pi_i, \quad Var(Y_i) = \pi_i(1-\pi_i) \quad \text{per } i = 1, ..., n.$$

Il modello usato per dati dicotomici è il modello con legame *logit*, ossia

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

con x_{ij} variabili esplicative del modello e β_j parametri di regressione del modello, per $j = 1, ..., p$ e $i = 1, ..., n$.

Con i dati a disposizione si stima il seguente modello (riassunto in Tabella 2.7)

$$logit(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Eta + \hat{\beta}_2 Sesso + \hat{\beta}_3 Somma + \hat{\beta}_4 Conc.$$

La devianza residua del modello è 102.86 con 240 gradi di libertà. Le variabili Sesso e Conc risultano non significative.

Si eliminano, quindi, attraverso una procedura backward le variabili Sesso e Concordanza e si ottiene il seguente modello (riassunto in Tabella 2.8)

$$logit(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Eta + \hat{\beta}_2 Somma.$$

Coeff	Stima	St. Error	Statistica z	$Pr(>\|z\|)$
β_0	-7.44925	2.44172	-3.051	0.00228 **
β_1	-0.03415	0.01656	-2.062	0.03917 *
β_2	0.74038	0.60531	1.223	0.22128
β_3	1.19558	0.48624	2.459	0.01394 *
β_4	0.29013	0.62526	0.464	0.64263

Tabella 2.7: Modello iniziale per la variabile Stato.

Coeff	Stima	St. Error	Statistica z	Pr(> |z|)
β_0	-6.54110	2.06267	-3.171	0.00152 **
β_1	-0.02977	0.01604	-1.855	0.06355 .
β_2	1.36760	0.27605	4.954	< 0.001 ***

Tabella 2.8: Modello semplificato per la variabile Stato.

I parametri risultano significativi al livello α=0.10. La devianza residua risulta 104.57 con 242 gradi di libertà. Risulta quindi un buon modello.

Dato che sono due modelli annidati si può effettuare un test ANOVA per confrontarli. La differenza tra le due devianze residue risulta pari a 1.71 (p-value=0.425), che porta a concludere che il modello con meno parametri è preferibile.

La tabella di corretta classificazione che si ottiene con il secondo modello[1] è la Tabella 2.9, che porta ad una probabilità di corretta classificazione pari a 0.92. Risulta quindi un buon modello.

Stato	$\hat{\pi}_i \leq 0.5$	$\hat{\pi}_i > 0.5$
0	45	16
1	3	181

Tabella 2.9: Valori previsti dal modello $logit(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Eta + \hat{\beta}_2 Somma$ e valori osservati.

Si ottiene, quindi, il seguente modello

$$logit(\hat{\pi}_i) = -6.54 - 0.03 \times Eta_i + 1.37 \times Somma_i, \quad i = 1, ..., n,$$

$$\Leftrightarrow \hat{\pi}_i = \frac{e^{-6.54-0.03\times Eta_i+1.37\times Somma_i}}{1 + e^{-6.54-0.03\times Eta_i+1.37\times Somma_i}}.$$

[1] Assumendo che si preveda la presenza di tumore maligno se la probabilità stimata verifica $\hat{\pi}_i > 0{,}5$

Le stime dei parametri sono interpretabili come rapporto di quote. Ad esempio, si ottiene $OR_{\beta_1} = e^{\hat{\beta}_1} = e^{-0.03} = 0.97$. La probabilità di avere un tumore maligno diminuisce (anche se di poco dato che 0.97 è molto vicino ad 1) ad ogni aumento di un anno di età.

2.4 Conclusioni

In questo capito sono state analizzate le variabili presenti nel dataset.

Attraverso l'analisi bivariata si è notato che non c'è nessuna relazione tra lo Stato del paziente e il Sesso. La probabilità di tumore maligno è quindi la stessa nei due sottogruppi di pazienti.

Esiste invece una dipendenza con l'età del paziente. È quindi verificato che esiste una fascia d'età più a rischio di altre. In particolare, le personecon meno di 50 anni sono più soggette a tumore maligno.

C'è anche un'ovvia relazione tra la variabile Concordanza e lo Stato del paziente. Più i patologi sono concordi tra loro, più la probabilità di tumore maligno aumenta.

Dopo aver affrontato il problema della dipendenza tra le variabili, si è proposto un modello per la probabilità di riscontrare un tumore maligno.

Tale probabilità è stata modellata con un modello lineare generalizzato. Attraverso l'interpretazione dei parametri, si conclude che l'aumento di un anno d'età fa diminuire, anche se di poco, la probabilità di avere un tumore maligno. L'aumento, invece, di un giudizio di malignità da parte dei patologi fa aumentare di 4 volte la probabilità di avere un tumore maligno.

Nel prossimo capitolo si descriveranno alcune misure di concordanza usate per valutare l'accordo tra i giudizi degli otto patologi presenti nel dataset.

Capitolo 3

Misure di concordanza

In questo capitolo si analizzano i giudizi dei patologi mediante un diverso approccio rispetto ai metodi usati nei capitoli precedenti. Infatti, nel dataset sono presenti otto variabili dicotomiche, che rappresentano il giudizio dato da otto patologi tra loro indipendenti ai 245 casi di tumore adrenocorticale. In questo capitolo viene calcolata la concordanza usando degli indici generalmente usati in ambito psicologico. In particolare, ci si concentra sugli indici di concordanza per variabili dicotomiche. Alcuni riferimeti ai metodi sono Cronbach (1951), Cohen (1960), Fabbris (1996) e Quattro (2004).

3.1 Attendibilità dei giudizi

Quando le misurazioni sulle unità statistiche derivano dalla valutazione di due o più osservatori, occorre verificare che l'accordo fra questi osservatori nel determinare il punteggio o la categoria di appartenenza sia il più alto possibile.

Nello studio generale della concordanza, durante la codifica si possono verificare due tipi di errori: casuali o sistematici. Gli *errori casuali* sono dovuti ai problemi pratici incontrati durante la realizzazione della ricerca e in particolare nella fase di codifica (fatica, livello di attenzione, fretta, stress ...). Si può però immaginare che, se la codifica potesse essere eseguita infinite volte, gli errori casuali tenderebbero a compensarsi reciprocamente. Il secondo tipo di errore, l'*errore sistematico*, si verifica quando, per qualsiasi ragione, un osservatore sistematicamente attribuisce un determinato evento ad una categoria diversa da quella in cui esso rientra. Supponiamo che a tutti gli osservatori venga fornito un manuale relativo ad un sistema di codifica in cui la definizione di un determinato comportamento è sbagliata: tutti gli osservatori, adeguandosi alla definizione erronea, forniranno una codifica sbagliata nella stessa direzione, ovvero viziata dall'errore sistematico.

La *validità* di un sistema di codifica rappresenta il grado in cui esso misura realmente ciò che si propone di misurare, mentre la sua *attendibilità*

corrisponde al grado di accordo fra codifiche effettuate indipendentemente dall'osservatore. In altre parole, l'attendibilità si riferisce alla coerenza interna al sistema di codifica, mentre la validità si riferisce alla capacità del sistema di codifica di riflettere realmente il processo.

Altri due concetti diversi, sebbene collegati tra loro sono l'accordo e attendibilità. L'accordo si riferisce al grado in cui due osservatori concordano tra loro. Questo tipo di accordo non previene le molteplici fonti d'errore che possono alterare la ricerca. L'attendibilità, invece, è un concetto più generale ed intende idealmente far fronte a tutte le possibili fonti di errore. In generale, l'attendibilità è definita come il grado in cui i dati sono esenti da errori di misura: minore è l'errore, maggiore è la coerenza dei dati.

Facendo riferimento all'osservatore come fonte di errore, si possono distinguere tre tipi di attendibilità:

- *attendibilità intra-osservatore*: un osservatore può non essere attendibile rispetto a se stesso;
- *attendibilità inter-osservatore*: un osservatore può non essere attendibile rispetto ad un altro osservatore;
- *attendibilità dell'osservatore*: un osservatore può non essere attendibile rispetto ad un osservatore ideale, che si assume abbia codificato perfettamente.

3.2 Attendibilità intra-osservatore e relativi indici

L'*attendibilità intra-osservatore* corrisponde al grado con cui un osservatore, che giudica lo stesso fenomeno in condizioni identiche in momenti diversi, produce gli stessi risultati di codifica, realizzando così un buon livello di consistenza interna.

Siccome questo tipo di approccio implica che il medesimo osservatore codifichi ripetutamente gli stessi dati, la valutazione dell'attendibilità intra-osservatore può essere viziata da problemi legati a stanchezza o noia.

Per calcolare questa attendibilità si ricorre a due osservatori diversi che però vengono considerati come forme parallele di un singolo osservatore.

In questo caso il coefficiente da utilizzare per valutare l'attendibilità intra-osservatore tra due forme parallele, cioè tra due osservatori che codificano in momenti diversi, è l'indice ρ *di Pearson* o, per dati dicotomici, il *coefficiente di correlazione tetracorico* (si veda il Paragrafo 3.2.1) e il *coefficiente* α *di Cronbach* (si veda Paragrafo 3.2.2). Questi coefficienti esprimono la proporzione di varianza vera, ossia la varianza dovuta ai soggetti che vengono osservati, rispetto alla varianza totale, e forniscono un indice di quanto i dati sono liberi dall'errore casuale, senza tener conto dell'errore sistematico dovuto all'osservatore.

3.2.1 Analisi di una tabella tetracorica

Si considerino due variabili X e Y osservate su n unità statistiche. Le frequenze congiunte si dispongono in una tabella di frequenze 2×2, detta tetracorica, dove a, b, c, d sono le frequenze dell'osservazione congiunta di x_i e y_j (i,j=1,2).

		Y		
		1	0	Tot
X	1	a	b	a+b
	0	c	d	c+d
	Tot	a+c	b+d	n

Tabella 3.1: Tabella tetracorica 2×2.

Il coefficiente di correlazione tetracorico ρ è il coefficiente di correlazione dato da (Fabbris, 1996)

$$\rho = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{3.1}$$

Questo coefficiente varia tra -1 e 1 e raggiunge i valori estremi quando la dipendenza tra X e Y è massima. Si ha quindi che ρ=1 se b e c sono nulli, ρ=-1 se a e d sono nulli, mentre ρ=0 se concordanze e discordanze si bilanciano, e dunque se X e Y sono indipendenti.

Dato che il test χ^2 di indipendenza per una tabella 2×2 assume l'espressione

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

si ha che

$$\rho = \sqrt{\frac{\chi^2}{n}}.$$

Per verificare la significatività dell'indice ρ si può quindi ricorrere allo studio della significatività del test χ^2 di indipendenza.

3.2.2 Il coefficiente α di Cronbach

Un indice molto usato per il calcolo della consistenza interna è l'α *di Cronbach* (Cronbach, 1951) che corrisponde alla misura dell'affidabilità basata sulla coerenza delle risposte ai singoli item[1] del test e rappresenta quindi un indice di omogeneità degli item.

[1] È la singola unità di cui è costituito un test. In psicologia, i problemi, le domande, i compiti sottoposti agli individui vengono genericamente chiamati in questo modo.

Siano $x_1, ..., x_n$ i soggetti sottoposti al test, $i_1, ..., i_K$ gli item del test e p_{ij} il punteggio relativo alla risposta j del soggetto i, $j = 1...K$, $i = 1...n$. Siano s_j^2 la varianza dei singoli item e s_{tot}^2 la varianza della somma dei punteggi. Si ottiene la Tabella 3.2.

	Item					
Sogg.	i_1	$\cdots$	i_j	$\cdots$	i_K	Somma
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1K}	$\sum_{k=1}^{K} p_{1j}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{iK}	$\sum_{k=1}^{K} p_{ij}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_n	p_{n1}	$\cdots$	p_{nj}	$\cdots$	p_{nK}	$\sum_{k=1}^{K} p_{nj}$
	s_1^2	$\cdots$	s_j^2	$\cdots$	s_K^2	

Tabella 3.2: Tabella per il calcolo dell' α di Cronbach.

Quando le risposte ai test prevedono più di due alternative, la formula da usare è

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} s_j^2}{s_{tot}^2}\right). \tag{3.2}$$

Questo indice varia da 0 a 1 ed esprime il rapporto tra la somma delle varianze degli item e la varianza totale della scala, ossia la varianza dei punteggi sommati. Per ottenere una buona consistenza interna, e quindi un α elevato, è necessario che la varianza relativa ai singoli item sia piuttosto bassa in relazione alla varianza della scala.

In caso di risposte dicotomiche si ha come riferimento la Tabella 3.3, dove d_{ij} è la risposta j relativa al soggetto i che può assumere solo il valore 0 o 1. Si ha che p_j rappresenta la proporzione con la quale viene scelta l'alternativa codificata con 1 e $q_j = 1 - p_j$, $j = 1, ..., K$. Si ottiene quindi la formula KR-20 (Kuder-Richardson Formula 20), ossia (Kuder e Richardson, 1937)

$$\alpha_{KR-20} = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} p_j q_j}{s_{tot}^2}\right).$$

L'α di Cronbach e il KR-20 vengono interpretati come riportato in Tabella 3.4.

Lo svantaggio di questo indice è che che dipende da due fattori:

- la lunghezza della scala (numero degli item). Infatti, a parità di altre condizioni, all'aumentare del numero degli item, aumenta il valore dell'indice;

	Item					
Sogg.	i_1	$\cdots$	i_j	$\cdots$	i_K	Somma
x_1	d_{11}	$\cdots$	d_{1j}	$\cdots$	d_{1K}	$\sum_{k=1}^{K} d_{1k}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	d_{i1}	$\cdots$	d_{ij}	$\cdots$	d_{iK}	$\sum_{k=1}^{K} d_{ik}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_n	d_{n1}	$\cdots$	d_{nj}	$\cdots$	d_{nK}	$\sum_{k=1}^{K} d_{nk}$
	$\frac{\sum_{i=1}^{n} d_{i1}}{n} = p_1$	$\cdots$	$\frac{\sum_{i=1}^{n} d_{ij}}{n} = p_k$	$\cdots$	$\frac{\sum_{i=1}^{n} d_{i1}}{n} = p_K$	

Tabella 3.3: Tabella per il calcolo del KR-20.

valore α	Livello di accordo
< 0.60	problematico
0.60-0.70	appena sufficiente
0.70-0.80	discreto
0.80-0.90	buono
>0.90	ottimo/eccellente

Tabella 3.4: Interpretazione dell'α di Cronbach.

- la correlazione tra gli item. Infatti, maggiore è la correlazione tra gli item, maggiore sarà l'indice α.

Questo indice risente anche della troppa omogeneità dei dati. Se i giudici sono sempre concordi, l'indice risulterà non calcolabile.

Un intervallo di confidenza per α viene ottenuto con metodi boostrap (Li Chan e Cui, 2011).

3.3 Attendibilità tra osservatori e relativi indici

L'*attendibilità inter-osservatore* corrisponde al grado in cui due osservatori producono risultati di codifica simili quando osservano lo stesso fenomeno.

Essa può essere interpretata come il grado in cui i due osservatori possono essere considerati intercambiabili e indica quanto i dati sono liberi da errore casuale e sistematico legato alla codifica eseguita dagli osservatori.

Tuttavia, non è in grado di distinguere i due tipi di errore.

Di fronte a dei dati riportati in una matrice di confusione come quella in Tabella 3.5, la soluzione più diffusa e semplice per calcolare l'attendibilità inter-osservatore è il ricorso alla *percentuale di accordo*. La percentuale di

	Osservatore 1					
Oss. 2	i_1	$\cdots$	i_k	$\cdots$	i_K	Somma
i_1	f_{11}	$\cdots$	f_{1k}	$\cdots$	f_{1K}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i_k	f_{k1}	$\cdots$	f_{kk}	$\cdots$	f_{kK}	$f_{k\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i_K	f_{K1}	$\cdots$	f_{Kk}	$\cdots$	f_{KK}	$f_{K\cdot}$
	$f_{\cdot 1}$	$\cdots$	$f_{\cdot k}$	$\cdots$	$f_{\cdot K}$	n

Tabella 3.5: Matrice di confusione.

accordo (o indice di concordanza) è data da

$$\frac{Accordi}{Accordi + Disaccordi} \times 100 = \frac{\sum_k f_{kk}}{n} \times 100. \tag{3.3}$$

Sebbene questo indice abbia il vantaggio di essere intuitivo e facile da calcolare, ha due difetti che non possono essere eliminati.

Il primo è che la percentuale di accordo risulta gonfiata, rispetto al vero accordo, in quanto non viene corretta per il cosiddetto *accordo dovuto al caso*. Infatti, se si assegna a due osservatori indipendenti il compito di generare a caso una sequenza di codici appartenenti allo stesso sistema di codifica, le loro codifiche mostrano lo stesso un certo livello di accordo, quello dovuto al caso.

Il secondo difetto della percentuale di accordo è che essa dipende dalla frequenza del comportamento osservato, ossia dalla distribuzioni marginali della matrice di confusione. Dato che la grandezza della percentuale di accordo può essere aumentata indebitamente dall'accordo dovuto al caso, che, a sua volta, dipende dalla distribuzione marginale dei comportamenti, non ha senso fornire una soglia della percentuale di accordo sopra la quale si può dire che l'indice è accettabile, né possono essere paragonate percentuali di accordo provenienti da studi diversi, che hanno ragionevolmente una diversa probabilità marginale (Nussbeck, 2005). In più, dato che il valore di accordo osservato, posto sia al numeratore sia al denominatore nella formula per il calcolo della percentuale di accordo, contiene in sé l'errore dovuto al caso, il numeratore non fornisce un indice di varianza vera, né il denominatore un indice di varianza totale. Di conseguenza, poiché un indice tradizionale di attendibilità si ottiene a partire dal rapporto tra varianza vera e varianza totale, la percentuale di accordo non può essere considerata ad alcun titolo un indice di attendibilità.

Per calcolare un intervallo di confidenza si può usare la seguene formula

$$\hat{\rho} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\rho}(1 - \hat{\rho})}{n}}, \tag{3.4}$$

dove $\hat{\rho}$ è la stima della percentuale di accordo e $z_{1-\frac{\alpha}{2}}$ è il quantile di livello $1-\frac{\alpha}{2}$ di una normale standard.

Un altro possibile indice è il *k di Cohen* (Cohen, 1960).

3.3.1 Il *k* di Cohen

Il k di Cohen è un indice per il calcolo dell'accordo tra gli osservatori che ha il notevole vantaggio di correggere l'indice di accordo per l'accordo dovuto al caso. La tabella a riferimento per il calcolo del k di Cohen è la Tabella 3.5.

Per il calcolo del k di Cohen si devono calcolare le frequenze attese come nel calcolo del test χ^2 di indipendenza, ossia

$$\hat{f}_{ij} = \frac{f_{\cdot j} \times f_{i\cdot}}{n},$$

e si ottiene la Tabella 3.6.

	Osservatore 1					
Oss. 2	1	$\cdots$	j	$\cdots$	J	
1	$\hat{f}_{11}$	$\cdots$	$\hat{f}_{1j}$	$\cdots$	$\hat{f}_{1J}$	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	$\hat{f}_{j1}$	$\cdots$	$\hat{f}_{jj}$	$\cdots$	$\hat{f}_{jJ}$	$f_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
J	$\hat{f}_{J1}$	$\cdots$	$\hat{f}_{Jj}$	$\cdots$	$\hat{f}_{JJ}$	$f_{J\cdot}$
	$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot J}$	n

Tabella 3.6: Tabella delle frequenze attese.

L'indice k di Cohen è dato da

$$k = \frac{F_{oss} - F_{att}}{n - F_{att}}, \tag{3.5}$$

dove

- n è il numero totale dei casi;
- F_{oss} è il numero di accordi osservati dato da $F_{oss} = \sum_{j=1}^{J} f_{jj}$;
- F_{att} è il numero di accordi attesi dato da $F_{att} = \sum_{j=1}^{J} \hat{f}_{jj}$.

Questo indice varia da -1 a 1. Il valore nullo indica che gli osservatori vanno d'accordo come due persone che assegnano una codifica casuale; quando l'indice è negativo, i due osservatori sono sistematicamente in disaccordo; quando l'indice è positivo, i due osservatori vanno d'accordo, indipendentemente dall'accordo dovuto al caso.

Diversi autori hanno proposto differenti interpretazione del k di Cohen. In Tabella 3.7 viene riportata l'interpretazione più usata di Landis e Koch (Landis e Koch, 1977).

valore k di Cohen	Livello di accordo
<0.20	Accordo pessimo
0.20-0.40	Accordo modesto
0.40-0.60	Accordo moderato
0.60-0.80	Accordo buono
0.80-1	Accordo ottimo

Tabella 3.7: Interpretazione k di Cohen.

Nel caso di grandi campioni ($n \geq 100$), per calcolare un intervallo di confidenza per k, è possibile il ricorso alla distribuzione normale standardizzata

$$k \pm z_{1-\frac{\alpha}{2}} \sigma_k,$$

dove σ_k può essere calcolato come

$$\sigma_k = \sqrt{\frac{F_{oss}(n - F_{oss})}{n(n - F_{att})^2}} = \frac{\sqrt{F_{oss}\left(1 - \frac{F_{oss}}{n}\right)}}{n - F_{att}}$$

Per calcolare un intervallo di confidenza per l'indice k quando n è piccolo si usano i metodi boostrap (Efron B. e Tibshirani R.J., 1993).

Per il test di significatività $H_0 : k = 0$ contro $H_1 : k > 0$, per $n \geq 100$ si può utilizzare la statistica test

$$z = \frac{k}{\sqrt{\sigma_{k0}^2}},$$

con

$$\sigma_{k0} = \sqrt{\frac{F_{att}}{n(1 - F_{att})}}.$$

Anche questo indice, come la percentuale di accordo, è influenzato dalle distribuzioni marginali nella tabella di confusione. Ad esempio, la Tabella 3.8 riporta due situazioni in cui la proporzione di accordo fra i giudici è altissima (0.90), ma la diversa distribuzione della frequenze marginali produce due valori diversi di k. Vengono riportati tra parentesi i valori delle frequenze attese.

Il k di Cohen riferito alla Tabella 3.8(a) è

$$k_{(a)} = \frac{F_{oss} - F_{att}}{n - F_{att}} = \frac{90 - 82}{100 - 82} = 0.44,$$

(a)

	Osservatore A		
Osserv. B	Incluso	Escluso	Tot
Incluso	85(81)	5(9)	90
Escluso	5(9)	5(1)	10
Tot.	90	10	100

(b)

	Osservatore A		
Osserv. B	Incluso	Escluso	Tot
Incluso	45(25)	5(25)	50
Escluso	5(25)	45(25)	50
Tot.	50	50	100

Tabella 3.8: Esempio di tabelle per il calcolo del k di Cohen.

mentre il k di Cohen riferito alla Tabella 3.8(b) è

$$k_{(a)} = \frac{F_{oss} - F_{att}}{n - F_{att}} = \frac{90 - 50}{100 - 82} = 0.8.$$

Questo esempio mostra come il k è maggiore quando gli accordi sono equamente distribuiti sulla diagonale della tabella di contingenza.

3.3.2 Il *k* di Fleiss

L'indice k di Fleiss (Fleiss, 1971) viene utilizzato quando gli esaminatori sono più di due. In questo caso infatti la tabella di contingenza sarà a più entrate e il k di Cohen non è più calcolabile.

Al fine di valutare l'accordo tra le classificazioni espresse da più esaminatori, si considerano n soggetti, ciascuno dei quali viene classificato mediante K categorie esaustive e mutuamente esclusive da un gruppo di M ($M > 2$) esaminatori, i quali possono non essere gli stessi per ogni soggetto.

Indicato con x_{ij} il numero di esaminatori che hanno assegnato l'i-esimo soggetto ($i = 1, ..., n$) alla k-esima categoria (k=1,...,K), le assegnazioni possono essere rappresentate come nella Tabella 3.9.

	Categorie					
Soggetti	1	$\cdots$	k	$\cdots$	K	Tot
1	x_{11}	$\cdots$	x_{1k}	$\cdots$	x_{1K}	$x_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	x_{i1}	$\cdots$	x_{ik}	$\cdots$	x_{iK}	$x_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
n	x_{n1}	$\cdots$	x_{nk}	$\cdots$	x_{nK}	$x_{n\cdot}$
Tot	$x_{\cdot 1}$	$\cdots$	$x_{\cdot k}$	$\cdots$	$x_{\cdot K}$	

Tabella 3.9: Generica tabella per il calcolo del k di Fleiss.

Definita la proporzione di coppie di esaminatori che hanno assegnato il soggetto i alla categoria k

$$P_{ik} = \frac{x_{ik}(x_{ik} - 1)}{M(M-1)}$$

è possibile calcolare la proporzione delle coppie di assegnazioni concordanti relative al soggetto i, data da

$$P_i = \sum_{k=1}^{K} P_{ik} = \frac{1}{M(M-1)} \sum_{k=1}^{K} x_{ik}^2 - 1,$$

e misurare l'accordo osservato tramite la media

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i = \frac{1}{(M-1)} \left(\frac{1}{Mn} \sum_{i,k} x_{ik}^2 - 1 \right).$$

Sia

$$p_k = \frac{x_{\cdot k}}{Mn} = \frac{1}{Mn} \sum_{i=1}^{n} x_{ik}$$

una stima della probabilità di assegnazione casuale alla categoria k, allora l'accordo atteso per effetto del caso è dato da (Scott, 1955, Fleiss, 1971)

$$\bar{P}_e = \sum_{k=1}^{K} p_k^2.$$

Sottraendo dall'accordo osservato l'accordo atteso casuale e normalizzando, si ottiene la statistica

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{3.6}$$

proposta da Fleiss come generalizzazione dell'indice k di Cohen. È opportuno sottolineare che la statistica k di Fleiss rappresenta l'estensione dell'indice π di Scott[2] al caso in cui gli esaminatori sono più di due e costituisce uno degli strumenti più usati per valutare l'accordo tra M esaminatori.

I valori della statistica k di Fleiss sono compresi tra -1 e 1. Il valore -1 indica il massimo disaccordo, il valore 0 indica che l'accordo osservato è uguale all'accordo atteso per il caso ed il valore 1 indica il massimo accordo. Per l'interpretazione dell'indice k in funzione del grado di accordo secondo Landis e Koch (Landis e Koch, 1977) si utilizza la stessa tabella usata per l'interpretazione del k di Cohen (Tabella 3.7).

[2]Indice statistico per misurare l'attendibilità inter-osservatore su scala nominale, simile al k di Cohen (Scott, 1955)

3.4 Attendibilità dell'osservatore

Assumiamo che un ricercatore predisponga un manuale di codifica, definito protocollo standard, che rappresenti il prodotto della codifica eseguita da un osservatore ideale e infallibile. Questa versione della codifica preparata da esperti, che si presume accurata, viene detta *gold standard* (Bakemn e Quera, 2011). Assumiamo che questo flusso di codifica sia perciò considerabile "vero" e che venga confrontato con il prodotto della codifica di uno o più osservatori.

Tramite questa procedura il ricercatore può:

1. controllare che il codificatore esegua correttamente la codifica;
2. calibrare i codificatori;
3. ottenere una codifica che riflette il contenuto di ciò che è suo interesse codificare.

Ne consegue che l'attendibilità dell'osservatore corrisponde al grado con cui l'osservatore concorda con quanto stabilito da un protocollo standard assunto come vero. Questa procedura permette di eliminare qualsiasi tipo di errore, purché il protocollo standard sia formulato correttamente.

Tuttavia, un semplice metodo per il calcolo dell'attendibilità dell'osservatore consiste nel riportare la codifica del protocollo standard nelle colonne corrispondenti all'osservatore in questione. L'indice di attendibilità risultante dice se l'osservatore testato è accurato e se aderisce alle definizioni delle categorie comportamentali riportate nel manuale di codifica.

In questo lavoro questo tipo di attendibilità non verrà approfondita ulteriormente in quanto non si dispone della codifica *gold standard*.

3.5 L'attendibilità con i dati a disposizione

Per quanto riguarda il calcolo del coefficiente di correlazione tetracorico (Formula 3.1), i risultati sono riportati nella Tabella 3.10.

Dalla Tabella 3.10 si può notare che tutte le correlazioni tetracoriche sono positive e molte di esse si avvicinano ad uno. Si possono notare infatti molti valori maggiori di 0.95. La media dei coefficienti risulta pari a 0.93. La correlazione interna risulta quindi molto buona.

Nella Tabella 3.11 sono riportati i valori utili per il calcolo del coefficiente α di Cronbach (Formula 3.2). Si ha

$$\alpha_{cr} = \frac{8}{7}\left(1 - \frac{0.148 + 0.148 + ... + 0.114}{6.778}\right) = 0.953,$$

	Med1	Med2	Med3	Med4	Med5	Med6	Med7	Med8
Med1	1							
Med2	0.99	1						
Med3	0.99	0.99	1					
Med4	0.99	0.99	0.99	1				
Med5	0.94	0.94	0.95	0.95	1			
Med6	0.87	0.87	0.88	0.87	0.91	1		
Med7	0.92	0.92	0.91	0.91	0.764	0.81	1	
Med8	0.92	0.92	0.93	0.92	0.87	0.93	0.89	1

Tabella 3.10: Coefficiente tetracorico tra le coppie di patologi.

	Med1	Med2	Med3	Med4	Med5	Med6	Med7	Med8	Tot
Media	0.820	0.820	0.824	0.816	0.894	0.857	0.718	0.869	6.620
S.E.	0.148	0.148	0.145	0.151	0.095	0.123	0.203	0.114	6.778

Tabella 3.11: Media e Standard Error delle risposte dei vari patologi.

che risulta molto alto e vicino alla perfetta concordanza. Questo significa che il sistema di codifica proposto ai vari patologi è chiaro e di facile interpretazione, anche per i patologi meno esperti.

La proporzione di accordo (Formula 3.3), calcolata grazie alla libreria `irr` del software *R* risulta pari a 0.759 ($\pm$0.027). Nel 75,9% dei casi i patologi sono concordi nella stessa opionione.

A scopo illustrativo vengono riportati in Tabella 3.12 gli indici k di Cohen (Formula 3.5) calcolati tra le coppie di patologi. Grazie alla Tabella 3.12 si può notare anche l'accordo tra i patologi più esperti e i patologi più giovani.

	Med1	Med2	Med3	Med4	Med5	Med6	Med7	Med8
Med1	1							
Med2	0.97	1						
Med3	0.99	0.99	1					
Med4	0.99	0.96	0.97	1				
Med5	0.67	0.67	0.68	0.66	1			
Med6	0.62	0.62	0.63	0.61	0.68	1		
Med7	0.65	0.65	0.64	0.64	0.37	0.48	1	
Med8	0.69	0.69	0.70	0.68	0.61	0.74	0.51	1

Tabella 3.12: k di Cohen calcolato tra i vari patologi.

Tutti i *p-values* associati all'ipotesi nulla di non concordanza sono minori

di 0.001 e, quindi, c'è una concordanza significativa tra tutte le coppie di patologi a tutti i livelli di α usuale.

La statistica k di Fleiss (Formula 3.6) con i dati a disposizione relativi al giudizio di 8 patologi su 245 soggetti risulta pari a 0.702. Il test z associato è pari a 58.21 che porta al rifiuto, a tutti i i livelli di α usuali, dell'ipotesi nulla di assenza di concordanza tra i patologi.

In Tabella 3.13 vengono riportati gli indici di concordanza interna e tra osservatori per quanto riguarda la stratificazione territoriale, la divisione tra maschi e femmine, la stratificazione per classi di età e la divisione tra tumore benigno e maligno. Vengono riportati tra parentesi gli intervalli di confidenza di livello $1-\alpha$=0.05 per quanto riguarda l'α di- Cronbach (metodi bootstrap) e la proporzione di accordo (Formula 3.4).

		Concordanza interna		Conc. tra osservatori	
	n	Media coef. tetracorico	α di Cronbach	Prop. di accordo	k di Fleiss
Firenze	61	0.848	0.933 (0.893, 0.959)	0.607 (0.484, 0.730)	0.6100
Padova	42	0.956	0.965 (0.939, 0.982)	0.714 (0.577, 0.851)	0.7530
Torino	117	0.959	0.956 (0.937, 0.979)	0.829 (0.761, 0.897)	0.7130
Altro[3]	25	0.999	0.847	0.88 (0.753,1)	0.3870
Maschi	96	0.904	0.964 (0.944, 0.981)	0.802 (0.722,0.882)	0.759
Femmine	149	0.960	0.945 (0.922, 0.962)	0.732 (0.661,0.803)	0.666
0-30 anni	26	0.887	0.928 (0.195, 0.982)	0.731 (0.561, 0.901)	0.598
30-60 anni	154	0.921	0.947 (0.925, 0.964)	0.735 (0.665,0.805)	0.675
60-100 anni	65	0.971	0.966 (0.947, 0.982)	0.785 (0.685, 0.885)	0.765
T. benigno	61	0.841	0.918 (0.883, 0.942)	0.377 (0.255,0.499)	0.509
T. maligno	184	0.115	0.477 (0.253, 0.597)	0.886 (0.840, 0.932)	0.0915

Tabella 3.13: Indici di concordanza stratificati.

Dalla Tabella 3.13 risulta che la concordanza interna è ottima in tutte la stratificazioni dei dati. L'unica situazione in cui la concordanza risulta bassa è la stratificazione riguardante il tumore maligno. Questo è dovuto all'alta omogeneità dei dati. In questo caso è più corretto usare le tecniche proposte nel Capitolo 2.

Per quanto riguarda la proporzione di accordo si nota che in tutti i casi essa è maggiore di 0.50. Questo significa che in nessun caso i dati possono essere considerati come dei dati assegnati casualmente tra le varie modalità.

Gli indici k di Fleiss sono tutti positivi. I *p-values* associati ai coefficienti k sono tutti minori di 0.001, anche per quanto riguarda il coefficiente legato al tumore maligno. Questo significa che si rifiuta, per ogni stratificazione e a tutti i livelli di α usuali, l'ipotesi di non concordanza tra gli esaminatori.

3.6 Conclusioni

In questo capitolo si è proposto una metodologia per verificare se l'algoritmo reticolare, usato per capire la malignità o meno del tumore alle ghiandole surrenali, può essere utilizzato da patologi con differenti esperienze.

Per far questo si sono usati degli indici che di solito vengono usati in campo psicologico, ossia gli indici di attendibilità o concordanza.

Esistono tre tipi di attendibilità:

- l'attendibilità interna, che misura quanto un osservatore è attendibile. Misura quindi se davanti allo stesso fenomeno l'osservatore produce la stessa codifica, ovvero se da lo stesso risultato ottenuto in precedenza;
- l'attendibilità tra gli osservatori, ovvero quanto gli osservatori concordano nei vari casi ai quali vengono sottoposti;
- l'attendibilità rispetto ad un osservatore standard, che misura quanto un osservatore sia in grado di rispettare una codifica data per vera.

Per quanto riguarda il primo tipo di attendibilità sono stati proposti due indici: il *coefficiente di correlazione tetracorico* e l'α *di Cronbach*.

Nei dati a disposizione entrambi gli indici si possono considerare molto buoni o addirittura eccellenti. Questo significa che ogni patologo è coerente con se stesso e quindi codifica allo stesso modo casi simili.

Questo dato è importante perché i vari patologi, che avevano esperienze diverse, erano chiamati a dare un giudizio; l'aver notato una buona concordanza interna significa che anche i patologi più giovani riescono a dare un giudizio corretto, in accordo con i patologi più esperti.

[3]La statificazione Altro è riferita ai dati di Treviso e Milano, ossia alle due città con meno dati e con omogeneità più alta.

Per quanto riguarda il secondo tipo di attendibilità sono stati considerati tre tipi di indici: la *proporzione di accordo*, il *k di Cohen* (da usare quando i giudici sono due) e il *k di Fleiss* (da usare con più di due giudici).

Anche in questo caso i dati hanno dimostrato una buona o addirittura ottima concordanza. Questo significa che anche i patologi meno esperti sono d'accordo con i patologi più esperti nella maggioranza dei casi. Questo a testimonianza del fatto che questo nuovo metodo è di facile interpretazione ed è facile da usare.

Conclusioni

All'inizio di questa relazione viene spiegato che cos'è l'algoritmo reticolare. È una nuova tecnica usata per diagnosticare il tumore adrenocorticale.

Nel dataset sono presenti 245 casi di tumore adrernocorticale. Dopo l'analisi esplorativa, si è verificato innanzitutto quali variabili influenzano questo tipo di tumore. Si è scoperto che il sesso del paziente non è una variabile che influenza il tumore al surrene, mentre esiste una fascia d'età più a rischio di altre. Le persone con meno di 50 anni, e quindi anche i bambini, sono più a rischio.

Nell'ultimo capitolo si è spiegato cos'è e come si misura la concordanza. I dati a disposizione dimostrano una buona concordanza interna ai patologi e una buona concordanza tra i patologi. Questo significa che tutti i patologi hanno ben interpretato il sistema di codifica, che risulta quindi facilmente interpretabile. Questa risulta quindi una buona tecnica perchè è facilmente interpretabile ed è più veloce ed economica dei sistemi usati precedentemente.

Bibliografia

- Azzalini A. (2001). *Inferenza statistica. Una presentazione basata sul concetto di verosimiglianza.* Springer, Milano.

- Bakeman R., Quera V. (2011). Sequential Analysis and Observational Methods for the Behavioral Sciences. *Cambridge University Press, New York.*

- Chiorri C. (2011). *Teoria e tecnica psiconometrica. Costruire un test psicologico.* Mc-Graw-hill, Milano.

- Cohen J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 213-220.

- Cronbach L.J. (1951). Coefficient α and the internal structure of tests. *Psychometrika*, **16**, 297-333.

- Duregon E. *et al.* (2013). The reticulin algorithm for adrenocortical tumors diagnosis: a multicentric validation study on 245 unpublished cases. *American Journal of Surgical Pathology*, to appear.

- Efron B. e Tibshirani R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, 178-201.

- Fabbris L. (1996). *STATREE 1.0: sistema esperto per la scelta del metodo di analisi statistica.* Edizioni Summa, Padova.

- Fleiss J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378-382.

- Kuder G. F., Richardson M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, **2**, 151-160.

- Landis J. R., Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

- Li Chan W., Cui Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology*, **64**, 367-387.

- Nussbeck F.W. (2005). *Assessing multimethod association with categorical variables*. Handbook of Multimethod Assessment in Psychology, 231-247.

- Pace L., Salvan A. (2010). *Introduzione alla Statistica. Inferenza, verosimiglianza, modelli*. Cedam, Padova.

- Piccolo D. (2010). *Statistica*. Il Mulino, Milano.

- Quattro P. (2004). Un test di concordanza tra più esaminatori. *Statistica, anno LXIV*, **1**.

- Weiss L.M. (1984). Comparative histologic study of 43 metastasizing and nonmetastasizing adrenocortical tumors. *American Journal of Surgical Pathology*, **8**, 163-169.

Siti consultati

- www.corriere.it/salute/sportello_cancro/tiroide-surrene/index.shtml.

- www.medicitalia.it

- www.R−project.org.

- www.treccani.it