

---

*Parte ottava*

***Analisi Statistica sui determinanti della qualità proteica  
del grano duro in Sicilia 2000-2005: alcuni risultati preliminari***

V. Muggeo<sup>1</sup>, V. Brasile<sup>1</sup>, B. Messina<sup>2</sup>, G. Russo<sup>2</sup>

---

<sup>1</sup>Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”  
Università degli Studi di Palermo

<sup>2</sup>Consorzio di Ricerca “Gian Pietro Ballatore” - Assoro (EN)



## 1. Introduzione

Il Consorzio di Ricerca “*Gian Pietro Ballatore*” dal 2000 porta avanti un progetto, oggi denominato “Qualità e Tracciabilità del Grano Duro in Sicilia” che rientra in un più ampio progetto a base nazionale “Stoccaggio Differenziato del Frumento Duro”; quest’ultimo, avviato nel 1998 dal Ministero delle Politiche Agricole, assegna all’Istituto Sperimentale per la Cerealicoltura di Roma il compito di coordinare le attività di monitoraggio di tutte le regioni italiane coinvolte nel progetto.

Il progetto “Qualità e Tracciabilità del Grano Duro in Sicilia” ha previsto la realizzazione di una rete di monitoraggio del grano duro, che coinvolge diversi centri di stoccaggio dislocati in diversi punti della Sicilia, e prevede l’impiego di strumenti ad alta tecnologia (Infratec mod. 1241 della FOSS). Durante la campagna granaria (da fine maggio agli inizi di agosto), tale strumentazione Infratec analizza il grano duro conferito dai granicoltori, determinandone rapidamente cinque caratteri legati alla qualità merceologica: contenuto percentuale di proteine (% s.s.), contenuto percentuale di glutine (% s.s.), contenuto di umidità (%), peso specifico (Kg/hl) e indice di giallo (coordinata b).

In questo modo è possibile conoscere in tempi rapidi la qualità del grano prodotto dal singolo granicoltore in modo tale da disporre di elementi sia per un’adeguata formulazione del prezzo, sia per un immagazzinamento del prodotto in partite omogenee in funzione della qualità. Risulta infatti un requisito sempre più richiesto dall’industria di trasformazione (molini e pastifici) l’offerta di grosse partite di grano omogenee per determinate caratteristiche merceologiche.

La procedura di controllo della qualità merceologica del grano duro, che è inserita in un sistema qualità certificato secondo la norma UNI EN ISO 9001:2000 (Cartabellotta *et alii* 2002), prevede diverse fasi, il cui espletamento avviene tramite tecnici addestrati in tal senso. All’arrivo del grano presso il centro di stoccaggio vengono registrati i dati in ingresso, costituiti da informazioni rilasciate dal conferitore stesso. Le informazioni riguardano i riferimenti del granicoltore, la zona di provenienza del prodotto, la varietà, le unità di azoto apportato (concimazione in presemina e in copertura), il precedente colturale e la resa ottenuta.

Segue la pesata ed il campionamento rappresentativo (eseguito tramite specifiche sonde di campionamento) la cui procedura è standardizzata in accordo alla norma UNI 10243:1994 (campionamento delle granaglie).

Infine si procede con l’analisi del campione per mezzo dell’Infratec; i risultati così ottenuti, vengono registrati e afferiscono ad un data base centralizzato gestito dal Consorzio Ballatore; dopo un controllo sulla presenza di valori aberranti, i dati vengono analizzati, elaborati e divulgati.

Il numero dei centri che aderiscono al progetto “Qualità e Tracciabilità del Grano Duro in Sicilia” è cresciuto nel corso dei vari anni passando da 9 del 2000 a 27 del 2002, fino ad arrivare a 34 nella campagna granaria 2005-2006 (Russo *et alii*, 2007)

I dati sui quali è stata impostata l’indagine statistica, di seguito descritta, sono relativi al quinquennio 2000-2005 (totali rilievi 42.801); tali dati, disaggregati per anno e registrati in maniera differente da un anno all’altro, hanno subito un corposo processo di omogeneizzazione e validazione, prima di essere aggregati in unico data base, contenente 31.759 rilevazioni omogenee dell’intero periodo in esame.

L’obiettivo principale dell’indagine è vedere quali sono state le aree e le varietà che hanno determinato le migliori performance produttive in merito al contenuto proteico (uno dei caratteri qualitativi più rilevanti per le aziende di trasformazione), nonché quello di individuare quali sono stati i fattori che maggiormente hanno influenzano lo stesso.

## 2. Materiali e Metodi

### *I dati utilizzati*

Le analisi merceologiche sono effettuate su un campione di grano di circa 600g, rappresentativo della partita di granella conferita al centro di stoccaggio. Si intende per singola partita la quantità di grano contenuta in un singolo cassone di carico.

Le analisi effettuate dall'Infratec sono relative ai valori dell'*umidità*, delle *proteine*, del *glutine*, del *peso ettolitrico* e dell'*indice di giallo*.

I dati sulla *varietà*, *provenienza*, *concimazione in presemina e copertura*, *precedenti colturali* e *resa* vengono raccolti tramite interviste dirette ai conferitori di grano, effettuate dal personale addestrato che effettua le analisi.

### *Descrizione delle Variabili del dataset*

- *Varietà*: indica la varietà di grano analizzato. Per varietà (coltivata) si intende un “insieme” di piante coltivate che si colloca ad un livello tassonomico inferiore rispetto a quello rappresentato dalla “specie” e che si distingue chiaramente sulla base di caratteri morfologici, fisiologici, chimici o altro. Tale insieme quando viene coltivato mantiene i suoi caratteri distintivi e risponde alle condizioni di identità, omogeneità, stabilità (genotipica e fenotipica), valore agronomico e di utilizzazione. Le varietà previste dal data base sono: Appio, Appulo, Arcangelo, Arcobaleno, Bronte, Cannizzo, Canyon, Ciccio, Colorado, Claudio, Colosseo, Conca D'oro, Creso, Daunia, Duilio, Durbel, Flavio, Gianni, Gemme, Grazia, Ionio, Iride, Italo, Meridiano, Neodur, Mongibello, Norba, Ofanto, Perseo, Platani, Plinio, Preco, Rusticano, S. Carlo, Simeto, Svevo, Tresor, Valbelice, Varano, Vesuvio, Vitromax. Le varietà non incluse nel suddetto elenco sono registrate nel dataset con la dicitura “altre”.
- *Provincia*: indica la provincia nella quale è presente il centro di stoccaggio. Le province interessate sono: Agrigento, Caltanissetta, Catania, Enna, Palermo, Ragusa, Siracusa, Trapani.
- *Concimazione in presemina*: indica le unità di azoto apportate al terreno prima della semina ed è espressa in kg/ha di azoto. Non sempre viene effettuata. I valori di questa variabile sono raggruppati nelle seguenti classi: fino a 25, da 26 a 50, da 51 a 75, da 76 a 100, oltre 100, sconosciuto, biologico.
- *Concimazione in copertura*: indica le unità di azoto apportate al terreno dopo la semina ed è espressa in kg/ha di azoto. Non sempre viene effettuata. I valori di questa variabile sono raggruppati nelle seguenti classi: fino a 25, da 26 a 50, da 51 a 75, da 76 a 100, oltre 100, sconosciuto, biologico.
- *Precedente colturale*: indica la coltura che ha preceduto quella del grano. I precedenti colturali previsti dal data base sono: maggese, leguminose, oleaginose, cereali a paglia, barbabietole, mais, ortaggi.
- *Peso ettolitrico (Kg/hl)*: indica il peso specifico apparente del prodotto. È un carattere importante per la qualità merceologica del grano e il suo valore oscilla normalmente tra 76 e 82-83 kg/hl. Buone granigioni hanno peso specifico 81-82.
- *Umidità (%)*: indica la percentuale di acqua contenuta nella cariosside. Un basso contenuto in umidità è importante per la conservabilità del prodotto. Per una buona conservabilità il contenuto in umidità non dovrebbe superare il 12-13%; valori superiori a tale percentuale possono determinare deprezzamento e in taluni casi anche rifiuto del prodotto.
- *Proteine (% s.s.)*: indica il contenuto in proteine calcolato in percentuale sulla sostanza secca. Il contenuto proteico è la qualità merceologica del grano maggiormente apprezzata dalle attività di trasformazione. Generalmente esso rappresenta un carattere importante nella formazione del prezzo. In Sicilia tale contenuto oscilla prevalentemente tra il 9% ed il 15%.
- *Contenuto in glutine (% s.s.)*: indica il contenuto in glutine calcolato in percentuale sulla sostanza secca. Il glutine è un complesso proteico che si forma durante l'impasto della semola con acqua. Un buon contenuto in glutine risulta fondamentale nella trasformazione della semola in pasta e pane.
- *Indice di giallo*: indica l'intensità di colorazione gialla delle cariossidi e delle semole (*coordinata b –*

*metodica Sgrulletta e De Stefanis, 1996*). È correlato alla quantità di pigmenti carotenoidi presenti nella cariosside. L'industria di trasformazione attualmente orienta le sue scelte in funzione del maggiore colore giallo delle semole al fine di ottenere pasta intensamente colorata. Questo carattere non ha alcuna influenza sulle caratteristiche tecnologiche degli impasti, ma solo sull'aspetto finale dei prodotti.

### **Metodi Statistici**

Tecniche di analisi descrittiva, quali tabelle di *frequenze, medie e deviazioni standard* nonché *diagrammi a barre e boxplot* (Iacus, 2005), sono state utilizzate per avere informazioni sintetiche sulle distribuzioni delle variabili osservate.

Per quantificare il ruolo dei diversi fattori nella determinazione del contributo proteico è stato costruito un *modello di regressione multiplo*. Tale strumento viene frequentemente utilizzato per l'analisi di dipendenza, ovvero quando tra le variabili osservate è possibile individuare le 'variabili in entrata (variabili *input* o *fattori*) e quelle in uscita (*output, outcome* o semplicemente *risposta*). In tal caso è di principale interesse valutare e quantificare l'effetto di tali fattori sulla *risposta*. Tralasciando i dettagli che possono trovarsi in un qualsiasi testo di Statistica Applicata (qui citiamo Weisberg, 2005), la logica è quella di supporre che possa esistere una relazione tra la risposta ed uno o più fattori. Nel nostro caso noi assumiamo:

$$\text{PROTEINE} = \text{VARIETÀ} + \text{AZOTO IN COPERTURA} + \text{AZOTO IN PRESEMINA} + \text{PROVINCIA} + \text{PRECEDENTE COLTURALE} + E$$

La scrittura di sopra esprime sinteticamente un modello di regressione: il membro di sinistra rappresenta la *variabile risposta*, mentre nel membro di destra sono inclusi i fattori che possono influenzare la risposta stessa: tali fattori sono tipicamente detti *covariabili* o *variabili esplicative*. In altri termini noi postuliamo che il contenuto proteico possa dipendere sia dai valori delle variabili esplicative inserite nel modello sia naturalmente da una componente stocastica indicata convenzionalmente con *E* che rappresenta la variabilità fisiologica del fenomeno in studio.

Se indichiamo con  $X_1, X_2, \dots, X_p$  le variabili esplicative del modello (cioè la varietà, provincia, ecc.), un modo più rigoroso per esprimere un modello di regressione utilizzato nel presente lavoro è il seguente:

$$\text{media (PROTEINE)} = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p \quad (1)$$

in cui ciascun *parametro*  $b_1, \dots, b_p$  quantifica l'apporto delle covariabili  $X_1, \dots, X_p$  nell'aumentare (se il coefficiente  $b$  è positivo) o diminuire (se  $b$  è negativo) la media generale delle proteine indicata con  $b_0$ . Si osservi che quando nel modello vengono inserite variabili *categoriali*, ovvero variabili qualitative con due o più modalità (ad esempio la variabile provincia è una variabile categoriale che nel caso della regione Sicilia ha nove modalità), nella costruzione del modello è necessario assumere una categoria di riferimento e valutare l'effetto delle altre modalità come differenza rispetto a quella di riferimento.

In generale i 'valori numerici' dei coefficienti  $b$  che si ottengono attraverso opportuni metodi computazionali, sono delle *stime campionarie*, perché relative soltanto a un campione, e non all'intera popolazione, di osservazioni. Questo significa che ogni  $b$  sarà una *stima* e quindi inevitabilmente affetto da errore. Per decidere quanta fiducia porre nelle stime ottenute nel campione, è necessario sempre accompagnare i risultati da una misura dell'*incertezza*, il cosiddetto *errore standard*. Quanto più piccolo è l'errore standard, tanto più attendibile è la stima ottenuta; più specificatamente, quando il rapporto (in valore assoluto) " $|stima/errore\ standard| < 2$ " allora la rispettiva variabile risulterà *non significativamente* associata con la risposta. Di contro, variabili per cui " $|stima/errore\ standard| > 2$ " potranno essere considerate *significativamente associate* con la risposta; in tal caso più alto sarà tale rapporto, maggiore è l'evidenza di una associazione con la risposta. Infine si osservi che unitamente alla stima dei coefficienti di

regressione  $b$ , il modello comprenderà anche una stima della varianza residua (tipicamente indicata con  $\sigma^2$ ) che misura la variabilità fisiologica del fenomeno.

L'analisi statistica dei dati, presentata e discussa in seguito, verrà utilizzata per quantificare l'influenza di ciascuna variabile esplicativa e allo stesso tempo per individuare quali covariabili risultano essere predittive del contenuto proteico.

### 3. Risultati

#### *Analisi Descrittiva*

Le tabelle 8.1 e 8.2 mostrano le distribuzioni dei rilievi tra le province di appartenenza dei centri di stoccaggio e la distribuzione percentuale dei rilievi per varietà.

Per quanto riguarda le province si osservi come per Ragusa e Siracusa sono disponibili i dati degli ultimi anni, e ciò perchè inizialmente al progetto non hanno aderito centri ubicati nelle suddette province. La provincia con il maggior numero di rilievi è Palermo (25,7%) seguita da Enna (18,8) e da Agrigento (17,6%).

Dalla tabella 8.2 si osserva come tra le varietà più rilevate troviamo Simeto (23,5%), seguita da Ciccio (15,4%), Duilio (15,1%), ed Arcangelo (8,6%) e queste quattro varietà insieme formano circa il 63% dei rilievi totali. È anche interessante osservare la diminuzione delle percentuali legate alla varietà Colosseo, da 3,2% del 2000 a meno dell'1% per gli anni 2004-2005, e gli aumenti percentuali dei rilievi delle varietà Mongibello (da 0,1 a 5,9%) e Rusticano (da 1,5 a 5,1%). Queste variazioni comunque rispecchiano la dinamica della diffusione delle suddette varietà sul territorio regionale ed infatti la produzione di seme di Colosseo è passata da 34.846 q del 2000 a 10.653 q del 2004, quella di Mongibello da 1.681 q del 1999 a 21.624 q del 2004, quella di Rusticano da 4.782 q del 1998 a 33.440 q del 2004 (fonte: E.N.S.E.).

**Tab. 8.1** - Distribuzione (valori assoluti e percentuali sul totale delle  $n=42.801$  osservazioni) dei rilievi tra le province interessate nel progetto

Anno	AG		CL		CT		EN		PA		RG		SR		TP	
	Ass.	%	Ass.	%	Ass.	%	Ass.	%	Ass.	%	Ass.	%	Ass.	%	Ass.	%
2000	1.022	24.0	733	17.2	148	3.5	606	14.3	732	17.2	0	0.0	0	0.0	1.011	23.8
2001	1.387	25.6	424	7.8	863	15.9	585	10.8	899	16.6	0	0.0	0	0.0	1.258	23.2
2002	705	15.2	432	9.3	235	5.1	668	14.4	2.125	45.7	0	0.0	177	3.8	306	6.6
2003	1.683	19.3	832	9.5	996	11.4	2.761	31.7	1.726	19.8	0	0.0	378	4.3	346	4.0
2004	2.043	19.3	1.339	12.6	657	6.2	2.213	20.9	2.650	25.0	234	2.2	586	5.5	884	8.3
2005	702	7.7	1.792	19.6	717	7.8	1.226	13.4	2.849	31.1	607	6.6	326	3.6	938	10.2
<b>Totale</b>	<b>7.542</b>	<b>17.6</b>	<b>5.552</b>	<b>13.0</b>	<b>3.616</b>	<b>8.4</b>	<b>8.059</b>	<b>18.8</b>	<b>10.981</b>	<b>25.7</b>	<b>841</b>	<b>2.0</b>	<b>1.467</b>	<b>3.4</b>	<b>4.743</b>	<b>11.1</b>

**Tab. 8.2** - Distribuzione percentuale (sul totale delle  $n=42.801$  osservazioni) dei rilievi relativi alle varietà registrate per anno di osservazione

ANNO	simeto	ciccio	duilio	arcangelo	valbelice	rusticano	mongibello	colosseo	creso	totale
2000	19.8	19.1	18.5	4.3	4.9	1.5	0.1	3.2	1.3	72.7
2001	22.1	17.8	15.1	10.8	3.6	4.3	0.3	2.6	3.0	79.4
2002	25.3	16.5	18.3	8.9	3.1	2.3	0.5	1.3	1.7	77.9
2003	27.6	15.4	15.2	12.6	1.9	2.9	1.9	1.2	0.4	79.0
2004	23.3	16.7	14.8	9.2	4.2	2.9	2.1	0.4	0.7	74.4
2005	21.2	10.2	12.0	4.8	3.4	5.1	5.9	0.6	1.8	65.1
<b>Totale</b>	<b>23,5</b>	<b>15,4</b>	<b>15,1</b>	<b>8,6</b>	<b>3,5</b>	<b>3,4</b>	<b>2,3</b>	<b>1,3</b>	<b>1,3</b>	<b>74,2</b>

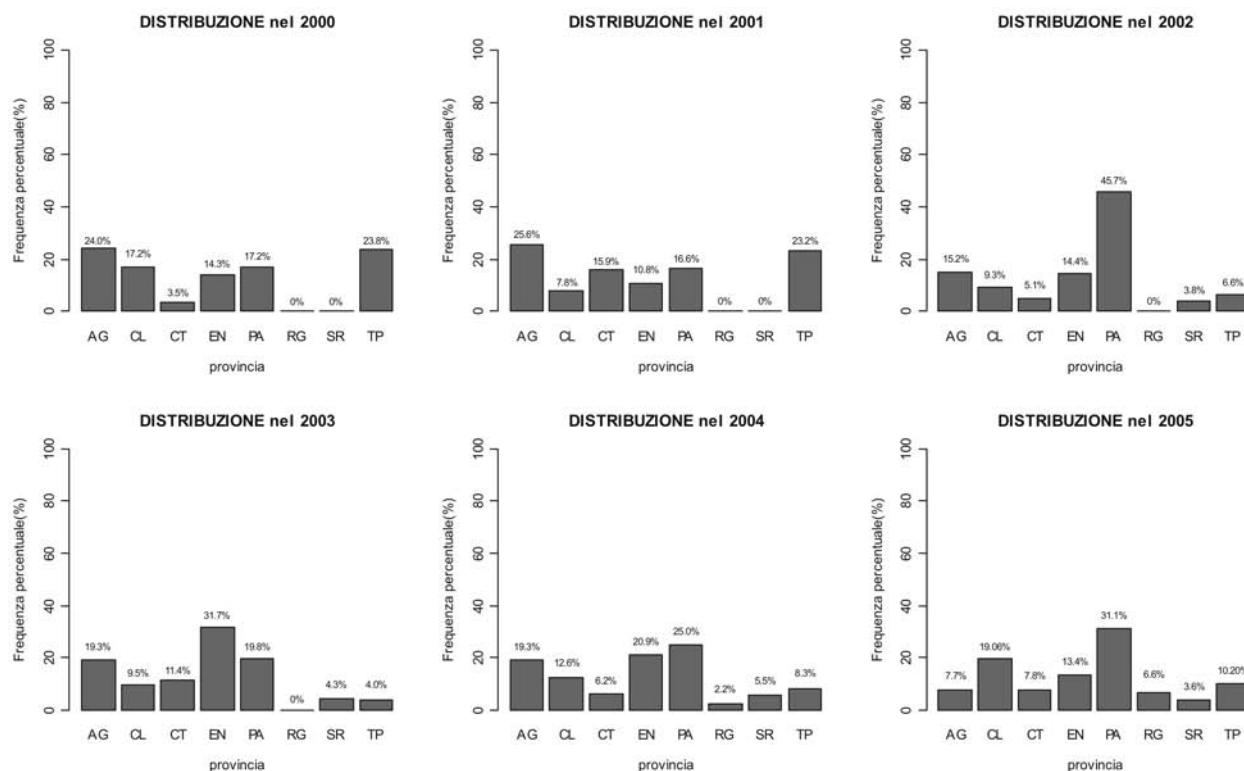


Fig. 8.1 Diagramma a barre della distribuzione dei rilievi per provincia di appartenenza per singolo anno.

Statistiche descrittive relative al contenuto proteico, variabile di maggiore interesse in questo lavoro, sono riportate nelle tabelle 8.3 e 8.4, dove i valori medi e le deviazioni standard sono stati calcolati per *anno e provincia* (tabella 8.3) ed *anno e varietà* (tabella 8.4). È forse utile evidenziare che mentre la ‘media’ definisce l’apporto proteico medio, la deviazione standard rappresenta la variabilità con cui tale valore medio è stato ottenuto. In tal senso la deviazione standard può essere considerata come una misura di ‘affidabilità’ del valore. Ad esempio, da un esame più approfondito della tabella 8.4 si nota come la varietà Creso e soprattutto Mongibello siano associate ad un maggiore contenuto proteico (ad esempio nel 2000 i valori medi risultano rispettivamente pari a 8% e 15.50%) ma anche ad una maggiore variabilità con deviazioni standard palesemente superiori a quelle delle altre varietà. È interessante evidenziare la diminuzione dei valori medi e di deviazioni standard che si registra per tutte le varietà nel corso degli anni 2004 e 2005.

Tab. 8.3 - Contenuto proteico, media (%) e deviazione standard (d.s.), della granella stoccata per province interessate e anno di osservazione

Provincia	2000		2001		2002		2003		2004		2005	
	media	d.s.	media	d.s.	media	d.s.	media	d.s.	media	d.s.	media	d.s.
AG	12.84	1.73	11.63	1.31	13.73	1.92	12.40	1.85	10.83	0.93	11.34	0.85
CL	13.15	1.83	11.68	1.13	12.74	1.61	11.99	1.86	11.07	0.99	11.32	1.00
CT	12.45	1.36	10.98	1.92	13.38	1.48	12.54	1.77	11.30	0.88	10.95	0.92
EN	11.54	1.38	10.68	1.59	12.94	1.79	12.28	2.05	10.65	1.10	11.05	1.01
PA	13.44	1.72	12.21	1.30	13.11	1.87	12.24	1.79	11.20	0.99	11.33	0.98
RG	-	-	-	-	-	-	-	-	11.63	0.98	11.37	1.06
SR	-	-	-	-	12.57	1.04	11.94	1.31	11.08	0.67	11.22	1.12
TP	12.14	1.35	11.37	1.24	13.25	1.63	12.16	1.72	11.26	0.96	11.14	0.96

**Tab. 8.4** - Contenuto proteico, media (%) e deviazione standard (d.s.) della granella stoccata per varietà e anno di osservazione.

	2000		2001		2002		2003		2004		2005	
	media	d.s.	Media	d.s.	media	d.s.	media	d.s.	media	d.s.	media	d.s.
SIMETO	12.40	1.59	11.59	1.55	13.36	1.83	12.46	1.81	11.10	0.96	11.37	0.94
ARCANGELO	12.69	1.73	11.21	1.27	12.54	1.76	11.53	1.89	10.66	1.05	10.85	1.01
CICCIO	12.15	1.68	10.88	1.55	12.88	1.73	12.06	1.95	10.90	0.98	10.75	0.84
DUILIO	13.13	1.85	11.65	1.35	13.69	1.77	12.57	1.73	11.13	0.98	11.36	0.86
COLOSSEO	12.39	1.54	11.82	1.61	12.47	1.17	12.41	1.74	10.62	0.88	11.51	0.86
VALBELICE	12.47	1.79	12.20	1.35	13.26	2.03	12.95	1.87	10.95	0.99	11.45	0.92
CRESO	13.38	2.13	12.10	1.33	13.05	1.65	12.63	1.73	11.84	0.89	11.60	0.96
MONGIBELLO	15.50	2.60	12.54	1.63	13.67	2.17	13.49	2.23	11.33	1.08	11.65	0.96
ALTRI	12.92	1.87	11.84	1.42	13.12	1.67	12.32	1.72	11.07	0.98	11.24	1.00

### *Un tentativo di analisi multipla: i modelli di regressione*

I valori riportati nelle tabelle 8.3 e 8.4, unitamente ai relativi grafici in figura 8.2, sono molto utili perché danno indicazioni di massima e di facile lettura delle caratteristiche riguardanti i diversi fattori coinvolti nello studio.

Tuttavia l'interpretazione di tali statistiche descrittive necessita una qualche attenzione: ad esempio, considerando il *boxplot* relativo alla provincia di appartenenza del centro, si può osservare come il grano proveniente dai centri di stoccaggio della provincia di Palermo abbia, in media, un contenuto proteico superiore a quello del grano proveniente dalle altre province. Sebbene tale risultato possa essere degno di una qualche attenzione a scopi descrittivi, risulta altresì importante capire se questa differenza è reale. Il maggiore contenuto proteico riscontrato nel campione potrebbe infatti non riflettere le vere differenze relative alla popolazione, ovvero a tutto il grano prodotto (e producibile) in Sicilia. È altresì importante individuare se tale eventuale differenza è realmente dovuta alla provincia di "Palermo" (intesa come *proxy* di altri fattori legati al territorio e al microclima) o può essere confusa da altri fattori. Ad esempio, se i coltivatori della provincia di Palermo avessero la propensione a concimare il proprio terreno con dosi più elevate rispetto a quelle utilizzate nelle altre province, allora i valori elevati riscontrati per Palermo verosimilmente potrebbero essere dovuti ad una elevata concimazione piuttosto che da altri fattori afferenti al territorio.

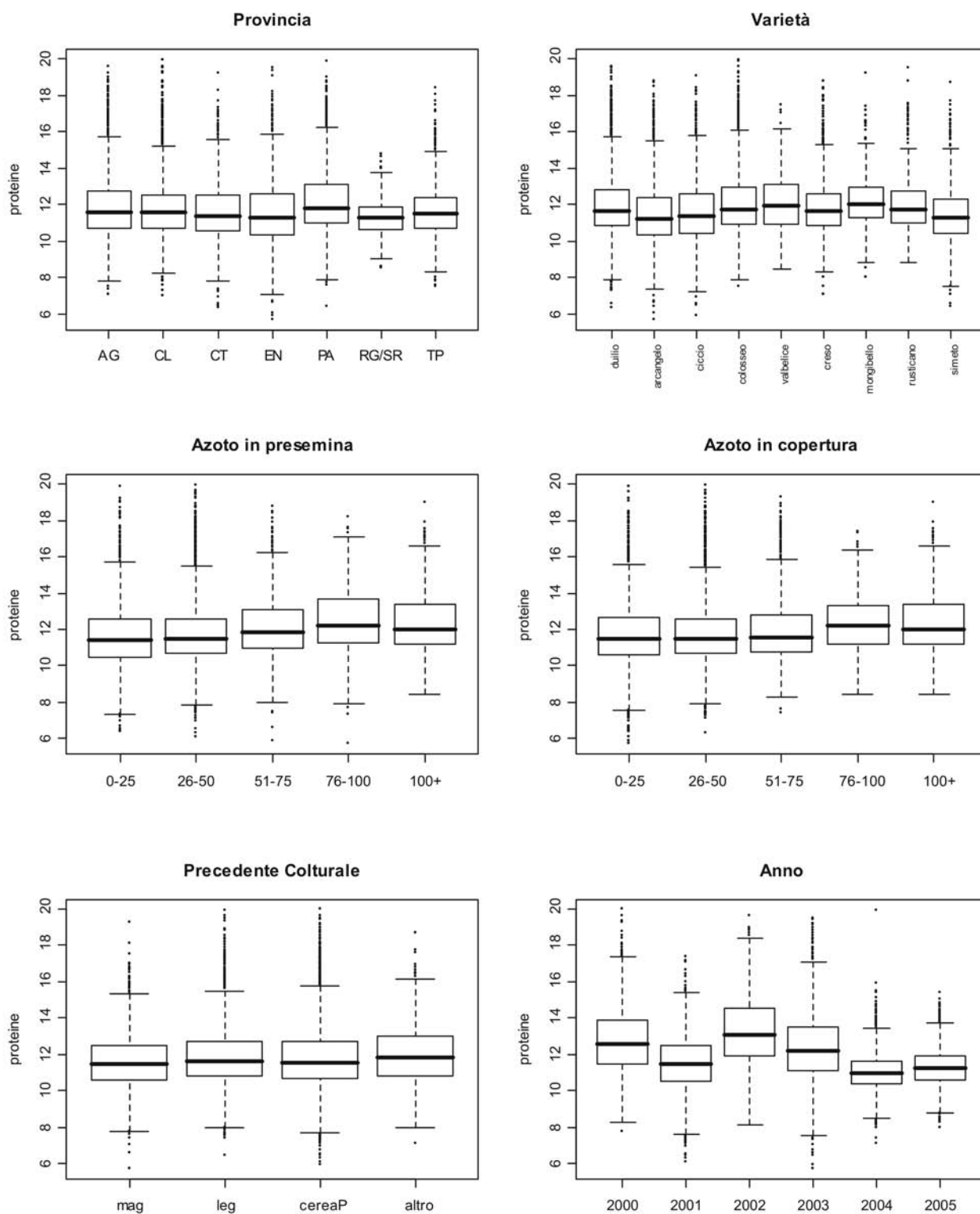
I modelli di regressione descritti sopra possono rispondere ai quesiti posti: ovvero essi forniscono stime delle variazioni del contenuto in proteine imputabili ai diversi fattori, considerando e *depurando* dal possibile effetto *confondente* degli altri fattori *inclusi* nel modello. Variabili non incluse, se legate in qualche modo sia alla risposta sia ai fattori, possono influenzare i risultati.

Per la notevole influenza che le condizioni climatiche hanno sulla produzione del grano e per la forte variabilità che c'è tra gli anni (vedi relativo *boxplot* in figura 8.2) sono state condotte analisi di regressione stratificate per anno. Ovvero per ogni anno (dal 2000 al 2005) è stato stimato un modello di regressione, con variabili esplicative descritte in precedenza.

Ottenute le stime specifiche per anno è stato verificato se il contributo di ogni fattore è *omogeneo* negli anni, ovvero se i valori di uno stesso fattore ottenuti negli anni, sono (approssimativamente) costanti. Ed inoltre, poiché le stime specifiche per anno sono soggette a fluttuazioni campionarie, è stato verificato se è possibile ottenere un unico valore che sintetizzi ed esprima l'effetto di ogni fattore sul contenuto proteico.

Le tabelle 8.5 e 8.6 sintetizzano i concetti sopra esposti. Avendo ottenuto per ogni anno le stime dei coefficienti delle variabili esplicative, successivamente si è proceduto a sintetizzare tali valori per ottenere delle misure sintetiche. Lo scopo è stato quello di ottenere, per ogni variabile inserita nel modello, un unico valore che potesse essere meno affetto da incertezza e quindi di più facile lettura. Tali stime sintetiche vengono dette di 'meta-analisi'. Queste vengono ottenute considerando una media delle stime specifiche per anno ponderate per l'inverso del loro errore standard: più piccolo è l'errore standard, maggiore sarà il peso nella determinazione della stima complessiva.





**Fig. 8.2** – Boxplot<sup>1</sup> del contenuto proteico rispetto alle variabili esplicative

<sup>1</sup>Il boxplot è una rappresentazione grafica che consente di evidenziare sinteticamente gli aspetti di una distribuzione. All'interno della 'scatola' è racchiuso il 50% delle osservazioni (con valori compresi tra i percentili 0.25 e 0.75) e la linea orizzontale all'interno della scatola rappresenta la mediana della distribuzione stessa. Le linee orizzontali in alto ed in basso rappresentano i limiti che considerazioni teoriche e pratiche suggeriscono possano essere interpretati come valori di "tolleranza". Le osservazioni più grandi e più piccole di tali estremi, che nei grafici vengono rappresentate con dei pallini, costituiscono valori molto diversi dal resto della distribuzione.

Oltre a fornire un unico indice quantitativo per ogni variabile, le tecniche di meta-analisi consentono di verificare l'omogeneità (cioè l'uguaglianza) delle stime anno-specifiche. Naturalmente una stima di meta-analisi sarà più affidabile e 'sensata', quanto meno diverse saranno le stime individuali.

Nella prima colonna sono riportate le stime sintetiche di ogni effetto, escludendo le categorie assunte come riferimento che hanno valore fissato pari a zero. Le successive due colonne riportano gli estremi degli intervalli di confidenza (IC) al livello di fiducia del 95%, *IC (95%)*. Ciò è possibile affermare che il vero valore ignoto dell'effetto ricada nell'IC con una fiducia del 95%. È chiaro che IC che escludono lo zero necessitano di particolare attenzione perché individuano i fattori che, secondo il modello, sono risultati significativamente associati (positivamente o negativamente) con il contenuto proteico.

Così rispetto alla varietà "Simeto", "Mongibello" è quella che, in media, garantisce il contenuto proteico più alto (0.766, *IC(95%)* 0.253, 1.280), mentre "Duilio" e soprattutto "Ciccio" mostrano *performance* peggiori. La concimazione con azoto in presemina sembra associata ad un migliore contenuto in proteine solo con dosi elevate, mentre una chiara evidenza esiste in favore della concimazione azotata in copertura: si nota come le due categorie "26-50" e "51+" abbiano incrementi significativamente diversi da zero, 0.20 (*IC 95%* 0.001, 0.398) e 0.36 (*IC 95%*, 0.102, 0.618) rispettivamente. Il ruolo della provincia, inteso come insieme di fattori territoriali e microclimatici, risulta anche abbastanza chiaro: le province di TP ed EN risultano associate a contributi proteici più bassi. Molto incerti sono i risultati relativi al precedente colturale, dove le stime sono prossime allo zero che è compreso nei relativi *IC(95%)*.

Infine l'ultima colonna della tabella 8.6 riporta il risultato di un *test statistico* finalizzato a verificare l'omogeneità negli anni delle stime. Righe contrassegnate da una '√' individuano le stime che sono risultate omogenee negli anni: la quasi totalità degli effetti stimati (15 su 20 stime) emergono essere differenti negli anni.

#### 4. Conclusioni

Il presente studio costituisce un primo tentativo di analisi "multivariata", in cui si è cercato di spiegare il contenuto in proteine in funzione di fattori potenzialmente influenti: varietà, azoto in presemina ed in copertura, provincia e precedente colturale. Naturalmente a causa della natura *osservazionale* dello studio (opposto ai cosiddetti studi *sperimentali* o *controllati*) particolare attenzione è necessaria ad interpretare i risultati perché i fattori che possono influenzare il raccolto sono molteplici e talvolta di difficile rilevazione e misurazione. Nel presente lavoro, ad esempio, i fattori pedoclimatici (piovosità, esposizione, giacitura, ecc.) non vengono contemplati; tuttavia, nonostante la mancanza di dette informazioni, i risultati di tale studio possono risultare importanti e costituire un punto di partenza per ricerche successive.

Riteniamo comunque che questa prima esperienza possa in futuro essere perfezionata apportando alcuni miglioramenti indirizzati ad affinare l'analisi statistica e nello specifico detti miglioramenti potrebbero riguardare uno o più dei punti seguenti:

- analisi multivariata, ovvero tener conto delle correlazioni esistenti tra due o più variabili risposta, (ad esempio contenuto in proteine e resa);
- effetto *clustering* dovuto alla differente qualità ed affidabilità dei dati rilevati dai singoli centri di stoccaggio;
- interazioni (sinergia) tra fattori, ad esempio è plausibile ritenere che esista un'azione sinergica di azoto in presemina e azoto in copertura sul contenuto proteico;
- aspetti metodologici legati alla costruzione del modello, quali possibili assunzioni sulla forma della densità della risposta e modellazione congiunta della varianza.

In conclusione si ritiene che, sebbene alcuni risultati risultino già noti, essi sono comunque interessanti perché costituiscono stime quantitative (con associati intervalli di confidenza) 'aggiustate' per gli altri fattori confondenti inseriti nel modello. Inoltre esse si riferiscono a situazioni non-sperimentali, ma ad osservazioni di pieno campo e per questo sono potenzialmente di notevole interesse per gli addetti ai lavori.

Tab. 8.5 - Stime dei coefficienti e relativi errori standard nei modelli di regressione nei diversi anni di osservazione.

VARIABILI	2000		2001		2002		2003		2004		2005	
	Stima	Err:St	Stima	Err:St	Stima	Err:St	Stima	Err:St	Stima	Err:St	Stima	Err:St
Varietà												
<i>Simeto</i>	0	-	0	-	0	-	0	-	0	-	0	-
<i>Ciccio</i>	-0.440	0.122	-0.593	0.098	-0.844	0.095	-0.732	0.069	-0.331	0.037	-0.759	0.043
<i>Dutilio</i>	-0.712	0.161	-0.712	0.086	-0.523	0.12	-0.405	0.075	-0.149	0.048	-0.594	0.060
<i>Arcangelo</i>	0.380	0.096	-0.22	0.082	0.258	0.097	0.165	0.073	0.041	0.035	-0.163	0.043
<i>Colosseo</i>	-0.055	0.195	0.126	0.145	-0.813	0.267	-0.147	0.205	-0.118	0.162	-0.190	0.241
<i>Valbelice</i>	-0.630	0.156	0.386	0.134	-0.373	0.200	0.226	0.156	0.132	0.062	0.191	0.069
<i>Creso</i>	-0.210	0.286	0.529	0.133	-0.134	0.222	0.071	0.336	0.713	0.124	0.165	0.093
<i>Mongibello</i>	1.872	0.923	1.608	0.576	0.172	0.407	1.682	0.182	0.108	0.095	0.285	0.058
<i>Rusticano</i>	-1.055	0.341	-0.609	0.127	-0.066	0.192	-0.563	0.137	-0.003	0.07	-0.379	0.075
Presemina												
<i>0-25</i>	0	-	0	-	0	-	0	-	0	-	0	-
<i>26-50</i>	0.170	0.156	-0.004	0.102	0.154	0.115	0.022	0.073	-0.026	0.041	0.102	0.045
<i>51+</i>	0.183	0.181	-0.188	0.123	0.110	0.116	0.046	0.083	0.081	0.062	0.17	0.058
Azoto Copertura												
<i>0-25</i>	0	-	0	-	0	-	0	-	0	-	0	-
<i>26-50</i>	0.223	0.106	0.111	0.084	0.534	0.093	0.382	0.071	-0.129	0.038	0.119	0.046
<i>51+</i>	0.421	0.110	0.334	0.125	0.686	0.113	0.745	0.099	-0.02	0.05	0.053	0.052
Provincia												
<i>AG</i>	0	-	0	-	0	-	0	-	0	-	0	-
<i>CL</i>	0.253	0.143	0.048	0.155	-1.329	0.156	-0.421	0.098	0.191	0.049	0.356	0.058
<i>CT</i>	-0.200	0.337	-0.481	0.114	-0.2	0.251	0.203	0.109	0.442	0.059	-0.067	0.075
<i>EN</i>	-1.052	0.136	-0.841	0.117	-0.888	0.148	-0.099	0.086	-0.253	0.053	-0.024	0.057
<i>PA</i>	0.366	0.122	0.518	0.137	-0.435	0.129	-0.286	0.089	0.334	0.042	0.238	0.047
<i>RG/SR</i>	-											
<i>TP</i>	-1.325	0.125	-0.545	0.103	-1.075	0.154	-0.679	0.121	0.319	0.053	0.123	0.073
Preced. Colturale												
<i>maggese</i>	0	-	0	-	0	-	0	-	0	-	0	-
<i>leguminose</i>	0.145	0.140	0.085	0.118	-0.290	0.134	-0.078	0.117	0.017	0.05	-0.005	0.051
<i>cereali a paglia</i>	-0.023	0.121	0.153	0.102	-0.108	0.131	0.557	0.109	0.051	0.045	-0.253	0.046
<i>altro</i>	-0.058	0.169	0.474	0.151	-0.148	0.17	0.024	0.154	0.01	0.068	0.074	0.083

**Tab 8.6** - Stime complessive e relativi intervalli di confidenza per l'effetto dei fattori nella determinazione del contenuto proteico negli anni.

	Stima	IC(95%)		Omogeneità tra anni
		lim-inf	lim-sup	
<b>Varietà</b>				
<i>Simeto</i>	0			
<i>Ciccio</i>	-0.617	-0.818	-0.416	
<i>Duilio</i>	-0.504	-0.713	-0.294	
<i>Arcangelo</i>	0.066	-0.091	0.224	
<i>Colosseo</i>	-0.109	-0.260	0.043	✓
<i>Valbelice</i>	0.016	-0.215	0.248	
<i>Creso</i>	0.245	-0.043	0.533	
<i>Mongibello</i>	0.766	0.253	1.280	
<i>Rusticano</i>	-0.389	-0.644	-0.133	
<b>Presemina</b>				
<i>0-25</i>	0			
<i>26-50</i>	0.038	-0.012	0.089	✓
<i>51+</i>	0.091	0.025	0.157	✓
<b>Azoto Copertura</b>				
<i>0-25</i>	0			
<i>26-50</i>	0.200	0.001	0.398	
<i>51+</i>	0.360	0.102	0.618	
<b>Provincia</b>				
<i>AG</i>	0			
<i>CL</i>	-0.138	-0.508	0.232	
<i>CT</i>	-0.026	-0.349	0.297	
<i>EN</i>	-0.507	-0.806	-0.208	
<i>PA</i>	0.125	-0.106	0.356	
<i>TP</i>	-0.530	-1.042	-0.019	
<b>Preced. Colturale</b>				
<i>maggese</i>	0			
<i>leguminose</i>	-0.003	-0.064	0.058	✓
<i>cereali a paglia</i>	0.057	-0.155	0.270	
<i>altro</i>	0.053	-0.033	0.139	✓

### **Riferimenti Bibliografici**

Iacus SM (2005) Statistica, McGraw-Hill

Weisberg S. (2005), Applied Linear Regression, III ed., Wiley

Sgrulletta D., De Stefanis E., Conciatori A. (1999). *Analisi veloce del colore giallo sulla granella. Il sistema qualità nella filiera frumento duro*. Tecnica molitoria, 8: 861-866.

Cartabellotta D., Sgrulletta D., Monastero E., Tusa P., Messina B., Russo G., Catalano R., Nobile G., (2002). *Sistema di gestione per la qualità del grano duro in Sicilia. Conforme alla norma UNI EN ISO 9001:2000*. L'Informatore Agrario, 31: 31-32.

Russo G., Messina B., Campisi D., Fonti A., Monastero E., (2007). *Monitoraggio della qualità del grano duro prodotto in Sicilia: risultati delle campagne di raccolto biennio 2005-2006 e confronto con i risultati registrati nel quinquennio 2000-2004*. Osservatorio della filiera cerealicola siciliana – Terzo Rapporto – La filiera del grano duro in Sicilia.

