Lezione 1

RICLASSIFICARE E LEGGERE CONGIUNTAMENTE

In questa lezione ...

In questa lezione introdurremo le procedure elementari di rilevazione e analisi congiunta di due caratteri, presupposto dell'analisi statistica bivariata. In particolare:

- Familiarizzeremo attraverso esempi con l'apparato di base della contabilità bivariata: le tabelle a **doppia entrata**, le frequenze **congiunte** e le frequenze **marginali**.
- Ci soffermeremo in particolare sul modo di **leggere una tabella** a partire dal confronto delle distribuzioni vincolate.
- A partire da distribuzioni congiunte torneremo alla costruzione di variabili singole ottenute per **somma o media di** altre **variabili**.
- Infine concentrandoci su variabili quantitative discrete, esamineremo una **rappresentazione grafica** corrispondente alla tabella a doppia entrata, verificando le corrispondenze tra i due strumenti.

Ripartiamo da una matrice dati

Ripartiamo allora dalle informazioni raccolte a una cena di classe (la classe A delle prime lezioni..) e riclassifichiamo i presenti in base non a un solo carattere, ma due insieme: il genere e il titolo di studio S.

Costruiamo la corrispondente **tabella a doppia entrata** conteggiando non le singole osservazioni ma le coppie di osservazioni riferite a un individuo:

G\S	D	L	P	N(G)
M	//=2	/=1	/=1	4
F	0	////=4	//=2	6
N(S)	2	5	3	10

Anche qui la somma delle numerosità dei caratteri congiunti di una riga/colonna dà la numerosità di un carattere singolo

G	S
M	D
M	L
M	P
M	D
F	L
F	L
F	P
F	P
F	L
F	L

La tabella a doppia entrata

Esprimiamo in linguaggio formale la tabella a doppia entrata:

x_i \ y_j	y_1	y_2	…	y_j	…	y_s	n_i
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_{2*}
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_{i*}
…	…	…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_{r*}
n_j	n_{*1}	n_{*2}	…	n_{*j}	…	n_{*s}	N

- n_{ij} = numerosità congiunte di osservazioni $\{X=x_i \wedge Y=y_j\}$
- n_{i*} = numerosità marginali di osservazioni $\{X=x_i\}$ qualunque sia il valore di y
- n_{*j} = numerosità marginali di osservazioni $\{Y=y_j\}$ qualunque sia il valore di x
- $n_{i*} = \Sigma_{J=1..s}\, n_{ij}$ **CONDIZIONI**
- $n_{*J} = \Sigma_{i=1..r}\, n_{ij}$ **DI QUADRO**
- $N = \Sigma_{i=1..r}\, n_{i*} = \Sigma_{J=1..s}\, n_{*j}$

Le numerosità marginali

Isoliamo righe e colonne 'ai margini' della tabella:

X_i	n_i
x_1	n_{1*}
x_2	n_{2*}
…	…
x_i	n_{i*}
…	…
x_r	n_{r*}
	N

Y_j	y_1	y_2	…	y_j	…	y_s	
n_j	n_{*1}	n_{*2}	…	n_{*j}	…	n_{*s}	N

- Le due colonne ai margini sinistro e destro della tabella, ricompattate insieme, riproducono la distribuzione 'univariata' di X, per la quale vale $\Sigma_{i=1..r} n_{i*} = N$
- Le due righe ai margini alto e basso della tabella, ricompattate insieme, riproducono la distribuzione 'univariata' di Y, per la quale vale $\Sigma_{j=1..s} n_{*j} = N$
- La tabella a doppia entrata contiene dunque almeno tre distribuzioni distinte:
  - La distribuzione congiunta
  - Le due distribuzioni semplici (**marginali**) di X e Y

Dalle numerosità alle frequenze relative

Come per le distribuzioni univariate, le numerosità possono essere relativizzate, dividendole per la numerosità totale: $\mathbf{f_{ij} = n_{ij} / N}$

X_i \ Y_j	y_1	y_2	…	y_j	…	y_s	f_i
x_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1s}	f_{1*}
x_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2s}	f_{2*}
…	…	…	…	…	…	…	…
x_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{is}	f_{i*}
…	…	…	…	…	…	…	…
x_r	f_{r1}	f_{r2}	…	f_{rj}	…	f_{rs}	f_{r*}
f_j	f_{*1}	f_{*2}	…	f_{*j}	…	f_{*s}	1

- f_{ij} = frequenze **congiunte** di osservazioni $\{X=x_i \wedge Y=y_j\}$
- f_{i*} = frequenze **marginali** di osservazioni $\{X=x_i\}$ qualunque sia il valore di y
- f_{*j} = frequenze **marginali** di osservazioni $\{Y=y_j\}$ qualunque sia il valore di x
- $f_{i*} = \Sigma_{j=1..s} f_{ij}$ **CONDIZIONI**
- $f_{*j} = \Sigma_{i=1..r} f_{ij}$ **DI QUADRO**
- $1 = \Sigma_{i=1..r} f_{i*} = \Sigma_{j=1..s} f_{*j}$

Variabili doppie quantitative e per classi

Quando poco fa abbiamo ricuperato la matrice dati del gruppo di compagni di classe, abbiamo costruito la distribuzione congiunta tra due caratteri qualitativi, uno nominale-dicotomico (G), l'altro ordinale (S titolo di studio).

Ma possiamo produrre con identica procedura distribuzioni congiunte di variabili quantitative o di qualitative e quantitative abbinate insieme.

Per esempio costruiamo la variabile doppia (S,E): siamo curiosi di capire se c'è qualche legame tra studio e performance economica.

S	E
D	4,5
L	2,2
P	3,7
D	6,8
L	4,7
L	4,2
P	1,5
P	5,0
L	5,2
L	7,2

A suo tempo (vol. I, lezione 1) avevamo riaggregato per classi E in due modi diversi. Qui usiamo un terzo criterio (è lecito!): da 0 a 4, da 4 a 6, oltre 6.

Chi studia guadagna di più (o no?)

Costruiamo la tabella col solito noioso lavoro di conteggio.

E' vero: "lo fa il calcolatore"; ma dobbiamo sapere come lavora!.

Il risultato è nella tabella qua sotto. Cosa ci dice? Proseguendo faremo qualche passo in più per 'leggere' una tabella.

S\E	0-|4	4-|6	6-|8	N(S)
D	0	/=1	/=1	2
L	/=1	///=3	/=1	5
P	//=2	/=1	0	3
N(E)	3	5	2	10

Ma già qui possiamo imparare una cosa: una 'spia' della relazione che studiamo è data dalla diagonale (se esiste) in cui si concentra la maggior parte delle osservazioni.

Se l'ellisse (blu) si colloca sulla diagonale principale, c'è una relazione diretta tra S e E.

Qui l'ellisse si colloca sulla controdiagonale: c'è quindi una relazione inversa tra S e E (chi studia non piglia pesci?)

S	E
D	4,5
L	2,2
P	3,7
D	6,8
L	4,7
L	4,2
P	1,5
P	5,0
L	5,2
L	7,2

Leggere una tabella con le frequenze vincolate

C'è una qualche relazione tra livello di scolarità (S) e livello delle entrate (E)? Il modo migliore per capirlo è confrontare le distribuzioni di frequenze vincolate.

S \ E	Basse	Medie	Alte	
Diploma	0	1	1	2
Laurea	1	3	1	5
PostLaurea	2	1	0	3
	3	5	2	10

La nostra testa ha un modo automatico per valutare l'eventuale relazione tra S e E. Consiste nell'analizzare la 'distribuzione dei redditi' separatamente per diplomati laureati e post: cioè separatamente **riga per riga**.

S \ E	Bas	Med	Alt	
Dipl	0/2	1/2	1/2	2
Lau	1/5	3/5	1/5	5
Post	2/3	1/3	0/3	3
	3/10	5/10	2/10	10

Tra i **diplomati** (prima riga) nessuno ha bassi redditi, uno su due (50%) ha medi redditi, uno su due (50%) alti redditi.
Tra i **laureati** uno su cinque (20%) ha bassi redditi, tre su cinque (60%) medi redditi, uno su cinque (20%) alti redditi.
Tra i **postlaurea** solo uno su tre (33%) ha medi redditi, nessuno su tre alti redditi.
Sintetizziamo: **la % con alti redditi è del 50% tra i diplomati, del 20% tra i laureati e dello 0% tra i post**...

Il rapporto tra le numerosità marginali e il loro totale ci dà le frequenze relative marginali.
Lo stesso rapporto per una singola riga (o colonna) ci dà la frequenze vincolate.

Cosa sono le frequenze vincolate

Definiamo quindi frequenza vincolata $f_{j/i}$ il rapporto tra la numerosità congiunta n_{ij} e la marginale di riga corrispondente n_{i*}, o – indifferentemente - il rapporto tra la frequenza congiunta f_{ij} e la marginale di riga corrispondente f_i.

$$f_{j|i} = n_{ij} / n_{i*} = f_{ij} / f_{i*}$$

Y_j / X_i	y_1	…	y_j	…	y_s	f_i
x_1	f_{11}/f_{1*}	…	f_{1j}/f_{1*}	…	f_{1s}/f_{1*}	1
…	…	…	…	…	…	…
x_i	f_{i1}/f_{i*}	…	f_{ij}/f_{i*}	…	f_{is}/f_{i*}	1
…	…	…	…	…	…	…
x_r	f_{r1}/f_{r*}	…	f_{rj}/f_{r*}	…	f_{rs}/f_{r*}	1

- Per ogni riga vale la condizione di quadro 1= $\Sigma_{j=1..s} f_{j|i}$
- Le frequenze vincolate possono essere calcolate per colonna, rapportando una numerosità (o frequenza) congiunta alla corrispondente marginale di colonna: $f_{i|j} = n_{ij}/n_{*j} = f_{ij}/f_{*j}$
- Ovviamente anche per ogni colonna vale la condizione di quadro 1= $\Sigma_{i=1..r} f_{i|j}$

Un altro esempio: tabelle tetracoriche

Facciamo un altro esempio, riclassificando le venti regioni italiane secondo il colore della giunta regionale e il tasso % di occupazione maschile tra i 25 e i 34 anni.

Piemonte	S	88
Val d'Aosta	S*	88
Lombardia	D	88
Trentino	S	91
Veneto	D	89
Friuli	D	85
Liguria	D	80
Emilia	S	88
Toscana	S	84
Umbria	S	83
Marche	S	84
Lazio	D	70
Abruzzi	D	76
Molise	D	71
Campania	S	58
Puglie	D	70
Basilicata	S	64
Calabria	D	55
Sicilia	D	63
Sardegna	D	65

In questo caso una variabile (G) è già dicotomica (sinistra–destra), l'altra (T) è quantitativa discreta, ma nulla ci impedisce di dicotomizzarla, fissando per esempio una soglia a T=80.

G \ T	≥ 80	< 80	
Sinistra	///////=7	// = 2	9
Destra	//// = 4	/////// = 7	11
	11	9	20

Una tabella due per due (si dice tabella tetracorica) è la forma più semplice di distribuzione congiunta.

G T	alto	bas	
Sin	7	?	9
De	?	?	?
	11	?	20

Tanto semplice che basta una frequenza congiunta e 2 marginali per ricostruirla (provate!!)

Per riga o per colonna è lo stesso

L'esempio ridotto all'osso consente di rifare esercizio di lettura di una distribuzione congiunta. Ci chiediamo:c'è relazione tra colore della giunta e tasso di occupazione?

Alla domanda rispondiamo calcolando le frequenze vincolate $f(t_j|g_i)=n(t_j,g_i)/n(g_i)$

G T	alto	bas	
Sin	8	2	10
De	4	6	10
	12	8	20

G T	alto	bas	
Sin	0,8	0,2	1
De	0,4	0,6	1
	0,6	0,4	1

G T	alto	bas	
Sin	0,67	0,25	0,5
De	0,33	0,75	0,5
	1	1	1

Nelle giunte di sinistra la % di regioni con alto livello di occupazione è diversa e maggiore di quella riscontrata nelle regioni con giunte di destra. I due caratteri regionali dunque non risultano indipendenti tra loro (ma non pensate subito a un nesso tra occupazione e tendenze politiche!!)

Va detto invece che, costruendo le frequenze vincolate per colonna, la lettura della tabella ci deve portare alle stesse considerazioni. Vediamo che nelle regioni a più alta occupazione la % di giunte di sinistra è maggiore che nelle regioni a bassa occupazione: ma questo è solo un modo diverso per dire la stessa cosa detta sopra.

Dunque: sia che si legga (tramite frequenze vincolate) una tabella a doppia entrata per colonna oppure per riga, se c'è una qualche connessione la si coglierà comunque!

Distribuzioni congiunte e variabili somma

Nel 1974 Peter Townsend (non il cantante!) studiando la carta della povertà nel Regno Unito ha costruito un Indice di Deprivazione Relativa Multipla, sommando senza alcuna ponderazione (unweighted) 12 indicatori Sì/No di disagio:

Sei indicatori di risorse materiali:	
* *Regime dietetico*	* *Mezzi/attrezzature*
* *Vestiario*	* *Lavoro*
* *Riscaldamento*	* *Abitazione*

Sei indicatori di capabilities:	
* *Salute*	* *Istruzione*
* *Attività familiari*	* *Ambiente*
* *Relazioni sociali*	* *Tempo libero*

Poniamo che ogni indicatore sia dicotomico. Ad esso corrisponderà una distribuzione (di Bernoulli) così fatta:

Risorsa $X=\begin{cases} X=0 \text{ (sufficiente)} & X=1 \text{ (insuff)} \\ \text{Frequenza } 1-f & \text{Frequenza } f \end{cases}$

Sommando due indicatori per 100 persone (per esempio C=cibo e V=vesti, a parità di distribuzioni 'marginali' (cioè dei singoli indicatori semplici) otterrò tante possibili distribuzioni della variabile Somma quante le forme della distribuzione congiunta:

C \ V	0	1	
0	**70**	**0**	**70**
1	**10**	**20**	**30**
	80	**20**	**100**

$X=\begin{cases} 0 & 1 & 2 \\ 70 & 10 & 20 \end{cases}$

C \ V	0	1	
0	**50**	**20**	**70**
1	**30**	**0**	**30**
	80	**20**	**100**

$X=\begin{cases} 0 & 1 & 2 \\ 50 & 50 & - \end{cases}$

Marginali identiche, combinazioni variabili

Per capire il meccanismo della somma di variabili, replichiamo l'esempio. Supponiamo di aver testato le 100 persone per tutti e sei gli indicatori dicotomici di carenze materiali, e di avere costruito un indice M, che va da 0 a un massimo (potenziale) di 6 con questa distribuzione: 50 persone presentano M=0, 20 persone M=1, 20 M=2, le ultime 10 M=3 (nessuno assomma punteggi superiori).

Applicata alle stesse 100 persone la batteria delle *capabilities* si trova per l'indicatore composto C (tra 0 e 6) la stessa distribuzione di M, ci si domanda: **qual è la distribuzione dell'indice complessivo P di povertà, ottenuto sommando C e V?**

La risposta è: **dipende**. Dipende infatti dalla distribuzione congiunta (C,V) (e quindi dalla relazione che intercorre tra C e V). Vediamo due scenari possibili.

	0	1	2	3	
0	50				50
1		20			20
2			20		20
3				10	10
	50	20	20	10	100

$X_I = \begin{cases} 0 & 2 & 4 & 6 \\ 50 & 20 & 20 & 10 \end{cases}$

I° scenario

In questo caso bisogni materiali (M) e carenze di capabilities (C) vanno di pari passo: dove c'è l'uno c'è l'altra.

(segue)

Lezione 15
Elementi di analisi della varianza

In questa lezione ...

In questa lezione e nella prossima le parole "associare" ed "inferire" del titolo del volume trovano finalmente il loro punto d'incontro.

Nella sua versione di base lo scopo dell'analisi della varianza (ANOVA) è quello di studiare la dipendenza di una variabile quantitativa (Y) da una variabile categoriale (A). Le modalità della variabile categoriale (A1, A2, …, Ak) identificano dei gruppi. Si tratta allora di valutare se, in media, le differenze tra i gruppi - rispetto alla variabile Y - possono essere considerate statisticamente significative.

In questa lezione, in particolare, vedremo:

- Come condurre l'ANOVA (one-way)
- Le caratteristiche principali della v.c. F di Fisher-Snedecor
- Una riformulazione del modello utile per generalizzazioni

Confrontare gruppi

Per introdurre la logica dell'ANOVA, riprendiamo l'**esempio** della performance nel salto in alto.

Supponiamo di valor **valutare se una nuova tecnica di salto consente di ottenere miglioramenti significativi** rispetto a quella tradizionale.

I nostri atleti saranno quindi divisi in due gruppi. Un gruppo salterà con il metodo "standard" e l'altro con un nuovo metodo ("avanzato").

Indichiamo con **A1** il primo gruppo ed **A2** il secondo. Con **Y** la performance (salto in cm).

Un modo di procedere potrebbe essere quello di confrontare la media di Y nei due gruppi (μ_1 ed μ_2).

Salto (Y)	Metodo di salto
212	A1
218	A2
215	A1
218	A1
220	A2
218	A1
224	A2
220	A1
226	A2
229	A1

Ovvero, verificare se **H_0: $\mu 1=\mu 2=\mu$**
Contro l'ipotesi alternativa
H_1: H_0 non è vera

Il termine **ANOVA** è acronimo di *ANalysis Of Variance*, ed è stato introdotto nel 1918 da Ronald Fisher

Questione di varianze

L'ANOVA generalizza in ambito inferenziale la logica della scomposizione della varianza (lezione 4).

Supponiamo di voler testare l'efficacia di un nuovo farmaco sul livello di colesterolo. Un gruppo di pazienti lo trattiamo col nuovo farmaco, un gruppo con caratteristiche equivalenti non lo trattiamo.

L'**obiettivo** è di valutare quanta parte della variabilità di Y è da attribuire a differenza tra i gruppi (*varianza spiegata* dal trattamento) rispetto alla differenza interna ai gruppi (*varianza residua*).

Per comprendere meglio, facciamo un secondo esempio. Supponiamo che Y (variabile dipendente: explanandum) sia la velocità che la Ferrari raggiunge in un dato circuito. Si vuole testare l'efficacia di un nuovo tipo di gomme. Quindi il tipo di gomme è la variabile esplicativa (explanans). Vengono fatti 6 giri del circuito, 3 con le gomme usuali (gruppo A1) e 3 con le gomme nuove (gruppo A2).
Se in tutti i giri del gruppo A1 la velocità risulta pari a 330 km orari, ed in tutti i giri del gruppo A2 pari a 335, si ottiene una variabilità di Y nulla **all'interno** dei due gruppi e tutta da ricondurre alla differenza **tra** i due gruppi. Abbiamo quindi buone ragioni per considerare migliori le nuove gomme.
Se invece nel gruppo A1 avessimo osservato le velocità: 310, 330, 350 e nel gruppo A2: 325, 335, 345, nonostante le medie siano ancora rispettivamente 330 e 335, avremmo ora meno certezze sulla validità delle nuove gomme.

Il modello generale

Più in generale i gruppi sono k (j=1,2..k). Pertanto $H_0: \mu_1=\mu_2=\ldots=\mu_k=\mu$

Siano i **valori campionari** osservati:
Trattamento A1: $y_{11}, y_{12}, \ldots, y_{1n1}$
Trattamento A2: $y_{21}, y_{22}, \ldots, y_{2n1}$
...
Trattamento Ak: $y_{k1}, y_{k2}, \ldots, y_{kn1}$

Indichiamo le **medie campionarie** con $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$. Con $\bar{y}$ la media complessiva.

La distanza tra la generica osservazione *i* del gruppo *j* e la media totale può essere scomposta nel seguente modo:

$$(y_{ji} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ji} - \bar{y}_j)$$

Differenza tra media del gruppo j e media totale

Differenza tra osservazione y_{ij} e media del gruppo j

Passando poi alle sommatorie, si dimostra valere la seguente relazione:

$$\underbrace{\sum_{i,j}(y_{ij} - \bar{y})^2}_{SST} = \underbrace{\sum_{i,j}(\bar{y}_j - \bar{y})^2}_{SSM} + \underbrace{\sum_{i,j}(y_{ij} - \bar{y}_j)^2}_{SSE}$$

Due casi particolari

La logica della scomposizione della varianza l'avevamo trovata in ambito descrittivo nella lezione 4. Qui non ragioneremo con le varianze ma con le "devianze" (scostamenti al quadrato, in inglese "*Sum of Squares: SS*), che corrispondono al numeratore delle varianze.
La devianza totale (*Total:* **SST**), viene quindi scomposta in
- una parte "spiegata" dalla differenza tra i gruppi (*Model*: **SSM**),
- una parte residua (*Error*: **SSE**) interna ai gruppi.

Come si era visto nella lezione 4, partiamo da due casi limite

In ciascun gruppo le osservazioni hanno tutte stesso valore (pari alla media di gruppo), mentre le medie dei gruppi sono diverse tra di loro.
In tal caso SSE=0 ed **SST = SSM**. Quindi **la variabilità di Y è tutta riconducibile alle differenze tra gruppi**.

Le medie dei gruppi sono tutte uguali (e pari alla media totale), mentre all'interno di ciascun gruppo le osservazioni differiscono.
In tal caso SSM=0 ed **SST=SSE**. Quindi **le differenze tra gruppi non "spiegano" nulla della variabilità di Y**.

Test di significatività delle differenze tra gruppi

Dividendo le devianze per i corrispondenti gradi di libertà si ottengono le **stime delle varianze**:

Tra i gruppi ("spiegata"):
SSM / (k – 1)

Tanto più è elevata tanto più la varianza totale è spiegata dalla differenza tra i gruppi.

Illustre statistico

Entro i gruppi ("residua"):
SSE / (n – k)

Quanto più è elevata tanto meno la differenza tra i gruppi "spiega" la varianza totale.

Come **statistica test** utilizziamo il rapporto F. **Se è vera l'ipotesi nulla** (medie dei gruppi tutte eguali tra di loro e pari alla media totale) allora **tale rapporto si distribuisce come una v.c. $F_{g1,g2}$ di Fisher-Snedecor con k-1 ed n-k gradi di libertà**.

$$F = \frac{SSM/(k-1)}{SSE/(n-k)}$$

La v.c. F di Fisher-Snedecor

Uffa, una nuova distribuzione. Ma almeno è parente di quelle viste finora?

Proprio così. Cerchiamo allora di farne la conoscenza. Siano $S_1 \sim \chi^2_{g1}$ e $S_2 \sim \chi^2_{g2}$ due v.c. Chi-quadrato indipendenti con gradi di libertà rispettivamente g_1 e g_2. Si definisce v.c. F di Fisher-Snedecor il seguente rapporto:

$$F_{g1,g2} = [S_1/(g_1)] / [S_2/(g_2)]$$

La funzione di densità della v.c. F è asimmetrica e unimodale. Converge ad una v.c. Normale se sia g_1 che g_2 tendono ad infinito. Converge al Chi-quadrato fissando uno dei due gradi di libertà e facendo tendere l'altro all'infinito.

Una volta calcolato il valore della statistica test (F_{oss}) si utilizza la funzione di densità della v.c. $F_{g1,g2}$ per ricavare la probabilità di ottenere, sotto H_0, un valore del test ancora più estremo (per tale probabilità, chiamata **p-value**, useremo la notazione Pr>F). Il criterio generale è quello di rigettare l'ipotesi nulla quando p-value ha valore molto basso (ad es. <0,05)

Tabulazione della v.c. F

La "**tabulazione**" dei valori della v.c. $F_{g1,g2}$ è più ardua rispetto alle distribuzioni già viste, perché entrano in gioco ben due gradi di libertà. Una soluzione è quella di prefissare uno specifico livello ($\alpha = Pr > F_\alpha$) di significatività e tabulare i valori F_α corrispondenti. I valori della tabella sottostante corrispondono ad α **= 0,05** (di fatto i più utilizzati).
Ad **esempio**, per $g_1=1$ e $g_2=8$, si ha $F_{0,05}=5,32$. Al livello di significatività del 5% rigetteremo l'ipotesi nulla se la statistica test ci fornisce un valore $F_{oss} \geq 5,32$; propenderemo invece per l'ipotesi nulla se $F_{oss} < 5,32$.

$g_2 \backslash g_1$	1	2	3	4	5	6	7	8	9	10	20	50	100	1000
1	161	199	216	225	230	234	237	239	241	242	248	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.58	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.70	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.44	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.75	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.32	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.02	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.80	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.64	2.59	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	1.97	1.91	1.85
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.78	1.60	1.52	1.45
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.68	1.48	1.39	1.30
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.58	1.36	1.26	1.11

Oramai tabelle come questa sono in larga parte superate. Persino i fogli di calcolo più comuni contengono la distribuzione F e sono quindi in grado di fornirvi l'esatto valore Pr>F che corrisponde alla vostra statistica test.

La tabella dell'ANOVA

I risultati dell'Anova vengono tipicamente presentati secondo lo schema seguente (proposto da Fisher nel 1925).

Fonte di variazione	*gradi di libertà*	*devianza*	*varianza stimata*	*valore del test F*
Tra i gruppi	k-1	SSM	SSM/(k-1)	[SSM/(k-1)]/[SSE/(n-k)]
Entro i gruppi	n-k	SSE	SSE/(n-k)	
Totale	n-1	SST		

Ora, finalmente, abbiamo tutto ciò che ci serve per la soluzione dell'**esempio** sulla performance nel **salto in alto**. Otteniamo:

Fonte di variazione	*gradi di libertà*	*devianza*	*varianza stimata*	*valore del test F*
Tra i gruppi	1	26,7	26,7	1,03
Entro i gruppi	8	207,3	25,9	
Totale	9	233,0		Pr>F = 0,34

Come si vede **la varianza tra gruppi ha lo stesso ordine di grandezza della varianza interna** (F vicino ad 1). La differenza tra le medie di gruppo può quindi essere semplicemente casuale. I dati osservati non forniscono pertanto sufficiente evidenza empirica per affermare che il metodo "avanzato" è migliore.

Stesse medie di gruppo ma minor dispersione interna

Nell'esempio appena visto
i **dati di base** erano i seguenti.
Metodo tradizionale:
212; 215; 218; 218; 220; 229
Metodo avanzato:
218; 220; 224; 226
Le **medie campionarie**
erano rispettivamente 218,7 e 222.

In base a tali dati l'ipotesi nulla che le vere medie di gruppo siano in realtà uguali tra di loro (e quindi le medie campionarie siano diverse solo come conseguenza dell'incertezza campionaria) non è stata rigettata.

Cosa sarebbe successo se invece i dati fossero stati i seguenti?
Metodo tradizionale: 218; 218; 218; 218; 220; 220
Metodo avanzato: 220; 220; 224; 224

Le medie campionarie di gruppo sono ancora 218,7 e 222.

Rispetto al caso precedente la varianza interna ai gruppi è però ora molto più bassa: SSE/8 = 2,7 mentre quella tra gruppi è ancora SSM/1 = 26,7.
Risulta quindi F=10 a cui corrisponde un p-value (Pr>F) pari a 0,013.

La differenza tra le medie può ora essere considerata significativa (quindi rifiutiamo H_0, il metodo avanzato può essere considerato migliore).

Un secondo esempio

Vediamo un altro esempio utilizzando dati tratti da un'indagine (telefonica) sulle famiglie con bambini tra i 5 ed i 14 anni. Il campione (casuale semplice) è costituito da n=300 unità.

Vogliamo verificare se esiste una relazione tra tempo trascorso davanti alla televisione dai bambini e titolo di studio del padre. Distinguiamo il livello di istruzione in tre categorie: alto, medio e basso. I gruppi sono quindi ora k=3.

Le medie campionarie dei tre gruppi risultano: 2,1 per il titolo alto; 3,15 per il titolo medio; 4 per il titolo basso. Possiamo considerare tali differenze statisticamente significative?

La risposta, che affidiamo all'ANOVA, è affermativa.

Fonte di variazione	*gradi di libertà*	*devianza*	*varianza stimata*	*valore del test F*
Tra i gruppi	2	159,65	79,825	53,1
Entro i gruppi	297	446,68	1,504	
Totale	299	606,33		

Il p-value (Pr>F) risulta inferiore a 0,0001

Un esempio più articolato

Per valutare se il proprio metodo di apprendimento interattivo *on-line* della lingua inglese per italiani è migliore rispetto a quello di una scuola concorrente, la Virtual Oxford Int. organizza il seguente esperimento.
Seleziona un gruppo (A1) di 10 propri studenti appena iscritti. Recluta poi altre 10 persone (A1) interessate ad imparare l'inglese, con caratteristiche analoghe (in termini di sesso, età e livello di istruzione) al gruppo A1.

Prima di applicare il metodo di insegnamento viene valutato il livello di conoscenza di base della lingua attraverso una prova scritta ed orale (pre-test). I punteggi acquisiti nel pre-test sono sono riportati nella tabella a fianco (seconda colonna).
Viene condotta una prima ANOVA su tali punteggi per verificare che il livello di partenza sia lo stesso nei due gruppi.

Group	Pre	Post
A1	3.1	6.2
A1	2.7	5.3
A1	4.1	8.1
A1	6.2	8.4
A1	5.5	7.9
A1	5.1	8.2
A1	3.4	6.4
A1	4.7	6.2
A1	4.3	5.8
A1	1.8	6.2
A2	2.3	3.1
A2	3.6	4.2
A2	4.8	5.4
A2	7.2	7.8
A2	3.3	5.2
A2	2.5	3.2
A2	5.2	5.8
A2	3.2	3.6
A2	5.4	5.8
A2	4.5	4.4

Fonte di variazione	*gradi di libertà*	*devianza*	*varianza stimata*	*Valore del test F*
Tra i gruppi	1	0,06	0,06	0,03
Entro i gruppi	18	37,07	2,06	
Totale	19	37,13		

Pr>F = 0,87

segue

La media sul punteggio del pre-test è pari a 4,09 per il gruppo A1 e a 4,2 per il gruppo A2. Tale differenza non è significativa. Il p-value ottenuto dall'ANOVA risulta infatti molto elevato. Possiamo pertanto considerare analogo il livello di conoscenza di partenza della lingua inglese nei due gruppi.

Verificata la parità di condizioni iniziali, al gruppo A1 viene applicato il metodo di insegnamento *on-line*, mentre le persone del gruppo A2 vengono iscritte alla scuola concorrente CEPUS. Dopo 6 mesi di apprendimento viene somministrata una nuova prova di valutazione del livello raggiunto (colonna "post" nella tabella precedente). La media sui punteggi finali risulta 6,87 per A1 e 4,85 per A2.

I risultati dell'ANOVA (riportati qui sotto) ci dicono che ora la differenza tra i due gruppi è significativa. Il metodo *on-line* sembra quindi dare risultati migliori (quantomeno rispetto alla scuola CEPUS).

Fonte di variazione	*gradi di libertà*	*devianza*	*varianza stimata*	*Valore del test F*
Tra i gruppi	1	20,402	20,402	11,9
Entro i gruppi	18	30,766	1,709	
Totale	19	51,168		

Pr>F = 0,0028

Una riformulazione del modello

La seguente relazione vista sui dati campionari:
$(y_{ji} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ji} - \bar{y}_j)$

Riflette la seguente a livello di popolazione:
$(y_{ji} - \mu) = (\mu_j - \mu) + (y_{ji} - \mu_j)$

Da cui otteniamo: $\mathbf{y_{ji} = \mu + (\mu_j - \mu) + (y_{ji} - \mu_j) = \mu + \beta_j + \varepsilon_{ij}}$

I parametri β_j misurano la differenza tra medie, mentre ε_{ij} è l'"errore casuale" e coglie la variabilità residua di Y (quella interna ai gruppi). Da questa specificazione del modello dell'ANOVA si posso apprezzare ancor meglio i due casi estremi:

se la variabilità interna ai gruppi è nulla risulta: $\mathbf{y_{ji} = \mu + \beta_j}$
ovvero gli scostamenti di Y dalla propria media μ (su cui si fonda la varianza totale) sono unicamente da ricondurre alle differenze tra gruppi.

se le differenze tra gruppi sono nulle si ha: $\mathbf{y_{ji} = \mu + \varepsilon_{ij}}$
ovvero gli scostamenti di Y dalla propria media μ sono unicamente da ricondurre alla componente di variabilità non spiegata.

Il modello lineare

L'ipotesi nulla del test F era: $\mathbf{H_0: \mu_1 = \mu_2 = \ldots = \mu_k = \mu}$

Con il modello $y_{ji} = \mu + \beta_j + \varepsilon_{ij}$ l'ipotesi nulla del test F diventa:

$$\mathbf{H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

Come vedremo nel prossimo capitolo, tale formulazione troverà diretta corrispondenza nel modello di regressione. E' vero che in questa lezione abbiamo considerato l'ANOVA ad un solo fattore (one-way), ovvero con un'unica varia-bile categoriale (esplicativa) le cui modalità corrispondono ai gruppi con diverso trattamento. Ma è intuibile che il modello può essere generalizzato al caso di più variabili esplicative (multi-way). A sua volta l'ANOVA multi-way trova corri-spondenza nella regressione multipla (lezione 8). In tutti questi casi comunque

sono due le ipotesi fondamentali dell'ANOVA:

1. **le osservazioni sono indipendenti tra loro e seguono una distribuzione normale;**
2. **la varianza è costante nei vari gruppi (corrisponde all'ipotesi di omoscedasticità del modello di regressione).**

Lezione 16
Inferenza e regressione

In questa lezione ...

In questa lezione ci agganceremo direttamente agli argomenti trattati nella lezione 6, dove è stato introdotto il modello di regressione lineare semplice in ambito descrittivo. Ci eravamo cioè posti nella situazione in cui si dispone dei dati su tutta la popolazione di interesse. Ma spesso i dati provengono da un campione. Come la media campionaria è una stima della media della popolazione, allo stesso modo i parametri della regressione calcolati sui dati campionari saranno una stima (affetta da errore) dei parametri riferiti a tutta la popolazione. In questa lezione:

- Vedremo anzitutto **come è costruito il modello**, su quali ipotesi si poggia, come ottenere stime dei parametri del modello e come misurare l'errore di stima sui parametri
- Presenteremo i principali **test** sulla significatività dei parametri e sull'adattamento complessivo del modello
- Infine estenderemo gli strumenti al modello di **regressione multipla**, riprendendo il concetto di relazione spuria e distinguendo le operazioni mentali richieste a fini esplicativi e a fini previsivi.

Ripartiamo dalla variabilità

Supponiamo che le n osservazioni della variabile di interesse (Y) siano estratte (tramite campionamento casuale semplice) da una popolazione normale con media μ e varianza σ^2 .

$Y_i \sim N(\mu, \sigma^2)$

Dove, in particolare $E(Y_i) = \mu$

Equivalentemente, potremmo anche considerare il seguente semplice modello:

$\mathbf{Y_i = \mu + \varepsilon_i}$
$\varepsilon_i \sim N(0, \sigma^2)$

Vale ancora, infatti:
$E(Y_i) = \mu + E(\varepsilon_i) = \mu$

*Abbiamo in tal modo scomposto Y_i in una componente "di fondo" (comune alle varie unità i: **μ**) ed una componente "residuale" (specifica di ciascuna unità i: $\boldsymbol{\varepsilon_i}$).*

Le intuizioni alla base del modello di regressione si devono a F. Galton (1822-1911), primo cugino di Darwin. Oltre al termine regressione, si deve a lui anche quello di "anticiclone".

L'errore di previsione

Se μ è il parametro di interesse che vogliamo misurare, allora

ε_i rappresenta l'errore che commettiamo nel prendere la singola osservazione Y_i come misura dell'ignoto parametro di interesse:

$\varepsilon_i = Y_i - \mu$ (componente residuale o errore casuale).

L'interpretazione può anche essere ribaltata. Possiamo infatti affermare che, in assenza di altre informazioni, il valor medio è la miglior previsione di Y (vedi lez. 6 del 1° volume). Quindi:

ε_i può anche essere interpretato come l'errore che si commette nel prevedere Y_i attraverso μ.

Ad esempio. Se un mio amico ha preso 22 all'esame di statistica ma so che il voto che mediamente si prende a tale esame è 28, mi aspetto più facilmente di prendere 28 che 22. Ovvero conta più la media che l'esperienza del mio amico. In assenza di altre informazioni sul fenomeno la media è quindi il miglior predittore.

Alla ricerca di predittori efficaci

La media mi sembra un po' poco come predittore. Possiamo fare di meglio?

Se so che in media gli studenti studiano un mese e mezzo per preparare l'esame, sono autorizzato a ritenere che se studio due mesi (a parità di tutto il resto) prenderò un voto più alto di 28?

Per rispondere a domande di questo tipo dobbiamo far fare un salto di qualità al nostro modello. Ovvero dobbiamo inserire l'informazione disponibile su una seconda variabile (X) potenzialmente correlata alla nostra variabile di studio (Y).

Se Y ed X sono correlate possiamo sfruttare la conoscenza di X per meglio prevedere Y.

X $\xrightarrow{?}$ Y

Il modello lineare e sue ipotesi

Facendo scendere in campo X possiamo arricchire il modello nel seguente modo:

$\mathbf{Y_i = \mu_i + \varepsilon_i}$

$\mu_i = f(X_i)$ *componente sistematica*

$\varepsilon_i \sim N(0, \sigma^2)$ *errore casuale*

La componente sistematica coglie la dipendenza di Y da X, mentre la componente casuale rappresenta la variabilità di Y che rimane "non spiegata" da X.

Più sinteticamente il modello può anche essere scritto nel seguente, più familiare, modo:

$$\mathbf{Y_i = \alpha + \beta X_i + \varepsilon_i}$$

ove si è posto $f(X_i) = \alpha + \beta X_i$, ovvero si è assunto che la dipendenza di Y da X sia di tipo lineare.

Riassumiamo le **ipotesi** alla base del modello:

1. $E(\varepsilon_i) = 0$ (gli errori casuali oscillano attorno allo 0)
2. $Var(\varepsilon_i) = \sigma^2$ per ogni *i* (ipotesi di omoscedasticità)
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ (gli errori sono incorrelati tra di loro)
4. ε_i hanno distribuzione normale

Inoltre i valori x_i sono considerati fissi e misurati con precisione.

Stima dei parametri

I parametri α e β si riferiscono alla popolazione.
Con i dati campionari, ricorrendo al criterio dei minimi quadrati (fatte salve le precedenti ipotesi da 1 a 3), possiamo ottenere i seguenti **stimatori** (il teorema di Gauss-Markov ci garantisce essere **corretti e di minima varianza** nella classe degli stimatori lineari non distorti):

$$a = \bar{y} - b\bar{x}$$

$$b = \mathrm{cov}_{YX} / \mathrm{var}_X = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Queste formule coincidono nella sostanza con quelle già viste nella regressione descrittiva (lezione 6).
La differenza è che qui quello che otteniamo non sono i parametri veri della popolazione ma loro stime.

Stimati i parametri, i valori previsti dal modello saranno: $\hat{y} = a + bx$

Scostamenti (residui) tra osservazioni e modello: $e_i = y_i - (a + bx_i) = y_i - \hat{y}_i$

Inoltre, stimatore di σ^2 : $s^2 = \sum e_i^{\,2} / (n-2) = \sum (y_i - \hat{y}_i)^2 / (n-2)$

Un esempio già esplorato

Riprendiamo l'esempio visto nella lezione in cui è stata introdotta la regressione in ambito descrittivo.
I dati si riferivano a 10 atleti e l'interesse era quello di valutare la relazione tra età e performance nel salto in alto.
Supponiamo ora che tali 10 atleti non siano tutta la nostra popolazione di interesse, ma ne costituiscano solamente un suo campione casuale semplice.

X (Età)	Salto (Y)
18	212
18	218
18	215
19	218
19	220
20	218
20	224
21	220
21	226
22	229

Con le formule appena viste otteniamo:
b=3,04 ; a=160,35 ; s^2 = 7,95

Dal punto di vista tecnico finora, dunque, nulla di nuovo.

Dobbiamo però ancora trattare la parte relativa all'inferenza sui parametri…

I valori sono gli stessi già calcolati nella lezione 6. Ora però tali valori vanno intesi come le stime campionarie dei parametri (ignoti) della popolazione α e β.

(SEGUE)