

# Introduzione all'Analisi della Varianza (ANOVA)

Marcello Gallucci

[marcello.gallucci@unimib.it](mailto:marcello.gallucci@unimib.it)

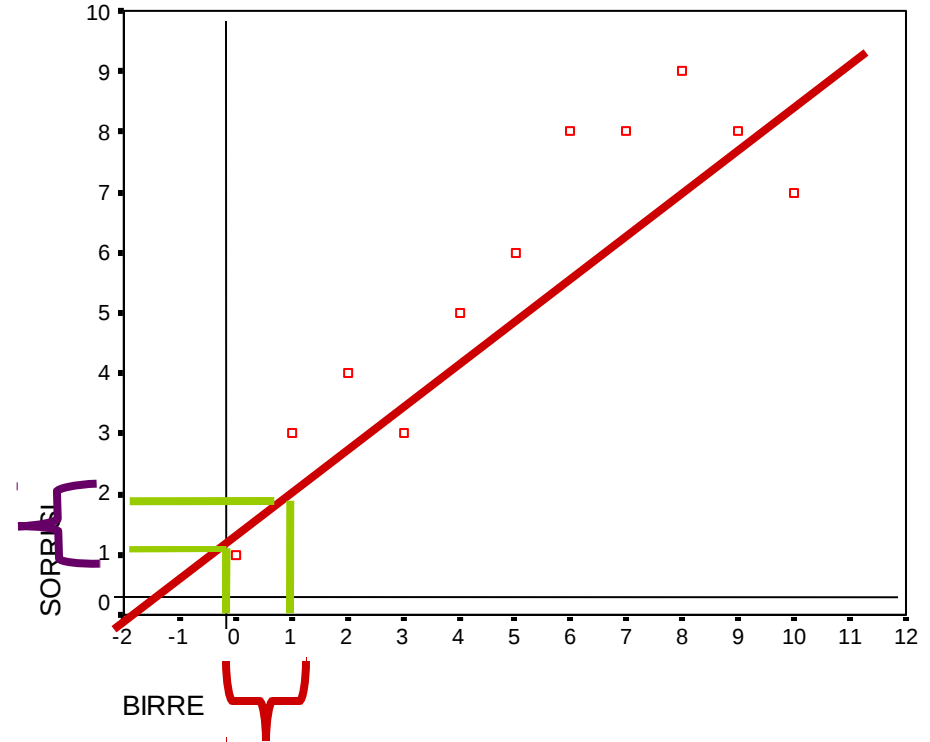
AMD

# Variabili nella Regressione

- Nella regressione, la variabile dipendente è sempre quantitativa e, per quello che abbiamo visto, anche le variabili indipendenti

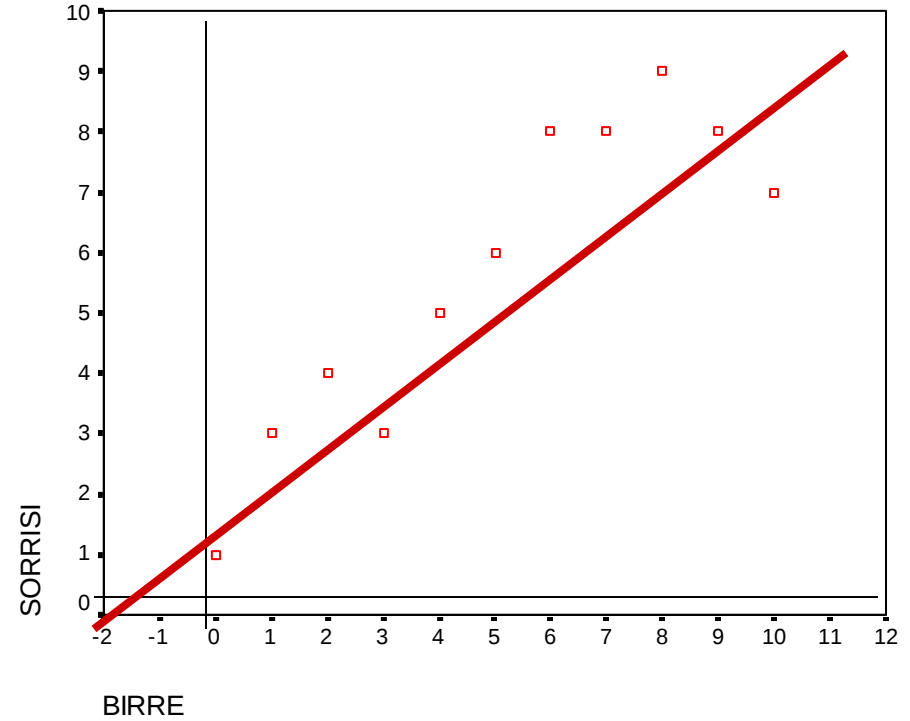
Variabili quantitative

$$\hat{y}_i = a + b_{yx} \cdot x_i$$



# Regressione NS

- La retta di regressione rappresenta la predizione lineare (o dipendenza lineare) tra una variabile indipendente ed una dipendente, espressa nelle unità di misura originali



# Variabili Nominali (categoriche)

- Nelle variabili nominali i valori indicati sono delle qualità, cioè sono equivalenti a delle etichette

N di birre      Regione di Provenienza

Punteggi

Aumentare di una unità significa solo cambiare quantità

Media, varianza non hanno senso

3	nord
4	nord
2	nord
1	nord
3	centro
4	centro
6	centro
7	centro
2	nord
3	nord

# Variabili Nominali

- Esempi di variabili nominali:
  - Sesso (Donne=1 Uomini=2)
  - Nazionalità (es. Francese=1, Tedesco=2 e Italiano=3)
  - Facoltà (es. Psicologia=1, Informatica=2, Medicina=3)
  - Gruppi sperimentali (es. Trattamento=1, Placebo=2)

In generale, le variabili nominali indicano in quale gruppo il soggetto appartiene

# Variabili Nominali

- Nelle variabili nominali, l'unità di misura non ha la stessa interpretazione ad ogni livello della scala di misura
- Passare da Italiano a Francese, non è la stessa cosa che passare da Francese a Tedesco
- Qualunque numero che distingue i gruppi rappresenta la stessa qualità

In generale, i valori numerici dati ai soggetti sono arbitrari

# Esperimento

- Variabili nominali si trovano spesso negli esperimenti
  - Nel seguente esperimento testiamo l'effetto di un ancoraggio cognitivo sulla stima delle quantità numeriche:

- Domanda 1 a tutti i soggetti: Secondo te, le nazioni africane alle nazioni unite sono più o meno del X %.

Ancora

Var.  
Dipendente

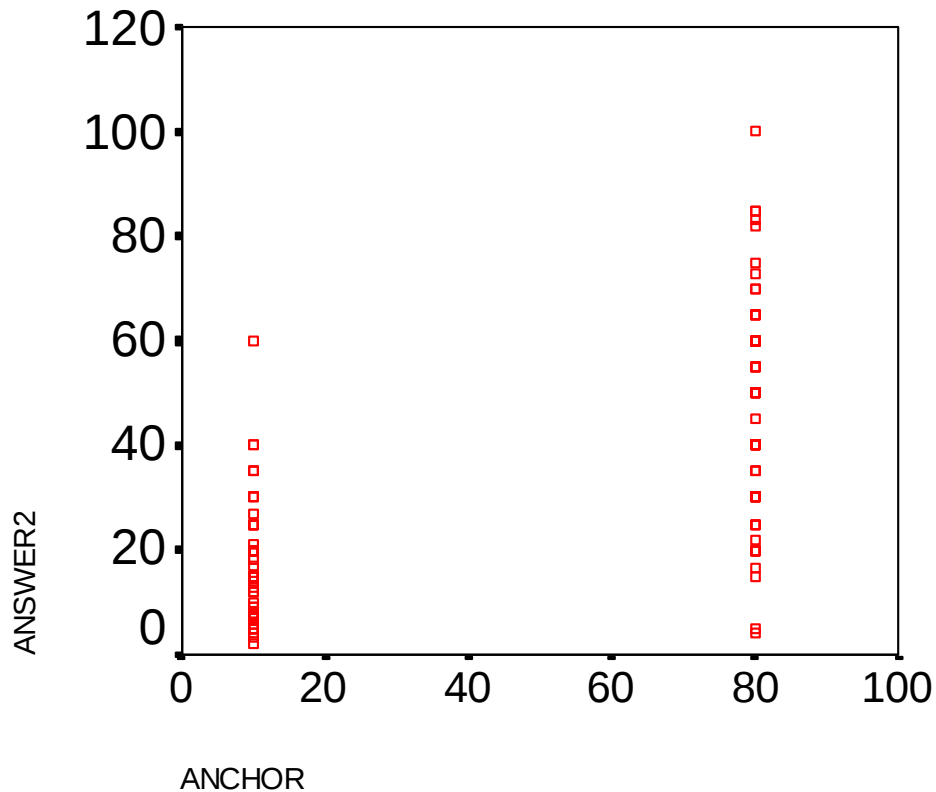
- Domanda 2: Quante sono le nazioni africane in percentuale alle nazioni unite

- Gruppo 1: ancora 10%. Gruppo 2: ancora 80%

Condizione  
sperimentale

# Esperimento: Risultati

- Quesito sperimentale: L'ancoraggio ha un effetto significativo sulla stima numerica?



## Report

ANSWER2

ANCHOR	Media	N	Deviazione std.
10	17,0447	85	10,08970
80	44,8288	85	19,68864
Totale	30,9368	170	20,91424

C'e' un differenza significativa?

Quanto spieghiamo con tale differenza?



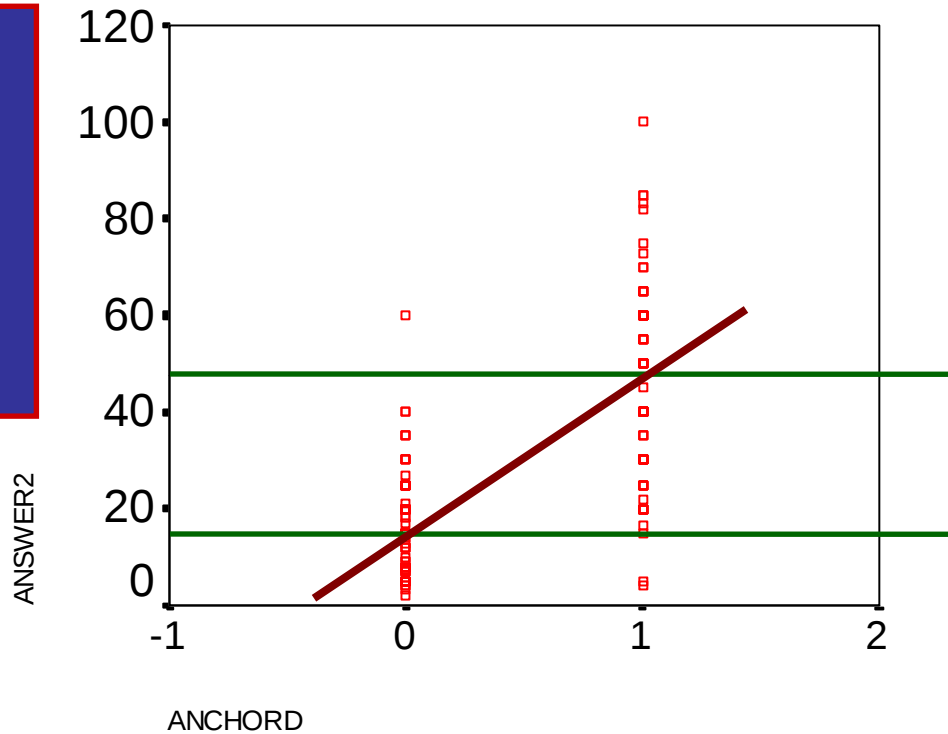
# Regressione

- Partiamo da ciò che sappiamo: Regressione

Codifichiamo la variabile indipendente come

Gruppo ancora 10%  $X=0$

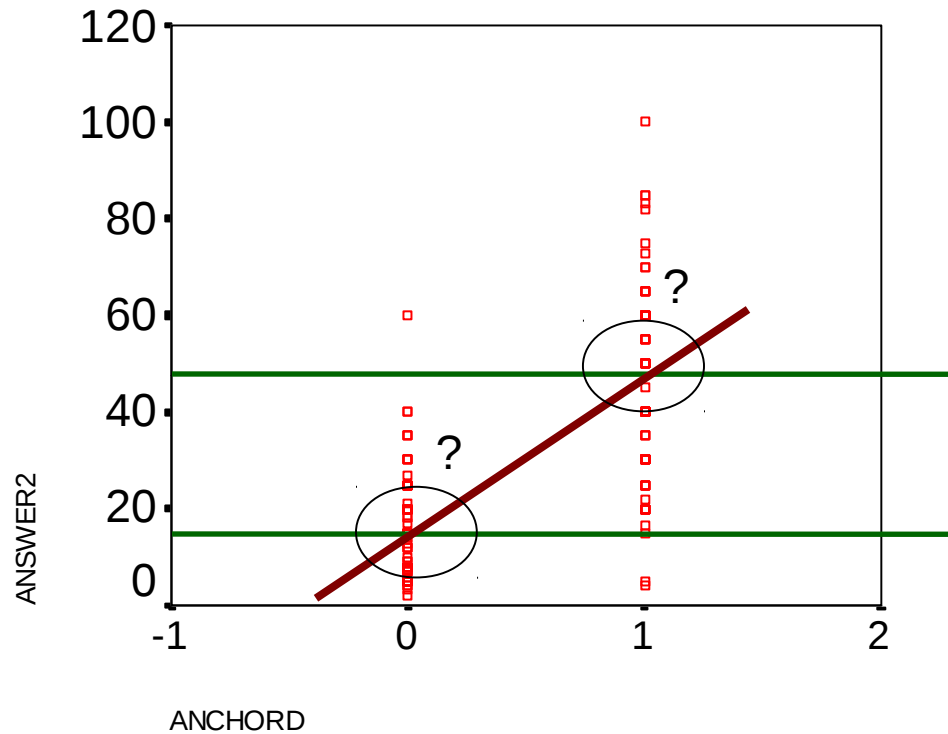
Gruppo ancora 80%  $X=1$



- Stimiamo la regressione

# Regressione

- Dove passerà la regressione? Quali valori di Y saranno indicati dalla regressione

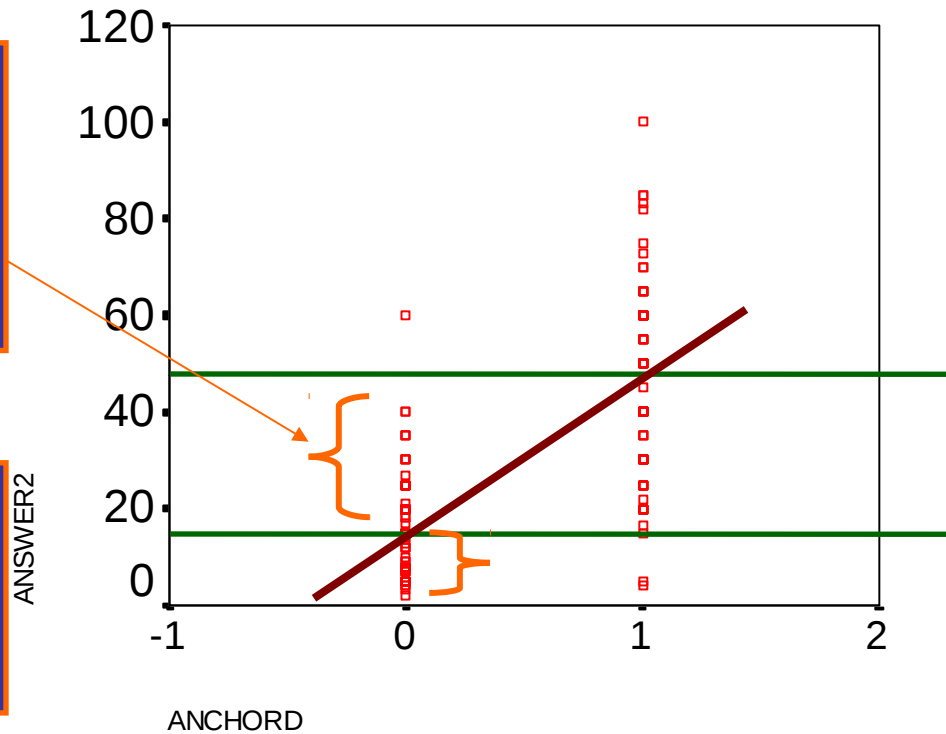


# Regressione

- Dove passerà la regressione? Quali valori di Y saranno indicati dalla regressione

Ricordiamo che la regressione è la retta che minimizza la distanza tra i punti osservati e quelli stimati

Ricordiamo che in una distribuzione, la media è il valore che minimizza la distanza tra i punteggi osservati e se stessa

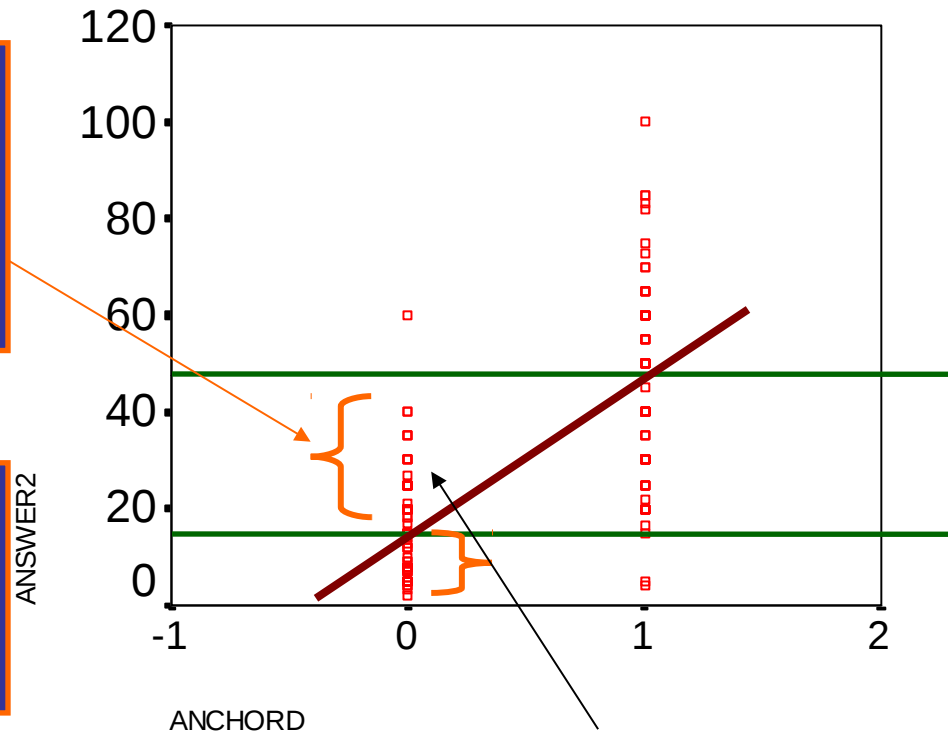


# Regressione

- Dove passerà la regressione? Quali valori di Y saranno indicati dalla regressione

Ricordiamo che la regressione è la retta che minimizza la distanza tra i punti osservati e quelli stimati

Ricordiamo che in una distribuzione, la media è il valore che minimizza la distanza tra I punti osservati e se stessa



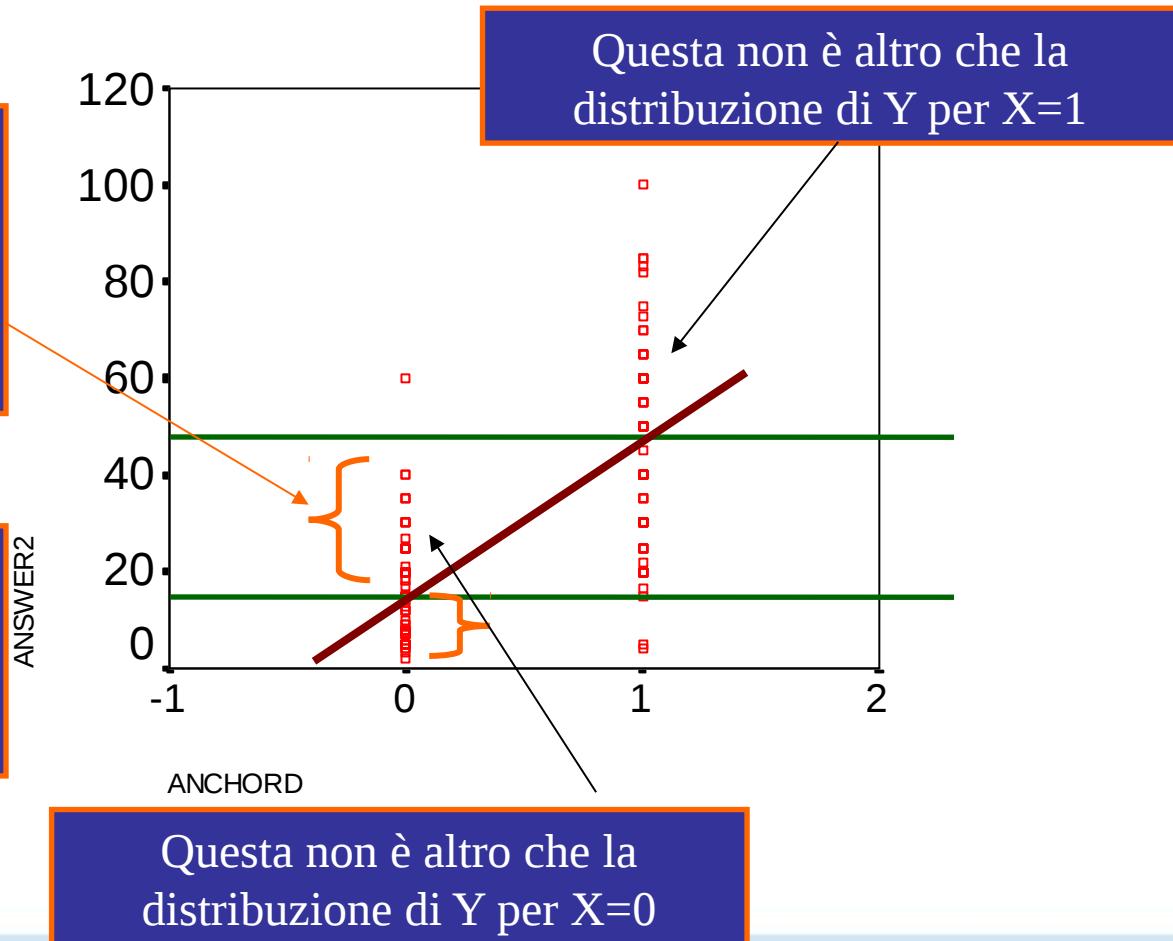
Questa non è altro che la distribuzione di Y per  $X=0$

# Regressione

- Dove passerà la regressione? Quali valori di Y saranno indicati dalla regressione

Ricordiamo che la regressione è la retta che minimizza la distanza tra i punti osservati e quelli stimati

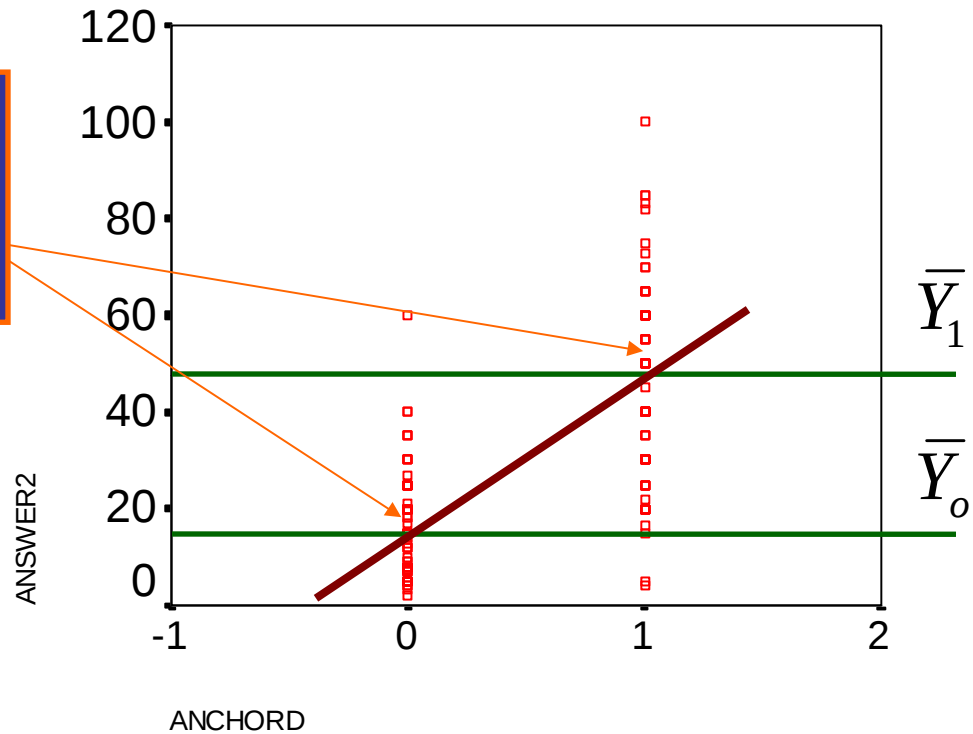
Ricordiamo che in una distribuzione, la media è il valore che minimizza la distanza tra i punti osservati e se stessa



# Regressione

- Dove passerà la regressione? Quali valori di Y saranno indicati dalla regressione

Dunque la regressione passerà necessariamente per la media di Y per il gruppo X=0 e per la media di Y per il gruppo X=1



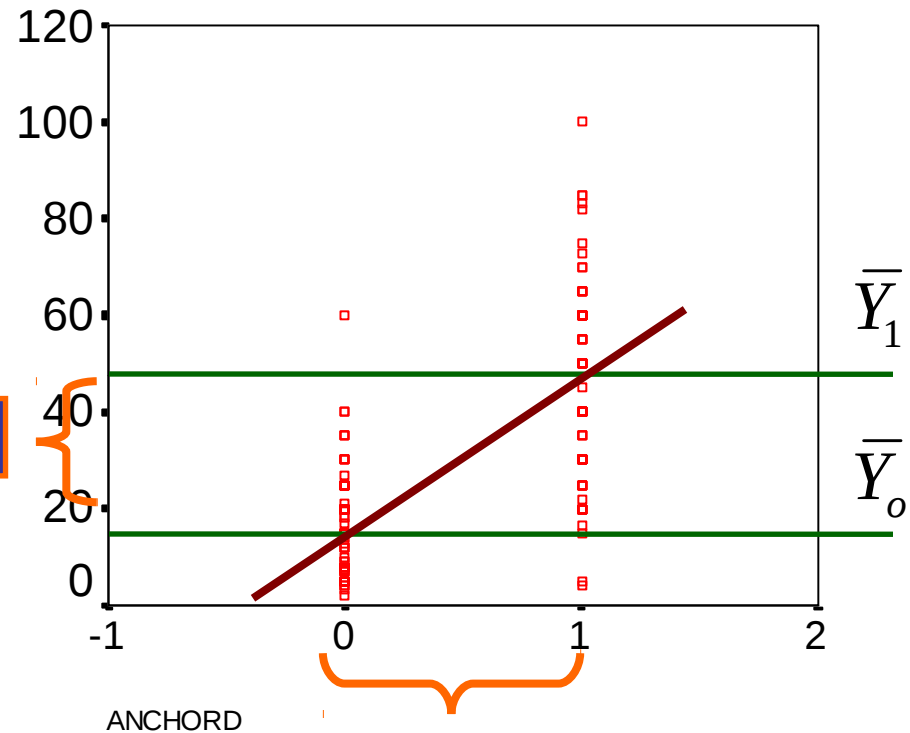
# Coefficiente di Regressione

- Quale valore indicherà il coefficiente di regressione?

Ricordiamo che B indica il cambiamento atteso in Y al variare di X di una unità

Cambiamento in Y

ANSWER2



Aumento X di una unità

# Coefficiente di Regressione

- Quale valore indicherà il coefficiente di regressione?

$$Y = a + b \cdot X$$

X=0 cioè gruppo 10%

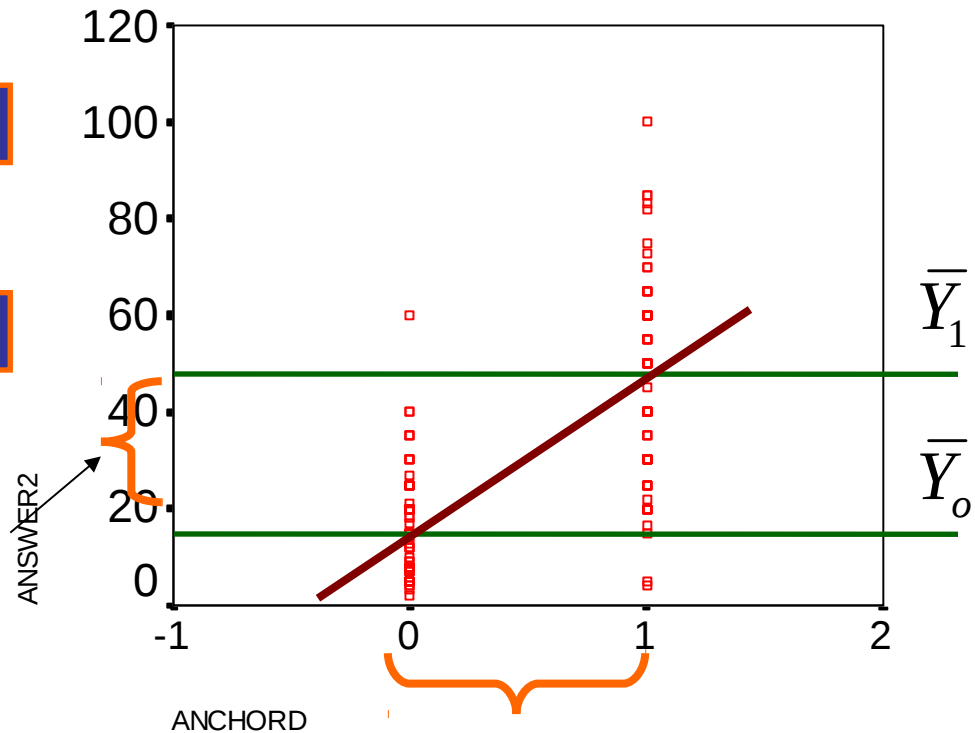
$$Y = a + B \cdot 0 = a = \bar{Y}_0$$

X=01 cioè gruppo 80%

$$Y = \bar{Y}_0 + B \cdot 1 = \bar{Y}_1$$

Differenza

$$B = \bar{Y}_1 - \bar{Y}_0$$



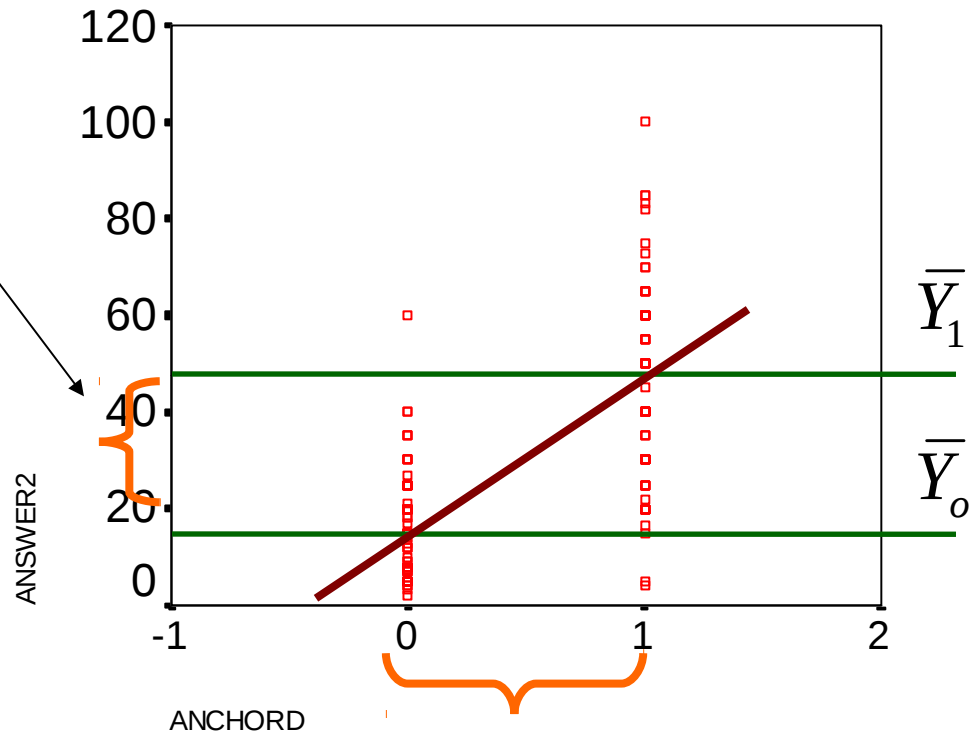


# Coefficiente di Regressione

- Se la variabile indipendente è dicotomica, il coefficiente ci indica la differenza tra i gruppi

Differenza tra le medie dei gruppi

$$B = \bar{Y}_1 - \bar{Y}_0$$



# Inferenza statistica

- Se il coefficiente è significativo (possiamo rifiutare l'ipotesi nulla  $b=0$ ), allora diremo che esiste una differenza significativa tra i gruppi

## Report

ANSWER2

ANCHOR	Media	N	Deviazione std.
10	17,0447	85	10,08970
80	44,8288	85	19,68864
Totale	30,9368	170	20,91424

**Test inferenziale**

## Coefficienti<sup>a</sup>

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	17,045	1,697		10,045	,000
	ANCHOR	27,784	2,400	,666	11,579	,000

a. Variabile dipendente: ANSWER2

**Significatività**

# Inferenza statistica

- Concluderemo che esiste un effetto dell'ancora sulle stime delle quantità

## Report

ANSWER2

ANCHOR	Media	N	Deviazione std.
10	17,0447	85	10,08970
80	44,8288	85	19,68864
Totale	30,9368	170	20,91424

C'è un differenza significativa tra le medie dei gruppi

Test inferenziale

## Coefficienti<sup>a</sup>

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	17,045	1,697		10,045	,000
	ANCHOR	27,784	2,400	,666	11,579	,000

a. Variabile dipendente: ANSWER2

Significatività

# Test sulla varianza

- Guardiamo ora il test fatto sulla varianza spiegata

Riepilogo del modello				
Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,666 <sup>a</sup>	,444	,441	15,64360

a. Stimatori: (Costante), ANCHOR

		ANOVA <sup>b</sup>				
Modello		quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	32808,181	1	32808,181	134,063	,000 <sup>a</sup>
	Residuo	41113,332	168	244,722		
	Totale	73921,513	169			

a. Stimatori: (Costante), ANCHOR  
b. Variabile dipendente: ANSWER2

**Test inferenziale**

ANOVA<sup>b</sup>

Significatività

Ricordiamo il significato della F?

Significatività

# Test per il coefficiente $R^2$

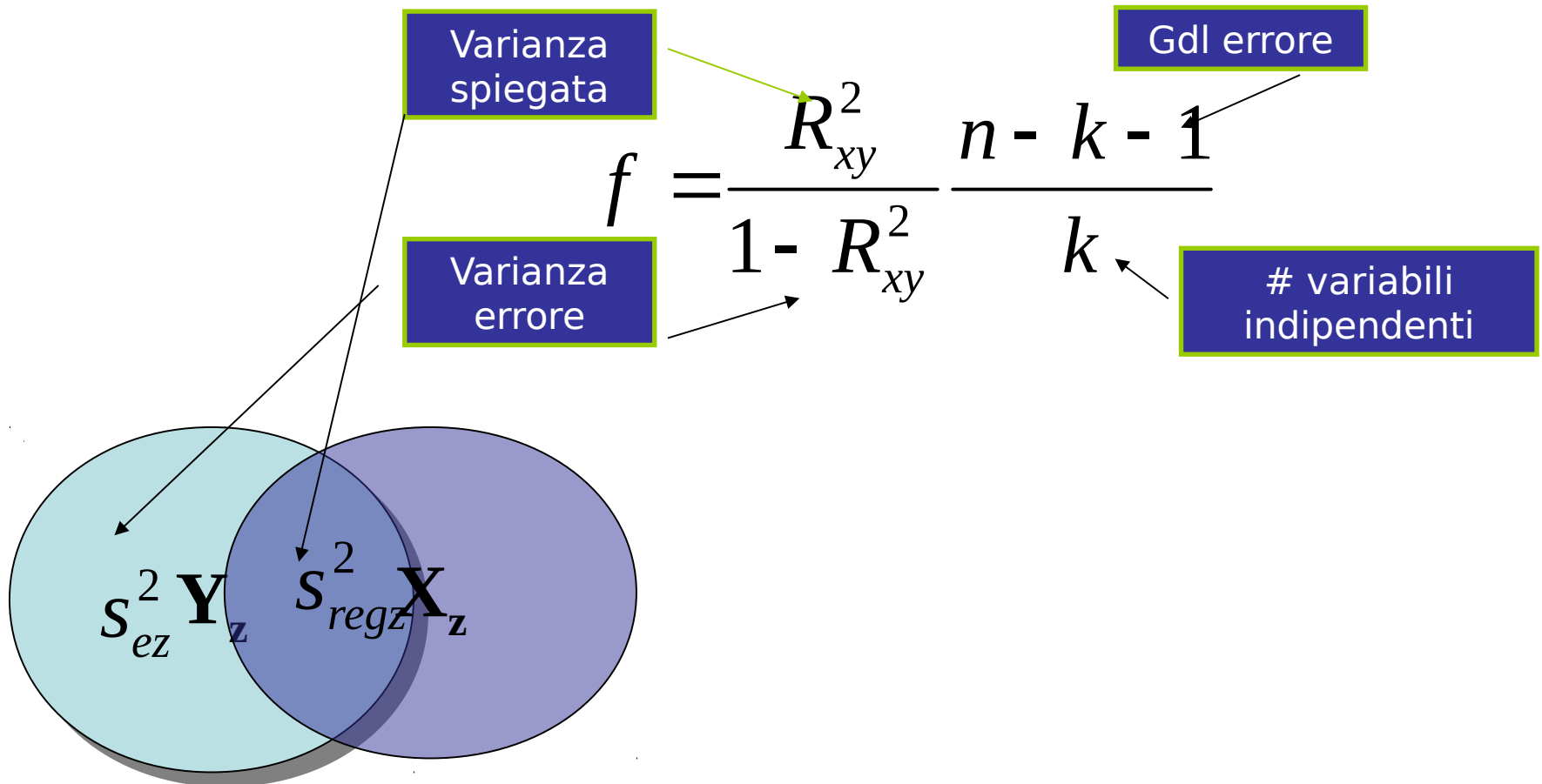
- Come si ottiene la F

$$F = \frac{\text{Varianza spiegata} \cdot R_{xy}^2}{\text{Varianza errore} \cdot (1 - R_{xy}^2)} \cdot \frac{\text{Gdl errore} \cdot (n - k - 1)}{k \cdot \text{\# variabili indipendenti}}$$

Il rapporto fra varianza spiegata e varianza di errore (moltiplicato per il rapporto fra gradi di libertà) si distribuisce secondo una distribuzione F con k e n-k-1 gdl

# Test per il coefficiente $R^2$

- Come si ottiene la F



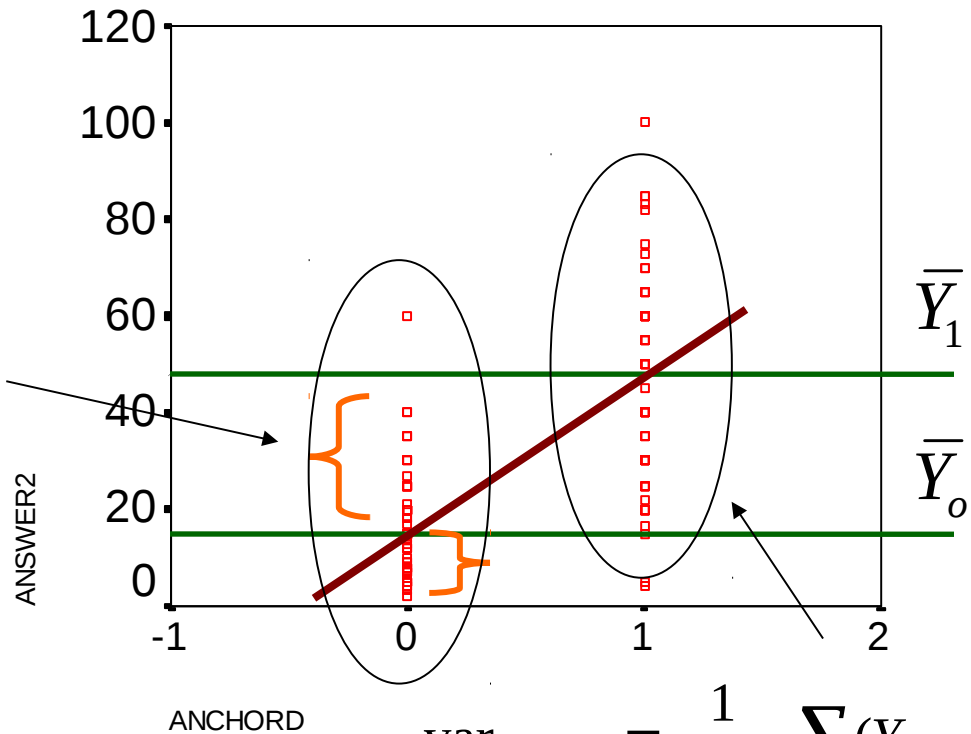
# Rapporto varianza

- Se la variabile indipendente è dicotomica, la varianza di errore sarà la somma delle varianze dentro i gruppi

Varianza non catturata dalla regressione

$$\text{var}_{\text{gruppo0}} = \frac{1}{n_0 - 1} \sum (Y_{0i} - \bar{Y}_0)^2$$

Varianza dentro i gruppi  
(within groups)



$$\text{var}_{\text{gruppo2}} = \frac{1}{n_1 - 1} \sum (Y_{1i} - \bar{Y}_1)^2$$

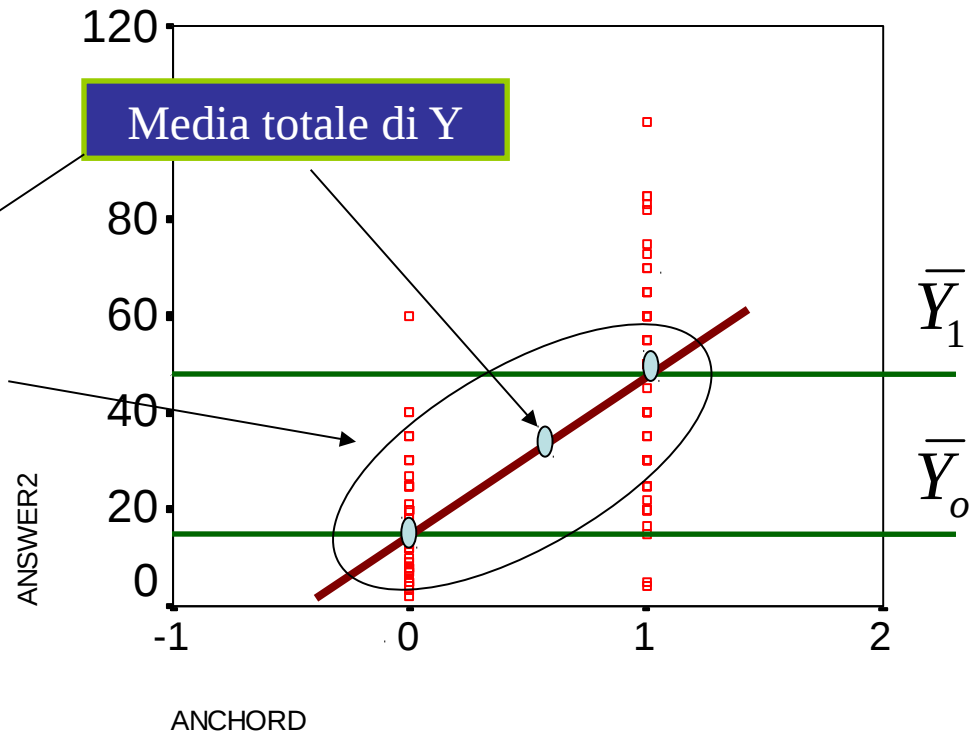
# Rapporto varianza

- Se la variabile indipendente è dicotomica, la varianza spiegata è la varianza tra i punteggi predetti

Varianza catturata  
regressione

$$\text{var}_{bet} = \frac{\sum (\bar{Y}_g - \bar{Y})^2}{df}$$

Varianza tra i gruppi  
(between groups)





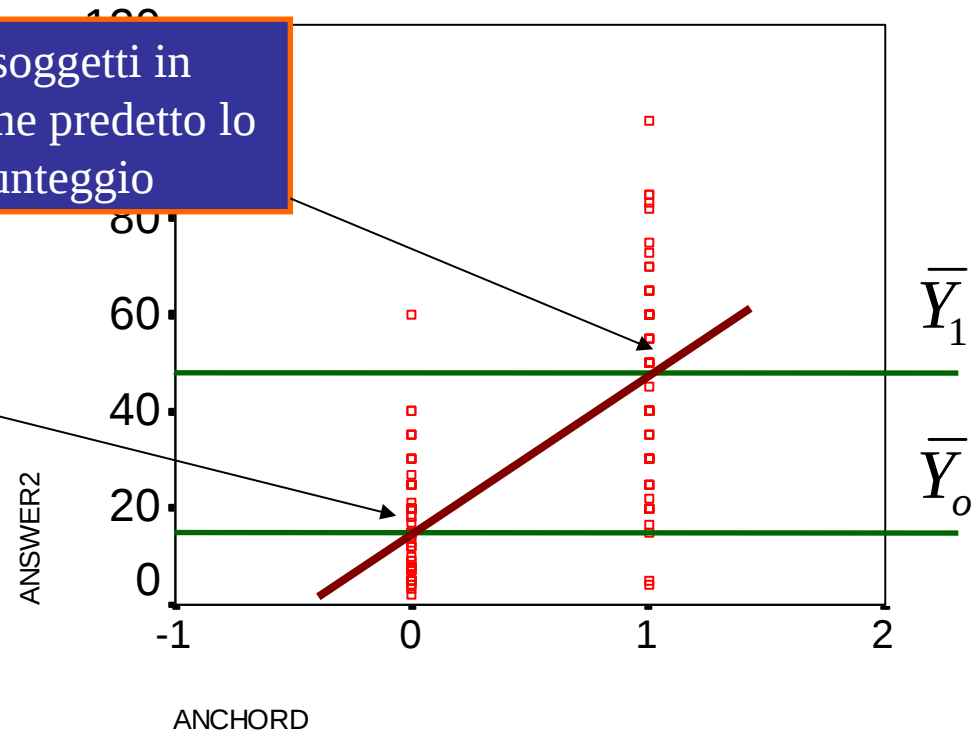
# Varianza tra i gruppi

- Equivalentemente, testiamo la bontà della predizione che noi facciamo dando ad ogni soggetto il punteggio medio del gruppo (variabilità dovuta ai gruppi), rispetto all'errore che facciamo

Per tutti i soggetti in gruppo=1 viene predetto lo stesso punteggio

Per tutti i soggetti in gruppo=0 viene predetto lo stesso punteggio

Ciò è il punteggio che otterremmo se le differenze fossero solo dovute al gruppo di appartenenza

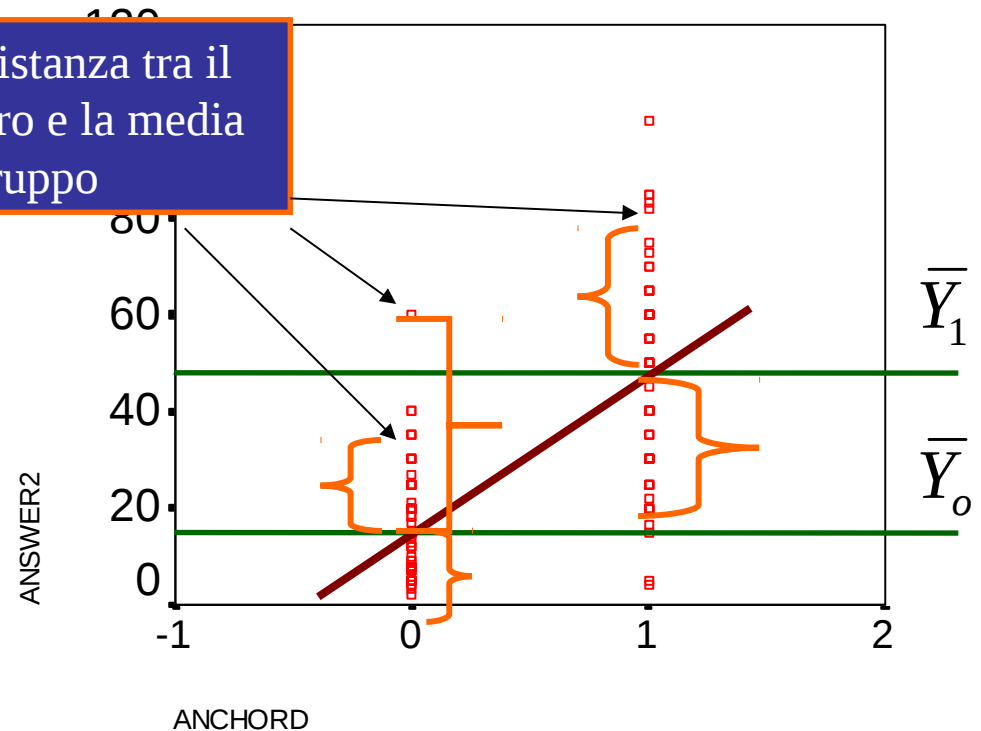


# Varianze nei gruppi

- Equivalentemente, stimiamo la bontà della predizione che noi facciamo dando ad ogni soggetto il punteggio medio del gruppo (variabilità dovuta ai gruppi), rispetto all'errore che facciamo

Errore è la distanza tra il punteggio vero e la media del gruppo

La variabilità dentro i gruppi è l'errore che commettiamo



# Rapporto varianze

- Il rapporto tra le varianze, pesato per i gradi di libertà, equivale al test F

$$\frac{\text{var}_{between} = \sum (\bar{Y}_g - \bar{\bar{Y}})^2 / df_{bet.}}{\text{var}_{within} = \sum (Y - \bar{Y}_g)^2 / df_{wit.}} = F$$

**ANOVA<sup>b</sup>**

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	32808,181	1	32808,181	134,063	,000 <sup>a</sup>
	Residuo	41113,332	168	244,722		
	Totale	73921,513	169			

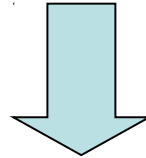
a. Stimatori: (Costante), ANCHOR  
b. Variabile dipendente: ANSWER2

*var<sub>between</sub>* (pointing to the 'Media dei quadrati' column for Regression)  
*var<sub>within</sub>* (pointing to the 'Media dei quadrati' column for Residual)  
*df<sub>within</sub>* (pointing to the 'df' column for Residual)  
*df<sub>bet</sub>* (pointing to the 'df' column for Regression)

# Variabili Nominali

- Quando la variabile indipendente ha più di due gruppi, lasceremo la rappresentazione in termine di retta, ma manteniamo i concetti di varianza tra i gruppi e varianza nei gruppi

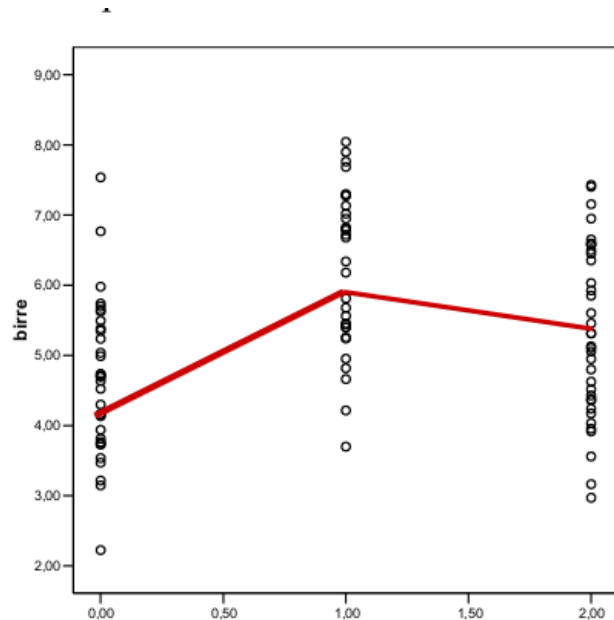
Testeremo le differenze tra i gruppi  
in termini di F



Ciò si chiama  
**analisi della varianza**  
**(Analysis Of Variance)**

# Perdiamo la retta

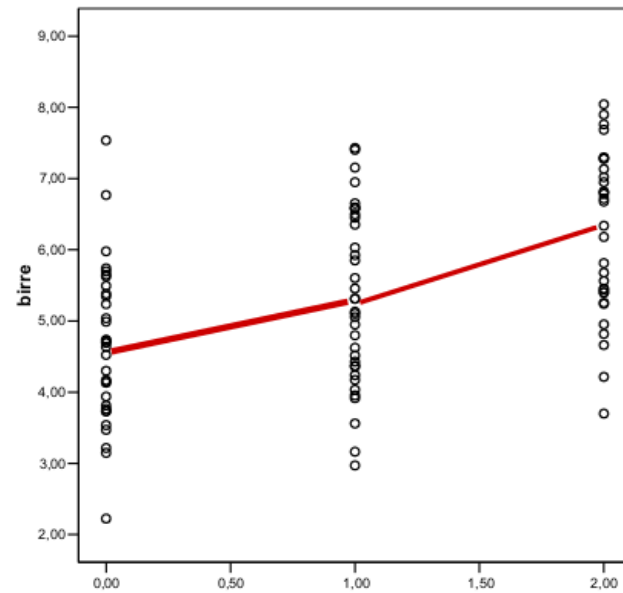
- Perdiamo la rappresentazione in termini di una retta in quanto la variabile indipendente non è ordinabile in maniera univoca



Italia

Germania

Francia



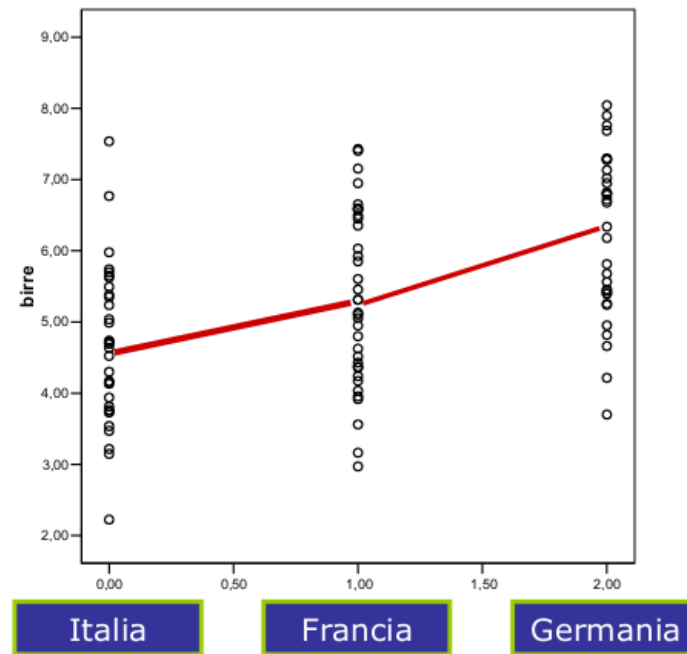
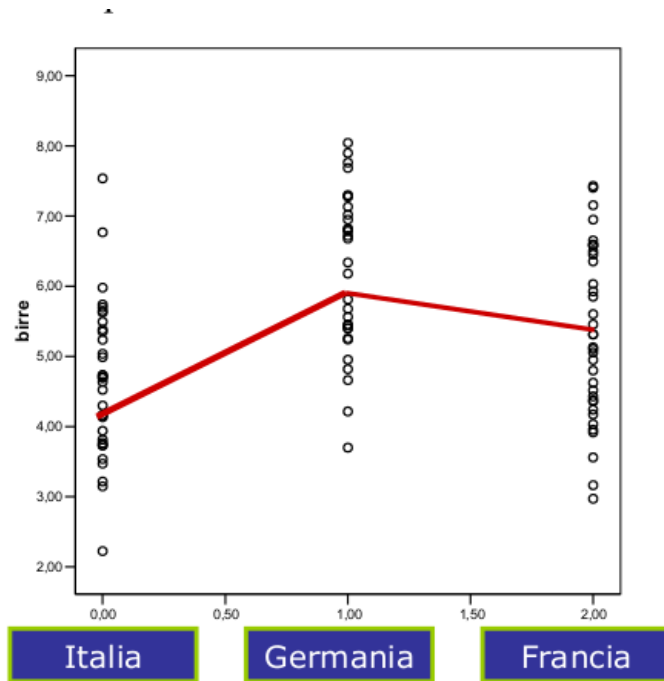
Italia

Francia

Germania

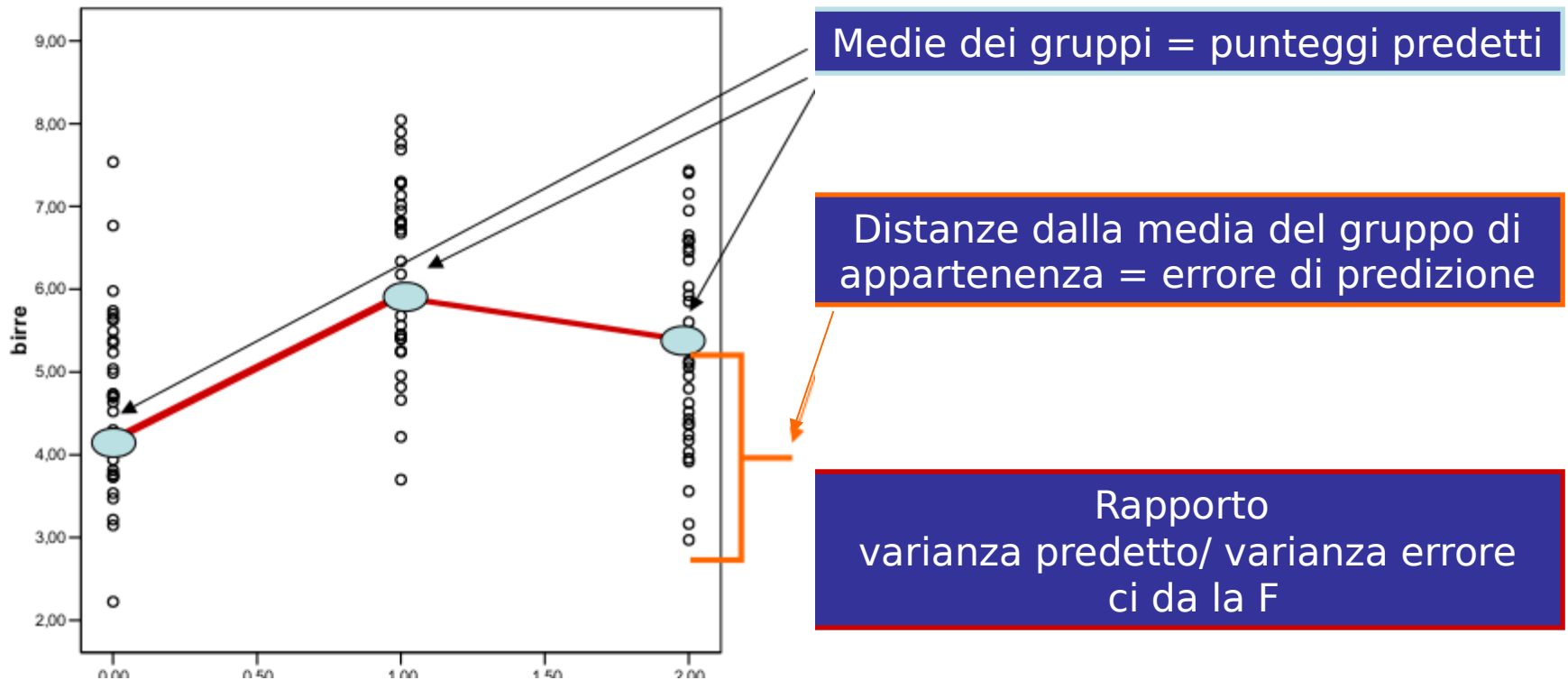
# Perdiamo la retta

- Ed un'unica retta (non spezzata) non potrebbe rappresentare i dati a nostra disposizione



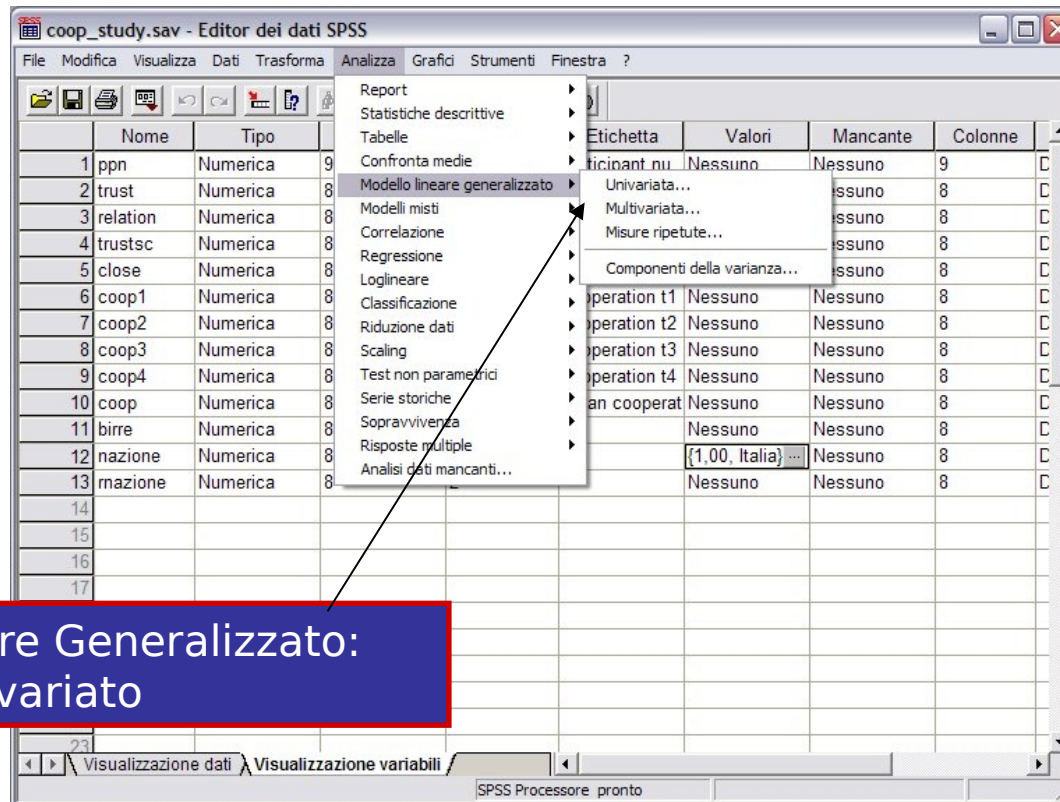
# Manteniamo i rapporti di varianza

- Possiamo comunque testare le differenze tra le medie mediante il rapporto F



# ANOVA in pratica

- In SPSS useremo una routine più generale della regressione, che consente di analizzare anche le variabili indipendenti nominali (**modello lineare generalizzato**)



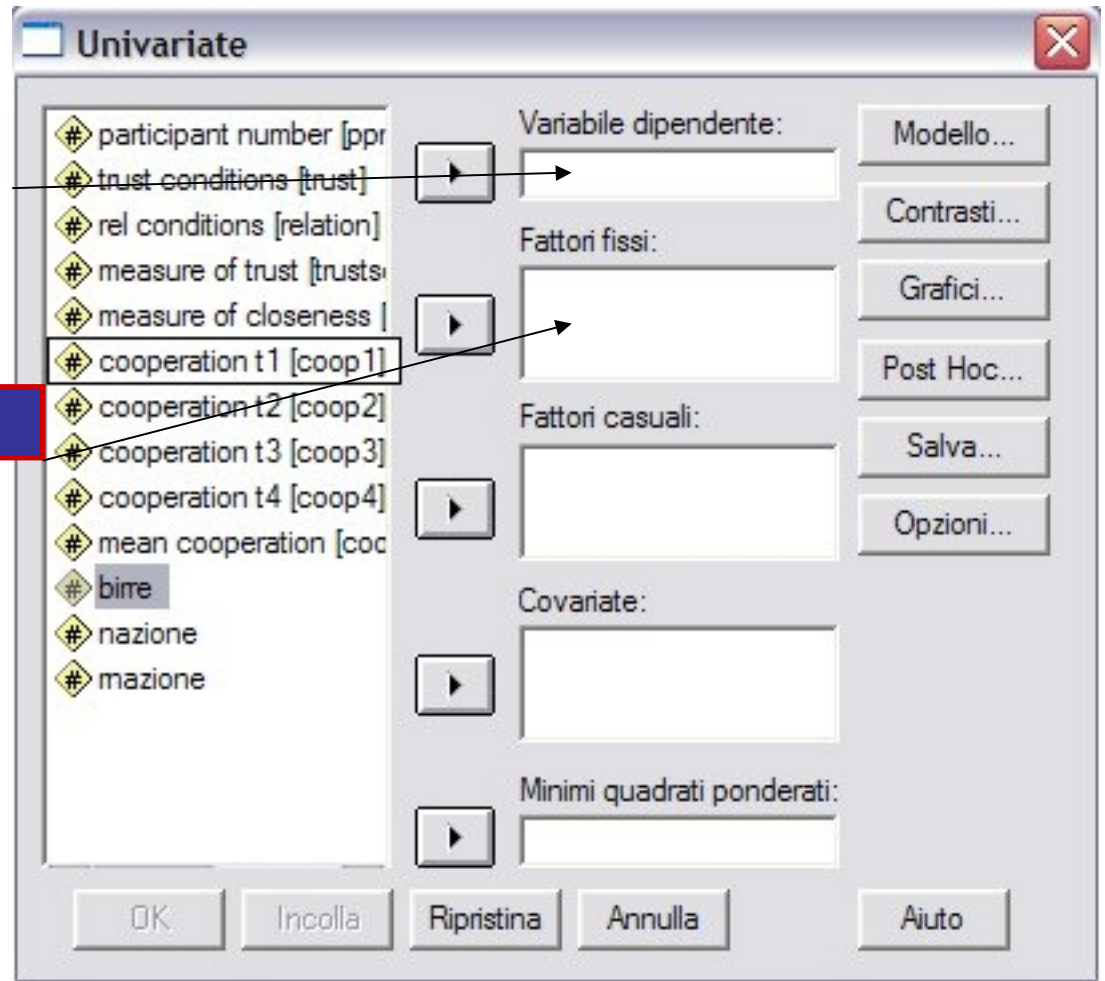
**Modello Lineare Generalizzato:  
Univariato**



# ANOVA in pratica

Variabile dipendente

Variabile indipendente



# Output

Gradi di libertà

Test F

Variabilità dovuta a..

Test degli effetti fra soggetti

variabile dipendente: birra

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.
Modello corretto	38,114 <sup>a</sup>	2	19,057	14,248	,000
Intercetta	2897,454	1	2897,454	2166,282	,000
nazione	38,114	2	19,057	14,248	,000
Errore	129,740	97	1,338		
Totale	3056,398	100			
Totale corretto	167,854	99			

a. R quadrato = ,227 (R quadrato corretto = ,211)

Concluderemo che tra i tre gruppi c'è una differenza significativa

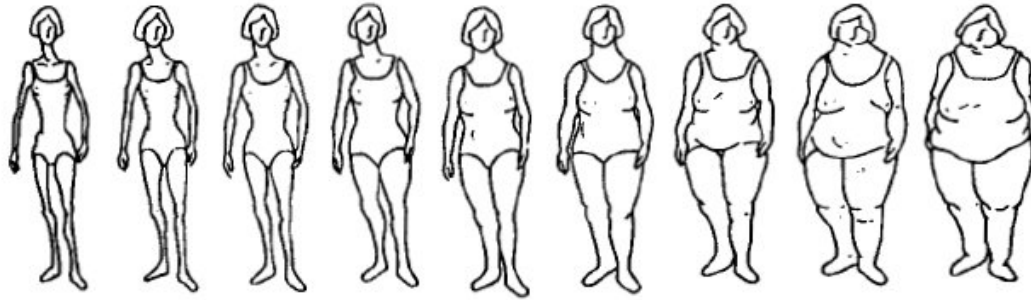
Valore-p: significatività

Almeno due medie sono diverse

# Ricerca Anoressia

- In ricerca sull'anoressia\* sono state misurate su un campione di 85 donne la propria “figura reale”, la “figura ideale” e l'autostima.

## Pictorial Body Image Scale



- Il campione era formato da donne normo-peso e donne anoressiche

# Ricerca Anoressia

## Group

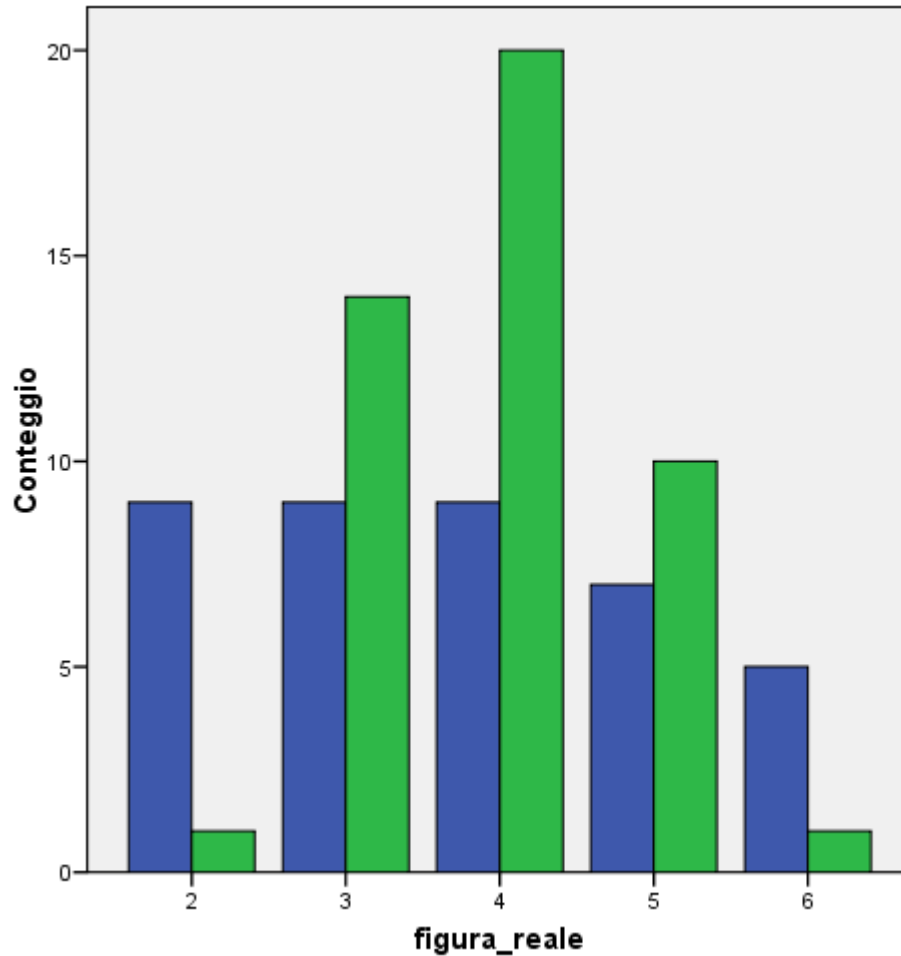
		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	A	39	45.9	45.9	45.9
	C	46	54.1	54.1	100.0
	Totale	85	100.0	100.0	

Vogliamo stabilire se ci sono delle differenze tra i due gruppi nel BIPS attuale e ideale

## Report

Group		figura reale	figura ideale
A	Media	3.74	2.92
	N	39	39
	Deviazione std.	1.352	.807
C	Media	3.91	3.35
	N	46	46
	Deviazione std.	.839	.482
Totale	Media	3.84	3.15
	N	85	85
	Deviazione std.	1.100	.681

# Distribuzione per gruppo



Group  
A  
C

BIPS reale

Le medie di queste due distribuzioni sono uguali?

# ANOVA

## Test degli effetti fra soggetti

Variabile dipendente: figura\_reale

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	.606 <sup>a</sup>	1	.606	.498	.483	.006
Intercetta	1237.312	1	1237.312	1015.915	.000	.924
Group	.606	1	.606	.498	.483	.006
Errore	101.088	83	1.218			
Totale	1352.000	85				
Totale corretto	101.694	84				

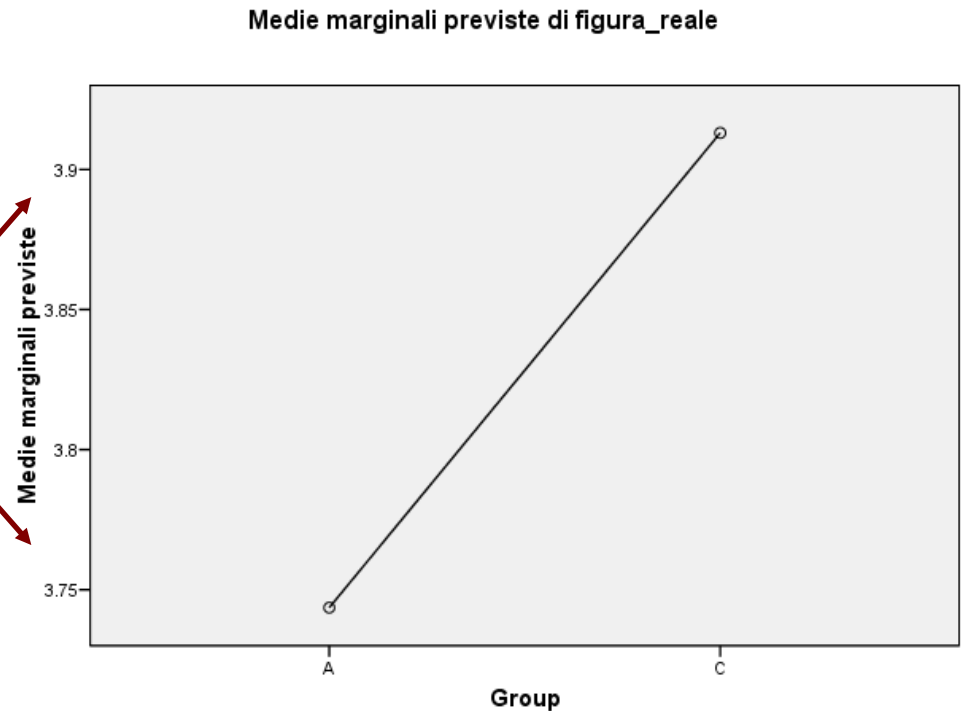
a. R quadrato = .006 (R quadrato corretto = -.006)

Non vi sono differenze tra le medie dei due gruppi

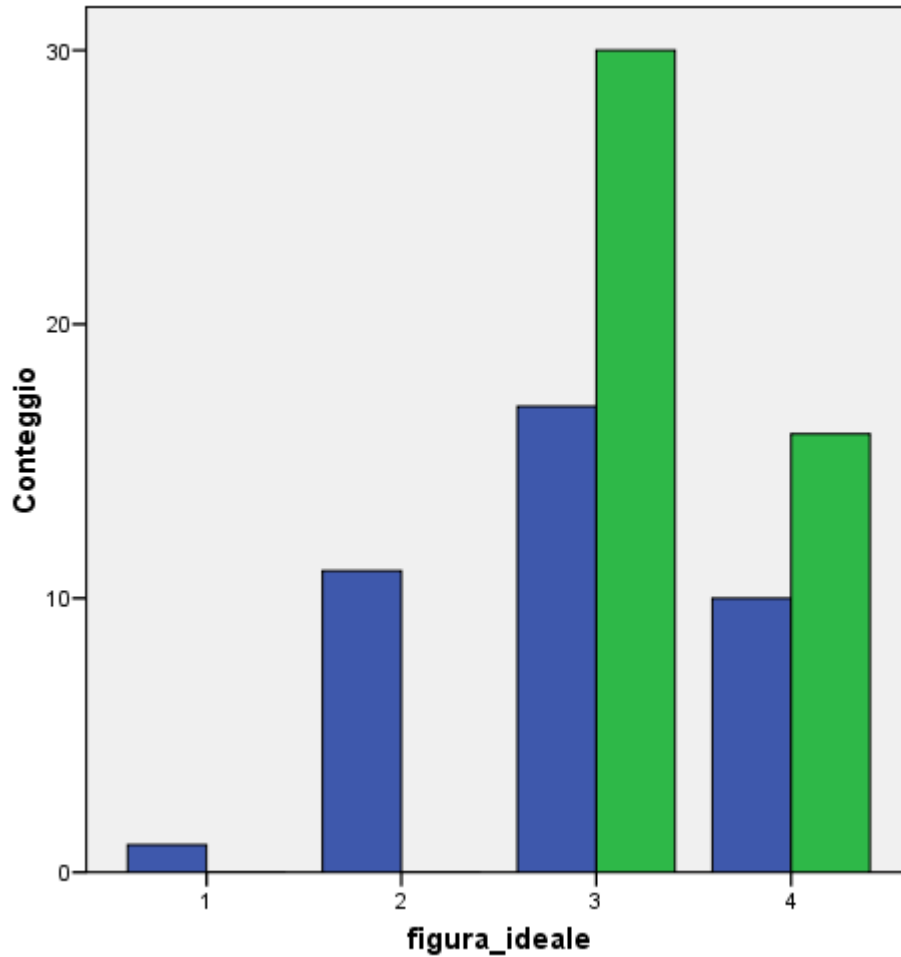
# Attenzione al grafico

- Il software produce un grafico delle medie che è sempre “espanso” e potrebbe far sembrare che vi sia una differenza
- Se l'effetto non è significativo, il grafico non va guardato

Notare la scala



# BIPS ideale



Group  
■ A BIPS ideale  
■ C

Le medie di queste due distribuzioni sono uguali?



# ANOVA

## Test degli effetti fra soggetti

Variabile dipendente: figura\_ideale

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	3.808 <sup>a</sup>	1	3.808	8.977	.004	.098
Intercetta	829.972	1	829.972	1956.814	.000	.959
Group	3.808	1	3.808	8.977	.004	.098
Errore	35.204	83	.424			
Totale	884.000	85				
Totale corretto	39.012	84				

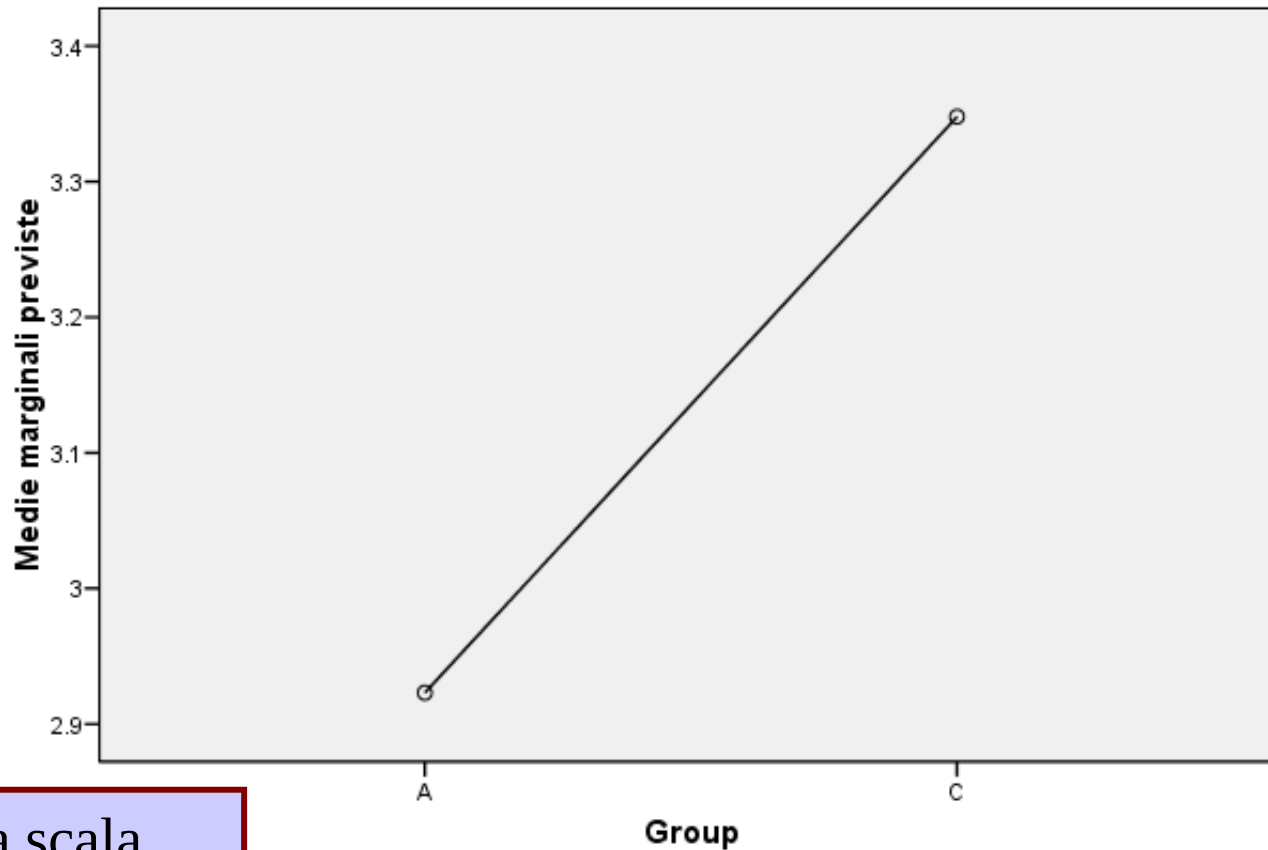
a. R quadrato = .098 (R quadrato corretto = .087)

Vi sono differenze significative fra i due gruppi

Che spiegano circa 10% della varianza

# Grafico delle medie

Medie marginali previste di figura\_ideale



Notare la scala

# Analisi della covarianza

- Si potrebbe dubitare che la differenza sia dovuta alla differenza tra autostima possibile nei due gruppi
- Testiamo prima se vi è una differenza tra i gruppi nell'autostima
- Poi testiamo le differenze nei tra PIBS ideale al netto dell'autostima

# ANOVA

## Test degli effetti fra soggetti

Variabile dipendente: Autostima

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	54.782 <sup>a</sup>	1	54.782	33.877	.000	.290
Intercetta	1306.344	1	1306.344	807.839	.000	.907
Group	54.782	1	54.782	33.877	.000	.290
Errore	134.218	83	1.617			
Totale	1549.000	85				
Totale corretto	189.000	84				

a. R quadrato = .290 (R quadrato corretto = .281)

## Group

Variabile dipendente: Autostima

Group	Media	Errore std.	Intervallo di confidenza 95%	
			Limite inferiore	Limite superiore
A	3.128	.204	2.723	3.533
C	4.739	.187	4.366	5.112

Le ragazze anoressiche hanno una autostima più bassa

# ANCOVA

- Ora proviamo a testare BIPS ideale tenendo costante la autostima
- Cioè facciamo una ANOVA ed inseriamo anche una variabile indipendente continua (come nella regressione)
- L'effetto di GROUP sarà dunque la differenza tra Anoressiche e Controllo nel BIPS ideale tenendo costante le differenze in autostima (come se Anoressiche e Controllo fossero uguali nell'autostima)

$$\hat{y}_i = a + b_1 \text{GROUP} + b_2 \text{AS}$$

GROUP=[0 1]

AS=continua

# ANCOVA

## Test degli effetti fra soggetti

Variabile dipendente: figura\_ideale

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	4.255 <sup>a</sup>	2	2.127	5.019	.009	.109
Intercetta	66.538	1	66.538	156.978	.000	.657
Autostima	.447	1	.447	1.054	.308	.013
Group	1.650	1	1.650	3.892	.052	.045
Errore	34.757	82	.424			
Totale	884.000	85				
Totale corretto	39.012	84				

a. R quadrato = .109 (R quadrato corretto = .087)

Al netto della autostima, le differenze tra gruppi si riducono

# Riduzione delle differenze

Eta-quadro per i gruppi

Sig.	Eta quadrato parziale
.004	.098
.000	.959
.004	.098

Eta-quadro gruppi dopo parzializzato autostima

Sig.	Eta quadrato parziale
.009	.109
.000	.657
.308	.013
.052	.045

# ANCOVA

- Ora proviamo a testare BIPS ideale tenendo costante BISP reale

## Test degli effetti fra soggetti

Variabile dipendente: figura\_ideale

Sorgente	Somma dei quadrati Tipo III	df	Media dei quadrati	F	Sig.	Eta quadrato parziale
Modello corretto	8.433 <sup>a</sup>	2	4.216	11.306	.000	.216
Intercetta	34.220	1	34.220	91.763	.000	.528
figura_reale	4.625	1	4.625	12.402	.001	.131
Group	3.167	1	3.167	8.492	.005	.094
Errore	30.579	82	.373			
Totale	884.000	85				
Totale corretto	39.012	84				

a. R quadrato = .216 (R quadrato corretto = .197)



# Dunque

- Le ragazze anoressiche non si percepiscono come più magre delle ragazze normopese
- Le ragazze anoressiche hanno una immagine ideale molto più magra delle normopeso
- La differenza nella immagine ideale non dipende dalla percezione della immagine reale
- La differenza nella immagine ideale dipende in parte dalla propria autostima

Fine

Fine della Lezione VI

