UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Analisi Multivariata dei Dati

Introduzione al corso e al modello statistico

Marcello Gallucci

Milano-Bicocca

A
M
D

Programma Odierno

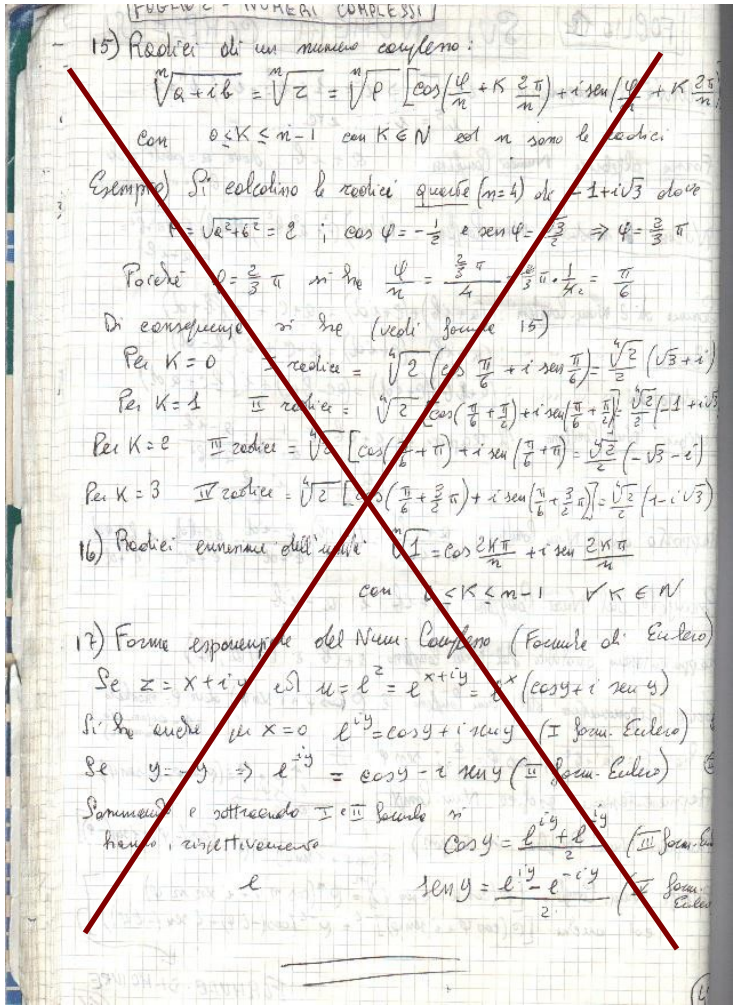
- I numeri del corso
- Programma del corso
- Concetti Statistici Introduttivi

Numeri del Corso

- 21 lezioni, 2 ore l’una (a parte le pause)
- 18 ore di esercitazioni
- 2 appelli a sessione + recuperi
- 8 CFU

Stile del Corso

- Enfasi sui concetti
- Enfasi sul ragionamento statistico
- Minimo utilizzo di formule
- Utilizzo di software per i calcoli
- Enfasi sulla interpretazione dei risultati



Software Statistico

- Il programma SPSS verra’ usato per condurre i calcoli necessari alle analisi
- L’utilizzo di SPSS verra’ insegnato nelle esercitazioni
- Varie fonti offrono guide pratiche a SPSS
- Materiale online del libro di testo

Esercitazioni

- Il corso si completa di 16 ore di esercitazione da tenersi nelle aule informatiche

- Lo scopo delle esercitazioni e’ di imparare ad eseguire ed interpretare praticamente le analisi statistiche studiate a lezione

- Le esercitazione inizieranno tra due settimane. I gruppi verranno fatti dopo l'iscrizione al gruppo via pagina web del corso

Libri di Testo

Libri di Testo

Il libro di testo è *Gallucci, Leone, Berlingeri (2011). Modelli statistici nelle scienze sociali, seconda edizione*

Le “lezioni” possono essere scaricate dalla pagina del corso

I capitoli da studiare sono elencati nel materiale del corso

http://elearning.unimib.it/course/view.php?id=5601

Esami

- L’esame e’ scritto e si svolge nei laboratori informatici dove lo studente potrà utilizzare il software statistico per rispondere alle domande

1) Domande a scelta multipla

2) Domande aperte riguardanti una ricerca empirica su cui lo studente condurrà le analisi usando SPSS.

- Tutti possono integrare il voto con l’esame orale
- L’esame orale potrebbe abbassare il voto!!

Scopi del Corso

- Lo studio **approfondito** di alcune importanti Tecniche Statistiche Univariate
  - Analisi della varianza
  - Regressione lineare e logistica
  - Modelli lineari generalizzati

Inteso come ripasso di corsi precedendi

- Studio di tecniche **multivariate** per l'analisi di grandezze psicologiche osservate ripetutamente nel tempo o in modalità ripetute
  - Analisi della Varianza a misure ripetute
  - Modelli misti
  - Analisi fattoriale

Focus centrale del corso

Tecniche Multivariate

- Cosa sono ?
  - Per capire le tecniche multivariate dobbiamo ricordare cosa sono le tecniche univariate
  - Per ricordare le tecniche univariate dobbiamo ricordare la logica delle tipo tecniche statistiche che andremo a studiare
  - Tecniche volte allo studio delle relazioni tra variabili

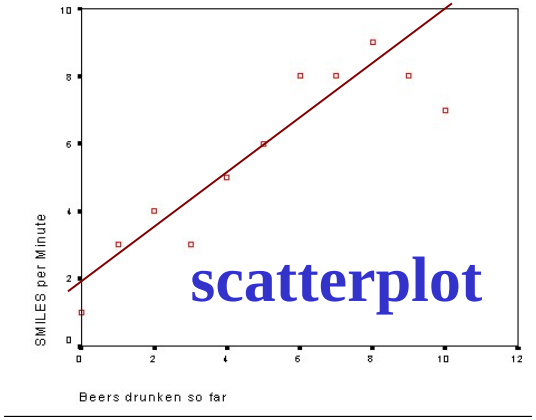
Tecniche Univariate

Techniche volte a studiare e quantificare gli effetti di una o più variabili indipendenti (variabili esplicative o predittori) su una variabile dipendente (variabile di interesse)

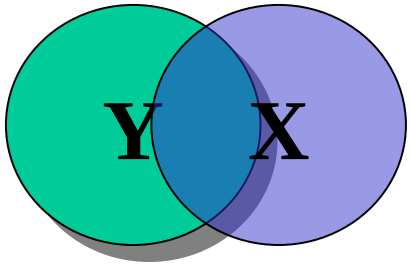
- Cosa intendiamo per “effetti”
- Cosa intendiamo per “variabile dipendente”
- Cosa intendiamo per “variabili indipendenti”

Relazioni statistiche

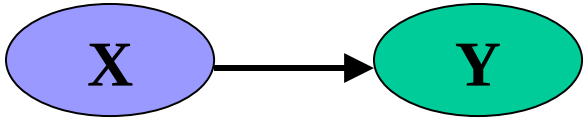
La maggior parte delle tecniche statistiche che conosciamo (e incontreremo) definiscono un **modello statistico** delle **relazioni** fra variabili di interesse



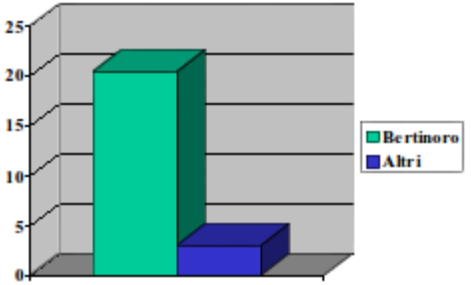
scatterplot



Path Diagram

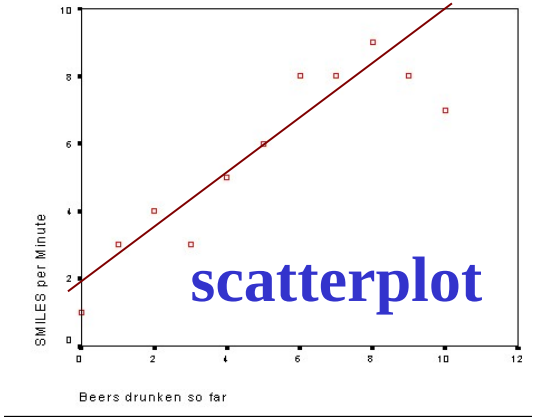


Differenze medie

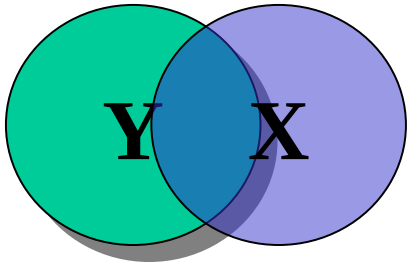


Modello Statistico

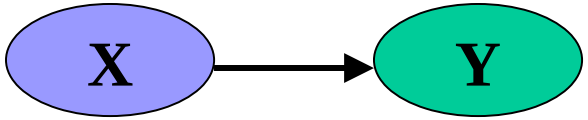
Un semplice **modello statistico** è una rappresentazione efficiente e compatta dei dati raccolti per descrivere un fenomeno empirico



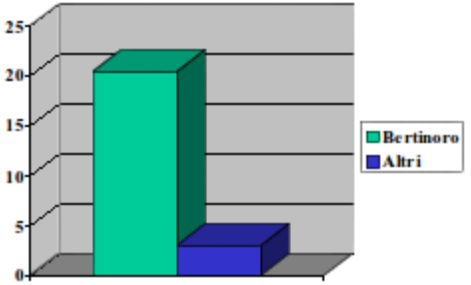
scatterplot



Path Diagram



Differenze medie

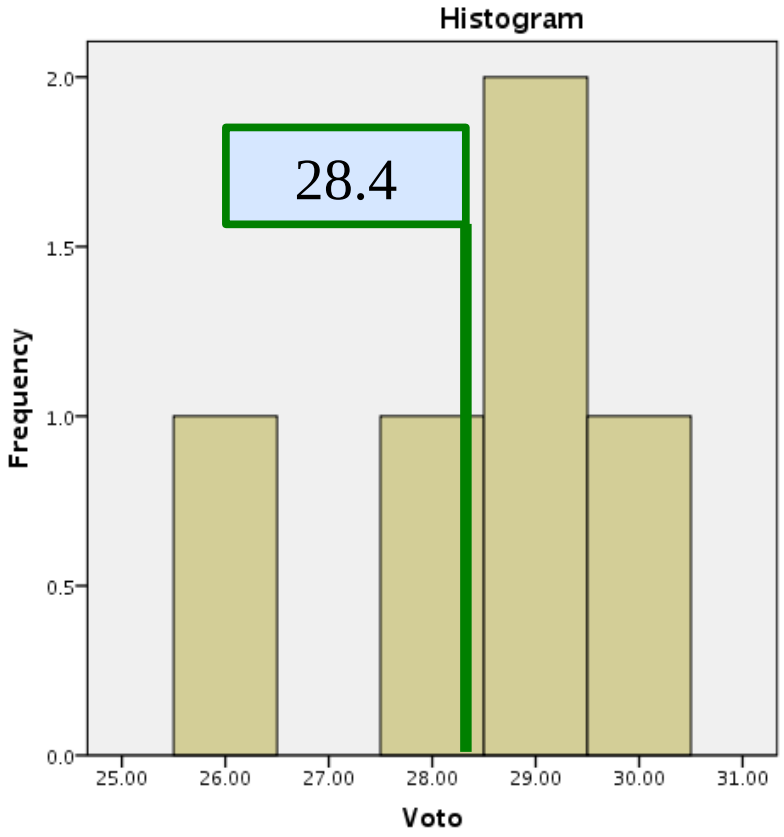


Esempio: la media

Q: “Come vanno gli studenti al mio corso?”

R: “Hanno una media del 28.4”

$$\frac{\sum_i X_i}{N} = \bar{X}$$

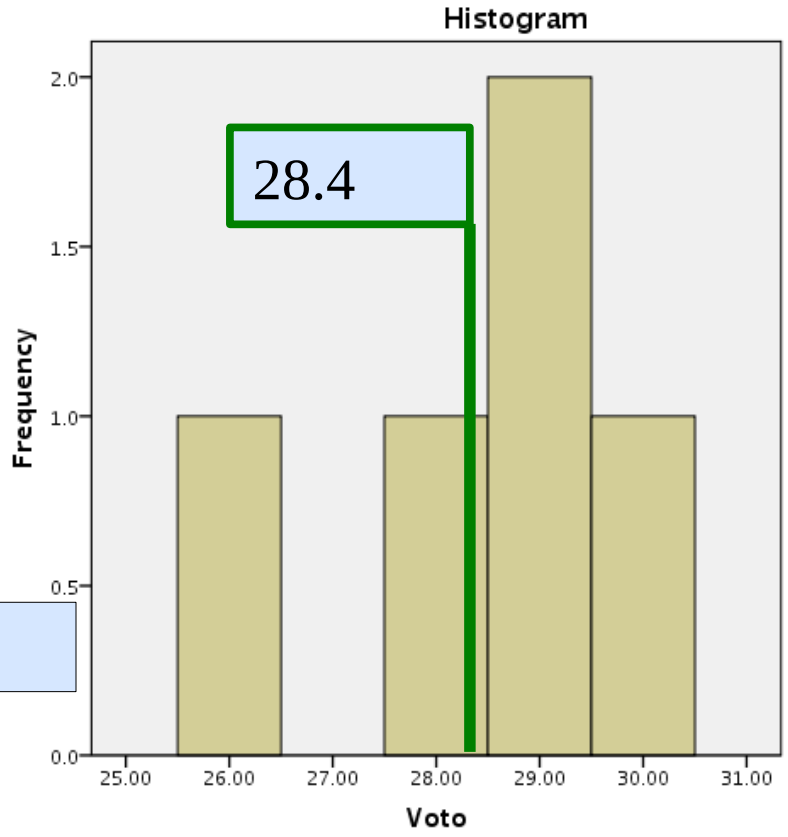


Introduzione

Il modello statistico e la rappresentazione che ne facciamo serve (tra l'altro) a tre scopi:

- Descrizione efficiente e compatta
- Predizione del futuro
- Inferenza sulla popolazione

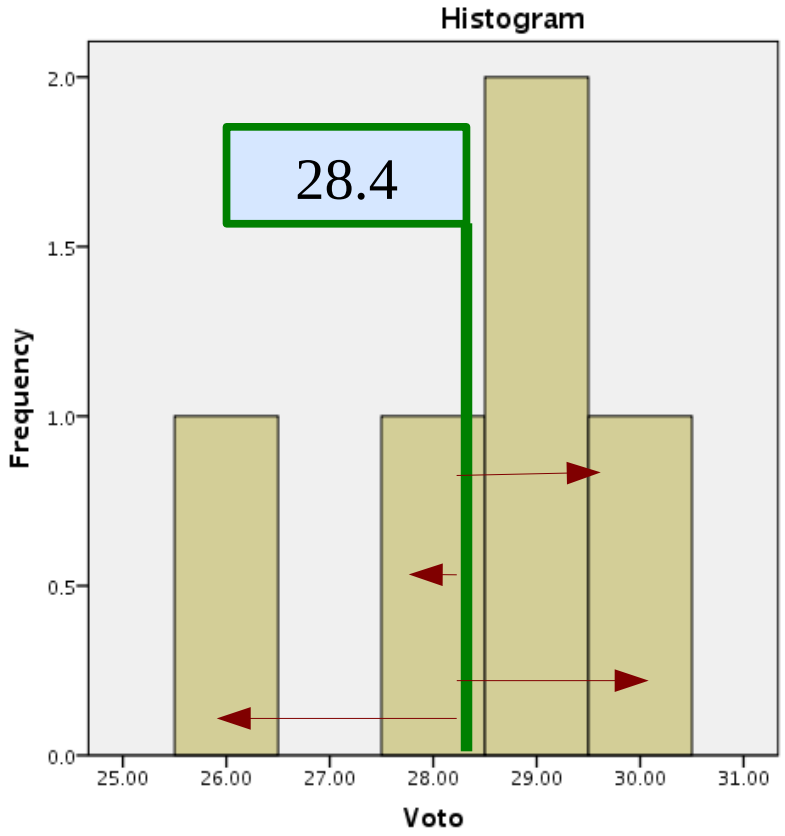
Cioè: comprensione del fenomeno



Errore di approssimazione

Come tutte le rappresentazioni compatte ed efficienti, anche quella statistica è una approssimazione dei dati rappresentati

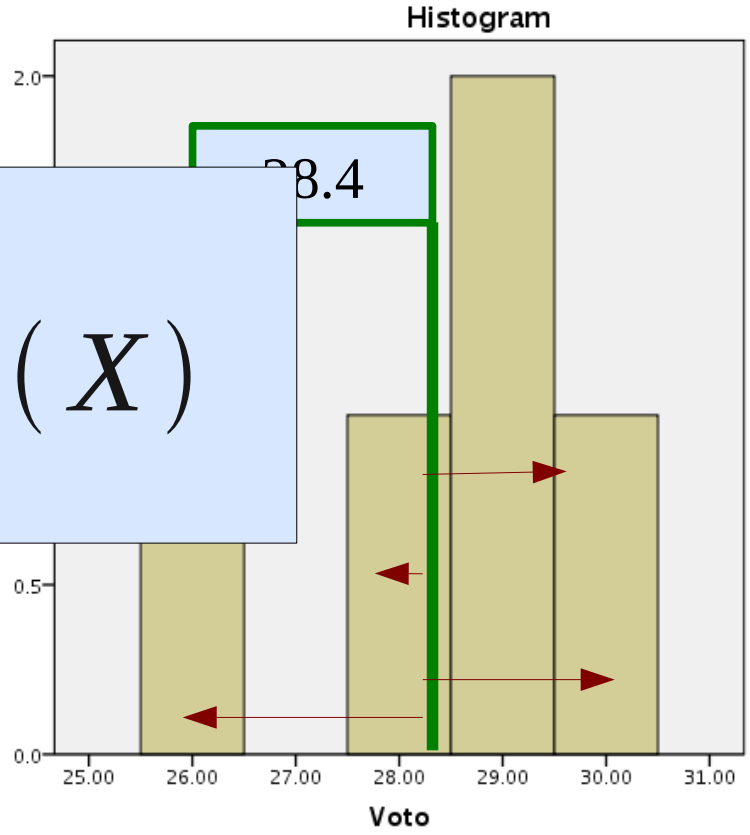
Se, per semplificare, diremo che la performance è di 28.4, mis-rappresenteremo alcuni dei voti effettivi



Errore di approssimazione

Calcolando questo errore per ogni caso (ogni studente), elevandolo al quadrato (sbagliare in più o in meno è uguale) e facendo la media per ogni caso, quantifichiamo l'errore medie associato alla media

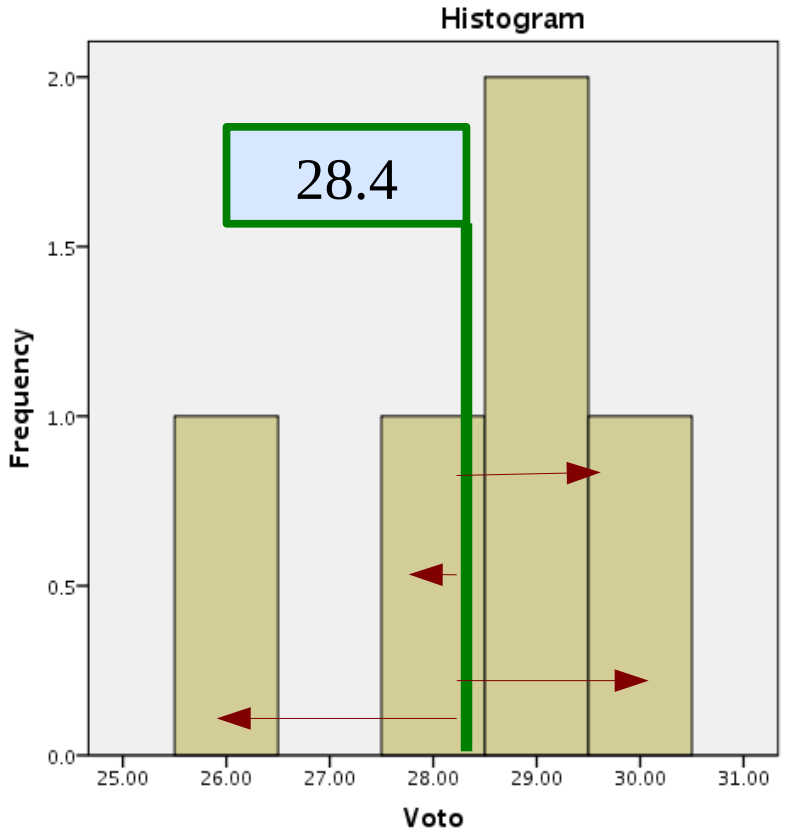
$$\frac{\sum_i (X_i - \bar{X})^2}{N-1} = Var(X)$$



Inferenza statistica

Il modello statistico è associato ad una serie di test inferenziali che ci consentono di trarre conclusioni sulla popolazione di riferimento

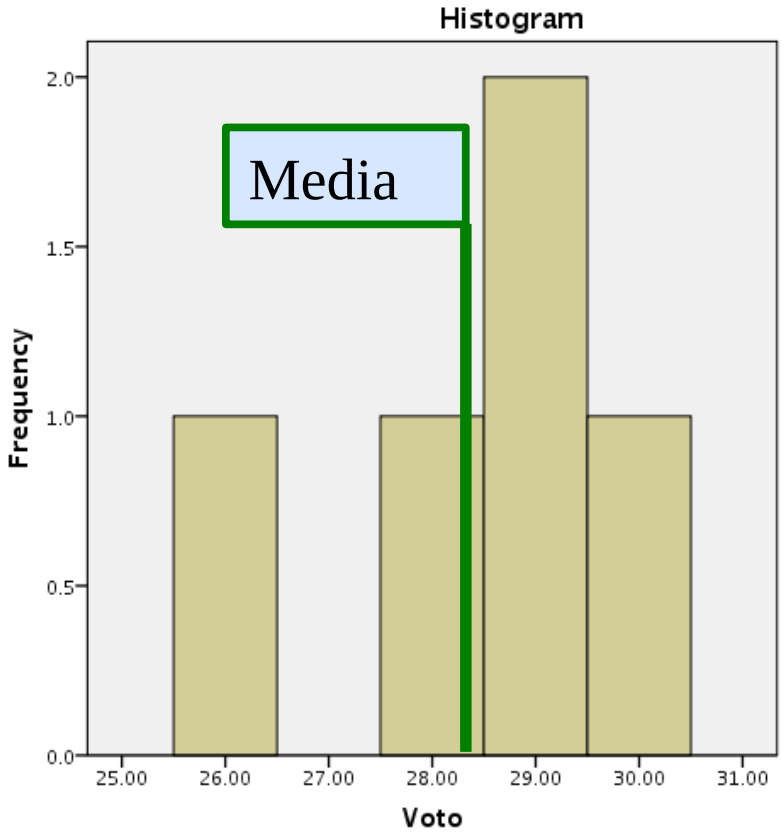
$$\frac{\bar{X} - \mu}{\sqrt{\frac{Var(X)}{N}}} = ttest$$



Modello statistico

Il modello statistico sarà una buona rappresentazione dei dati se:

- I parametri sono modellati correttamente
- Gli errori sono modellati correttamente
- La struttura dei dati è rispettata



Scegliere un modello statistico

Per costruire un corretto modello statistico dei nostri dati dobbiamo sapere una serie di cose:

- Cosa ci serve il modello (lo scopo dell'analisi)
- Che tipo di variabili abbiamo
- Che tipo di relazioni vogliamo studiare
- Quali sono le unità di misurazioni dei dati
- Come sono strutturati i nostri dati

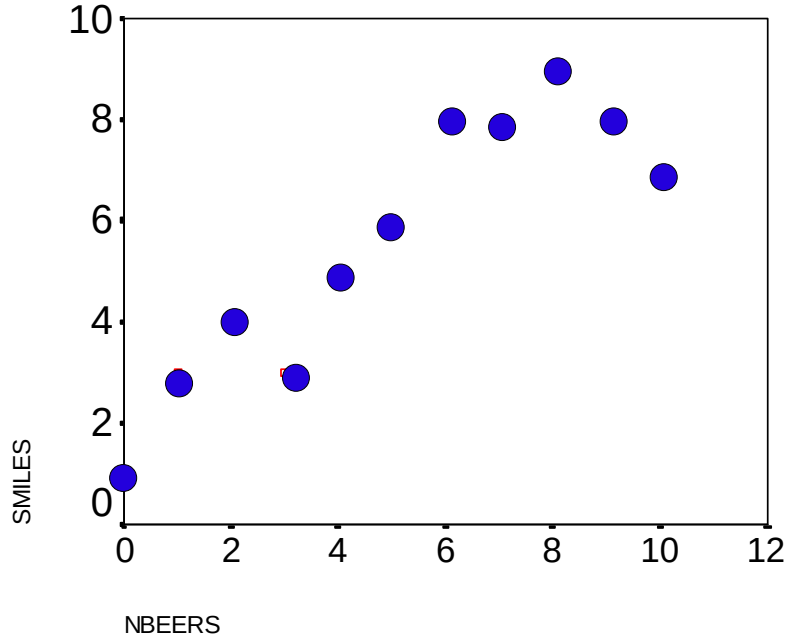
Il modello di regressione

(Capitolo 2)

Concentti fondamentali

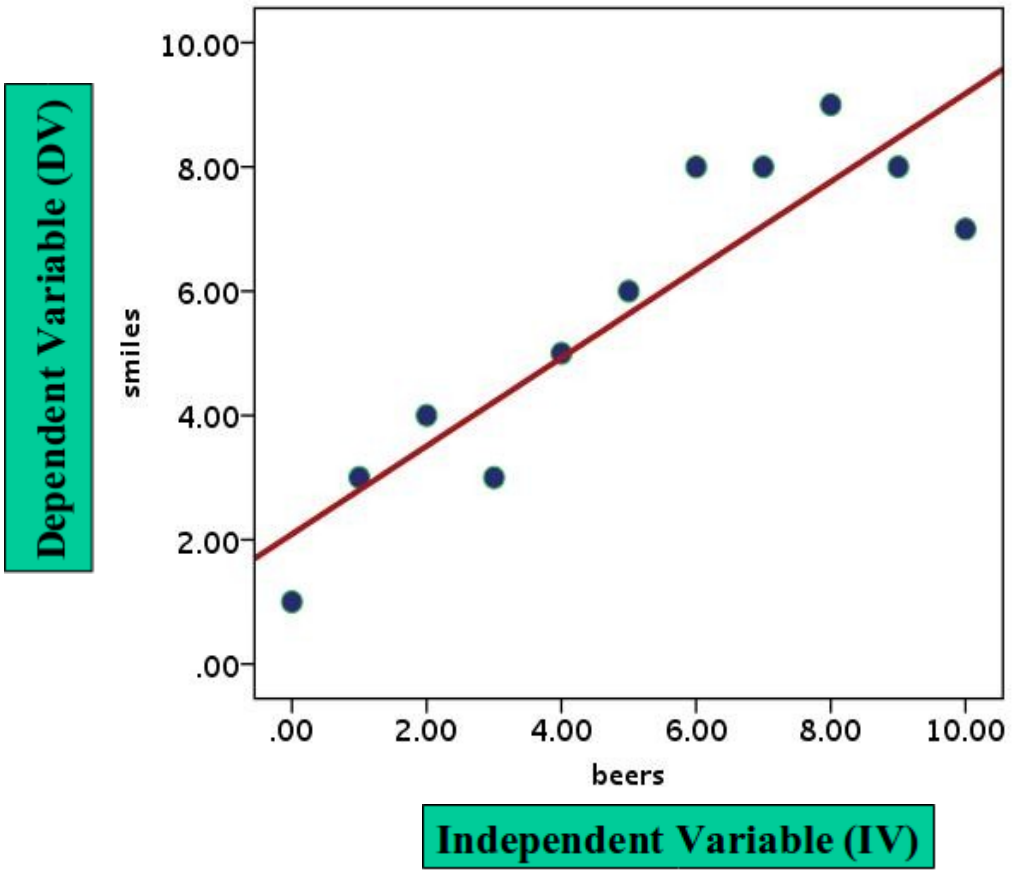
Consideriamo ora questa ipotetica ricerca: siamo andati in un pub ed abbiamo contato quanti sorrisi le persone ai tavoli producevano (ogni 10 minuti) e quante birre avevano bevuto fino a quel momento

Birre	Sorrisi
0	1
1	3
2	4
3	3
4	5
5	6
6	8
7	8
8	9
9	8
10	7



Concentti fondamentali

Lo scopo della retta di regressione è di rappresentare la relazione lineare tra la variabile indipendente e la dipendente



Nel caso più semplice, abbiamo una retta senplice

$$y_i = a + b \cdot y_i + e_i$$

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

Concetti fondamentali

La retta può essere descritta mediante due coefficienti: il termine costante ed il coefficiente angolare

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

a. Dependent Variable: SMILES

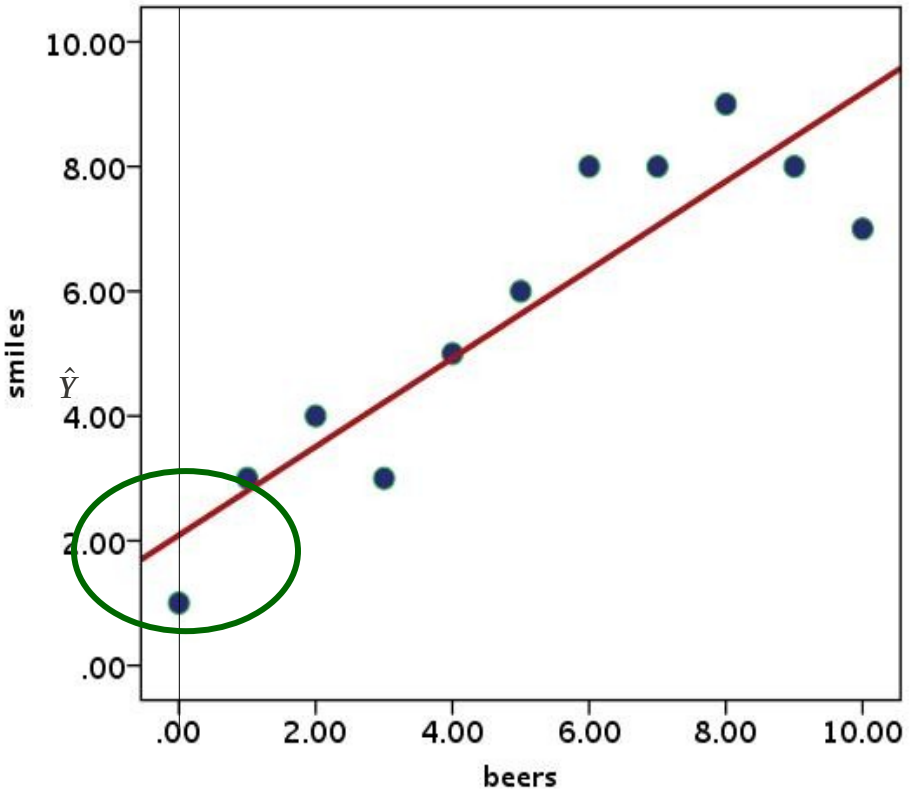
Termine constante (o intercetta)

Coefficiente di regressione (angolare)

Coefficiente costante

a l'intercetta della linea: indica il valore atteso (medio) della VD per la VI=0

$$\hat{y} = a + b \cdot 0$$

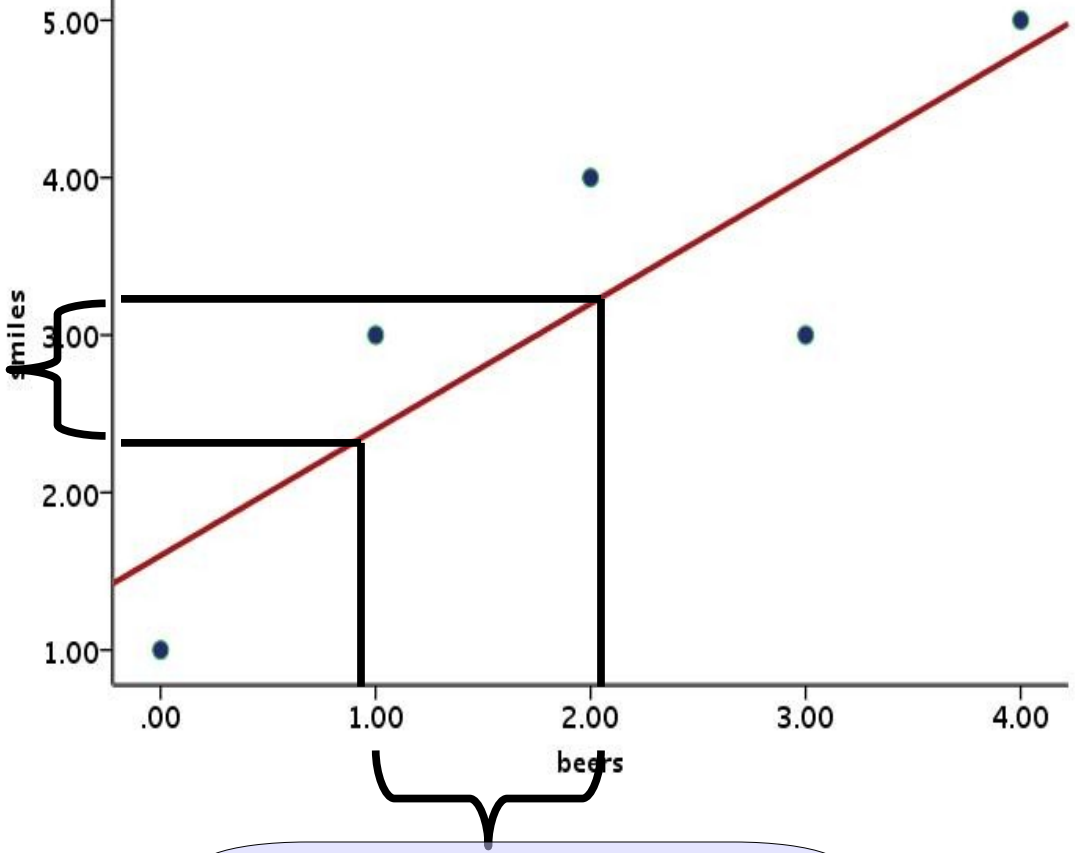


Quando un partecipante ha bevuto zero birre, mostra (in media) 2.09 sorrisi

Coefficiente di regressione

B è il coefficiente angolare della retta: indica il cambiamento atteso nella VD al variare di una unità della VI

I sorrisi aumentano di B unità

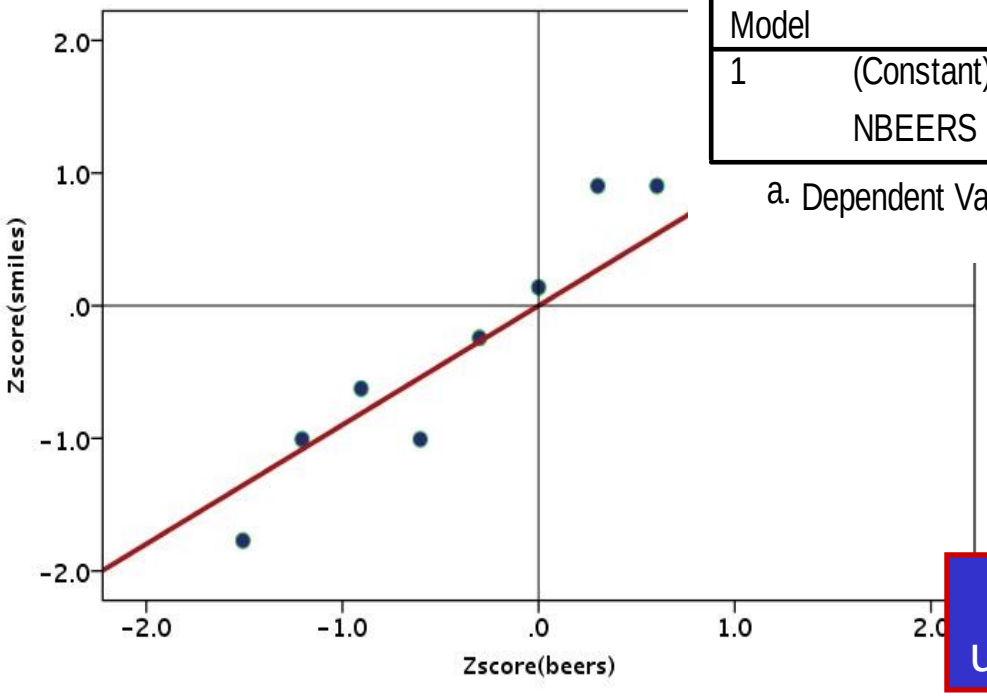


Per ogni birra che si beve, i sorrisi aumentano in media di .709 unità

Per una unità in più della VI: una birra in più

Coefficienti standardizzati

Il coefficiente **Beta** equivale al coefficiente di regressione calcolato dopo aver standardizzato tutte le variabili



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

a. Dependent Variable: SMILES

Il coefficiente standardizzato è uguale al coefficiente r di Pearson

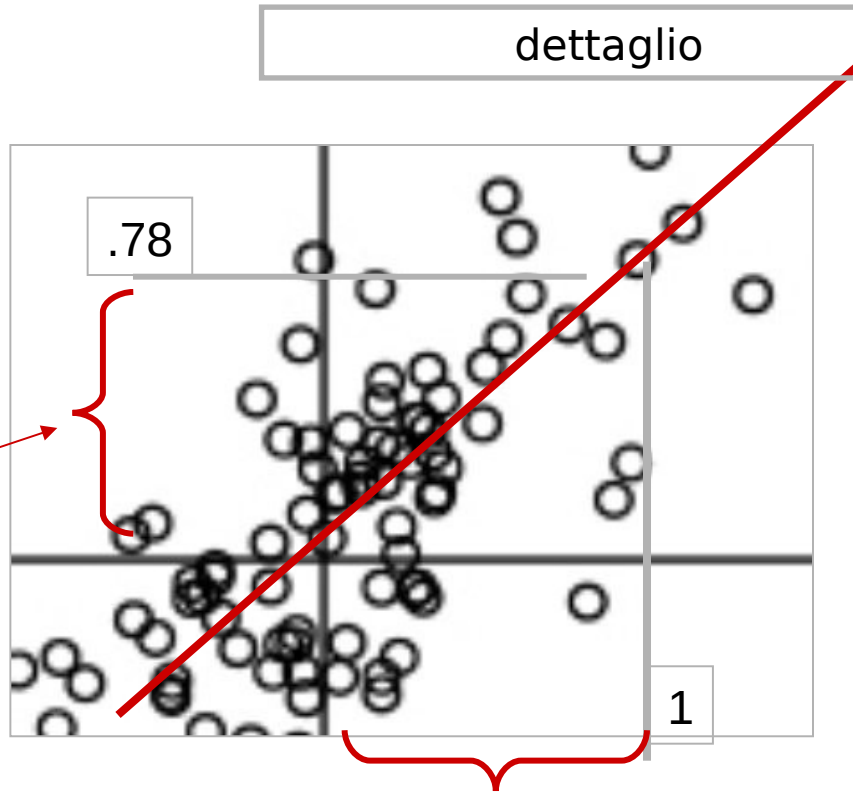
Correlazione: Interpretazione

- La correlazione indica il cambiamento atteso in **v**, al variare di **x** di una deviazione standard

Legge di relazione r=0.78

$$\hat{v}_z = r_{xv} x_z$$

Mi aspetto una scostamento pari a 78% della dev.std di **v**



Test inferenziale

I coefficenti vengono testati per la loro significatività statistica mediante il t-test **t test**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

a. Dependent Variable: SMILES

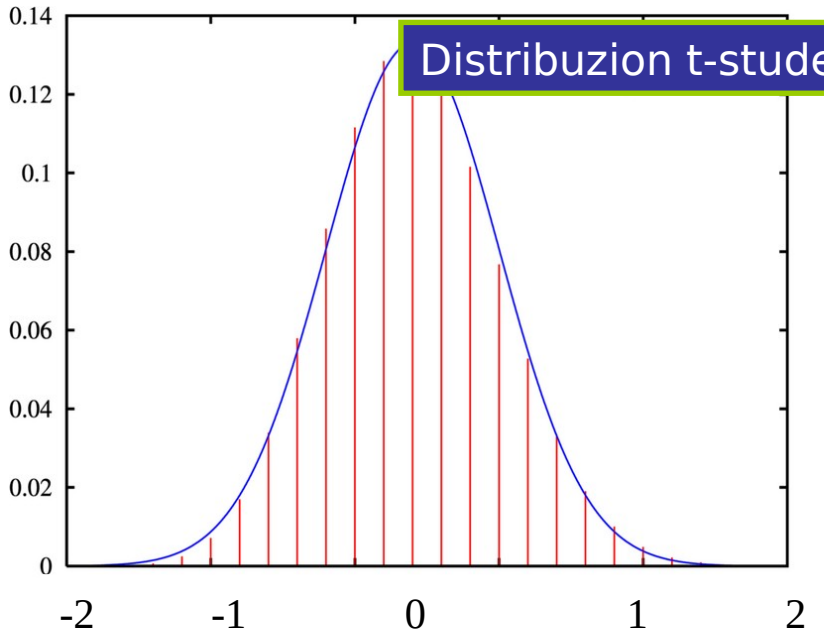
Se Sig. < 0.05, diremo che B (e β) sono significativamente diversi da zero

Valore “p”

Test Inferenziale

- Uno dei test più usati è il t-test

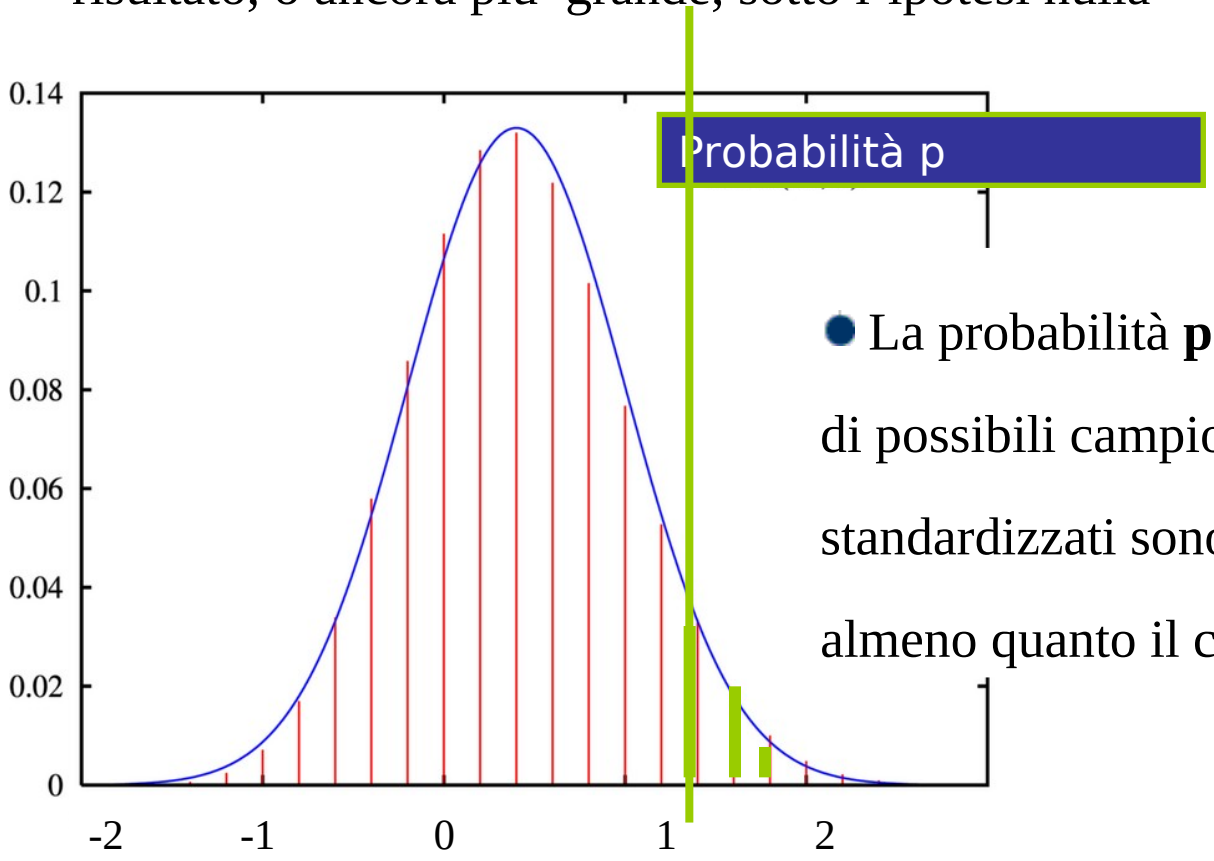
$$\frac{S}{\sqrt{\frac{Var(S)}{N}}} = ttest$$



S=parametro stimato

Test Inferenziale

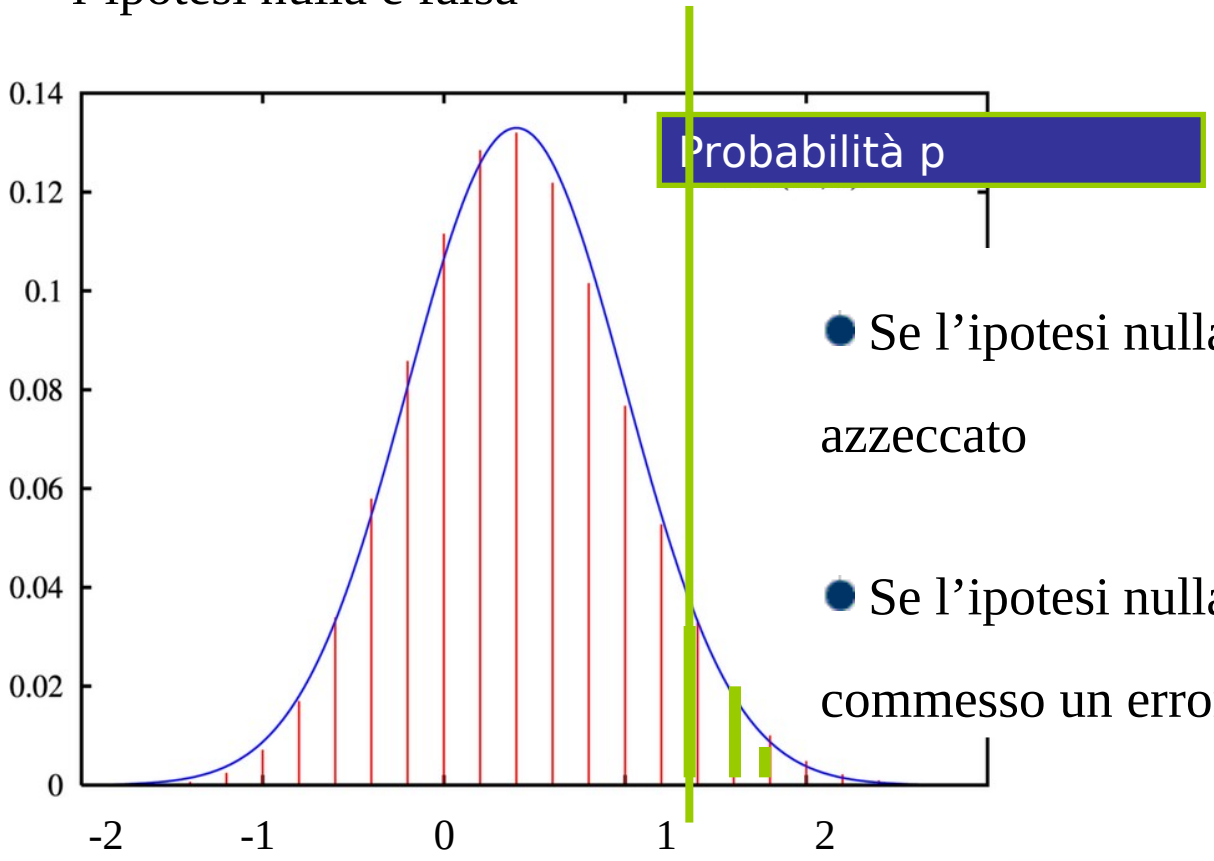
- In generale, il valore **p** indica la probabilità di ottenere il nostro risultato, o ancora più grande, sotto l'ipotesi nulla



- La probabilità **p** equivale alla proporzione di possibili campioni i cui scostamenti standardizzati sono distanti dall'ipotesi nulla almeno quanto il campione da noi osservato

VALORE-P

- Il valore **p** indica il rischio che noi prendiamo quando affermiamo che l'ipotesi nulla è falsa



- Se l'ipotesi nulla è falsa, ci abbiamo azzeccato
- Se l'ipotesi nulla è vera, abbiamo commesso un errore, detto del Tipo I

Test inferenziale e valore p

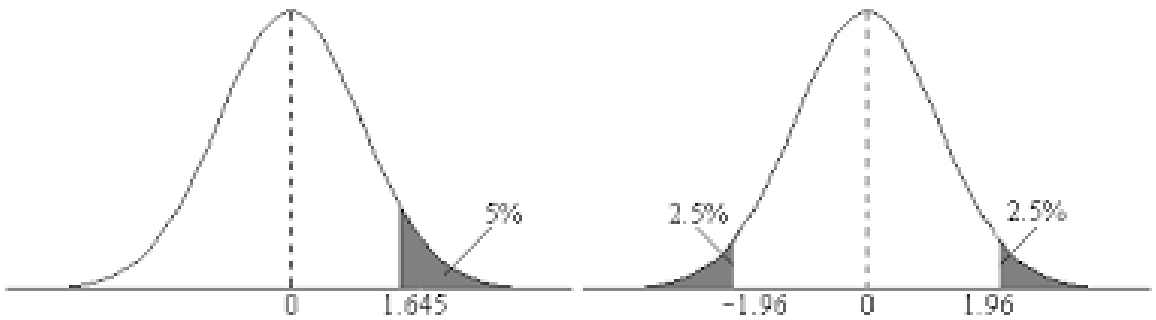
- Sulla base del valore del test inferenziale, possiamo ricavarci la probabilità corrispondente nella distribuzione (valore p)

- Nelle scienze sociali il valore **convenzionale** di significatività è 0.05 (5% di errore) o 0.01 (1 % errore)

- Il test può essere direzionale (una via) o di diseguaglianza (due vie)

- Test ad una via (one-tail): il parametro (ad es., la media) =0 vs. >0 (o <)

- Test a due vie (two-tails): il parametro (ad es., la media) =0 vs. ≠0



(a) One-tailed test (b) Two-tailed test

Errore Standard

- La maggior parte dei test inferenziali confrontano la stima con il suo errore standard **errore standard**

$$\frac{S}{\sqrt{\frac{Var(S)}{N}}} = ttest$$

parametro stimato

Errore standard

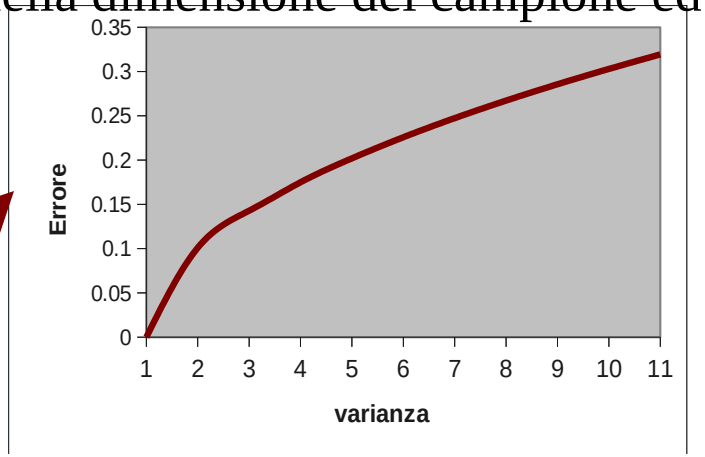
Indica quanto variabilità ci aspettiamo nei valori della stima se ripetessimo la stima su tanti campioni presi dalla stessa popolazione

L’errore standard

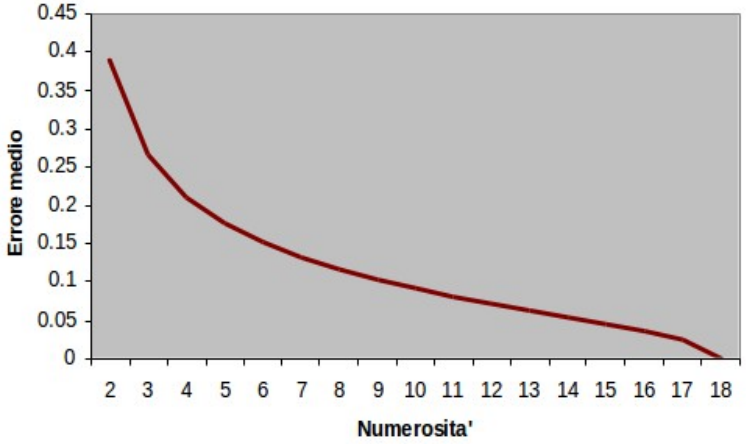
Errore standard diminuisce all’aumentare della dimensione del campione ed aumenta all’aumentare della variabilità

Aumenta all’aumentare della varianza

$$ES = \sqrt{\frac{var(S)}{N}}$$



Diminuisce all’aumentare della numerosità



L’errore standard: conseguenze

- Se il carattere stimato ha poca variabilità, campioni piccoli possono dare buone stime

- Se il carattere stimato ha molta variabilità, campioni grandi sono necessari

Numerosità e tipi di fenomeni studiati

- Un conseguenza importante di questo principio e' che tanto più generale (uguale per tutti) e' il fenomeno che stiamo studiando, tanto meno casi ci serviranno (e viceversa)

Fenomeni generali:

Fenomeni neurologici

Fenomeni chimici

Studi di morfologia del cervello

Studi di funzionalità

Fenomeni specifici:

Opinioni

Risposte comportamentali a stimoli complessi

Atteggiamenti

Intervallo di confidenza

- L'errore standard consente di calcolare l'intervallo di confidenza di una stima (come la media, la correlazione, il coefficiente di regressione)

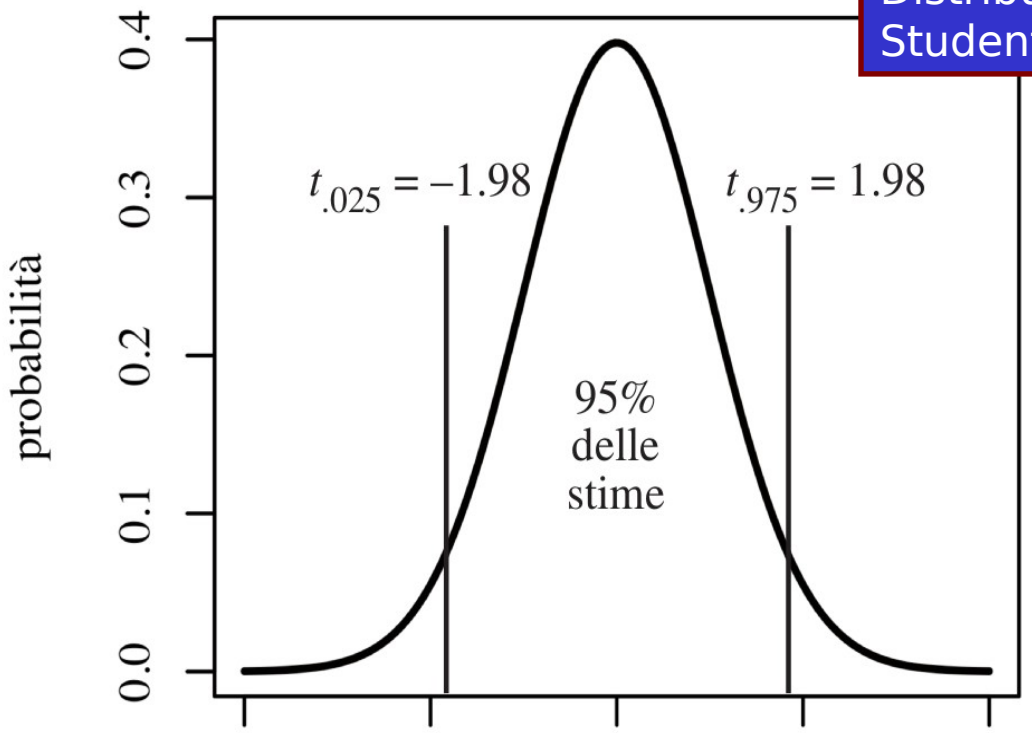
L’ intervallo di confidenza è un intervallo di valori plausibili per quel parametro (ad es., media) nella popolazione (ad es., bevitori di birra)

Dato che la nostra stima varia da campione a campione, IC indica in quale intervallo di valori è ragionevole che cada la stima ripetendo il campionamento

Intervallo di confidenza (IC)

L'IC è molto importante per capire i risultati ottenuti e cattura il concetto di ***accuratezza*** nella stima del parametro.

Distribuzione t di Student



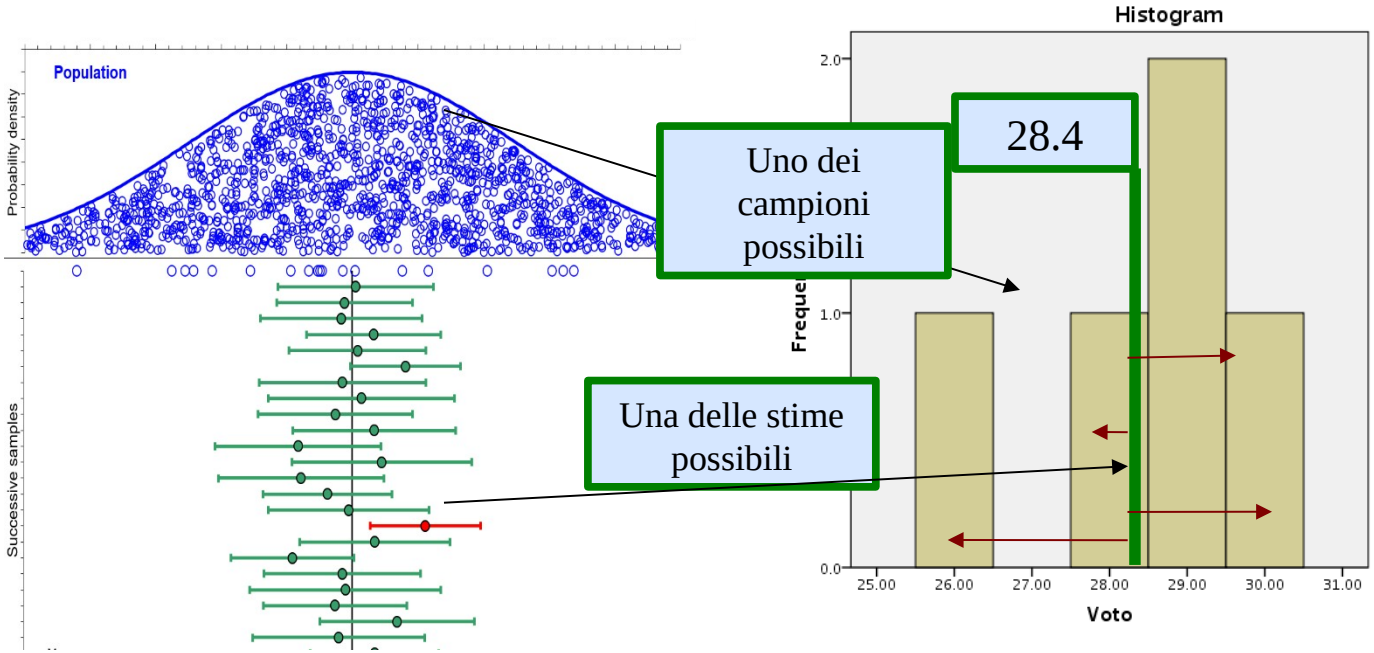
$$y_{.025} = \overline{y} + t_{.025} \times \sigma_{\overline{y}}$$

Errore standard

medio

Variabilità delle stime

Se continuiamo ad estrarre campioni dalla popolazione, le stime del parametro (media) varieranno in funzione della variabilità dei dati della numerosità campionaria del campione



Intervallo di confidenza

- Viene prodotto dal software

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	2.091	.684		3.057	.014	.543	3.638
	Birre	.709	.116	.898	6.132	.000	.448	.971

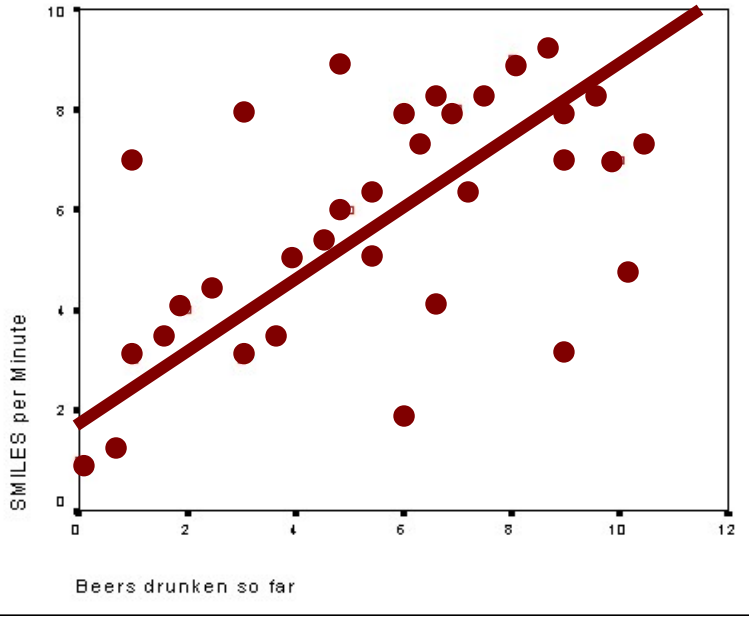
a. Dependent Variable: Sorrisi

Il coefficiente di regressione dell'esempio è "ragionevolmente", cioè con fiducia al 95% nell'intervallo .448-.971

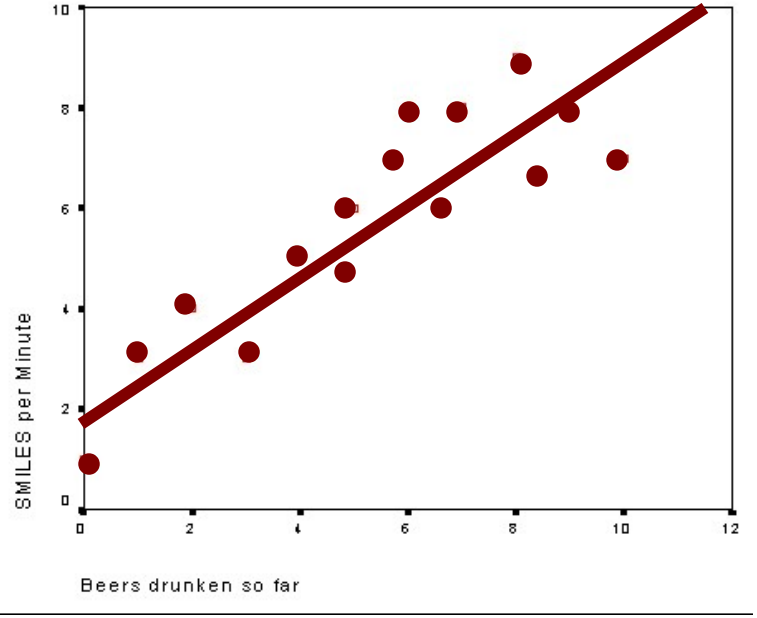
Bonta’ di adattamento

Non tutte le rette di regressione hanno lo stesso potere predittivo, cioè la stessa capacità di adattarsi ai dati osservati

bassa



alta



Errore di regressione

Notiamo che la predizione non corrisponde di norma ai valori osservati

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

predetti

errore

$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$

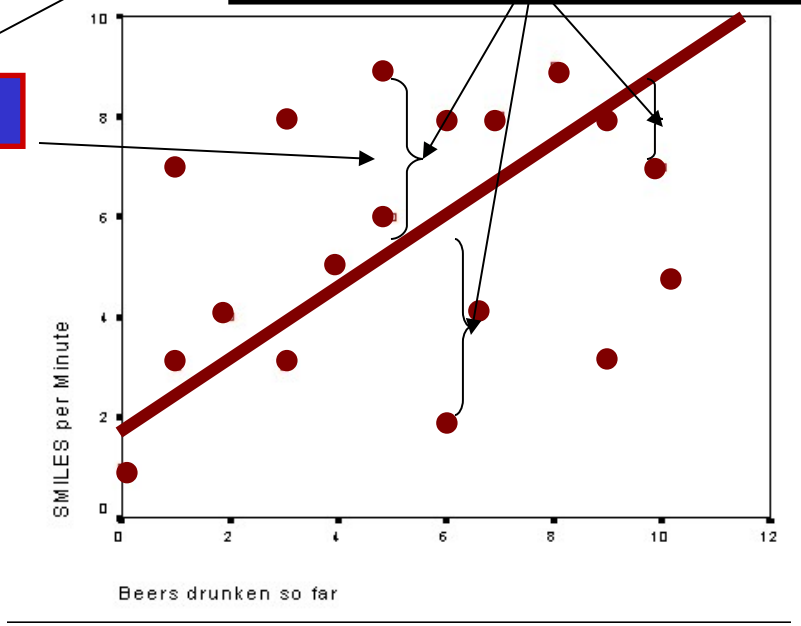
Dunque i valori osservati di Y possono essere espressi come somma dei valori predetti e l'errore

$$y_i = (a + b_{yx} x_i) + (y_i - \hat{y}_i)$$

retta

errore

Discrepanza osservati-predetti



Quanto e' grande l'errore di regressione

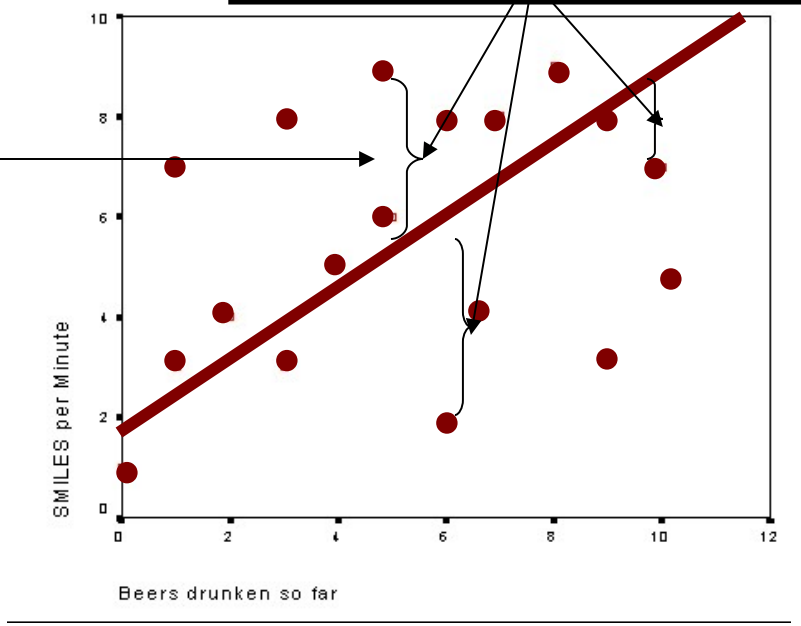
Calcoliamoci la distanza media tra i punti osservati e la retta

Le distanze si calcolano mediante le differenze al quadrato

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-1} = s_e^2$$

Notiamo che questa e' una varianza, che chiameremo varianza di errore

Discrepanza osservati-predetti



Proporzione riduzione errore

Il modello si adatterà ai dati tanto più riduce l'errore di predizione rispetto a non usare tale modello

- La logica è di confrontare due casi:
  - L'errore calcolato per la regressione data
  - L'errore associato alla media, cioè errore associato a non utilizzare la regressione

Proporzione riduzione errore

Senza regressione l'unica predizione plausibile di Y e' la media di Y

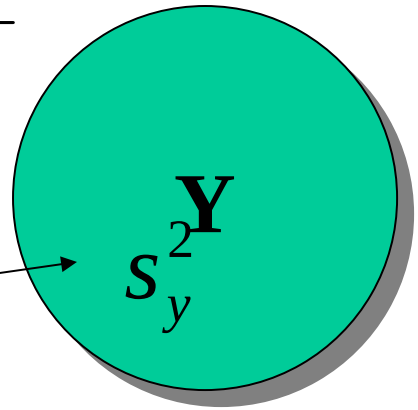
Predizione senza regressione

$$\hat{y}_i = M_y$$

Errore senza regressione

$$s_y^2 = \frac{\sum (y_i - M)^2}{n - 1}$$

Le deviazioni dalla media (la varianza) non siamo in grado di spiegarle



Proporzione riduzione errore

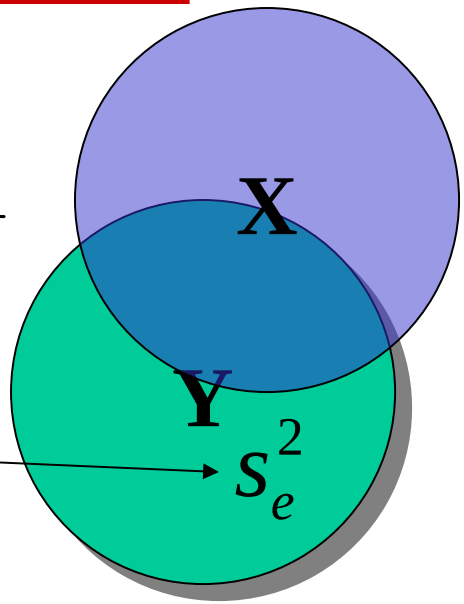
Con la regressione faremo una certa predizione

Predizione con regressione

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

Errore con regressione

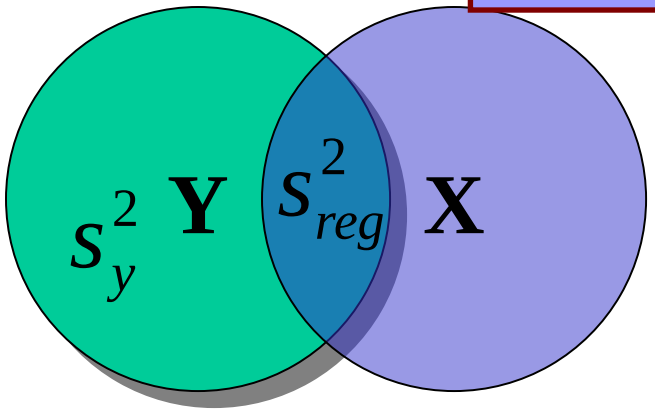
$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 1}$$



Le deviazioni dalla regressione (varianza di errore) non siamo in grado di spiegarle

R-quadro

Dunque il fit della regressione è tanto buono quanto riesce a migliorare la predizione, cioè a diminuire l'errore



Errore senza regressione

$$\frac{s_y^2 - s_e^2}{s_y^2} = \frac{s_{reg}^2}{s_y^2} = R_{yx}^2$$

Errore con regressione

Cioe': Quanto si riduce l'errore di predizione grazie al fatto che usiamo la regressione

Effetti Statistici

- Per effetto statistico si intende quanto il cambiamento di una o più variabili ha effetto sul cambiamento di un'altra variabile

- Interpretazione esplicativa: quanto siamo in grado di spiegare della variabilità di una variabile sulla base della variabilità delle altre (**basata sulle varianze**)

- Interpretazione predittiva: quanto siamo in grado di predirre della variabilità di una variabile basandoci sulla variabilità delle altre (**basata sui coefficienti**)

Fine della Lezione I