Statistica descrittiva

Corso di Laurea in Informatica

Corso di Statistica

Cristina Zucca

zucca@dm.unito.it

http://www2.dm.unito.it/paginepersonali/zucca/index.htm

Obiettivi della lezione:

- Statistica descrittiva: le variabili
- Frequenze: tabelle e grafici
- Indici di posizione, di dispersione e di forma
- Media e varianza di dati raggruppati
- Correlazione tra variabili
- Retta di regressione

Statistica

Descrittiva

Ho un insieme di dati e li voglio descrivere, sintetizzare e commentare

Induttiva

Ho un insieme di dati e li utilizzo per fare inferenza

Statistica descrittiva:

Abbiamo un insieme di dati che vogliamo sintetizzare e descrivere

Esempio 1: # mensile di interventi di manutenzione per un macchinario

1 6 3 1 3 2 2 1 2 6 3 0 1 4 3 2 1 3 1 2 2 1 2 4 3

Esempio 2: precipitazioni in pollici a Torino nel mese di aprile (20 giorni)

2.9	3.7	3.2	4.0	3.9	2.1	2.9	2.9	1.1
0.4	3.0	3.3	3.2	1.0	2.2	5.4	3.5	3.6
4.0	0.7							

Esempio 3: # di donne con un'occupazione professionale nel 1986 negli USA

A: Ingegneria/Informatica:	347
B: Sanità:	1937
C: Istruzione:	2833
D: Area sociale/legale:	698
E: Arte/Sport:	901
F: Altro:	355

N osservazioni = dati che vogliamo analizzare

Statistica descrittiva: le variabili

- **Variabili**
  - **Numeriche**
    - **Discrete** Esempio 1
    - **Continue** Esempio 2
  - **Categoriche** Esempio 3

Frequenze

Si considerino N dati da analizzare. I dati vengono suddivisi in un opportuno numero di classi; per ogni classe si ha:

Frequenza assoluta: ν_i numero di oggetti del tipo *i*-esimo

$$0 \le \nu_i \le N$$

$$\sum_i \nu_i = N$$

Frequenza relativa: $f_i = \frac{\nu_i}{N}$

$$0 \le f_i \le 1$$

$$\sum_i f_i = \sum_i \frac{\nu_i}{N} = 1$$

Frequenza percentuale: è la freq. relativa moltiplicata per 100

$$f_i \cdot 100$$

Frequenza cumulativa assoluta: N_i è la somma della freq. assoluta + la freq. cumulativa assoluta del dato precedente.

$$N_i = N_{i-1} + \nu_i = \sum_{k=0}^{i} \nu_k \qquad 0 \le N_i \le N$$

Frequenza cumulativa relativa: F_i è la somma della freq. relativa + la freq. cumulativa relativa del dato precedente.

$$F_i = F_{i-1} + f_i = \sum_{k=0}^{i} f_k \qquad 0 \le F_i \le 1$$

Tabella di distribuzioni di frequenze:

Caso discreto: Esempio 1

	A	B	C	D	E	F	G	H	I
1	Dati		Classi	freq ass	freq rel	freq cumul	freq cumul relativa	freq perc	freq perc cumul
2	1		0	1	0,04	1	0,04	4	4
3	6		1	7	0,28	8	0,32	28	32
4	3		2	7	0,28	15	0,6	28	60
5	1		3	6	0,24	21	0,84	24	84
6	3		4	2	0,08	23	0,92	8	92
7	2		5	0	0	23	0,92	0	92
8	2		6	2	0,08	25	1	8	100
9	1		TOT	25	1			100	

Le classi sono: $A_k = \{x_i \mid x_i = k\}$

N.B. : la somma delle freq.ass. = n° tot di osservazioni
la somma delle freq.rel. =1
la somma delle freq.perc.=100

Tabella di distribuzioni di frequenze:

Caso continuo: Esempio 2

Dati	Classi		freq ass	freq rel	freq cumul	freq cumul relativa	freq perc	freq perc cumul
2.9	0<x<=0,5	0.5	1	0.05	1	0.05	5	5
3.7	0,5<x<=1	1	2	0.1	3	0.15	10	15
3.2	1<x<=1,5	1.5	1	0.05	4	0.2	5	20
4	1,5<x<=2	2	0	0	4	0.2	0	20
3.9	2<x<=2,5	2.5	2	0.1	6	0.3	10	30
2.1	2,5<x<=3	3	4	0.2	10	0.5	20	50
2.9	3<x<=3,5	3.5	4	0.2	14	0.7	20	70
2.9	3,5<x<=4	4	5	0.25	19	0.95	25	95
1.1	4<x<=4,5	4.5	0	0	19	0.95	0	95
0.4	4,5<x<=5	5	0	0	19	0.95	0	95
3	5<x<=5.5	5.5	1	0.05	20	1	5	100
3.3		TOT	20	1			100	

Le classi non sono scelte in modo univoco.

Ogni osservazione deve appartenere a 1 sola classe!

N.B. Nelle tabelle di frequenza nel caso di variabili continue perdo dell'informazione ma ho un guadagno nella leggibilità dei dati. Nel caso di variabili discrete non c'è perdita di informazione!

Tabella di distribuzioni di frequenze:

Caso variabili categoriche: Esempio 3

	Classi	freq ass	freq rel	freq perc
2	Ing/Informatica	347	0,04907	4,90737
3	Sanità	1937	0,27394	27,3936
4	Istruzione	2833	0,40065	40,0651
5	Area sociale/legale	698	0,09871	9,87131
6	Arte/Sport	901	0,12742	12,7422
7	Altro	355	0,05021	5,02051
8	TOT	7071	1	100

I dati sono già raggruppati in classi

N.B. Per variabili categoriche NON ha senso parlare di frequenze cumulative!!!

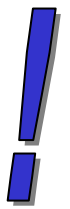
Tabella di distribuzioni di frequenze: Excel

La funzione FREQUENZA calcola la frequenza relativa di occorrenza dei valori di un intervallo e restituisce una matrice verticale di numeri.

Sintassi:

FREQUENZA(matrice_dati; matrice classi)

tale istruzione viene inserita come formula matrice dopo aver selezionato un intervallo di celle adiacenti nel quale dovrà apparire il risultato.



Il numero di elementi nella matrice restituita è maggiore di un'unità rispetto al numero di elementi contenuti in matrice_classi

EXCEL: Formule in forma di matrice

Una formula in forma di matrice può eseguire più calcoli e restituire uno o più risultati.

Procedura:

selezionare la cella o le celle in cui si desidera immettere la formula, creare la formula e premere CTRL+MAIUSC+INVIO

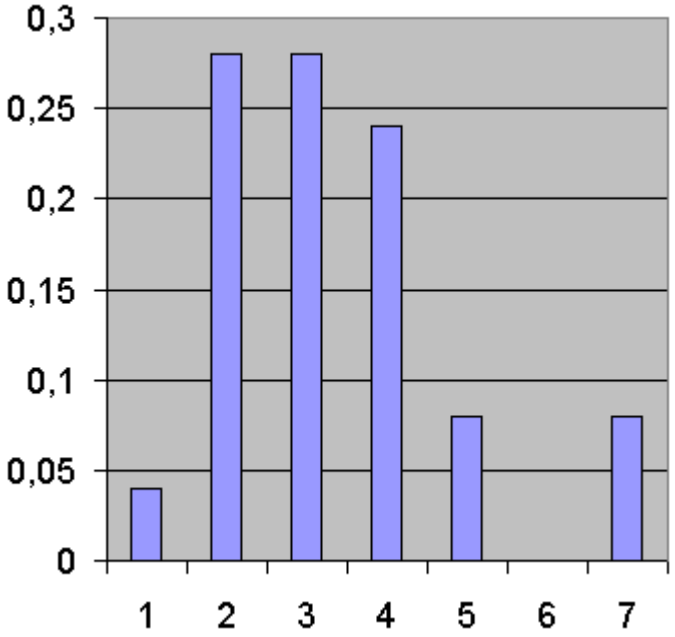
Grafici di distribuzioni di frequenze:

Istogramma

Caso discreto: Esempio 1

Si fissano sull'asse delle ascisse i valori delle classi e, in corrispondenza, si disegna una barra la cui altezza è pari alla frequenza (relativa o assoluta)

L'altezza ha la stessa unità di misura della probabilità teorica



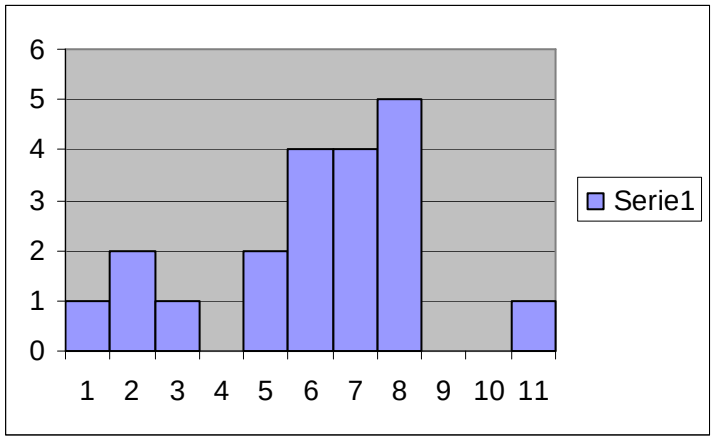
Istogramma

Caso continuo: Esempio 2

Si disegnano rettangoli adiacenti, le cui basi sono gli intervalli che definiscono le classi e le altezze sono date dalle frequenze (relative o assolute)

L'altezza NON ha la stessa unità di misura della probabilità teorica

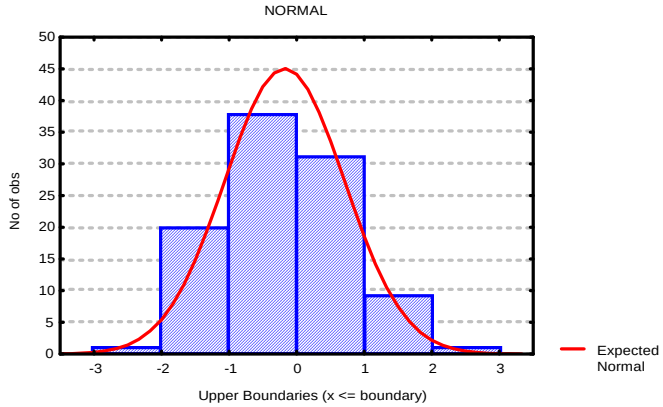
L'AREA ha la stessa unità di misura della probabilità ⇒ l'altezza del rettangolo deve essere proporzionale al quoziente tra la frequenza della classe e l'ampiezza dell'intervallo che la definisce



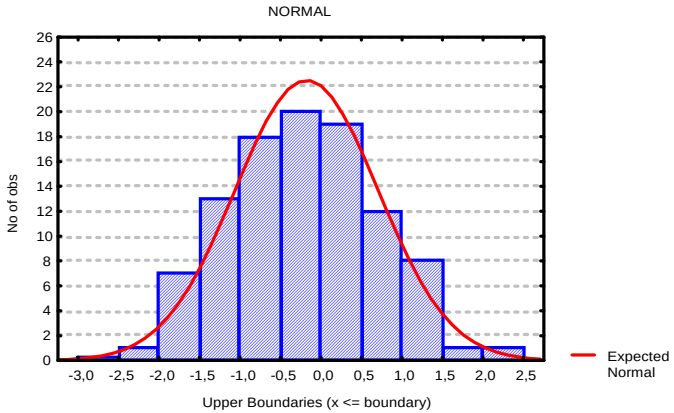
Istogramma delle frequenze assolute

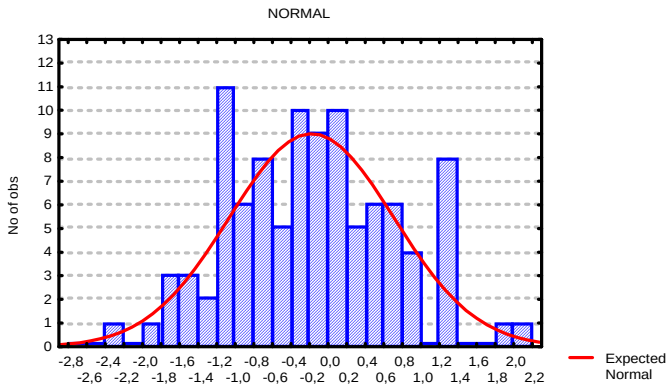
Istogramma: quante classi scelgo?
Taglia campione: n=100

6 classi



10 classi





36 classi

Regola pratica:

n.classi ≈ √n

Grafici di distribuzioni di frequenze:

Diagramma a barre /di Pareto: Esempio 3

Viene utilizzato nel caso di distribuzioni categoriche, ad ogni classe corrisponde una barra la cui altezza ne indica la frequenza mentre la base (uguale per ogni classe) non ha significato.

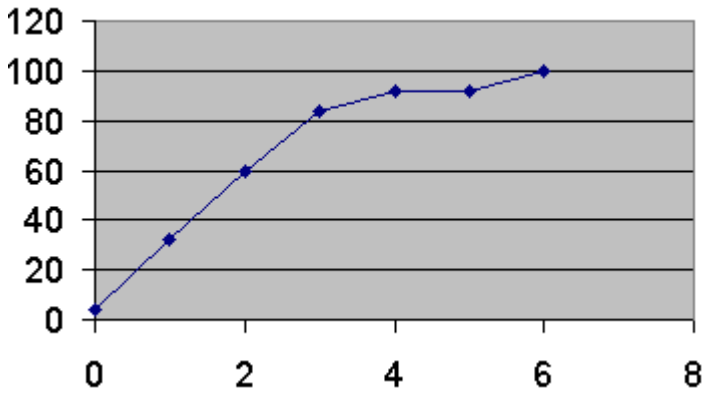
Ogiva: grafico delle frequenze cumulative di v. numeriche (discrete o continue)

Sulle ordinate si riportano le frequenze cumulative

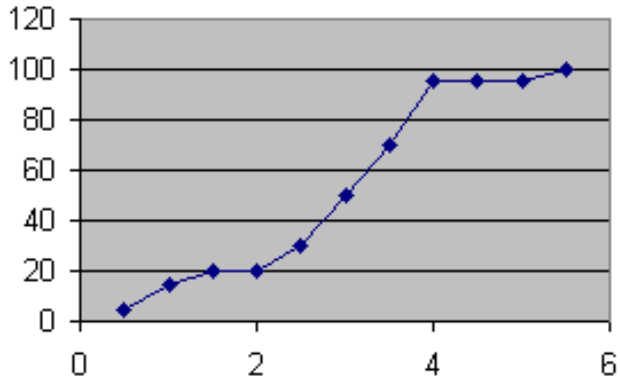
Sulle ascisse si riportano (caso discreto) i valori osservati

(caso continuo) gli estremi degli intervalli di variabilità

Si uniscono con una spezzata i punti ottenuti.



Es1



Es2

EXCEL: ANALISI DATI

STRUMENTI DI ANALISI è un insieme di strumenti di analisi dei dati che consente di ridurre i passaggi necessari allo sviluppo di complesse analisi statistiche. Forniti i dati e i parametri per ciascuna analisi, lo strumento utilizzerà le funzioni macro statistiche appropriate, visualizzando i risultati in una tabella di output.

Per visualizzare un elenco degli strumenti di analisi:

scegliere Analisi dati dal menu Strumenti. Se tale comando non è visualizzato, dal menu Strumenti selezionare Aggiunte… e scegliere Analisi dati.

EXCEL: Strumento di analisi Istogramma

Consente di calcolare le frequenze individuali e cumulative per un intervallo di celle e di classi di dati.

Opzioni della finestra di dialogo Istogramma:

- **intervallo di input**: immettere il riferimento di cella per l'intervallo di dati da analizzare
- **intervallo di classe (facoltativo):** immettere un intervallo di celle contenente un insieme di valori limite che definiscano gli intervalli delle classi
- **intervallo di output**: immettere il riferimento della cella superiore sinistra della tabella di output

Principali indici statistici

I grafici finora analizzati ci danno informazioni qualitative; possiamo quantificarle ricorrendo ai seguenti indici.

Siano $x_1, x_2, \ldots, x_n$ n osservazioni numeriche

INDICI

- **di posizione**
  - **MODA**
  - **MEDIANA**
  - **MEDIA**
- **di dispersione**
  - **SCARTO QUADRATICO MEDIO**
  - **VARIANZA**
  - **RANGE**
- **di forma**
  - **ASIMMETRIA (SKEWNESS)**
  - **CURTOSI (KURTOSIS)**

Indici di posizione:

MODA

E' definita come il valore che ha la frequenza più alta.

MEDIA

E' quel valore che corrisponde alla somma di tutti i valori diviso il numero dei valori stessi.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

dove:
X_i = esito i-ma misura
n = numero dei dati (taglia del campione)

MEDIANA

E' quel valore al di sotto del quale cadono la metà dei valori campionari.

Gli indici di posizione indicano attorno a quale valore il campione dei dati e' posizionato ⇒ mi interessa la dispersione dei dati intorno a tali valori

N.B. NELLA DISTRIBUZIONE NORMALE
MEDIA= MODA = MEDIANA

Indici di dispersione:

$x_{max} - x_{min}$ RANGE (Campo di variazione)

$\frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|$ SCARTO MEDIO ASSOLUTO

$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ MEDIA DEI QUADRATI DEGLI SCARTI

$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ VARIANZA

$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ DEVIAZIONE STANDARD

p_esimo quantile/ 100p_esimo percentile: si considera np

Se np non è intero, considero k l'intero successivo e il p_esimo quantile è x_k

Se $np=k$ è intero, il p_esimo quantile è $(x_k + x_{k+1})/2$

Q_1=primo quartile =25°percentile

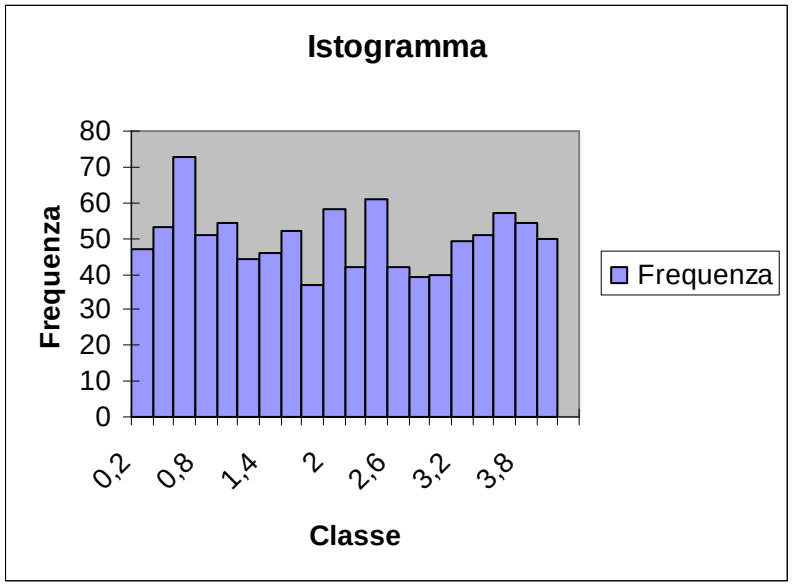
Q_2=secondo quartile =50°percentile=mediana

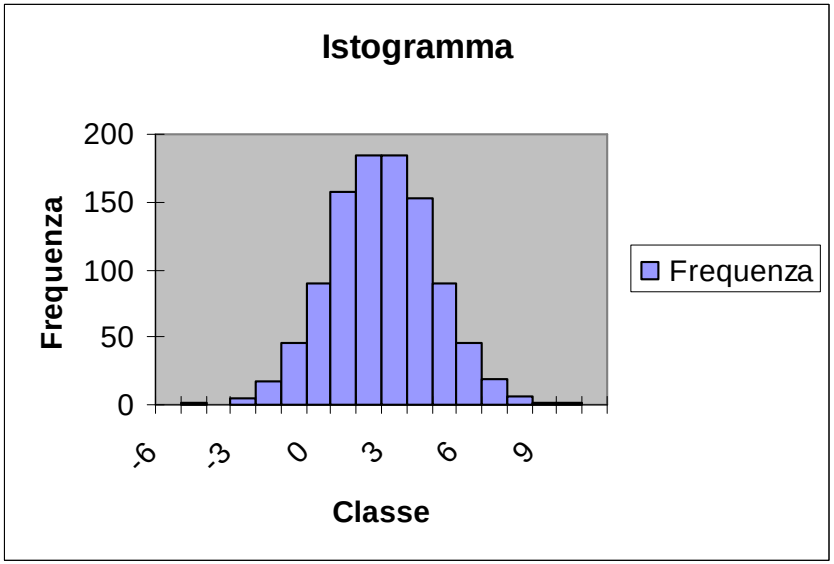
Q_3=terzo quartile =75°percentile

Media e varianza:

Media uguale

Deviazione Standard Diversa





Media=2
Varianza=1.33

Media=2
Varianza=4

Indici di forma

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{n\sigma^3}$$

INDICE DI ASIMMETRIA (Skewness)

>0 coda a destra

<0 coda a sinistra

=0 simmetrica

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{n\sigma^4}$$

CURTOSI

Misura quanto la distribuzione è appuntita

>3 poco appuntita

=3 caso della distribuzione normale

<3 molto appuntita

N.B. In molti software il coeff. di curtosi viene confrontato con il valore 0

Indici: Schema riassuntivo

di posizione

- **media:** $\bar{x} = \frac{\sum_i x_i}{N}$
- **moda:** punto di max della distribuzione
- **mediana:** valore sotto al quale cadono la metà dei valori campionari. Si dispongono i dati in ordine crescente e si prende quello che occupa la posizione centrale (*N* dispari) o la media dei 2 valori in posizione centrale (*N* pari)

di dispersione

- **varianza** $\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{N-1}$
- **deviazione standard** σ
- **range** $R = x_{\max} - x_{\min}$

di di forma

- **skewness (coeff. di asimmetria)** $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^3}{N}$

>0 coda a ds

<0 coda a sin

=0 simmetrica

- **curtosi:** misura quanto la distribuzione è appuntita $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^4}{N}$

>3 poco appuntita <3 molto appuntita

EXCEL: Strumento di analisi Statistica descrittiva

Fa un'analisi statistica dei dati selezionati fornendo informazioni sulla tendenza e dispersione dei dati

Opzioni della finestra di dialogo Statistica descrittiva:

- **intervallo di input**: immettere il riferimento di cella per l'intervallo di dati da analizzare

- **intervallo di output**: immettere il riferimento della cella superiore sinistra della tabella di output

- **Riepilogo statistiche:** genera una tabella di output con le seguenti statistiche:Media, Errore standard (della media), Mediana, Moda, Dev. Standard, Varianza, Curtosi, Asimmetria, Intervallo, Min, Max, Somma Conteggio.

Media e varianza di dati raggruppati

Supponiamo di avere a disposizione solo la tabella di distribuzione delle frequenze (**dati raggruppati**) di dati continui.

Il calcolo diretto di media e varianza **NON** è più possibile!!!

Siano $t_1,...,t_k$ i punti medi degli intervalli che definiscono le classi e siano ν_i le frequenze assolute di ogni classe

Classi	ti	*v*i
0<x<=1	0,5	1
1<x<=2	1,5	0
,,,	,,,	,,,

Media $\bar{x} = \frac{\sum_{i=1}^{k} t_i \nu_i}{N}$

Varianza $\sigma^2 = \frac{\sum_{i=1}^{k} (t_i - \bar{x})^2 \nu_i}{N} = \frac{1}{N} \sum_{i=1}^{k} t_i^2 \nu_i - \bar{x}^2$

Media e varianza di dati raggruppati

Classi	freq assolute
0<x<=1	3
1<x<=2	1
2<x<=3	6
3<x<=4	9
4<x<=5	0
5<x<=6	1

Utilizzo i dati raggruppati dell'Esempio 2

Classi	ti	freq assolute: vi	ti*vi	ti^2	ti^2*vi
0<x<=1	0,5	3	1,5	0,25	0,75
1<x<=2	1,5	1	1,5	2,25	2,25
2<x<=3	2,5	6	15	6,25	37,5
3<x<=4	3,5	9	31,5	12,25	110,25
4<x<=5	4,5	0	0	20,25	0
5<x<=6	5,5	1	5,5	30,25	30,25
TOT		20	55		181
media=	2,75				
varianza=	1,4875				

Media campionaria = 2.85
Varianza campionaria = 1.61

Correlazione tra variabili

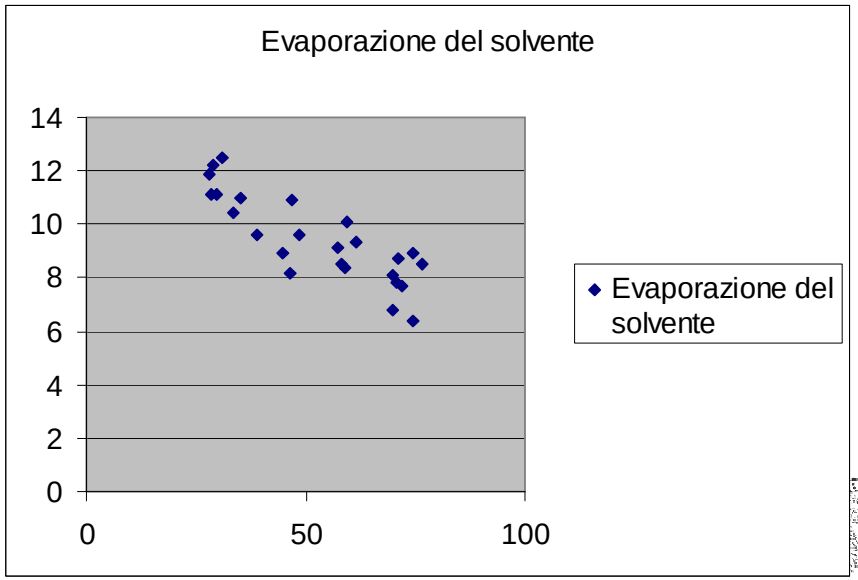
Finora abbiamo considerato una variabile alla volta, ora tratteremo analisi di tipo comparativo:

a. Osservo una variabile su piu' gruppi di individui
b. Osservo piu' variabili su un gruppo di individui
c. Entrambe le situazioni a. e b.

Esiste correlazione tra le variabili?

Scatterplot, diagramma a dispersione

Umidita'	Evaporazione del solvente
35,3	11
29,7	11,1
30,8	12,5
58,8	8,4
61,4	9,3
71,3	8,7
74,4	6,4
76,7	8,5
70,7	7,8
57,5	9,1
46,4	8,2
28,9	12,2



Indici di variazione bidimensionali

Date *n* osservazioni congiunte di 2 variabili $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Covarianza campionaria

$$c_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

•Se $c_{x,y}>0$ a valori grandi (piccoli) di *x* corrispondono valori grandi (piccoli) di *y*

x e *y* sono direttamente correlate

•Se $c_{x,y}<0$ a valori grandi (piccoli) di *x* corrispondono valori piccoli (grandi) di *y*

x e *y* sono inversamente correlate

•Se $c_{x,y}=0$ le variabili non sono correlate

Indici di variazione bidimensionali

Indice di correlazione

$$r = \frac{c_{x,y}}{\sigma_x \sigma_y}$$

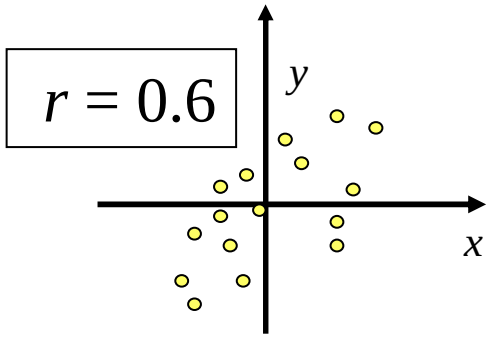
Date *n* osservazioni congiunte di 2 variabili $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

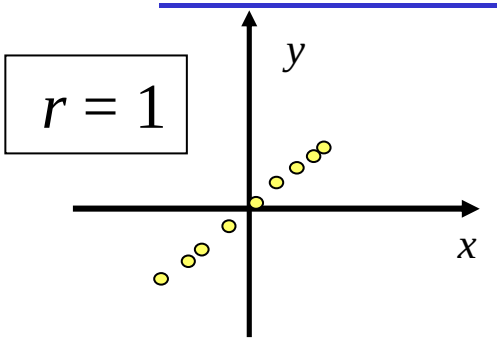
$$|r| \leq 1, \text{ cioè } -1 \leq r \leq 1$$

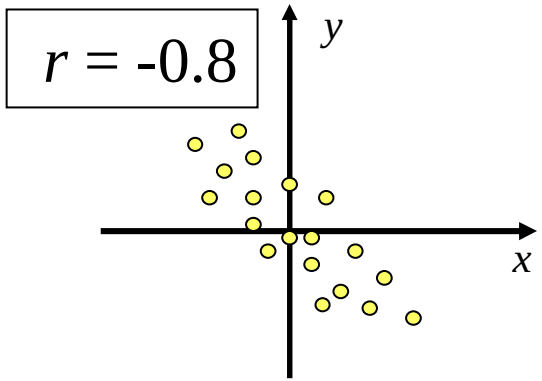
In particolare, $r = \pm 1 \Leftrightarrow \exists\, a, b$ costanti tali che $y_i = ax_i + b$

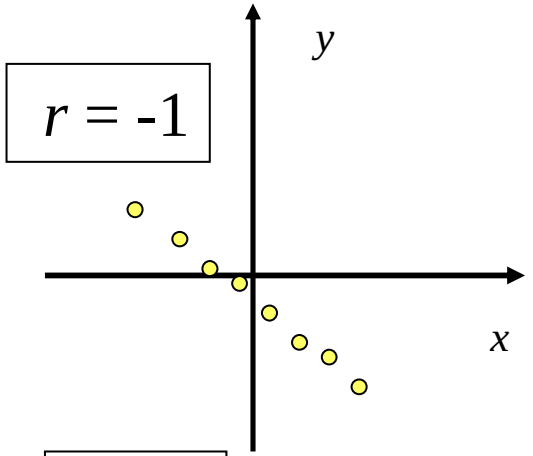
dove il segno di *r* =segno di *a*

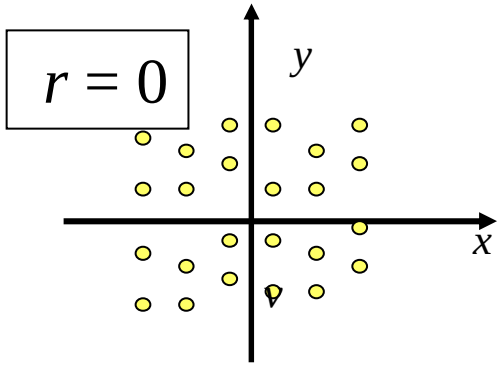
Diagramma di dispersione e indice di correlazione

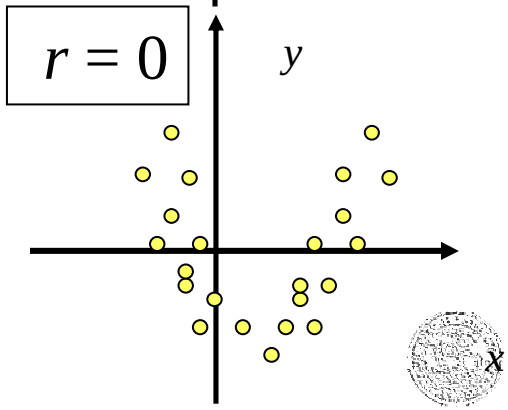








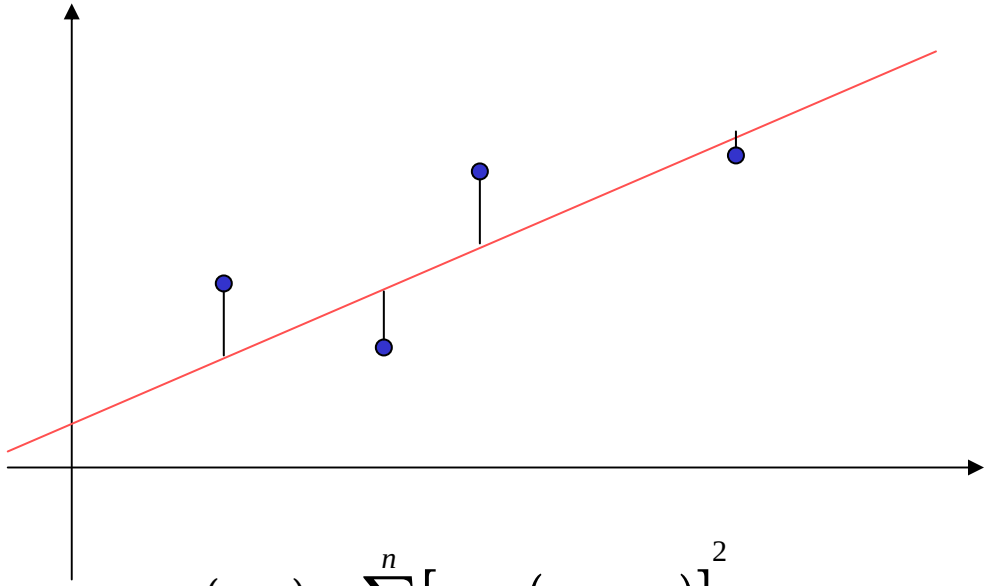




Regressione lineare: retta di regressione

Si vuole cercare la relazione lineare tra due variabili *x* e *y*.

Date *n* osservazioni congiunte di 2 variabili $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ cerco due coefficienti *a* e *b* tali che $y=ax+b$ passi il più possibile vicino a questi punti.



Cerco *a* e *b* tali che $f(a,b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2$ sia minima

(Metodo dei minimi quadrati)

Retta di regressione

$$y = \hat{a}x + \hat{b}$$

$$\hat{a} = \frac{c_{x,y}}{\sigma_x^{\,2}} \qquad \hat{b} = \bar{y} - \bar{x}\frac{c_{x,y}}{\sigma_x^{\,2}}$$

N.B. Il coefficiente angolare della retta ha il segno di $c_{x,y}$

Utilizzando le informazioni ottenute tramite lo scatterplot e il coefficiente di correlazione, parto dal presupposto che ci sia relazione lineare tra *x* e *y*

Valori stimati: $\hat{y}_i = \hat{a}x_i + \hat{b}$

Residui: $r_i = y_i - \hat{y}_i$

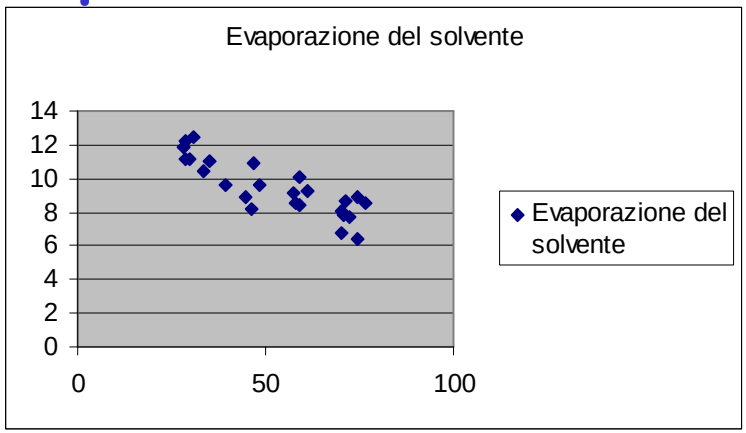
Utilizzando la retta di regressione posso fare delle previsioni

EXCEL: Retta di regressione

Umidita'	Evaporazione del solvente
35,3	11
29,7	11,1
30,8	12,5
58,8	8,4
61,4	9,3
71,3	8,7
74,4	6,4
76,7	8,5
70,7	7,8
57,5	9,1
46,4	8,2
28,9	12,2
28,1	11,9

Esercizio: Stabilire se c'e' dipendenza lineare tra l'umidita' del magazzino e l'evaporazione di un certo componente chimico.

Step1: Scatterplot



Step2: Coefficiente di correlazione

Utilizzando la funzione =CORRELAZIONE(dati_1;dati_2)

ottengo $r = -0.84695$

⇒ Ha senso determinare la retta di regressione

EXCEL: Retta di regressione

Step3: Retta di regressione

Avendo gia' lo scatterplot seleziono: Grafico-Aggiungi linea di tendenza

$$y = -0{,}0801x + 13{,}639$$