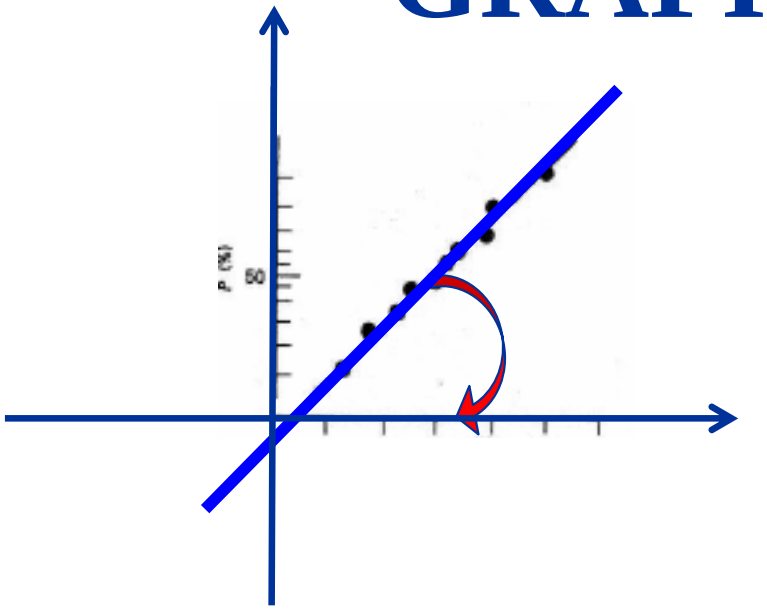
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

DIPARTIMENTO DI INGEGNERIA AEROSPAZIALE – D.I.A.S.

STATISTICA PER L'INNOVAZIONE

a.a. 2007/2008

GRAFICI DI PROBABILITÀ



Prof. Antonio Lanzotti

A cura di: Ing. Giovanna Matrone

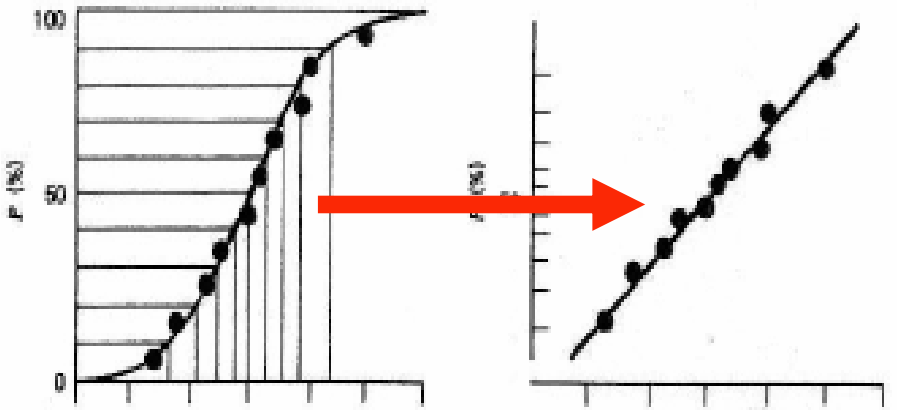
giovanna.matrone@unina.it

Analisi dei dati: la carta di probabilità

Una carta di probabilità rappresenta uno strumento grafico di grande utilità operativa. Essa consente di effettuare un'analisi "visiva" del set di dati a disposizione (osservazioni sperimentali), permettendo di testare la bontà di adattamento del modello di Cdf ipotizzato alle osservazioni sperimentali.

In particolare essa consente di vedere quanto le singole osservazioni sperimentali si discostino dal modello di Cdf ipotizzato.

Per una più agevole valutazione grafica degli scostamenti dei dati rispetto alla modello ipotizzato è possibile "*linearizzare*" il modello ipotizzato di Cdf che, esprimendo i valori cumulati della funzione distribuzione di probabilità, presenta un andamento sinusoidale.

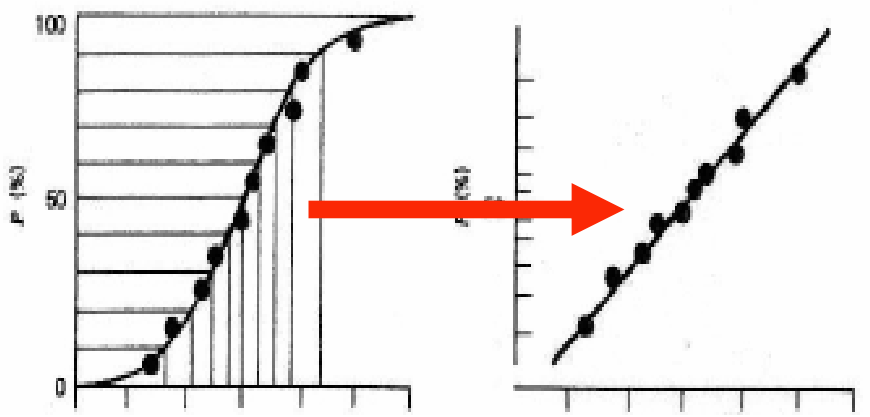


Analisi dei dati: la carta di probabilità

Il processo di linearizzazione è possibile grazie ad opportune relazioni funzionali, per poter poi valutare, successivamente, la capacità dei dati a disporsi in modo lineare sulla carta di probabilità.

L'idea alla base di una carta di probabilità è, infatti, la possibilità di applicare una opportuna trasformazione di scala dell'asse delle ordinate del grafico in modo tale da ottenere una Cdf che si distribuisca secondo una *linea retta*.

- Dati disomogenei
- Sistematica discordanza dei dati rispetto al modello ipotizzato
- "Peso" di alcuni specifici dati sulla stima dei parametri del modello
- Presenza di sottopopolazioni
- …



Linearizzazione della Cdf

Per linearizzare una Cdf è necessario individuare una opportuna relazione del tipo:

$$Y(x) = \frac{x-a}{b} \;\; ; \;\; b > 0$$

che esprime la linearizzazione della $F_X(x)$ in funzione delle determinazioni x di X (o di una sua trasformata); a e b sono rispettivamente parametro 'di posizione' e parametro 'di scala'.

- Supponiamo di voler linearizzare la Cdf Esponenziale:

$$F_X(x) = 1 - e^{-\lambda x}$$

Arriveremo ad un'espressione del tipo:

$$Y(x) = \ln\left[\frac{1}{1-F_X(x)}\right] = \lambda \cdot x \quad ; \quad \begin{matrix} a = 0 \\ b = \frac{1}{\lambda} \end{matrix}$$

Linearizzazione della distribuzione Gumbel (max)

$$F_X(x) = \exp\left[-\exp\left(-\frac{x-a}{b}\right)\right]$$

$$-\ln\left[F_X(x)\right] = \exp\left(-\frac{x-a}{b}\right);$$

$$-\ln\left[-\ln\left[F_X(x)\right]\right] = \frac{x-a}{b};$$

$$Y = \beta_1 \cdot X + \beta_0 \quad \text{con} \quad \beta_1 = \frac{1}{b} \text{ e } \beta_0 = -\frac{a}{b}$$

La Cdf Gumbel (dei max) è così linearizzabile ed il grafico corrispondente sarà ***semi-logaritmico*** poiché l'asse delle ascisse avrà una scala lineare e l'asse delle ordinate una scala logaritmica del tipo: $-\ln\left[-\ln\left[F_X(x)\right]\right]$

Costruzione della carta

La carta si caratterizza nel modo seguente:

- sull'asse delle ascisse vengono riportati i dati sperimentali $x_{(i)}$, ordinati in senso crescente $[x_{(1)} = x_{min} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n)} = x_{max}]$;
- sull'asse delle ordinate sono riportati i valori della Cdf campionaria, per la cui stima adottiamo la frazione:

$$\frac{i}{n+1}$$

(in letteratura sono suggerite anche altre frazioni di stima come : $\frac{i}{n}$ o $\frac{i-0.5}{n}$)

- si dispongono sulla carta le coppie di valori $\left(x_{(i)}, \hat{F}_X\left(x_{(i)}\right)\right)$;
- si traccia la retta che meglio approssima i punti. Il modello, allora, sarà tanto più adeguato quanto più la sequenza dei punti approssimerà la linea retta.

Costruzione della carta

A partire dal grafico è inoltre possibile ottenere delle stime grafiche approssimate dei parametri rappresentativi della Cdf ipotizzata.

D'altra parte, disponendo delle coppie di valori, è possibile calcolare la retta che minimizza i quadrati degli scarti dei singoli punti dalla retta con il Metodo dei Minimi Quadrati.

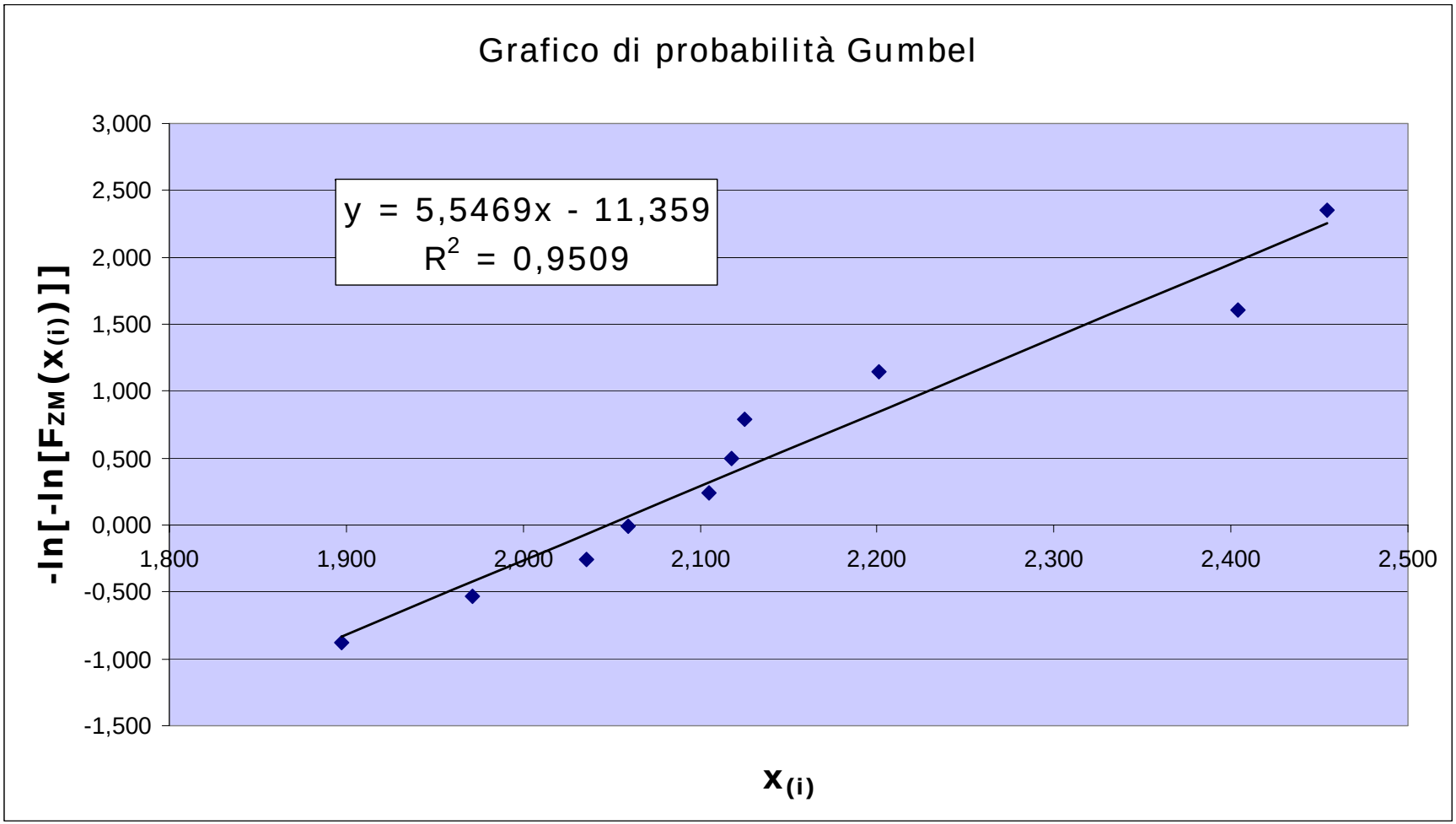
In questo modo sarà possibile calcolare il coefficiente di correlazione ρ attraverso il quale "quantificare" la bontà della scelta del modello.

Esempio N.1

	x_i Dati
1	2,1252313
2	2,0355379
3	2,4541777
4	2,2007804
5	2,4041944
6	1,9711134
7	2,1047979
8	1,8976315
9	2,0595329
10	2,1179674

N. d'ordine		$x_{(i)}$ Dati	*i*	$F_{ZM}(x_{(i)})$	$-\ln[-\ln[F_{ZM}(x_{(i)})]]$
(1)	8	1,898	*1*	0,091	-0,875
(2)	6	1,971	*2*	0,182	-0,533
(3)	2	2,036	*3*	0,273	-0,262
(4)	9	2,060	*4*	0,364	-0,012
(5)	7	2,105	*5*	0,455	0,238
(6)	10	2,118	*6*	0,545	0,501
(7)	1	2,125	*7*	0,636	0,794
(8)	4	2,201	*8*	0,727	1,144
(9)	5	2,404	*9*	0,818	1,606
(10)	3	2,454	*10*	0,909	2,351

Esempio N.1



b= 0,1803

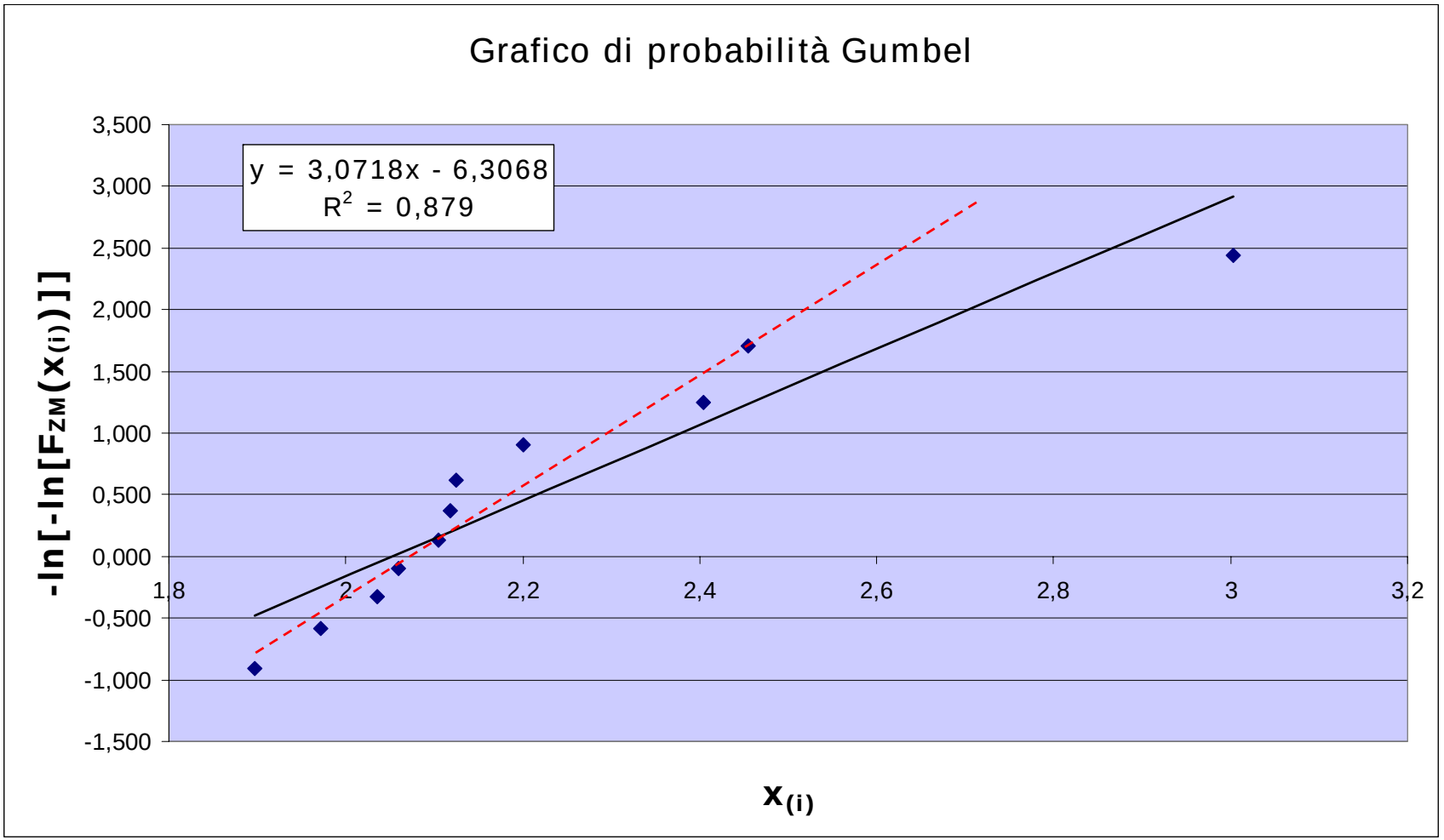
a= 2,0478

Esempio N.2

	x_i Dati
1	2,1252313
2	2,0355379
3	2,4541777
4	2,2007804
5	2,4041944
6	1,9711134
7	2,1047979
8	1,8976315
9	3,002381
10	2,0595329
11	2,1179674

N. d'ordine		$x_{(i)}$ Dati	*i*	$F_{ZM}(x_{(i)})$	$-\ln[-\ln[F_{ZM}(x_{(i)})]]$
(1)	8	1,8976315	*1*	0,083	-0,910
(2)	6	1,9711134	*2*	0,167	-0,583
(3)	2	2,0355379	*3*	0,250	-0,327
(4)	10	2,0595329	*4*	0,333	-0,094
(5)	7	2,1047979	*5*	0,417	0,133
(6)	11	2,1179674	*6*	0,500	0,367
(7)	1	2,1252313	*7*	0,583	0,618
(8)	4	2,2007804	*8*	0,667	0,903
(9)	5	2,4041944	*9*	0,750	1,246
(10)	3	2,4541777	*10*	0,833	1,702
(11)	9	3,002381	*11*	0,917	2,442

Esempio N.2



b= 0,3255

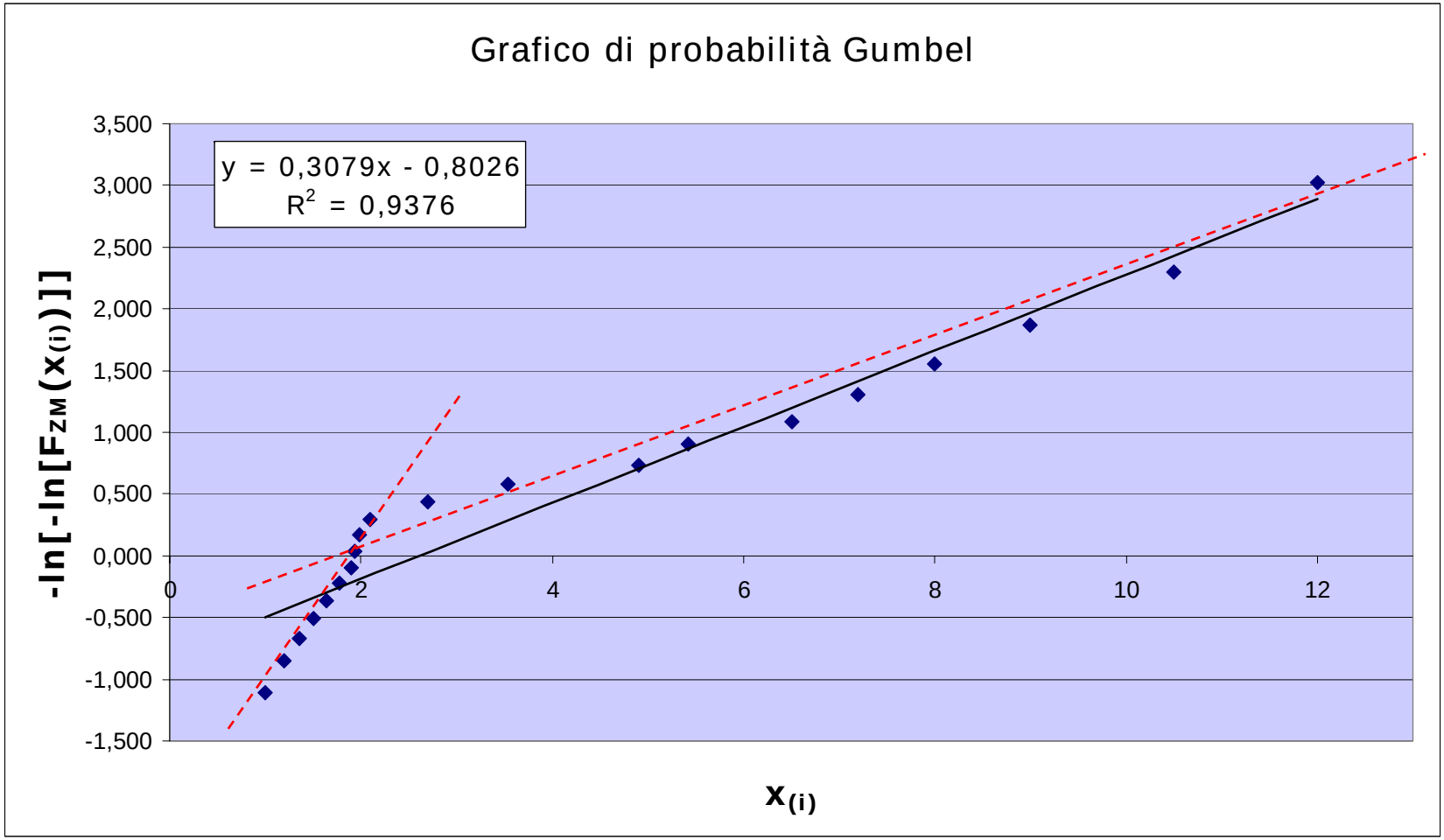
a= 2,0531

Esempio N.3

	x_i Dati
1	1,93
2	7,2
3	10,5
4	8
5	1,2
6	3,54
7	5,42
8	1,35
9	4,9
10	1,78
11	1,901
12	1,64
13	2,7
14	6,5
15	12
16	1,5
17	1
18	2,1
19	1,99
20	8,99

N. d'ordine		$x_{(i)}$ Dati	*i*	$F_{ZM}(x_{(i)})$	$-\ln[-\ln[F_{ZM}(x_{(i)})]]$
(1)	17	1	*1*	0,048	-1,113
(2)	5	1,2	*2*	0,095	-0,855
(3)	8	1,35	*3*	0,143	-0,666
(4)	16	1,5	*4*	0,190	-0,506
(5)	12	1,64	*5*	0,238	-0,361
(6)	10	1,78	*6*	0,286	-0,225
(7)	11	1,901	*7*	0,333	-0,094
(8)	1	1,93	*8*	0,381	0,036
(9)	19	1,99	*9*	0,429	0,166
(10)	18	2,1	*10*	0,476	0,298
(11)	13	2,7	*11*	0,524	0,436
(12)	6	3,54	*12*	0,571	0,581
(13)	9	4,9	*13*	0,619	0,735
(14)	7	5,42	*14*	0,667	0,903
(15)	14	6,5	*15*	0,714	1,089
(16)	2	7,2	*16*	0,762	1,302
(17)	4	8	*17*	0,810	1,554
(18)	20	8,99	*18*	0,857	1,870
(19)	3	10,5	*19*	0,905	2,302
(20)	15	12	*20*	0,952	3,020

Esempio N.3



b= 3,2478

a= 2,6067

Applicazione della carta (Gumbel max)

I dati riportati in tabella sono i valori massimi osservati delle portate annuali (piene) del fiume Feather (Nord California, USA) in ft^3/sec^1 dal 1902 al 1960, per un totale di $n = \mathbf{59}$ rilevazioni:

i	ANNO	PIENA	i	ANNO	PIENA	i	ANNO	PIENA	i	ANNO	PIENA
1	1902	42000	16	1917	80400	31	1932	22600	46	1947	45600
2	1903	102000	17	1918	28200	32	1933	8860	47	1948	36700
3	1904	118000	18	1919	65900	33	1934	20300	48	1949	16800
4	1905	81000	19	1920	23400	34	1935	58600	49	1950	46400
5	1906	128000	20	1921	62300	35	1936	85400	50	1951	92100
6	1907	230000	21	1922	36400	36	1937	19200	51	1952	59200
7	1908	16300	22	1923	22400	37	1938	185000	52	1953	113000
8	1909	140000	23	1924	42400	38	1939	8080	53	1954	54800
9	1910	31000	24	1925	64300	39	1940	152000	54	1955	13000
10	1911	75400	25	1926	55700	40	1941	84200	55	1956	203000
11	1912	16400	26	1927	94000	41	1942	110000	56	1957	83100
12	1913	16800	27	1928	185000	42	1943	108000	57	1958	102000
13	1914	122000	28	1929	14000	43	1944	24900	58	1959	34500
14	1915	81400	29	1930	80100	44	1945	60100	59	1960	135000
15	1916	42400	30	1931	11600	45	1946	54400			

Fonte: (Reiss and Thomas, 1997)

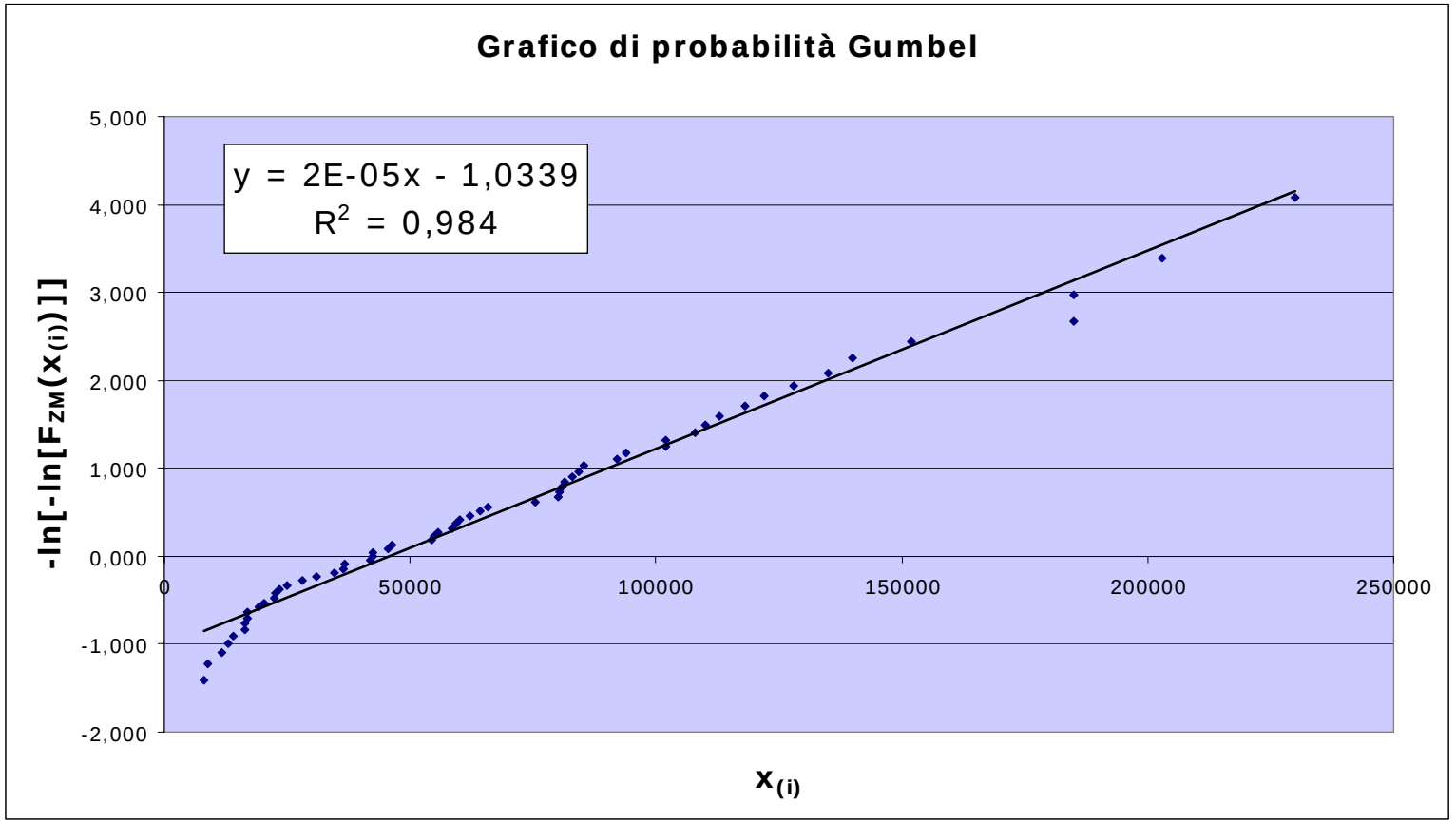
Nota: $1\ ft^3/sec = 0.02\ m^3/sec$

Applicazione della carta (Gumbel max)

- Supponiamo che i **59** dati x_i a disposizione siano i.i.d.
- Vogliamo:

1) Verificare che il modello Gumbel ben si adatti ai dati;
2) Determinare la probabilità che un generico **futuro valore massimo** di portata X_f sia superiore ad una fissata soglia x_{soglia}, supponendo che le x_f siano distribuite come le osservazioni precedenti;
3) Determinare il corrispondente periodo di ritorno T.

1) Grafico di probabilità Gumbel (max)



Poiché i punti nel grafico sono ben allineati, concludiamo che la distribuzione Gumbel è un modello adatto per i nostri dati (valori estremi massimi).

Siamo, quindi, in grado di ricavare anche le stime dei parametri:

$$\hat{a} \simeq 51695 \quad ; \quad \hat{b} \simeq 50000$$

2) Calcolo della probabilità di superamento

Fissiamo come valore soglia:

$$x_{soglia} = Max[x_i, \{i = 1,...,59\}] = 230000 \text{ ft}^3 / \text{sec};$$

Calcoliamo la probabilità che la v.a. X_f **futura portata annuale massima** superi il valore soglia fissato, utilizzando le stime dei parametri ricavate al punto 1):

$$\Pr\{X_f \le x_{soglia}\} = GumbelCdf\left(x_{soglia}; \hat{a}, \hat{b}\right) = 0.97213;$$

$$\Pr\{X_f > x_{soglia}\} = 1 - GumbelCdf\left(x_{soglia}; \hat{a}, \hat{b}\right) = 0.02787 \simeq 3.0\%$$

3) Il periodo di ritorno *T*

In campo ingegneristico è, spesso, necessario calcolare con un buon livello di approssimazione la probabilità di superamento di una certa soglia da parte di un evento.

Periodo di ritorno T

$$T = \frac{1}{1 - F_X\left(x_{soglia}\right)} \quad \text{con } F_X\left(x_{soglia}\right) = \text{Cdf di v.a. } X \text{ "valore massimo"}$$

Il periodo di ritorno *T* rappresenta il numero di unità di tempo (anni, settimane, ecc) che bisogna attendere prima che sia "normale" il verificarsi di un evento di entità superiore al valore soglia x_{soglia}.

3) Il periodo di ritorno *T*

$$T = \frac{1}{1 - F_X\left(x_{soglia}\right)} = \frac{1}{\Pr\left\{X_f > x_{soglia}\right\}} = \frac{1}{0.02787} \simeq 36 \text{ anni}$$

Era perciò necessario attendere 36 anni (a partire dal 1960) prima che fosse "normale" osservare una portata massima annuale del fiume Feather superiore alla piena massima osservata fino al 1960.

Riferimenti sul Libro

Pasquale Erto
“Probabilità e statistica per le scienze e l’ingegneria”
***McGraw Hill* – seconda edizione**

- **Capitolo 3**

  Trasformazioni di variabili aleatorie § 3.5 – pagg. 55-63

  Metodo dei minimi quadrati § 3.9.1 – pagg. 70-73

- **Capitolo 5**

  Modelli Gumbel e Weibull § 5.2.1 – pagg. 94-98

- **Capitolo 9**

  Metodo dei grafici di probabilità § 9.1.3 – pagg. 185-189

Per eventuali comunicazioni:

giovanna.matrone@unina.it
antonio.lepore@unina.it
andrea.colini@unina.it

Orario di ricevimento:

Giovedì ore 16:30-18:30
P.le Tecchio X piano
Stanza Dottorandi DIAS