Corso Integrato di Statistica Informatica e Analisi dei dati

Dr Carlo Meneghini

Dip. di Fisica "E. Amaldi"
via della Vasca Navale 84

meneghini@fis.uniroma3.it

http://webusers.fis.uniroma3.it/~meneghini

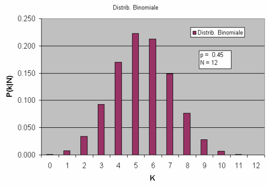
Confronto dati sperimentali e distribuzioni teoriche di probabilità

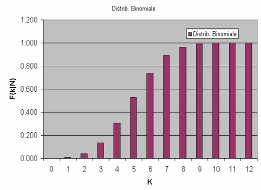
Principali distribuzioni di probabilità discrete

Binomiale

$$P(k) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$$

$$\mu_{bin} = Np$$
$$\sigma^2_{bin} = Np(1-p)$$
$$\sigma_{bin} = \sqrt{Np(1-p)}$$



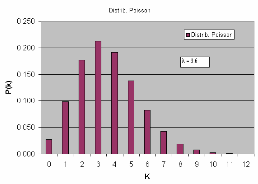


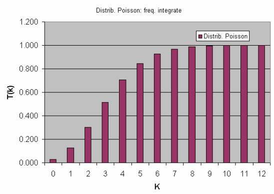
Principali distribuzioni di probabilità discrete

Poisson

$$P(x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

$$\mu_{pois} = \lambda$$
$$\sigma^2_{pois} = \lambda$$
$$\sigma_{pois} = \sqrt{\lambda}$$





Distribuzioni di probabilità continue

F(x) = funzione di distribuzione di probabilità

$$F(x) = \int_{-\infty}^{x} p(x)dx$$

Probabilità di osservare un valore della variabile X tra x_a e x_b

$$P(x_a \le X \le x_b) = F(x_b) - F(x_a)$$

p(x) = funzione di densità di probabilità

$$p(x) = \frac{dF(x)}{dx} = F'(x)$$

p(x)dx = probabilità di osservare un valore della variabile X compreso tra x e x+dx

$$dx\, p(x) = P(x \le X \le x + dx)$$

Principali distribuzioni di probabilità continue

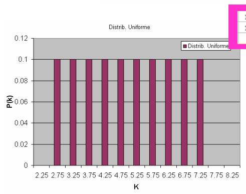
Uniforme

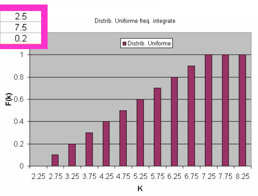
$$p_{uni}(x) = \begin{cases} \frac{1}{x_b - x_a} & x_a \le x \le x_b \\ 0 & (x < x_a), (x_b < x) \end{cases}$$

$$\mu_{uni} = \frac{x_a + x_b}{2}$$
$$\sigma^2_{uni} = \frac{(x_a - x_b)^2}{12}$$
$$\sigma_{uni} = \frac{x_a - x_b}{\sqrt{12}}$$

$$P_{uni}(x_1 \le X \le x_2) = \int_{x_1}^{x_2} p(x)dx = \frac{x_2 - x_1}{x_b - x_a}$$

x_a=	2.5
x_b=	7.5
p=	0.2





Principali distribuzioni di probabilità continue

Gauss (Normale)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

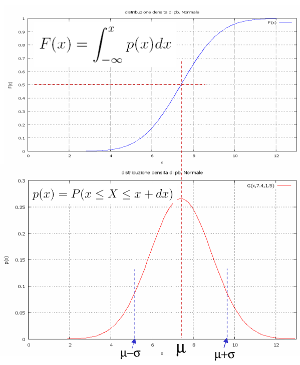
μ	7.4
σ	2.3

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0.682$$

$$P(\mu - 1.96\sigma \le X \le \mu + 1.96\sigma) = 0.95$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.954$$

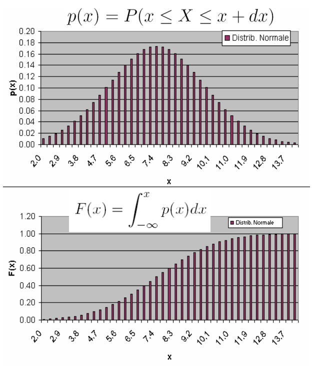
$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997$$



Principali distribuzioni di probabilità continue

Gauss (Normale)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

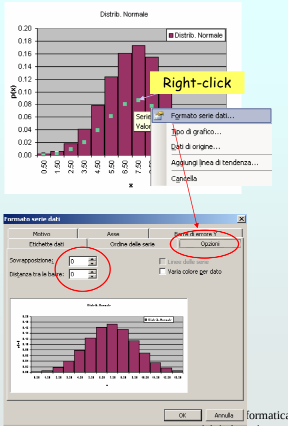


$$P(\mu - \sigma \le X \le \mu + \sigma) = 0.682$$

$$P(\mu - 1.96\sigma \le X \le \mu + 1.96\sigma) = 0.95$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997$$



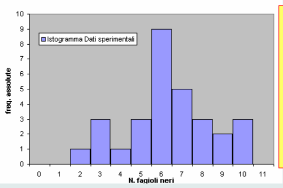
Istogramma:

L'area del rettangolo rappresenta la frequenza di una classe, l'altezza la densità di frequenza.

Il centro della base rappresenta il valore di riferimento.

incertezza di misura

La funzione di distribuzione binomiale è la funzione di distribuzione da cui provengono le frequenze di un istogramma costruito a partire dalle misure sperimentali.



In questo caso

Il numero totale di prove: $N_T = 30$

Consideriamo *successo* il fatto di osservare un valore nella classe 5 e *insuccesso* qualunque altro valore

Il numero di volte che abbiamo osservato il valore 5 è 3

quindi, in base ai dati sperimentali la probabilità di successo è:

$p_5 = 3/N_T = 0.1$

incertezza di misura

La funzione di distribuzione binomiale è la funzione di distribuzione da cui provengono le frequenze di un istogramma costruito a partire dalle misure sperimentali.

In base ai dati sperimentali il numero di volte che è lecito aspettarsi di trovare 5 fagioli neri pescando N_T fagioli è (valore atteso)

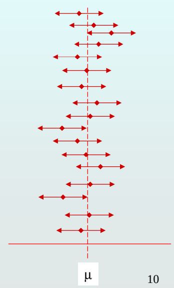
$\mu = N\, p_5$

L'incertezza è:

$\sigma^2 = N\, p_5(1-p_5)$

E' lecito aspettarsi che, ripetendo l'esperimento, le osservazioni sperimentali siano distribuite intorno al valore vero e che queste siano più frequenti nell'intervallo $\mu-\sigma$, $\mu+\sigma$

La deviazione standard σ è l'incertezza standard sul valore osservato.



L'incertezza standard sul numero di sul valore osservato in una data classe è σ_i

Dati sperimentali

Dati sperimentali

Pescata	n. fagioli neri
1	9
2	3
3	5
4	2
5	6
6	10
7	10
8	3
9	8
10	8
11	7
12	5
13	6
14	6
15	6
16	6
17	8
18	7
19	7
20	6
21	7
22	4
23	5
24	6
25	6
26	9
27	3
28	6
29	10
30	7

Riepilogo Statistiche

media:	6.4
dev.st	2.1
max	10
min	2
N. oss=	30

Istogramma di frequenze

Istogramma Dati sperimentali

classi	Freq. ass.	err. F	freq. rel	err. Freq. rel
0	0	0.0	0.000	0.000
1	0	0.0	0.000	0.000
2	1	1.0	0.033	0.033
3	3	1.7	0.100	0.058
4	1	1.0	0.033	0.033
5	3	1.7	0.100	0.058
6	9	3.0	0.300	0.100
7	5	2.2	0.167	0.075
8	3	1.7	0.100	0.058
9	2	1.4	0.067	0.047
10	3	1.7	0.100	0.058
11	0	0.0	0.000	0.000

N_i

$$f_i = \frac{N_i}{N_T} \sim p_i$$

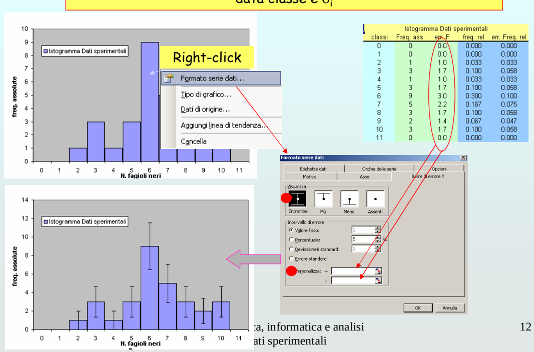
$$\sigma_i = \sqrt{f_i N_T (1 - f_i)}$$

$$\sigma_i \sim \sqrt{f_i N_T} \sim \sqrt{N_i}$$

$$\sigma_{f_i} = \frac{\sigma_i}{N_i} \sim \frac{1}{\sqrt{N_i}}$$

incertezza di misura

L'incertezza standard sul numero di sul valore osservato in una data classe è σ_i



Confronto dati sperimentali e distribuzioni teoriche

Distribuzioni discrete

Problema:

Confrontare la distribuzione sperimentale (N dati con distribuzione di frequenze assolute N_i) con la distribuzione teorica di probabilità P(x). Una volta costruito l'istogramma di frequenze sperimentali:

A

Calcolare la distribuzione delle frequenze relative f_{exp}

Calcolare l'incertezza sulle frequenze assolute: $\sigma_i = \sqrt{f_i N_T (1 - f_i)}$

$$\sigma_i \sim \sqrt{f_i N_T} \sim \sqrt{N_i}$$

Calcolare l'incertezza sulle frequenze relative $\sigma_{f_i} = \frac{\sigma_i}{N_i} \sim \frac{1}{\sqrt{N_i}}$

Riportare sul grafico P(x) e f_{exp} con le incertezze di misura

a, informati
ti sperimen

B

Calcolare l'incertezza sulle frequenze assolute: $\sigma_i = \sqrt{f_i N_T (1 - f_i)}$

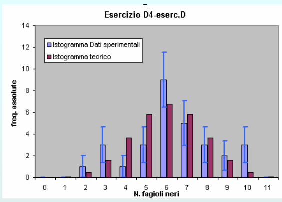
$$\sigma_i \sim \sqrt{f_i N_T} \sim \sqrt{N_i}$$

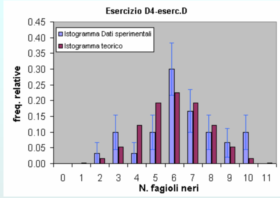
Calcolare le frequenze assolute teoriche:

$$N_{Th} = p_i N_T$$

Riportare sul grafico N_{Th} e N_i con le incertezze di misura

Confronto dati sperimentali e distribuzioni teoriche





Istogramma Dati sperimentali					Istogramma teorico		parametri Binomiale	
classi	Freq. ass.	err. F	freq. rel	err. Freq. rel	freq. ass.	freq. rel.	N	12
							p	0.5
0	0	0.0	0.000	0.000	0.007	0.000		
1	0	0.0	0.000	0.000	0.088	0.003		
2	1	1.0	0.033	0.033	0.483	0.016		
3	3	1.6	0.100	0.055	1.611	0.054		
4	1	1.0	0.033	0.033	3.625	0.121		
5	3	1.6	0.100	0.055	5.801	0.193		
6	9	2.5	0.300	0.084	6.768	0.226		
7	5	2.0	0.167	0.068	5.801	0.193		
8	3	1.6	0.100	0.055	3.625	0.121		
9	2	1.4	0.067	0.046	1.611	0.054		
10	3	1.6	0.100	0.055	0.483	0.016		
11	0	0.0	0.000	0.000	0.088	0.003		

14

Confronto dati sperimentali e distribuzioni teoriche: classi con più di un carattere (valore)

Dati sperimentali

Pescata	n. fagioli neri
1	9
2	3
3	5
4	2
5	6
6	10
7	10
8	3
9	8
10	8
11	7
12	5
13	6
14	6
15	6
16	6
17	8
18	7
19	7
20	6
21	7
22	4
23	5
24	6
25	6
26	9
27	3
28	6
29	10
30	7

Riepilogo Statistiche

media:	6.4
dev.st	2.1
max	10
min	2
N. oss=	30

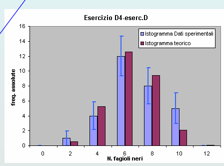
	G	H	I	J	K	L	M	N	O	P
1									parametri Binomiale	
2	Istogramma Dati sperimentali						Istogramma teorico		N	12
3	classi	Freq. ass.	err. F	freq. rel	err. Freq. rel		freq. ass.	freq. rel.	p	0.5
4	0	0	0.0	0.000	0.000		0.007	0.000		
5	2	1	1.0	0.033	0.033		0.571	0.019		
6	4	4	1.9	0.133	0.062		5.237	0.175		
7	6	12	2.7	0.400	0.089		12.568	0.419		
8	8	8	2.4	0.267	0.081		9.426	0.314		
9	10	5	2.0	0.167	0.068		2.095	0.070		
10	12	0	0.0	0.000	0.000		0.095	0.003		

=DISTRIB.BINOM(G5,P2,P3,0)+DISTRIB.BINOM(G5-1,P2,P3,0)

Per una la classe i-esima la funzione FREQUENZA di Excel calcola

$$F(i) - F(i-1) = N_i - N_{i-1}$$

Questo valore deve essere confrontato con la Pb di osservare il valore i-esimo oppure i-1 esimo

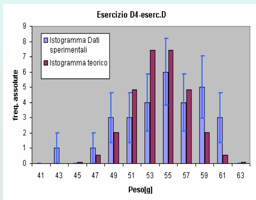


Vedere l'esempio nel file: confronto2.xls

15

Confronto dati sperimentali e distribuzioni teoriche: distribuzioni continue

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1								Istogramma Dati sperimentali									Parametri Gauss	
2	Dati sperimentali			media:	53.5		Classi								Istogramma teorico		μ	54
3	Campione	Peso[g]		dev.st	4.5		limiti		centro	Freq. ass.	err. F	freq. rel	err. Freq. rel		freq. ass.	freq. rel.	σ	3
4	1	41.11		max	60.27		40	42	41	0	0.0	0.000	0.000		0.001	0.000		
5	2	51.02		min	41.11		42	44	43	1	1.0	0.033	0.033		0.012	0.000		
6	3	51.93		N. oss=	30		44	46	45	0	0.0	0.000	0.000		0.102	0.003		
7	4	51.84					46	48	47	1	1.0	0.033	0.033		0.568	0.019		
8	5	50.94					48	50	49	3	1.6	0.100	0.055		2.054	0.068		
9	6	57.12					50	52	51	3	1.6	0.100	0.055		4.838	0.161		
10	7	54.95					52	54	53	4	1.9	0.133	0.062		7.425	0.248		
11	8	57.09					54	56	55	6	2.2	0.200	0.073		7.425	0.248		
12	9	58.80					56	58	57	4	1.9	0.133	0.062		4.838	0.161		
13	10	53.02					58	60	59	5	2.0	0.167	0.068		2.054	0.068		
14	11	50.12					60	62	61	3	1.6	0.100	0.055		0.568	0.019		
15	12	59.39					62	64	63	0	0.0	0.000	0.000		0.102	0.003		



=DISTRIB.NORM(H5,R2,R3,1)-DISTRIB.NORM(G5,R2,R3,1)

Per una distribuzione continua la probabilità associata ad un intervallo di valori (classe) si calcola a partire dalla distribuzione integrata:

$$P(x_a < X < x_b) = F(x_b) - F(x_a)$$

Vedere l'esempio nel file: confronto3.xls

16