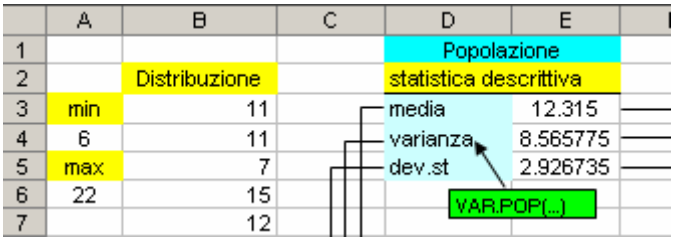
Ulteriori conoscenze di informatica – Elementi di statistica
Esercitazione3

Sui PC a disposizione sono istallati diversi sistemi operativi. All'accensione scegliere Windows.

Immettere	Nome utente	**b##**	(## : numero del pc)
	Pass.:	**biologia##**	(## : numero del pc)
		Stud_fisica	

Esercizio 1) Verifica del teoreme del limite centrale
Il file di dati: **popolazioni.dat** è un file ASCII a più colonne dove sono riportati dati distribuiti secondo diverse distribuzioni teoriche.

- Importare i dati in Excel utilizzando uno dei metodi visti nella esercitazioni precedenti
- Utilizzare un foglio per le popolazioni ed un foglio per i calcoli.
- Importare i dati di una delle popolazioni nel foglio del calcolo e calcolare i parametri descrittivi: media, varianza, dev. standard estremi (min e max). Utilizzare le funzioni VAR.POP() e DEV.ST.POP() per il calcolo della varianza e dev. standard della popolazione.



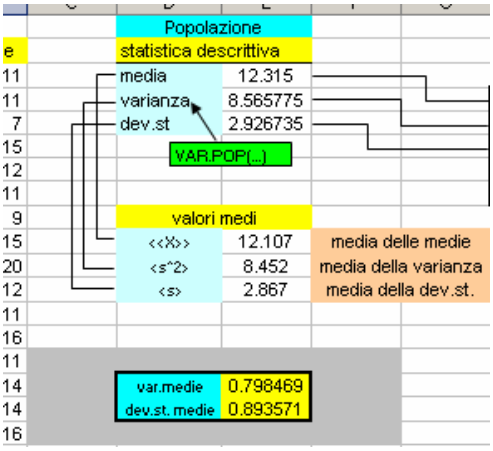
Effettuare un certo numero K di campionamenti di m elementi ciascuno:
per i campionamenti si possono utilizzare diversi metodi:
selezionare e copiare dati di regioni diverse,
utilizzare la funzione CAMPIONAMENTO delle funzioni avanzate di analisi dati, etc…

	Campioni									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
1	11.0000	12.0000	14.0000	11.0000	10.0000	9.0000	14.0000	11.0000	12.0000	8.0000
2	11.0000	8.0000	15.0000	12.0000	13.0000	12.0000	14.0000	17.0000	16.0000	12.0000
3	7.0000	14.0000	13.0000	8.0000	7.0000	22.0000	10.0000	17.0000	12.0000	7.0000
4	15.0000	12.0000	12.0000	9.0000	17.0000	14.0000	16.0000	12.0000	15.0000	13.0000
5	12.0000	11.0000	14.0000	15.0000	9.0000	11.0000	16.0000	10.0000	11.0000	16.0000
6	11.0000	12.0000	7.0000	16.0000	14.0000	11.0000	8.0000	10.0000	16.0000	12.0000
7	9.0000	10.0000	13.0000	11.0000	17.0000	14.0000	11.0000	9.0000	12.0000	7.0000
8	15.0000	14.0000	10.0000	12.0000	11.0000	11.0000	17.0000	11.0000	13.0000	11.0000
9	20.0000	14.0000	11.0000	8.0000	6.0000	13.0000	16.0000	8.0000	10.0000	13.0000
10	12.0000	11.0000	11.0000	14.0000	12.0000	14.0000	16.0000	17.0000	11.0000	10.0000
11	11.0000	9.0000	7.0000	11.0000	16.0000	12.0000	8.0000	9.0000	16.0000	11.0000
12	16.0000	10.0000	12.0000	13.0000	10.0000	17.0000	11.0000	8.0000	13.0000	9.0000
13	11.0000	14.0000	14.0000	9.0000	14.0000	16.0000	10.0000	10.0000	12.0000	15.0000
14	14.0000	9.0000	15.0000	7.0000	12.0000	14.0000	14.0000	8.0000	15.0000	12.0000

Per ognuno dei K campioni determinare i parametri della distribuzione (media, varianza e deviazione standard ed errore standard della media). Confrontare l'errore standard della media ottenuto dalle stime della varianza con quello calcolato dalla varianza della popolazione (Dev.St.Pop/ $m^{0.5}$) I valori risulteranno simili.

media	12.500	11.429	12.000	11.143	12.000	13.571	12.929	11.214	13.143	11.143
var	10.577	4.264	6.769	7.516	11.846	10.418	10.071	11.258	4.286	7.516
dev.st	3.252	2.065	2.602	2.742	3.442	3.228	3.174	3.355	2.070	2.742
err.st.med.	0.869	0.552	0.695	0.733	0.920	0.863	0.848	0.897	0.553	0.733
dev.st pop/m^0.5	0.782									

Confrontare i parametri statistici della popolazione con i valori medi ottenuti dai campionamenti: la media delle medie è prossima al valor medio della popolazione, la media delle varianze campionarie sono simili alla varianza della popolazione etc…



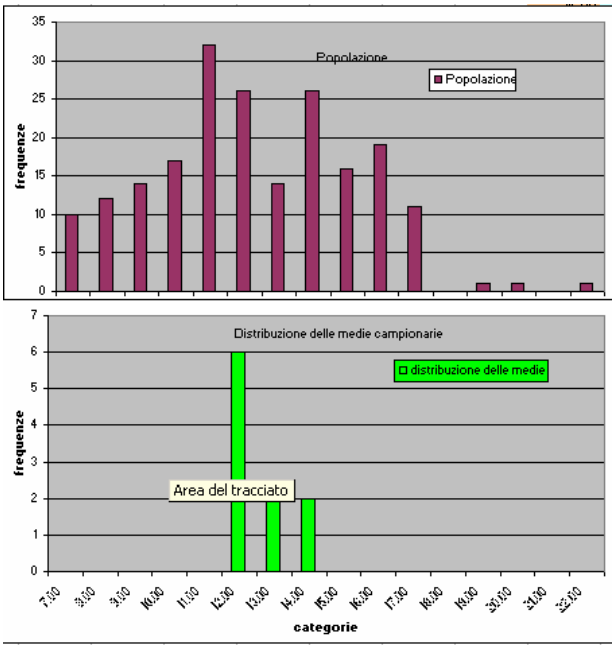
Calcolare varianza e deviazione standard della distribuzione delle medie campionarie e verificare che la deviazione standard della distribuzione delle medie campionarie e molto simile all'errore standard sulla media, cioè a Dev.St.Pop/ $m^{0.5}$

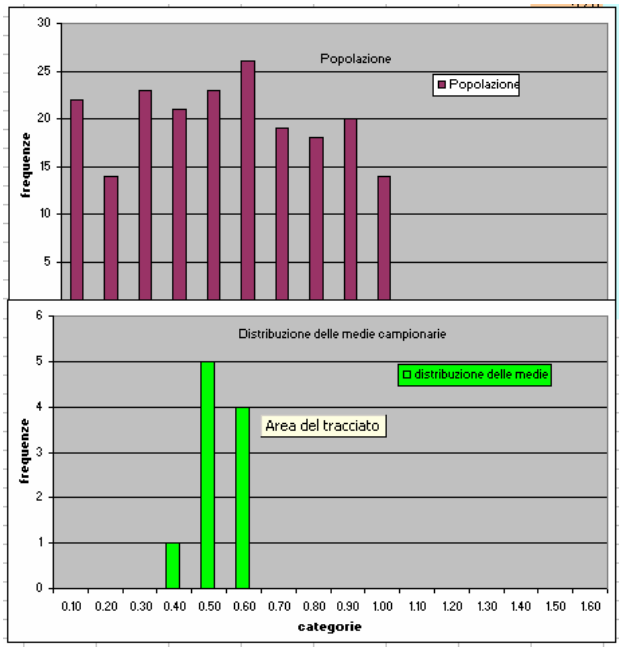
Con un po' di operazioni è possibile costruire grafici che mostrano in modo più intuitivo il comportamento dei campioni

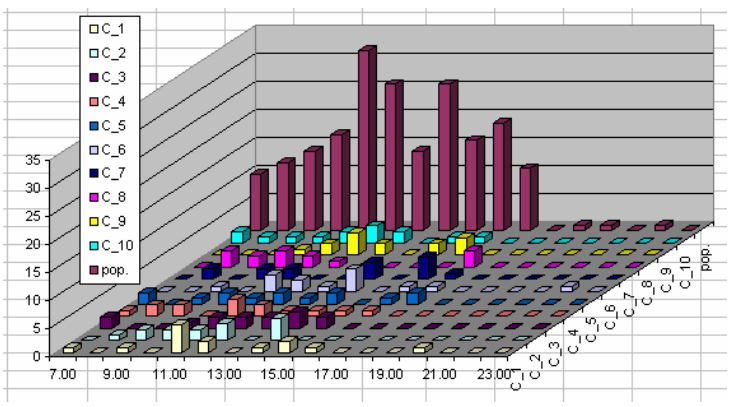
e della distribuzione delle media campionarie.
- definire le classi per cui calcolare le frequenze. Un modo per automatizzare il calcolo delle classi si può vedere nel file Esercizio7.xls: a partire da un minimo (da inserire) e da un intervallo (da inserire).
- Utilizzare la funzione FREQUENZA come già fatto per l'esercitazione 1 per il calcolo automatico delle frequenze.

Minimo	classi	pop.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	medie
6	7.00	10	1	0	2	1	2	0	0	0	0	2	0
delta	8.00	12	0	1	0	2	0	0	2	3	0	1	0
1	9.00	14	1	2	0	2	1	1	0	2	0	1	0
	10.00	17	0	2	1	0	2	0	2	3	1	1	0
	11.00	32	5	2	2	3	1	3	2	2	2	2	0
	12.00	26	2	3	2	2	2	2	0	1	4	3	6
	13.00	14	0	0	2	1	1	1	0	0	2	2	2
	14.00	26	1	4	3	1	2	4	3	0	0	0	2
	15.00	16	2	0	2	1	0	0	0	0	2	1	0
	16.00	19	1	0	0	1	1	1	4	0	3	1	0
	17.00	11	0	0	0	0	2	1	1	3	0	0	0
	18.00	0	0	0	0	0	0	0	0	0	0	0	0
	19.00	1	0	0	0	0	0	0	0	0	0	0	0
	20.00	1	1	0	0	0	0	0	0	0	0	0	0
	21.00	0	0	0	0	0	0	0	0	0	0	0	0

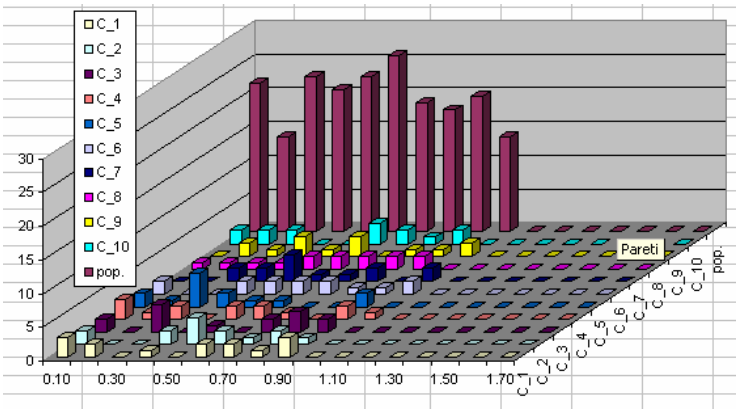
Preparare gli istogrammi con la distribuzione della popolazione originale, la distribuzione delle medie e le distribuzioni dei vari campionamenti.
Sostituendo i dati nella colonna della popolazione e aggiornano le opzioni (minimo e delta) per gli istogrammi si dovrebbeo ottenere i grafici relativi a popolazioni con diverse distribuzioni.





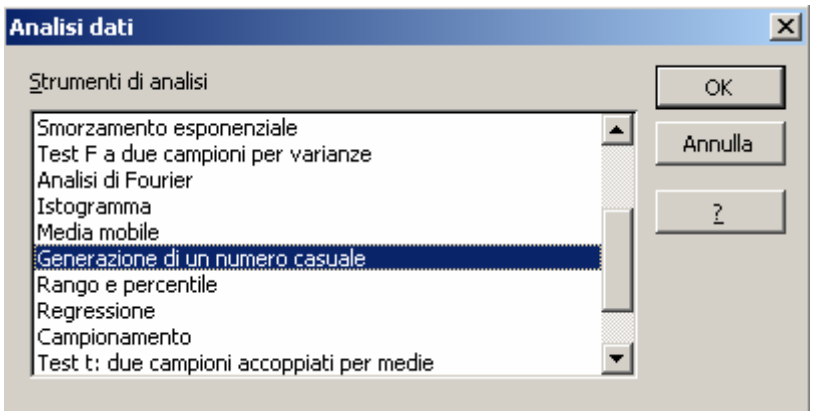


Distribuzione di Bernulli

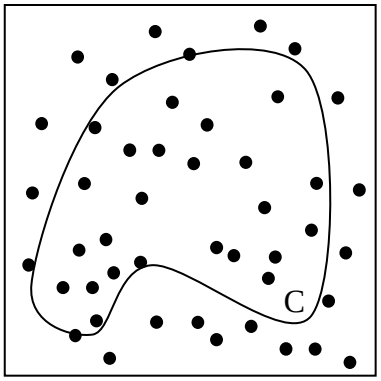


Distribuzione Uniforme

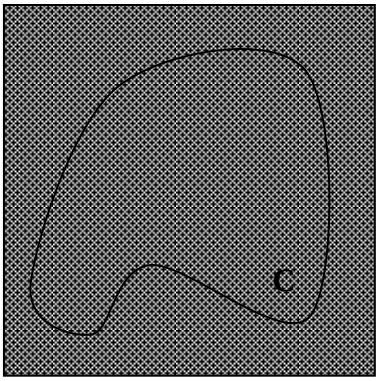
Provare a utilizzare popolazioni con diverse distribuzioni utilizzando la funzione “generazione di un numero casuale” delle funzioni avanzate di analisi dati di EXCEL.



Esercizio 2) Integrazione con il metodo MonteCarlo. Supponiamo di avere una figura complicata racchiusa da una curva C e di volerne calcolare l'integrale, cioè la superficie. Scegliamo una figura piana regolare, di cui conosciamo la superficie, che contenga la figura in questione, es. un quadrato di lato a. Se ora scegliamo a caso un certo numero di punti **a caso** all'interno del quadrato una parte di essi cadrà all'interno della curva C e una parte no.



a

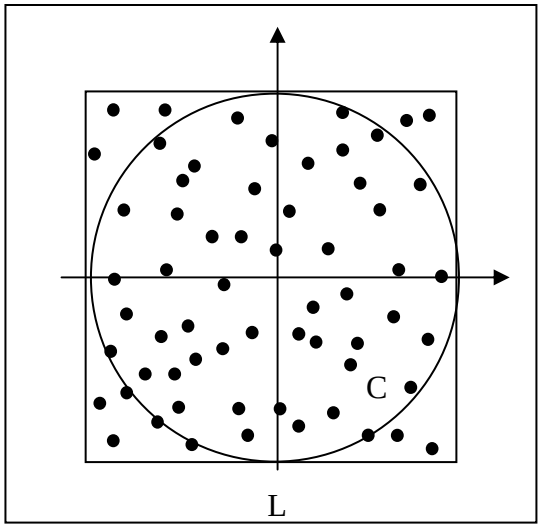


a

E' abbastanza intuitivo pensare che se la distribuzione dei punti è uniforme, il rapporto tra il numero di punti all'interno della curva C N_c e il numero di punti totali all'interno del quadrato N_t approssima il rapporto tra la superficie racchiusa dalla S_c curva C e la superficie del quadrato S_t.

Quindi possiamo stimare la superficie della curva come: $S_c \sim S_t\ N_c / N_t$ e tale approssimazione sarà tanto più precisa quanto più aumento la densità dei punti (N_t). Questo è un metodo statistico per la soluzione numerica di problemi irresolubili o molto complicati da risolvere in modo analitico noto come Metodo Monte Carlo.

Questo metodo è particolarmente potente per il calcolo di integrali a molte dimensioni. Proviamo ad utilizzarlo per il calcolo del valore di π. Se considero un cerchio di raggio R iscritto in un quadrato di lato L=2R il rapporto tra la superficie del cerchio ($S_c = \pi R^2$) e la superficie del quadrato ($S_q = L^2 = 4R^2$) è:



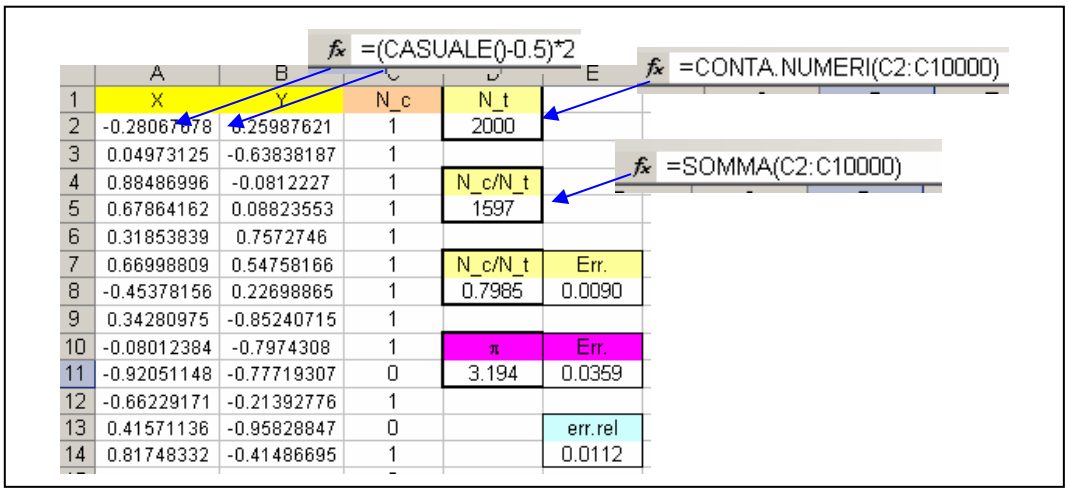
$$\frac{S_c}{S_q} = \frac{\pi}{4}$$

Utilizziamo un foglio elettronico er il calcolo: assumiamo R = 1
- generiamo due colonne di numeri casuali (X,Y) uniformemente distribuiti nell'intervallo -1,1. Per questo usiamo la funzione CASUALE() che genera un numero casuale nell'intervallo [0:1]:

(CASUALE()-0.5)*2

- In una terza colonna scriviamo 1 se x^2+y^2 è minore o uguale a 1
- calcoliamo $\pi = 4\ N_c/N_t$

Scegliamo prima 100, poi 1000 poi 2000 coppie di coordinate e vediamo che la precisione migliora.



	A	B	C	D	E
1	X	Y	N_c	N_t	
2	-0.28067678	-0.25987621	1	2000	
3	0.04973125	-0.63838187	1		
4	0.88486996	-0.0812227	1	N_c/N_t	
5	0.67864162	0.08823553	1	1597	
6	0.31853839	0.7572746	1		
7	0.66998809	0.54758166	1	N_c/N_t	Err.
8	-0.45378156	0.22698865	1	0.7985	0.0090
9	0.34280975	-0.85240715	1		
10	-0.08012384	-0.7974308	1	π	Err.
11	-0.92051148	-0.77719307	0	3.194	0.0359
12	-0.66229171	-0.21392776	1		
13	0.41571136	-0.95828847	0		err.rel
14	0.81748332	-0.41486695	1		0.0112

Quale è la precisione del metodo? Se denominiamo con "successo" il caso che un punto cada all'interno della superficie racchiusa dalla curva C, la probabilità di successo ad ogni estrazione (coppia x,y) è $p = S_c/S_t$. La variabile X può assumere solo valori 1 (successo: nel cerchi) 0 (insuccesso: fuori del cerchio) e segue la statistica di Bernulli con valore atteso

valore atteso $\mu = p$ e $\sigma^2 = \sqrt{p(1-p)}$ varianza:

La mia estrazione permette di ottenere una stima del valore atteso e della varianza

$$\bar{x} = \frac{N_c}{N_t} \qquad s^2 = \bar{x}(1-\bar{x})$$

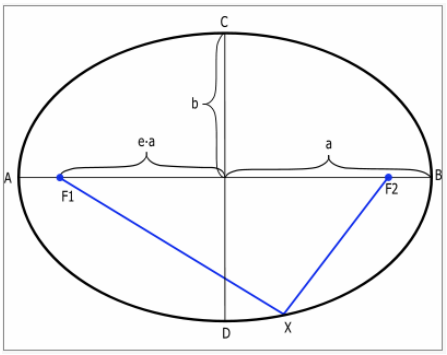
l'errore sulla media è l'errore standard della media:

$$s_{\bar{x}} = \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{N_t}}$$

Quindi l'errore standard sul valore di π stimato è 4 s_x. La precisione del metodo si può stimare utilizzando l'errore relativo:

$$\frac{s_{\bar{x}}}{\bar{x}} = \sqrt{\frac{1-\bar{x}}{\bar{x}}}\frac{1}{\sqrt{N_t}}$$

Vediamo quindi che la precisione del metodo migliora come l'inverso della radice del numero di prove. Attenzione: l'errore standard sulla media corrisponde ad un'intervallo di confidenza al 68%, significa che nel 32% circa delle prove otterrò risultati che si discostano dal valore vero più dell'errore stimato.



Ora proviamo ad applicare il metodo al calcolo del volume di un ellissoide in uno spazio N dimensioni (es. N=4). L'ellisse è una figura geometrica piana descritta dall'equazione:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Dove a e b sono i semiassi dell'ellisse.

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1$$

La sua superficie è il luogo dei punti per cui:

In uno spazio a n dimensioni un ellissoide è descritto dall'equazione:

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \ldots \frac{x_n^2}{a_n^2} = 1$$

Dove a_i sono i semiassi. Il volume è il luogo dei punti per cui:

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \ldots \frac{x_n^2}{a_n^2} \leq 1$$

Proviamo ad applicare il metodo MonteCarlo per calcolare il volume di un ellissi in uno spazio a n=4 dimensioni.
Suggerimento:
- L'ellisse è iscritta in un iper-parallelogramma di lati $2a_1$, $2a_2$, …
- per ogni dimensione generare numeri casuali nell'intervallo $-a_i$ - a_i
- attenzione al calcolo dell'errore
- provare prima a calcolare l'area del cerchio, poi il volume della sfera etc…

	A	B	C	D	E	F	G	H
1	semiassi:	1	1	1	1		N_t	V_c
2		X_1	X_2	X_3	X_4		2000	16
3		-0.57417037	-0.65803498	-0.89658912	0.30080695	0		
4		0.15746986	-0.61010017	0.13632381	0.44196492	1	N_c	
5		-0.52178165	0.62257833	0.28595947	0.63815568	0	627	
6		0.16791094	-0.04819943	0.18546853	0.76207686	1		
7		-0.36855577	0.985473	-0.67018164	0.95458443	0	N_c/N_t	Err.
8		0.07748379	0.222171	-0.65421697	0.52428191	1	0.3135	0.0104
9		-0.51155364	0.23682059	-0.50294731	-0.78082065	0		
10		0.01574244	-0.00924946	-0.16619762	-0.24938264	1	V_E	Err.
11		0.94860457	-0.28393836	0.66434359	0.2568892	0	5.016	0.1660
12		0.28557808	-0.31048349	-0.28884308	-0.40992885	1		
13		0.61242691	-0.80016047	0.20088888	-0.6704982	0		err.rel
14		-0.27258154	0.16353727	0.2770445	0.00720501	1		0.0331

Volume di una ipersfera in uno spazio n=4 dimensioni