Elementi di Inferenza Statistica

Variabili casuali

V Scuola Estiva AISV

La statistica come strumento di analisi nelle scienze umanistiche e comportamentali

Soriano nel Cimino (VT), 6 Ottobre 2009

Pier Francesco Perri
Dipartimento di Economia e Statistica - UNICAL
pierfrancesco.perri@unical.it

Il concetto di variabile casuale

- La definizione classica di probabilità fa riferimento a fenomeni aleatori i cui possibili risultati formano lo spazio campionario
- La descrizione degli eventi passa attraverso la formulazione di espressioni verbali che, in quanto tali, risultano poco idonee per essere trattate matematicamente
- Può essere utile, in tal senso, ``quantificare gli eventi'', ovvero considerare come possibili risultati di un esperimento casuale non più gli eventi - descritti tramite espressioni del linguaggio comune - bensì un insieme di numeri reali, in corrispondenza con gli eventi, mediante i quali la loro trattazione diventa più spedita.

Il concetto di variabile casuale

- Questa esigenza genera il concetto di variabile casuale (v.c.) con il quale si suole indicare una quantità variabile il cui valore è determinato dal risultato di un esperimento casuale.
- Questo legame di dipendenza del valore della variabile casuale dall'esito di un esperimento evoca l'idea di funzione:

Variabile Casuale

Una variabile casuale è la misura numerica del risultato di un esperimento (fenomeno) casuale.
Definire una variabile casuale significa associare, secondo una determinata regola, un numero ad ogni risultato dell'esperimento.

Il concetto di variabile casuale

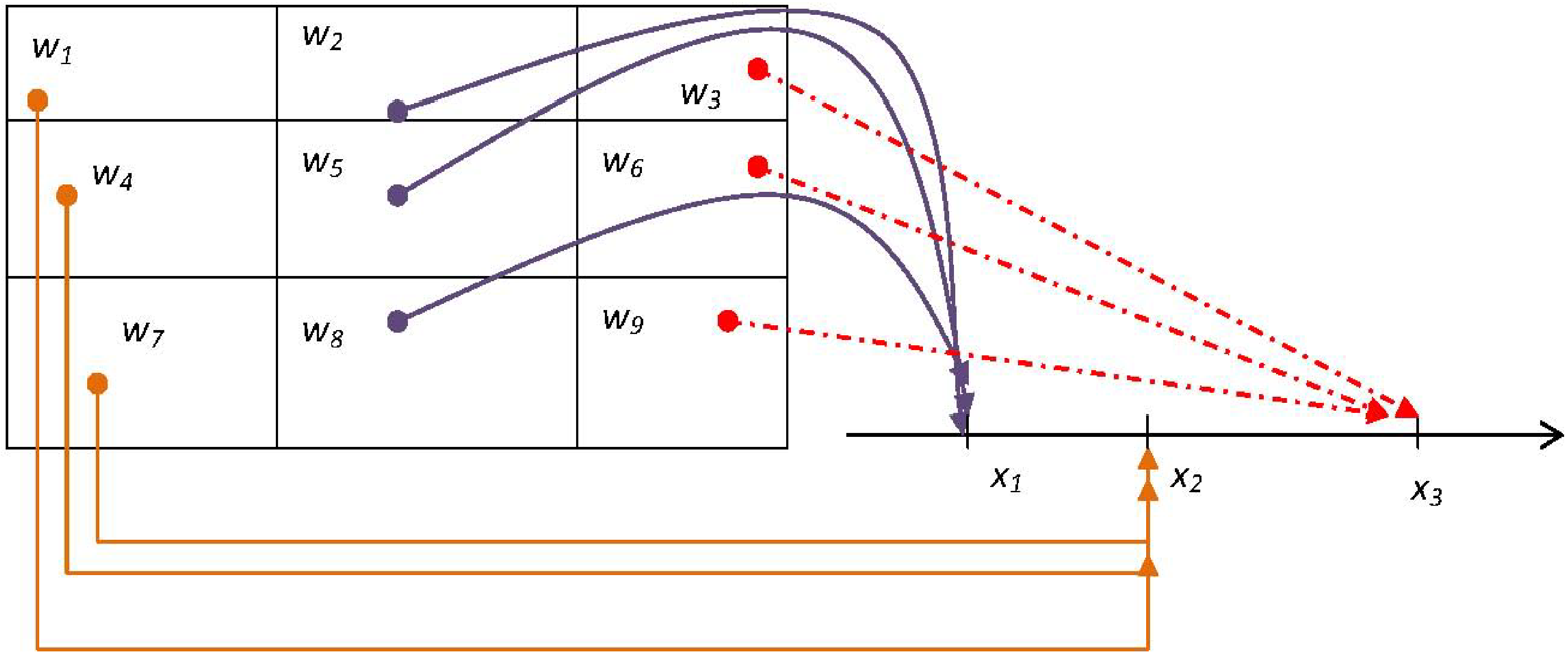
Variabile Casuale Discreta

Una variabile casuale è detta discreta se può assumere un numero finito o un'infinità numerabile di valori.

E' utilizzata per descrivere fenomeni che prevedono un'operazione di conteggio

Il concetto di variabile casuale

Spazio campionario



Variabile casuale: esempio

Consideriamo il lancio di tre monete regolari.
Se in corrispondenza di ogni possibile risultato conteggiamo il numero di volte che appare testa, avremmo definito la variabile casuale X: ``numero di teste uscite''.

La situazione può essere schematizzata nel seguente prospetto

Spazio campionario	Valori di X	Probabilità
ω_1={TTT}	3	1/8
ω_2={TTC}	2	1/8
ω_3={TCC}	1	1/8
ω_4={CCC}	0	1/8
ω_5={CTC}	1	1/8
ω_6={CTT}	2	1/8
ω_7={CCT}	1	1/8
ω_8={TCT}	2	1/8

Distribuzione di probabilità

La v.c. X: "numero di teste uscite" assume 4 diversi valori: 0,1,2,3.
Se tali valori si organizzano in un prospetto associando ad ognuno di essi la probabilità del loro verificarsi, si costruisce la **distribuzione di probabilità** della v.c. in esame

Valori di X	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Le probabilità sono tutte comprese tra 0 e 1 e la loro somma è pari 1

- ✓ *Qual è la probabilità di osservare almeno una testa?*
- ✓ *Qual è la probabilità di osservare meno di tre teste?*
- ✓ *Quante teste mediamente si osservano lanciando tante volte le tre monete?*

Distribuzione di probabilità

In generale, data una v.c. discreta X che può assumere i valori (distinti) $x_1, x_2, \ldots, x_k$ la sua distribuzione di probabilità è definita tramite una funzione P(.), detta funzione di probabilità che associa ad ogni singolo valore x_2 la sua probabilità

Valori di X	x_1	x_2	…	x_i	…	x_k
$P(X=x_i)$	p_1	p_2	…	p_i	…	p_k

La funzione P(.) viene espressa tramite un modello matematico

Media di X $\mu_x = \sum_{i=1}^{k} x_i p_i$

Varianza di X $\sigma_x^2 = \sum_{i=1}^{k} (x_i - \mu_x)^2 p_i$

Deviazione standard di X $\sigma_x = \sqrt{\sigma_x^2}$

La v.c. Binomiale

Il risultato di molti esperimenti casuali è spesso di tipo binario e i due possibili esiti (complementari) vengono generalmente detti "successo" e "insuccesso".
Un esperimento di questo tipo è detto **bernulliano**.

Esperimento	Esiti (successo, insuccesso)
Lancio di una moneta	testa, croce
Analisi clinica	positivo, negativo
Sostenimento di un esame	promosso, bocciato

L'evento "successo" viene definito dal ricercatore in base alle finalità dello studio. La sua probabilità è p.
La probabilità dell'evento "insuccesso" è 1-p.

Un esperimento bernulliano viene espresso tramite **la v.c. di Bernoulli** che presenta media p e varianza p(1-p).

La v.c. Binomiale

Supponiamo ora di ripetere n-volte un esperimento bernulliano e di volere determinare la probabilità di osservare un certo numero di successi.

La v.c. Binomiale consente di "modellare" tali probabilità, sotto determinate ipotesi di lavoro:

1. Le condizioni sperimentali rimangono invariate nel corso del ripetersi dell'esperimento, ovvero la probabilità di successo p non muta da esperimento ad esperimento
2. Gli esperimenti bernulliani sono indipendenti, ovvero l'esito di un esperimento non influenza in nessun modo quello che potrà essere il risultato di un altro esperimento

La v.c. Binomiale

Variabile Casuale Binomiale

Sia p la probabilità di successo in un esperimento benulliano e sia X la v.c. discreta "numero di successi in n esperimenti benulliani", X=0,1,…,n.

La funzione di probabilità della v.c. X è data da:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0,1,\dots,n$$

dove

$$\binom{n}{x} = \frac{n!}{(n-x)!\,x!}, \quad n = n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$$

La v.c. Binomiale: esempio

Un test di cultura generale è composto da 10 domande ciascuna a riposta multipla con 4 alternative di cui solo una corretta.

Il test si considera superato se si risponde correttamente ad almeno 6 domande.

Qual è la probabilità che un soggetto completamente "ignorante" degli argomenti oggetto d'esame superi il test con il minimo sforzo rispondendo casualmente alle domande?

- Esperimento bernulliano: rispondere ad una domanda multipla
- Evento successo: risposta corretta
- Probabilità di successo: $p=0.25$
- Numero di esperimenti bernulliani: $n=10$ risposte multiple
- $X \sim Bin(10, 0.25)$

La v.c. Binomiale: esempio

Siamo interessati a determinare la probabilità di rispondere correttamente a 6 domande, ovvero P(X=6):

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\Downarrow$$

$$P(X = 6) = \binom{10}{6} 0.25^6 (1 - 0.25)^{10-6} = 0.016$$

Di contro, la probabilità che il test non venga superato è pari a:

$$P(X<6)=P(X=0)+P(X=1)+....+P(X=5)=0.996$$

La v.c. Binomiale: esempio

Potremmo anche essere interessati a determinare il numero medio di risposte corrette che un soggetto darebbe se ripetesse il test più volte.

L'aspettativa della v.c. Binomiale è pari a np mentre la sua varianza la si ottiene come np(1-p)

Nel nostro esempio, il numero medio di risposte corrette è

$$\mu = 10 * 0.25 = 2.5$$

La deviazione standard è invece

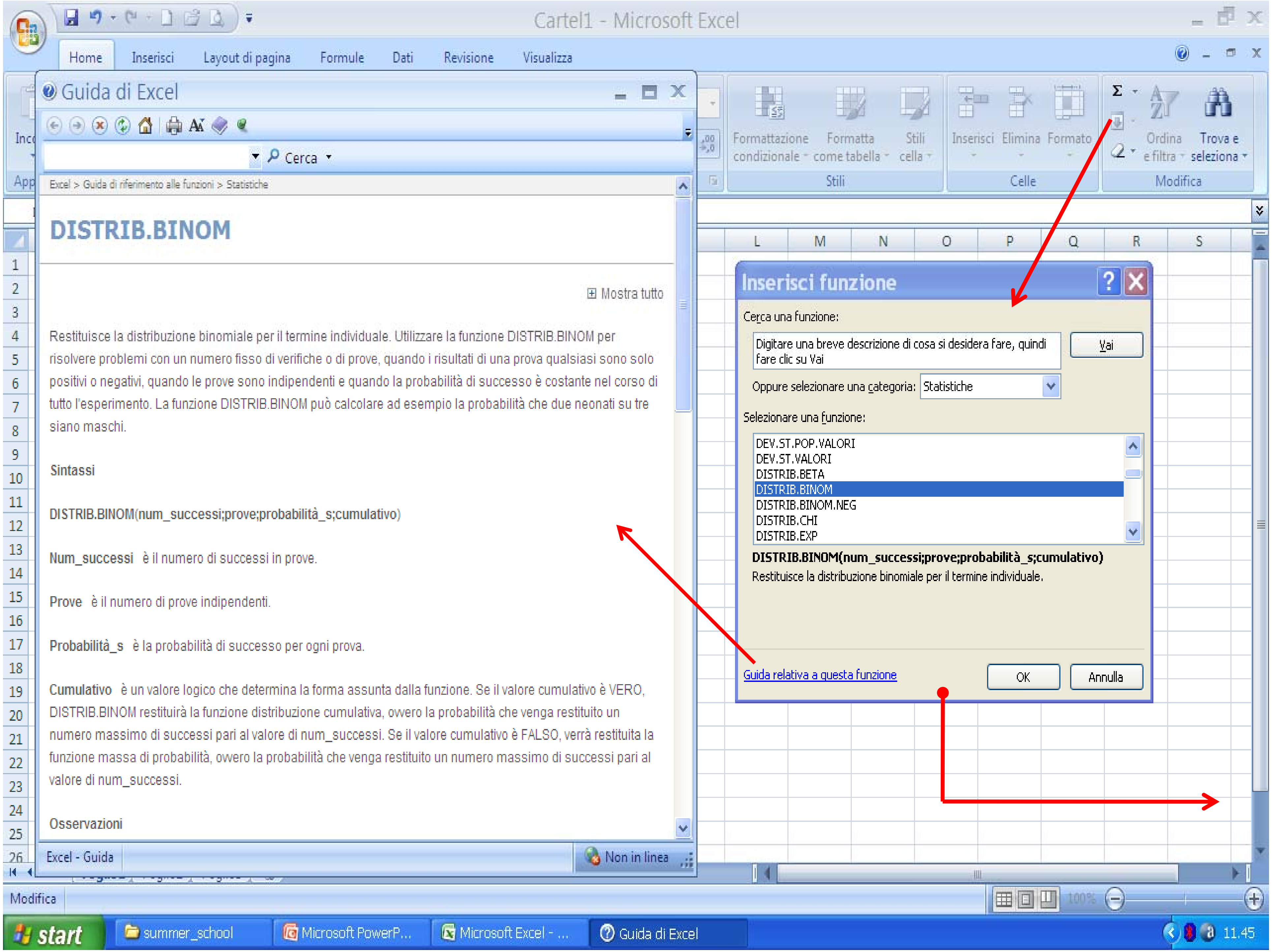
$$\sigma = \sqrt{10 * 0.25 * 0.75} = 1.37$$

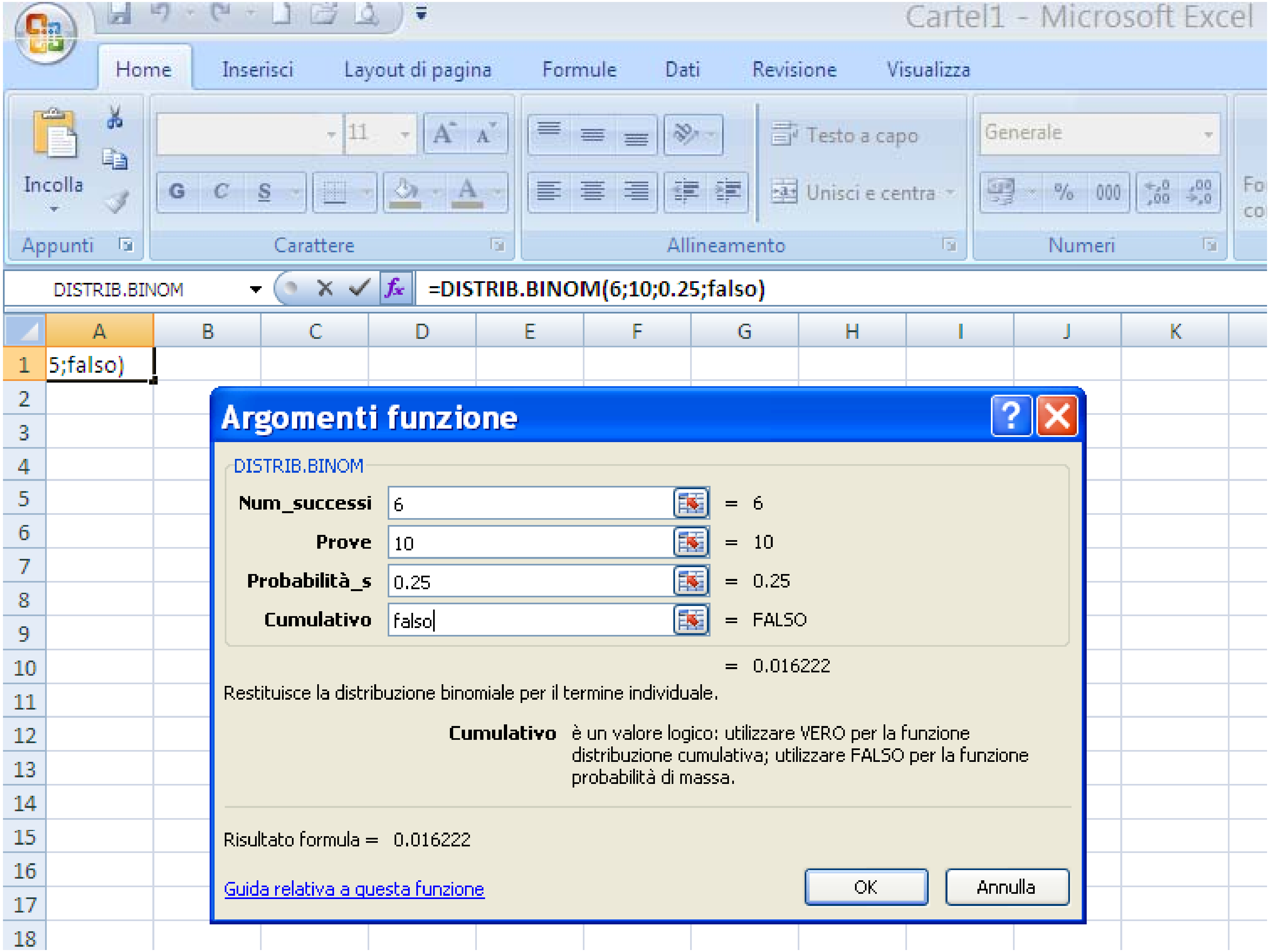
✓ Qual è il significato dei valori ottenuti?

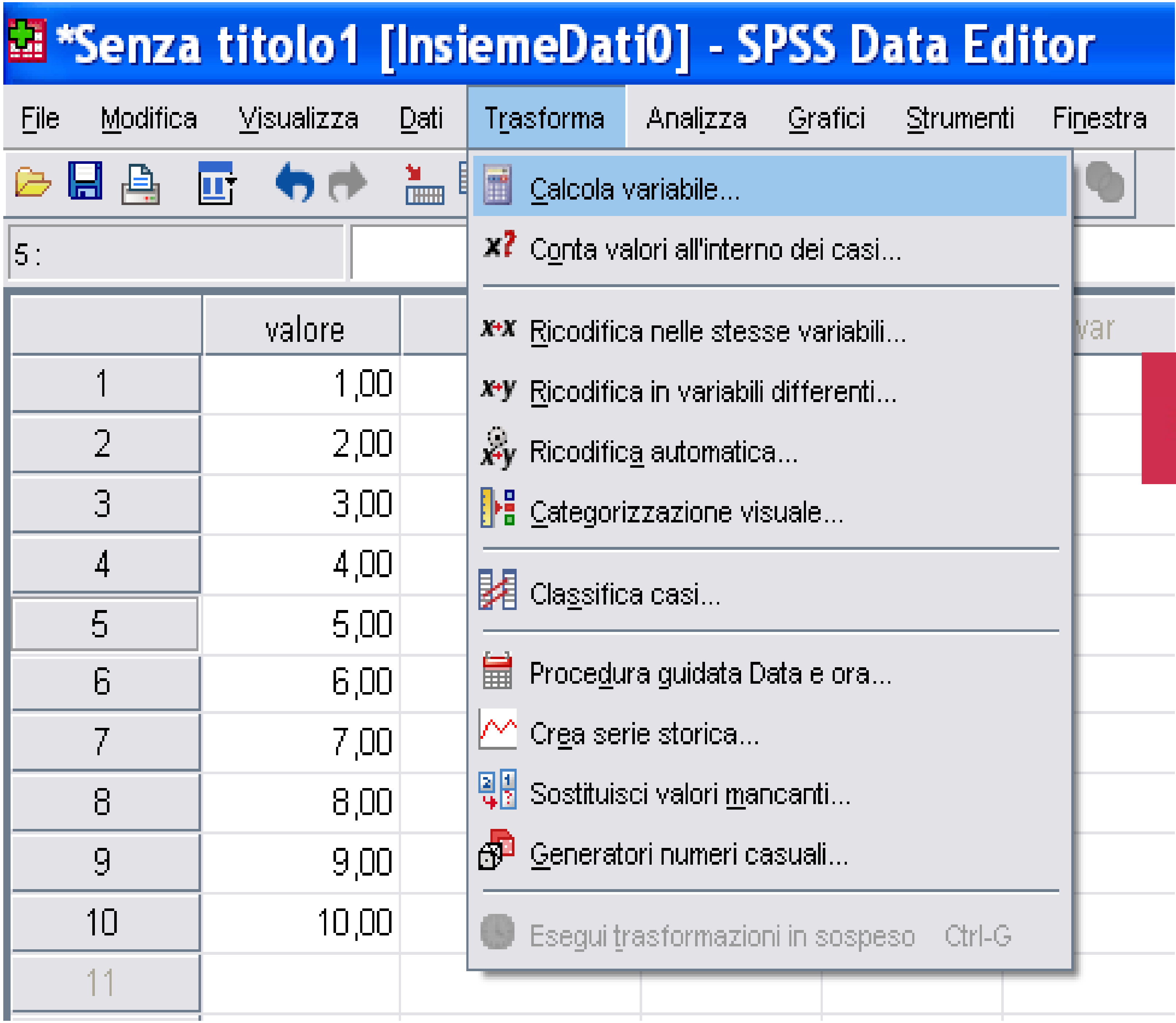
La v.c. Binomiale con MS Excel

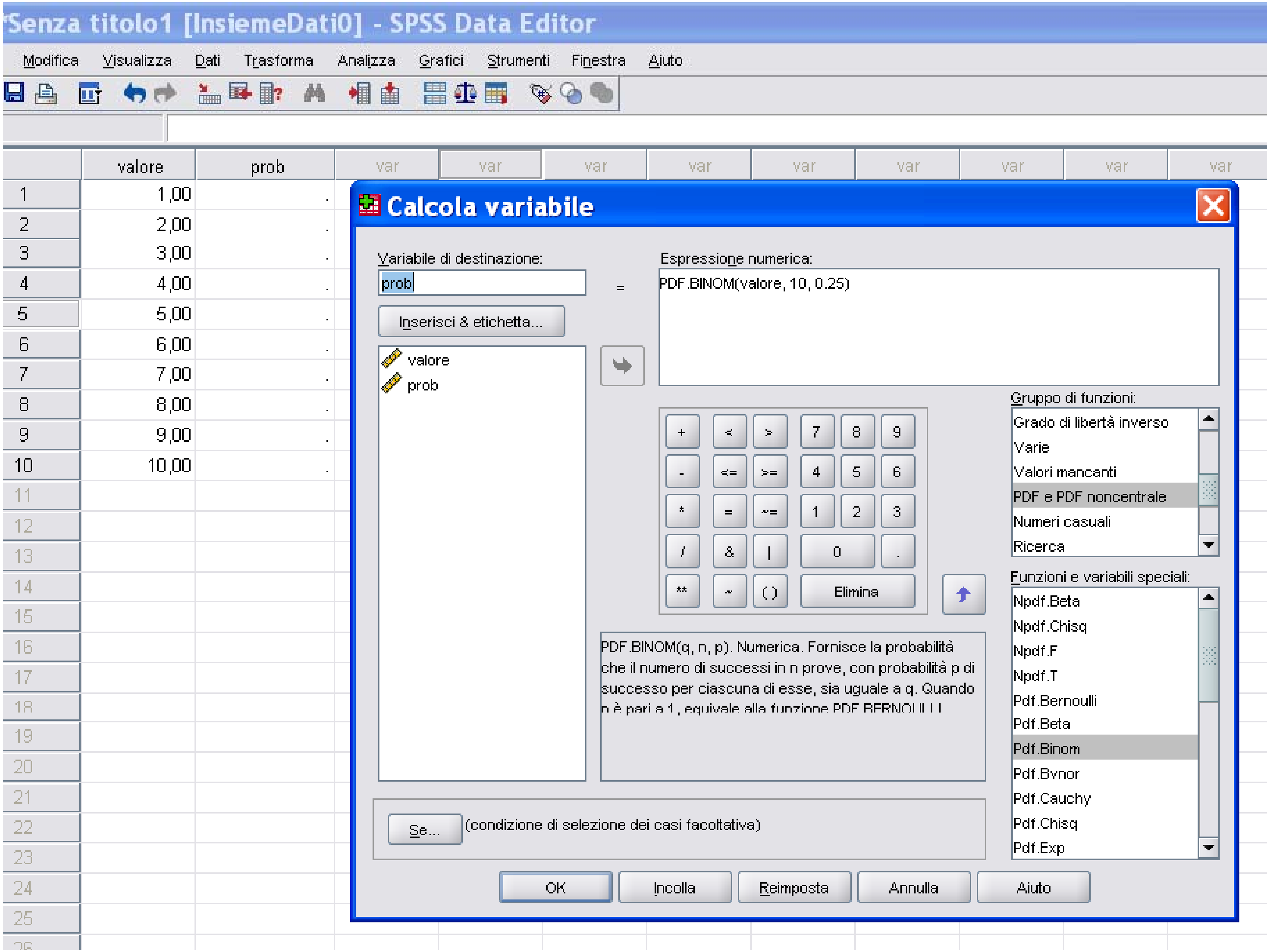
I calcoli necessari per determinare le probabilità di una v.c. Binomiale possono essere facilmente realizzati tramite MS Excel eseguendo alcune semplici operazioni:

1. Si apre un foglio di lavoro
2. Si posiziona il cursore su una cella
3. Dalla barra dei comandi si inserisce una funzione
   Inserisci funzione (∑) → Statistiche→ DISTRIB.BINOM
4. Si specificano gli argomenti della funzione


Cartel1 - Microsoft Excel
Home
Inserisci
Layout di pagina
Formule
Dati
Revisione
Visualizza
Guida di Excel
Cerca
Excel > Guida di riferimento alle funzioni > Statistiche
DISTRIB.BINOM
Mostra tutto
Restituisce la distribuzione binomiale per il termine individuale. Utilizzare la funzione DISTRIB.BINOM per risolvere problemi con un numero fisso di verifiche o di prove, quando i risultati di una prova qualsiasi sono solo positivi o negativi, quando le prove sono indipendenti e quando la probabilità di successo è costante nel corso di tutto l'esperimento. La funzione DISTRIB.BINOM può calcolare ad esempio la probabilità che due neonati su tre siano maschi.
Sintassi
DISTRIB.BINOM(num_successi;prove;probabilità_s;cumulativo)
Num_successi è il numero di successi in prove.
Prove è il numero di prove indipendenti.
Probabilità_s è la probabilità di successo per ogni prova.
Cumulativo è un valore logico che determina la forma assunta dalla funzione. Se il valore cumulativo è VERO, DISTRIB.BINOM restituirà la funzione distribuzione cumulativa, ovvero la probabilità che venga restituito un numero massimo di successi pari al valore di num_successi. Se il valore cumulativo è FALSO, verrà restituita la funzione massa di probabilità, ovvero la probabilità che venga restituito un numero massimo di successi pari al valore di num_successi.
Osservazioni
Excel - Guida
Non in linea
Formattazione condizionale
Formatta come tabella
Stili cella
Stili
Inserisci
Elimina
Formato
Celle
Ordina e filtra
Trova e seleziona
Modifica
Inserisci funzione
Cerca una funzione:
Digitare una breve descrizione di cosa si desidera fare, quindi fare clic su Vai
Vai
Oppure selezionare una categoria: Statistiche
Selezionare una funzione:
DEV.ST.POP.VALORI
DEV.ST.VALORI
DISTRIB.BETA
DISTRIB.BINOM
DISTRIB.BINOM.NEG
DISTRIB.CHI
DISTRIB.EXP
DISTRIB.BINOM(num_successi;prove;probabilità_s;cumulativo)
Restituisce la distribuzione binomiale per il termine individuale.
Guida relativa a questa funzione
OK
Annulla
Modifica
start
summer_school
Microsoft PowerP...
Microsoft Excel - ...
Guida di Excel
11.45


Cartel1 - Microsoft Excel
Home
Inserisci
Layout di pagina
Formule
Dati
Revisione
Visualizza
Incolla
Appunti
Carattere
Testo a capo
Unisci e centra
Allineamento
Generale
Numeri
DISTRIB.BINOM
=DISTRIB.BINOM(6;10;0.25;falso)
5;falso)
Argomenti funzione
DISTRIB.BINOM
Num_successi 6 = 6
Prove 10 = 10
Probabilità_s 0.25 = 0.25
Cumulativo falso = FALSO
= 0.016222
Restituisce la distribuzione binomiale per il termine individuale.
Cumulativo è un valore logico: utilizzare VERO per la funzione distribuzione cumulativa; utilizzare FALSO per la funzione probabilità di massa.
Risultato formula = 0.016222
Guida relativa a questa funzione
OK
Annulla


*Senza titolo1 [InsiemeDati0] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
5 :
valore
var
1 1,00
2 2,00
3 3,00
4 4,00
5 5,00
6 6,00
7 7,00
8 8,00
9 9,00
10 10,00
11
Calcola variabile...
Conta valori all'interno dei casi...
Ricodifica nelle stesse variabili...
Ricodifica in variabili differenti...
Ricodifica automatica...
Categorizzazione visuale...
Classifica casi...
Procedura guidata Data e ora...
Crea serie storica...
Sostituisci valori mancanti...
Generatori numeri casuali...
Esegui trasformazioni in sospeso Ctrl-G


*Senza titolo1 [InsiemeDati0] - SPSS Data Editor
Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
valore prob var var var var var var var var var
1,00 2,00 3,00 4,00 5,00 6,00 7,00 8,00 9,00 10,00
Calcola variabile
Variabile di destinazione:
prob
Inserisci & etichetta...
valore
prob
Espressione numerica:
PDF.BINOM(valore, 10, 0.25)
Elimina
Gruppo di funzioni:
Grado di libertà inverso
Varie
Valori mancanti
PDF e PDF noncentrale
Numeri casuali
Ricerca
Funzioni e variabili speciali:
Npdf.Beta
Npdf.Chisq
Npdf.F
Npdf.T
Pdf.Bernoulli
Pdf.Beta
Pdf.Binom
Pdf.Bvnor
Pdf.Cauchy
Pdf.Chisq
Pdf.Exp
PDF.BINOM(q, n, p). Numerica. Fornisce la probabilità che il numero di successi in n prove, con probabilità p di successo per ciascuna di esse, sia uguale a q. Quando n è pari a 1, equivale alla funzione PDF.BERNOULLI.
Se...
(condizione di selezione dei casi facoltativa)
OK Incolla Reimposta Annulla Aiuto

Variabili casuali continue

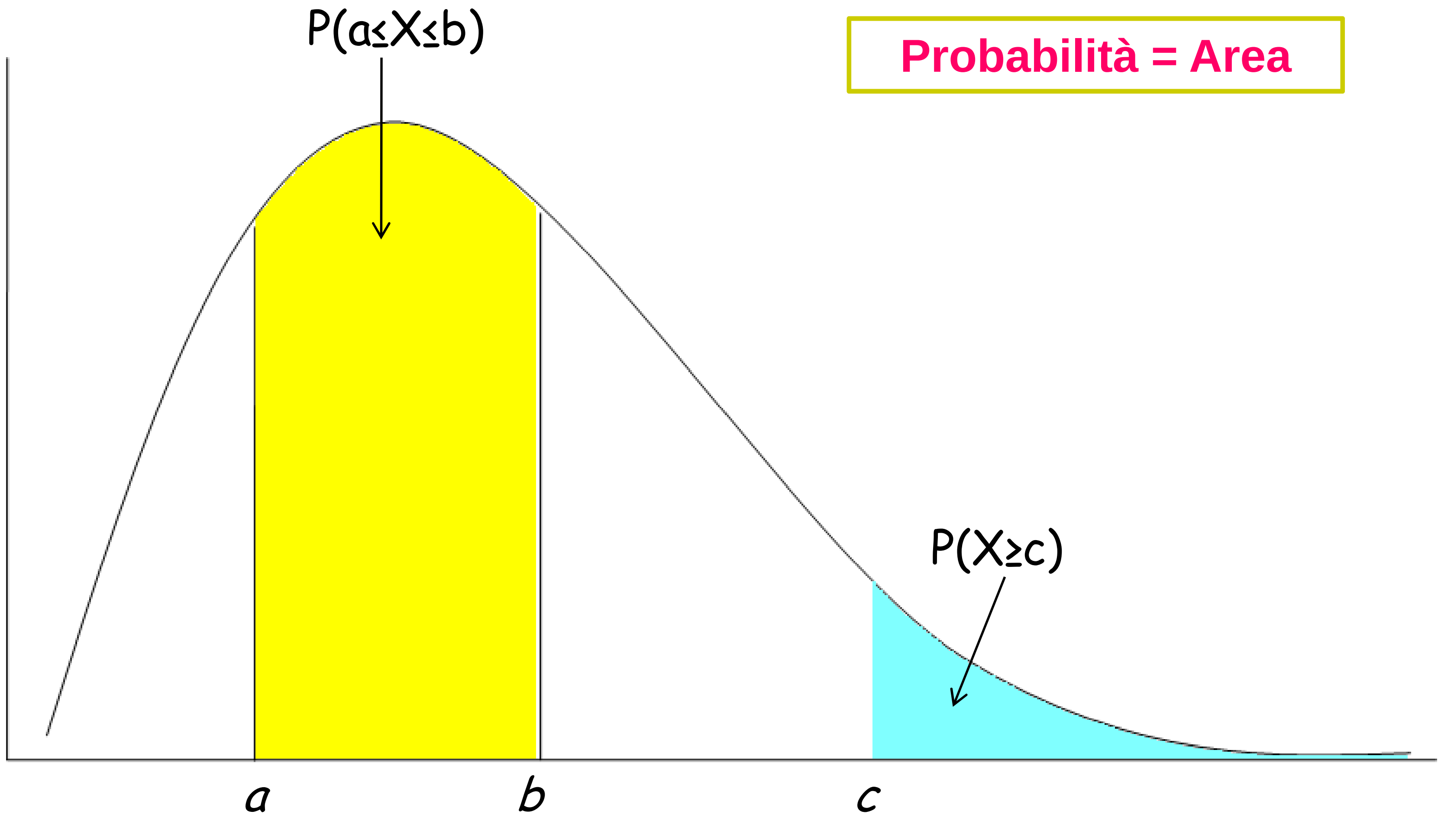
Variabile Casuale Continua
Una variabile casuale è detta continua se può assumere qualunque valore in un intervallo numerico. E' utilizzata per descrivere fenomeni che prevedono un'operazione di misurazione

La distribuzione di probabilità è specificata da una curva (**funzione di densità**) che consente di determinare la probabilità che la v.c. assuma valori in un prefissato intervallo

- Ogni intervallo ha una probabilità compresa tra 0 e 1 ed è l'area sotto la curva nell'intervallo
- L'intervallo contenente tutti i possibili valori della v.c. ha probabilità 1, ovvero l'area sotto la curva è pari a 1

Variabili casuali continue

Probabilità = Area



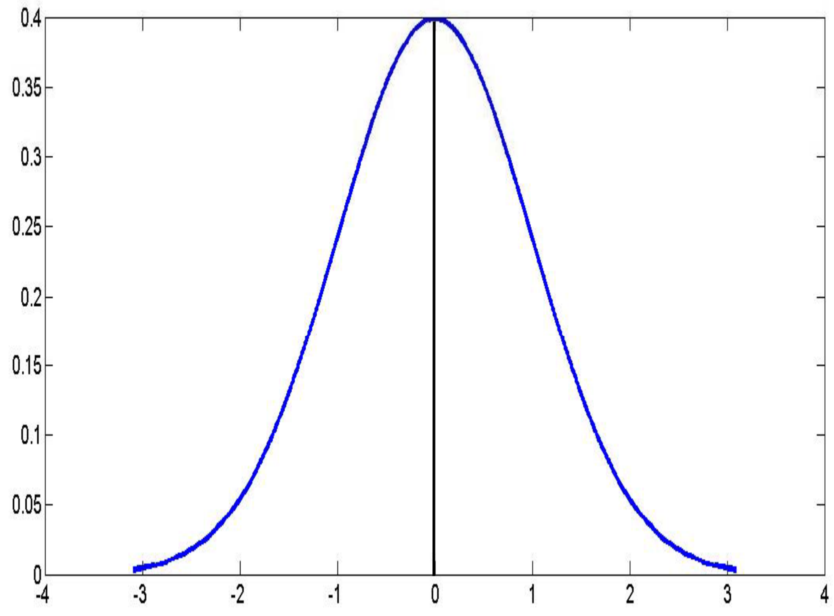
La v.c. Normale

La v.c. Normale svolge un ruolo di primaria importanza in Statistica:

1. fornisce una ragionevole approssimazione di molti fenomeni aleatori
2. rappresenta la distribuzione limite di altre v.c.
3. è essenziale nelle procedure inferenziali

E' di natura continua e la forma della sua distribuzione è

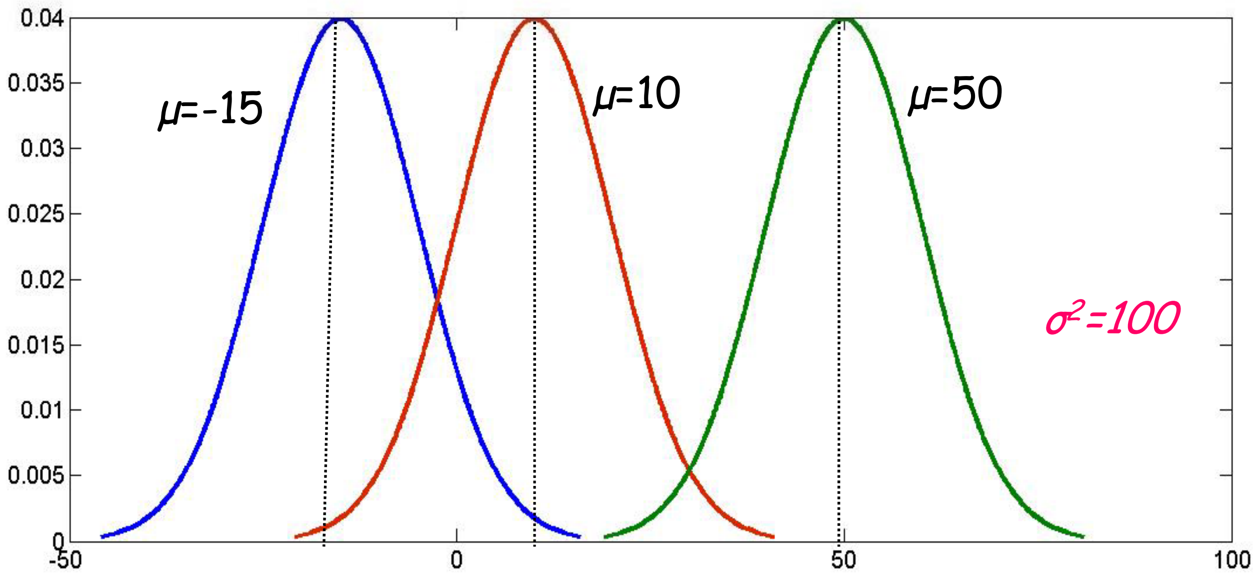
- unimodale
- **simmetrica intorno alla media**
- **campanulare**
- **determinata dalla media e dalla varianza**



La v.c. Normale

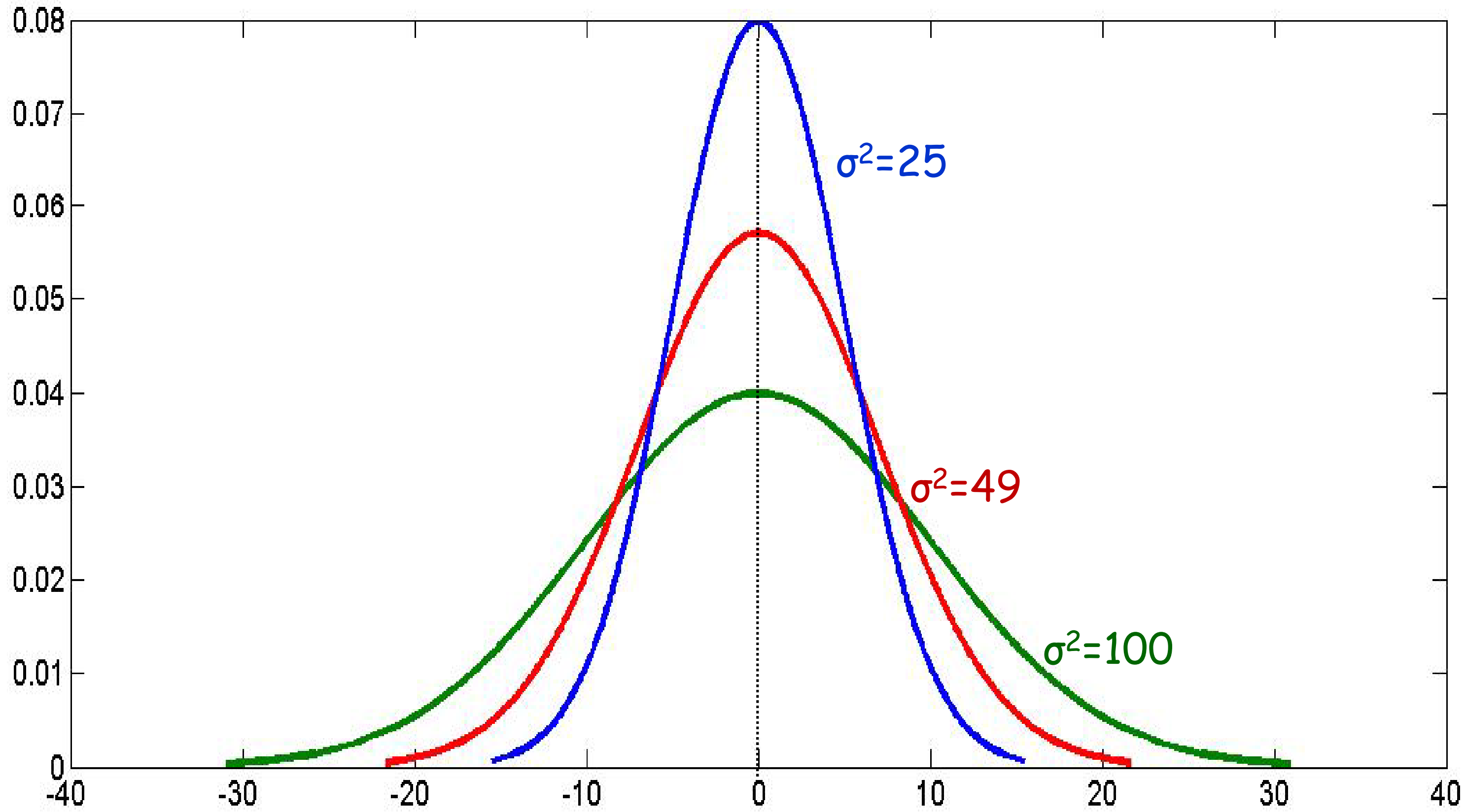
Non esiste una sola Normale ma tante curve a seconda della media e della varianza

Stessa varianza ma media diversa



La v.c. Normale

Stessa media ma varianza diversa



Come calcolare le probabilità

La variabile casuale Normale viene identificata attraverso la sua media e varianza:

$$X \sim N(\mu, \sigma^2)$$

e la probabilità che essa assuma un valore compreso tra a e b è l'area sottesa alla curva nell'intervallo (a,b).

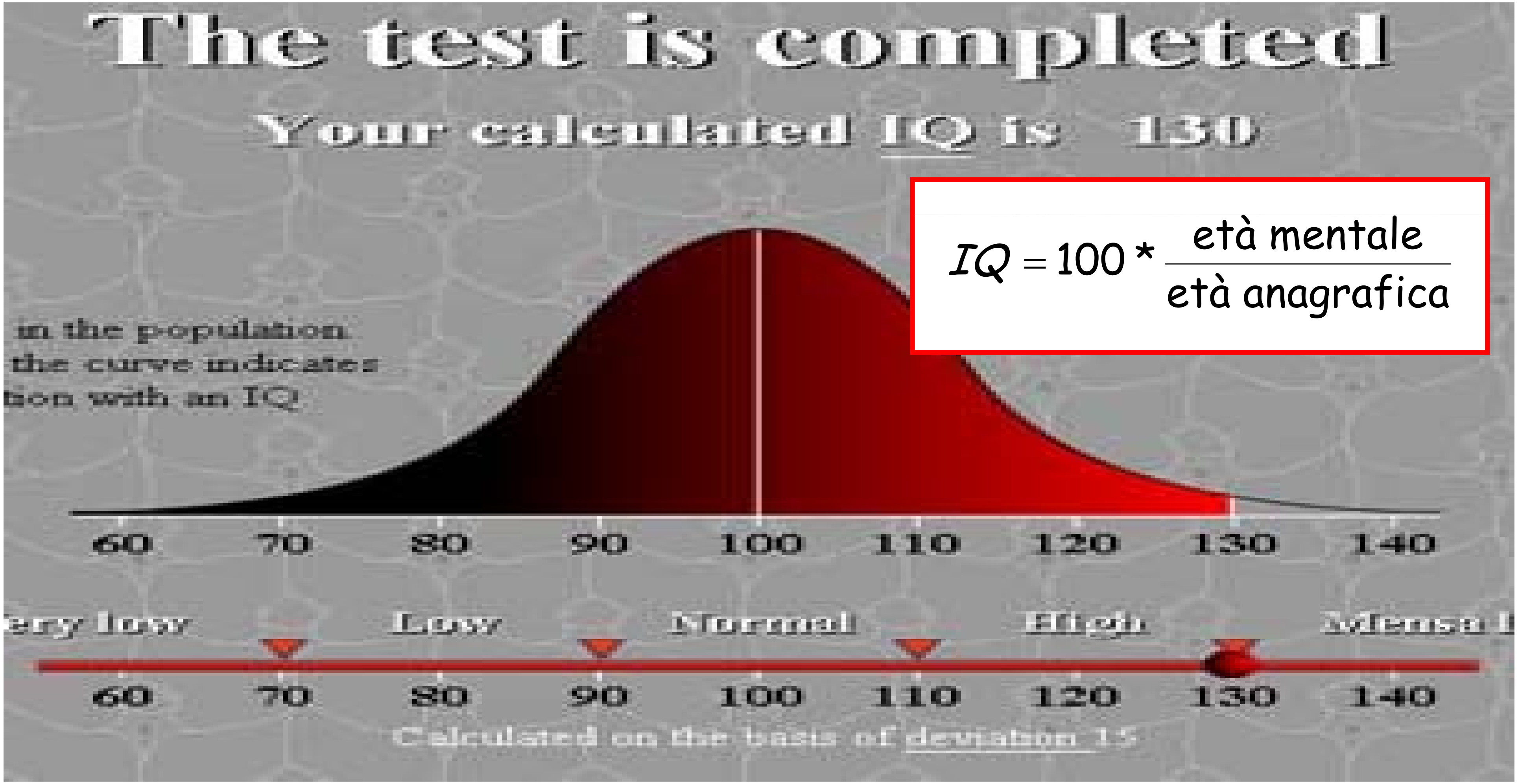
Il calcolo semplificato di una probabilità passa attraverso la definizione della v.c. Normale Standarizzata

$$Z = \frac{X - \mu}{\sigma}$$

la cui distribuzione è Normale con media zero e varianza unitaria

Come calcolare le probabilità

Il quoziente di intelligenza (IQ) ha una distribuzione normale con media 100 e deviazione standard 16



$$IQ = 100 * \frac{\text{età mentale}}{\text{età anagrafica}}$$

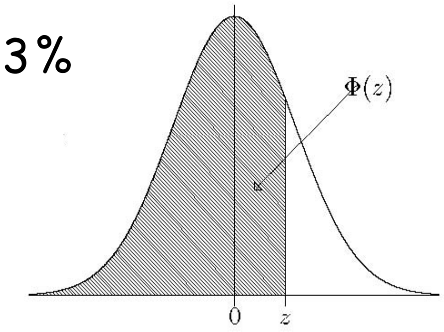
- Qual è la percentuale di persone con IQ "molto basso"?

$$P(X < 70) = P\left(\frac{X-100}{16} < \frac{70-100}{16}\right) = P(Z < -1.875)$$

$$= \Phi(-1.875) = 1 - \Phi(1.875) = 1 - 0.9696 = 0.0304 \rightarrow 3\%$$

Tavola 1 *(segue)*: Funzione di ripartizione della Variabile Casuale Normale Standardizzata

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	[illegible]
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	[illegible]
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	[illegible]
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	[illegible]
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	[illegible]
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	[illegible]
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	[illegible]
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	[illegible]
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

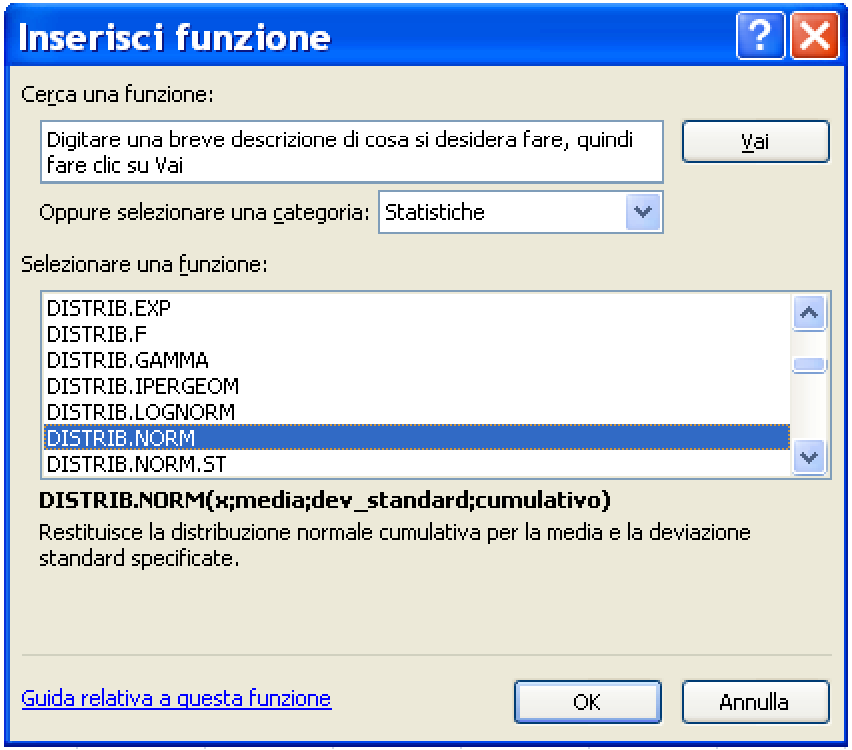


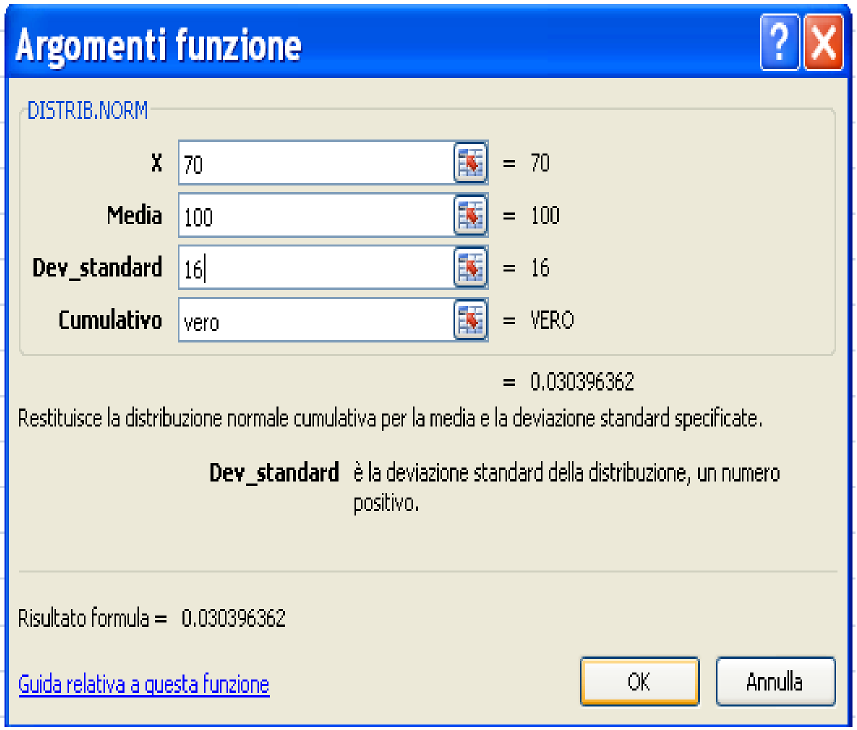
70 rappresenta il 3° percentile della distribuzione

- Qual è la percentuale di persone con IQ "molto basso" ?

$$P(X<70)=P\left(\frac{X-100}{16}<\frac{70-100}{16}\right)=P(Z<-1.875)$$

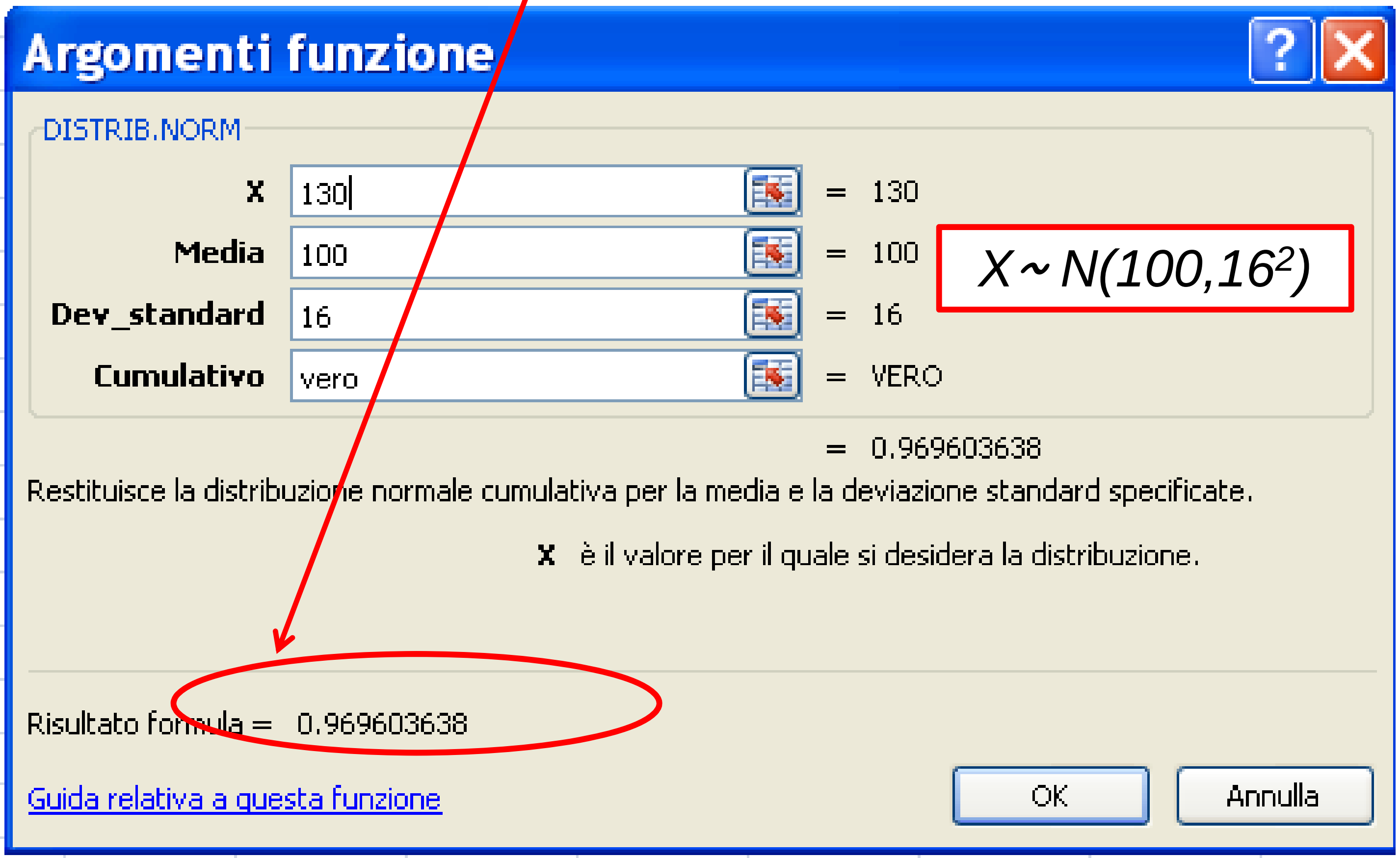
$$=\Phi(-1.875)=1-\Phi(1.875)=1-0.9696=0.0304\rightarrow 3\%$$





- Qual è la percentuale di persone con IQ "molto alto"?

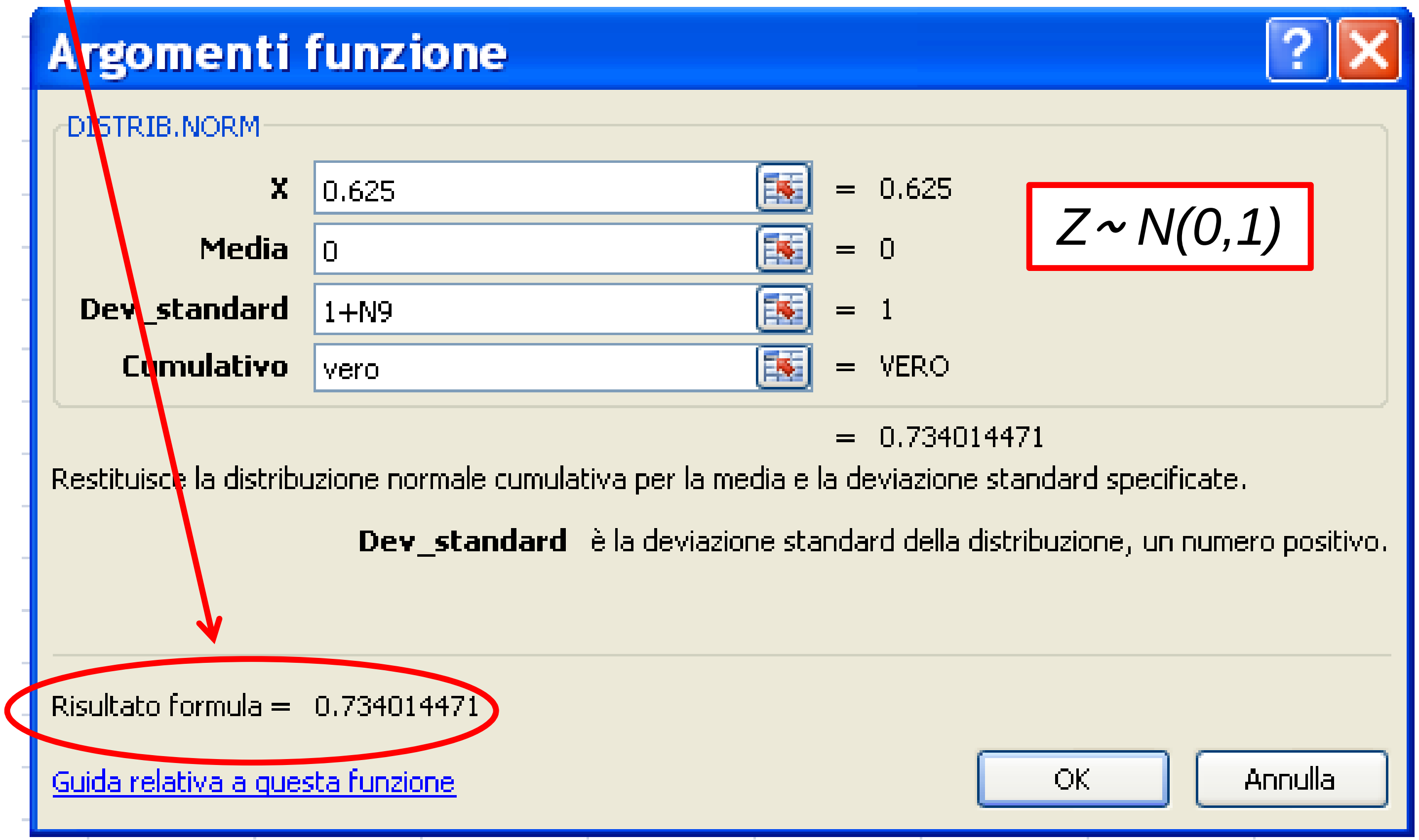
$$P(X > 130) = 1 - P(X < 130) = 1 - 0.9696 = 0.0304 \to 3\%$$



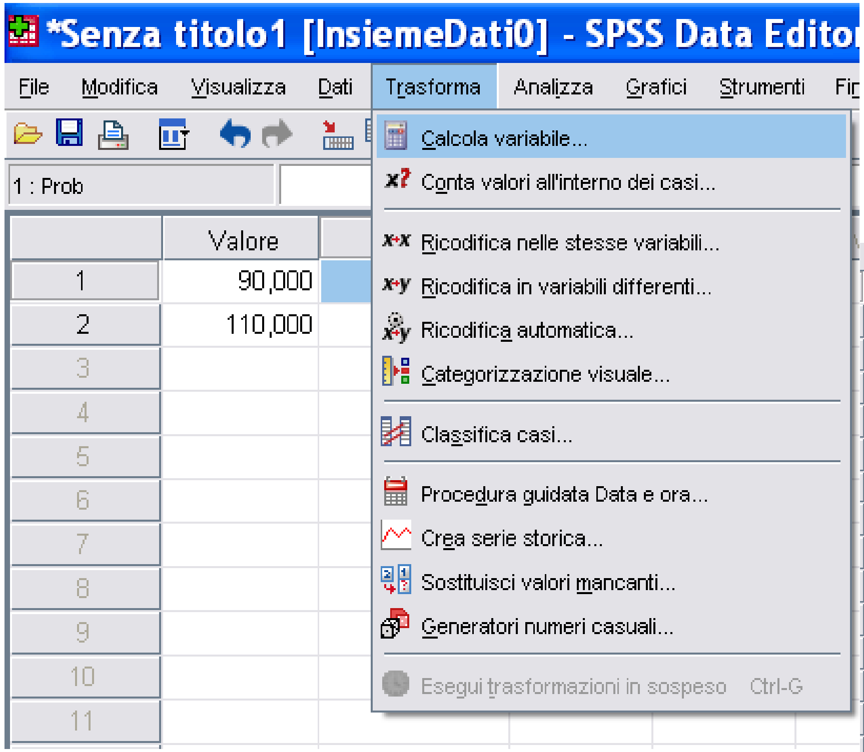
❑ Qual è la percentuale di persone con IQ "normale"?

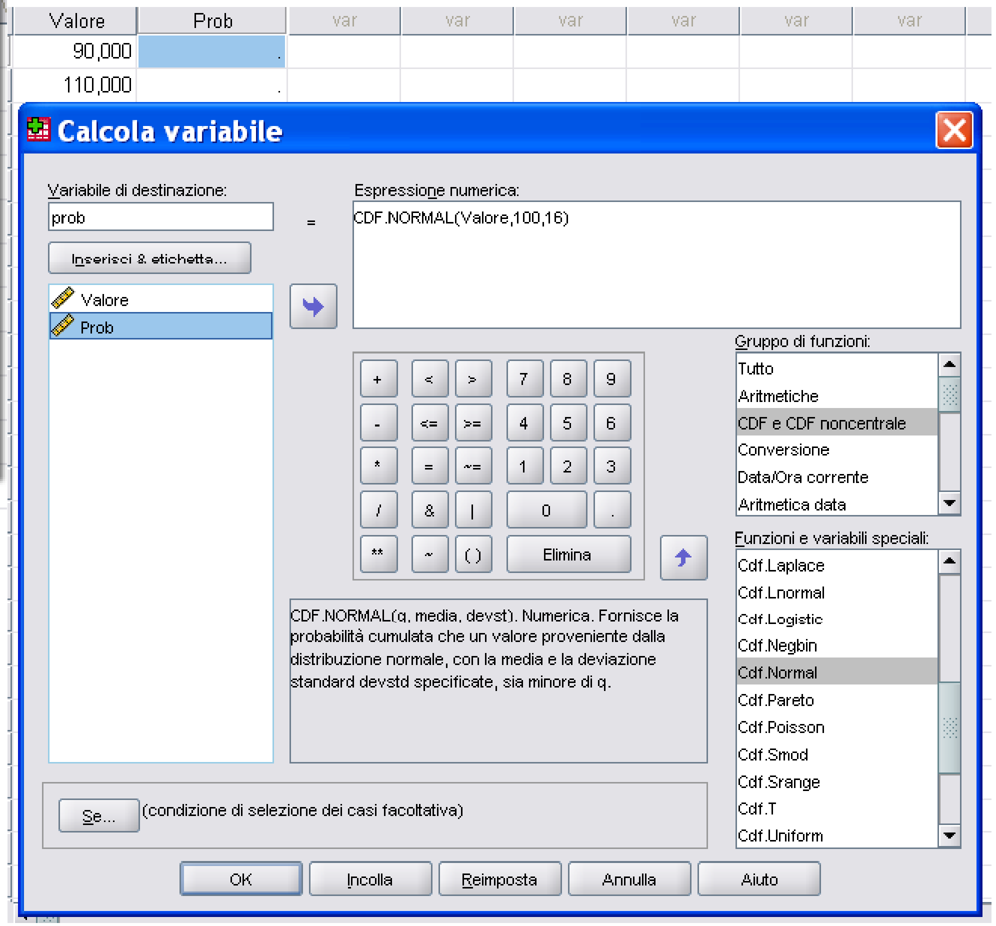
$$P(90 < X < 110) = \Phi\left(\frac{110-100}{16}\right) - \Phi\left(\frac{90-100}{16}\right) = \Phi(0.625) - \Phi(-0.625)$$

$$= 2 * \Phi(0.625) - 1 = 2 * 0.734 - 1 = 0.468 \rightarrow 46.8\%$$



$$P(90 < X < 110) = 0.468$$





L'ipotesi di Normalità

Numerose procedure statistiche si basano sull'assunto che il fenomeno in esame abbia una distribuzione Normale.

L'applicazione indiscriminata di metodi statistici che richiedono l'ipotesi di normalità conduce a risultati non conformi alla realtà quando i dati osservati non suffragano tale ipotesi.

Verificare la normalità di un fenomeno è quindi presupposto essenziale per la corretta applicazione dei metodi statistici.

A tal fine esistono diversi criteri: intervalli tipici, supporto grafico, verifica di ipotesi

Intervalli tipici

Sotto l'ipotesi di normalità è agevole definire le seguenti probabilità:

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$

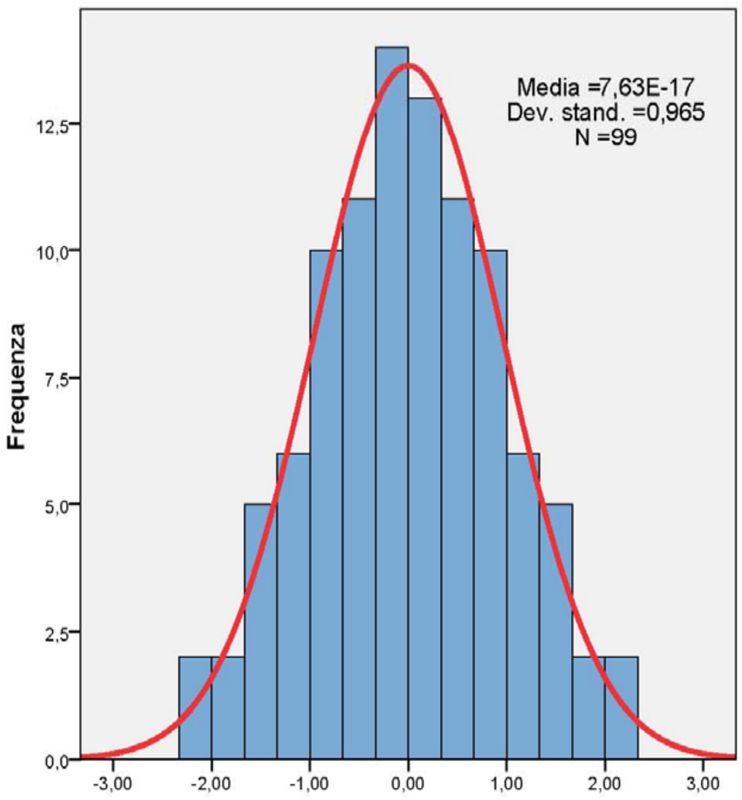
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

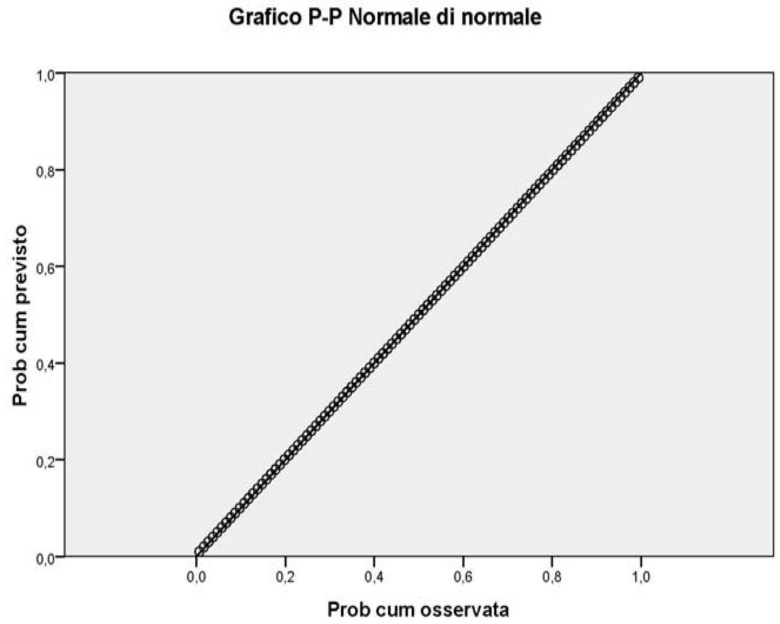
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

Regola empirica per verificare se i dati osservati supportano l'ipotesi di normalità

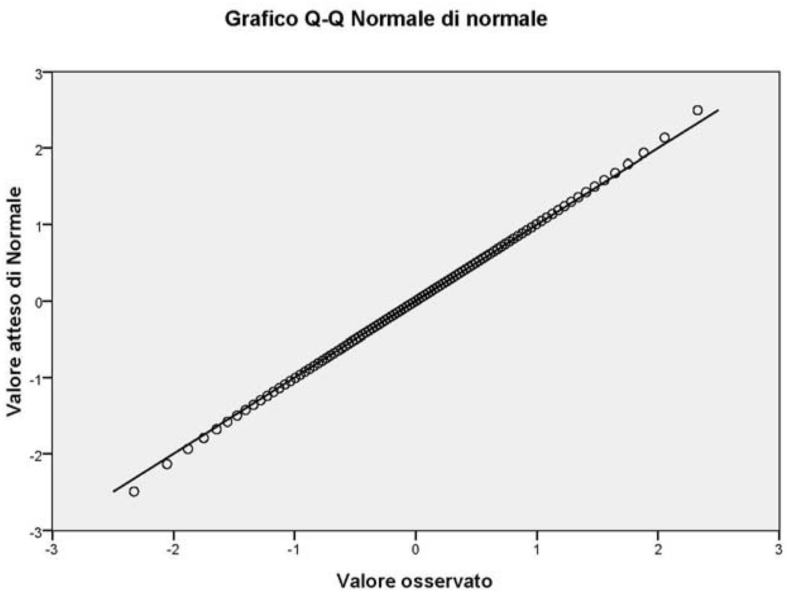
Metodi grafici

Istogramma



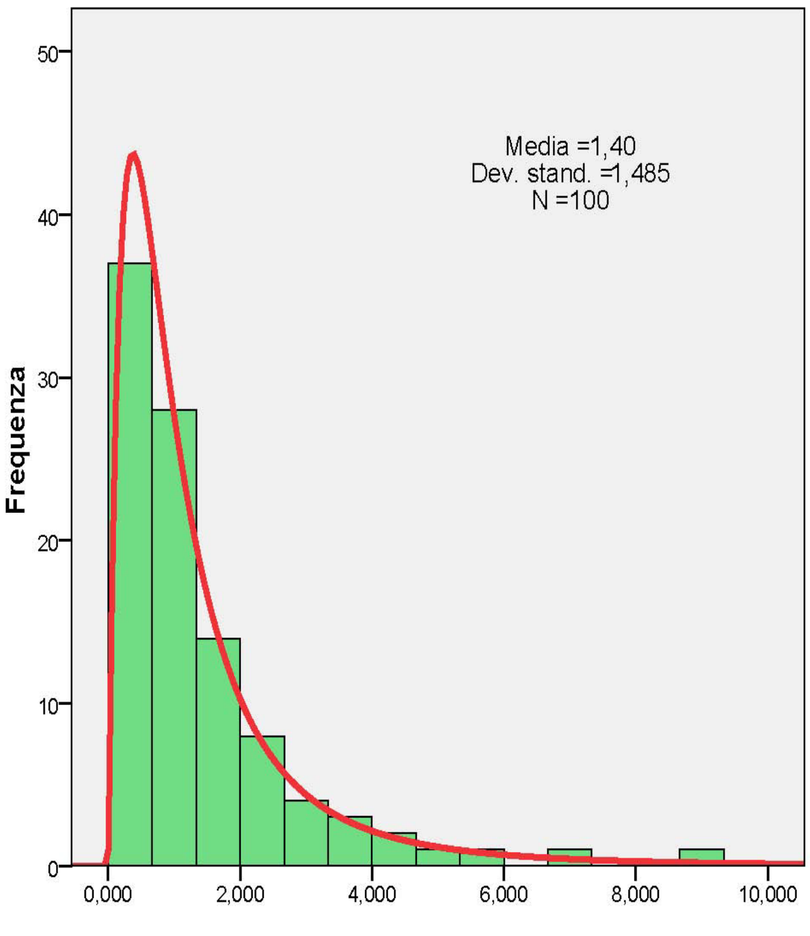


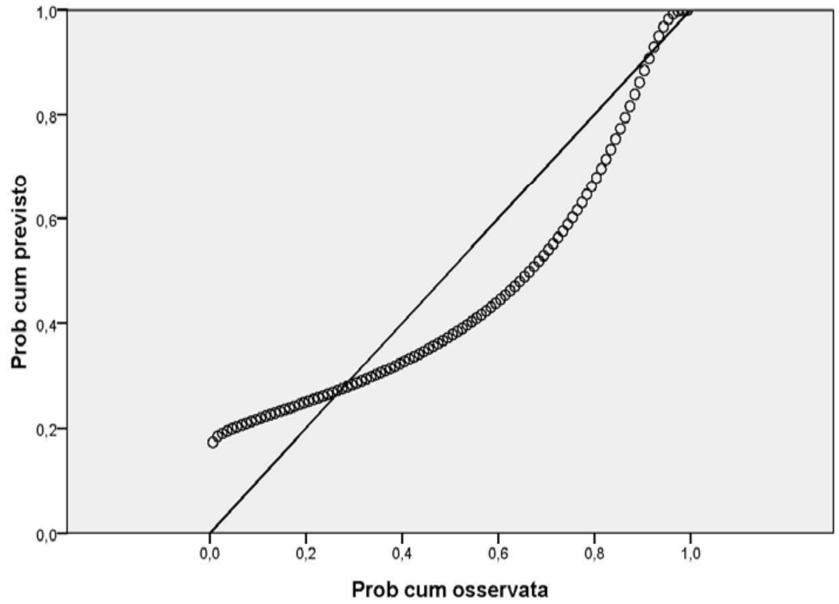
PP-Plot



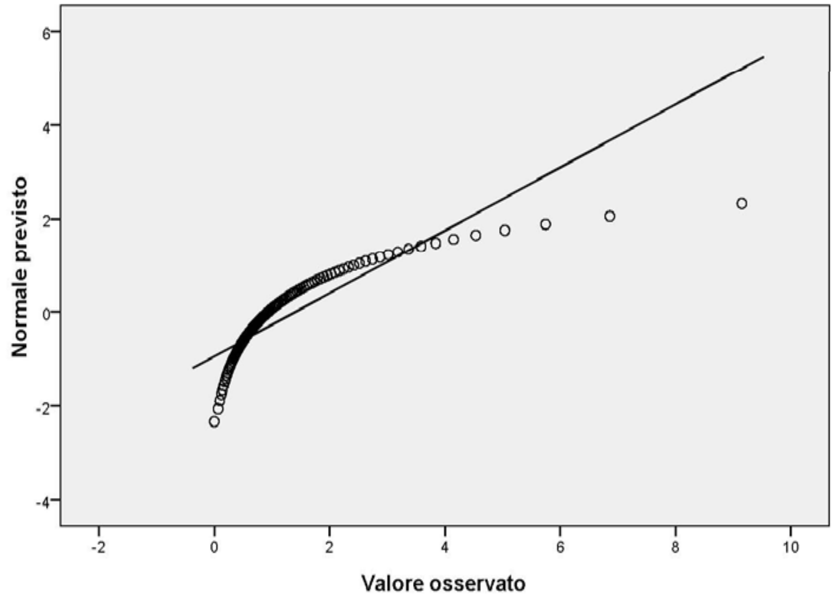
QQ-Plot

Istogramma



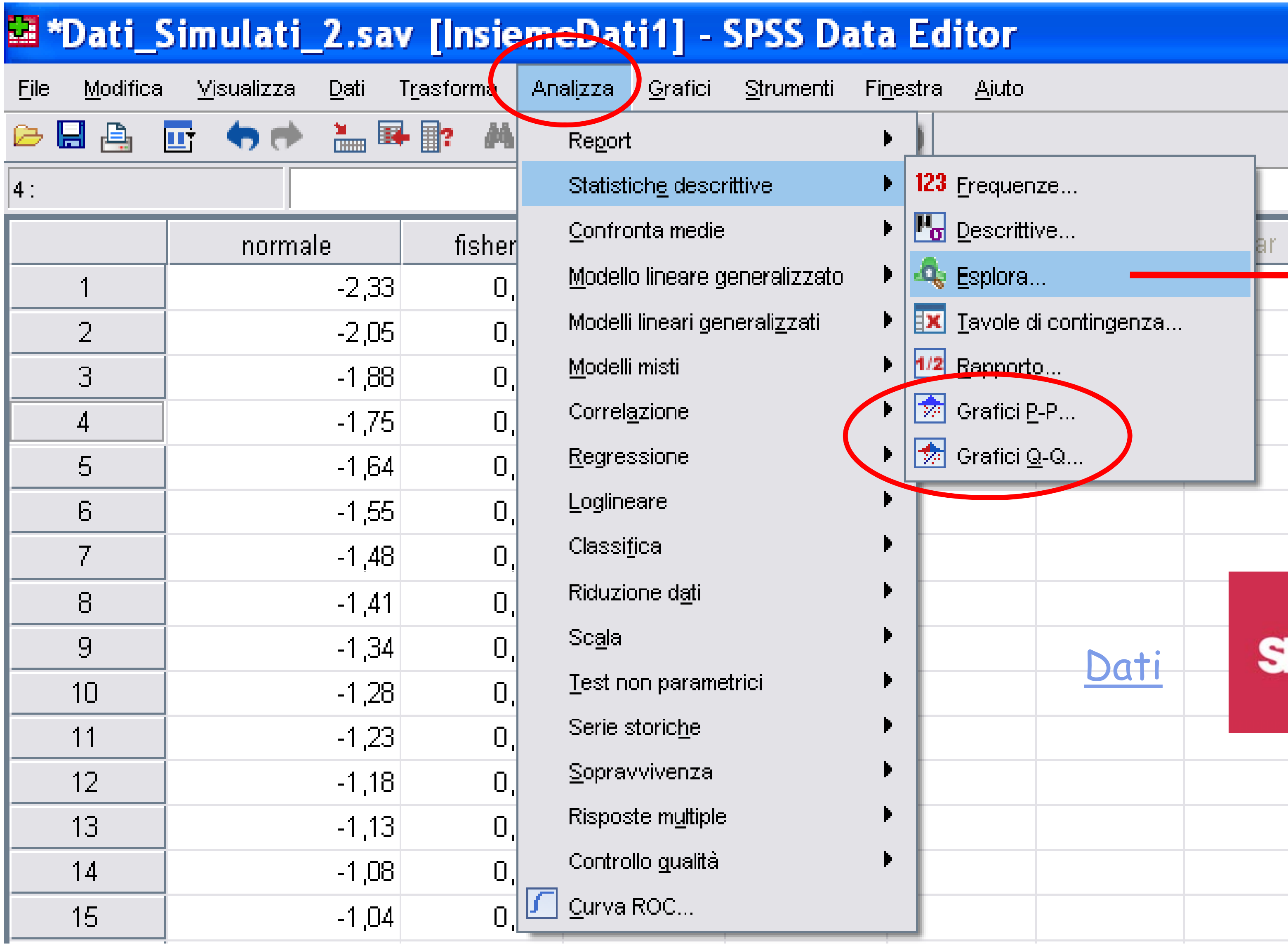


PP-Plot

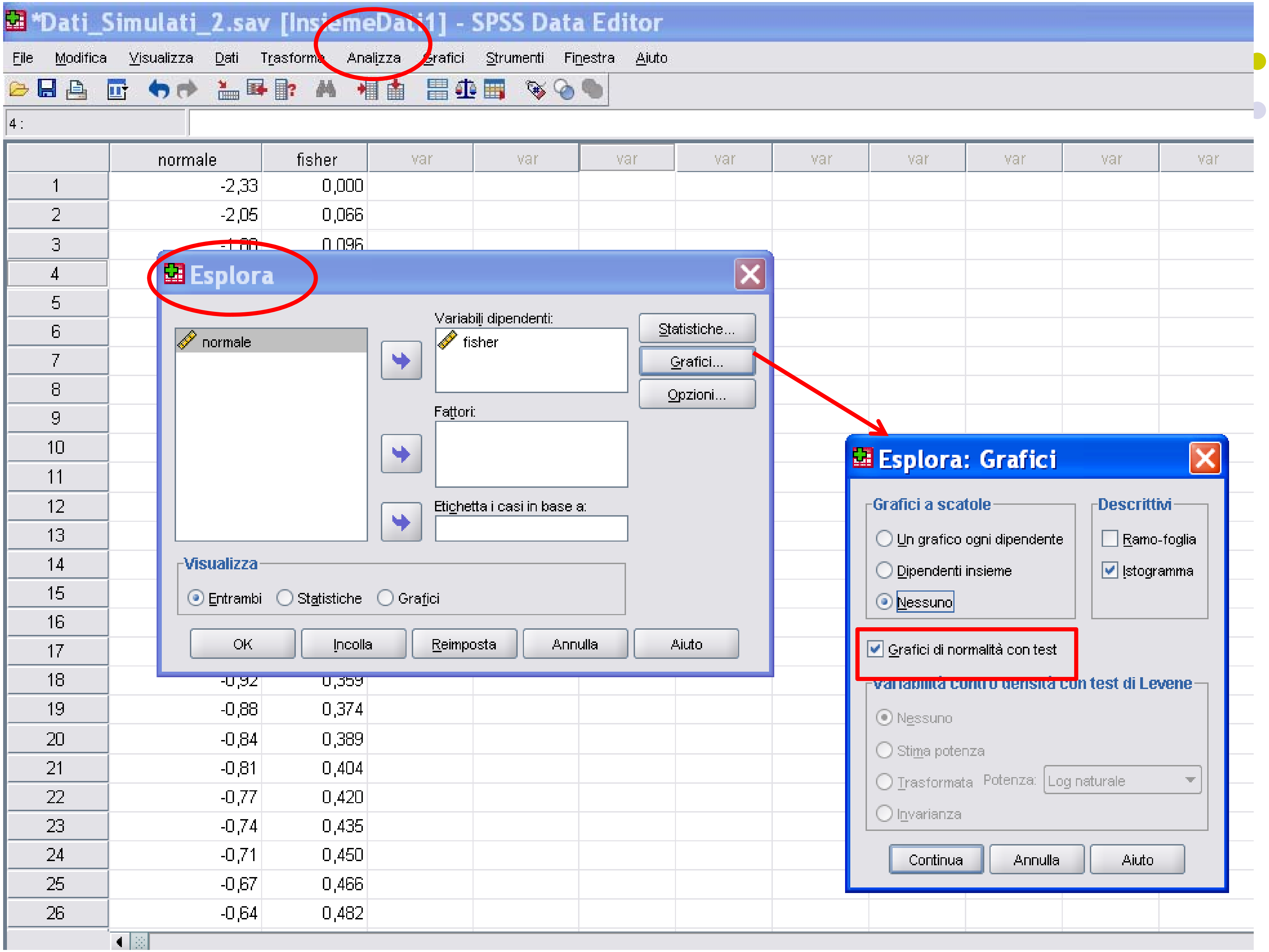


QQ-Plot

Accertare la normalità tramite SPSS

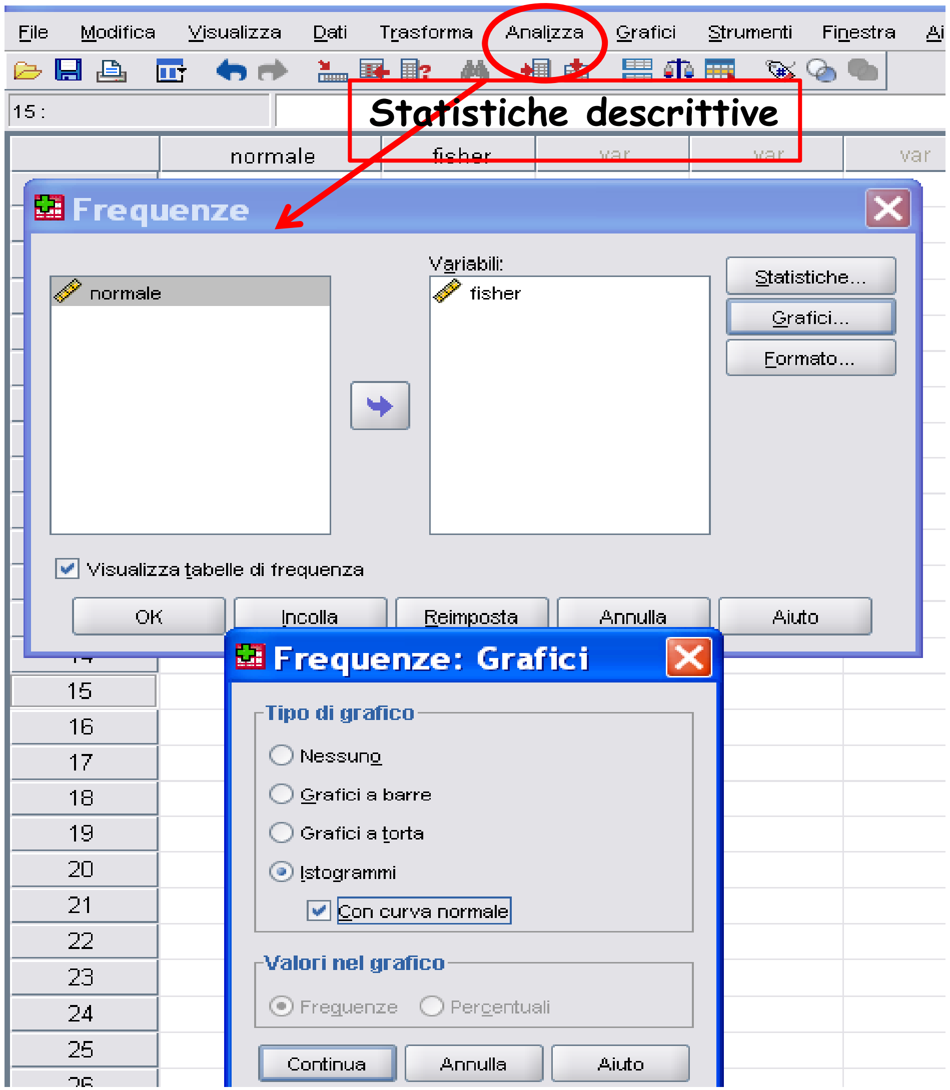


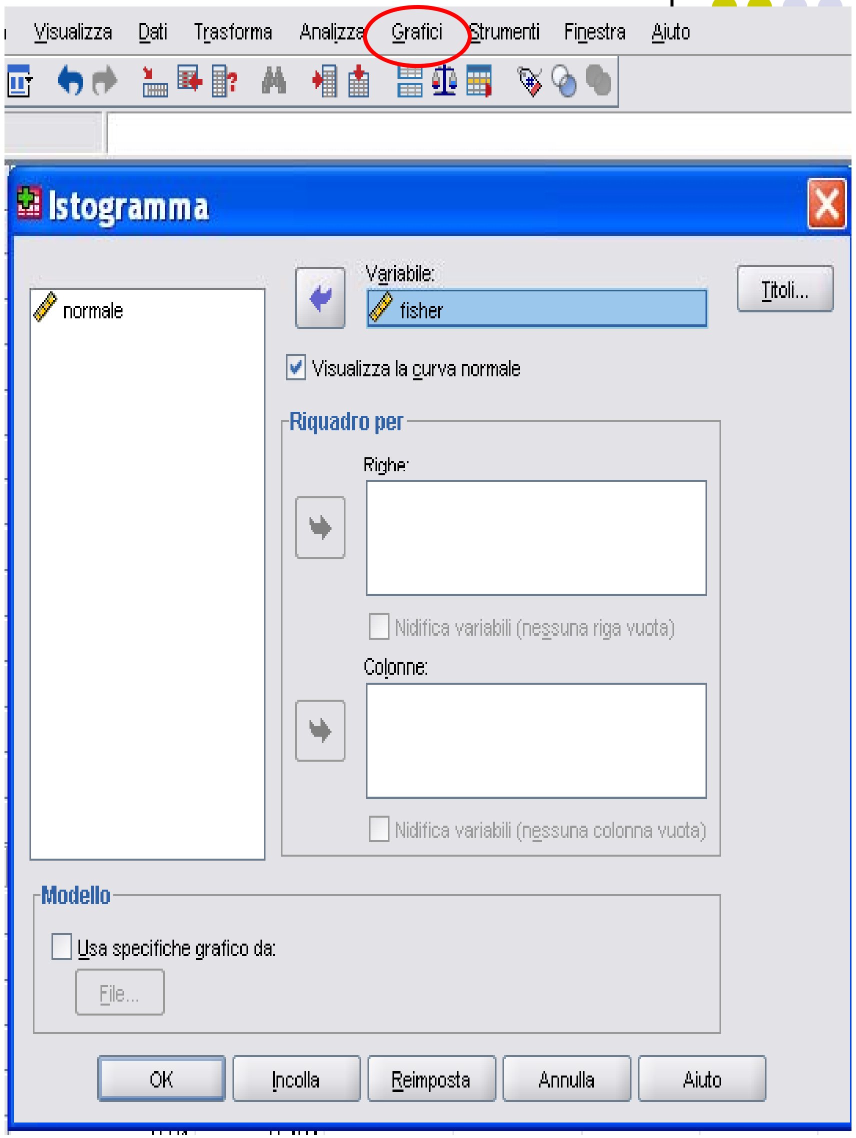
Dati



*Dati_Simulati_2.sav [InsiemeDati1] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
normale fisher var var var var var var var var var
Esplora
normale
Variabili dipendenti:
fisher
Fattori:
Etichetta i casi in base a:
Statistiche...
Grafici...
Opzioni...
Visualizza
Entrambi Statistiche Grafici
OK Incolla Reimposta Annulla Aiuto
Esplora: Grafici
Grafici a scatole
Un grafico ogni dipendente
Dipendenti insieme
Nessuno
Descrittivi
Ramo-foglia
Istogramma
Grafici di normalità con test
Variabilità contro densità con test di Levene
Nessuno
Stima potenza
Trasformata Potenza: Log naturale
Invarianza
Continua Annulla Aiuto

E' possibile sovrapporre la curva normale ad un istogramma scegliendo l'opzioni "con curva normale" oppure "visualizza la curva normale"





E’ anche possibile sovrapporre una qualsiasi curva ad un istogramma. Dopo aver visualizzato l’istogramma nel file di output, occorre attivare l’editor dei grafici cliccando due volte sul grafico stesso