Elementi di Calcolo delle Probabilità

- I dati che raccogliamo e su cui lavoriamo sono stati acquisiti con delle procedure in cui interviene **il caso**.
- Si pensi agli esperimenti e al campionamento

- Dobbiamo studiare "il comportamento" del caso
- La teoria della probabilità è basata sul concetto di ***esperimento casuale***; ovvero un esperimento il cui risultato non può essere previsto con certezza prima di eseguire l'esperimento.
- Di solito si assume che l'esperimento possa essere ripetuto all'infinito, essenzialmente sotto le stesse condizioni.
- Questa assunzione è importante poiché la teoria della probabilità si occupa dei risultati a lungo termine, al replicare dell'esperimento.

Il comportamento del caso e la valutazione della probabilità

- Perché le scommesse, che dipendono dalla faccia del dado, o dalla carta che uscirà che sono imprevedibili, sono un business redditizio per i casinò?
- Se si sottopongono tutti gli impiegati di un'azienda a un test sull'Aids, qual è la probabilità che almeno un test sia positivo se tutti gli individui sottoposti al test sono sani?
- Conoscendo il gruppo sanguigno di 100 individui, qual è la probabilità che, per un individuo scelto a caso, il gruppo non sia A?

I fenomeni aleatori e il comportamento del caso

- Per rispondere alle precedenti domande dobbiamo studiare i **fenomeni aleatori o casuali** e le regole che governano il **comportamento del caso**.
- Il comportamento del "Caso" non è prevedibile a breve termine, ma ha un **andamento regolare** e **probabilisticamente** prevedibile a lungo termine.

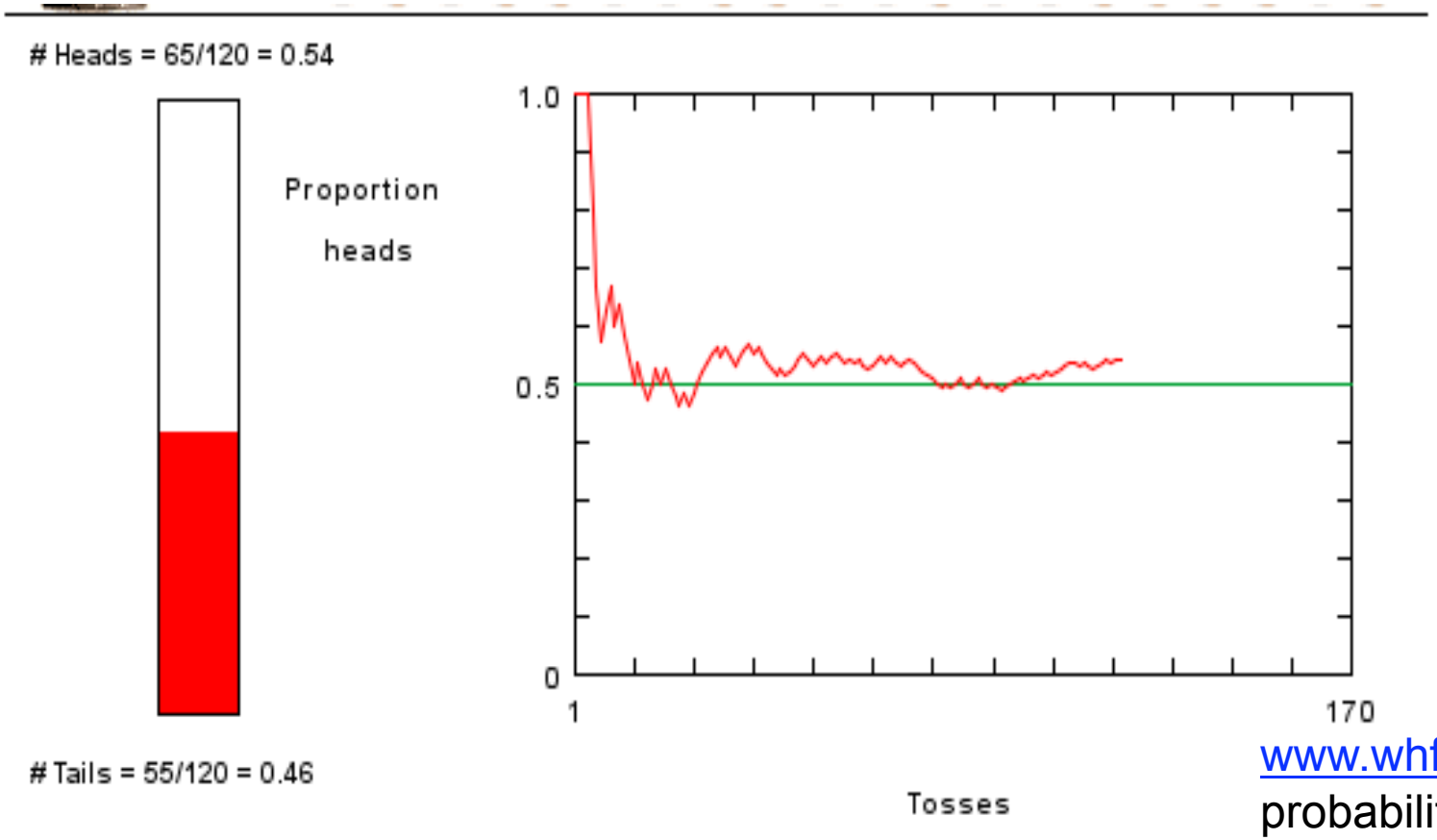
- I risultati dei fenomeni aleatori mostrano un andamento che rivela delle regolarità in modo chiaro in numerose ripetizioni anche se il risultato di ciascuna prova non è prevedibile.
- Questo fatto notevole è alla base **dell'idea di probabilità e della possibilità di valutazione probabilistica in molti casi di interesse.**

Un esempio

Cosa succede se lanciamo una moneta molte volte?

Nell'esempio: H H H H T T H T T H H H H T T…..

Si vede, dalla figura, come la proporzione (frequenza relativa) di teste sia molto variabile nei primi lanci…..



120 lanci

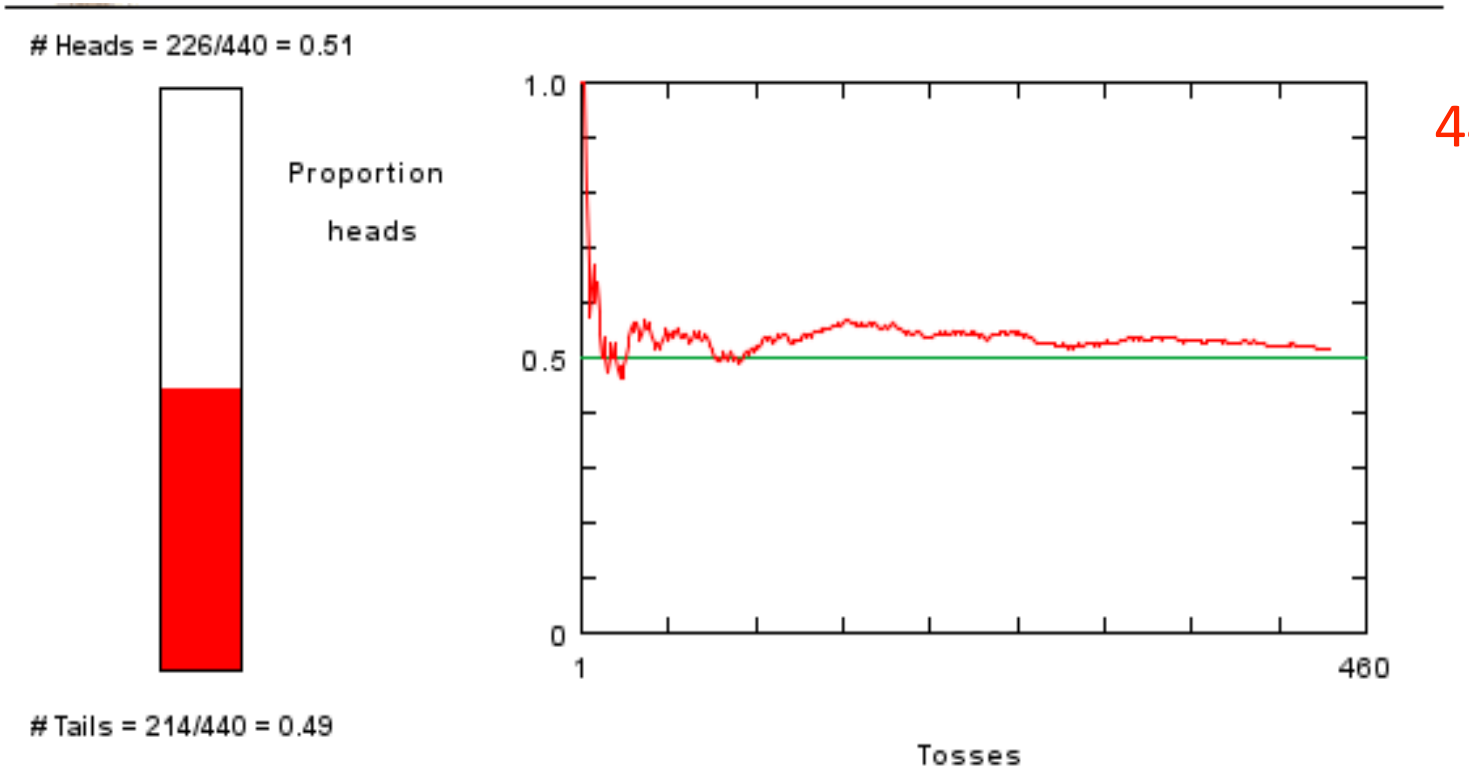
www.whfreeman.com/scc

probability applet

Un esempio (continua)

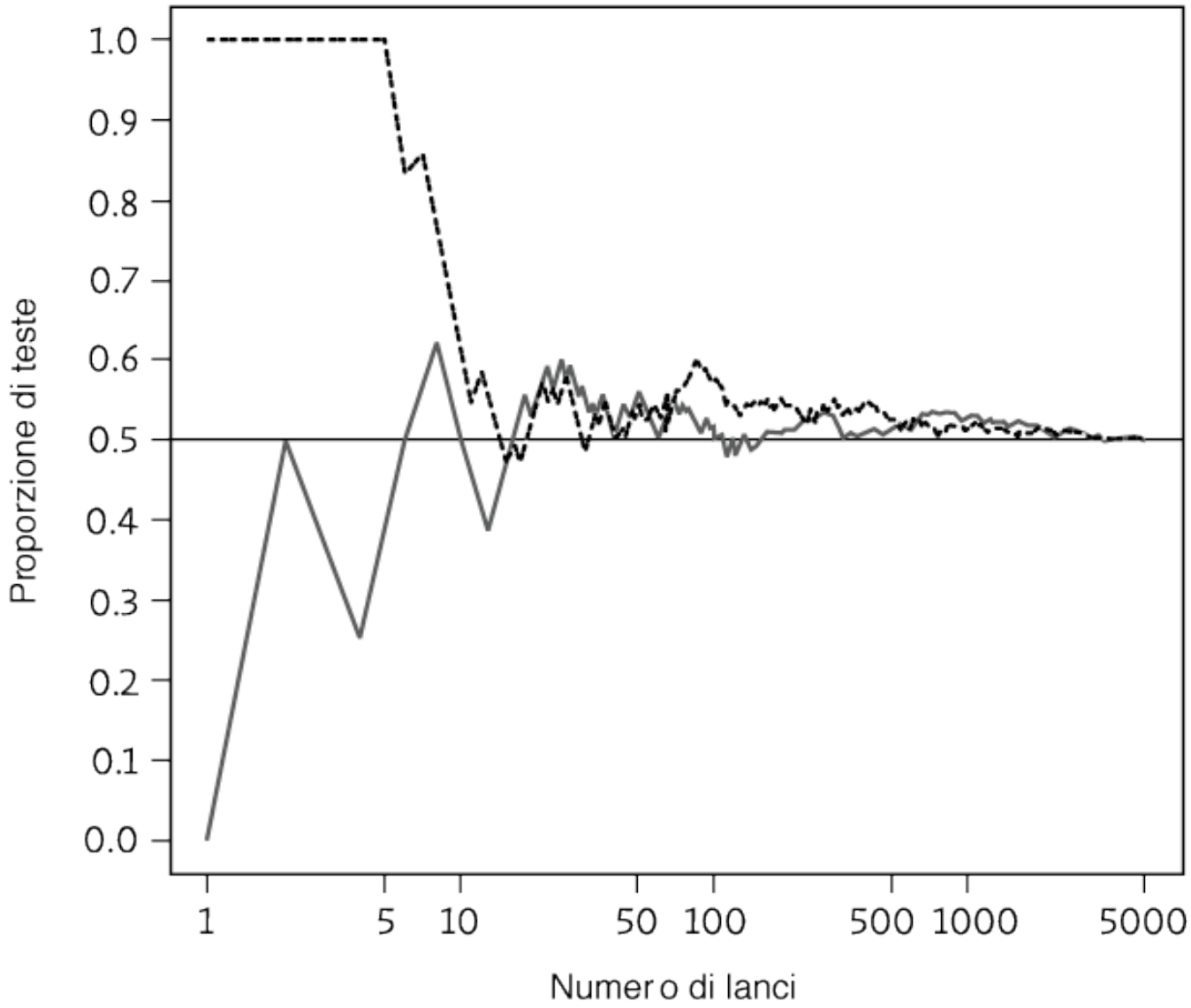
Aumentando il numero di lanci la proporzione (frequenza relativa) di teste si avvicina a 0.5

Diciamo che **0.5 è la probabilità di testa**



440 lanci

ESEMPIO



5000 lanci

Proporzione dei lanci il cui risultato è testa all'aumentare dei lanci. Per un numero elevato di lanci tale proporzione è molto vicina a 0.5. Il grafico mostra i risultati ottenuti con 2 esperimenti.

L'idea di probabilità

- **Probabilità 0.5** significa che ci aspetta che l'evento di interesse "si verifichi metà delle volte su un gran numero di prove".

- Si potrebbe sospettare che la probabilità che esca testa è 0.5 perché la moneta ha 2 facce, ma non basta (esistono le monete truccate).
- Analogamente anche i bambini possono essere di sesso M o F e le probabilità, calcolate statisticamente, non sono uguali: la probabilità di M è circa 0.51.

- Questa idea di probabilità è empirica. La probabilità descrive ciò che ci si aspetta che succeda in moltissime prove.

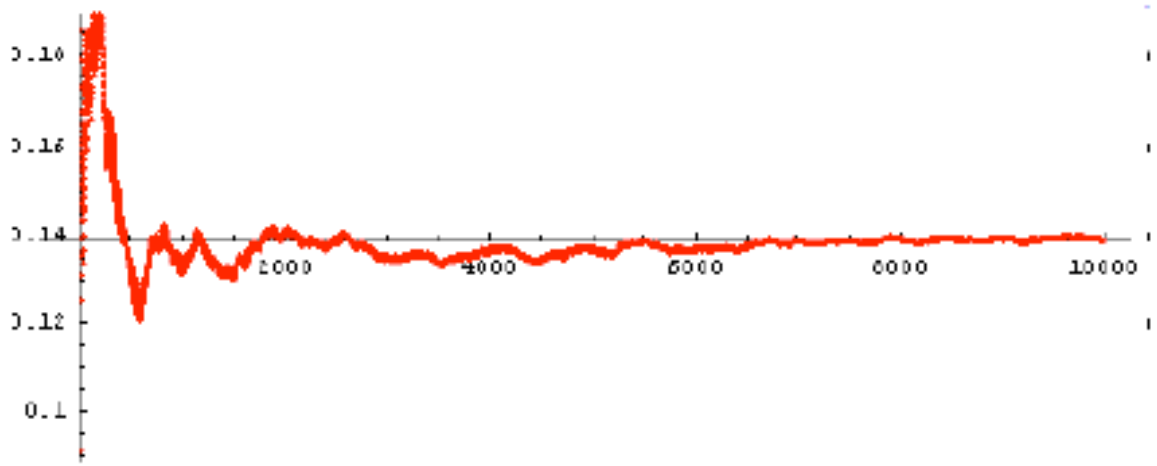
La probabilità di un evento e la legge empirica del caso

La legge empirica del caso

- Esperimento E = lancio due dadi.
- Evento A = esce 8
- In n =10000 prove si osservano k = 1386 successi di A

  $\Rightarrow k/n = 0.1386 = 13.86\%$

  $\Rightarrow$ Prob(A) $\cong$ 13.86%



P(A)= P("esce 8")
= 5/36
= 0.138

Miti e false credenze

- L'idea di probabilità si basa sul fatto che i fenomeni aleatori "tendono a regolarizzarsi su un gran numero di prove".

- Ma la nostra intuizione su ciò che è aleatorio ci inganna.
- Pensiamo che tale regolarità si manifesti anche su poche prove.

- Se ciò non accade cerchiamo delle spiegazioni… che sono, generalmente,in contraddizione con la teoria delle probabilità

Miti e false credenze

- Esempio 1

Lanciando 8 volte una moneta, quale sequenza è più probabile?

H T H T T H H T oppure

T T T H H H H H

- Esempio 2

Se in una famiglia nascono 8 maschi, cosa vi aspettate per il sesso del prossimo nascituro?

Es 1

A questa domanda la maggior parte risponde che è più probabile la prima. Ma sono entrambe ugualmente probabili. La moneta non ha memoria. Non sa quali fossero i risultati precedenti e non può cercare di creare una sequenza bilanciata.

Es 2

La maggior parte delle risposte è: femmina

La serie consecutiva di nascite maschili (che è prevista dal calcolo delle probabilità) sorprende. Poiché **non** siamo abituati a vedere lunghe serie di prove non abbiamo una buona percezione del comportamento del caso.

Molti credono, a torto, nella "legge dei piccoli numeri".

In realtà è dimostrato statisticamente che le sequenze di nascite di bambini di uno stesso sesso sono più frequenti di quanto il semplice caso suggerirebbe. In altre parole nel caso precedente è più probabile (la spiegazione è genetica) che si osservi un'altra nascita maschile.

Elementi di Calcolo delle Probabilità: modello probabilistico

Ad un esperimento casuale (v. pag 1) sono sempre associati:

1) un insieme di tutti i possibili risultati (spazio campionario Ω). I possibili risultati sono “punti” in questo “spazio”.

3) una famiglia di suoi sottoinsiemi (gli eventi)

5) una probabilità P associata ad ogni evento

Questi 3 elementi costituiscono un modello matematico adatto a descrivere un esperimento casuale.

Quale spazio campionario Ω?

- Es: esperimento aleatorio o casuale: lancio di un dado
- Es: esperimento casuale: misurazione del peso dei ragazzi di 11 anni in una determinata città.

  Prendiamo come spazio campionario Ω l'insieme [0,∞), anche se la maggior parte degli elementi di questo insieme sono impossibili all'atto pratico.
- Es: esperimento casuale: si vuole determinare il minimo dosaggio di un farmaco al quale un paziente reagisce positivamente.
- Si potrebbe pensare Ω = (0, ∞)
- Un evento E potrebbe essere il dosaggio è compreso fra 2 e 10→ E = (2,10)

Si noti che su molti testi **Ω è indicato con S**

La probabilità di un evento: definizione frequentista

- Intuitivamente, la probabilità di un evento dovrebbe misurare la **frequenza relativa** dell'evento a lungo termine. Specificamente, supponiamo di ripetere indefinitamente l'esperimento casuale.
- Per un evento *A* dell'esperimento la probabilità è la proporzione di volte nella quale l'evento si verificherebbe se si ripetesse l'esperimento infinite volte nelle stesse condizioni

- Ad es. se campioniamo casualmente un individuo da una popolazione composta per 2/3 da maschi, la prob. di campionare un maschio sarà 2/3

La probabilità di un evento

La probabilità matematica è, quindi, un'idealizzazione teorica di quello che potrebbe accadere in una serie infinita di prove.

Ma molto spesso si studiano eventi aleatori per i quali **non è possibile calcolare la probabilità teorica**.

In questi casi accettiamo come probabilità, la frequenza relativa di un evento che si ottiene da un numero abbastanza elevato di prove o di osservazioni, tutte effettuate nelle stesse condizioni.

La probabilità di un evento

Ad esempio, non è possibile sapere la probabilità che esca "testa" lanciando una moneta *truccata*. L'unico modo per conoscere tale probabilità è di lanciare un gran numero di volte la moneta registrando i risultati.

Ad esempio, su 1000 lanci otteniamo 612 volte testa.

Si può dire che la probabilità di ottenere testa con quella moneta è pari a 0.612.

Questa probabilità ottenuta empiricamente non può essere calcolata con precisione.

Quasi certamente se si facesse un'altra serie di 1000 lanci si otterrebbe un risultato lievemente diverso!

La probabilità di un evento

PROBABILITA'

- si esprime con un numero compreso fra 0 e 1
- la somma delle probabilità di tutti i possibili eventi è =1
- la probabilità che un evento NON si verifichi è pari a 1 meno la probabilità che l'eventi si verifichi

É intuitivo che:

probabilità che un certo carattere sia vero per un individuo estratto a caso da una popolazione $\simeq$ (equivale) **proporzione della popolazione che è provvista di quel carattere**

Ad esempio, se il 75% dei ceppi di *Enterococcus* è resistente alla tetraciclina, allora avremo una probabilità pari a 0.75 (75%) che un ceppo di *Enterococcus* preso a caso sia resistente.

La probabilità di un evento

In una serie di prove, ripetute un gran numero di volte ed eseguite tutte nelle stesse condizioni, la **frequenza relativa** tende ad assumere valori prossimi alla probabilità dell'evento stesso e l'approssimazione è tanto maggiore quanto più numerose sono le prove eseguite.

Esempio: Se negli ultimi 30 anni nella nostra città ha nevicato 18 volte, la probabilità che nevichi quest' anno è 18/30=3/5.

Esempio: Si vuole calcolare la probabilità che un neonato sia femmina. Su 100.000 nascite si sono avute 48.500 femmine. Essendo il numero di prove sufficientemente elevato ed ogni prova indipendente dall'altra, utilizziamo la definizione frequentista:

P(F) = 48500 / 100000 = 0,485 P(M) = 51500 / 100000 = 0,515

La probabilità: esempio

Un altro ***esempio***: *Qual è la probabilità che un lavoratore abbia un infortunio sul lavoro?*

Difficile rispondere! Certamente dipende dal lavoro che fa.
Un operaio minatore è sicuramente più a rischio di un impiegato alle Poste. Questo perché statisticamente ci sono più infortuni lavorando in una miniera che in un Ufficio Postale.

Esempio: Mi aspetto che un capretto che deve nascere abbia il vello a macchie o che l'abbia nero?
Come posso "misurare" la probabilità con cui si produce un evento o l'altro?

Definizione frequentista: rapporto tra il numero di volte in cui si è verificato l'evento e il numero di prove fatte

Avremo bisogno di far fare molti figli alle nostre capre

Se la probabilità che un figlio sia nero è ¾ quale è la prob. che sia a macchie?

Dal campione alla popolazione

Campione	Popolazione
Unità statistiche	Unità statistiche
Variabili osservate	Variabili aleatorie
Valori delle variabili	Valori delle v. a.
Statistiche descrittive o Indici riassuntivi	Stimatori dei parametri incogniti
Distribuz. di frequenze relative	Distr. di probabilità
Certezza	**Incertezza**

Dalle distribuzioni di frequenze alle distribuzioni di probabilità

Esempio. Si è osservato **un campione** di 730 nidi di una particolare specie di uccello in una determinata foresta e si è costruita la distribuzione di frequenze del numero di uova per nido

Quale esperim. casuale? Ω =? Quali eventi elementari?

N° uova	Frequenze	Frequenze relative
0	**90 nidi**	**0.12**
1	**165**	**0.23**
2	**209**	**0.29**
3	**187**	**0.26**
4	**67**	**0.09**
5	**12**	**0.01**
Totale	**730**	**1.00**

Continuazione dell'Esempio.

Volendo studiare l'intera **popolazione incognita di nidi** da cui proviene il campione, poiché le frequenze relative per tutta la popolazione non sono note si può pensare (usando la def. frequentista di probabilità) che, ad es., 0.12 rappresenti la probabilità di trovare 0 uova in un nido della popolazione.

In questo studio si vogliono calcolare le **probabilità** dei possibili valori della **variabile aleatoria n° di uova per nido** e la legge secondo cui tali probabilità evolvono, ossia **la distribuzione di probabilità della v. a.**

Continuazione dell'Esempio.

Si pensa a un Modello Probabilistico (Stocastico) per rappresentare il fenomeno aleatorio (esperimento casuale) osservato con riferimento all'intera popolazione

Un fenomeno aleatorio non è prevedibile e ha caratteristiche aleatorie

In altre parole, si cerca una **rappresentazione idealizzata della realtà** ossia di quello che si osserva

Distribuzioni di probabilità – Modelli probabilistici

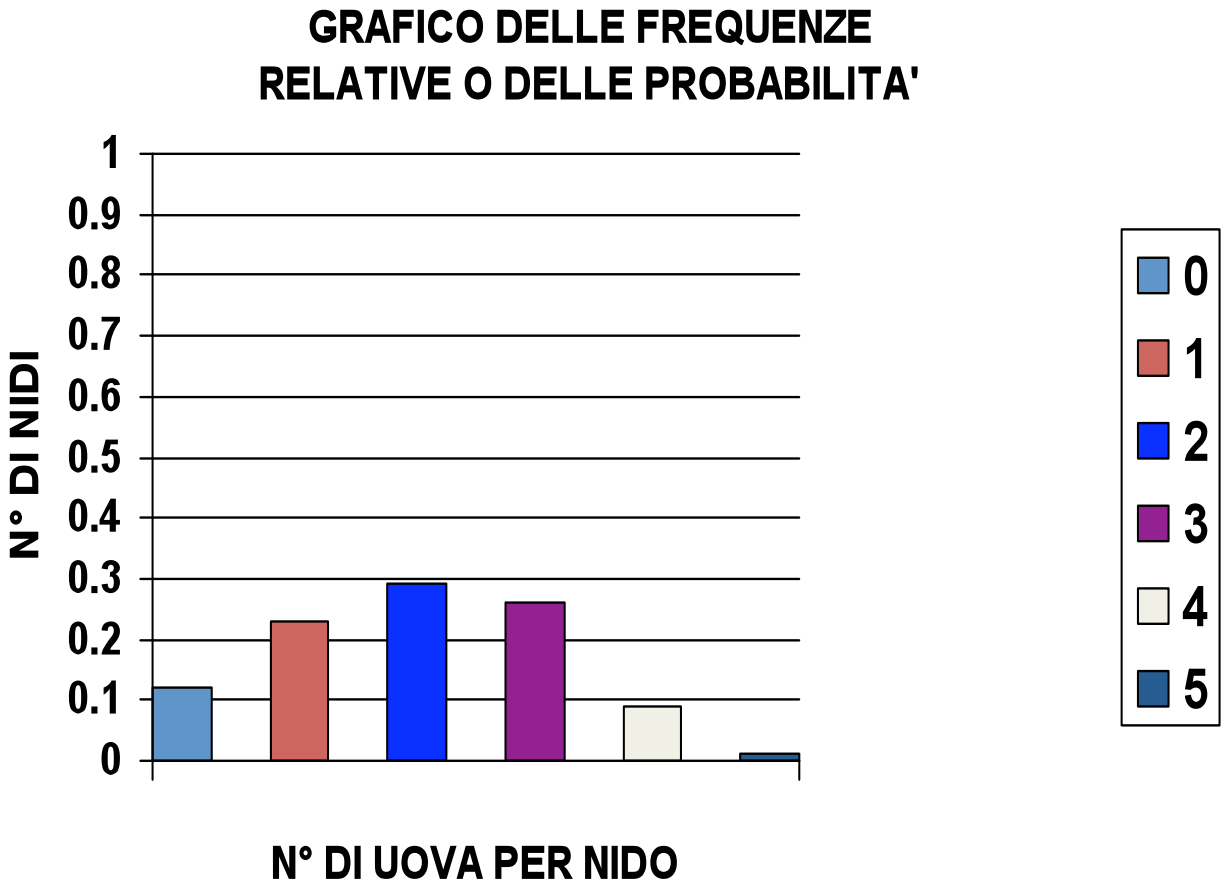
Distribuzioni di frequenze relative (**campione**)

Distribuzioni di probabilità (**popolazione**)

Esempio. Distribuzione del numero di uova per nido di una particolare specie di uccello in una foresta.

N° uova	Frequenze	Frequenze relative	Probabilità
0	**90**	**0.12**	**0.12**
1	**165**	**0.23**	**0.23**
2	**209**	**0.29**	**0.29**
3	**187**	**0.26**	**0.26**
4	**67**	**0.09**	**0.09**
5	**12**	**0.01**	**0.01**
Totale	**730**	**1.00**	**1.00**

grafico a segmenti della distribuzione di frequenze relative o probabilità della v. a. n° di uova per nido



Distribuzioni di probabilità discrete

- Grafico a segmenti delle frequenze relative

Grafico a segmenti delle probabilità: raffigura la distribuzione di probabilità della variabile aleatoria discreta: "n° di uova per nido".

La distribuzione di probabilità è la distribuzione di una **variabile aleatoria** nell'intera popolazione.

In realtà dovremmo ripetere il nostro esperimento un n° infinito di volte per ottenere esattamente questa distribuzione.

E'una distribuzione teorica che ci dice quali frequenze relative (probabilità) per ogni risultato dobbiamo aspettarci su un gran n° di prove.

Variabili aleatorie discrete

- Una variabile aleatoria o casuale X discreta assume diversi valori con probabilità specificate dalla sua funzione di distribuzione
- Le v. a. discrete assumono un numero finito o un'infinità numerabile di valori, Xi =xi i=1,2,...;
- Sono completamente descritte quando sia nota la probabilità con cui si può verificare ciascun valore: $P(Xi = xi) = pi$ con $\Sigma pi = 1$
- Media e Varianza sono indici riassuntivi delle proprietà di tali variabili

Distribuzioni di probabilità

Si osserva un fenomeno aleatorio.

Esempio. Numero di uova per nido di una particolare specie di uccello in una determinata foresta.

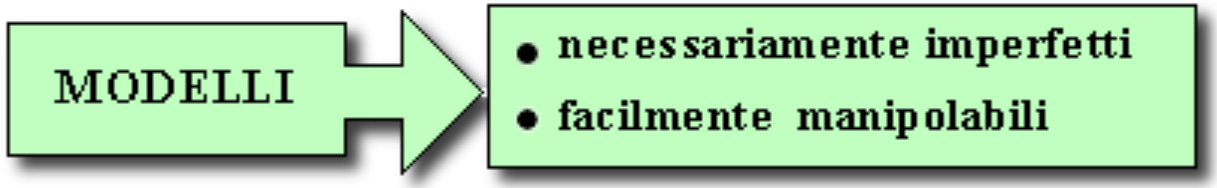
Come rappresentare questa aleatorietà?

Si cerca una **Rappresentazione idealizzata della realtà**

Per il fenomeno studiato si valutano le probabilità e la legge secondo cui tali probabilità evolvono

Si pensa a un Modello probabilistico (aleatorio)

Vedremo come diverse **distribuzioni di probabilità (modelli aleatori) possano essere adattate a diverse situazioni reali** → **MODELLIZZAZIONE**



Distribuzioni di probabilità discrete come modelli

- **modelli stocastici** (stocastico = dovuto al caso, aleatorio), tengono in considerazione le variazioni (casuali e non) delle variabili di *input*, e quindi forniscono risultati in termini di "probabilità".
- È importante sottolineare che ciò che differenzia i modelli **deterministici da quelli stocastici** è che in questi ultimi si tiene conto della **variabilità dei dati di input.**

- **Distribuzioni di probabilità discrete (modelli stocastici), utili nelle applicazioni, che studieremo:**
- **Distribuzione binomiale, Distribuzione di Poisson**

Distribuzioni **discrete** per variabili di tipo **discreto**

La funzione di distribuzione specifica la probabilità che la variabile aleatoria assuma uno specifico valore. Per esempio mi permette direttamente di calcolare:

P(X = 3), ovvero la probabilità che la v. a. X assuma il valore 3 se la variabile può assumere solo valori discreti come 0, 1, 2, 3,... La somma di tutte le probabilità calcolate per ogni valore che può assumere la v. a. deve essere pari a 1.

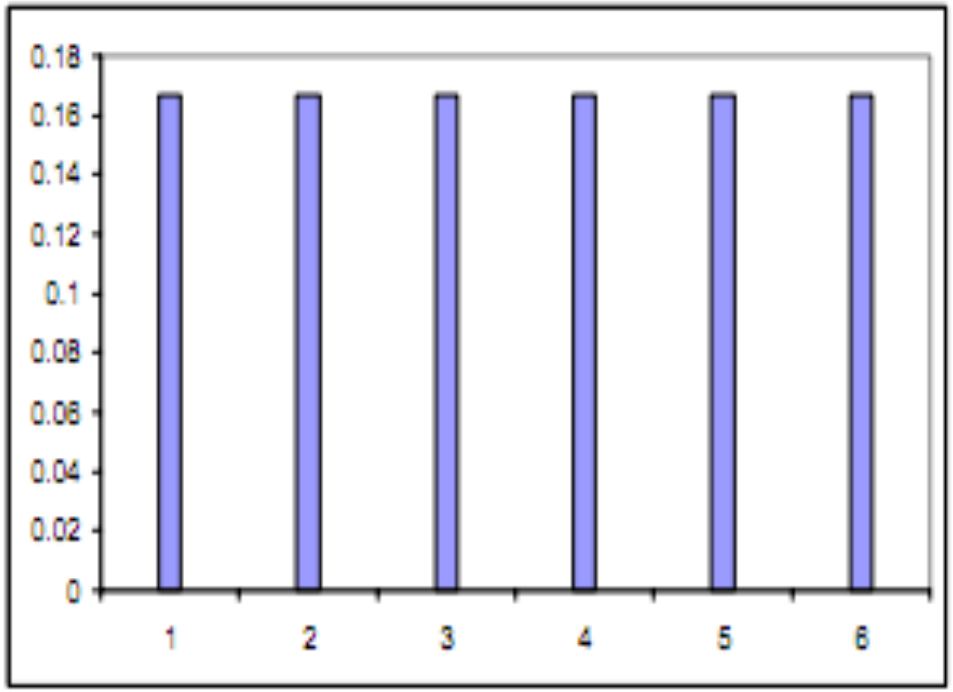
Le **caratteristiche principali** della funzione sono

$f(x) = P(X=x)$

$f(x) \geq 0$ per tutti i valori che può assumere x

$\Sigma_x f(x) = 1$

Distribuzione uniforme



Distribuzione teorica di probabilità dei valori possibili che si possono ottenere nel lancio di un dado equilibrato: è discreta e uniforme

Esempio di variabile aleatoria discreta: quale distribuzione (quale modello) ??

- Siamo interessati alla v. a. discreta X che conta il numero di femmine in una famiglia di 3 figli.

- Ossia “il numero di successi” su 3 prove.

- Per tale variabile quale **distribuzione di probabilità??**

Valori di X: 0 1 2 3

Esempio di modello probabilistico discreto: la distribuzione di probabilità binomiale

- Siamo interessati alla v. a. discreta X che conta il numero di femmine in una famiglia di 3 figli.
- Ossia "il numero di successi" su 3 prove.
- Il modello ideale per rappresentare tale situazione è dato dalla: distribuzione binomiale di parametri n, p, dove n=3 e p=0.5

Valori di X:	0	1	2	3
Probabilità di tali valori:	0.125	0.375	0.375	0.125

(o frequenze relative)

Il modello probabilistico binomiale è un buon modello per rappresentare il fenomeno aleatorio che conta il n° di femmine in una famiglia di 3 figli.

Facciamo uso di una distribuzione di probabilità di cui è nota l'espressione

La distribuzione di probabilità binomiale

- Esempio: Qual'e la probabilità che ci sia solo un maschio in una famiglia di 3 figli?
- Esperimento casuale costituito da **3 prove**, ossia le 3 nascite (eventi o risultati dell'esperimento), ripetute e **indipendenti** in ciascuna delle quali sono **possibili 2 risultati** che indicheremo con 1 e 0.

- Per ogni prova, è nota e costante **la probabilità di successo**. Nell'esempio:

1	0
$p \approx 0.5$	$1\text{-}p \approx 0.5$

Distribuzione binomiale di **parametri n e p**

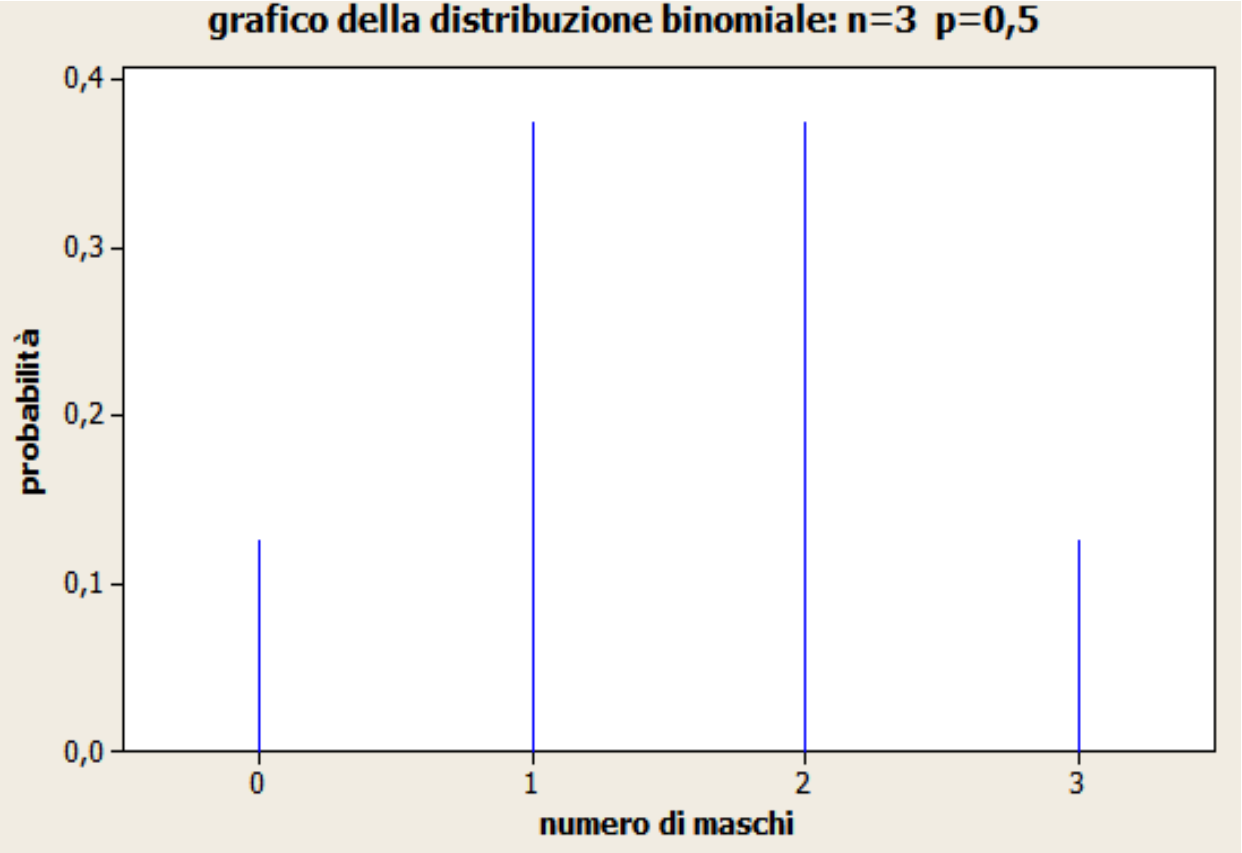
- Qual'e la probabilità che in una famiglia di 3 figli (3 prove) ci sia solo un maschio (n° successi X =1) ?
- Qual è la probabilità che su n prove il n° di successi X sia uguale a k?

k= 0,1,…,n

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

Media = ? Varianza = ?

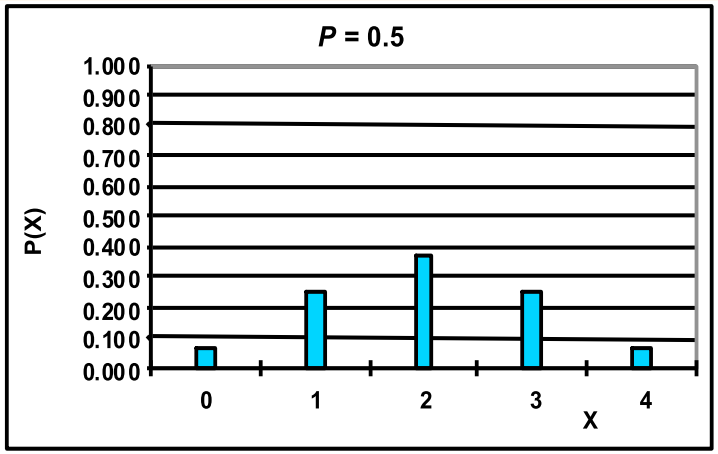
Grafico della distribuzione di probabilità binomiale di **parametri n = 3 e p = 0.5** (Es. precedente)

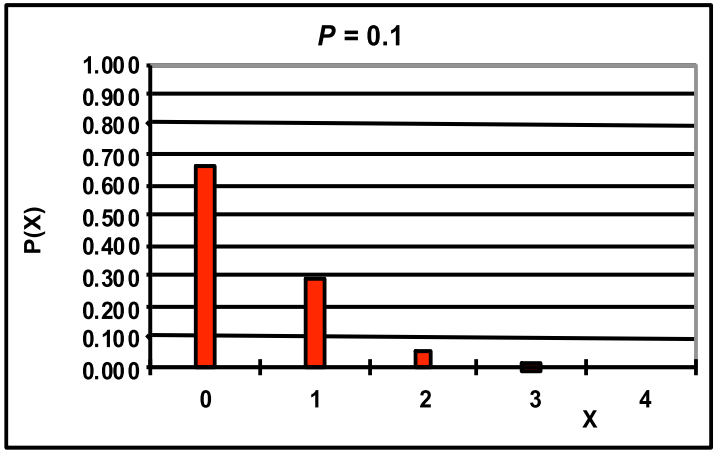


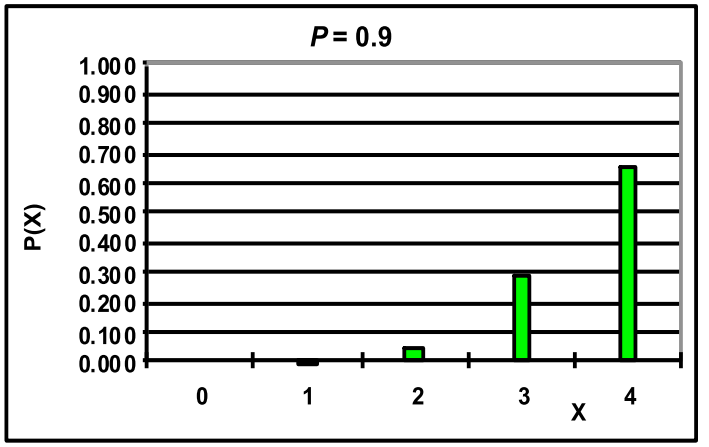
I parametri individuano in modo univoco la distribuzione binomiale

Graphs of Selected Binomial Distributions

n = 4 PROBABILITY			
X	0.1	0.5	0.9
0	0.656	0.063	0.000
1	0.292	0.250	0.004
2	0.049	0.375	0.049
3	0.004	0.250	0.292
4	0.000	0.063	0.656







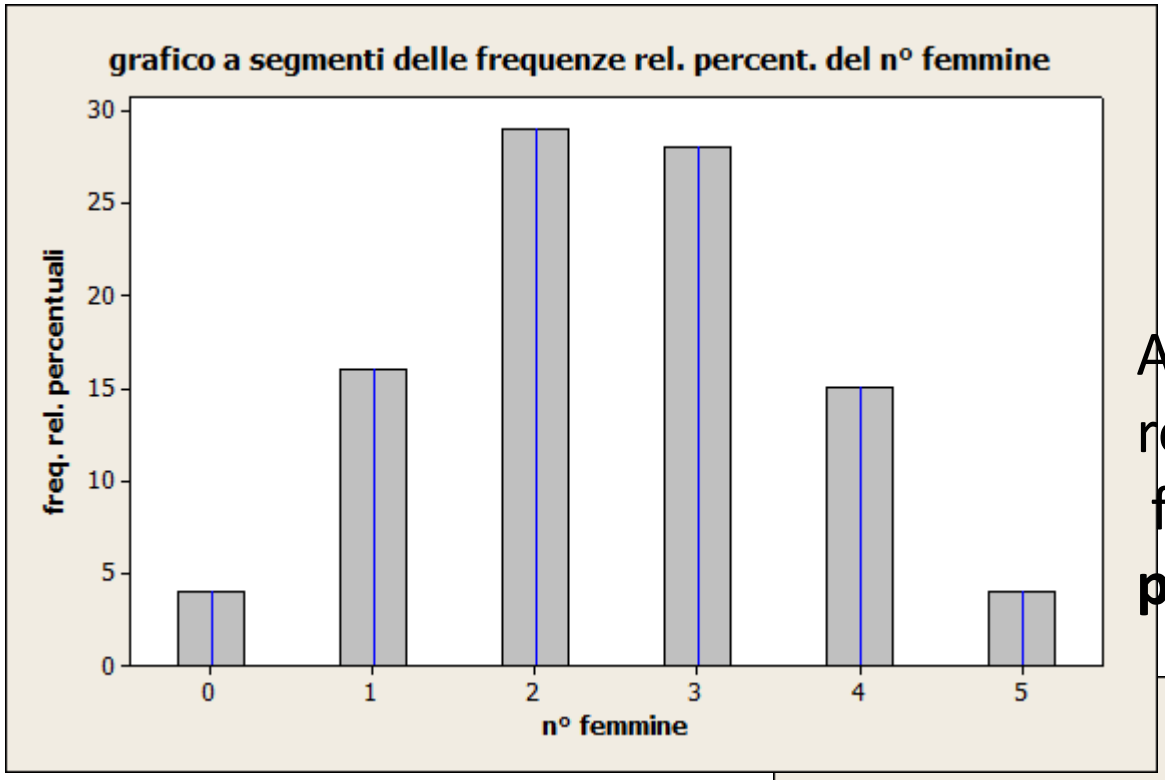
Modellizzazione di un fenomeno osservato: la distr. binomiale

Sono stati esaminati 480 nidi di tordo dove erano sopravissuti 5 uccellini. Si osserva il n° di femmine sopravissute in ogni nido.

femmine	maschi	probabilità binomiale	frequenze osservate	frequenze relative
5	0	0.038	21 nidi con 5 femmine	0.043
4	1	0.159	76	0.158
3	2	0.310	138	0.287
2	3	0.310	142	0.295
1	4	0.159	80	0.166
0	5	0.038	23	0.047

Modellizzazione di un fenomeno osservato: la distr. binomiale

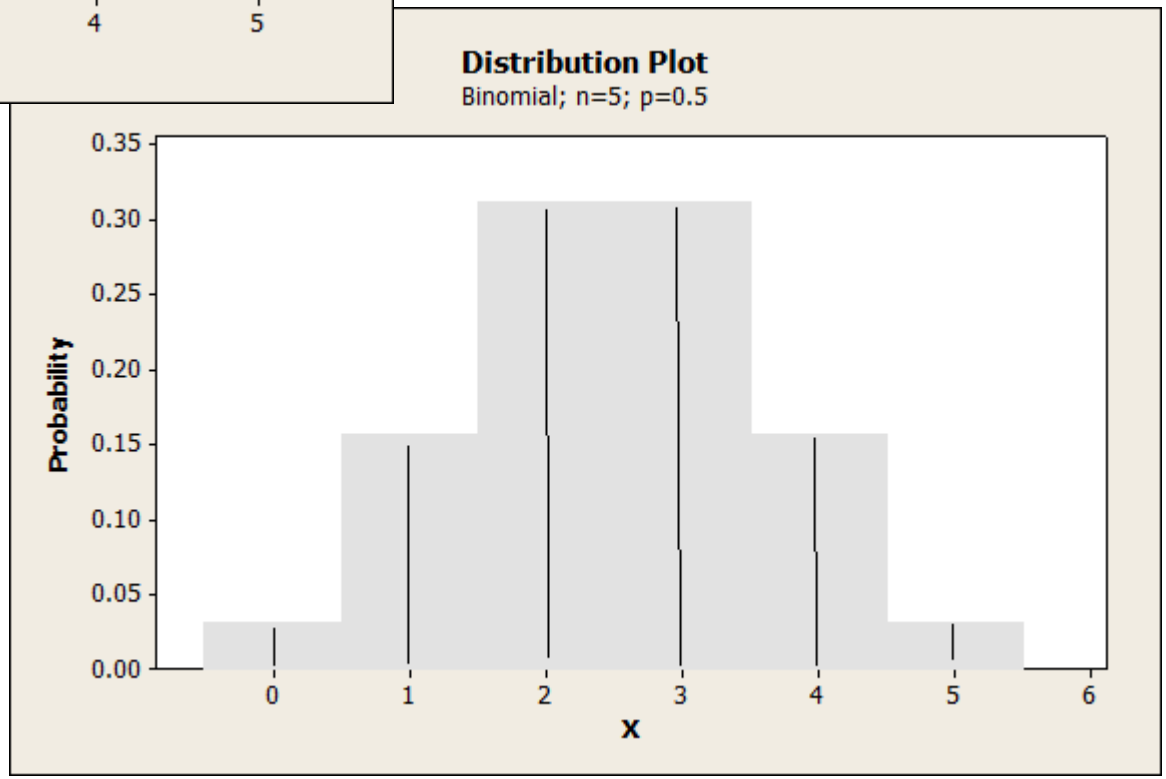
- **Attenzione**: In ogni nido ci sono un numero di prove fisso (n=5), con una certa probabilità di successo e insuccesso in ogni prova.
- Se in ogni nido la prob. di essere femmina fosse uguale e indipendente per ogni piccolo, quale distrib. mi aspetterei per il numero di femmine (v. a. X)?
- Si cerca di modellizzare il fenomeno con una distribuzione **binomiale con p=0.5** (il valore p=0.5 è ipotizzato perché non noto) e **n=5.**
- Il parametro **n** della binomiale è noto (5), **p** è la probabilità di successo (femmina) nella singola prova (uovo) in 5 prove (un nido).



Esempio di modellizzazione

A confronto il grafico freq. rel. perc. **osservate** per il fenomeno e il grafico delle **probab. binomiali <u>n=5; p= 0.5</u>**

Ottimo modello



Distribuzione binomiale: esempi

Esempi

a. Numero di ibridi su n osservazioni;

b. Numero di studenti su n che superano l'esame con un voto maggiore di 28.

Quale v. a.? Quali parametri?

- Modellizziamo fenomeni che non sono prevedibili e che hanno caratteristiche aleatorie.

Facciamo uso di una distribuzione di prob. di cui è nota l'espressione.

Distribuzione binomiale: esempio

- Nella pianta di tabacco *Nicotiana* c'è un allele recessivo di un gene coinvolto nella produzione della clorofilla che, in omozigosi, non produce clorofilla e quindi si avranno foglie bianche.

- In campioni casuali di dimensione n= 13, il n° X di piantine con foglie bianche sarà modellizzato da una binomiale con p=0,25.
- Perché?
- Qual è la prob. che X=0?

ESEMPIO

Una certa malattia ha un'evoluzione per cui non si conoscono terapie, tuttavia tra le persone colpite il 40% guarisce spontaneamente nell'arco di due mesi.

Non conoscendo particolarità della malattia, la possibilità di guarigione nell'arco di due mesi viene vista come puramente casuale.

Con quale probabilità tra 6 persone colpite dalla malattia 2 guariranno spontaneamente nell'arco di due mesi?

Quali parametri?

Qual è il numero medio di guarigioni spontanee? Quanto vale la varianza?

Con quale probabilità nessuno guarirà spontaneamente?

ESEMPIO

Quattro bambini vengono vaccinati contro il morbillo.
Il vaccino attecchisce con probabilità 0.8, garantendo l'immunità del bambino alla malattia.

Quale v. a.? Quali parametri?
Con quale probabilità tutti i bambini risultano immunizzati?
Se 100 bambini vengono vaccinati, qual è il numero medio di bambini immunizzati?
Quanto vale la varianza di tale numero?

Modellizzazione

In sostanza, la creazione di un modello inizia con lo studio del fenomeno nella realtà; le osservazioni derivanti dallo studio vengono interpretate per cogliere gli aspetti più importanti del fenomeno. Poi, si costruisce il modello, lo si fa "funzionare" e si controlla se e quanto i risultati ottenuti corrispondono con la realtà. Poi il modello viene riconsiderato e modificato per renderlo più efficiente, e così di seguito.

La distribuzione di probabilità di Poisson di parametro λ

ESEMPIO Supponiamo di ispezionare un campione di **20 nidi**

Supponiamo di osservare in questo campione per ogni nido il numero di piccoli sopravvissuti. Il **numero totale** di piccoli sopravvissuti risulta pari a **38**

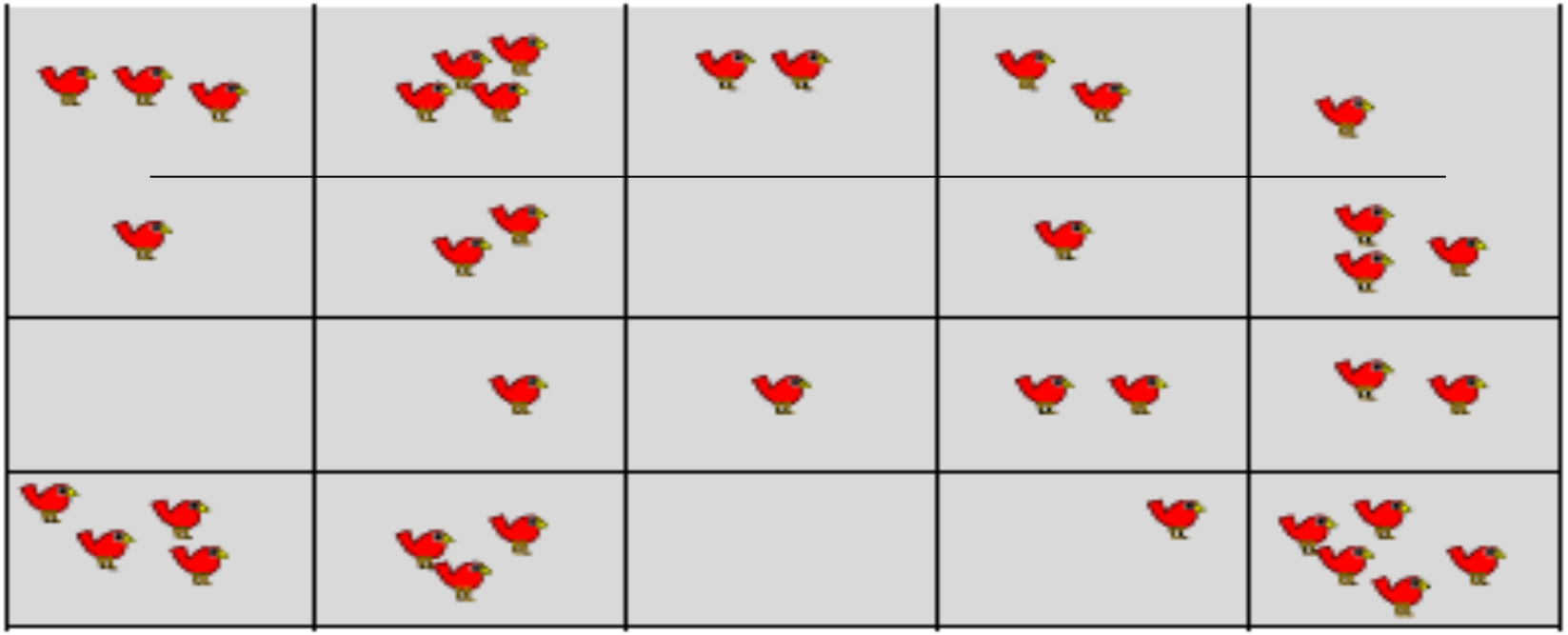
Il numero medio λ di piccoli sopravvissuti per nido è pari a 38/20.

λ è il parametro che individua ciascuna distrib. di Poisson

Possiamo pensare ai nidi come ad una griglia 4x5 con 20 caselle (ogni casella è un nido), ciascuna delle quali può contenere 0, 1, 2, …numeri di pulcini sopravvissuti

Qual è la variabile di interesse? E' noto il n° massimo che può assumere tale variabile?

La distribuzione di Poisson



x	n_i	n_i x
0	3	0
1	6	6
2	5	10
3	3	9
4	2	8
5	1	5
TOTALI	**20**	**38**

Come è distribuito il n° di piccoli sopravvissuti se l'unico effetto che agisce su questo numero è il caso?

La distribuzione di Poisson

La **differenza** rispetto alla distribuzione binomiale è chiara: il numero massimo che può assumere la variabile (numero di piccoli sopravvissuti per nido) non è noto.

Perciò ci si chiede come è distribuito il numero di piccoli sopravvissuti?

Ossia, quanti nidi mi aspetto con 0,1,2,….. piccoli se l'unico effetto che agisce su questo numero è il **caso**?

Non posso applicare la binomiale, appunto perchè la situazione è diversa

Distribuzione di probabilità di Poisson di parametro λ

$$P(X = k \quad successi) = e^{-\lambda} \frac{\lambda^k}{k!}$$

X =0, 1,2,….

La v. a. X può assumere un n° infinito di valori interi, e λ è il numero medio di successi indipendenti nel tempo o nello spazio, ossia numero totale di eventi/numero totale di intervalli (o blocchi) spaziali, volumetrici, o temporali.

Si noti che, quando si modellizza con una distribuzione di prob. un fenomeno osservato, i **parametri non sono noti** e vanno **stimati** sulla base del campione.

Distribuzione di probabilità di Poisson

- La distribuzione di Poisson descrive il numero di successi in intervalli (o blocchi) spaziali, volumetrici, o temporali quando
- i successi avvengono indipendentemente l'uno dall'altro
- i successi hanno la stessa probabilità di verificarsi in ogni punto dello spazio, di volume, o di tempo.
- Possiamo quindi usare questa distribuzione teorica di probabilità come modello per predire se le osservazioni che abbiamo fatto (nel tempo, nello spazio) sono compatibili con il semplice effetto del caso

[a differenza della binomiale, il numero di prove non è noto!

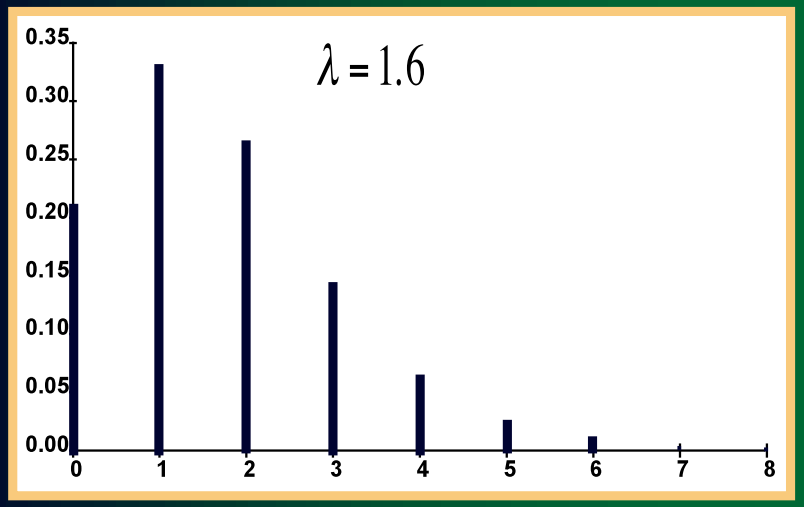
La distribuzione di Poisson

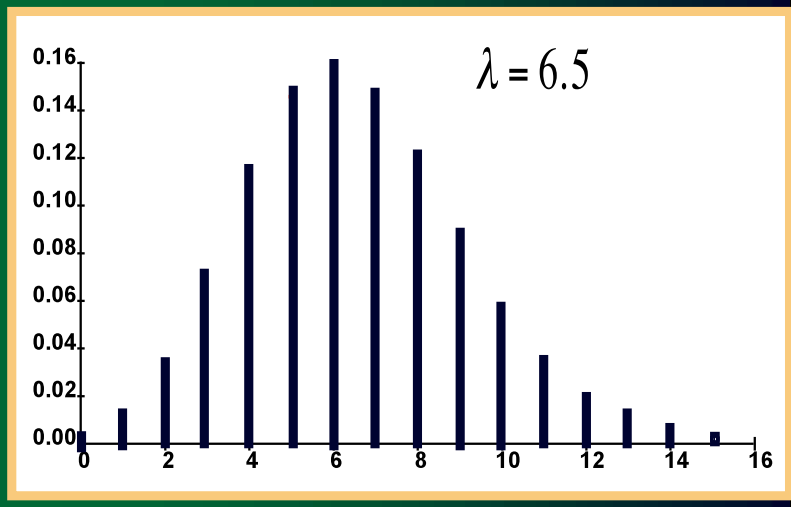
Altre esempi di variabili che potrebbero seguire, se interviene **solo la casualità**, la distribuzione di Poisson:

- numero di semi di una pianta infestante in un certo volume di terriccio in vendita
- numero di mutazioni in un certo intervallo di tempo
- numero di casi di influenza in un paese in una settimana
-numero di incidenti stradali mortali in un mese in una città
- numero di pezzi difettosi in una giornata di produzione
-numero di cetacei presenti in un tratto di mare

In tutti questi casi si può immaginare che nell'area, volume, tempo analizzati ci sia la possibilità teorica di osservare un numero elevatissimo di eventi tipo "presenza", ma che quelli realmente osservati siano invece "rari".

Poisson Distribution: Graphs





Al crescere del parametro λ la distribuzione diventa sempre più simmetrica e ha il massimo nel punto λ.

ESEMPIO

- n = 20 pazienti sono esaminati per vedere se un nuovo farmaco induce una prob. di ricovero del 40%

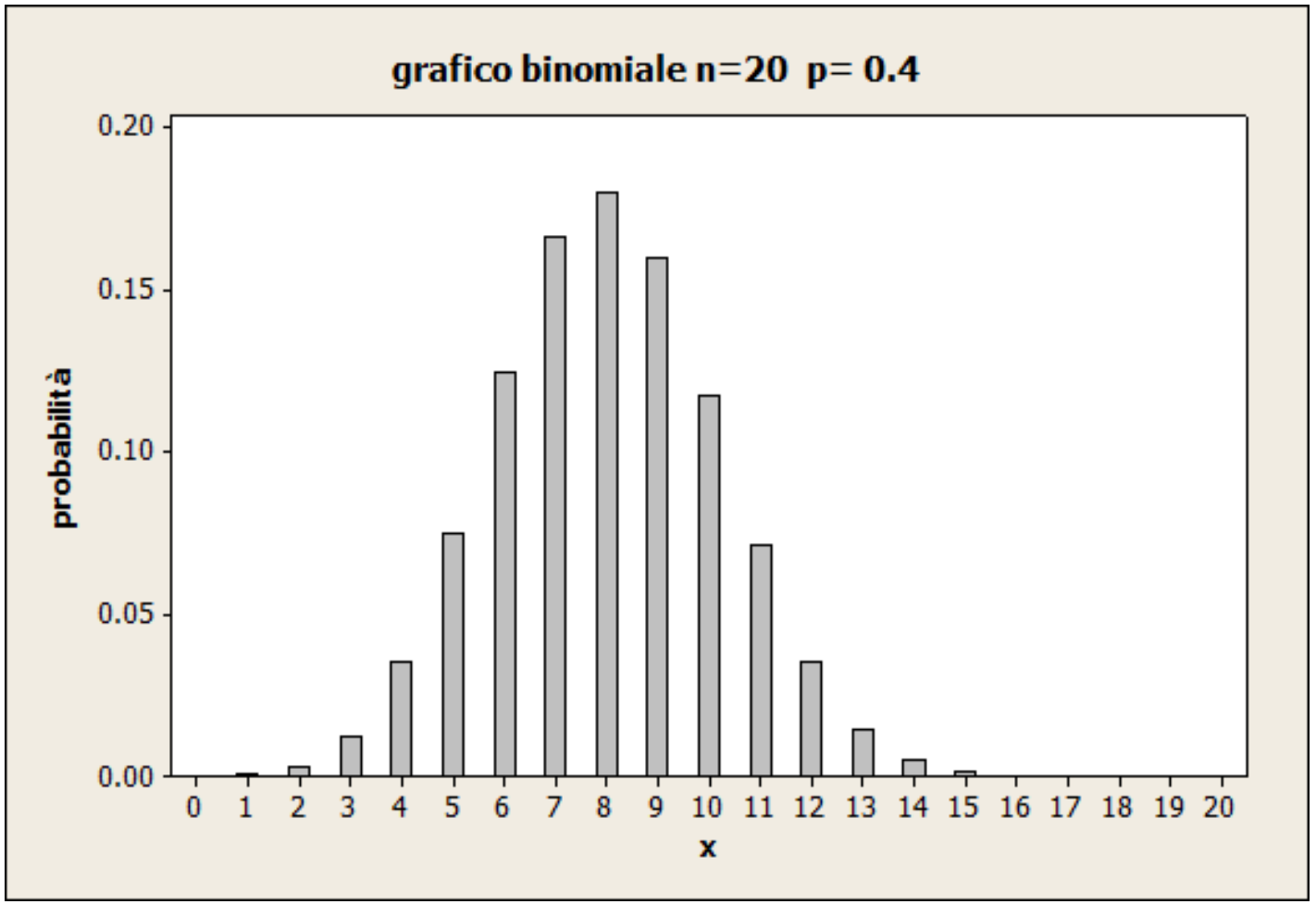
- Quale v. a.? Come modellizzare il fenomeno ?

  (specificare sempre i parametri della distribuzione)

= media di eventi per blocco, ossia numero totale di eventi/numero totale blocchi

-

Grafico relativo all'esempio precedente



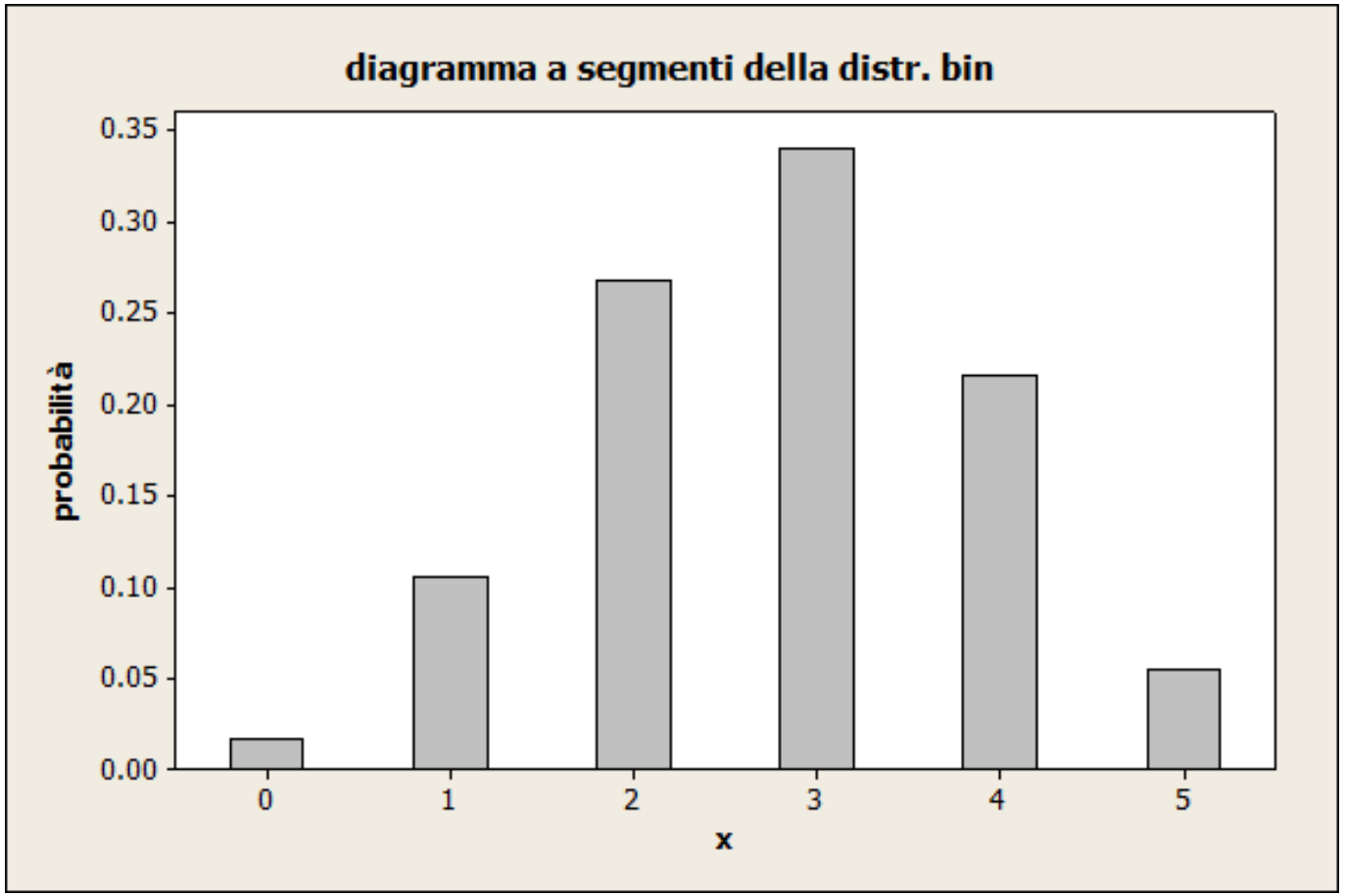
x	p(x)
0	0.000037
1	0.000487
2	0.003087
3	0.012350
4	0.034991
5	0.074647
6	0.124412
7	0.165882
8	0.179706
9	0.159738
10	0.117142
11	0.070995
12	0.035497
13	0.014563
14	0.004854
15	0.001294
16	0.000270
17	0.000042
18	0.000005
19	0.000000
20	0.000000

Esempio

In una pop. di soggetti affetti da tumore cerebrale il 56% dei malati non presenta crisi epilettiche come primo sintomo.
Si devono esaminare 5 nuovi soggetti e ci si chiede quale sia la prob. che 3 dei 5 non presentino una crisi epilettica come primo sintomo.

Quale v. a.? Come modellizzare il problema? Quali parametri?

Grafico relativo all'esempio precedente



Parametri n=5
p= 0.56

X	p(x)
0	0.016492
1	0.104947
2	0.267137
3	0.339993
4	0.216359
5	0.055073

E' molto importante che x e p si riferiscano all'esito identificato come "successo"

ESEMPIO

- Un professore di biologia programma di assegnare un quiz a sorpresa che consiste in 4 domande a risposta multipla, ognuna delle quali ha 5 risposte possibili (a,b,c,d,e) una sola delle quali è corretta. Se uno studente impreparato risponde in modo casuale, qual è la prob. che risponda in modo corretto a 3 delle 4 domande?

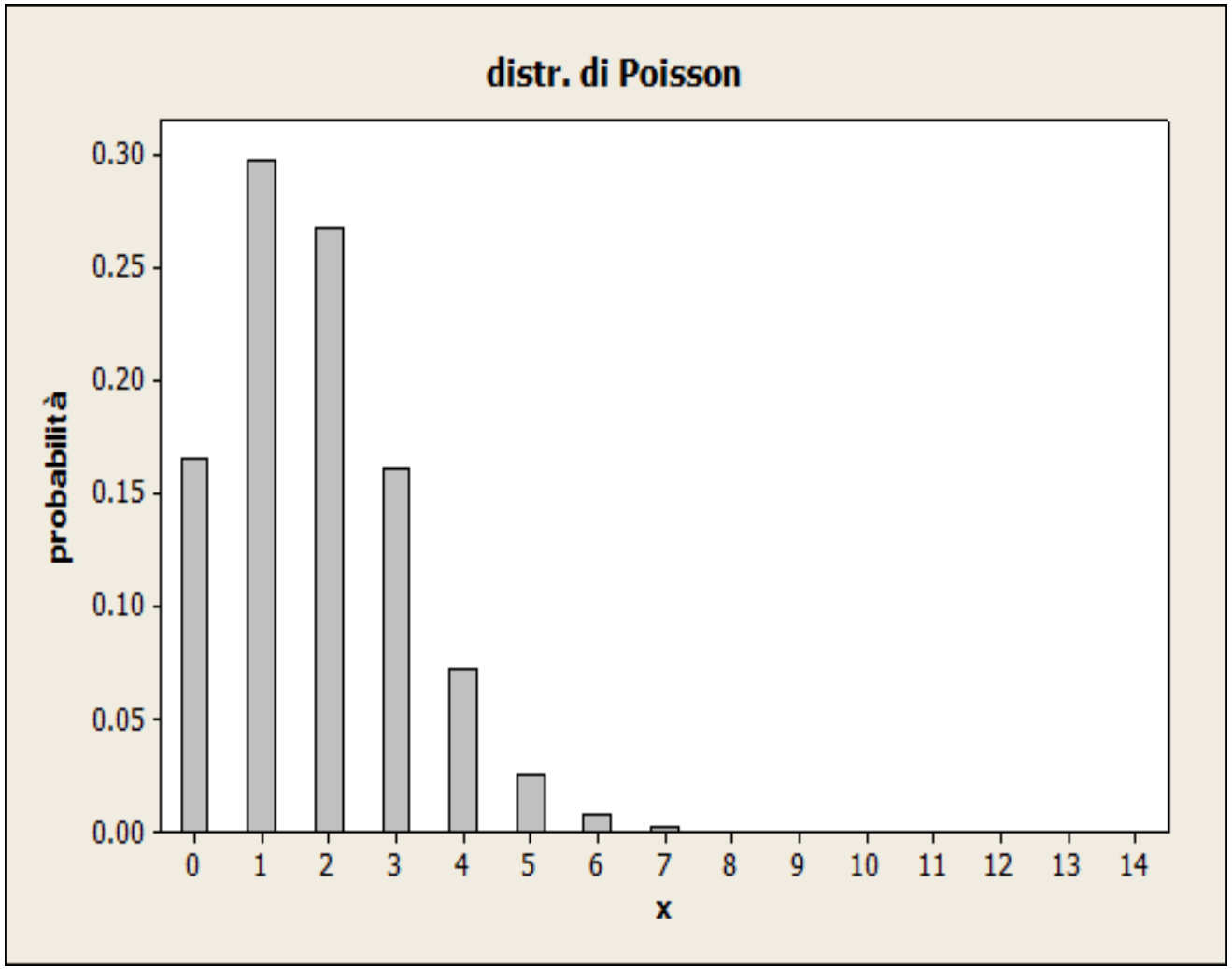
Quale v.a.? Quale modello? Quali parametri?

Esempio

In un ospedale le nascite avvengono casualmente e ci sono mediamente 1.8 nascite all'ora.
Qual è la prob di osservare 4 nascite fra le 21 e le 22 di un qualsiasi giorno?

Quale v.a.? Quale modello? Quali parametri?

Grafico relativo all'esempio precedente



parametro $\lambda = 1.8$

x	p(x)
0	0.165299
1	0.297538
2	0.267784
3	0.160671
4	0.072302
5	0.026029
6	0.007809
7	0.002008
8	0.000452
9	0.000090

Esempi

1) Per analizzare le tracce delle bombe V-I della seconda guerra mondiale, la zona meridionale di Londra è stata suddivisa in 576 regioni , ognuna delle quali di area 0.25 Km^2. Un totale di 535 bombe ha colpito l'area delle 576 regioni.

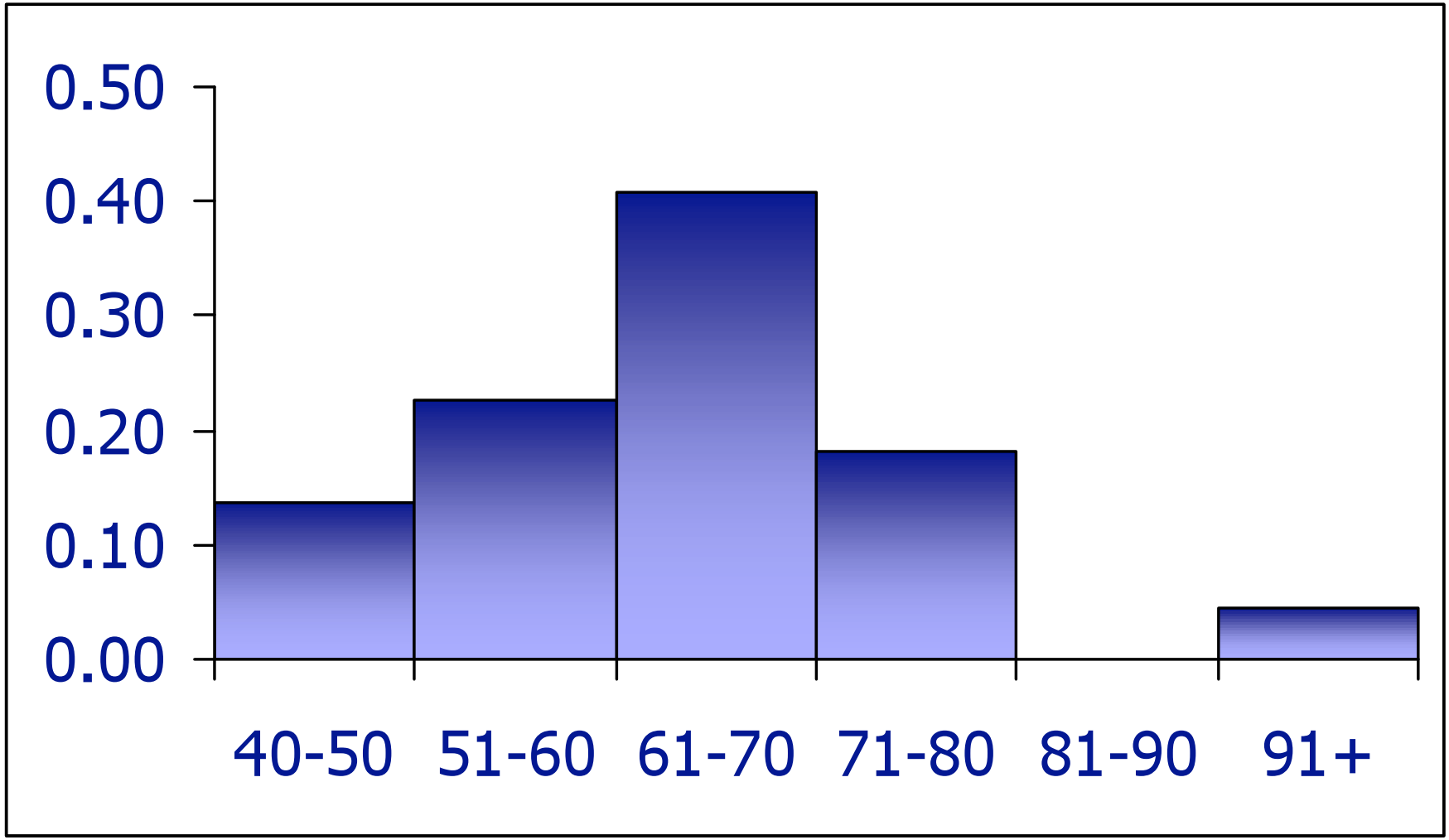
   Qual è la prob che, una regione scelta a caso, sia stata colpita 2 volte?

   Quale v.a.? Quale modello? Quale parametro?

2) Numero di auto in attesa al semaforo: è distribuita secondo Poisson?

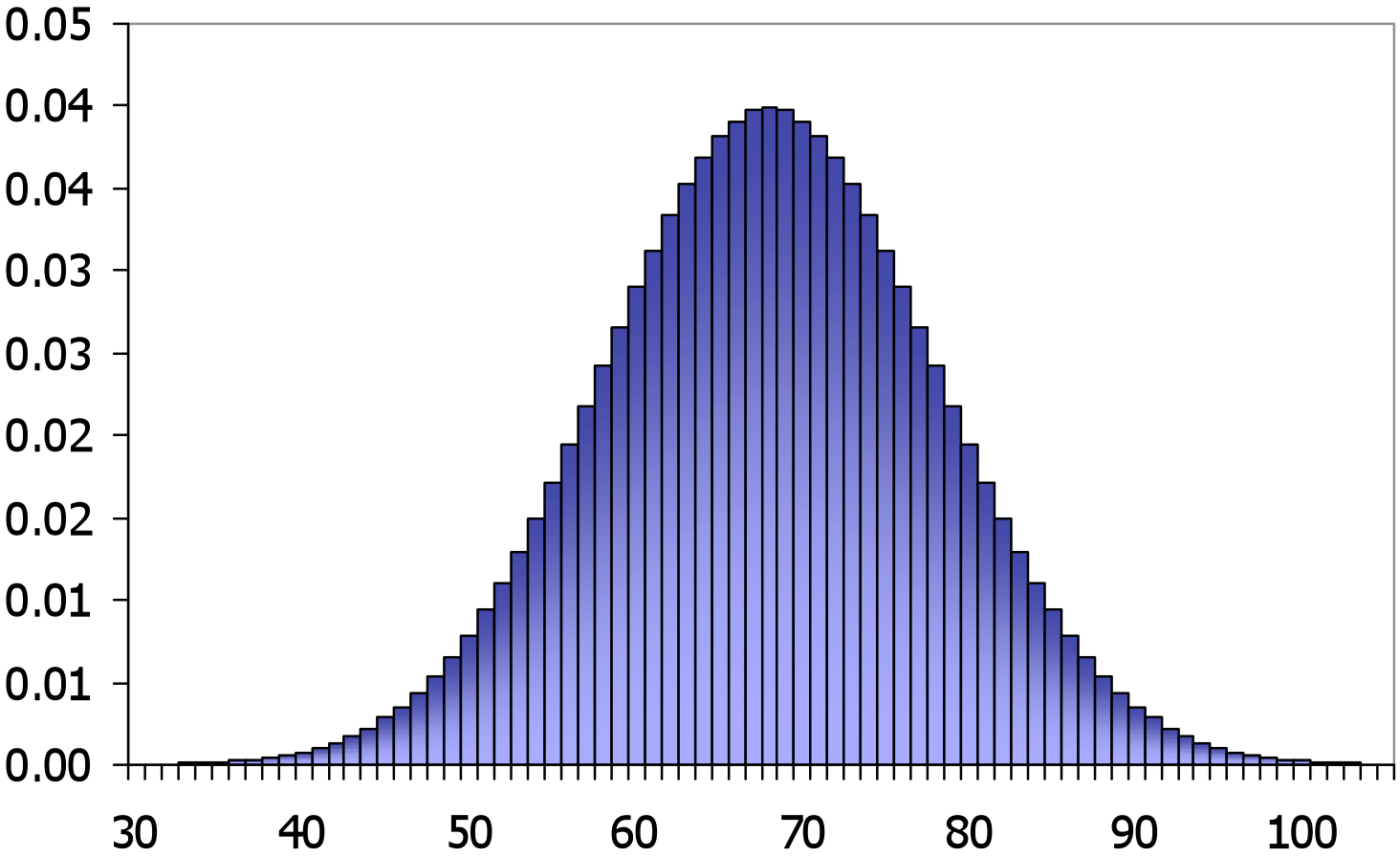
Variabili aleatorie continue: L'ISTOGRAMMA

Peso (kg) di 150 studenti tra i 12-18 anni



Variabili aleatorie continue: L'ISTOGRAMMA

Aumentiamo il n° delle osservazioni: Peso (g) di 150.000 studenti tra i 12-18 anni



Distribuzione di probabilità normale: Peso di studenti tra i 12-18 anni

Poiché le misure di peso sono su di una scala continua, è possibile aumentare il grado di precisione delle misurazioni in modo che le classi di frequenza siano a intervalli di 0.001 kg (1 g) invece che di 1 kg.

Se si considera un numero di osservazioni molto grande a un grado di precisione infinitamente elevato, i gradini dell'istogramma si trasformano in una curva continua simile a quella della distribuzione normale (con un andamento a campana).

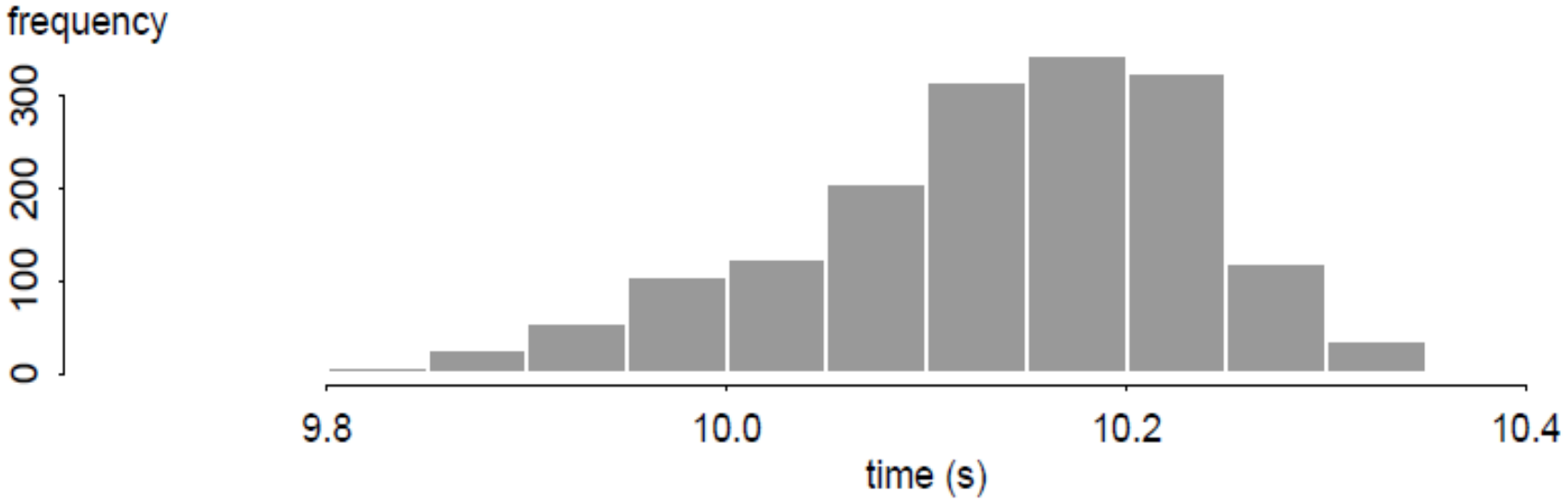
Distribuzione normale

- Se la scala del grafico è tale che **l'area totale** sotto la curva è pari a **1**, allora l'area rappresenta tutte le osservazioni e la curva è una **curva di densità**.
- L'area che sta sotto la curva e sopra un determinato intervallo di valori rappresenta la **proporzione (frequenza relativa)** di tutte le osservazioni che cadono in quell'intervallo.
- Si può, anche, dire che tale area rappresenta la **probabilità** che un individuo scelto a caso appartenga a quell'intervallo.

- Nell'es. dei pesi la distrib. Normale rappresenta un **buon modello** per la popolazione da cui proviene il campione osservato e può aiutare nello studio del fenomeno in questione.

Non tutte le distr. sono simmetriche. The histogram below shows the best 10 sprint times from the 168 all-time top male 100m sprinters. There are 1680 times in total, representing the top 10 times up to 2002 from each of the 168 sprinters. Ora il record è 9.58?

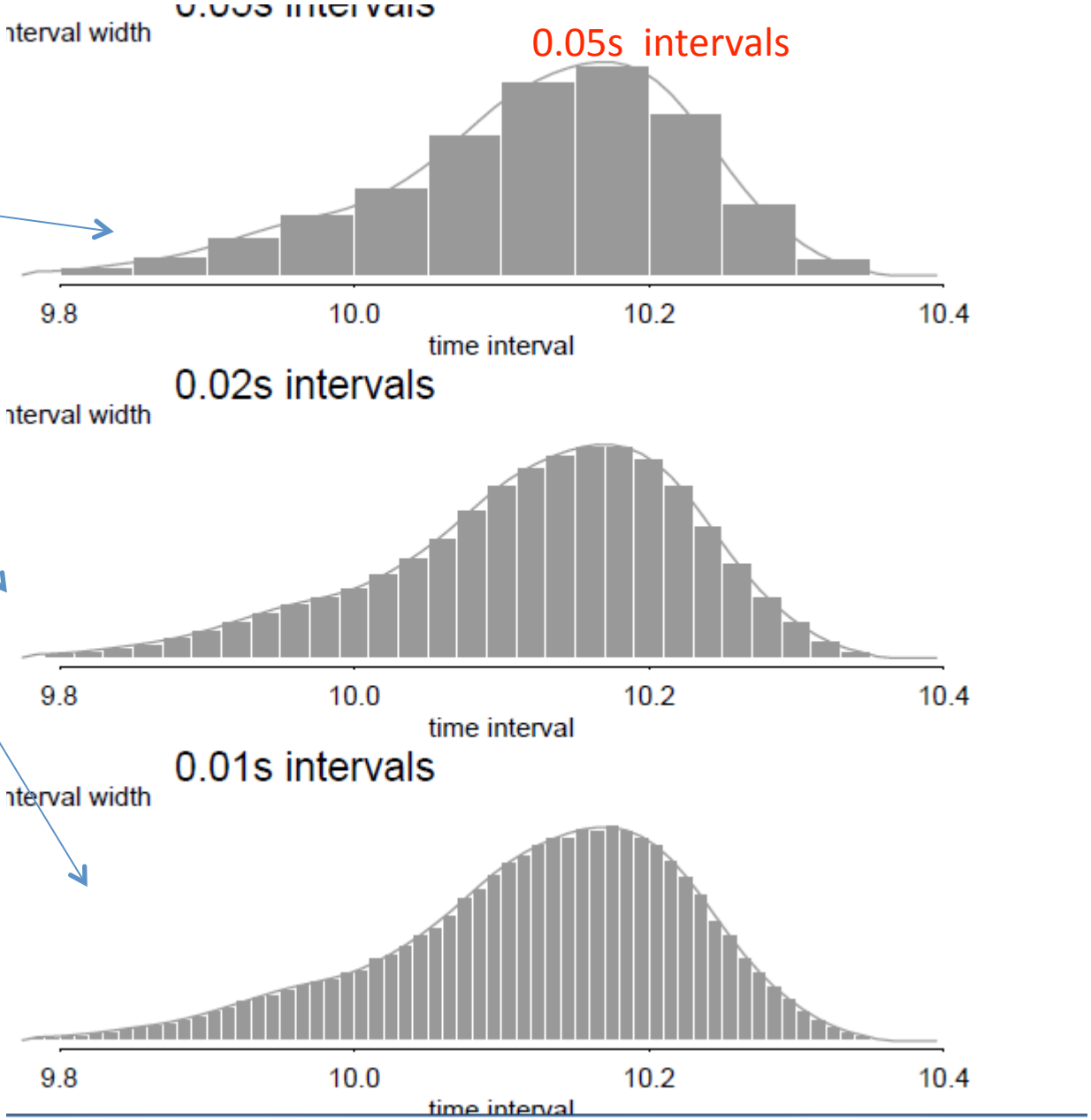
```
Min. 1st Qu. Median  Mean 3rd Qu.  Max.
9.78 10.08   10.15  10.14 10.21   10.41
```



the *most probable times are close to 10.2 seconds;*

- the distribution of times has a long left tail (left skew);
- times below 10.0s and above 10.3 seconds have low frequency (e freq relativa percentuale)

Densità


0.05s intervals
9.8
10.0
10.2
10.4
time interval
0.02s intervals
9.8
10.0
10.2
10.4
time interval
0.01s intervals
9.8
10.0
10.2
10.4

Istogramma e distribuzione normale: un altro esempio

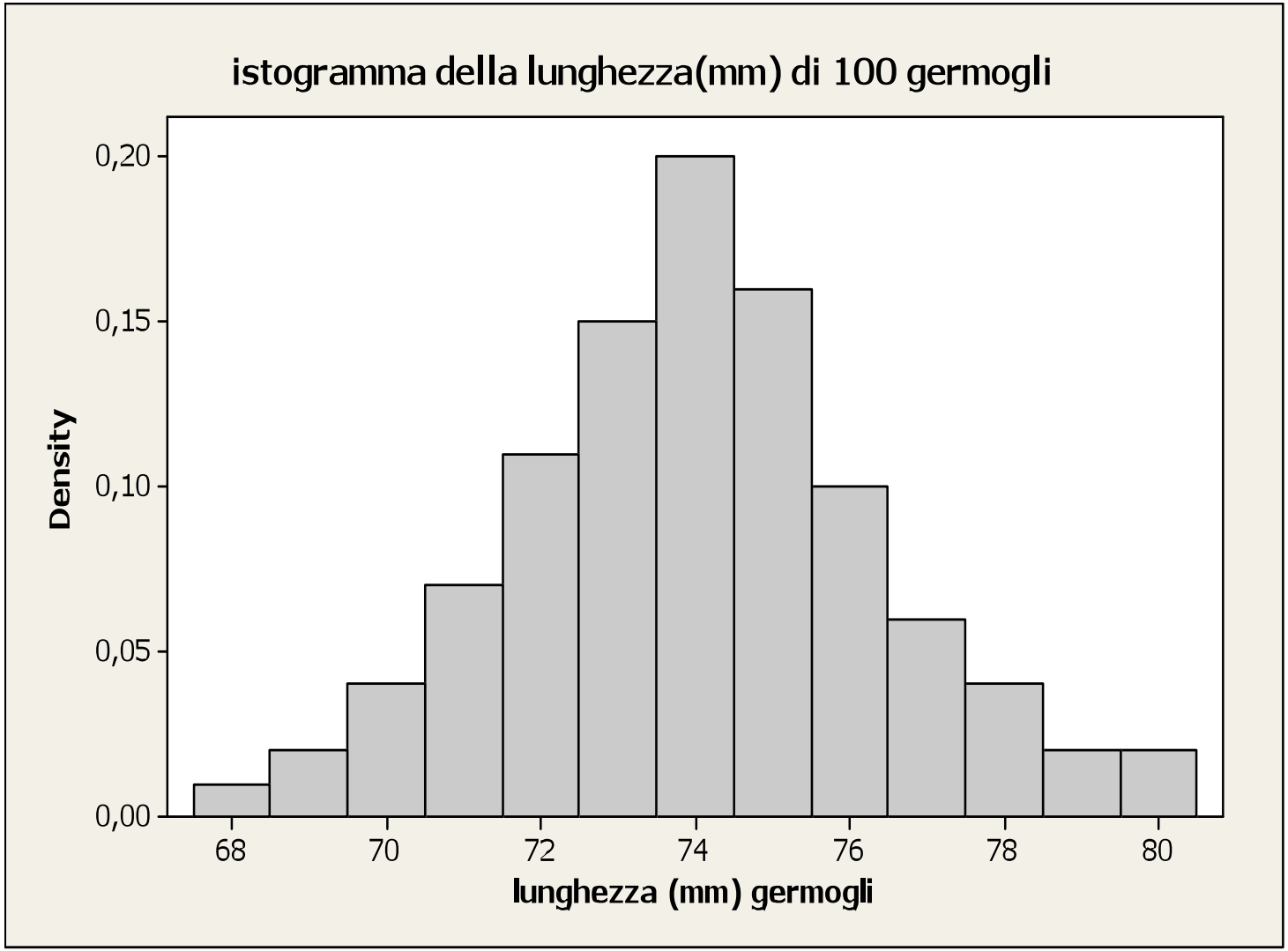
- **Esempio 13**. Nella tabella che segue sono riportate le lunghezze (al mm più vicino) di 100 germogli nati da semi piantati allo stesso tempo.
- Ad esempio, tutti i germogli di lunghezza tra 73.5mm e 74.5mm sono inseriti nella classe 74mm.
- Nella tabella sono, anche, riportate la distribuzione delle frequenze, delle frequenze relative e delle frequenze cumulate dei valori della variabile "lunghezza dei germogli".

Tabella delle frequenze

Intervallo classe	Lunghezza (mm)	Frequenza	Frequenza relativa	Frequenza rel. cumulata
67.5-68.5	68	1	0.01	0.01
68.5-69.5	69	2	0.02	0.03
69.5-70.5	70	4	0.04	0.07
70.5-71.5	71	7	0.07	0.14
71.5-72.5	72	11	0.11	0.25
72.5-73.5	73	15	0.15	0.40
73.5-74.5	74	20	0.20	0.60
74.5-75.5	75	16	0.16	0.76
75.5-76.5	76	10	0.10	0.86
76.5-77.5	77	6	0.06	0.92
77.5-78.5	78	4	0.04	0.96
78.5-79.5	79	2	0.02	0.98
79.5-80.5	80	2	0.02	1.00

Le frequenze relative cumulate

- La frequenza relativa cumulata per una data classe è ottenuta come somma della corrispondente frequenza relativa e di tutte quelle relative alle classi precedenti.
- In modo analogo si definisce la frequenza cumulata.
- Spesso si considera la frequenza percentuale cumulata pari alla frequenza relativa cumulata moltiplicata per 100.



E' unimodale (classe modale 73.5-74.5 mm)

$\overline{x}$ = **74.02mm** e deviazione standard **s = 2.39mm**.

L'esempio dei germogli

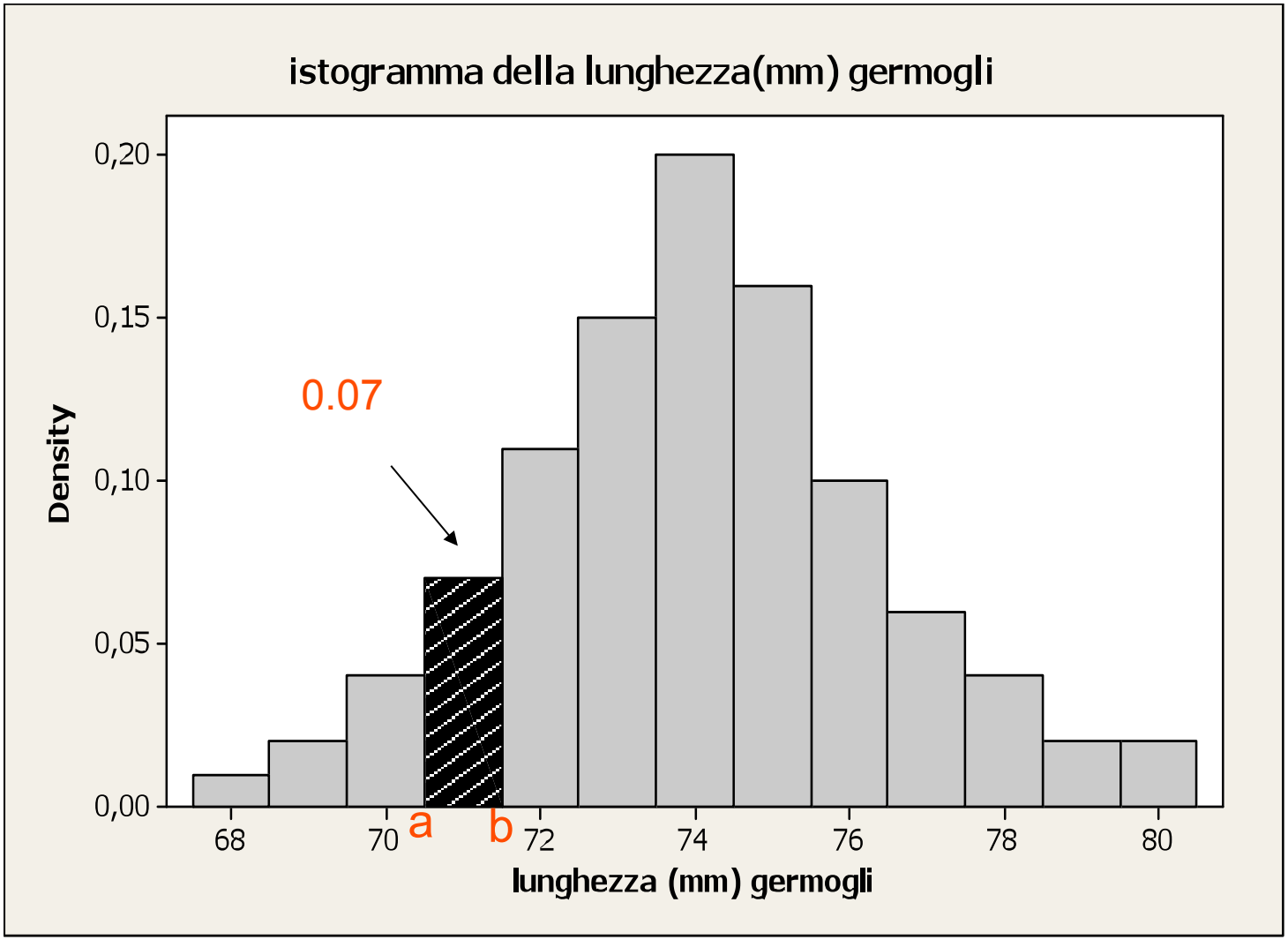
Sia X la lunghezza (mm) di un germoglio.

Si è visto che le lunghezze del campione di 100 germogli si distribuiscono approssimativamente come una distrib. simmetrica con media $\bar{x}$ = 74.02mm e deviazione standard s = 2.39mm →

Possiamo pensare che la **distr. Normale** rappresenti un buon modello per la popolazione da cui proviene il campione

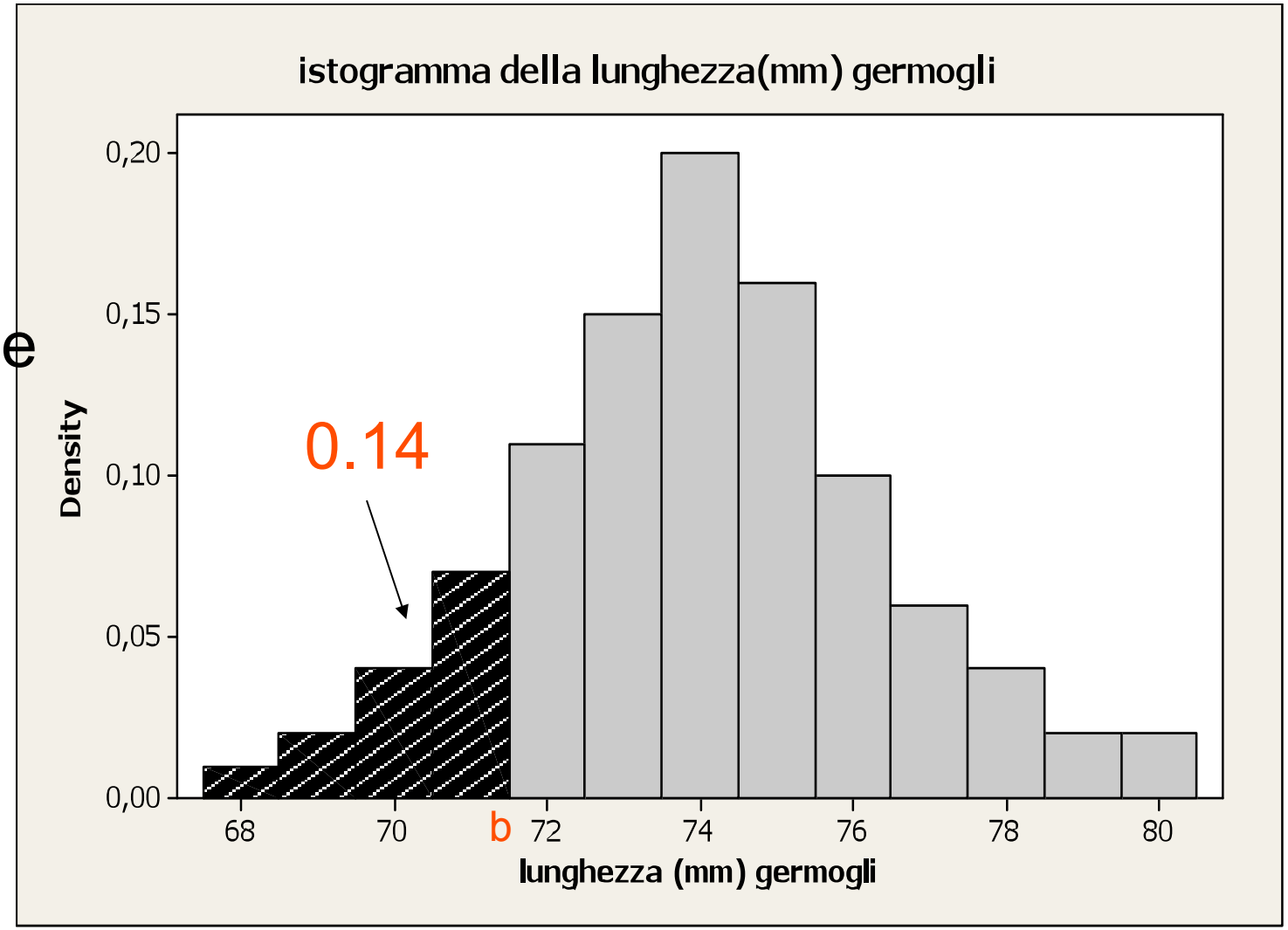
$$X \sim N(74.02, 2.39)$$

$\bar{x}$ e s forniscono le stime di μ e σ che sono incogniti per l'intera popolazione

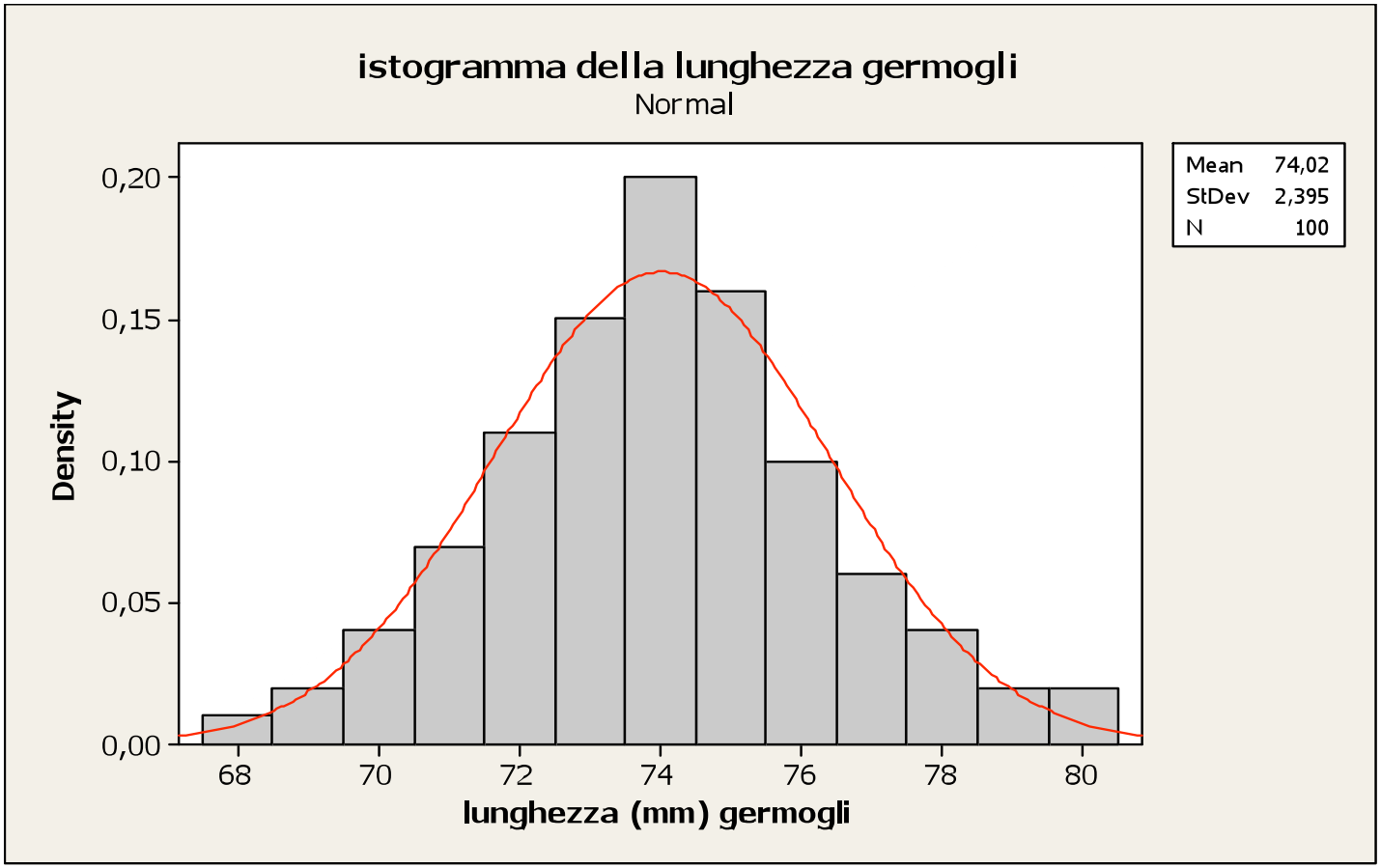


L'area <u>della barra</u> sull'intervallo a=70.5 e b=71.5 è pari a 0.07. Corrisponde al 7% di tutte le osservazioni. Ossia, nel campione di 100 germogli, il 7% ha lunghezza tra 70.5 e 71.5.

l'osservazione b a quale percentile corrisponde?



L'area tratteggiata in rosso rappresenta la frequenza relativa cumulata che fino al punto b è pari a 0.14 (14%).

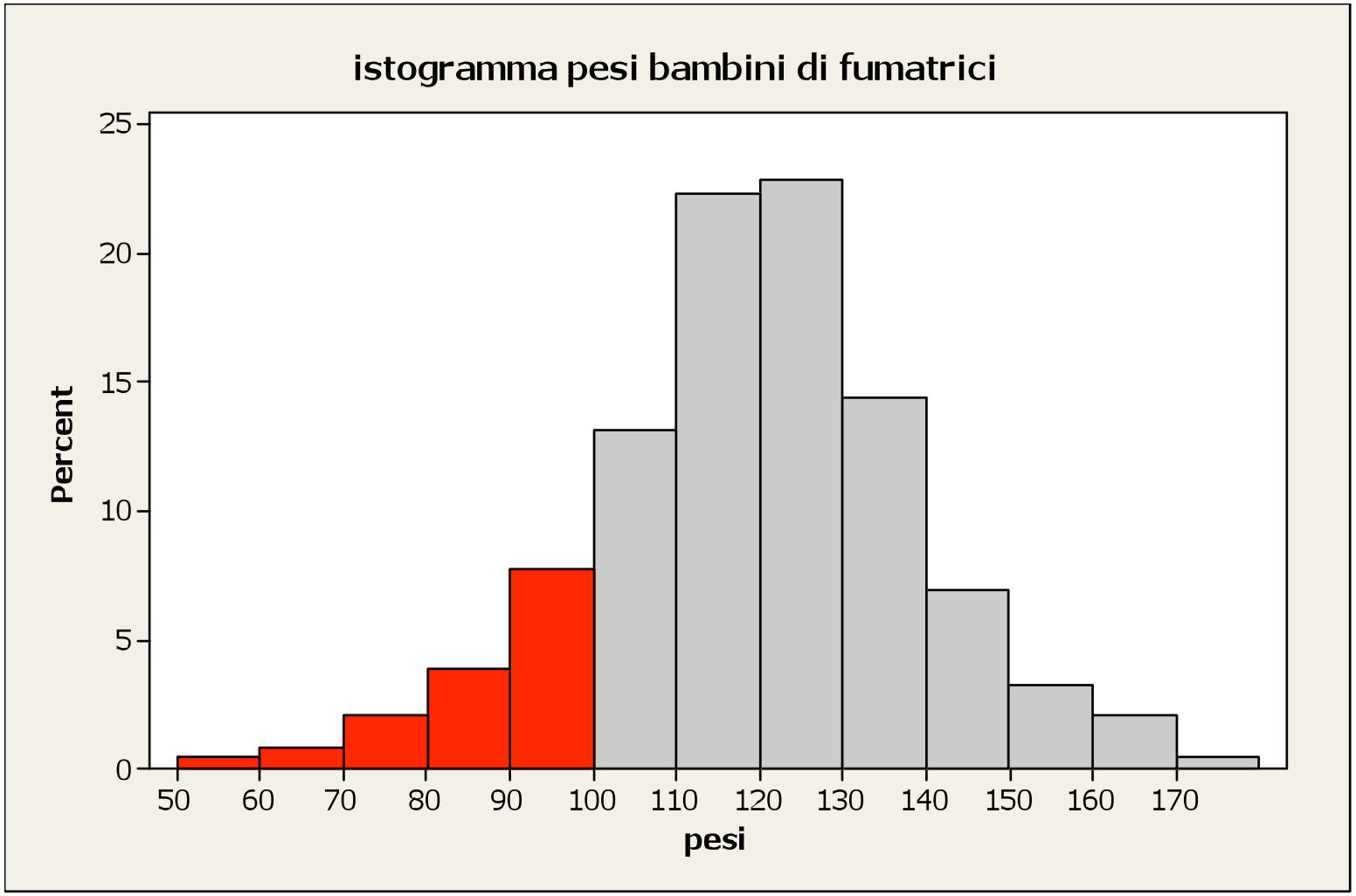


Se si aumenta il numero di osservazioni e si diminuisce l'ampiezza delle classi, l'istogramma si avvicina a una curva normale e l'area sotto la curva tra a=70.5 e b=71.5 è pari a 0.075 (si ricava dalle tavole della normale) ed è molto vicina alla percentuale (proporzione) osservata di germogli di lunghezza tra 70.5 e 71.5.

ESEMPIO: Pesi alla nascita

- Si considererà un sottoinsieme dei dati di un ampio studio condotto sulle donne in gravidanza tra il 1960 e il 1967 a San Francisco. Allo studio hanno partecipato 15000 famiglie con un livello di studio e di reddito medio-alto.
- Diverse misure del bambino venivano registrate alla nascita.
- Inizialmente considereremo 1236 maschi, nati tra il 1960 e il 1961, e che sono sopravissuti almeno 28 giorni. Per tali maschi si considereranno

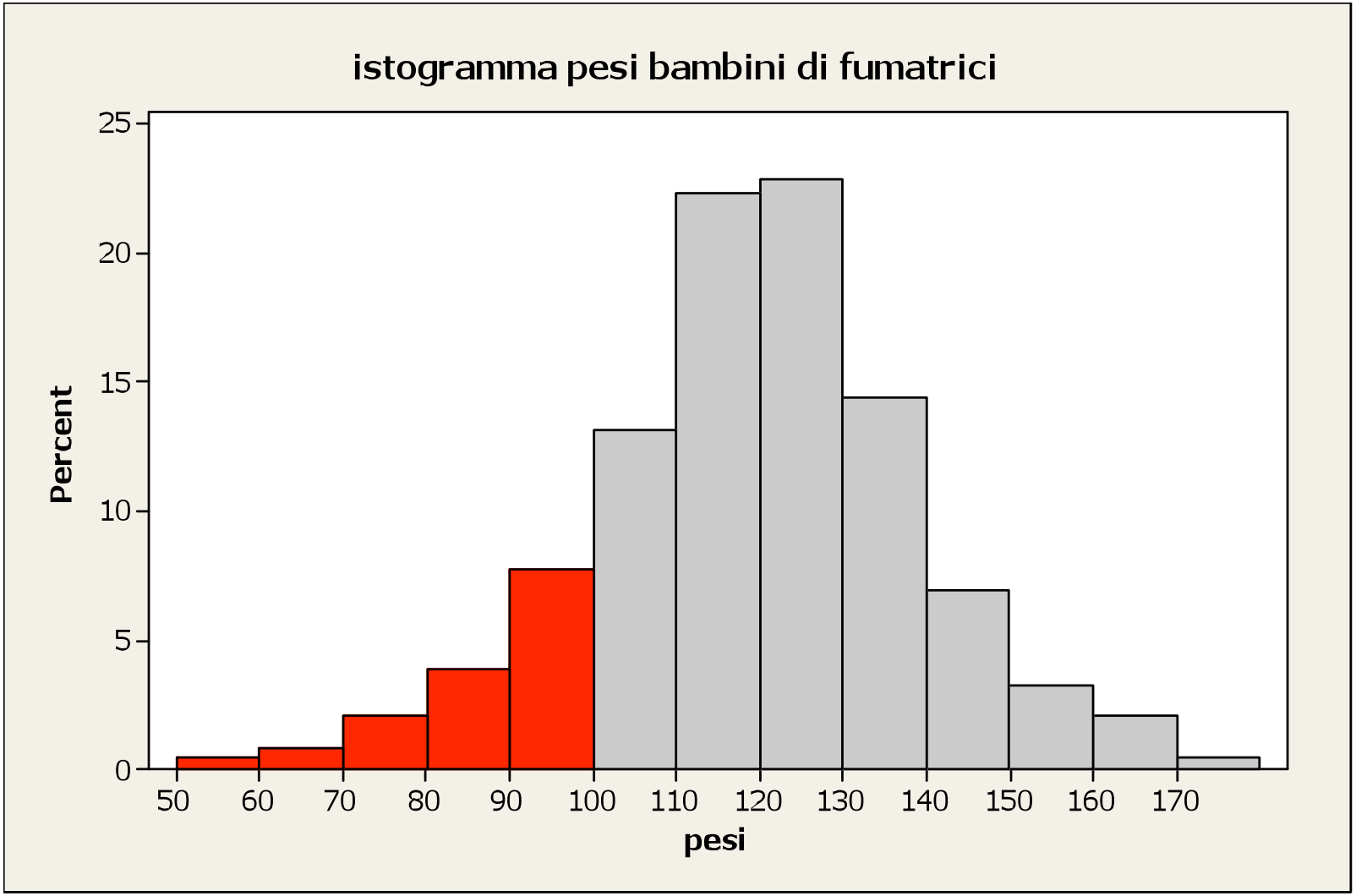
Variabile	Descrizione
Peso alla nascita	Peso alla nascita in once (0.035 once=1gr)
Abitudine al fumo della madre	Indicatore dell'abitudine al fumo in gravidanza. Fumo si (1), no (0)



Ampiezza classi=10

Campione n=1236

1) Interpretate l'istogramma. 2) A quale percentile corrisponde il peso di 100 once? 3) Come si interpreta tale percentile?



Qual è la probabilità che un bambino pesi tra 100 e 109 once?

Qual è la prob che un bambino pesi almeno 80 once?

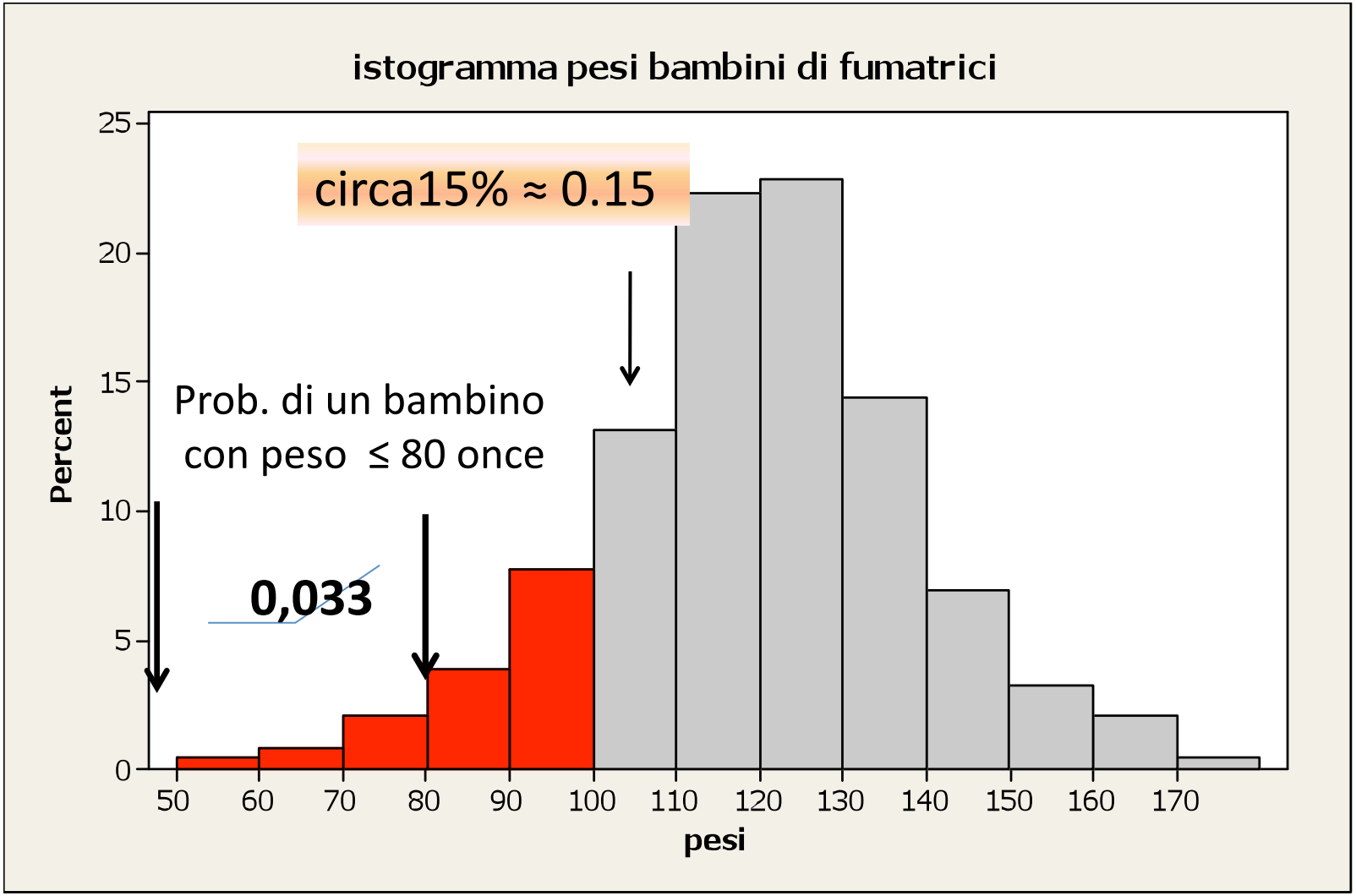
Pesi di bambini nati da fumatrici

fum	Count	Percent	CumPct
58	1	0,21	0,21
65	1	0,21	0,42
68	1	0,21	0,63
69	1	0,21	0,84
71	3	0,63	1,47
72	1	0,21	1,68
75	4	0,84	2,53
77	2	0,42	2,95
78	1	0,21	3,16
80	1	0,21	3,37
81	2	0,42	3,79
82	2	0,42	4,21
83	1	0,21	4,42
84	2	0,42	4,84
85	3	0,63	5,47
86	4	0,84	6,32
87	6	1,26	7,58
88	4	0,84	8,42
90	2	0,42	8,84
91	7	1,47	10,32
92	3	0,63	10,95
93	6	1,26	12,21
94	4	0,84	13,05
95	3	0,63	13,68
96	10	2,11	15,79
97	8	1,68	17,47
98	8	1,68	19,16
100	11	2,32	21,47
101	9	1,89	23,37
102	9	1,89	25,26
103	13	2,74	28,00
104	11	2,32	30,32
105	9	1,89	32,21
106	6	1,26	33,47
107	7	1,47	34,95
108	7	1,47	36,42
109	12	2,53	38,95

Prob che un bambino pesi tra 100 e 109 once → 100<peso<109

Pesi di bambini nati da fumatrici

fum	Count	Percent	CumPct
58	1	0,21	0,21
65	1	0,21	0,42
68	1	0,21	0,63
69	1	0,21	0,84
71	3	0,63	1,47
72	1	0,21	1,68
75	4	0,84	2,53
77	2	0,42	2,95
78	1	0,21	3,16
80	1	0,21	3,37
81	2	0,42	3,79
82	2	0,42	4,21
83	1	0,21	4,42
84	2	0,42	4,84
85	3	0,63	5,47
86	4	0,84	6,32
87	6	1,26	7,58
88	4	0,84	8,42
90	2	0,42	8,84
91	7	1,47	10,32
92	3	0,63	10,95
93	6	1,26	12,21
94	4	0,84	13,05
95	3	0,63	13,68
96	10	2,11	15,79
97	8	1,68	17,47
98	8	1,68	19,16
100	11	2,32	21,47
101	9	1,89	23,37
102	9	1,89	25,26
103	13	2,74	28,00
104	11	2,32	30,32
105	9	1,89	32,21
106	6	1,26	33,47
107	7	1,47	34,95
108	7	1,47	36,42
109	12	2,53	38,95



Qual è la probabilità che un bambino pesi tra 100 e 109 once? ≈ **0.15**

Qual è la prob che un bambino pesi almeno 80 once? →

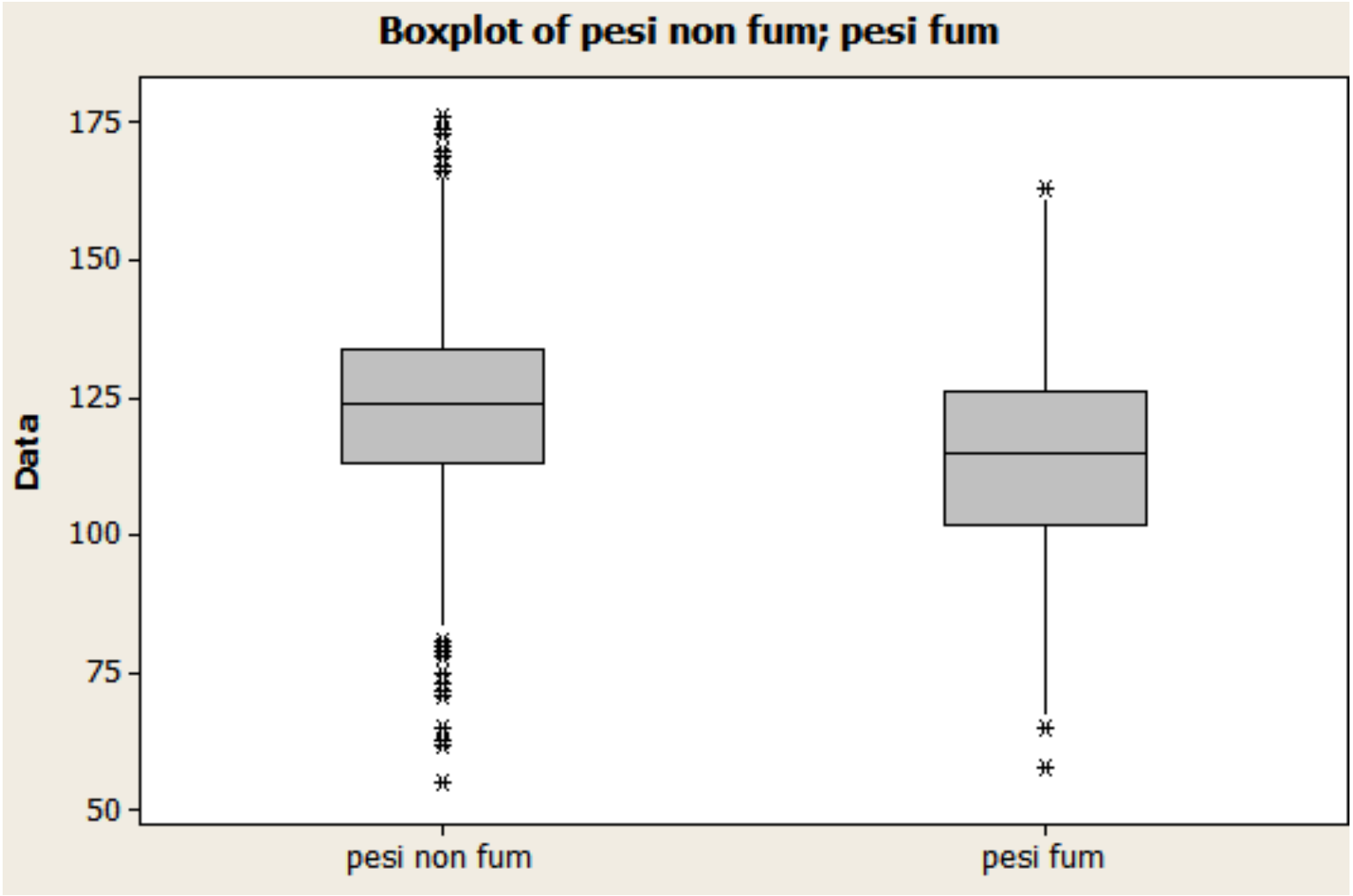
1- 0.033 = 0.967

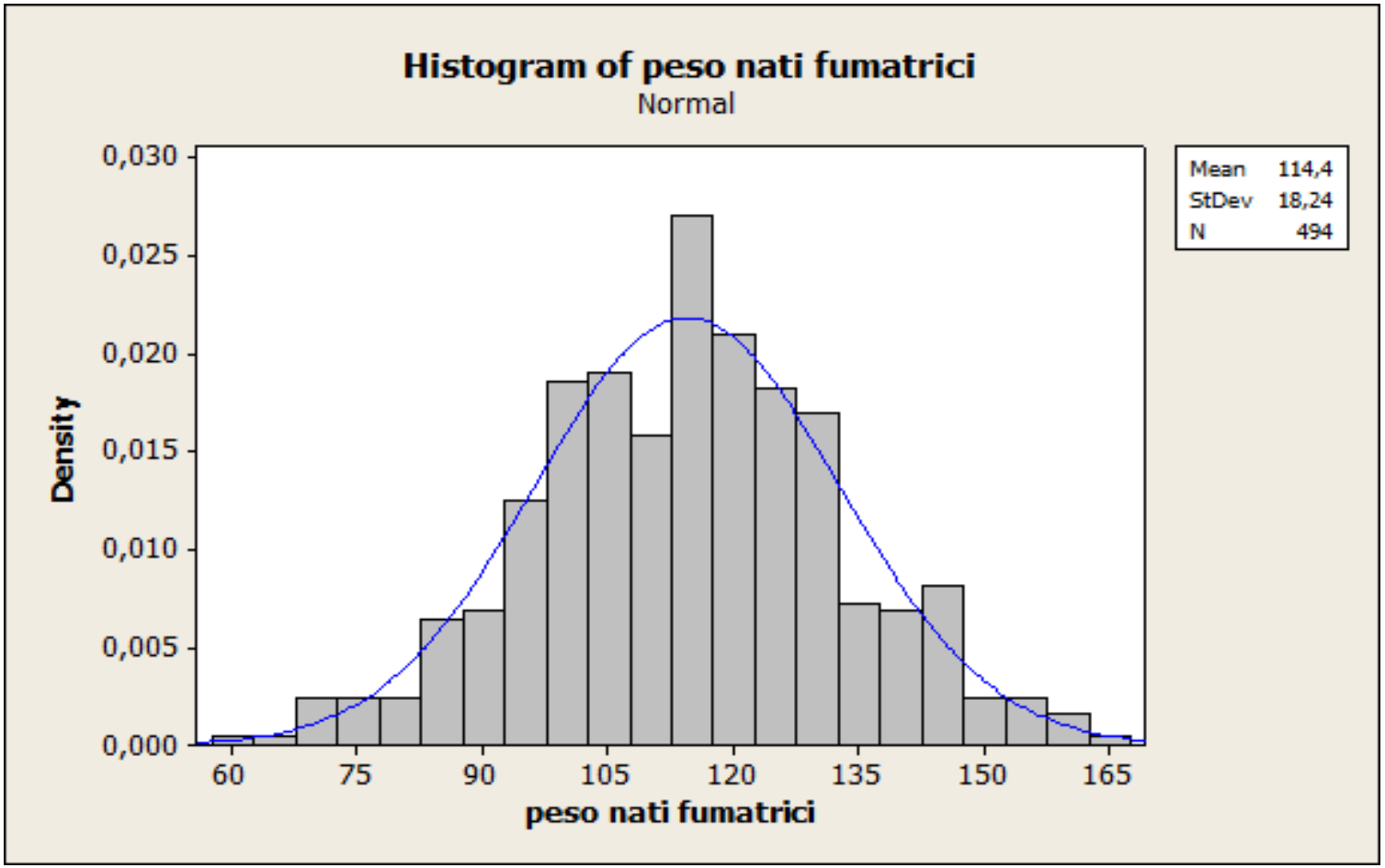
La classe mediana

Fumatrici e non fumatrici distinte

Classe birth weight	Cumulata percentuale **non F**	Cumulata percentuale **F**
50-59	0.13	0.21
60-69	0.53	0.83
70-79	1.61	3.10
80-89	3.36	8.27
90-99	7.54	20.66
100-109	17.78	40.09
110-119	40.16	**62.40**
120-129	**66.84**	81.00
130-139	85.71	90.91
140-149	94.07	97.11
150-159	97.70	99.18
160-169	99.19	100.00
170-179	100.00	100.00

Box-plot pesi non fum e pesi fum





Ampiezzac
lassi =5

campione
n=

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
114,36	18,24	58,00	102,00	115,00	126,25	163,00

Skewness	Kurtosis
-0,02	-0,02

Istogrammi e distribuzioni normali

- Entrambi gli istogrammi relativi ai campioni dei pesi dei neonati e delle lunghezze dei germogli suggeriscono per i dati osservati un andamento simile a quello di una distribuzione normale.
- Possiamo, pertanto, costruire un modello probabilistico normale per descrivere entrambi i fenomeni con riferimento **all'intera popolazione**.
- Si noti che abbiamo osservato in entrambi i casi **solo dei campioni.**
- La **curva di densità normale** rappresenta il modello complessivo per ciascuna delle due distribuzioni.

DISTRIBUZIONI DI FREQUENZA E DISTRIBUZIONI DI PROBABILITA'

- Ricordate:
- **Distribuzione di frequenza**: ricostruita a partire dai dati campionati
- **Distribuzione di probabilità**: ricostruita a partire dai dati di tutta la popolazione
- **Distribuzione teorica di probabilità**: è definita da una funzione matematica di cui conosco le caratteristiche e che mi permette di calcolare una probabilità associata a ciascun valore o intervallo di valori

La distribuzione normale

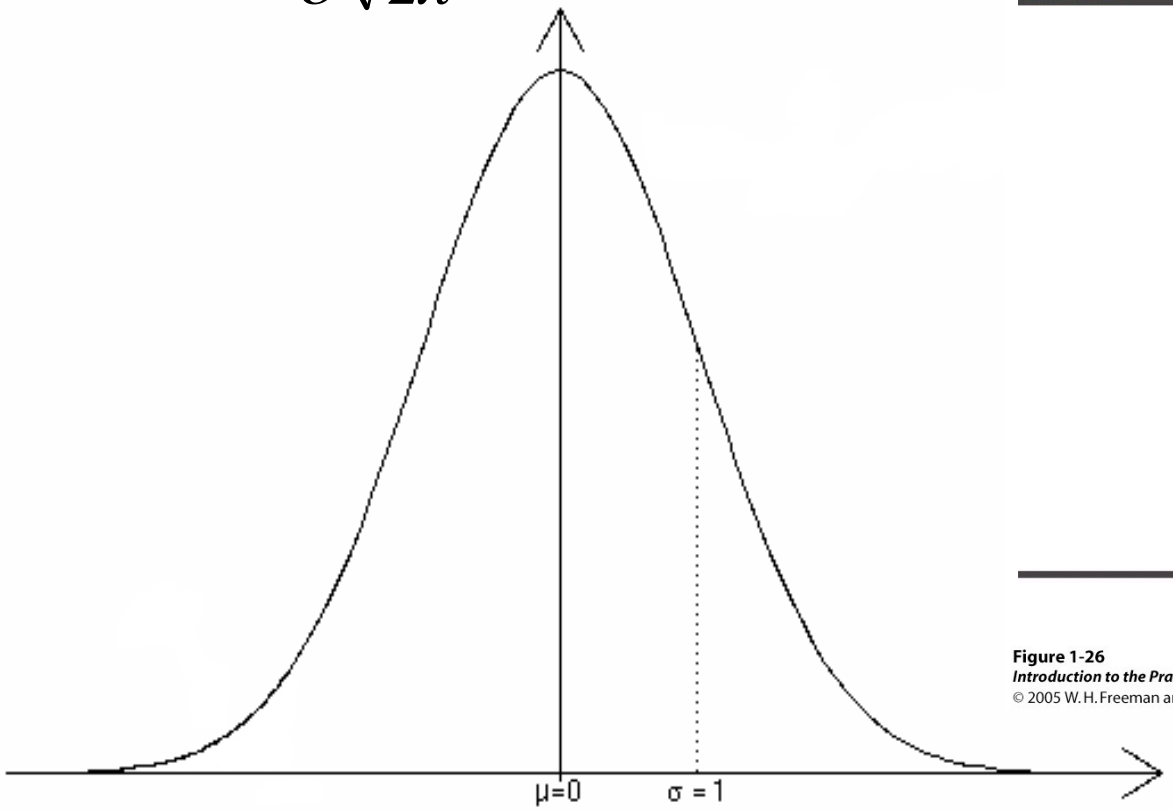
- Tutte le distribuzioni normali hanno la stessa forma generale. La curva di densità per una **particolare** distribuzione normale si ottiene specificando la sua **media μ** e la sua **deviazione standard σ** (o la sua varianza σ^2).

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

$$\left(-\infty < X < +\infty\right)$$

La Distribuzione Normale

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$



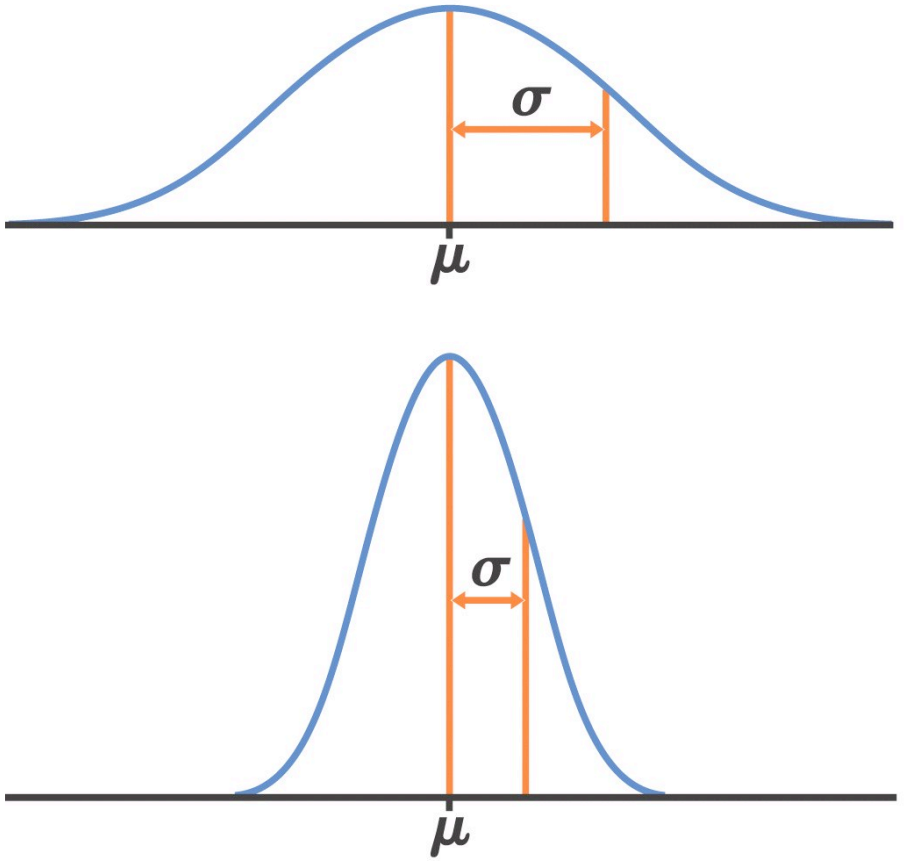
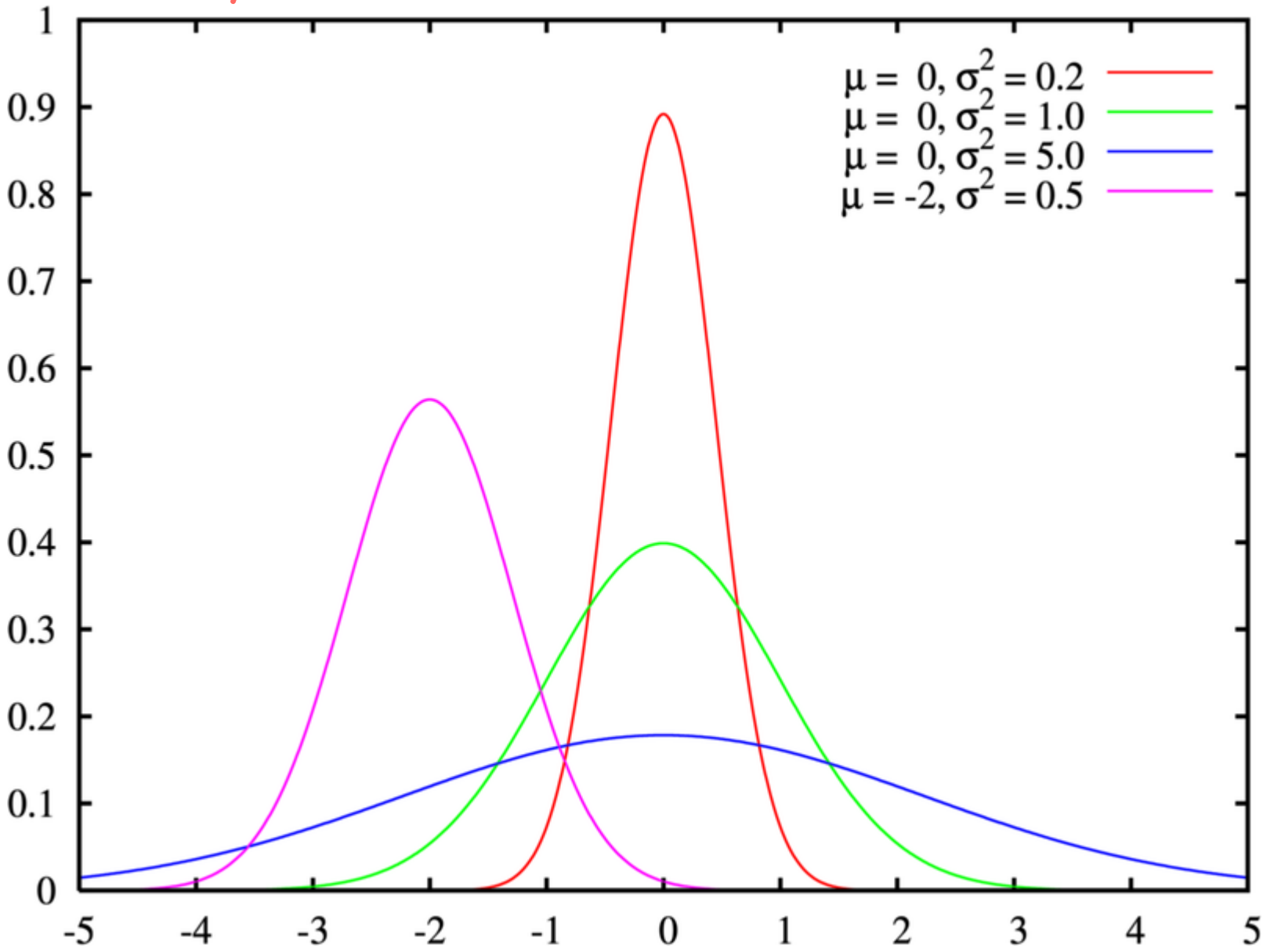


Figure 1-26
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Rappresentazione grafica di una distribuzione normale

Distribuzione normale N (μ, σ)- infinite distrib. al variare di μ e σ



Distribuzioni di v. a. continue: la distribuzione normale

Notate:

- L'altezza (asse delle y) di queste distribuzioni non fornisce **la probabilità** di osservare un valore (che è, per definizione, pari a 0)
- L'altezza della curva è invece una **densità di probabilità** (una probabilità divisa per un intervallo), e infatti si dovrebbero chiamare più precisamente distribuzioni di densità
- Quando ci serviremo di distribuzioni teoriche di probabilità per **variabili continue**, **sarà l'area** sottesa dalla curva, e non il valore di Y, a corrispondere alla **probabilità**.

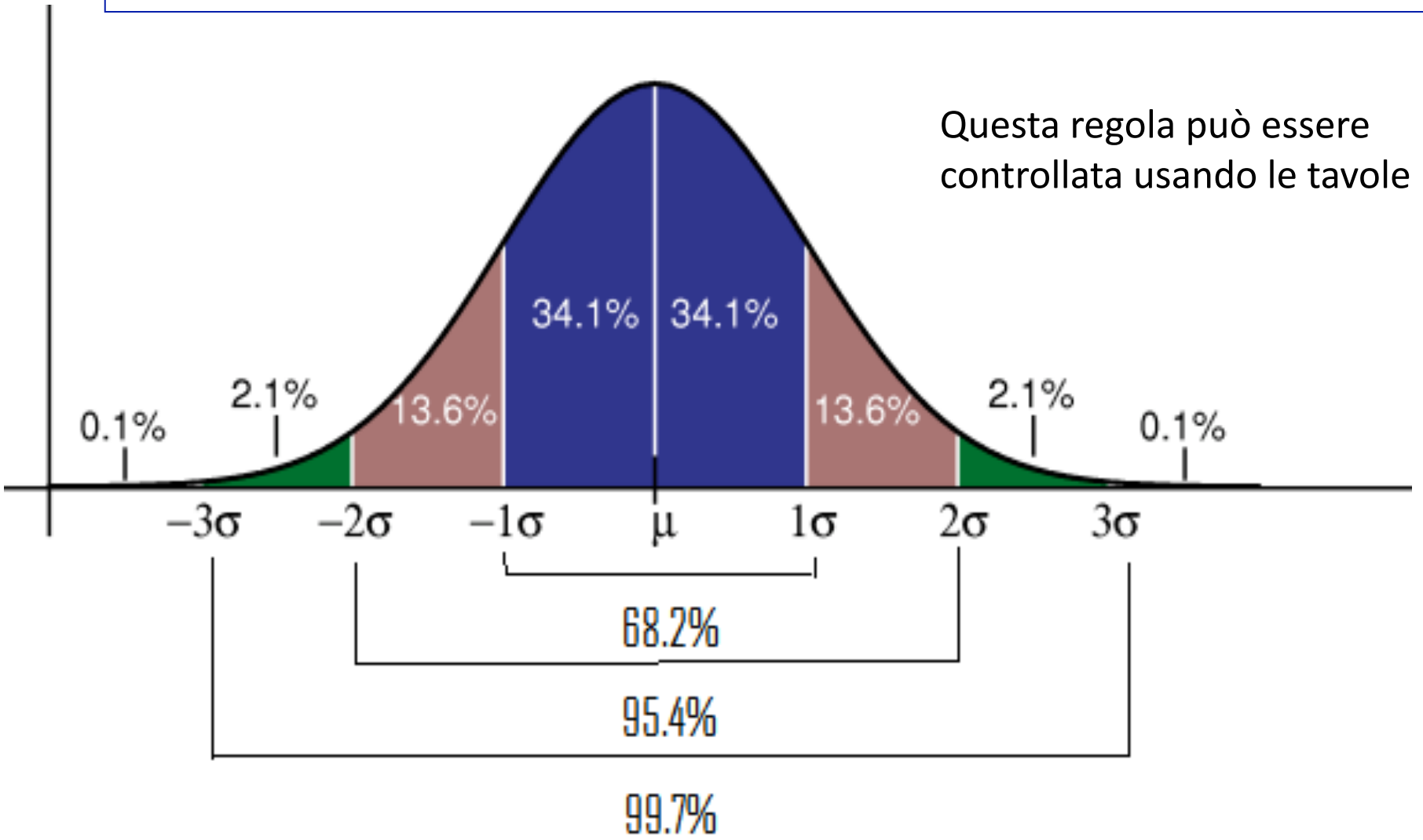
La distribuzione normale: una proprietà importante

La regola 68-95-99.7

Nella distribuzione Normale con media μ e deviazione standard σ:

- il **68%** delle osservazioni è compreso nell'intervallo
  $[\mu - \sigma, \mu + \sigma]$
- il **95%** delle osservazioni è compreso nell'intervallo
  $[\mu - 2\,\sigma, \mu + 2\sigma]$
- il **99.7%** delle osservazioni è compreso nell'intervallo
  $[\mu - 3\,\sigma, \mu + 3\sigma]$

Distribuzione Normale



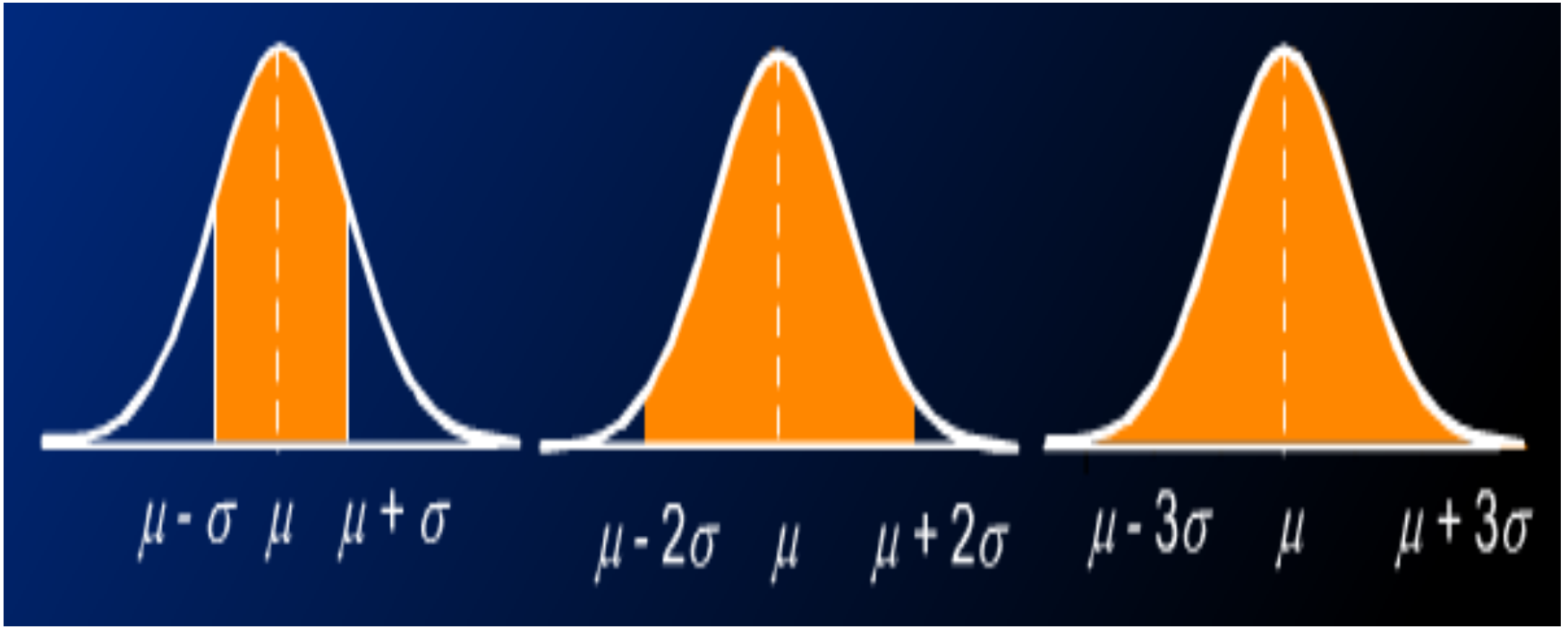
Questa regola può essere controllata usando le tavole

Questa regola è esattamente vera per una distribuzione normale. E' vera, con buona approssimazione, per le lunghezze dei germogli che sono approssimativamente normali.

Area colorata=
=0.683

Area colorata=
=0.954

Area colorata=
=0.997



C'è una probabilità pari al 68% di essere compresi tra μ-σ e μ+σ

C'è una probabilità pari al 95% di essere compresi tra μ-2σ e μ+2σ

C'è una probabilità pari al 99.7% di essere compresi tra μ-3σ e μ+3σ

La standardizzazione

Standardizzazione e valori *z*

Se *x* è un'osservazione da una distribuzione che ha media μ e deviazione standard σ, il **valore standardizzato** di *x* è

$$z = \frac{x - \mu}{\sigma}$$

Un valore standardizzato viene spesso chiamato **valore *z***.

La standardizzazione

- La standardizzazione trasforma la variabile aleatoria X che ha una distribuzione normale, con media μ e dev st σ, in una v. a. Z che ha una distribuzione normale standard con **media μ=0 e dev st =1**

- **Il valore z dice di quante deviazioni standard x dista dalla media**.
- Se x > μ → z è positivo
- Se x < μ → z è negativo

Le unità standard

1) **$X \sim N(\mu = 100, \sigma = 12)$**

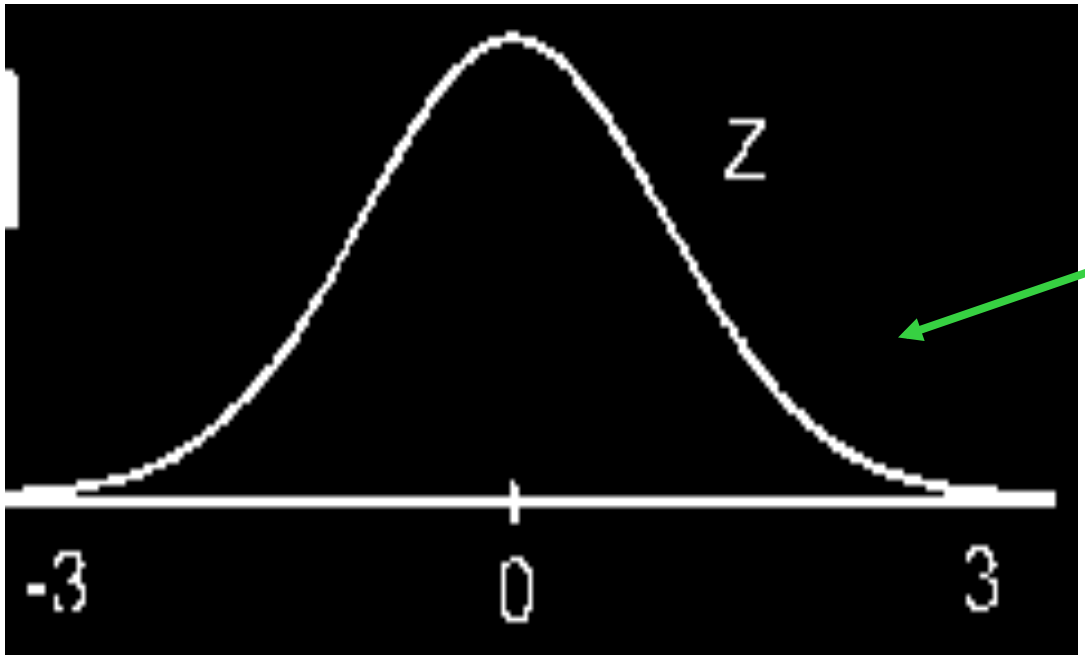
Trovare il valore z corrispondente a x = 128

$z = (128-100)/12$

$= \underline{2.333}$ (128 è maggiore della media di 2.3 dev. st.)

Se si standardizza una variabile aleatoria normale X si ottiene una nuova variabile aleatoria Z con distribuzione N (0,1).

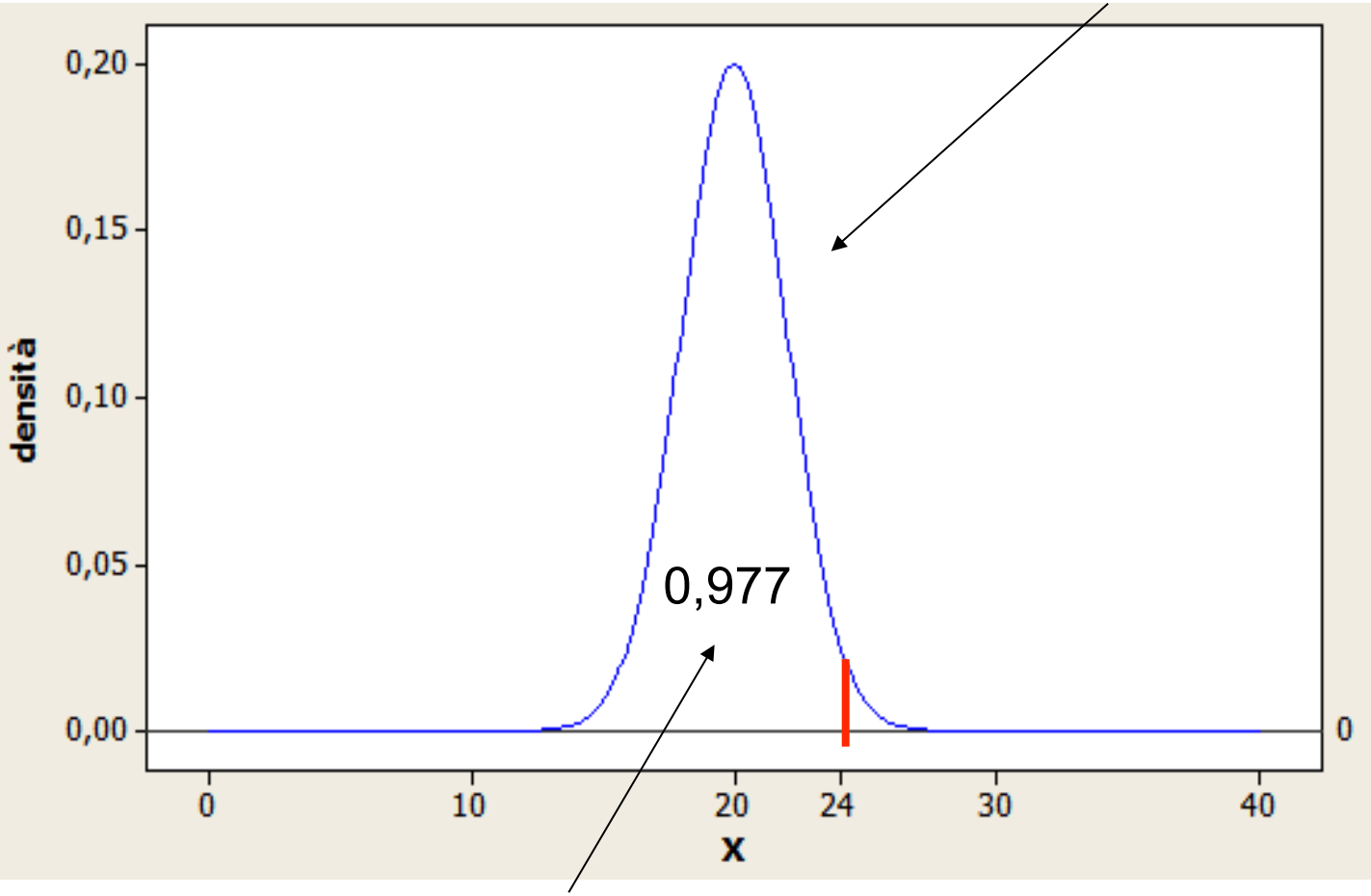
La distribuzione normale standard



$Z \sim N(0,1)$

La standardizzazione

N ~ (20, 2)



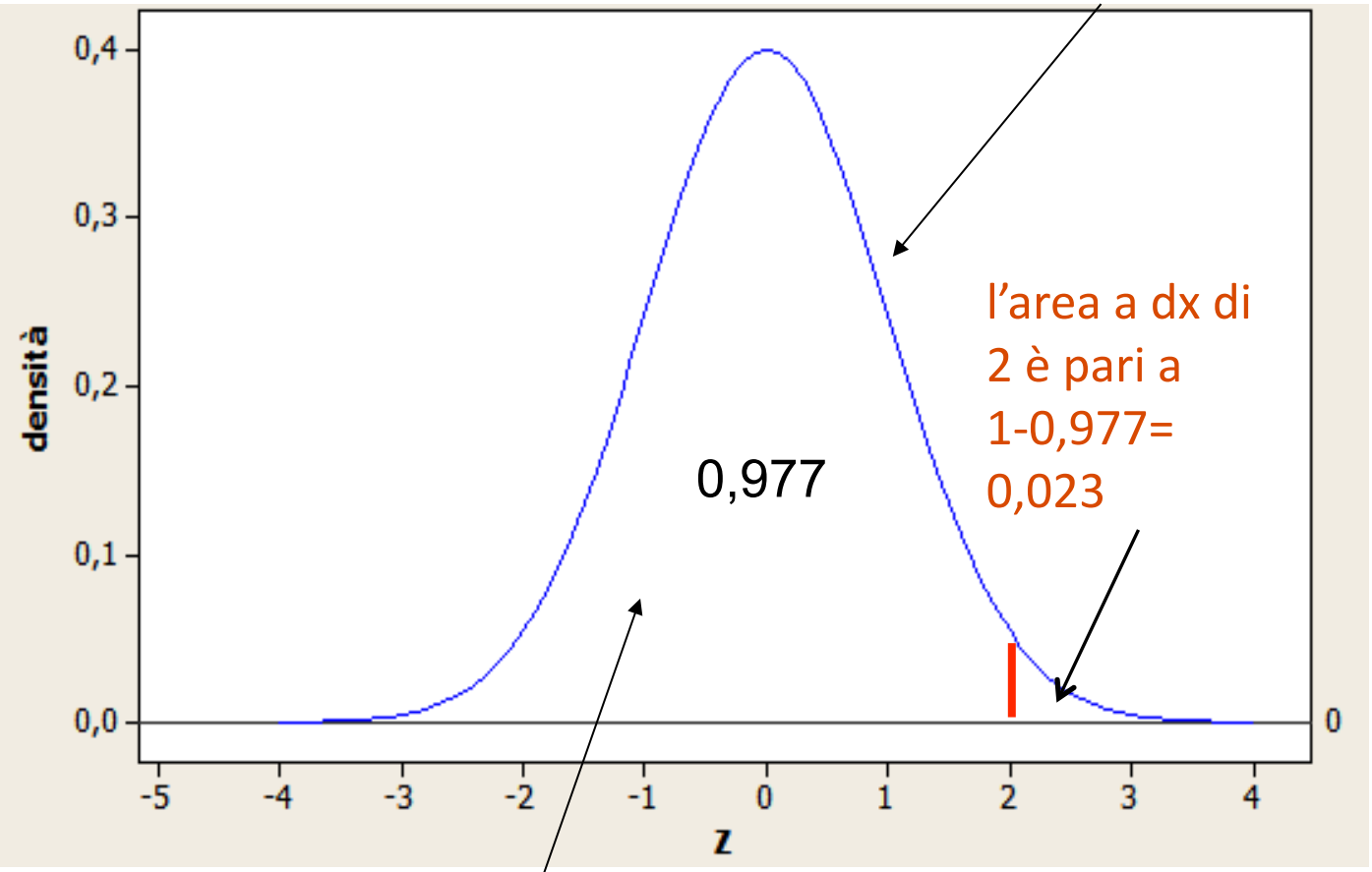
L'area sotto la curva normale fino al valore x=24 è pari a 0,977

La standardizzazione

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{24 - 20}{2} = 2$$

N ~ (0, 1)

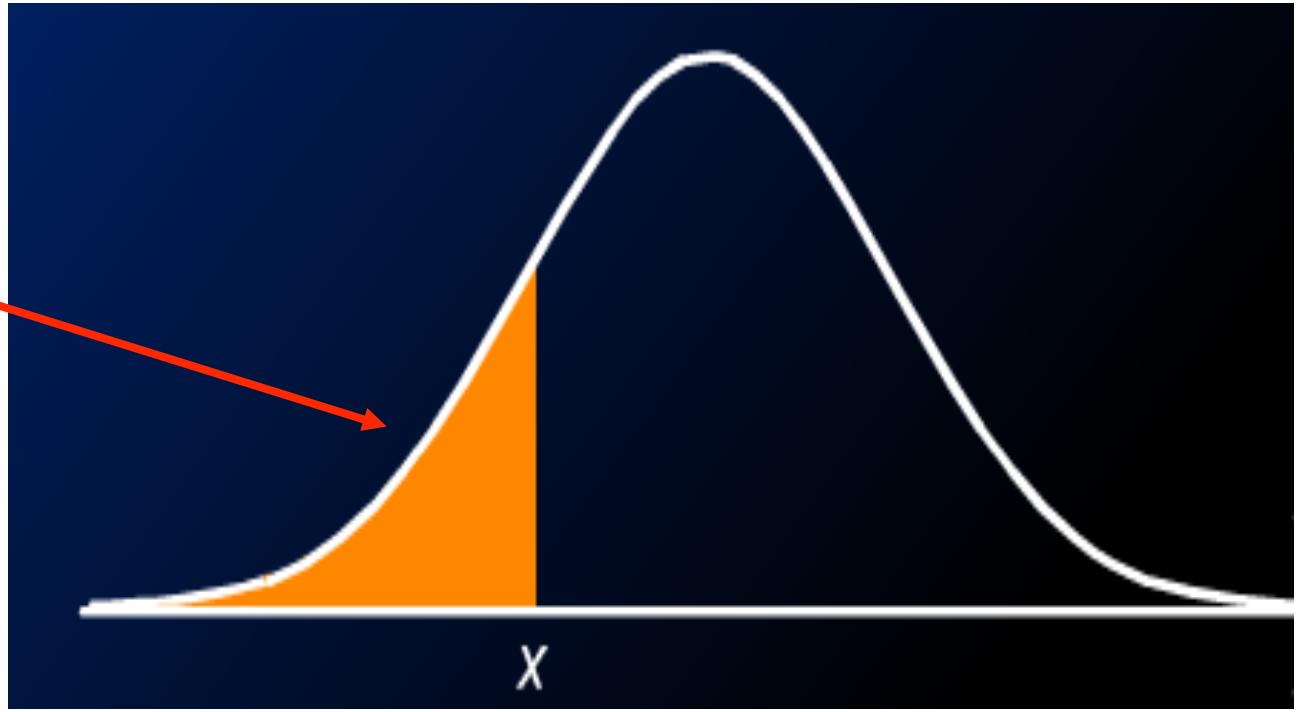


L'area sotto la curva normale standard fino al valore z = 2 è pari a 0,977

La distribuzione normale

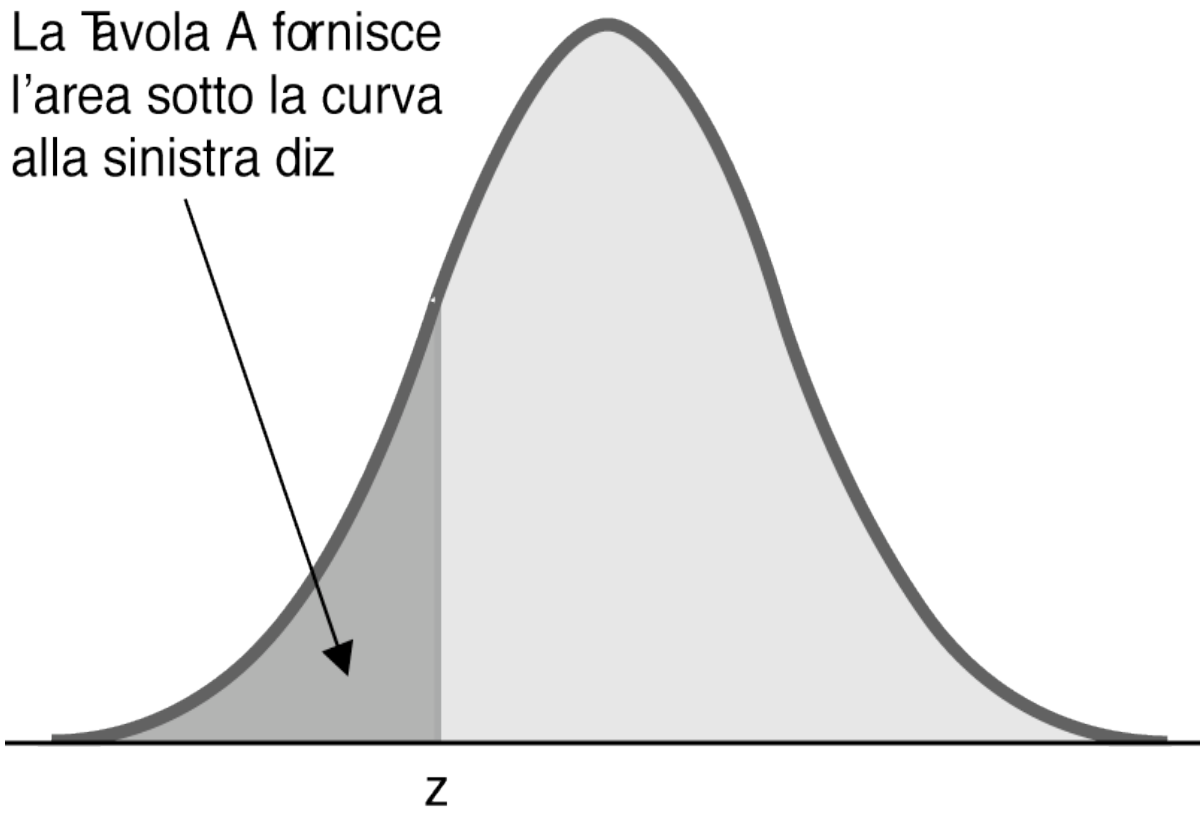
I **software statistici** calcolano l'area sotto la curva fino al punto x, ossia la proporzione di osservazioni che assumono valori ≤ x. Occorre precisare media e deviazione standard della distribuzione normale considerata.

Area =
= probabilità che un'unità scelta a caso abbia un valore ≤ x



La tavola della Normale standard

L'area sotto la curva alla sinistra di z corrisponde alla frequenza relativa cumulata nel punto z.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706

Area a sinistra di z = 1.47

TABLE A
STANDARD NORMAL PROBABILITIES (p.2)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.3										
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5										

NOTA: $P(a \leq z \leq b) = P(a<z<b)$

Tabelle on line:

http://econ.lse.ac.uk/ie/iecourse/ec220course_statstables0203.pdf

- Dato X ~ N (504,111)

  se x = 420 quanto vale z?

  z=(420-504)/111= -0.75

  Pr (Z<-0.75)?

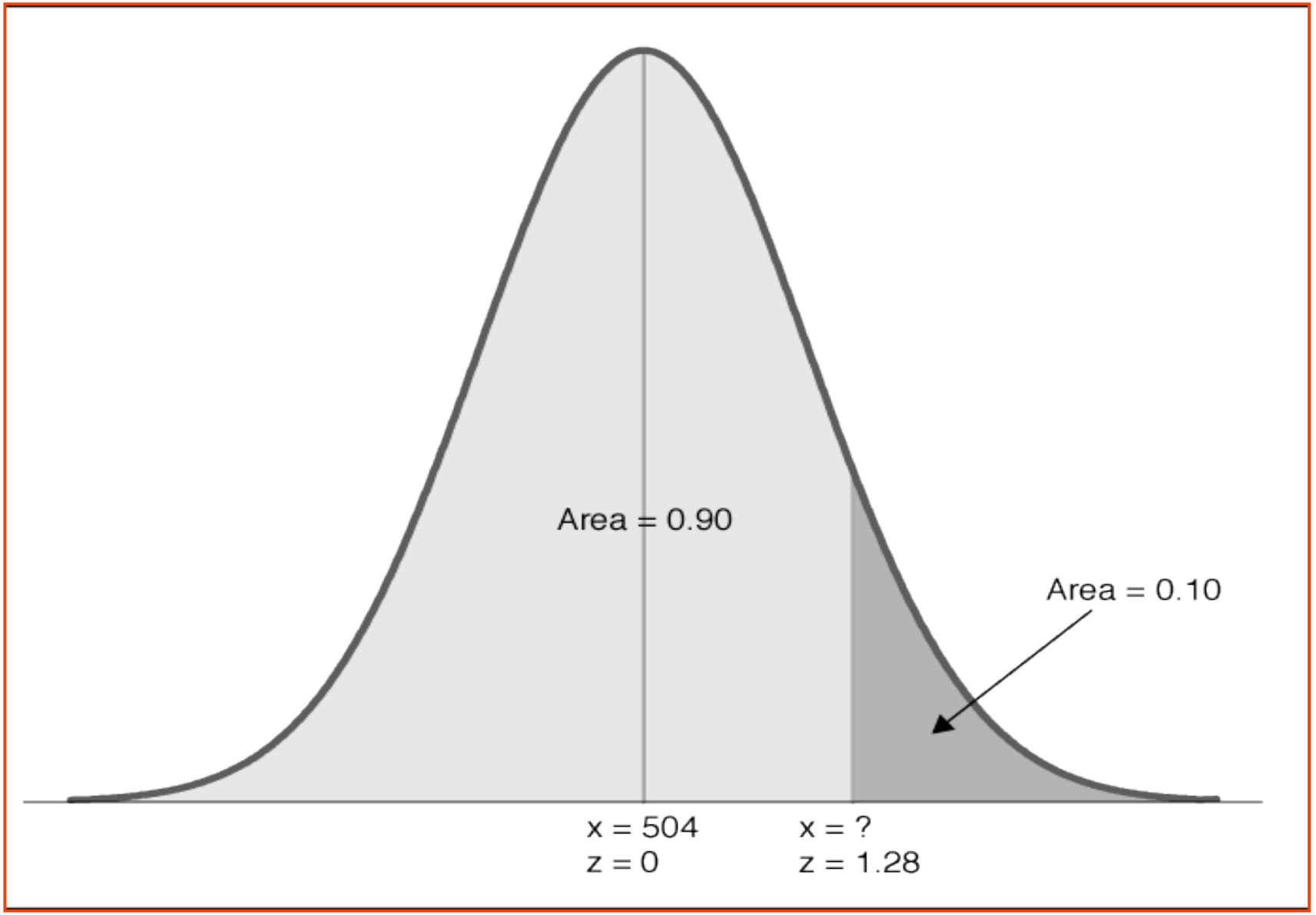
$$z = \frac{x - \mu}{\sigma}$$

  se z = 1.28

  quanto vale x?

  x = 504+(1.28)(111) = 646.1

  Qual è l'area a destra di 646.1?

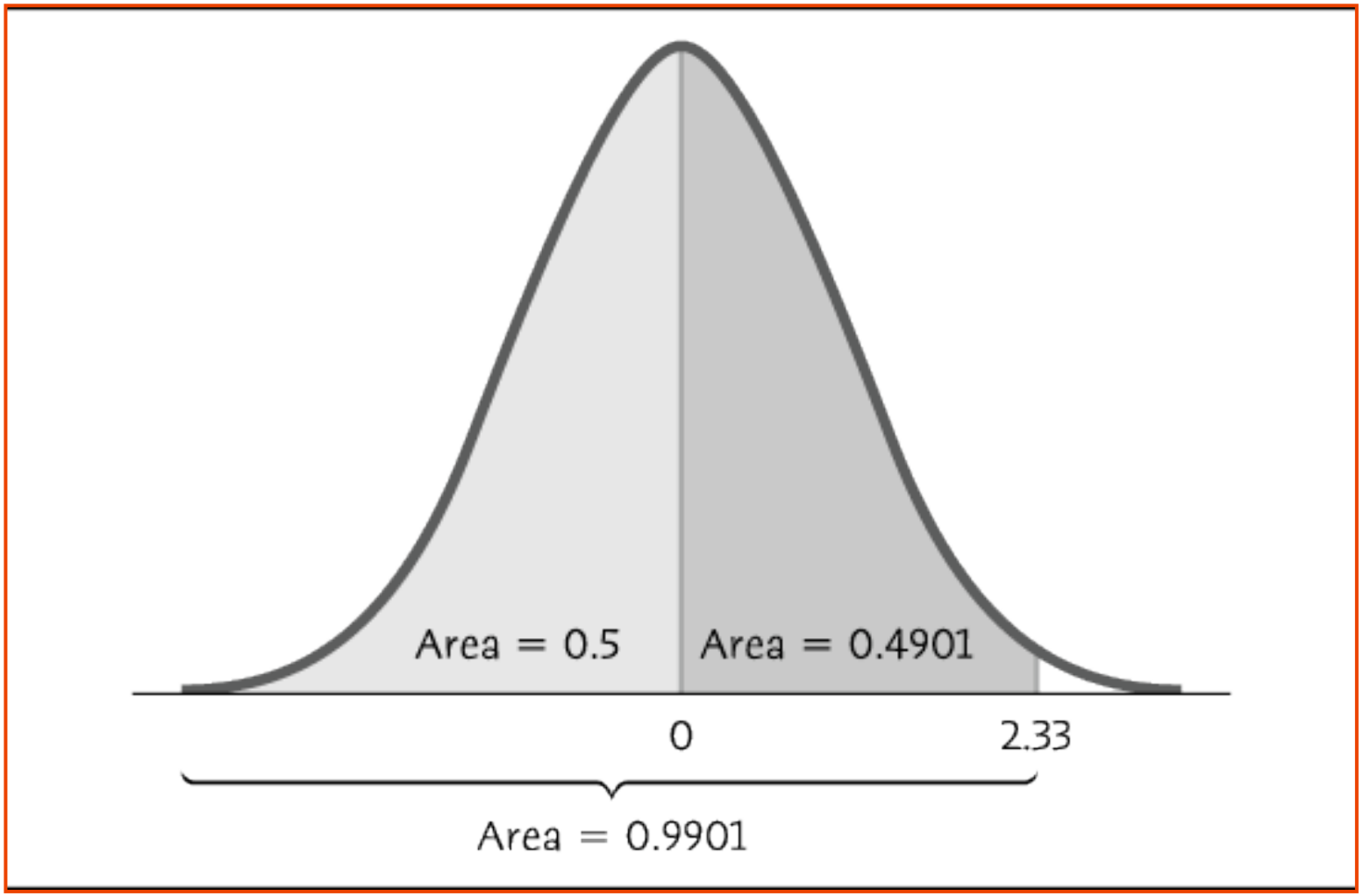


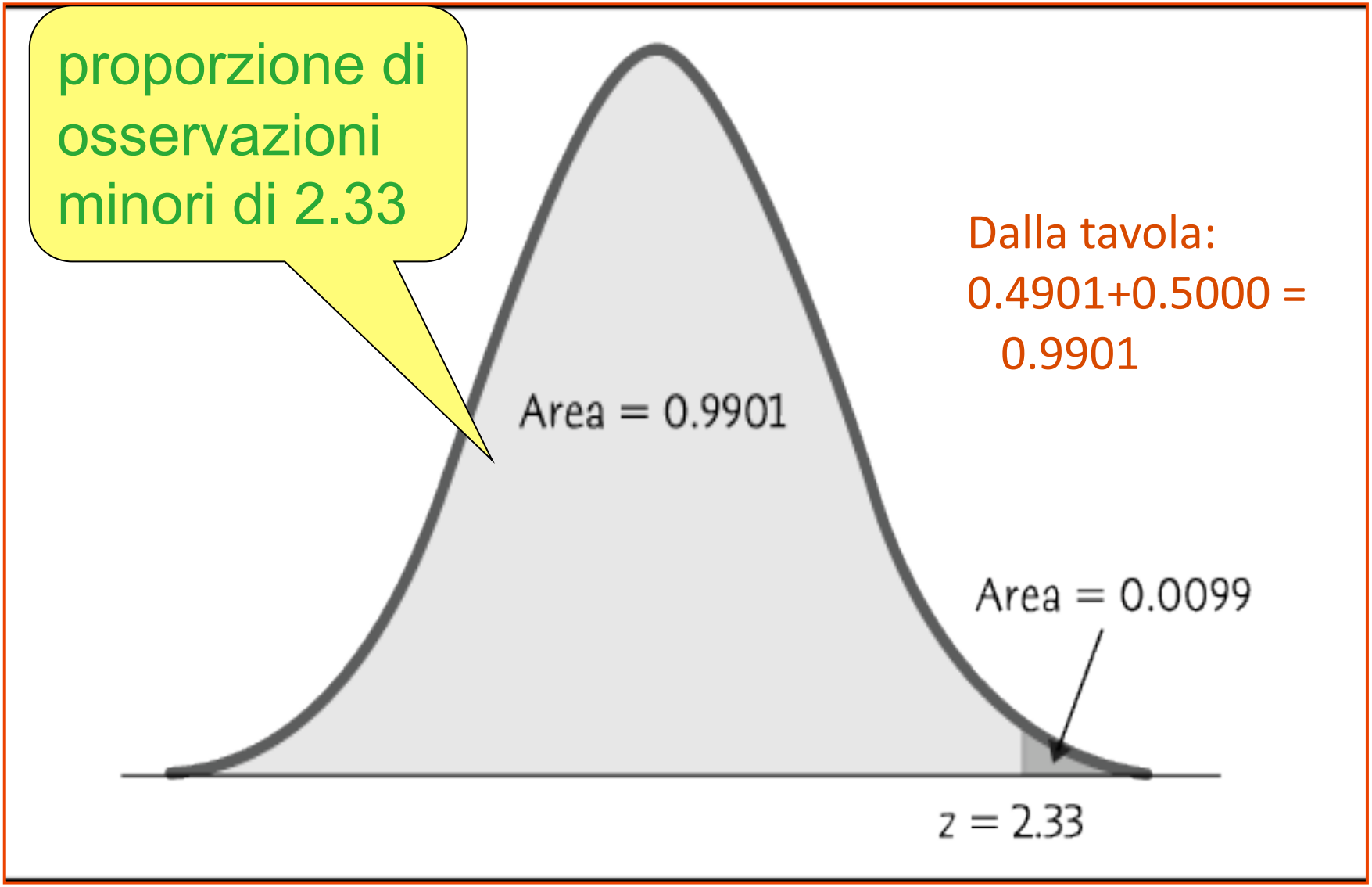
Dato X ~ N (504,111) se z = 1.28 →
x = 504+(1.28)(111) = 646.1

ESEMPIO

- **Quale proporzione di osservazioni di una variabile aleatoria Z assume un valore minore di 2.33?**
- **Ossia qual è la frequenza relativa (o probabilità) dei valori di**

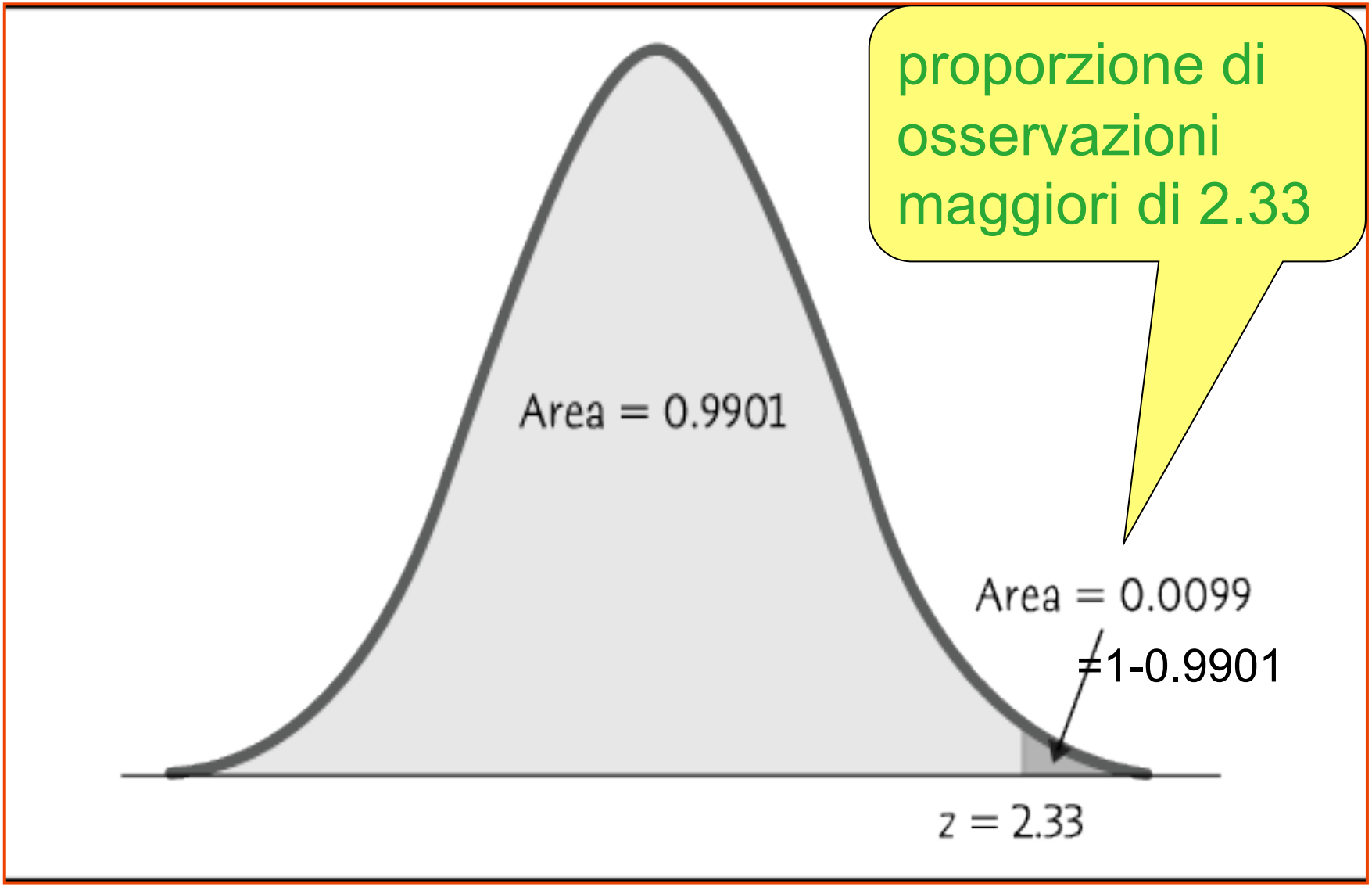
  $Z < 2.33$?


Area = 0.5
Area = 0.4901
0
2.33
Area = 0.9901

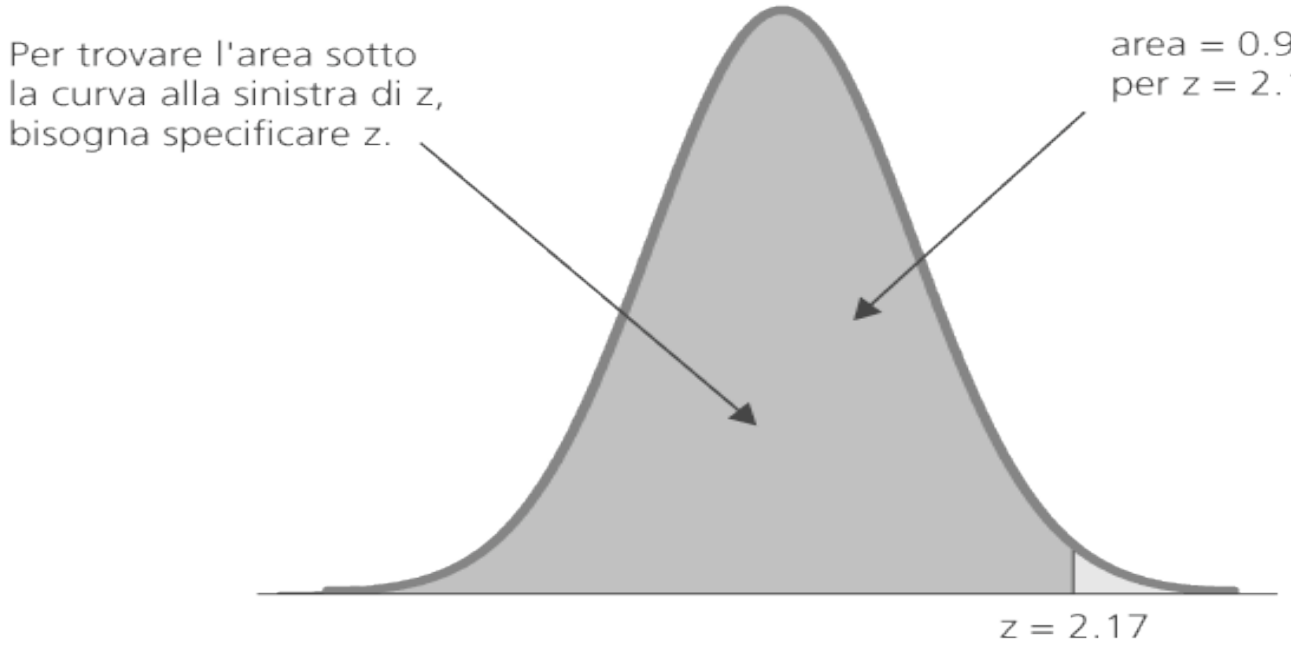


Dalla tavola:
0.4901+0.5000 =
0.9901

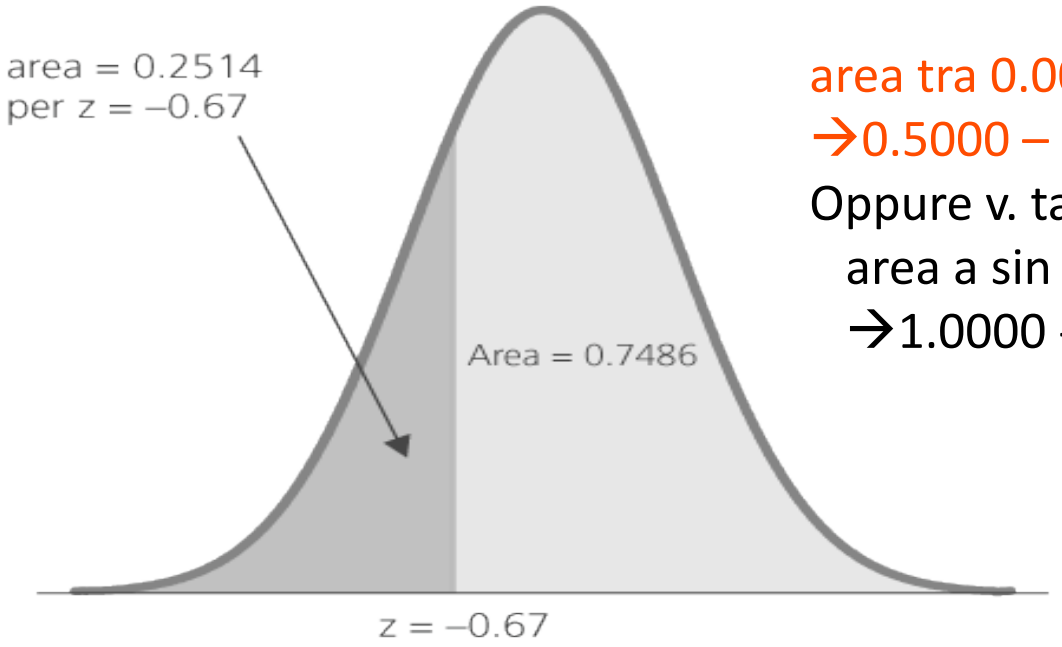
Quale proporzione di osservazioni di una variabile Z assume un valore minore di 2.33? Ossia qual è la frequenza relativa (o probab) della v. a. Z < 2.33?



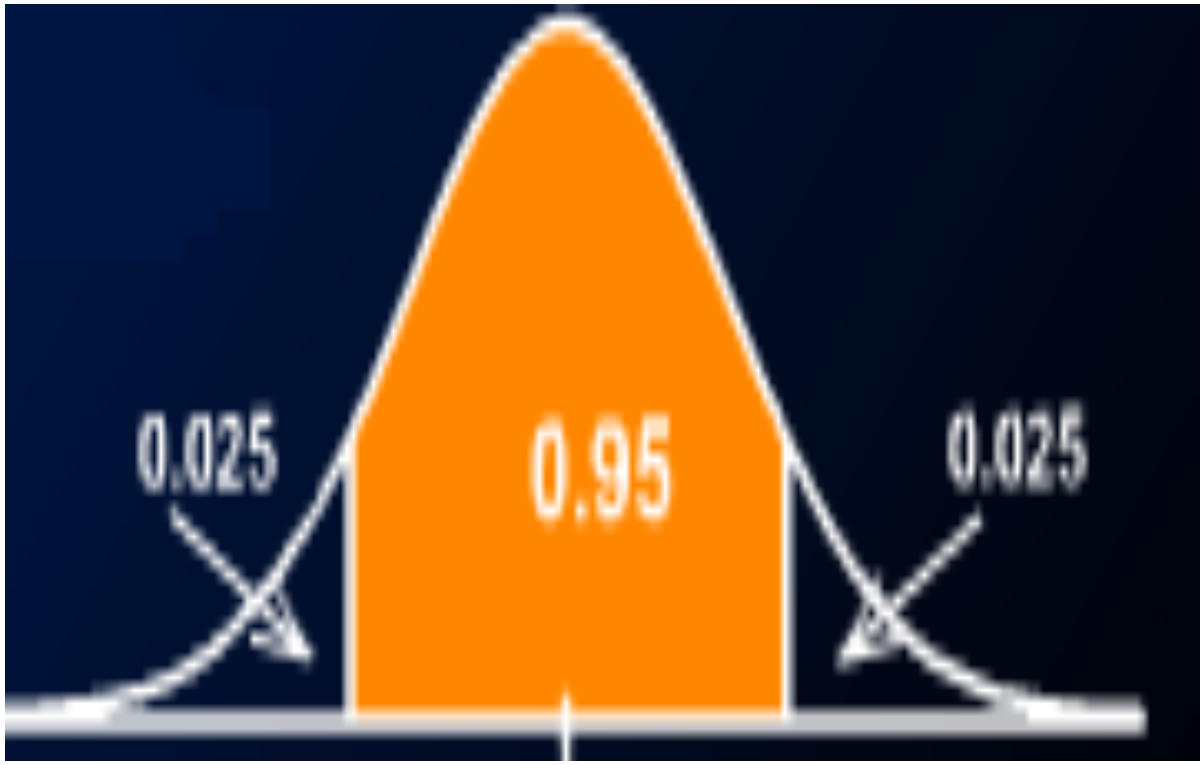
Quale proporzione di osservazioni della variabile Z assume un valore maggiore di 2.33? Ossia la probabilità che Z > 2.33?



(a)



area tra 0.00 e 0.67 = 0.2486
→0.5000 – 0.2486 = 0.2514
Oppure v. tavola rossa
area a sin di 0.67= 0.7486
→1.0000 - 0.7486 (b)



Quali sono gli estremi dell'intervallo che contiene il 95% dei valori centrali? Cosa si può dire in tal caso delle code della distribuzione?

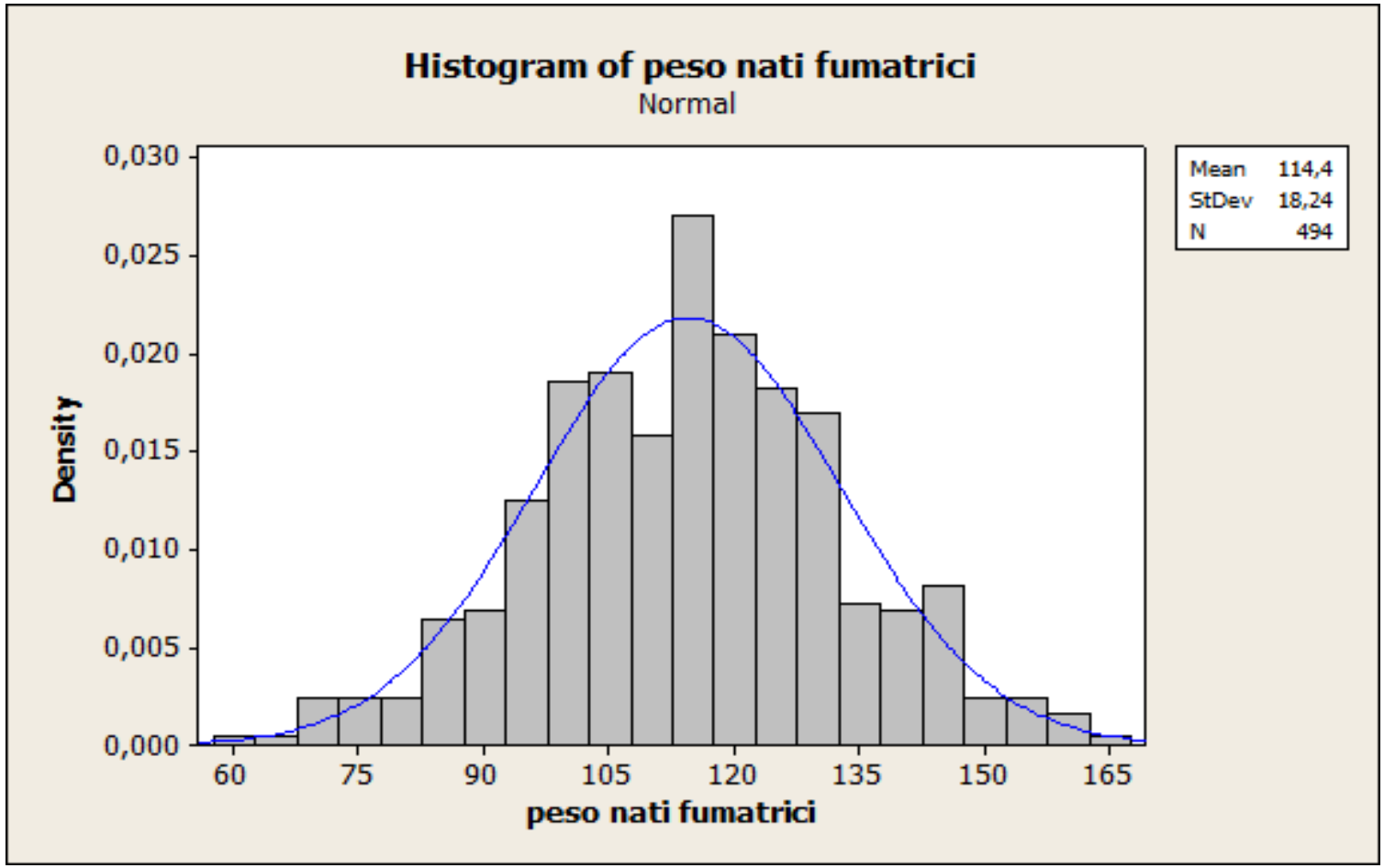
esercizio

- An exam is normally distributed with a mean of 200 points and a standard deviation of 25 points.

- (a) What percentage of the students score above 200 points?
- (b) What percentage of the students score below 175 points?
- (c) What percentage of the students score more than 250 points?
- a) 50% b) 16% c) 2%
- Quali sono I percentili corrispondenti a 200, 175 e 250?

  50-mo, 16-mo, 98-mo

ESERCIZIO



Ampiezza classi =5

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
114,36	18,24	58,00	102,00	115,00	126,25	163,00

Skewness	Kurtosis
-0,02	-0,02

ESERCIZIO

Pesi bambini nati da fumatrici

- Supponendo che i pesi (once) dei bambini si distribuiscano approx. secondo una normale N(114;18)

  i) a quale peso corrisponde il 5° percentile?

  ii) a quale peso corrisponde il 95° percentile?

  i) 84.39

  ii) 143.6

La distribuzione normale

- Esempio 15. La distribuzione del livello di colesterolo in un'ampia fascia di popolazione della stessa età e dello stesso sesso è approssimativamente normale. Per i ragazzi di 14 anni la media è $\mu = 170$mg di colesterolo per decilitro di sangue (mg/dl) e la dev. st. è $\sigma = 30$mg/dl. I livelli sopra 240mg/dl richiedono attenzione medica.
- Quale percentuale di ragazzi di 14 anni ha più di 240mg/dl di colesterolo?

- 1) Scriviamo il problema. Sia x il livello di colesterolo nel sangue. x ~ N(170,30). Bisogna trovare la percentuale di ragazzi con x>240.
- 2) Standardizziamo.

$$x > 240$$

$$\frac{x-170}{30} > \frac{240-170}{30}$$

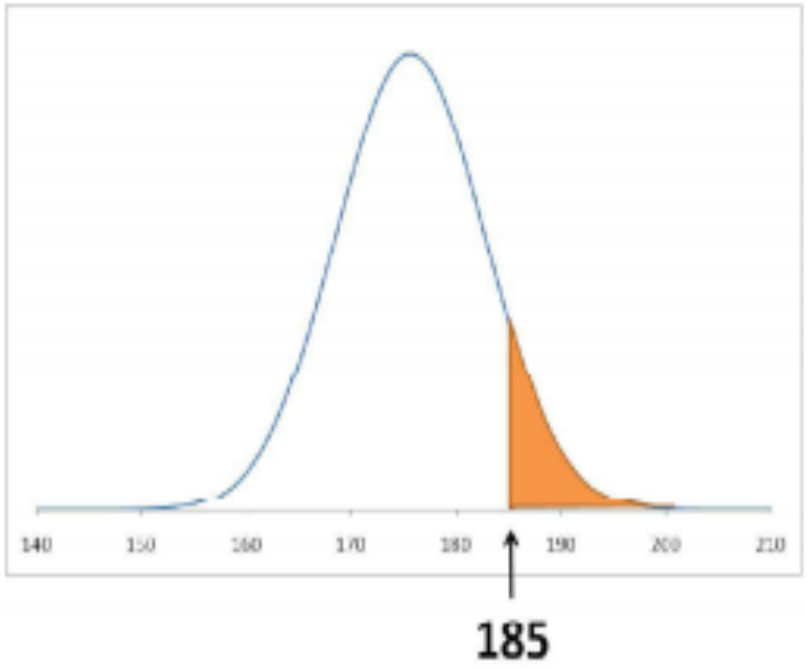
$$z > 2.33$$

- 3) Usiamo le tavole.

P(z < 2.33)= 0.9901. Poichè 1-0.9901=0.0099 diciamo che circa l'1% dei ragazzi ha un livello di colesterolo superiore a 240mg/dl

(tra 0.00 e 2.33→ 0.4893→ 0.5000-0.4893 = 0.01 =1%)

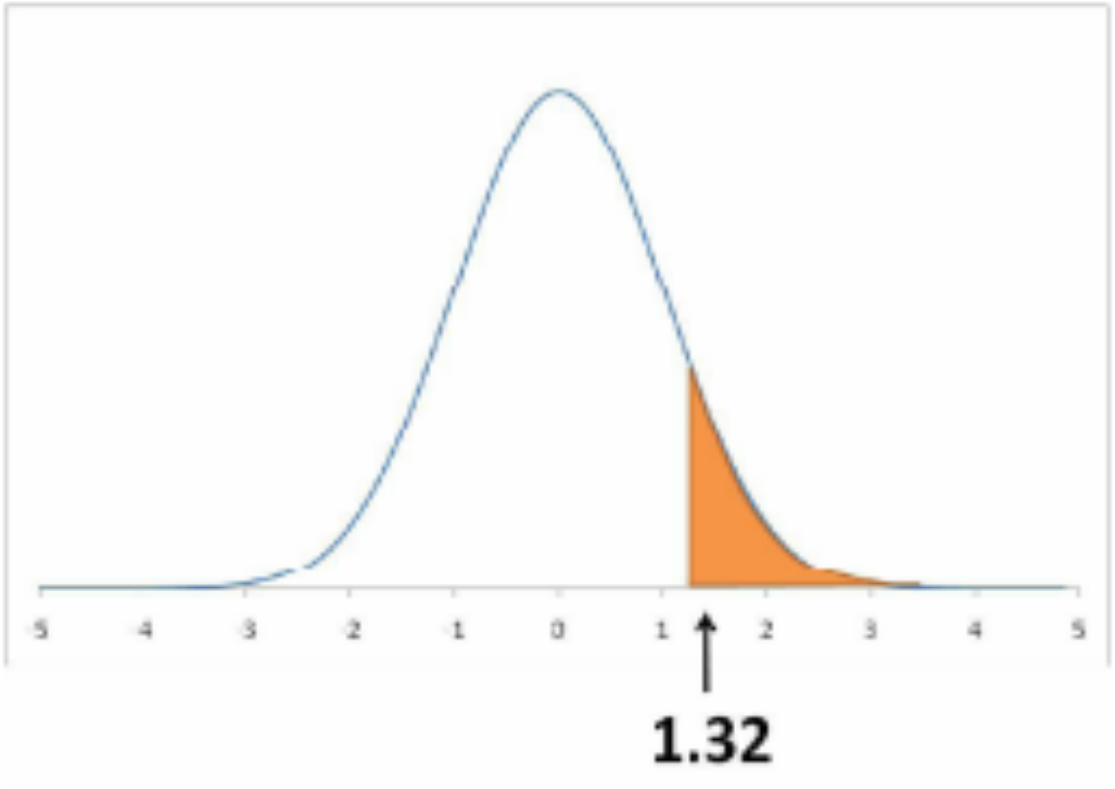
Esempio



Domanda: quale frazione di osservazioni saranno maggiori di 185 se la media è pari a 175.6 e la deviazione standard è pari a 7.1 (nella popolazione)?

Soluzione

$$Z = \frac{185 - 175.6}{7.1} = 1.32$$



- **esercizi con la normale**
- lunghezza ali mosche N(45.5, 3.90)
- Quale proporzione di mosche ha ali lunghe più di 51 decimi di mm?
- Quale proporzione ha ali lunghe tra 41 e 44?
- Quale proporzione ha ali lunghe meno di 39?
- Quale proporzione ha ali lunghe almeno 41 ma non più di 44?
- (occorre sempre standardizzare)
- **Problema inverso**
- Come trovare quel valore che ha una data proporzione di osservazioni al di sopra o al di sotto di esso?
- Quanto deve essere lunga l'ala per far si che solo il 10% delle mosche abbia ali più lunghe? (si tratta di trovare il 90 percentile)
- Qual è l'80-esimo percentile delle lunghezze delle ali?
- Qual è il quinto percentile delle lunghezze delle ali?
- (occorre destandardizzare: $x = \mu+\sigma z$)

SOLUZIONI ESERCIZI

- 1) (51-45.5)/3.90 = 1.41→1.000-0.920=0.08
- 2) (44-45.5)/3.90 = -0.38 →1.000-0.648=0.352

  (41-45.5)/3.90 = -1.153 →1.000-0.875=0.125

  →0.352-0.125=0.227

  4) 0.875-0.648=0.227

Problema inverso

Qual è l'80-esimo percentile?

z=0.84 x=45.5+(0.84*3.90)=48.77mm

Come si valuta la “normalità” di una distribuzione di dati?

- Come possiamo giudicare se i dati provengono da una distribuzione che può essere approssimata con una normale?
- Gli istogrammi, i diagrammi ramo-foglia e alcuni indici possono rivelare caratteristiche tipicamente non normali:

  outlier, asimmetria, interruzioni dei valori (gap), clusters.
- Se i grafici appaiono abbastanza simmetrici e unimodali occorre un metodo più sensibile, che possa rivelare l’adeguatezza del modello normale (simmetria, outlier, peso delle code).

Plot dei quantili normali. Come si costruisce

Vogliamo verificare se un determinato campione proviene da una distribuzione normale (con ugual media e scarto st.)

1. Si ordinano le osservazioni , e si calcolano i **percentili campionari x_i.**
2. Si considera la distribuzione normale standard e si trovano **i valori z_i che corrispondono agli stessi percentili (quantili normali standardizzati)**
3. Si costruisce un diagramma di dispersione con le osservazioni x_i sull'asse orizzontale e i corrispondenti quantili normali standardizzati z_i sull'asse verticale
4. Si verifica la normalità delle osservazioni controllando se i punti del diagramma si trovano approssimativamente su una retta

Tutti i software statistici riportano i plot dei quantili normali (normal plots).

Plot dei quantili normali: metodo grafico di controllo della normalità dei dati

Un campione ordinato di lunghezze di 19 germogli

(v. es. precedente)

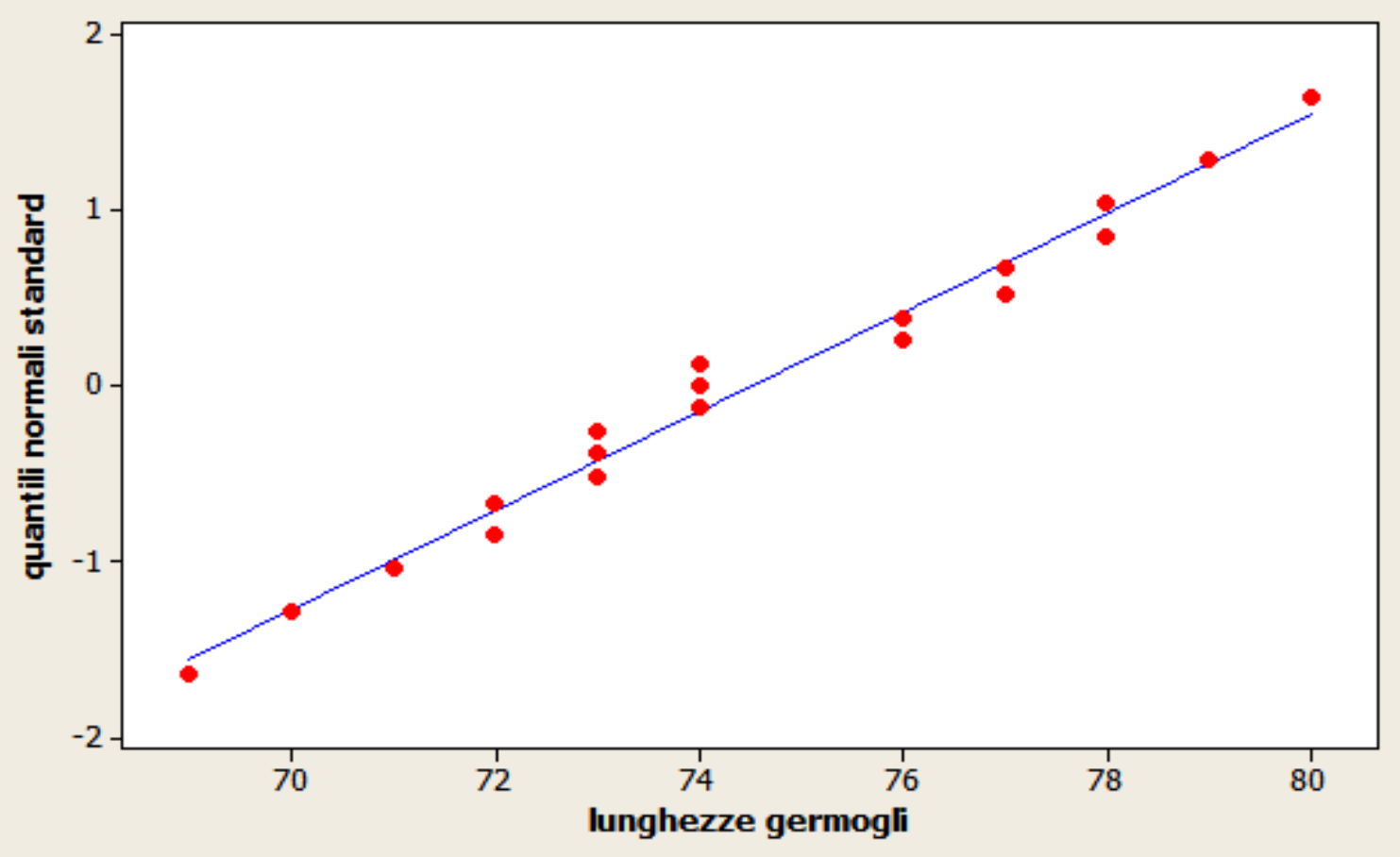
x_i	z_i	
69	-1,64485	
70	-1,28155	
71	-1,03643	
72	-0,84162	
72	-0,67449	26-mo → percentile
73	-0,52440	
73	-0,38532	
73	-0,25335	
74	-0,12566	lunghezze germogli
74	0,00000	
74	0,12566	
76	0,25335	
76	0,38532	
77	0,52440	
77	0,67449	
78	0,84162	
78	1,03643	
79	1,28155	
80	1,64485	

(quinto quantile)
26-mo percentile

quantili normali standardizzati

(diciannovesimo quantile)
95-mo percentile

Plot dei quantili normali per l'es. germogli: diagramma di dispersione

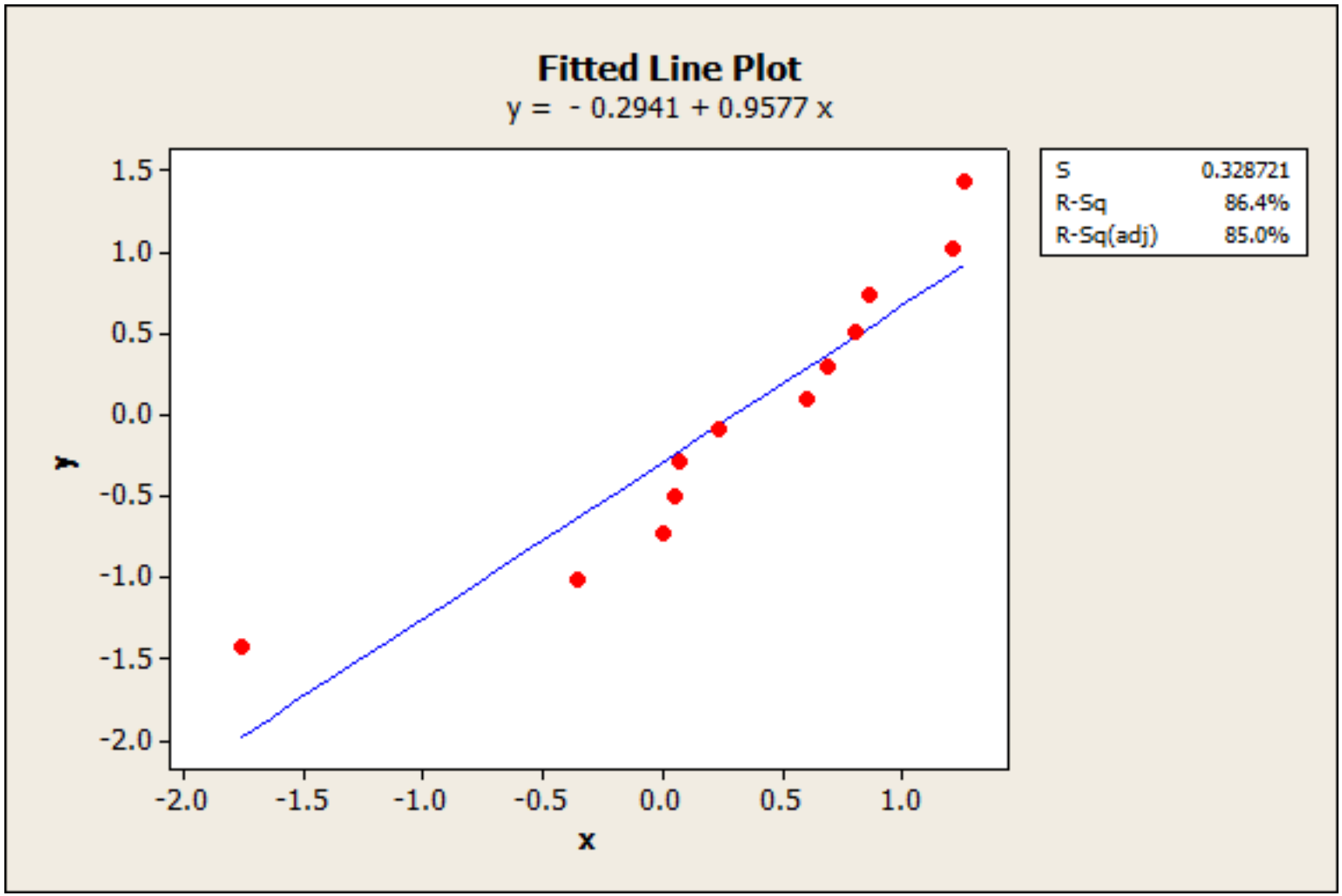


I punti del normal plot si dispongono approssimativamente su una retta inclinata positivamente. Le lunghezze dei 19 germogli hanno una distribuzione approssimativamente normale.

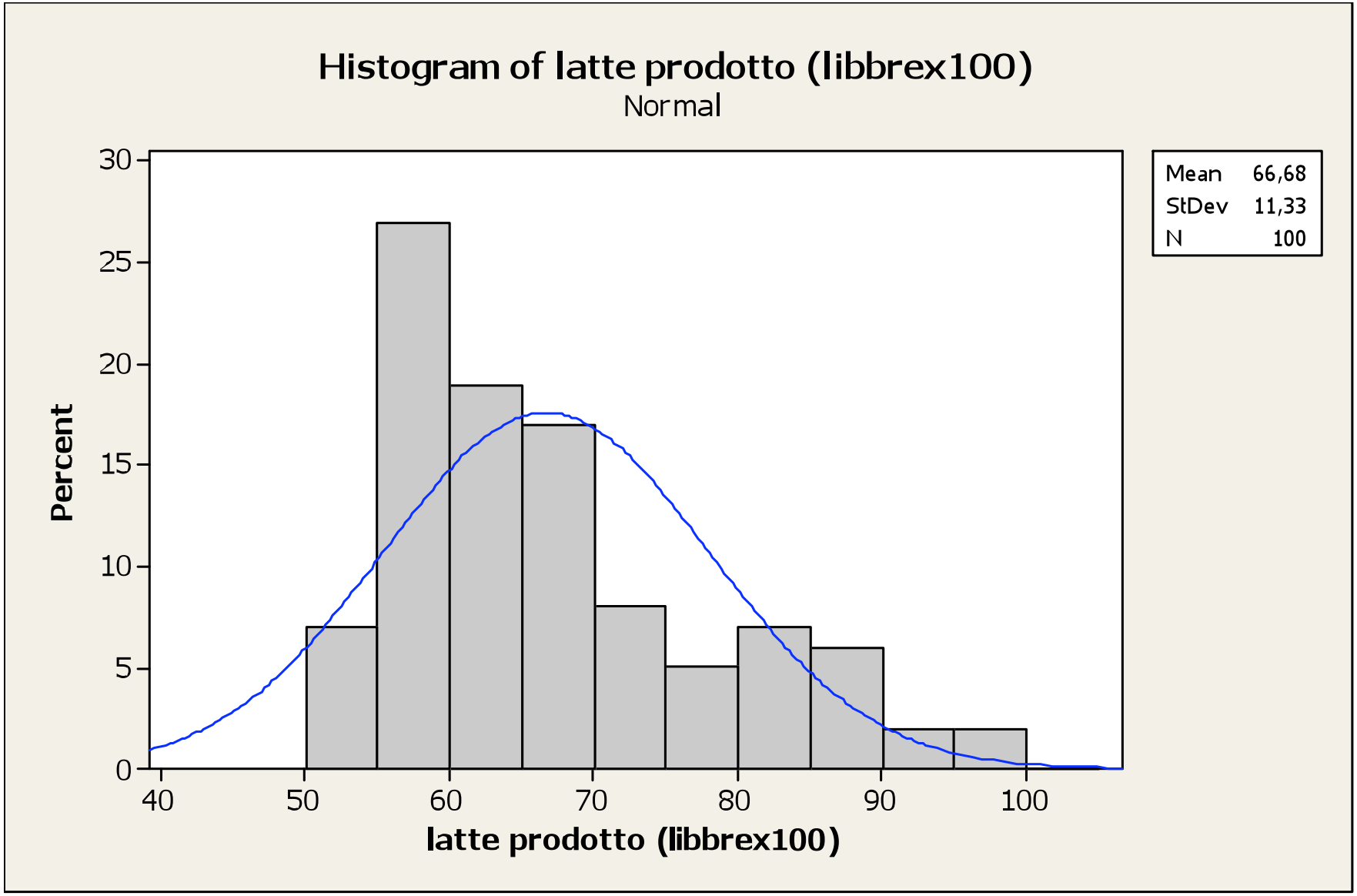
Plot dei quantili normali. Es.: campione di 13 dati estratto da una distribuzione normale standard

x	percentili	y=z
-1.75761	0.0769	-1.42624
-0.35848	0.1538	-1.02027
-0.00063	0.2308	-0.73621
0.04745	0.3077	-0.50238
0.06633	0.3846	-0.29342
0.23864	0.4615	-0.09666
0.60580	0.5385	0.09666
0.69193	0.6154	0.29342
0.81182	0.6923	0.50238
0.86228	0.7692	0.73621
1.21230	0.8462	1.02027
1.26512	0.9231	1.42624
1.42839	1.0000	*

Plot dei quantili normali. Es.: campione di 13 dati estratto da una distribuzione normale standard

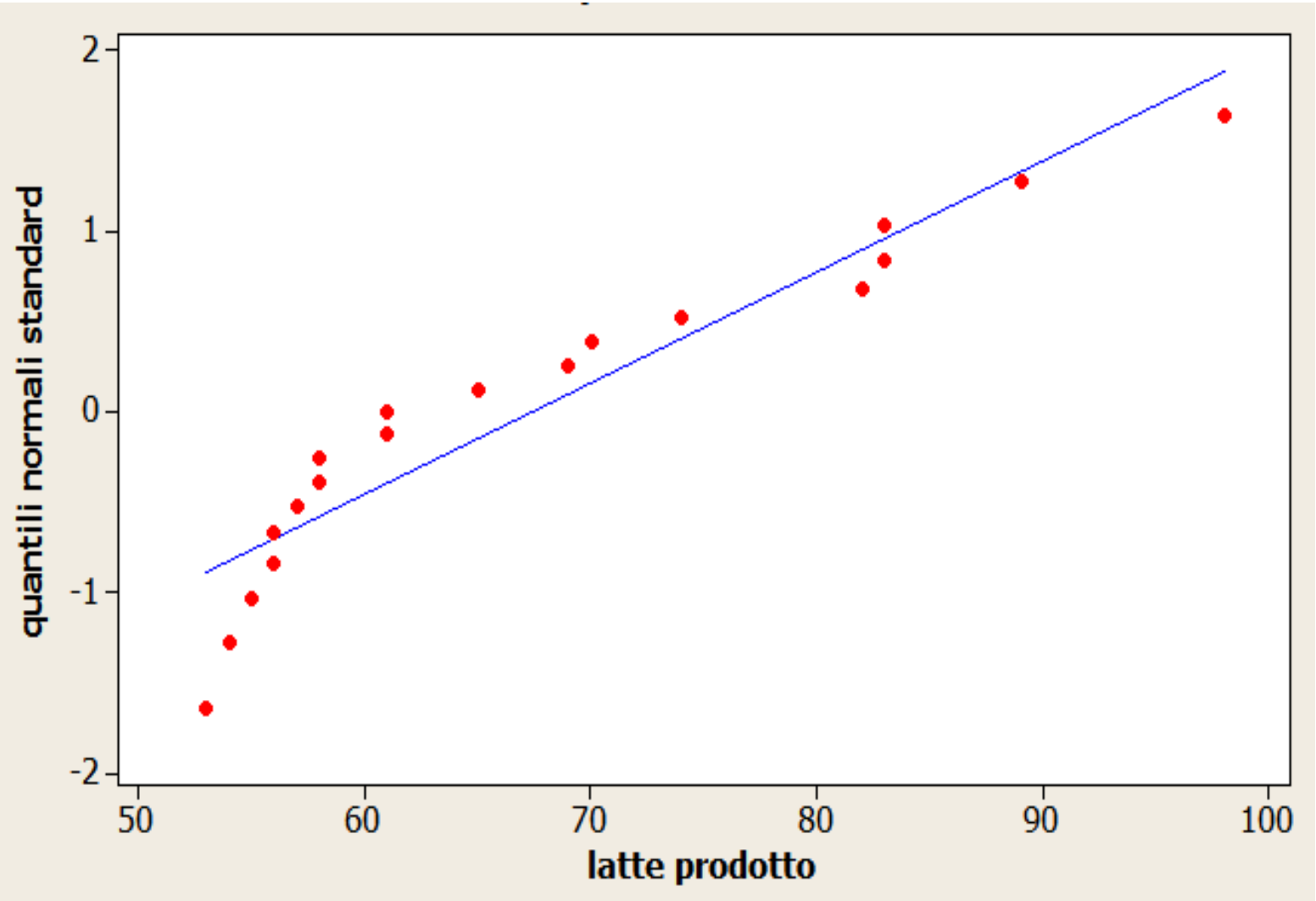


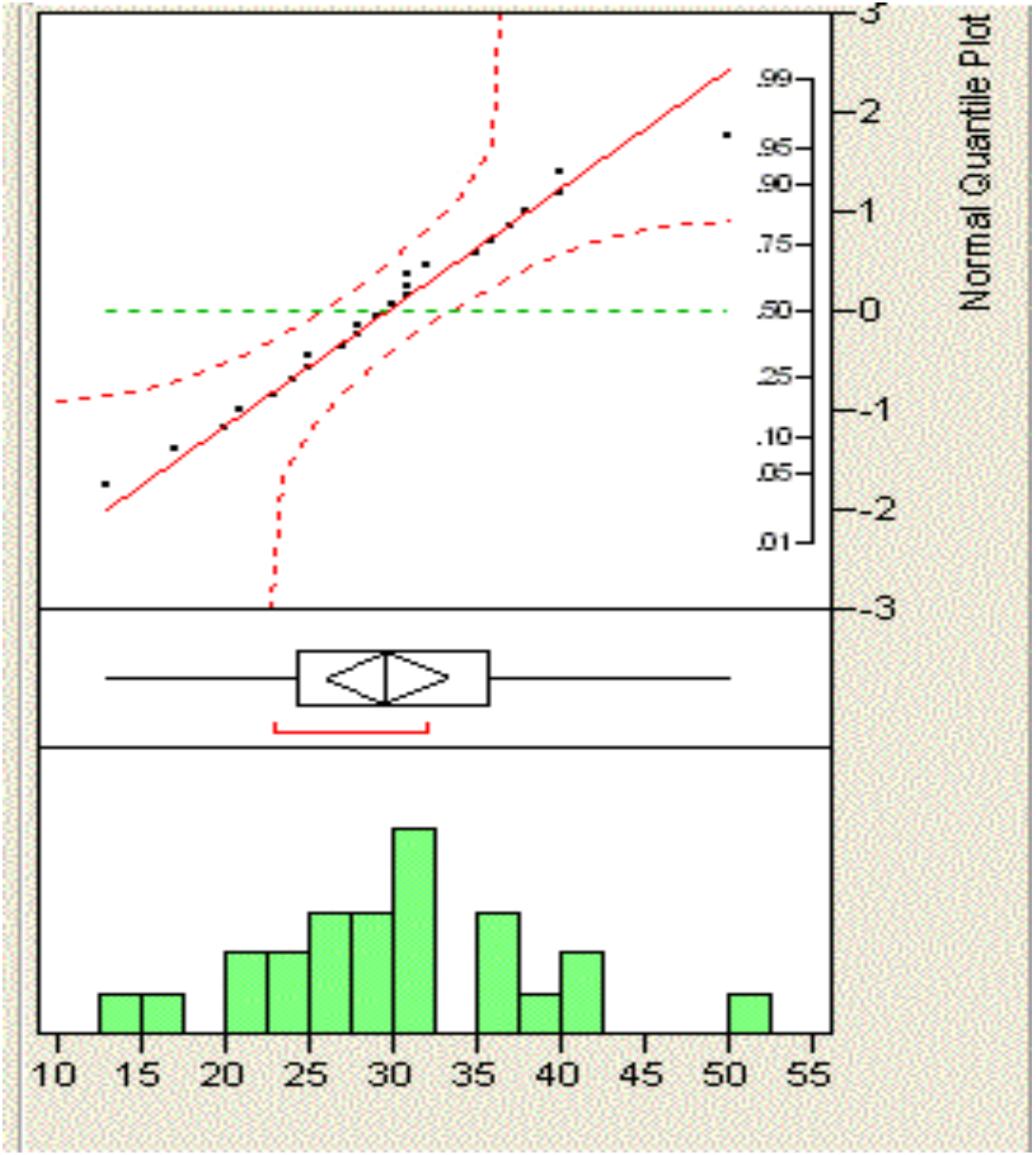
Produzione annuale di latte (libbre x 100) di 100 vacche Jersey di 2 anni (Sokal, Rohlf p.104)



Plot dei quantili normali
Diagramma di dispersione per 19 valori estratti dai 100

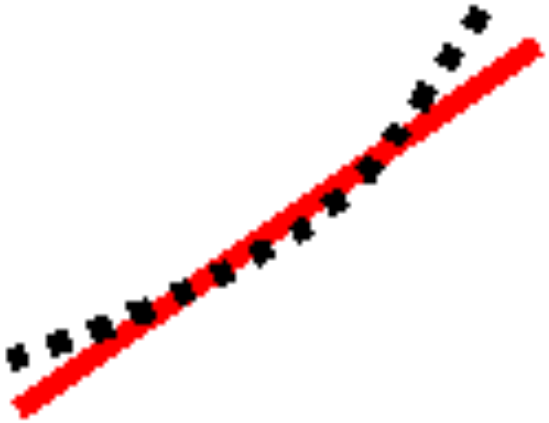
53
54
55
56
56
57
58
58
61
61
65
69
70
74
82
83
83
89
98





La variabilità naturale dei dati causa le fluttuazioni dei dati intorno alla Retta

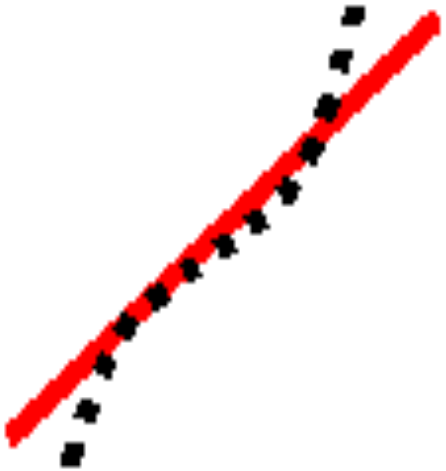
Per gli stessi dati sotto sono riportati il box-plot e l'istogramma



I dati presentano una lunga coda a sinistra

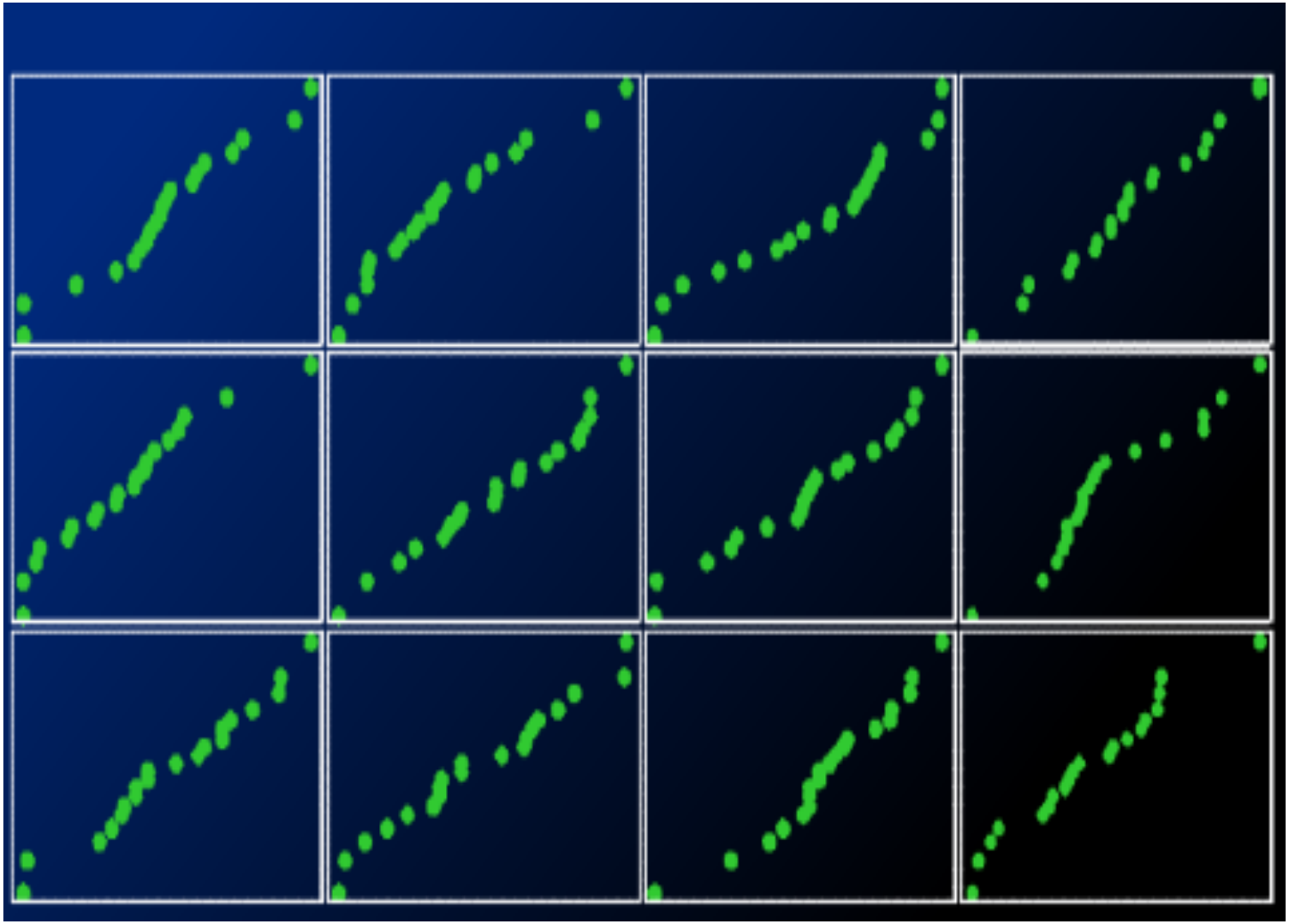
I dati presentano una lunga coda a destra

Code corte rispetto alla normale, ossia varianza minore rispetto a una normale



Code lunghe rispetto alla normale, ossia varianza maggiore rispetto a una normale

Plot dei quantili normali per campioni di dimensione 20 generati da una distribuzione normale

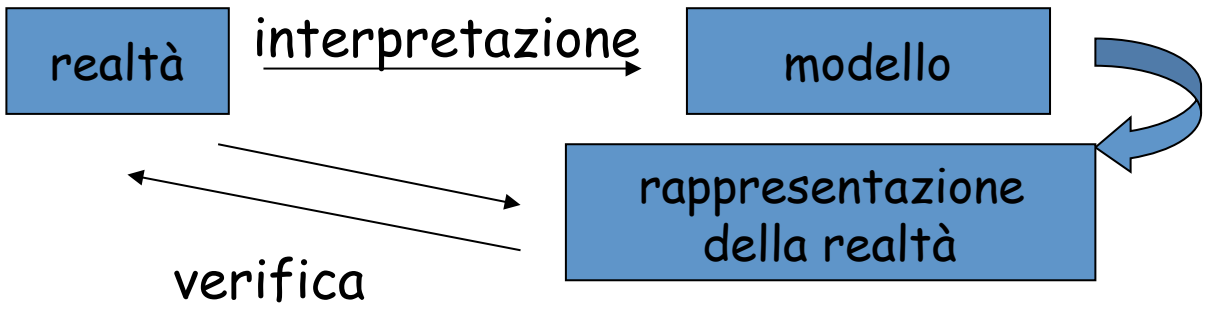


Modelli matematici

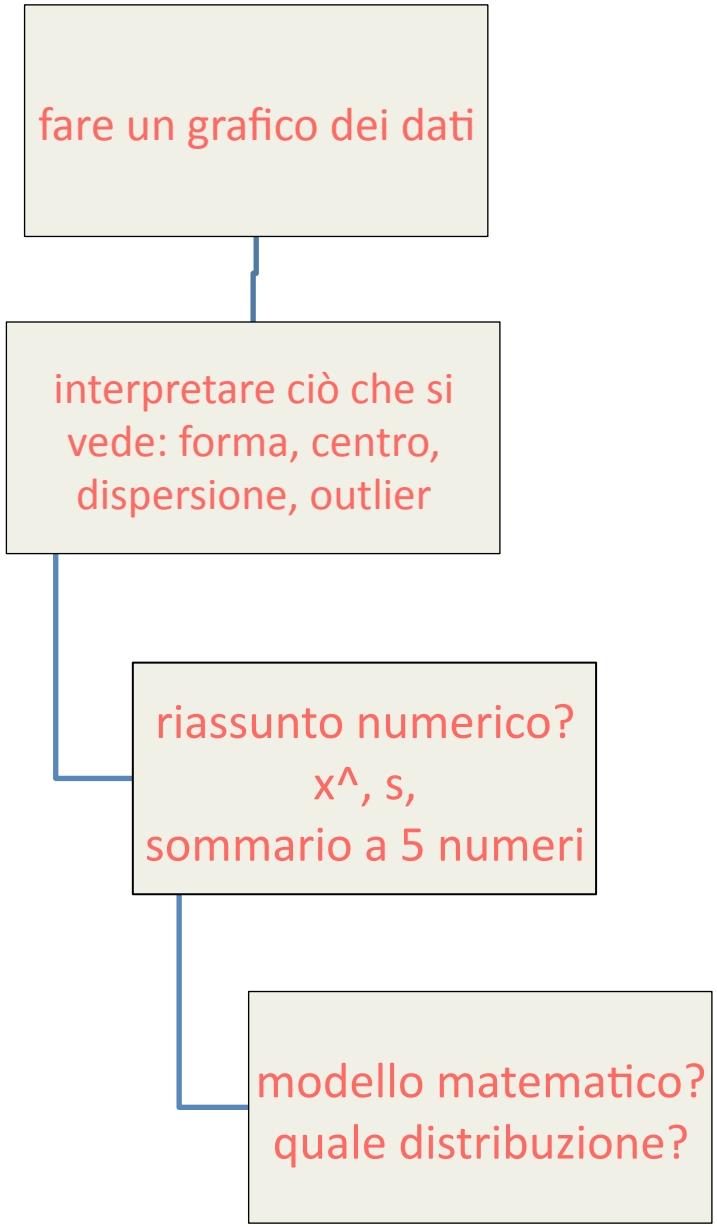
- Le distribuzioni di probabilità (con le loro formule) sono modelli matematici adatti a descrivere molti fenomeni naturali.
- Sono distribuzioni di frequenze teoriche per le popolazioni che forniscono una **rappresentazione idealizzata** dei fenomeni stessi. Riportano un'immagina compatta del modello complessivo dei dati, prescindendo da irregolarità minori.
- E' possibile incontrare distribuzioni di probabilità, generate empiricamente, che non possono essere descritte dai modelli noti.

I modelli

- In sostanza, la creazione di un modello inizia con lo studio del fenomeno nella realtà; le osservazioni derivanti dallo studio vengono interpretate per cogliere gli aspetti più importanti del fenomeno. Poi, si costruisce il modello, lo si fa "funzionare" e si controlla se e quanto i risultati ottenuti corrispondono con la realtà. Poi il modello viene riconsiderato e modificato per renderlo più efficiente, e così via



L'analisi dei dati con una variabile

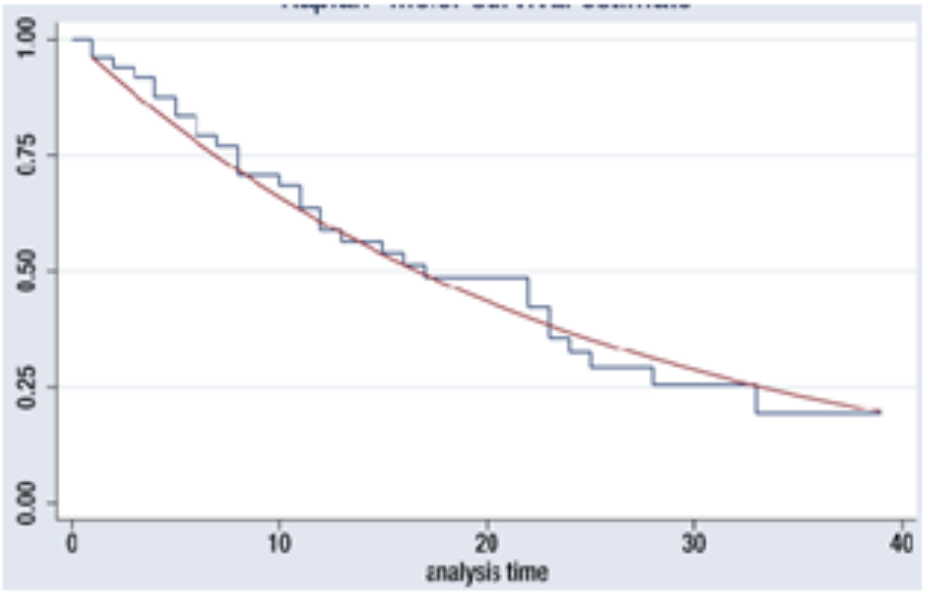


La distribuzione esponenziale come modello continuo per i fenomeni descritti dalla variabile aleatoria tempo di sopravvivenza

Anche in questo caso parliamo di curva di densità esponenziale

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

$x \geq 0$



. 5.4 Tempo di sopravvivenza in mesi per pazienti affetti da anemia aplastica grave e trattati con androgeni.