Seminari di didattica della matematica

Troppo bello per essere vero!

Dario Benedetto

Professore associato in Fisica Matematica, *Sapienza*

Genetica / Statistica

Genetica / Statistica

Mendel, il padre della genetica, ha “taroccato” i dati dei suoi esperimenti?

Genetica / Statistica

Mendel, il padre della genetica, ha “taroccato” i dati dei suoi esperimenti?

Fisher, uno dei padri della Statistica moderna, disse a proposito dei risultati di Mendel

“troppo belli per essere veri!”

Genetica / Statistica

Mendel, il padre della genetica, ha "taroccato" i dati dei suoi esperimenti?

Fisher, uno dei padri della Statistica moderna, disse a proposito dei risultati di Mendel

"troppo belli per essere veri!"

E perché uno dei padri della statistica si interessa delle leggi di Mendel?

la prospettiva storica I

Nelle moderne discipline probabilistico-statistiche confluiscono le ricerce che si sviluppano dal XVII secolo su

- gioco d'azzardo
- assicurazioni
- demografia

Nel XX secolo probabilità e statistica si sviluppano (anche) nell'interazione con la **genetica**.

la statistica e la divina provvidenza

J. Arbuthnot: *An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes* Philos. Transac. Royal Soc. 1710

Nota che per 82 anni consecutivi a Londra sono nati più maschi che femmine, dunque **sospetta che sia falsa** l'ipotesi che la probabilità di nascere di maschi e femmine sia la stessa.

Come analizzò questo fenomeno?

la statistica e la divina provvidenza

Se l'ipotesi fosse vera, la probabilità che in un anno nascano più maschi che femmine sarebbe (circa) 1/2.
Arbuthnot si chiese con quale probabilità, assumendo vera l'ipotesi $P(M) = P(F)/1/2$ si può osservare il fenomeno che si è verificato.

la statistica e la divina provvidenza

Se l'ipotesi fosse vera, la probabilità che in un anno nascano più maschi che femmine sarebbe (circa) 1/2.

Arbuthnot si chiese con quale probabilità, assumendo vera l'ipotesi $P(M) = P(F)/1/2$ si può osservare il fenomeno che si è verificato.

$P($ per 82 anni $M > F = 1/2^{82} \simeq 2 \times 10^{-25}$

la statistica e la divina provvidenza

Se l'ipotesi fosse vera, la probabilità che in un anno nascano più maschi che femmine sarebbe (circa) 1/2.
Arbuthnot si chiese con quale probabilità, assumendo vera l'ipotesi $P(M) = P(F)/1/2$ si può osservare il fenomeno che si è verificato.

$P($ per 82 anni $M > F = 1/2^{82} \simeq 2 \times 10^{-25}$

Arbuthnot ne deduce che la divina provvidenza compensa così la maggiore mortalità giovanile dei maschi rispetto alle femmine.

qualche osservazione

- Arbuthnot **non trova** la probabilità che l'ipotesi sia vera
- Arbuthnot calcola la probabilità di un evento "ipotetico" (quello reale si è verificato, dunque non c'è niente da calcolare)
- la procedura adottata somiglia al metodo scientifico: si osserva, si formula una teoria, se ne deducono delle conseguenze, si verificano le predizioni delle conseguenze.

genotipi e fenotipi

le caratteristiche degli individui (il **fenotipo**) sono ereditarie

genotipi e fenotipi

le caratteristiche degli individui (il **fenotipo**) sono ereditarie

il fenotipo è determinato dal **genotipo**, cioè dai geni presenti nei cromosomi

genotipi e fenotipi

le caratteristiche degli individui (il **fenotipo**) sono ereditarie

il fenotipo è determinato dal **genotipo**, cioè dai geni presenti nei cromosomi

ogni gene compare due volte, una nel cromosoma ereditato da un genitore, uno nel cromosoma *omologo* ereditato dall'altro genitore

genotipi e fenotipi

le caratteristiche degli individui (il **fenotipo**) sono ereditarie

il fenotipo è determinato dal **genotipo**, cioè dai geni presenti nei cromosomi

ogni gene compare due volte, una nel cromosoma ereditato da un genitore, uno nel cromosoma *omologo* ereditato dall'altro genitore

i geni si possono presentare in differenti versioni, le **varianti alleliche** che possono modificare il fenotipo

genotipi e fenotipi

le caratteristiche degli individui (il **fenotipo**) sono ereditarie

il fenotipo è determinato dal **genotipo**, cioè dai geni presenti nei cromosomi

ogni gene compare due volte, una nel cromosoma ereditato da un genitore, uno nel cromosoma *omologo* ereditato dall'altro genitore

i geni si possono presentare in differenti versioni, le **varianti alleliche** che possono modificare il fenotipo

il genotipo è dunque la coppia degli alleli del gene presenti nei due cromosomi

un caso semplice, il gene biallelico

indico con A e a due possibili varianti di un gene

i genotipi possibili sono 3: AA, Aa, aa
(Aa e aA sono indistinguibili)

un caso semplice, il gene biallelico

indico con A e a due possibili varianti di un gene

i genotipi possibili sono 3: AA, Aa, aa
(Aa e aA sono indistinguibili)

spesso un allele **domina** sull'altro e determina il fenopito;

se A domina su a:

genotipo		fenotipo
AA	omozigote dominante	A
Aa	eterozigote	A
aa	omozigote recessivo	a

un caso semplice, il gene biallelico

due esempi nelle piante dei piselli:

l'allele Y (seme giallo) è dominante sull'allele y (seme verde);

l'allele R (seme rotondo) è dominante sull'allele r (seme grinzoso)

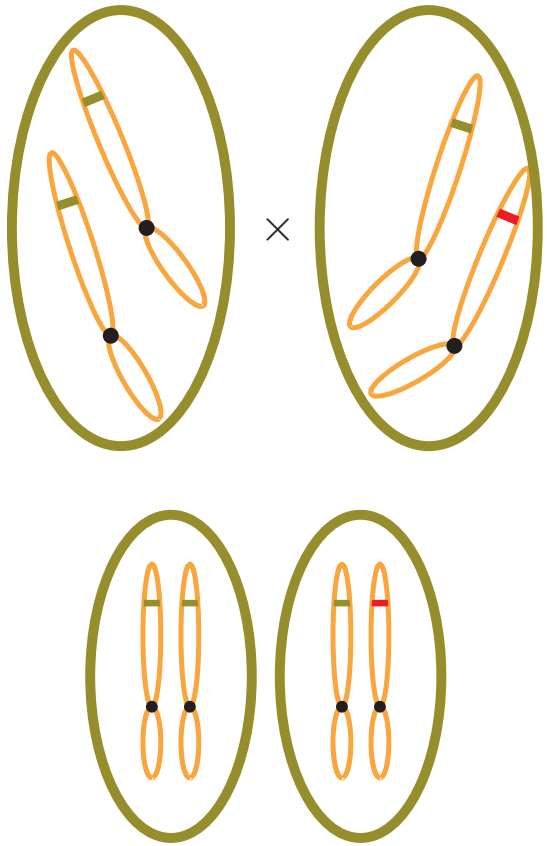
incrocio $YY \times Yy$

incrocio tra una pianta YY con una Yy (entrambe a semi gialli)

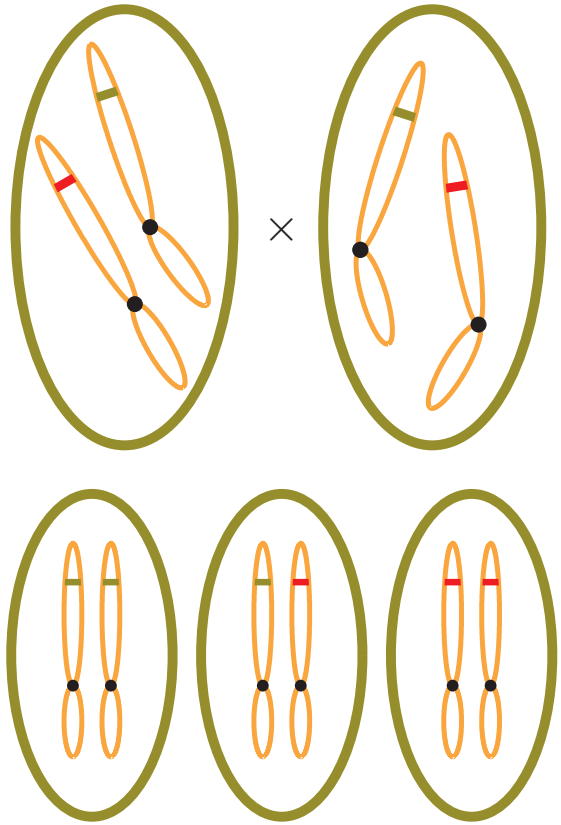
la pianta YY produce solo **gameti** Y
la pianta Yy produce metà gameti Y e metà y

solo due genotipi sono possibili, YY e Yy e hanno entrambi frequenza $1/2$, ma sono tutte a semi gialli

incrocio $YY \times Yy$



incrocio $Yy \times Yy$



incrocio $Yy \times Yy$

incrocio due piante Yy, entrambe a semi gialli

per entrambi i genitori, metà dei gameti sono Y, metà sono y

tutti i genotipi sono possibili

gen_2 \ gen_1	Y	y
Y	YY	Yy
y	yY	yy

gen_2 \ gen_1	$1/2$	$1/2$
$1/2$	$1/4$	$1/4$
$1/2$	$1/4$	$1/4$

incrocio $Yy \times Yy$

incrocio due piante Yy, entrambe a semi gialli

per entrambi i genitori, metà dei gameti sono Y, metà sono y

tutti i genotipi sono possibili

gen_2 \ gen_1	Y	y
Y	YY	Yy
y	yY	yy

gen_2 \ gen_1	$1/2$	$1/2$
$1/2$	$1/4$	$1/4$
$1/2$	$1/4$	$1/4$

dunque le frequenze dei genotipi sono

YY: $1/4$ Yy: $1/2$ yy: $1/4$

e le piante a seme giallo sono $1/4 + 1/2 = 3/4$

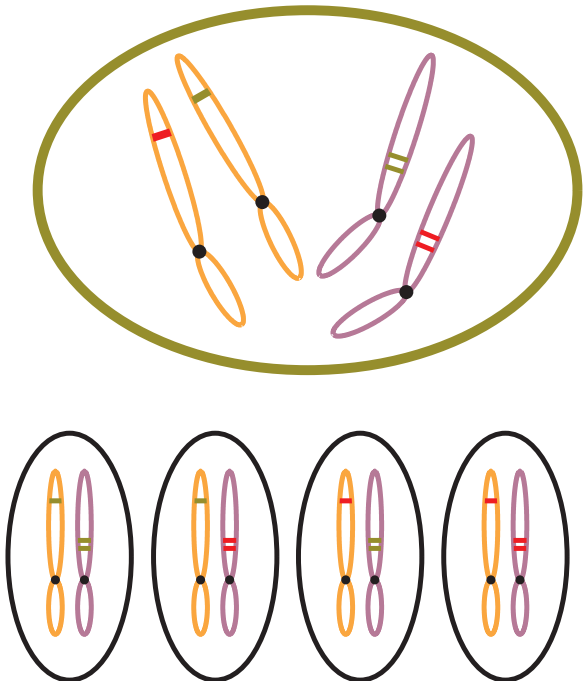
la legge di **segregazione**

Mendel selezionò delle piante **pure** per entrambi i due possibili colori del seme (generazione P)

le incrociò tra loro: $YY \times yy$ ottenendo la generazione F_1 di piante tutte a seme giallo, di genotipo Yy

autoincrociò le piante della generazione F_1 ottenendo la proporzione $1 : 3$ tra piante a semi gialli e piante a semi verdi

la legge dell'**assortimento indipendente**



la legge dell'**assortimento indipendente**

i geni per il colore e per l'aspetto del seme sono su due cromosomi diversi

consideriamo una pianta *YyRr*, a semi gialli e lisci, ma eterozigote per entrambi i caratteri

la pianta produce con uguali probabilità gameti dei 4 tipi *YR*, *Yr*, *yR*, *yr*

incrociamola con una pianta a semi verdi e grinzosi *rryy*, cioè eterozigote recessiva per entrambi i caratteri

la seconda produce solo gameti *yr*

la seconda legge di Mendel

i possibili risultati dell'incrocio sono equiprobabili e sono

$YR + yr = YyRr$	giallo e rotondo
$Yr + yr = Yyrr$	giallo e grinzoso
$yR + yr = yyRr$	verde e rotondo
$yr + yr = yyrr$	verde e grinzoso

in rapporto $1 : 1 : 1 : 1$

la prospettiva storica II

-
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
-
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
-
-
-
-
-

la prospettiva storica II

-
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
-
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
- 1902 Sutton-Boveri: teoria cromosomica
- 1915 T.H. Morgan et. al. *The Mechanism of Mendelian Inheritance*
-
-
-

la prospettiva storica II

-
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
-
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
- 1902 Sutton-Boveri: teoria cromosomica
- 1915 T.H. Morgan et. al. *The Mechanism of Mendelian Inheritance*
-
-
- 1953: J.Watson, F.Crick: la struttura del DNA

la prospettiva storica II

- 1859 C. Darwin *L'origine delle specie*
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
-
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
- 1902 Sutton-Boveri: teoria cromosomica
- 1915 T.H. Morgan et. al. *The Mechanism of Mendelian Inheritance*
-
-
- 1953: J.Watson, F.Crick: la struttura del DNA

la prospettiva storica II

- 1859 C. Darwin *L'origine delle specie*
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
- 1877 **F. Galton** "inventa" la retta di regressione studiando l'ereditarietà dell'altezza
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
- 1902 Sutton-Boveri: teoria cromosomica
- 1915 T.H. Morgan et. al. *The Mechanism of Mendelian Inheritance*
-
-
- 1953: J.Watson, F.Crick: la struttura del DNA

la prospettiva storica II

- 1859 C. Darwin *L'origine delle specie*
- 1865 G. Mendel *Verhandlungen des naturforschenden Vereins Brünn*
- 1877 **F. Galton** "inventa" la retta di regressione studiando l'ereditarietà dell'altezza
- 1900 le leggi di Mendel vengono riscoperte indipendentemente da 4 scienziati
- 1902 Sutton-Boveri: teoria cromosomica
- 1915 T.H. Morgan et. al. *The Mechanism of Mendelian Inheritance*
- 1918 **T. Fisher** *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*
- 1942: J. Huxley *Evolution: The Modern Synthesis*
- 1953: J.Watson, F.Crick: la struttura del DNA

analisi statistica degli esperimenti di Mendel

i dati di Mendel sono stati sottoposti a verifica statistica da Weldon nel 1902

descriviamo l'esperimento sul colore del baccello

ottiene 152 piante con baccello verde 428 piante con baccello giallo

la proporzione è circa $1:3$, che corrisponde a una distribuzione di $1/4$ del fenotipo recessivo e di $3/4$ del fenotipo dominante

la frazione esatta delle piante a baccello verde è

$$152/(152+428) = 152/580 \simeq 0.2621$$

analisi statistica degli esperimenti di Mendel

il valore misurato 0.2621 è sufficientemente vicino a 0.25, da avvalorare la teoria?

analisi statistica degli esperimenti di Mendel

il valore misurato 0.2621 è sufficientemente vicino a 0.25, da avvalorare la teoria?

l'approccio statistico non risponde a questa domanda

analisi statistica degli esperimenti di Mendel

il valore misurato 0.2621 è sufficientemente vicino a 0.25, da avvalorare la teoria?

l'approccio statistico non risponde a questa domanda

nell'approccio di Fisher, la statistica, attraverso i test, valuta la forza dell'evidenza dei dati **contro** l'ipotesi che si intende testare

i dati possono far rifiutare una teoria, ma non confermarla

analisi statistica degli esperimenti di Mendel

ipotesi da testare: la proporzione di piante a baccello verde è 1/4 del totale

(indicata con H_0, è l'*ipotesi zero*, spesso chiamata *ipotesi nulla*)

analisi statistica degli esperimenti di Mendel

ipotesi da testare: la proporzione di piante a baccello verde è 1/4 del totale

(indicata con H_0, è l'*ipotesi zero*, spesso chiamata *ipotesi nulla*)

la differenza tra 25% e il valore osservato 26.621% ci
può spingere a **rifiutare** H_0?

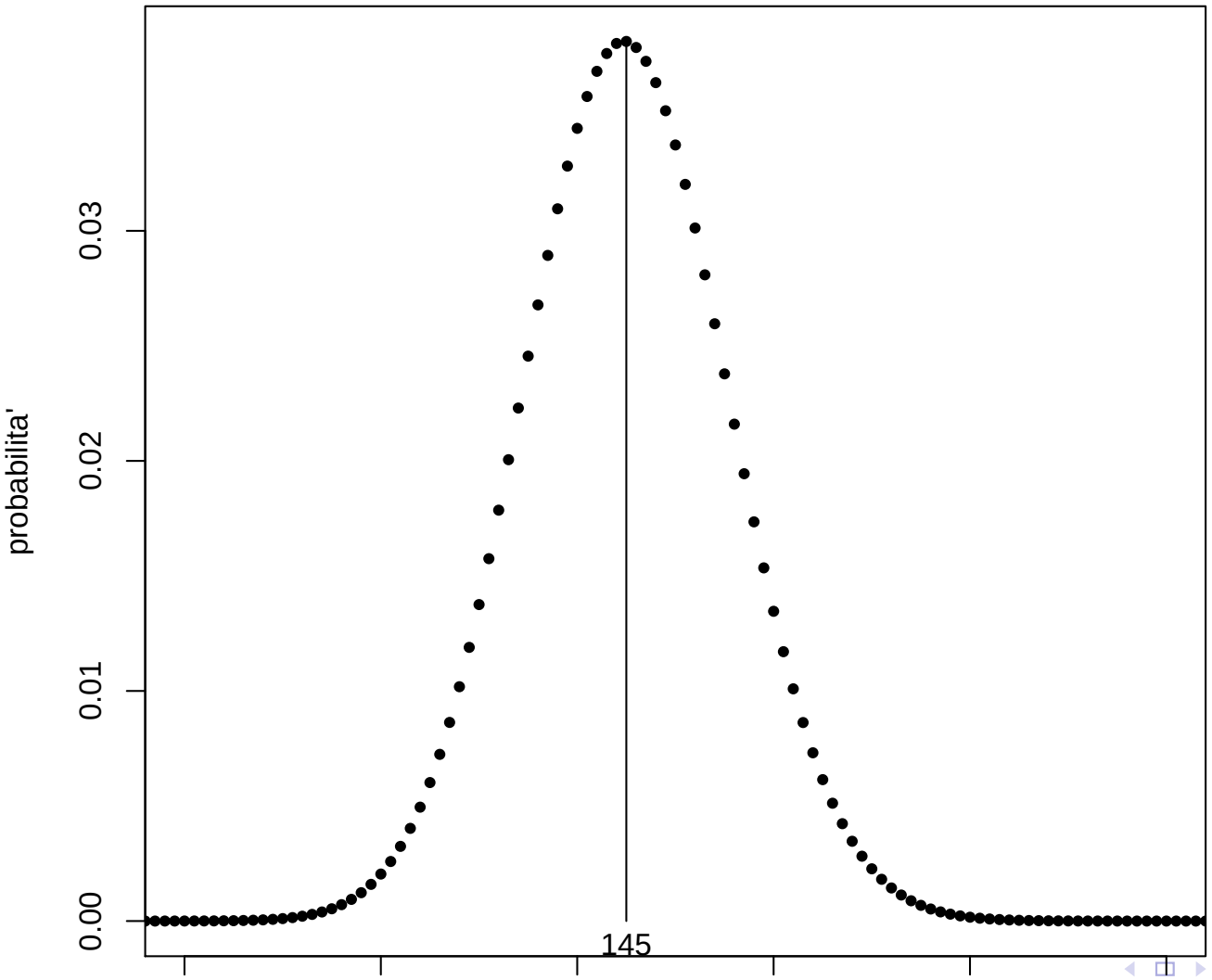
come si fa il test?

se l'ipotesi nulla fosse vera, il fenomeno osservato corrisponderebbe all'estrazione di una **variabile aleatoria binomiale**

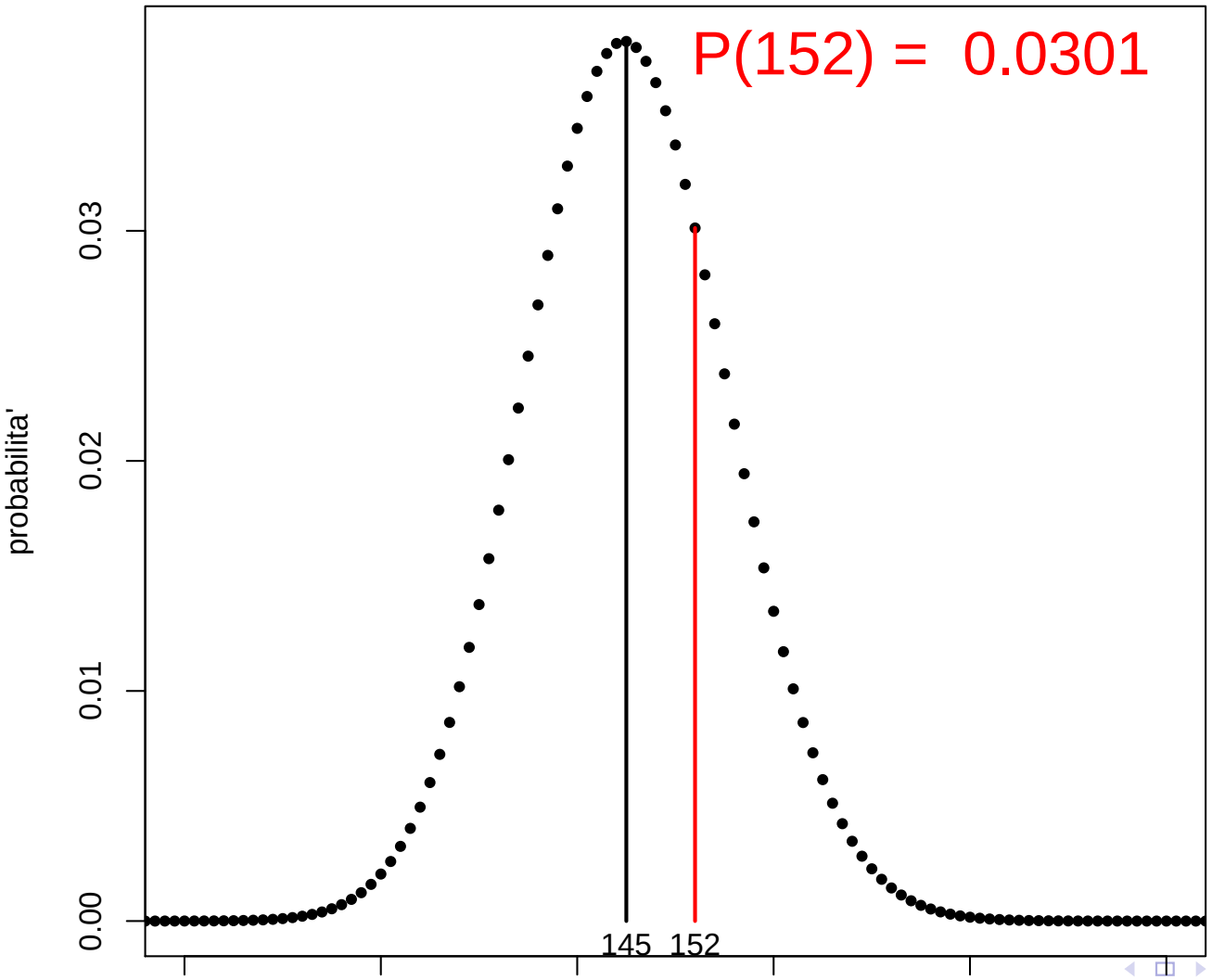
l'esperimento teorico di confronto è contare quante "teste" escono lanciando 580 volte una moneta truccata, per la quale la probabilità di testa sia 1/4.

il valore aspettato è $580/4 = 145$

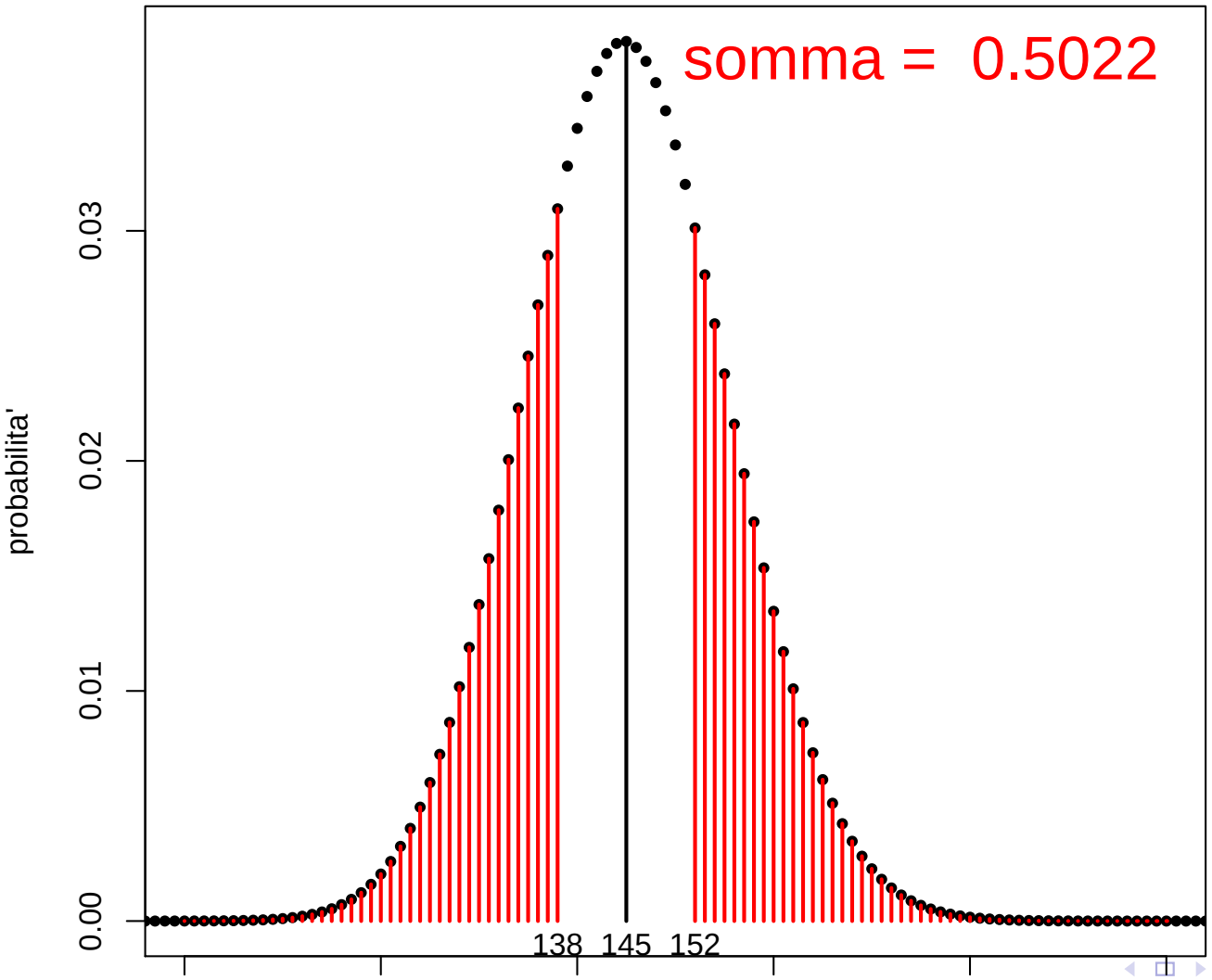
come si fa il test?



analisi statistica degli esperimenti di Mendel



analisi statistica degli esperimenti di Mendel



test binomiale esatto

H_0: la frazione di baccelli verdi è un quarto del totale

$7 = 152$ - 145 è la **statistica del test** e misura la distanza tra quanto osservato e il valore atteso

X variabile binomiale
$P(|X - 145| \geq 7) \simeq 0.5022$ è il $p-$**valore** del test

il $p-$valore

il $p-$valore del test **NON** è una misura della probabilità che H_0 sia vera

il $p-$valore

il $p-$valore del test **NON** è una misura della probabilità che H_0 sia vera

- è la probabilità di vedere uno scostamento pari o peggiore di quello osservato, nel caso H_0 fosse vera

il $p-$valore

il $p-$valore del test **NON** è una misura della probabilità che H_0 sia vera

- è la probabilità di vedere uno scostamento pari o peggiore di quello osservato, nel caso H_0 fosse vera
- è la probabilità di vedere lo scostamento osservato per una fluttuazione statistica (cioè per caso), ipotizzando che H_0 sia vera

il $p-$valore

il $p-$valore del test **NON** è una misura della probabilità che H_0 sia vera

- è la probabilità di vedere uno scostamento pari o peggiore di quello osservato, nel caso H_0 fosse vera
- è la probabilità di vedere lo scostamento osservato per una fluttuazione statistica (cioè per caso), ipotizzando che H_0 sia vera

dunque solo se il $p-$value fosse molto piccolo sarebbe lecito dubitare di H_0

che vuol dire piccolo?

H_0: proporzione 1/4

fisso la percentuale a 26.21% e cambio il numero n delle piante osservate

n	k	$p-$value	giudizio sullo scostamento
580	152	0.5022	non significativo
2000	524	0.2153	non significativo
4000	1048	0.08281	non significativo
8000	2097	0.01271	**significativo**
12000	3145	0.002397	**molto significativo**
16000	4194	0.0004252	**estremamente significativo**

che vuol dire piccolo?

è una valutazione convenzionale

$p > 0.05$	non significativo
$0.01 < p \leq 0.05$	significativo
$0.001 < p \leq 0.01$	molto significativo
$p \leq 0.001$	estremamente significativo

per la scoperta del bosone di Higgs, è stato osservato un fenomeno che, se fosse stato effetto di del caso, avrebbe avuto probabilità 10^{-6}

che vuol dire piccolo?

n	k	$p-$value	giudizio sullo scostamento
580	152	0.5022	non significativo
2000	524	0.2153	non significativo
4000	1048	0.08281	non significativo
8000	2097	0.01271	**significativo**
12000	3145	0.002397	**molto significativo**
16000	4194	0.0004252	**estremamente significativo**

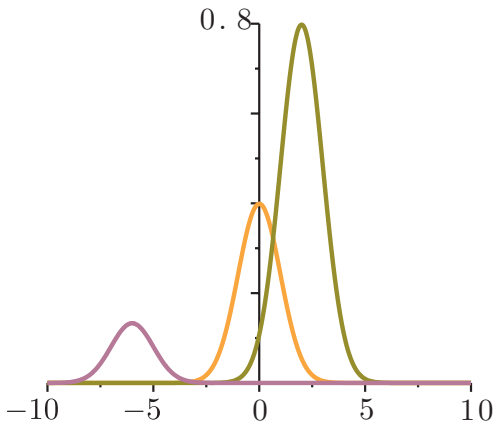
a parità di percentuale, è ragionevole aspettarsi un peggioramento di p all'aumentare di n

le variabili aleatorie normali

una variabile aleatoria **normale** o **gaussiana** ha densità di probabilià

$$N_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-(x-m)^2/(2\sigma^2)}$$

dove m è il suo valor medio e σ^2 la sua varianza



$m = -6, \sigma = 3, \quad m = 0, \sigma = 1, \quad m = 2, \sigma = 0.5$

le variabili aleatorie normali

σ è anche la distanza dei flessi da m e misura quanto la distribuzione è concentrata intorno alla media

indicherò con Z una variabile normale **standard**, cioè di media nulla e varianza 1.

il teorema del limite centrale

Sia X è la variabile binomiale con probabilità di successo p, e n è il numero di prove; la sua varianza è $np(1-p)$ e vale

$$\frac{X - pn}{\sqrt{np(1-p)}} \approx Z$$

cioè per n grande si ottiene una variabile gaussiana di media nulla e varianza 1

dunque $X/n - p$ è asintoticamente equivalente a una variabile gaussiana di media nulla e deviazione standard $\sqrt{\frac{p(1-p)}{n}}$

torinamo alla tabella

n	k	$p-$value	giudizio sullo scostamento
580	152	0.5022	non significativo
2000	524	0.2153	non significativo
4000	1048	0.08281	non significativo
8000	2097	0.01271	**significativo**
12000	3145	0.002397	**molto significativo**
16000	4194	0.0004252	**estremamente significativo**

al crescere di n la deviazione standard diminuisce come $1/\sqrt{n}$,

dunque osservare una deviazione dalla media di 0.2621 - 0.25 = 0.0121

diventa via via più improbabile!

alcuni esperimenti sull'indipendenza

dall'incrocio $RrYy \times rryy$ mi aspetto la proporzione $1:1:1:1$

seme giallo e rotondo	47
seme giallo e grinzoso	40
seme verde e rotondo	38
seme verde e grinzoso	41
totale	166

quando si **adattano** questi valori alla distribuzione esatta in cui tutte le frequenze sono uguali a $166/4 = 41.5$?

test di adattamento del $\mathcal{X}^2$

dobbiamo considerare una distribuzione **multinomiale**:

per $i = 1, \ldots n$ sia p_i la probabilità che accada l'evento e_i
(per esempio con prob. $1/4$ si ottiene uno dei 4 fenotipi del seme)

dopo n lanci, si osserva n_i volte l'evento e_i (n_i è la frequenza dell'evento e_i)

il valore atteso di n_i è np_i; come valuto complessivamente la distanza tra n_i e np_i?

test di adattamento del $\mathcal{X}^2$

teorema: per n grande la variabile

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

è distribuita come la somma dei quadrati di $k-1$ variabili gaussiane standardizzate indipendenti

questa somma è una variabile aleatoria che si chiama **chi quadro** a $k-1$ gradi di libertà, e si indica con

$$\mathcal{X}^2_{k-1}$$

test di adattamento del $\mathcal{X}^2$

- è la somma di k variabili, ma la somma delle n_i è n, dunque solo $k-1$ sono variabili indipendenti
- $(n_i - np_i)/\sqrt{np_i(1-p_i)}$ è circa gaussiana, dunque ogni $(n_i - np_i)^2/(np_i(1-p_i))$ è circa Z^2, ma nel test compare $(n_i - np_i)^2/(np_i)$

per $k = 2$ vale l'identità algebrica

$$\frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} = \frac{(n_1 - np_1)^2}{np_1(1-p_1)} \approx Z^2 = \mathcal{X}_1^2$$

analisi degli esperimenti sull'indipendenza

H_0: le percentuali dei 4 differenti fenotipi sono tutte pari a 1/4

statistica del test:

$$\left[(47-41.5)^2+(40-41.5)^2+(47-38)^2+(41-41.5)^2\right]/41.5 \simeq 1.084$$

$p-$valore del test:

$$P(\mathcal{X}_3^2 > 1.084) \simeq 0.78$$

analisi degli esperimenti sull'indipendenza

H_0: le percentuali dei 4 differenti fenotipi sono tutte pari a 1/4

statistica del test:

$$\left[(47-41.5)^2+(40-41.5)^2+(47-38)^2+(41-41.5)^2\right]/41.5 \simeq 1.084$$

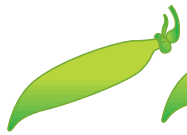
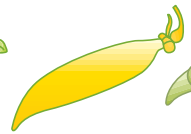
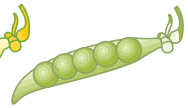
$p-$valore del test:

$$P(\mathcal{X}_3^2 > 1.084) \simeq 0.78$$

facendo molte prove, nel 78% dei casi si vedrebbero scostamenti dalla distribuzione teorica peggiori di quello osservato

lo scostamento osservato non smentisce l'ipotesi H_0

i risultati di Mendel

Carattere	Colore del seme	Aspetto del seme	Colore del fiore	Posizione del fiore	Altezza del fusto	Colore del baccello	Forma del baccello
Dominante	 Giallo	 Liscio	 Rosso	 Assiale	 Alto	 Verde	 Semplice
Recessivo	 Verde	 Rugoso	 Bianco	 Terminale	 Basso	 Giallo	 Concamerata

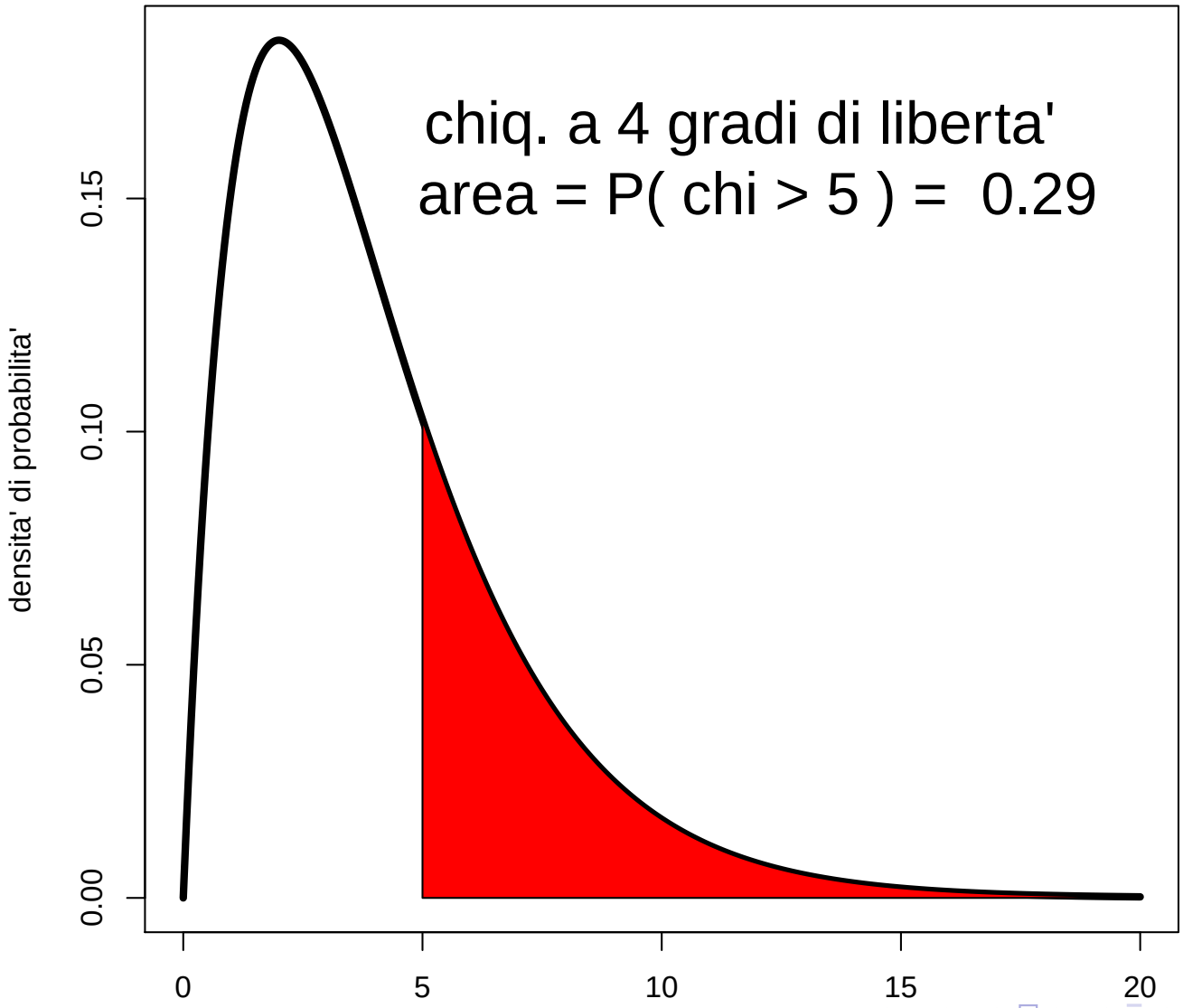
i dati di Mendel

il rapporto 3 : 1

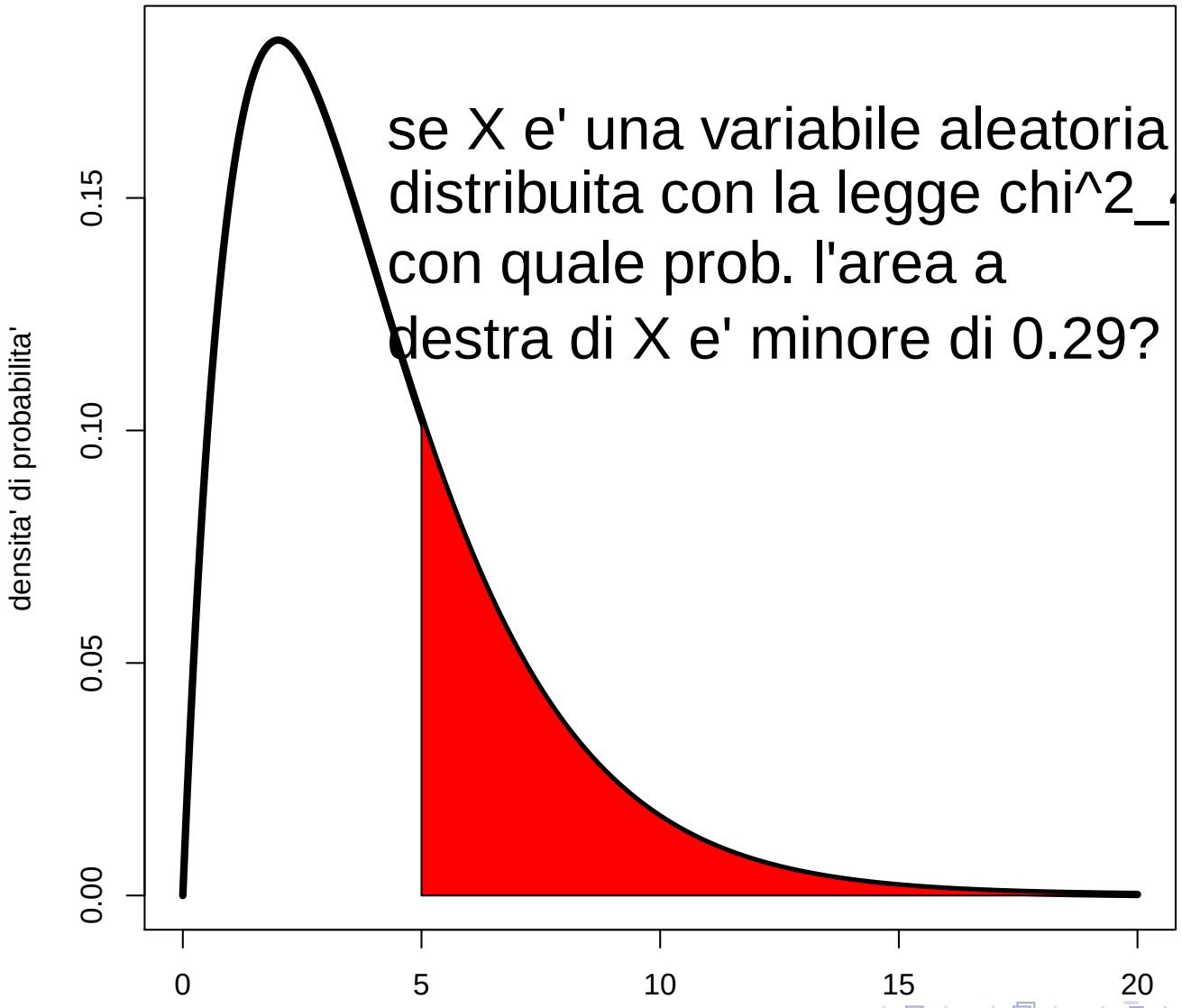
carattere	dominante	recessivo	$p-$value
forma del seme	5474	1850	0.608
colore del seme	6022	2001	0.908
colore del fiore	705	224	0.544
forma del baccello	882	299	0.814
colore del baccello	428	152	0.502
posizione del fiore	651	207	0.581
altezza del fusto	787	277	0.436

come dovrebbe essere distribuito il $p-$valore se vale l'ipotesi nulla?

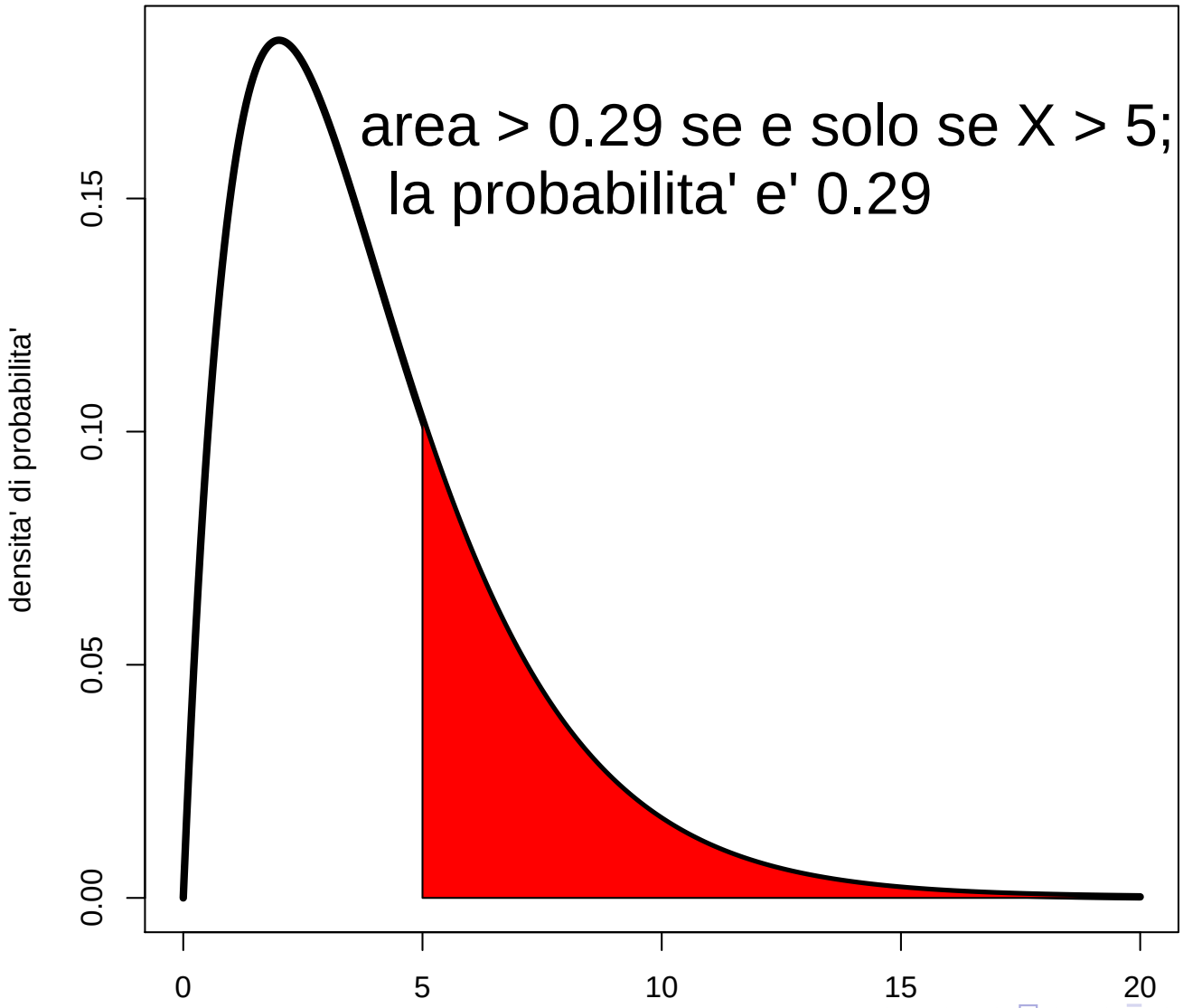
la distribuzione del p—value



la distribuzione del p—value



la distribuzione del p—value



la distribuzione del $p-$value

se H_0 è vera, e l'esperimento viene ripetuto molte volte, il $p-$value deve essere distribuito come una variabile aleatoria **uniforme** in $[0, 1]$

un indizio negli esperimenti sulla prima legge

carattere	dominante	recessivo	$p-$value
forma del seme	5474	1850	0.608
colore del seme	6022	2001	0.908
colore del fiore	705	224	0.544
forma del baccello	882	299	0.814
colore del baccello	428	152	0.502
posizione del fiore	651	207	0.581
altezza del fusto	787	277	0.436

un indizio negli esperimenti sulla prima legge

carattere	dominante	recessivo	$p-$value
forma del seme	5474	1850	0.608
colore del seme	6022	2001	0.908
colore del fiore	705	224	0.544
forma del baccello	882	299	0.814
colore del baccello	428	152	0.502
posizione del fiore	651	207	0.581
altezza del fusto	787	277	0.436

sei $p-$value su sette sono maggiori di $1/2\ldots$

quanto è probabile vedere un risultato come questo, o più estremo di questo, scegliendo 7 numeri a caso uniformemente in $[0, 1]$?

un indizio negli esperimenti sulla prima legge

carattere	dominante	recessivo	$p-$value
forma del seme	5474	1850	0.608
colore del seme	6022	2001	0.908
colore del fiore	705	224	0.544
forma del baccello	882	299	0.814
colore del baccello	428	152	0.502
posizione del fiore	651	207	0.581
altezza del fusto	787	277	0.436

sei $p-$value su sette sono maggiori di $1/2\ldots$

quanto è probabile vedere un risultato come questo, o più estremo di questo, scegliendo 7 numeri a caso uniformemente in $[0, 1]$?

$$7 \times 1/2^7 + 1/times1/2^7 = 0.0625$$

al pelo della significatività statistica!

raffiniamo l'analisi

carattere	dominante	recessivo	$p-$value
forma del seme	5474	1850	0.608
colore del seme	6022	2001	0.908
colore del fiore	705	224	0.544
forma del baccello	882	299	0.814
colore del baccello	428	152	0.502
posizione del fiore	651	207	0.581
altezza del fusto	787	277	0.436

con quale probabilità una variabile uniforme e' per 7 volte maggiore o uguale a 0.436?

$$(1 - 0.436)^7 = 0.018$$

significativamente improbabile!

secondo indizio...

esperimenti sui rapporti $1:1:1:1$

				$p-$valore
20	23	25	22	0.9015
31	26	27	26	0.8923
25	19	22	21	0.8346
24	25	22	27	0.9121
47	40	38	41	0.7809

con quale probabilità 5 numeri estratti uniformemente in $[0,1]$ risultano tutti maggiori di 0.78?

$$(1-0.78)^5 \approx 0.0005 = 1/2000$$

estremamente improbabile!

l'argomento di Fisher

I dati di Mendel erano stati sottoposti a verifica statistica da Weldon (zoologo e statistico) nel 1902, usando il test del $\mathcal{X}^2$, inventato da Pearson.

Fisher nel 1936 li rianalizza e fa un test finale: somma tutte le statistiche dei test di adattamento e confronta il valore ottenuto con la distribuzione $\mathcal{X}^2$ a 86 gradi di libertà: ottiene un $p-$valore di

$$7/100\ 000$$

l'argomento di Fisher

I dati di Mendel erano stati sottoposti a verifica statistica da Weldon (zoologo e statistico) nel 1902, usando il test del $\mathcal{X}^2$, inventato da Pearson.

Fisher nel 1936 li rianalizza e fa un test finale: somma tutte le statistiche dei test di adattamento e confronta il valore ottenuto con la distribuzione $\mathcal{X}^2$ a 86 gradi di libertà: ottiene un $p-$valore di

$$7/100\ 000$$

un buon motivo per rigettare l'ipotesi nulla che i dati degli esperimenti di Mendel sono corretti e concludere che sono troppo belli per essere veri!

una possibile spiegazione

nel caso dei dati sulla legge di segregazione, forse Mendel ha fatto due volte ogni esperimento, scegliendo il risultato migliore dei due

una possibile spiegazione

nel caso dei dati sulla legge di segregazione, forse Mendel ha fatto due volte ogni esperimento, scegliendo il risultato migliore dei due

in tal caso, come dovrebbe essere distribuito il $p-$valore?

una possibile spiegazione

nel caso dei dati sulla legge di segregazione, forse Mendel ha fatto due volte ogni esperimento, scegliendo il risultato migliore dei due

in tal caso, come dovrebbe essere distribuito il $p-$valore?

come il massimo tra due numeri scelti uniformemente in $[0, 1]$

una possibile spiegazione

se X e Y sono uniformi e indipendenti in $[0,1]$

$$P(\max(X,Y) < a) = P(X < a \text{ e } Y < a) = P(X < a) \cdot P(Y < a) = a^2$$

ma vale anche

$$P(\sqrt{X} < a) = P(X < a^2) = a^2$$

una possibile spiegazione

se X e Y sono uniformi e indipendenti in $[0, 1]$

$$P(\max(X, Y) < a) = P(X < a \text{ e } Y < a) = P(X < a) \cdot P(Y < a) = a^2$$

ma vale anche

$$P(\sqrt{X} < a) = P(X < a^2) = a^2$$

dunque il $p-$valore è distribuito come la radice quadrata di una variabile uniforme e quindi il quadrato del $p-$valore deve essere uniforme

una possibile spiegazione

testiamo l'ipotesi

carattere	dominante	recessivo	$p-$value	$p-$value^2
forma del seme	5474	1850	0.608	0.370
colore del seme	6022	2001	0.908	0.824
colore del fiore	705	224	0.544	0.297
forma del baccello	882	299	0.814	0.663
colore del baccello	428	152	0.502	0.252
posizione del fiore	651	207	0.581	0.338
altezza del fusto	787	277	0.436	0.190

la probabilità che 7 numeri scelti uniformemente in $[0, 1]$ siano maggiori di 0.19 è

$$(1 - 0.19)^7 \simeq 0.23$$

valore che non permette di rifiutare l'ipotesi che abbiamo fatto sul modo di operare di Mendel

Fonti

• D. Costantini **I Fondamenti storico-filosofici delle discipline statistico-probabilistiche** Bollati Boringhieri 2004 isbn 88-339-1528-x

• Griffiths et al. **Genetica - principi di analisi formale** quinta ed. 2002 Zanichelli isbn 88-08-08909-6

• Benedetto, Degli Esposti, Maffei **Matematica per le scienze della vita** seconda ed. 2012 CEA isbn 978-88-08-28238-5

• Pires A.M. Branco J.A. *A Statistical Model to Explain the Mendel-Fisher Controversy* Stat. Sci. 25 (4) 545–565 doi: 10.1214/10-STS342