Elementi di inferenza

Alessandra Nardi alenardi@mat.uniroma2.it

Lo schema della misurazione "precisa": quanto è veloce la luce?

Tra il 1879 ed 1882 Newcomb e Michelson condussero una serie di esperimenti che sono oggi considerati le prime misurazioni accurate sulla velocità della luce. Newcomb misurò il tempo necessario alla luce per percorrere la distanza tra il suo laboratorio ed uno specchio situato sul *Washington Monument* e per tornare all'origine, compiendo una distanza complessiva di 7.44373 km. I dati che seguono sono i 66 valori osservati da Newcomb

```
(28 22 36 26 28 28 26 24 32 30 27 24 33 21 36 32 31 25 24 25 28 36
27 32 34 30 25 26 26 25 -44 23 21 30 33 29 27 29 28 22 26 27 16 31
29 36 32 28 40 19 37 23 32 29 -2 24 25 27 24 16 29 20 28 27 39 23)
```

Il tempo effettivo misurato da Newcomb si ottiene a partire da questi valori $+24800 \times 10^{-9}$ (il primo valore osservato corrisponde a 0.000024828 secondi)

Come stimiamo la velocità della luce?

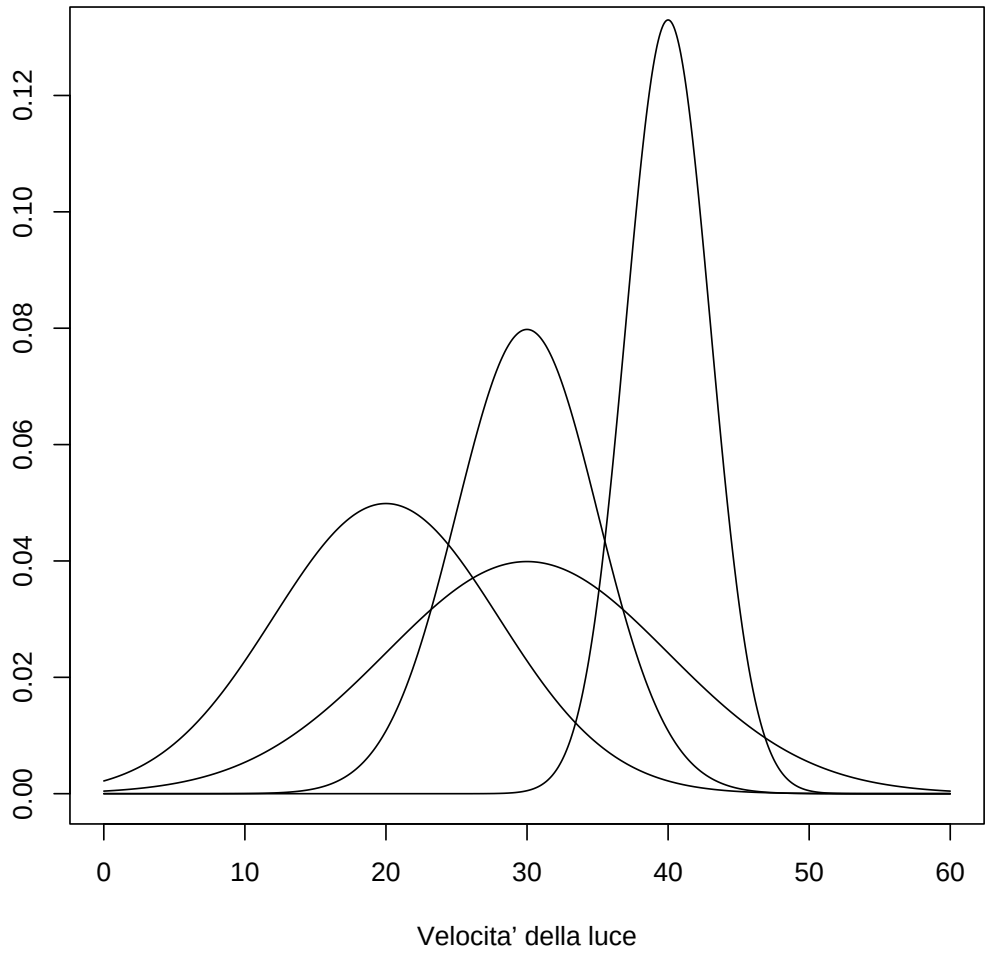
Ipotizziamo che valga il seguente modello

$$Y_i = \mu + \varepsilon_i \;\; \varepsilon_i \sim N(0, \sigma^2)$$

dove Y_i, il risultato della misurazione sperimentale, è uguale al vero tempo impiegato dalla luce, rappresentato dal **parametro** μ, più un errore accidentale ε_i che riflette qui l'imprecisione del metodo di Newcomb. Assumiamo inoltre che tale metodo non conduca ad errori sistematici, in eccesso o in difetto, ma piuttosto che l'errore commesso sia del tutto casuale e come tale che segua una legge normale di media nulla. Ne deriva che i singoli risultati delle misurazioni seguiranno anch'essi una legge normale che ha come media proprio la vera velocità della luce e di cui non conoscosciamo la variabilità σ^2 essendo legata alla precisione dello strumento utilizzato

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

Il nostro modello contiene infinte possibilità (al variare di μ e σ)



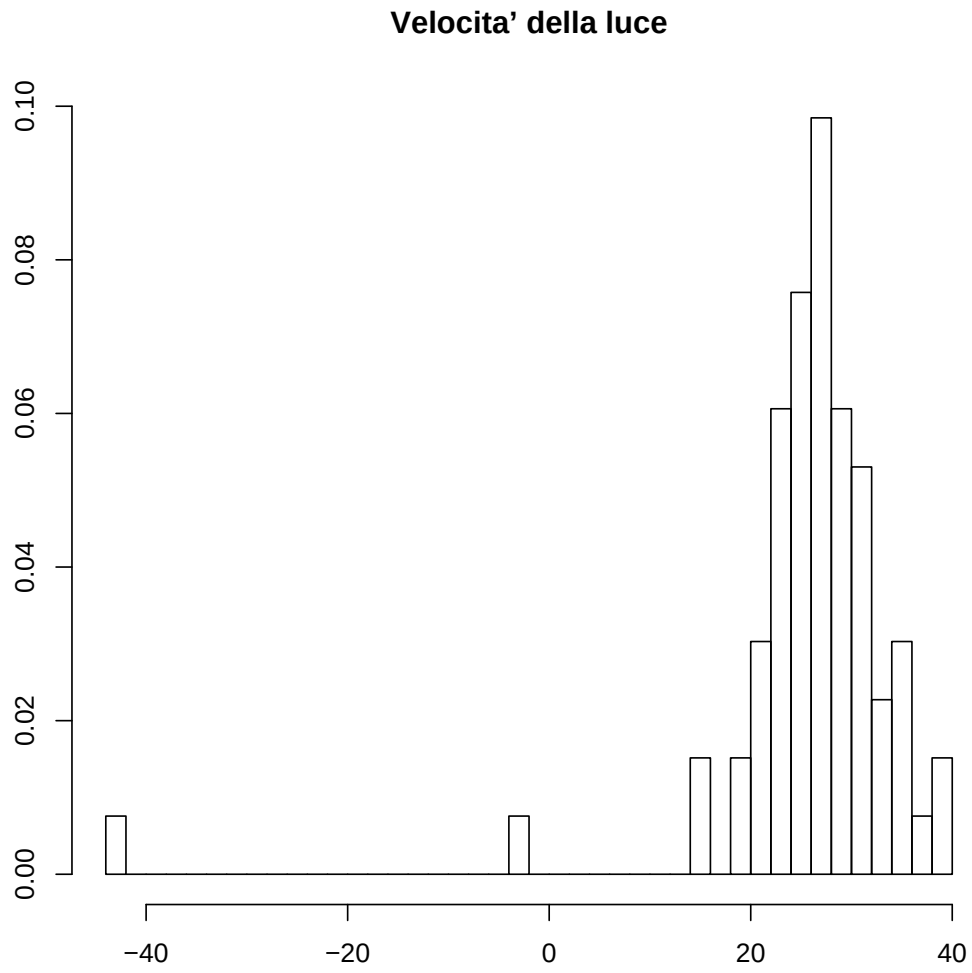
Notiamo come il modello è sintesi delle nostre conoscenze e del processo che ha generato i dati osservati.

Il modello semplifica le realtà ed in tal senso quasi mai è vero; tuttavia, se ben formulato, può aiutare a comprenderne alcuni aspetti.

È importante per un *buon modello* che i parametri siano chiaramente interpretabili. Nel nostro caso μ è il parametro di interesse, rappresentando il vero tempo impiegato dalla luce a percorrere una distanza complessiva di 7.44373 km; σ^2 misura l'imprecisione del metodo usato da Newcomb e, pur non essendo centrale nello studio, dovrà essere stimato.

Entrambi i parametri incogniti, sono delle costanti; l'unico elemento aleatorio nel nostro modello è l'errore ε_i e di conseguenza il risultato della misurazione Y_i. Attenzione che, mentre Y_i è ovviamente osservabile, lo stesso non è vero per ε_i.

Diamo un'occhiata ai dati sperimentali



Lo **stimatore** più naturale (e quello ottimale dato il nostro modello) per μ è la media campionaria $\bar{X}$.

Calcolandolo sul nostro campione otteniamo una **stima** pari a 26.21 che porta ad una stima del vero tempo impiegato dalla luce pari a 0.00002482621 secondi . Fin qui sembra tutto facile; tuttavia se affermassimo a questo punto che il valore stimato è sicuramente quello vero nessuno ci crederebbe (giustamente). Resta infatti da valutare quanta incertezza è legata al nostro risultato.

A distinctive function of statistics is this: it enables the scientist to make a numerical evaluation of the uncertainty of his conclusions (Snedecor 1950)

L'idea di fondo è che, se ripetessimo l'esperimento nelle medesime condizioni, otterremmo un valore probabilmente diverso. Quanto diverso? Con quale probabilità ?

Immaginiamo di ripetere infinite volte l'esperimento *(spazio dei campioni)* e di calcolare ogni volta una stima del tempo impiegato dalla luce basata sulla media del campione osservato.

Avremo così un insieme di medie campionarie. Potremo allora analizzare la loro distribuzione *(distribuzione campionaria di $\bar{Y}$)* e valutare la loro variabilità , calcolando ad esempio la deviazione standard, in modo da avere una misura del grado di incertezza dei risultati *(standard error)*.

Non potendo osservare concretamente tutti i possibili campioni, impostiamo il problema in modo teorico e ci facciamo aiutare dal calcolo delle probabilità . La media campionaria è semplicemente la somma delle singole osservazioni divisa per il loro numero

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Se il nostro modello è valido, se cioè è vero che il risultato di ogni singola misurazione segue la legge normale, e se le diverse osservazioni sono tra loro indipendenti, allora anche loro somma seguirà la stessa legge.

$$\sum_{i=1}^{n} Y_i \sim N(n\mu, n\sigma^2)$$

Dividere per la numerosità del campione ha il solo effetto di scalare media e varianza.

Otteniamo così

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

La media campionaria viene qui interpretata come una variabile aleatoria che assumerà valori diversi al variare del campione osservato.

Sempre se il nostro modello è valido allora la sua distribuzione campionaria è normale

Scopriamo inoltre che la media di $\bar{Y}$ coincide con il valore vero μ, cioè in media, potendo osservare tutti i possibili campioni, otterremmo il vero tempo impiegato dalla luce (purtroppo noi esserveremo soltanto uno dei possibili campioni).

Notiamo come la variabilità del nostro stimatore $\bar{Y}$ dipende direttamente dalla variabilità del fenomeno ed inversamente dalla numerosità del nostro campione

Questo implica che aumentando la numerosità del campione possiamo concentrare la legge normale intorno al suo valor medio μ. Ancora non abbiamo garanzie assolute circa il campione che andremo ad osservare, ma possiamo aumentare la probabilità che la sua media sia vicina al valor vero.

Nel nostro caso la varianza a numeratore σ^2 è espressione dell'imprecisione del metodo usato da Newcomb per misurare la velocità della luce. Possiamo stimarla sulla base della varianza dei dati sperimentali $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$, che nel nostro caso risulta pari a 115.46. Poiché abbiamo in tutto 66 osservazioni, la variabilità della media campionaria sarà pari a $115.46/66 = 1.75$

La corrispondente deviazione standard è nota come *standard error* ($SE(\bar{Y})$) e nel nostro caso è pari a $\sqrt{1.75} = 1.32$. L'interpretazione del valore ottenuto 1.32 non è intuitiva e richiede qualche commento:

- la variabilità che abbiamo stimato si riferisce alla media campionaria intesa come metodo di stima, immaginando di ripetere l'esperimento nelle medesime condizioni. In nessun modo possiamo riferire il valore 1.32 alla stima del tempo impiegato dalla luce (26.21) ottenuta nel nostro campione; quest'ultima è soltanto un numero e quindi non soggetto a variabilità ;
- nell'ipotesi di un modello normale, sappiamo che la media campionaria segue anch'essa una legge normale la cui media coincide con il valore vero μ (e non con il valore stimato di 26.21!). In una curva normale l'area compresa tra $\mu - 1.96 \times \sigma$ e $\mu + 1.96 \times \sigma$ e pari a 0.95. Ne segue che, con una probabilità di 0.95, l'errore massimo che commettiamo (in valore assoluto) è pari a $1.96 \times 1.32 = 2.59$. Questo ovviamente non esclude che il campione da noi osservato possa essere uno di quel 5% di campioni per i quali l'errore è più elevato.

In modo appena più elegante potremmo costruire un intervallo di confidenza per il vero tempo impiegato della luce. Scriviamo più formalmente quanto detto al punto precedente

$$\begin{aligned} Prob\{|\bar{Y} - \mu| < 1.96 \times SE(\bar{Y})\} &= 0.95 \\ Prob\{\mu - 1.96 \times SE(\bar{Y}) < \bar{Y} < \mu + 1.96 \times SE(\bar{Y})\} &= 0.95 \\ Prob\{\bar{Y} - 1.96 \times SE(\bar{Y}) < \mu < \bar{Y} + 1.96 \times SE(\bar{Y})\} &= 0.95 \end{aligned}$$

Così , nel nostro esempio l'intervallo di confidenza al 95% per il tempo impiegato dalla luce
è $(26.21 - 1.96 \times 1.32, 26.21 + 1.96 \times 1.32) = (23.62, 28.80)$

L'ultimo passaggio ha generato molta confusione tra i non statistici, convincendo alcuni che **con una probabilità del** 95% **il valore vero di** μ **si trova tra i due estremi dell'intervallo** o, in altri termini, che **con una probabilità di 0.95 il vero tempo impiegato dalla luce è compreso nell'intervallo** $(23.62, 28.80)$. In realtà il tempo impigato dalla luce è un numero, seppure incognito, e parlare di probabilità riferendosi ad un unico valore non ha alcun significato. Una volta stimato l'intervallo di confidenza esistono solo due possibilità : esso contiene μ oppure no.

L'unico intervallo rispetto al quale ha senso parlare di probabilità è $(\bar{Y} - 1.96 \times SE(\bar{Y}), \bar{Y} + 1.96 \times SE(\bar{Y}))$, i cui estremi dipendono dal campione osservato. Allora possiamo affermare che, in ipotetiche ripetizioni dell'esperimento, il 95% degli intervalli così stimati conterrà il vero tempo impiegato dalla luce. Ancora una volta la garanzia attiene al metodo utilizzato piuttosto che all'intervallo stimato.

Nel calcolare l'intervallo di confidenza abbiamo commesso una piccola imprecisione sostituendo $SE(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$ con una sua stima pari a 1.32. Nel nostro esempio infatti la variabilità delle misurazioni σ, legata alla precisione del metodo usato da Newcomb, non era nota ed è stata stimata sulla base della deviazione standard osservata nel nostro insieme di dati. Questo comporta il passaggio dalla variabile aleatoria

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

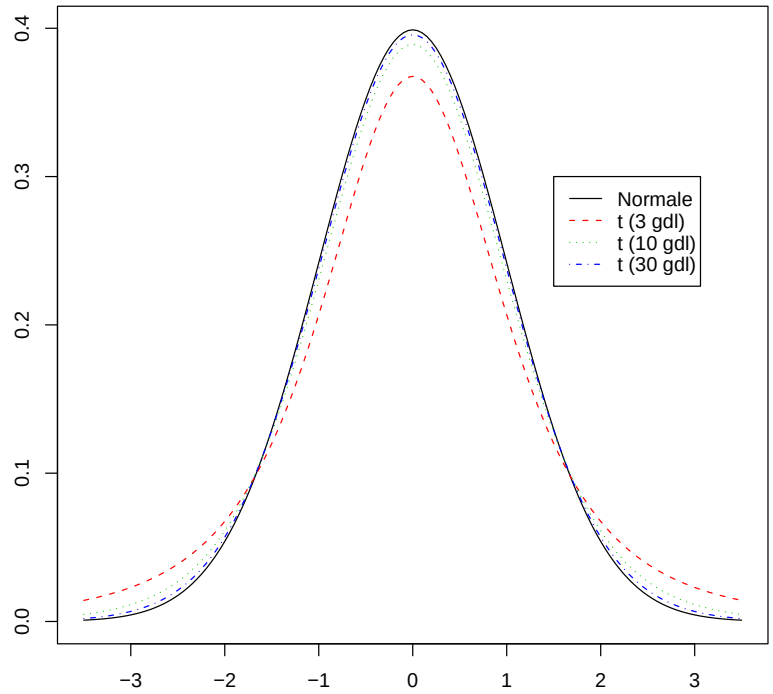
che segue una distribuzione normale standardizzata, alla variabile aleatoria

$$T = \frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}}$$

che segue una legge nota come t di Student con n-1 gardi di libertà

La stima di un secondo parametro (oltre μ) ha introdotto una ulteriore incertezza che si traduce in una maggiore probabilità per i valori estremi.

Infatti, confrontando la distribuzione t con la legge normale, osserviamo nella prima la presenza di code più pesanti, soprattuto per piccoli campioni, laddove la stima di σ risulta meno precisa. Al contrario, già per n=30, la distribuzione t è praticamente indistinguibile dalla normale.



Nel calcolo degli intervalli di confidenza dovremmo sostituire il percentile della legge normale 1.96, con il corrispondente valore calcolato per la distribuzione t di Student.

Riportiamo nela tabella seguente, i valori che delimitano un'area pari a 0.025 sulla coda destra della distribuzione.

gdl (n-1)	1	2	5	10	30	60	120	∞
t	12.706	4.303	2.571	2.228	2.042	2	1.98	1.96

L'intervallo di confidenza al 95% diventa
$(26.21 - 1.996 \times 1.32, 26.21 + 1.996 \times 1.32) = (23.57, 28.84)$,
leggermente più ampio del precente $((23.62, 28.80))$

Resta da chiederci cosa accadrebbe se il modello normale che abbiamo ipotizzato per le nostre osservazioni non fosse valido.

Immaginiamo ad esempio che il metodo usato da Newcomb per misurare il tempo impiegato dalla luce tendesse a sottostimarlo piuttosto che a sovrastimarlo, conducendo con maggiore probabilità a misure per difetto.

Assumiamo ancora che μ sia il valore atteso per il risultato della singola misurazione anche se adesso la sua distribuzione non è più normale.

Grazie ad un risultato noto come **teorema del limite centrale**, possiamo ancora affermare che la media campionaria segue una legge normale centrata intorno a μ, a condizione che il campione abbia una numerosità *elevata* (e che la variabilità delle misurazioni sia finita).

I piselli e il caso

Il caso compare la prima volta nelle scienze bio-mediche con Mendel, nei suoi studi sull'ereditarietà , intorno alla meta' del 1800. Al monaco di Brno non sfugge il fatto che, a fronte di regolarità di fondo riscontrabili nei meccanismi che regolano la trasmissione dei fattori ereditari, resta una certa imprevidibilità sul singolo individuo. Scrive Mendel che regolarità nei rapporti numerici tra tipi di piante "possono risultare solamente dalla media tratta dal maggior numero possibile di casi individuali: più grande è il loro numero, più facilmente si elimineranno le irregolarità accidentali". Così compare per la prima volta quella che in probabilità prenderà il nome di *legge dei grandi numeri*

Attraverso tecniche di impollinazione incrociata Mendel ottenne, a partire da due piante *progenitrici* una i cui piselli erano verdi ed una a piselli gialli, piantine figlie che, nella prima generazione produssero tutte piselli gialli (carattere dominante). Nella seconda generazione ricomparvero invece alcune piantine di piselli verdi.

Se accettiamo l'idea di un carattere dominante ed uno recessivo, esiste una regolarità nella frequenza con cui il carattere recessivo tende a rispresentarsi?

Detto altrimenti, qual'è la probabilità che si ripresenti una pianta a piselli verdi nella seconda generazione?

Ammettiamo come ragionevole l'idea che questa probabilità resti costante da esperimento ad esperimento (cerchiamo una regolarità) e che gli esperimenti siano stati condotti in modo indipendente.

Cerchiamo di formalizzare questa idea chiedendo aiuto proprio alla probabilità . Indichiamo con X il risultato aleatorio di un esperimento, dove gli esiti possibili sono soltanto due: successo (1) e insuccesso (0). Nel nostro caso il successo è rappresentato dal ripresentarsi di una piantina di piselli verdi.

Inoltre indichiamo con p la probabilità che l'esperimento si concluda con un successo. Chiaramente la probabilità di un insuccesso sarà 1-p. Scriviamo in simboli (tra breve non potremo farne a meno) quello che abbiamo appena detto a parole

$$Prob(X = x; p) = p^x(1-p)^{(1-x)}$$

Questa legge prende il nome di distribuzione Bernoulliana

Complichiamo appena un pò le cose e indichiamo con Y il numero di successi che otteniamo in 3 esperimenti tra loro indipendenti e condotti nelle medesime condizioni, dove la probabilità di successo resta invariata. Y potrà assumere il valori (0,1,2,3) Con quali probabilità ? ...

Y	X_1	X_2	X_3	Prob
0	0	0	0	$(1-p)^3$
1	1	0	0	$(1-p)^2 p$
1	0	1	0	$(1-p)^2 p$
1	0	0	1	$(1-p)^2 p$
2	1	1	0	$(1-p)p^2$
2	1	0	1	$(1-p)p^2$
2	0	1	1	$(1-p)p^2$
3	1	1	1	p^3

Quanto appare nella tabella può essere riscritto più generale come

$$Prob(Y = y; p, n) = \frac{n!}{y!(n-y)!} p^y (1-p)^{(n-y)}$$

dove la frazione a secondo membro *conta* in quanti modi possiamo ottenere y successi in n esperimenti ed n! (n fattoriale) equivale al prodotto dei primi n numeri naturali

Questa distribuzione discreta è nota come legge **binomiale**.

Una volta scritto il nostro modello, ci chiediamo quale **stimatore**, cioè quale funzione dei dati osservati, sia più adatto per arrivare a conoscere p. Anche in statistica esistono fortunatamente delle regolarità ; così il modo migliore per stimare una probabilità è quasi sempre la corrispondente proporzione osservata nel campione. Notate come ancora una volta stiamo usando una media campionaria anche se particolare.

In simboli

$$\hat{p} = Y/n = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mendel osservò in un totale di 8023 esperimenti, 6022 piantine a piselli gialli e 2001 a piselli verdi, ottenendo una stima di p pari a $2001/8023 = 0.249$, molto vicina a quel rapporto di 1 a 4 (0,25) che si dimostro poi vero. Stime molto simili si ottennero studiando il colore dei fiori e la rugosità o meno dei piselli.

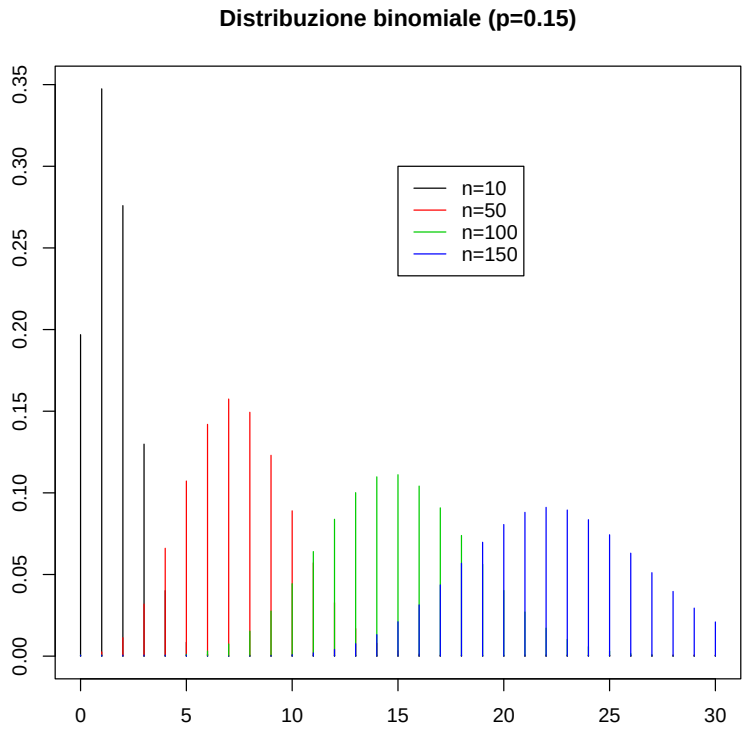
Resta da valutare l'incertezza della stima ottenuta da Mendel. Possiamo infatti riproporre la domanda: se altri ricercatori avessero effettuato ciascuno lo stesso numero di esperimenti (8023) nelle medesime condizioni quale variabilità ci ettenderemmo nelle loro stime.

Si dimostra che in un modello binomiale la precisione dei risultati ottenuti utilizzando la stimatore Y/n è legata sia al numero di esperimenti effettuati, sia al valore di p secondo la formula $Var(\hat{p} = Y/n) = \frac{p(1-p)}{n}$

Come è intuitivo, maggiore è il numero di esperimenti condotti, maggiore è la precisione del risultato finale (e Mendel ne sembrava già consapevole). Quello che è meno immediato osservare è che la variabilità delle stime è maggiore se la probabilità di un successo è prossima a 0.5, mentre si riduce se l'evento è quasi certo (p prossima ad 1) o quasi impossibile (p prossima a 0). Del resto è del tutto ragionevole che la precisione del risultato dipenda dall'incertezza sul successo nel singolo esperimento.

Nel caso di mendel la stima della varianza di Y/n è pari a $(0.249(1 - 0.249))/8023 = 0.0000233$, molto bassa come faceva supporre l'elevato numero di esperimenti condotti dal monaco.

Possiamo tradurre i risultati ottenuti in un intervallo di confidenza al 95% per p. Per far questo ricorriamo ad un risultato della teoria della probabilità che ci garantisce la convergenza della legge binomiale a quella normale al crescere di n e a condizione che p (e 1-p) non siano prossimi a 0.



Allora, sotto le medesime condizioni, anche il nostro stimatore Y/n seguirà una legge normale di media p e varianza $\frac{p(1-p)}{n}$ e potremo scrivere

$$Prob\{-1.96 < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < +1.96\} = 0.95$$

$$Prob\{\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\} = 0.95$$

Nel nostro esempio, data l'elevata numerosità , possiamo applicare questo risultato ottenendo un intervallo di confidenza stimato pari a

$$(0.249 - 1.96 \times 0.0048, 0.249 + 1.96 \times 0.0048) = (0.240, 0.258)$$

che conferma l'idea di estrema precisione del risultato finale