STANISLAO CANNIZZARO
Liceo Scientifico - ROMA

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Esperienza 1: la statistica

La statistica è una scienza che nasce tra il XVII e il XVIII secolo come una teoria dei giochi: ad oggi è uno strumento essenziale in molte discipline scientifiche e non solo.
Infatti viene utilizzata ogni qualvolta esiste la possibilità di poter replicare un determinato evento (dalla semplice indagine telefonica all'esperimento) sempre sotto le stesse condizioni, con due finalità:

- ricavare informazioni su una particolare caratteristica di ciò che si sta studiando (altezza degli italiani…lunghezza di tutti banchi di una scuola). Si parla in tal caso di *statistica descrittiva*;
- riuscire a predire l'evoluzione del fenomeno in esame (aumento entro 5 anni dell'uso di Internet in Italia…aumento della temperatura globale). Questa è la *statistica induttiva*.

Supponiamo allora di voler andare a determinare qual è l'altezza degli italiani: per farlo un centro di ricerche comincia a raccogliere dati, con telefonate e domande in giro per l'Italia.
Supponiamo che i primi 5 dati siano 170 cm, 166 cm, 179 cm, 173 cm e 181 cm: la migliore stima che possiamo dare per il valore dell'altezza degli italiani è data dalla **media aritmetica**. Detti x_i i valori dei dati raccolti dall'indagine, con i che va da 1 al numero N totale di dati, la media aritmetica $\bar{x}$ è:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

dove il simbolo Σ indica la somma su tutti i valori x_i. Perciò nel caso precedente:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{5} x_i = \frac{1}{5}(170 + 166 + 179 + 173 + 181) = 173{,}8\ cm$$

Il valore medio dell'altezza degli italiani è in realtà 175 cm, un po' più grande di quello appena trovato. Ma l'indagine va avanti e vengono registrati altri 5 valori: 179 cm, 174 cm, 164 cm, 185 cm, 178 cm. Questa volta il valore medio dell'altezza è di 174.9 cm
Come si è visto, aumentando il numero di persone intervistate, la media tende ad avvicinarsi al reale valore che chiameremo **valore vero**.

Quindi effettivamente la qualità della descrizione statistica che possiamo tirar fuori dai dati dipende da quanti se ne hanno a disposizione: si riesce ad ottenere un'informazione completa, ossia il valore vero di una grandezza, solo quando questo è una rappresentazione di tutte le osservazioni possibili (cioè se si riesce ad intervistare tutti gli italiani).
Perciò si è soliti fare una distinzione tra:

- la *popolazione* ossia l'insieme dei dati disponibili: la stima statistica di una grandezza che caratterizza una popolazione è detta *valore vero*.
- un *campione* ossia l'insieme dei dati che effettivamente viene analizzato. È sempre una parte della popolazione e la rappresentazione statistica di una sua grandezza è detta *valore medio* (il valore medio non è necessariamente la media aritmetica).

Di conseguenza, la precisione della descrizione statistica di un fenomeno dipenderà dalle dimensioni del campione e la sua affidabilità da quanto lo stesso è rappresentativo di tutta la popolazione. Nell'esempio precedente qualora il campione sia costituito da ragazzi e non da adulti, esso risulterebbe poco caratteristico dell'intera popolazione italiana in quanto l'altezza media viene valutata sull'età adulta: di conseguenza un valore medio non sarebbe indicativo della grandezza studiata.
All'aumentare quindi della rappresentatività e delle dimensioni del campione, questi diventerà sempre più indicativo della popolazione ed il suo valore medio tenderà verso il valore vero.

L'istogramma:

Una volta che si sono raccolti i dati è utile, al fine di una loro migliore visualizzazione, rappresentarli sotto forma di un grafico particolare detto **istogramma**.
Per realizzarlo si dividono l'intervallo dei dati sperimentali in sottointervalli detti **classi** e si calcola quanti risultati cadono in ciascuna classe, ovvero la **frequenza** di ciascuna classe f_i definita come:

$$f_i = \frac{N_i}{N}$$

dove N_i è il numero di dati che cadono nell'i-esima classe ed N il numero totale di dati raccolti.
Nel costruire le classi dell'istogramma bisogna assicurarsi che tutti i valori trovino un loro posto all'interno della suddivisione e che ciascuno di esso appartenga ad una sola classe.
Nell'esempio dell'altezza media degli italiani potremmo allora costruire 7 classi tali per cui:

- la prima classe contiene tutte le altezze inferiori a 160 cm
- la seconda da 160 cm a 165 cm
- la terza da 165 cm a 170 cm

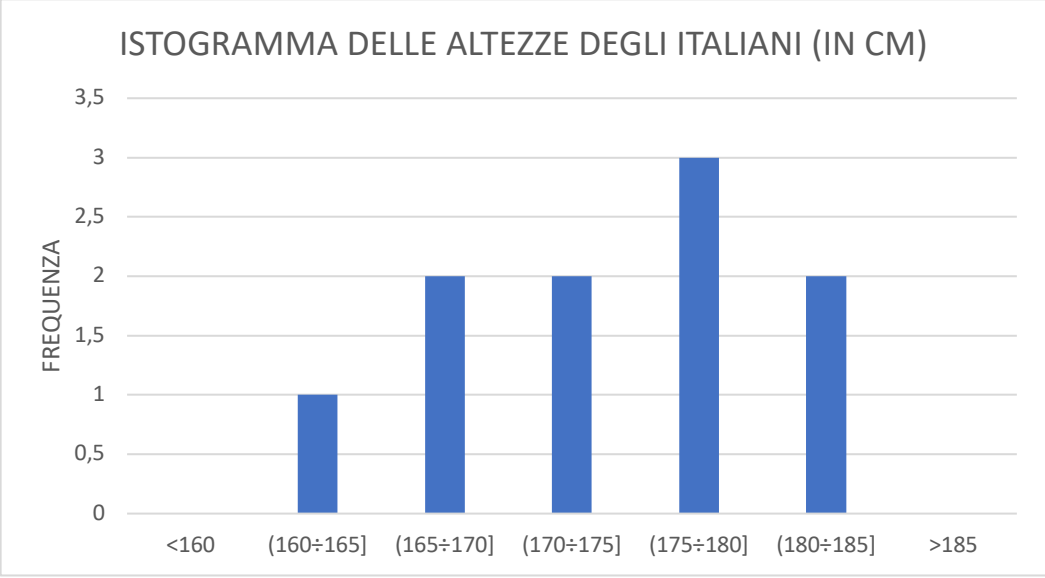
- la quarta da 170 cm a 175 cm
- la quinta da 175 cm a 180 cm
- la sesta da 180 cm a 185 cm
- la settima tutte le altezze maggiori di 185 cm

Per i dati che si trovano sul confine della classe, possiamo decidere arbitrariamente dove metterli: così ad esempio l'altezza 170 cm possiamo metterla nella terza classe. Graficamente si indica questo intervallo come $(165 \div 170]$ intendendo con la parentesi tonda che il valore di altezza 165 cm non è incluso nella classe mentre con la parentesi quadra che il valore 170 cm è invece incluso. Per tutti gli altri elementi è necessario seguire lo stesso criterio.

Possiamo allora costruire una tabella dove riportiamo le classi, il loro intervallo, quali elementi vi appartengono e qual è la frequenza di quella classe:

Numero classe	Intervallo della classe	Elementi	Frequenza
1	$\leq 160\ cm$	-	0
2	$160\ cm \div 165\ cm$	$164\ cm$	1/10
3	$165\ cm \div 170\ cm$	$166\ cm, 170\ cm$	2/10
4	$170\ cm \div 175\ cm$	$173\ cm, 174\ cm$	2/10
5	$175\ cm \div 180\ cm$	$178cm, 179\ cm, 179\ cm$	3/10
6	$180\ cm \div 185\ cm$	$181\ cm, 185\ cm$	2/10
7	$> 185\ cm$	-	0

Questo è ciò che otteniamo allora:



Ovviamente non è necessario riportare sempre la frequenza per ogni classe: può essere anche riportato il numero di elementi che cadono in ogni classe.

Esiste una regola empirica che consente di regolare il numero n delle classi:

$$\frac{N}{n} \geq 10$$

Ciò consente di far sì che il numero n delle classi sia adeguato ovverosia non ci sono classi con troppi dati al proprio interno e viceversa.

Abbiamo quindi visto che, davanti ad una distribuzione di dati come quella ottenuta da un'indagine statistica, la loro media aritmetica è una stima del valore medio del campione. Questo è vero in genere anche per i dati che si raccolgono quando si eseguono misure ripetute di una grandezza fisica soggette ad **errori casuali** ovvero ad errori che variano in modo imprevedibile da una misura all'altra e che ne influenzano il risultato qualche volta per eccesso, altre per difetto. In questo caso quindi è possibile trattare i dati in modo statistico.
La media statistica rappresenta quindi una buona stima del valore vero della grandezza che stiamo misurando alla quale tuttavia dobbiamo associare un'indeterminazione che nasce dalle fluttuazioni casuali dei risultati che è la **deviazione standard della media** $s_{\bar{x}}$:

$$s_{\bar{x}} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N(N-1)}}$$

La probabilità:

Nell'analisi statistica di un esperimento un risultato non ha tanto importanza in sé quanto in relazione a tutti i possibili risultati dell'esperimento, il cui insieme costituisce lo **spazio dei risultati**.
In questo spazio ogni possibile risultato dell'esperimento è rappresentato da un punto: ad esempio nel lancio di un dado lo spazio sarà costituito da 6 punti, tanti quanti sono i possibili esiti del lancio. Ogni possibile risultato è detto **evento**.
Via via che l'esempio viene complicato, lo spazio dei risultati aumenta di dimensione.
Supponiamo che tutti i punti dello spazio dei risultati siano equivalenti tra loro, nel senso che nessun evento si presenta più frequentemente di un altro: le prove sono tra loro equivalenti e identiche. Se i punti dello spazio dei risultati sono N, si assegna a ciascuno di essi la **probabilità a priori**:

$$p_i = \frac{1}{N}$$

Allora quando si lancia un dado non truccato ogni possibile evento ha probabilità $1/6$ di verificarsi mentre nel lancio di una moneta l'esito testa e l'esito croce hanno entrambe probabilità $1/2$.

Quando però non tutti i casi non sono tra di loro equivalenti la probabilità di un evento si calcola andando a contare i punti dello spazio dei risultati che lo rappresentano (**casi favorevoli**) e farne il rapporto con la dimensione dello spazio (**casi possibili**):

$$p_i = \frac{casi\ favorevoli}{casi\ possibili}$$

Ad esempio in un dado non truccato, la probabilità che in un lancio esca il numero 4 è pari a $1/6$ in quanto i casi favorevoli sono 1 ("esce il numero 4") mentre i casi possibili sono 6 ("esce il numero 1", "esce il numero 2", …, "esce il numero 6").

Supponiamo ora che io sia interessato a calcolare la probabilità che in un unico lancio di due dadi esca come punteggio totale sulle facce il numero 5.

Allora dovrò andare prima a vedere quali sono le combinazioni delle facce dei due dadi che mi danno come somma 5 e quali sono tutti i casi possibili. Costruiamo la seguente tabella:

	DADO 1	DADO 2
CASI FAVOREVOLI = 4	1	4
	2	3
	3	4
	4	5
CASI POSSIBILI = $6 \cdot 6 = 36$	1	1, 2, 3, 4, 5, 6
	2	1, 2, 3, 4, 5, 6
	3	1, 2, 3, 4, 5, 6
	4	1, 2, 3, 4, 5, 6
	5	1, 2, 3, 4, 5, 6
	6	1, 2, 3, 4, 5, 6

I casi possibili sono calcolati considerando che per ognuno dei possibili esiti su una delle due facce (1, 2, 3, 4, 5, 6) ce ne sono 6 per l'altra faccia: in totale quindi ce ne sono $6 \cdot 6 = 36$.

La probabilità è quindi di $4/36$ cioè $1/9 \simeq 11\%$: questo significa che su 9 lanci è probabile che un evento sia tale per cui la somma dei numeri sulle facce sia 5.

Questi due esempi mostrano due situazioni differenti:

- nel lancio di un dado tutti i possibili esiti hanno la stessa probabilità (1/6): si parla di *eventi equiprobabili*;

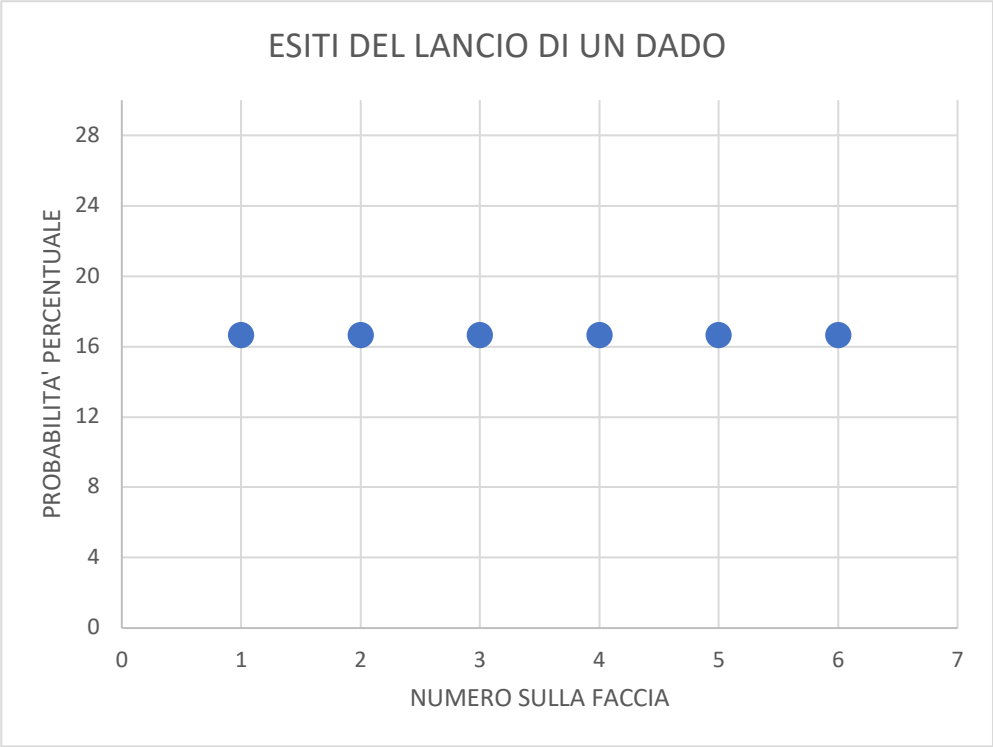


Figura 1

- nel lancio di due dadi, gli esiti della somma dei numeri sulle facce non hanno la stessa probabilità. Ci sono eventi più probabili (ottenere un 7) ed eventi meno probabili (ottenere un 2) perché ci sono più combinazioni che permettono di ottenere quell'evento. Ad esempio per il 7 ci sono 6 possibili casi (1+6, 2+5, 3+4, 4+3, 5+2, 6+1) mentre per ottenere un 2 si può solo avere l'evento 1+1.

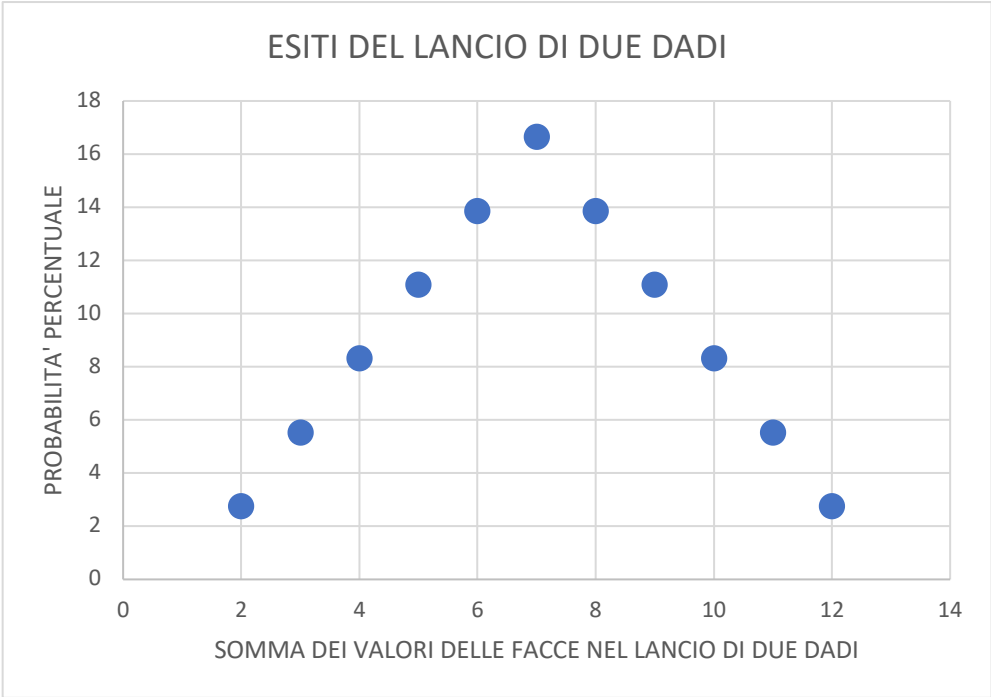


Figura 2

I grafici di questi due esempi sono naturalmente dei casi teorici in quanto quando si lanciano due dadi non è detto che entro 9 lanci si ottenga che la somma dei numeri sulle facce è 5 in quanto l'evento è aleatorio cioè è probabilistico. La probabilità che facendo un lancio si ottenga come somma 5 è del 11% ma ciò non esclude il fatto che questo evento si verifichi subito oppure non si verifichi affatto nei 9 lanci.

La definizione appena data di probabilità ha un limite in quanto per applicarla è necessario comunque conoscere tutto lo spazio dei risultati e, nel caso in cui si stanno eseguendo delle misure sperimentali in fisica, questo è infinito ed i suoi punti non sono tutti equiprobabili.

Possiamo allora dare un'altra definizione di probabilità che è puramente empirica: consideriamo ad esempio un dado truccato ovvero un dado per il quale ogni possibile esito ("uscita del numero X") non è equiprobabile. Non si può assegnare una probabilità a priori pari a 1/6 a ciascuno dei risultati del lancio e non esiste un metodo per valutarla teoricamente.

È possibile stimare la probabilità di ottenere in un lancio 1, 2, 3,...,6 attraverso un esperimento che consiste nel lanciare il dado un numero molto elevato di volte e calcolare la frequenza con cui si sono presentati i diversi risultati.

Possiamo fare un esempio andando a simulare con il computer il lancio di un dado truccato: supponiamo di effettuare 45 lanci. L'esito della simulazione è riportato in Figura 3:

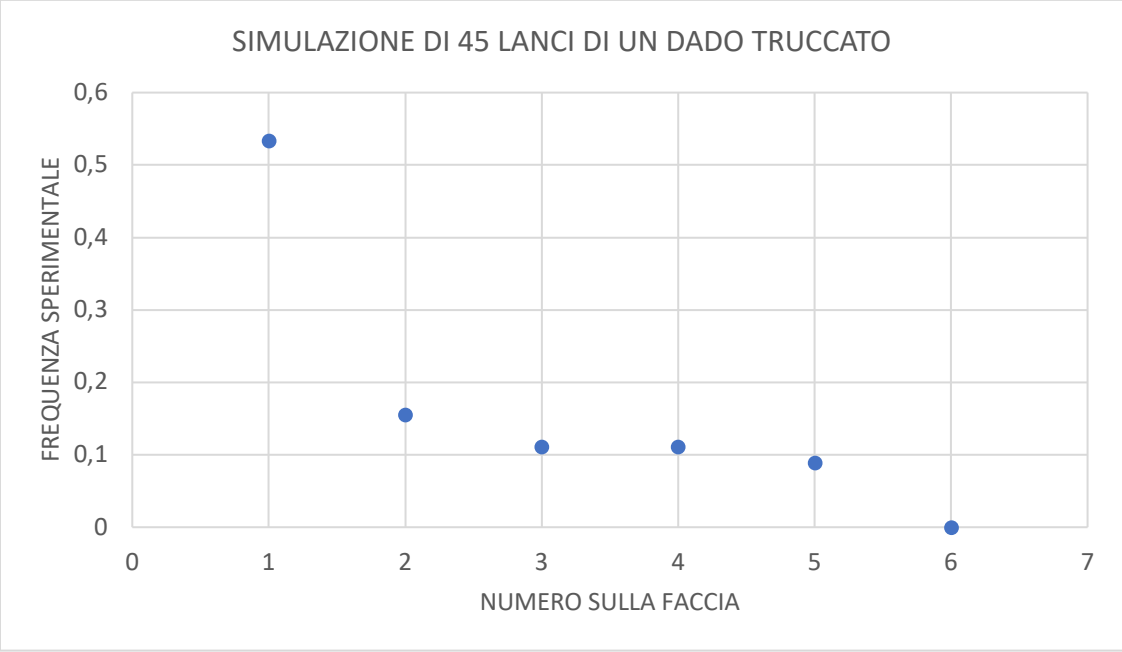


Figura 3

Come si vede la **distribuzione** dei risultati, ovvero il modo con il quale i risultati si dispongono all'interno del grafico, non è uniforme cioè i possibili eventi non sono tra loro equiprobabili.

Se aumentiamo il numero di lanci nella simulazione, ci accorgiamo che:

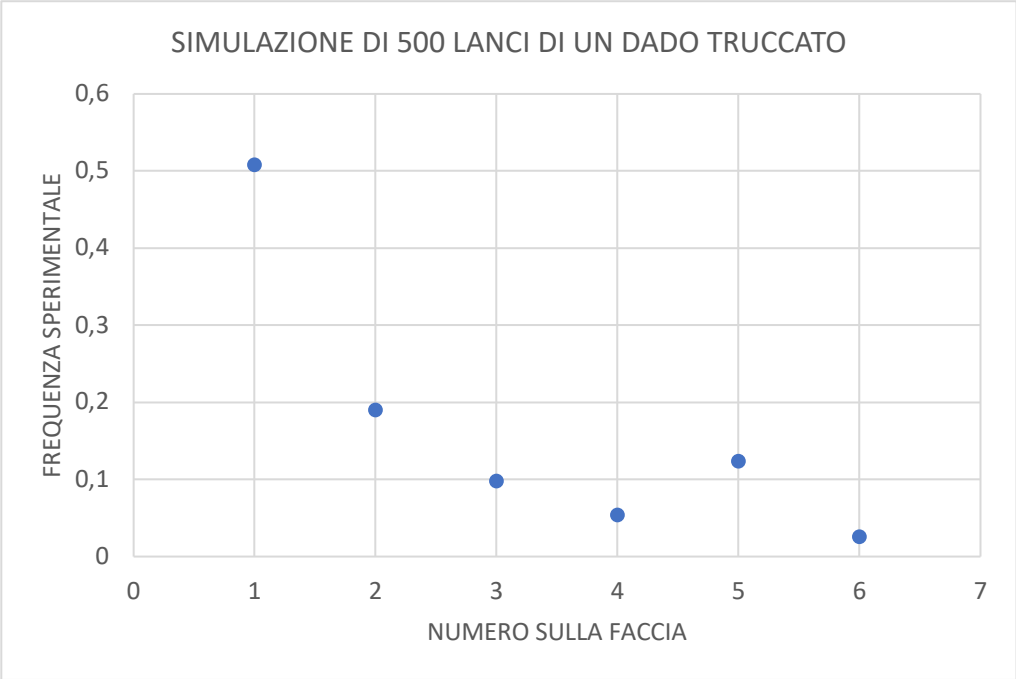


Figura 4

Le frequenze di uscita di ciascuna faccia cambiano radicalmente rispetto al caso precedente nel quale si erano effettuati pochi lanci. Ciò si vede meglio se confrontiamo i due esiti:

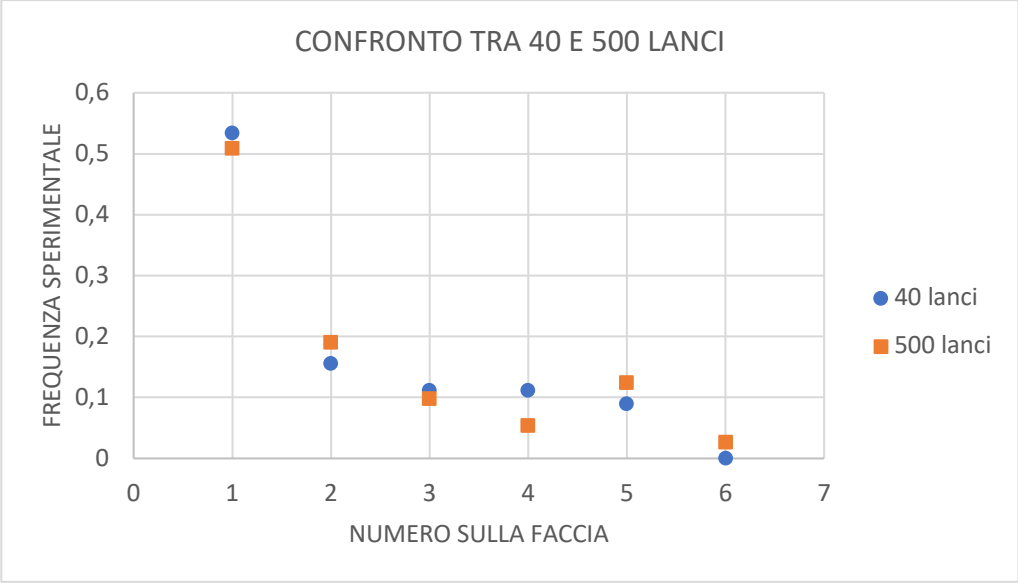


Figura 5

Quando eseguiamo quindi un esperimento, non appena il numero di prove è sufficientemente alto, la frequenza sperimentale tende a stabilizzarsi su un valore: questo valore viene assunto come la probabilità teorica del risultato stesso.

Quest'esito è noto come **legge dei grandi numeri**.

La correttezza di questa legge può esser vista confrontando una simulazione con un dato teorico. Supponiamo quindi di lanciare due dadi non truccati: abbiamo già visto in Figura 2 qual è la probabilità di ottenere i diversi valori della somma delle due facce.

Proviamo quindi a simulare 100 lanci:

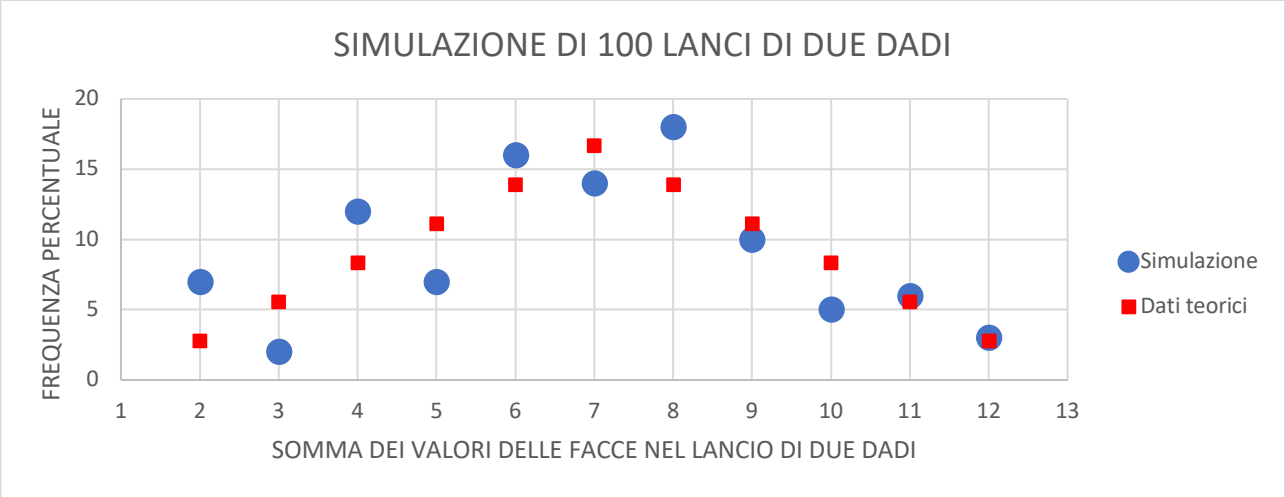


Figura 6

Il pallino blu più grande indica la frequenza percentuale di ciascun evento ovvero quante volte durante la simulazione si è presentato l'evento rispetto allo spazio dei risultati. Il quadrato rosso più piccolo indica la frequenza percentuale teorica mostrata già in Figura 2. Ciò che si nota è che la frequenza sperimentale non coincide con quella teorica. Se però proviamo ad aumentare il numero dei lanci della simulazione si osserva qualcosa di diverso.

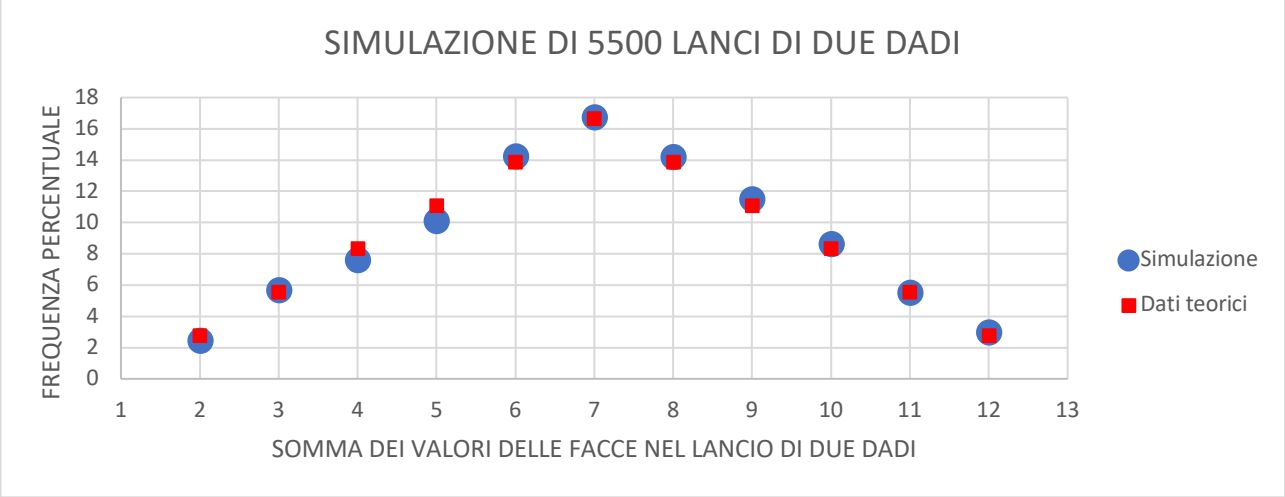


Figura 7

La simulazione dà esiti differenti rispetto al caso precedente: le frequenze sperimentali tendono a sovrapporsi ai dati teorici. Ma questo accade solo quando il numero di prove è elevato.

La validità di questa legge è puramente empirica e gli esempi riportati in Figura 6 e 7 non costituiscono una dimostrazione della legge ma servono a mostrare come la sovrapposizione tra dato teorico e sperimentale aumenti al crescere del numero di prove effettuate.