Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 6 7

Piano della presentazione

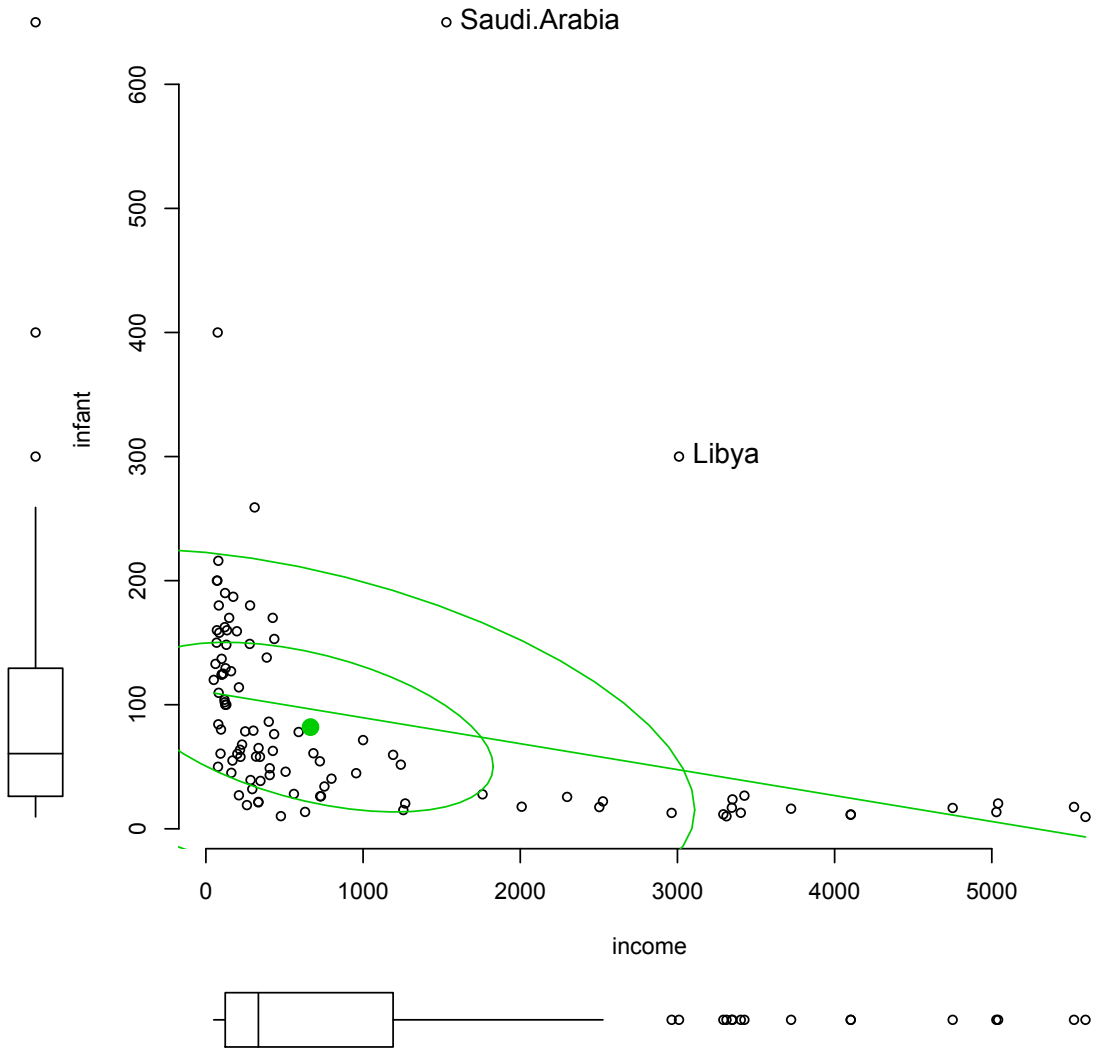
Metodi Grafici

In precedenza abbiamo descritto i box-plot che forniscono una rappresentazione grafica di cinque indici di una distribuzione (valore minimo, primo quartile, mediana, terzo quartile, valore massimo). Le proprietà di una distribuzione si possono però rappresentare graficamente anche in modi diversi.

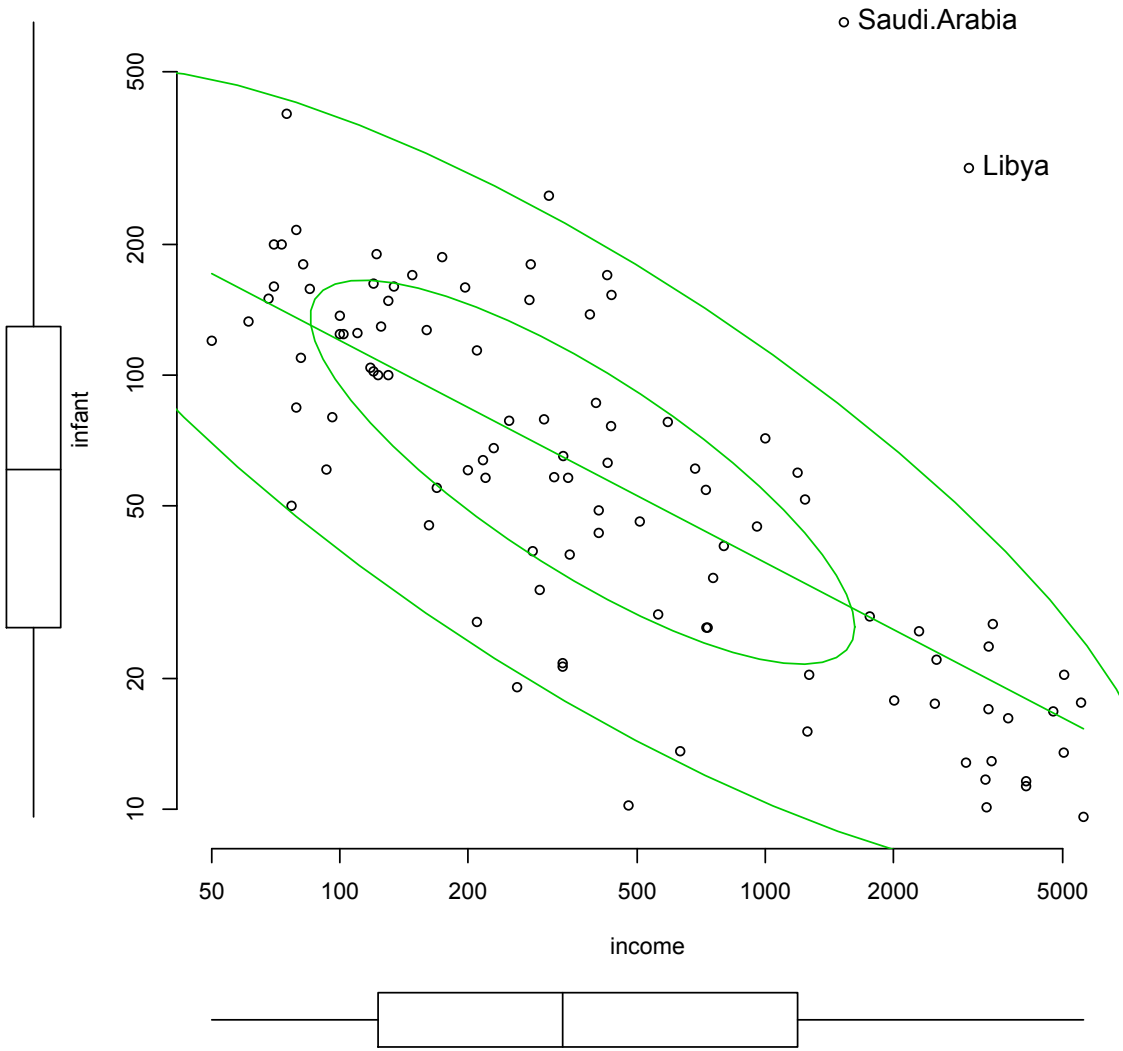
Diagrammi di dispersione Il grafico più comune per dati bivariati è il diagramma di dispersione. Ciascun dato viene rappresentato da un punto in un diagramma cartesiano.

istogrammi Grafici a barre che rappresentano la densità di dati numerici discreti o continui.

Diagramma di dispersione



Trasformazione dei dati



Istogramma

- L'istogramma fornisce una rappresentazione grafica di una distribuzione empirica (campione di osservazioni).

Dati discreti un istogramma per dati discreti è un grafico a barre che rappresenta la frequenza o la frequenza relativa dei valori di una variabile.

Dati continui un istogramma per dati continui si costruisce dividendo la risposta in un insieme di intervalli, solitamente di grandezza uguale.

- R dispone del comando `hist`. Se x è un vettore in cui sono stati salvati i dati, `hist(x)` restituisce l'istogramma di x.

Isotgramma

- La variabile x viene suddivisa in classi e su ciascuna classe viene calcolata la frequenza assoluta (opzione di default, equivale a `freq=TRUE`) o relativa (opzione `freq=FALSE`) di osservazioni che ricadono in ogni classe.
  - L'ampiezza di ciascun intervallo viene stabilita da R in modo automatico.
  - Si può comunque agire sul numero di classi oppure sull'ampiezza degli intervalli (argomenti `breaks` e `nclass`)

Istogramma

- Solitamente, gli istogrammi sono costruiti in modo tale che l'altezza delle barre sia uguale al numero di osservazioni (frequenza assoluta) in ciascun intervallo.
- Possiamo però anche scalare l'asse verticale in modo tale che l'area di ciascuna barra sia uguale alla frequenza relativa di osservazioni comprese in quella classe (cfr. figura). In questo secondo caso, l'area totale delle barre dell'istogramma sarà uguale a 1.0 e l'asse verticale rappresenterà delle densità.

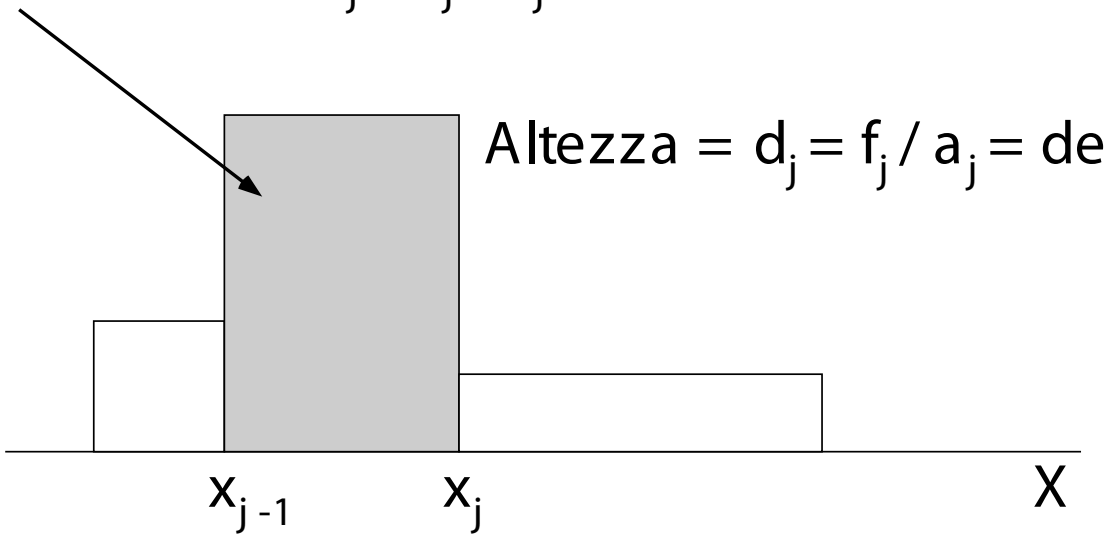
Istogramma

Classi	Freq. rel	Amp.cl	densità
$x_0 - x_1$	f_1	a_1	d_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{j-1} - x_j$	f_j	a_j	d_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{h-1} - x_h$	f_h	a_h	d_h
Totale	1		

- Densità di probabilità: $d_j = f_j/a_j$
- Ampiezza della classe: $a_j = x_j - x_{j-1}$

Istogramma

Area rettangolo= $a_j * d_j = f_j$ = frequenza relativa classe



Base rettangolo = $x_j - x_{j-1} = a_j$ = ampiezza classe

Istogramma

```
par(mfrow=c(1,2))
x <- c(0.5, 1, 1, 1.5, 1.6, 2, 2.4, 3.1, 4, 4.3, 4,
                7.7, 8.1, 9.9, 11.2)
 hist(x, ylab="Frequenze")
 hist(x, freq=FALSE, ylab="Densita'")
```

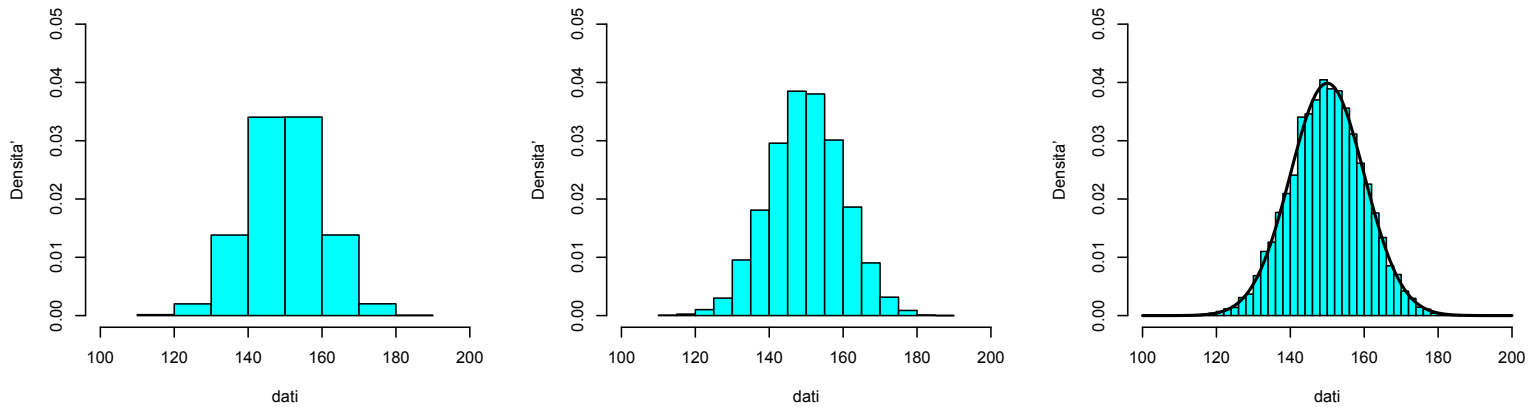
Consideriamo il calcolo della densità per la prima barra dell'istogramma: 6 di 15 osservazioni cadono nell'intervallo $(0 - 2$, con ampiezza $2)$. La frequenza relativa di osservazioni rappresentate dalla prima barra è uguale a $6/15 = 0.4$ e la densità è uguale a $0.4/2 = 0,2$.

Istogramma

- Supponiamo di raccogliere un grande numero di osservazioni di una variabile quantitativa continuae immaginiamo di rendere gli intervalli Δx sempre più piccoli.
- L'altezza dei singoli rettangoli dell'istogramma cambierà certamente, di poco o di molto, ma rimarrà comunque un numero finito allorché Δx tende a zero.
- Ciò che tenderà a zero con Δx è l'area di ogni rettangolo che è uguale alla frequenza relativa delle misure che cadono nell'intervallo Δx.

Istogramma

- In queste circostanze, i lati corti superiori dei vari rettangoli tenderanno a confondersi in una curva continua $f(x)$.
- La quantità $f(x)$ è detta densità continua di frequenza relativa (o probabilità).



Istogramma

- È possibile, anche se non frequente, usare ampiezze diverse per gli intervalli di un istogramma.
- In tali casi è importante che l'area di ciascuna barra sia proporzionale alla frequenza o frequenza relativa. In caso contrario, l'istogramma fornisce delle informazioni fuorvianti.

Limitazioni dell'istogramma

L'istogramma ha le seguenti limitazioni.

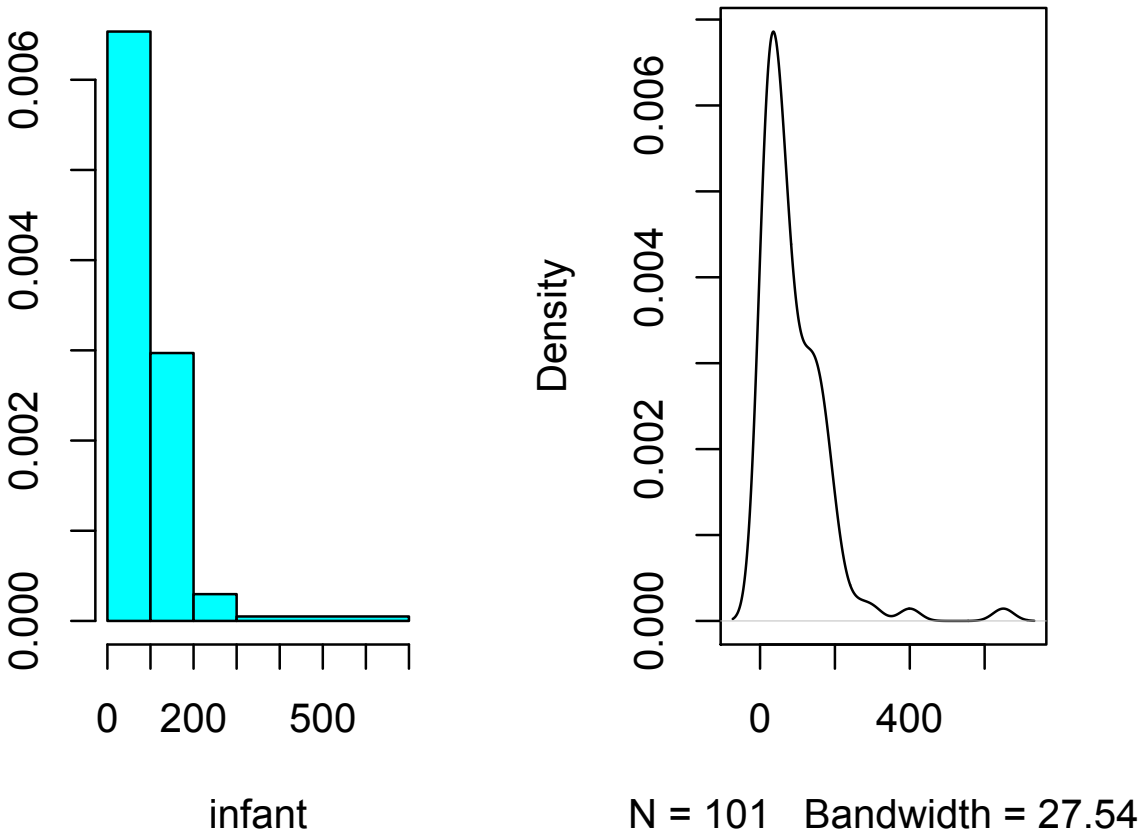
- L'impressione visiva che trasmette può dipendere dall'origine della sequenza delle classi.
- L'istogramma è discontinuo anche se la variabile rappresentata è continua.
- La forma dell'istogramma dipende dalla grandezza delle classi.
- Se vengono scelte classi piccole per rappresentare in maniera adeguata le proprietà della distribuzione intorno al suo centro, solitamente queste classi sono troppo piccole per evitare un eccesso di rumore dove i dati sono sparsi – di solito, nelle code della distribuzione.

Rimedi

- Per queste ragioni è preferibile rappresentare l'andamento di una distribuzione discreta, cioè di un istogramma, per mezzo di una curva continua, come quella generata dalla funzione `density()`.

```
library(MASS)
par(mfrow=c(1,2))
attach(Leinhardt)
truehist(infant,breaks=c(0, 100, 200, 300, 700))
plot(density(infant,na.rm=TRUE), main="")
detach(Leinhardt)
```

Funzione continua di densità di probabilità



Probabilità e inferenza statistica

- La motivazione ultima dell'analisi dei dati è il desiderio di inferire determinate proprietà della popolazione da cui è stato estratto il campione di dati.
- Le proprietà della popolazione che sono l'oggetto della statistica inferenziale sono detti parametri. Sono detti parametri, ad esempio, la media, la varianza e la mediana della popolazione.

Prima di descrivere alcuni metodi inferenziali che possono essere usati per stimare i parametri della popolazione dobbiamo però esaminare alcune proprietà della teoria della probabilità da cui dipende la struttura della popolazione.

Probabilità e inferenza statistica

La teoria della probabilità è una branca della matematica e si occupa dei fenomeni aleatori.

- Per definizione, un singolo evento aleatorio non è prevedibile. Tuttavia, le ripetizioni dei fenomeni aleatori esibiscono delle regolarità. Lo scopo della teoria della probabilità è quello di descrivere queste regolarità.
- La maggior parte della matematica moderna ha origini antiche. La teoria della probabilità, invece, non esisteva prima del rinascimento (le sue origini possono essere individuate nel diciassettesimo secolo).

Parametri e statistiche

- Uno degli usi della teoria della probabilità è quello di fornire le basi necessarie per l'inferenza statistica.
- L'inferenza statistica è il processo che consente di trarre delle conclusioni a proposito delle caratteristiche di una popolazione utilizzando le informazioni fornite da un campione di osservazioni tratto da quella popolazione.

Parametri e statistiche

- Le caratteristiche della popolazione sono dette parametri e sono simbolizzate dalle lettere greche. La media della popolazione, per esempio, si indica con μ.
- Le caratteristiche del campione sono dette statistiche e sono simbolizzate dalle lettere latine. La media del campione, per esempio, si indica con $\bar{x}$.

Parametri e statistiche

- Una caratteristica fondamentale delle statistiche è che variano da campione a campione.
  - Supponete, ad esempio, di raccogliere due campioni di studenti universitari e di calcolare l'altezza media degli studenti di ciascun campione.
  - È improbabile che le medie dei due campioni siano uguali, o che una o l'altra media sia uguale alla media della popolazione μ.

Parametri e statistiche

- Per usare $\bar{x}$ per fare delle inferenze relative a μ dobbiamo sapere che caratteristiche assumerebbe $\bar{x}$ se il processo di campionamento venisse ripetuto.
- Per esempio, quanto ci aspettiamo che $\bar{x}$ sia simile a μ quando raccogliamo un campione di n casi?
- La teoria della probabilità ci consente di fornire una risposta a domande di questo tipo.

Elementi di teoria della probabilità

Nella teoria della probabilità

- un esperimento aleatorio è una procedura ripetibile attraverso la quale viene compiuta un'osservazione;
- l'esito di un esperimento (*outcome*) è una possibile osservazione risultante da un esperimento;

lo spazio campione Ω dell'esperimento è l'insieme di tutti i possibili esiti. Ciascuna specifica realizzazione di un esperimento produce un particolare esito nello spazio campione.

Spazio campione discreto

- Lo spazio campione può essere discreto o continuo.
- I seguenti sono esempi di uno spazio campione discreto.

  Lancio di una moneta Lo spazio campione è $\Omega = \{T; C\}$. L'evento testa è $\{T\}$

  Lancio di due monete Lo spazio campione è $\{TT; TC; CT; CC\}$. L'evento che consiste nell'osservare una volta l'esito testa è $\{TC; CT\}$.

  Lancio di un dado Lo spazio campione è $\{1, 2, 3, 4, 5, 6\}$. L'evento dispari è $\{1, 3, 5\}$. L'evento < 3 è $\{1, 2\}$.

Spazio campione continuo

- Se registriamo con uno strumento molto preciso il tempo necessario per l'apprendimento di un certo compito motorio, allora lo spazio campione dell'esperimento sarà continuo e sarà costituito dai numeri reali positivi: $\Omega = \{x : x > 0\}$.

Eventi

- Un evento è un sottoinsieme dello spazio campione – ovvero, un insieme di esiti di un esperimento aleatorio.
- Un evento ha luogo se uno dei suoi elementi costituenti viene osservato.
  - Per esempio, per $\Omega = \{TT; TC; CT; CC\}$, l'evento $E = \{TT; TC\}$ testa nel primo lancio ha luogo se viene osservato l'esito TT o l'esito TC.
- Un evento si dice semplice se consiste in uno solo degli esiti possibili di un esperimento; si dice composto se consiste in un insieme di esiti possibili.
  - $E = \{TT; TC\}$ è un evento composto
  - $E = \{TT\}$ è un evento semplice.

Probabilità

Le probabilità sono dei numeri che vengono assegnati agli eventi in maniera coerente con i seguenti assiomi.

assioma 1 la probabilità di un evento E è un numero compreso tra 0 e 1: $0 \leq P(E) \leq 1$.

assioma 2 lo spazio campione Ω è esaustivo – quando l'esperimento è eseguito qualche esito deve venire osservato $P(\Omega) = 1$.

assioma 3 due eventi A e B sono disgiunti se non hanno alcun esito in comune – e*venti disgiunti non possono verificarsi simultaneamente*. La probabilità di osservare uno o l'altro di due eventi disgiunti è la somma delle loro separate probabilità: $P(A \cup B) = P(A) + P(B)$. In generale, se $A_1, A_2, \ldots, A_n$ è un insieme di n eventi mutuamente esclusivi (. . . disgiunti. . .), allora

$$P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{n} P(A_i)$$

.

Modelli probabilistici

Modello probabilistico di un esperimento descrizione dello spazio campione dell'esperimento e delle probabilità associate agli eventi definiti sullo spazio campione in maniera coerente con gli assiomi precedenti.

- Gli assiomi della teoria della probabilità non sono sufficientemente restrittivi da imporre un'unica attribuzione di probabilità agli eventi di uno spazio campione.
- Ciascun esperimento implica dunque infiniti modelli probabilistici diversi.

Illustrazione

- Supponiamo che gli eventi dello spazio campione $\Omega = \{TT; TC; CT; CC\}$ siano equiprobabili:

$$P(TT) = P(CC) = P(TC) = P(CT) = .25$$

  - La probabilità dell'evento $E = \{TT; TC\}$ è $P(E) = 0.25 + 0.25 = .50$
  - Siano $A = \{CT; CC\}$ l'evento croce nel primo lancio e $B = \{TT\}$ due teste. Gli eventi A e B sono disgiunti e l'evento A o B è $\{TT; CT; CC\}$.

$$P(A \cup B) = P(A) + P(B) = 0.50 + 0.25 = 0.75$$

Interpretazione frequentista

Nella statistica classica (e in questo corso) le probabilità sono interpretate come frequenze relative calcolate per un grande numero di ripetizioni dell'esperimento.

- Se la probabilità di un evento è 0.5, allora l'evento verrà osservato in circa metà delle ripetizioni dell'esperimento. Le frequenze relative, inoltre, si approssimano sempre di più alle probabilità al crescere del numero di ripetizioni dell'esperimento.
- L'interpretazione frequentista della probabilità fornisce una procedura empirica per stimare le probabilità: ripetere molte volte l'esperimento e calcolare la frequenza relativa con cui l'evento di interesse si è verificato.

- Gli eventi possono essere visualizzati utilizzando i diagrammi di Venn (così nominati in onore del matematico inglese del diciannovesimo secolo John Venn – anche se Leibnitz e Eulero avevano già in precedenza utilizzato rappresentazioni simili).
  - Un diagramma di Venn rappresenta lo spazio campione di un esperimento come uno spazio rettangolare.
  - Al suo interno, gli eventi definiti su questo spazio campione sono rappresentati con delle regioni chiuse.
  - In talune versioni dei diagrammi di Venn, la probabilità di un evento è proporzionale all'area della regione che lo rappresenta.

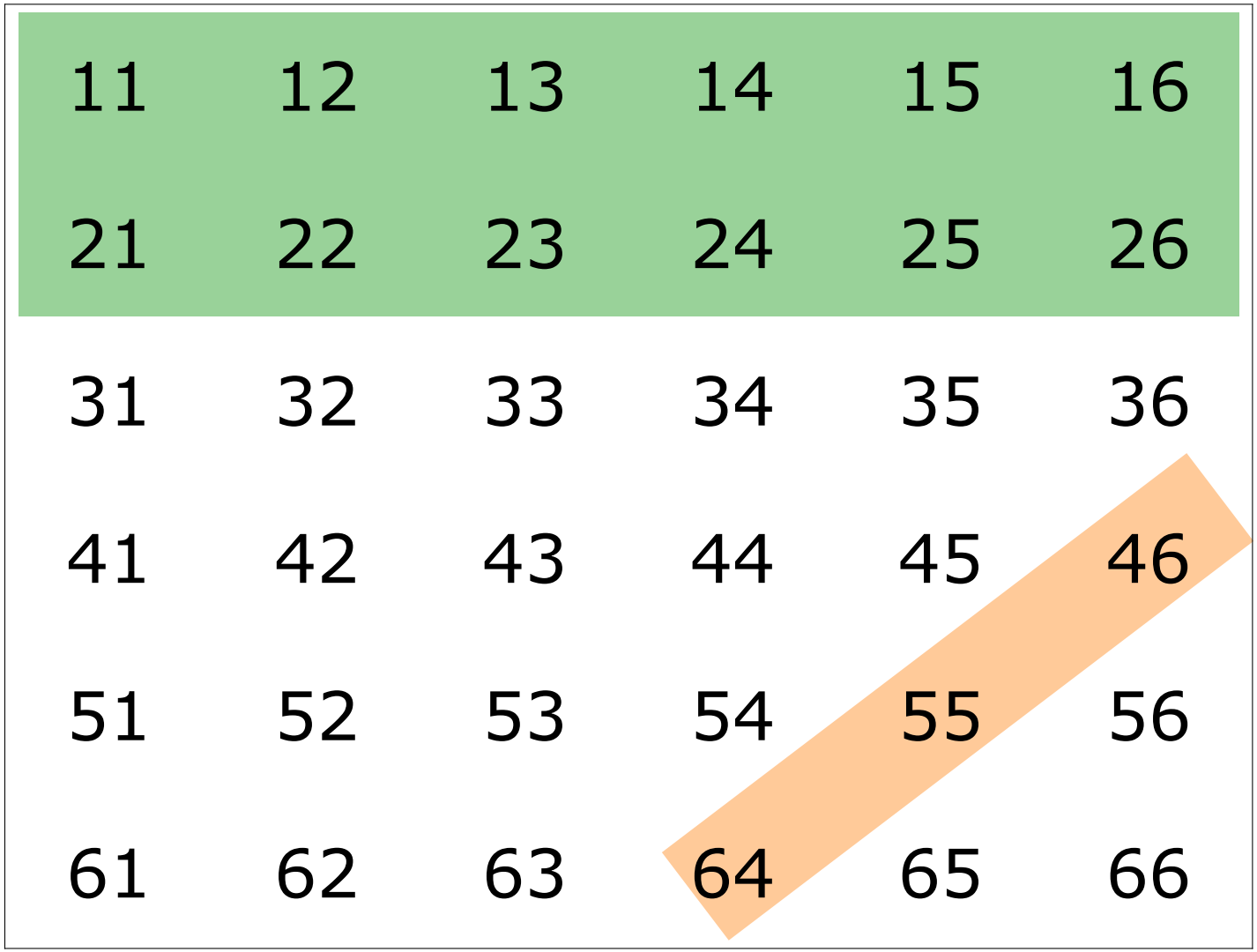
Diagrammi di Venn

Illustrazione

Si consideri l'esperimento consistente nel lancio di due dadi

- Nel successivo diagramma di Venn viene rappresentato lo spazio campione di questo esperimento. Tale spazio campione è costituito da $6 \times 6 = 36$ eventi semplici.
- Per esempio, (23) rappresenta l'esito *2 con il primo lancio e 3 con il secondo*.
- L'evento A rappresenta l'esito per cui si osserva un 1 o un 2 nel primo lancio.
- L'evento B rappresenta l'esito per cui i due lanci producono un punteggio totale uguale a 10.
- Si noti che questi due eventi sono disgiunti (non hanno alcun esito in comune).

Eventi disgiunti



Eventi disgiunti

- Supponiamo che i 36 esiti possibili E_i di questo esperimento siano *equiprobabili*: $P(E_i) = 1/36$. Dunque

$$P(A) = 12/36 = 1/3.$$

$$P(B) = 3/36 = 1/12.$$

- Inoltre, dato che A e B sono disgiunti

$$P(A \cup B) = 15/36 = 5/12, \text{ovvero}$$

$$\boxed{P(A \cup B) = P(A) + P(B)} = \frac{1}{3} + \frac{1}{12} = \frac{15}{36}$$

Eventi non disgiunti

Se due eventi A e B non sono disgiunti, allora quando sommiamo le loro probabilità dobbiamo evitare che la loro parte comune $A \cap B$ venga contata due volte, quindi:

$$P(A \cup B) = P(A) + P(B) - A \cap B$$

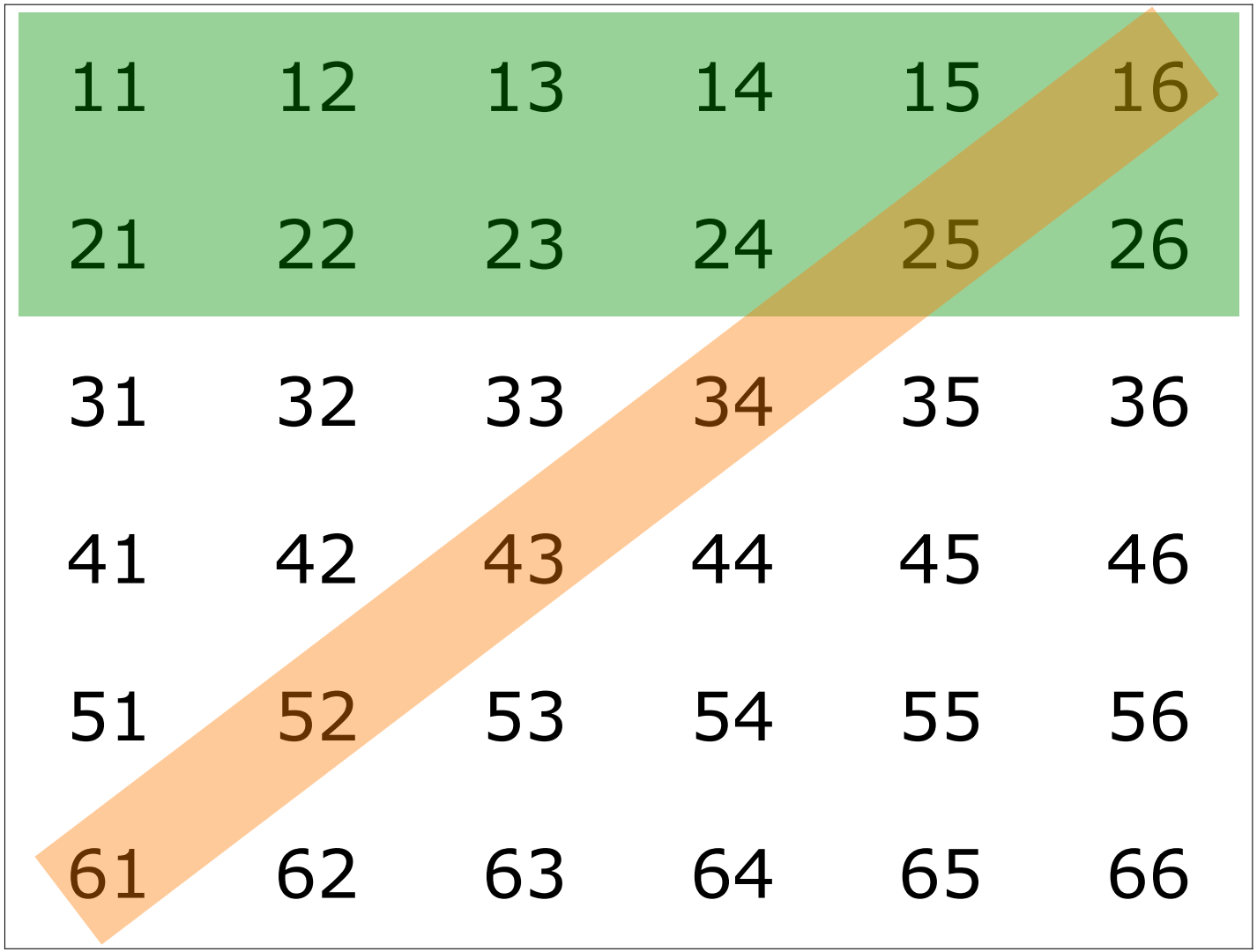
Eventi non disgiunti

Illustrazione

Si considerino gli eventi definiti dal successivo diagramma di Venn.

- L'evento A rappresenta l'esito per cui si osserva un 1 o un 2 nel primo lancio. $P(A) = 12/36$
- L'evento C rappresenta l'esito per cui i due lanci producono un punteggio totale uguale a 7.
  L'evento C si verifica in 6 casi su 36 possibili e dunque $P(C) = 6/36$
- L'evento $A \cup C$ corrisponde a 16 casi su 36 e ha probabilità $P(A \cup C) = 16/36$.
- L'evento $A \cap C$ corrisponde a 2 casi su 36 e ha probabilità $P(A \cap C) = 2/36$.

Eventi non disgiunti



Eventi non disgiunti

In base alla regola precedente, dunque

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - A \cap B \\ &= \frac{12}{36} + \frac{6}{36} - \frac{2}{36} = \frac{16}{36} \end{aligned}$$

Eventi indipendenti

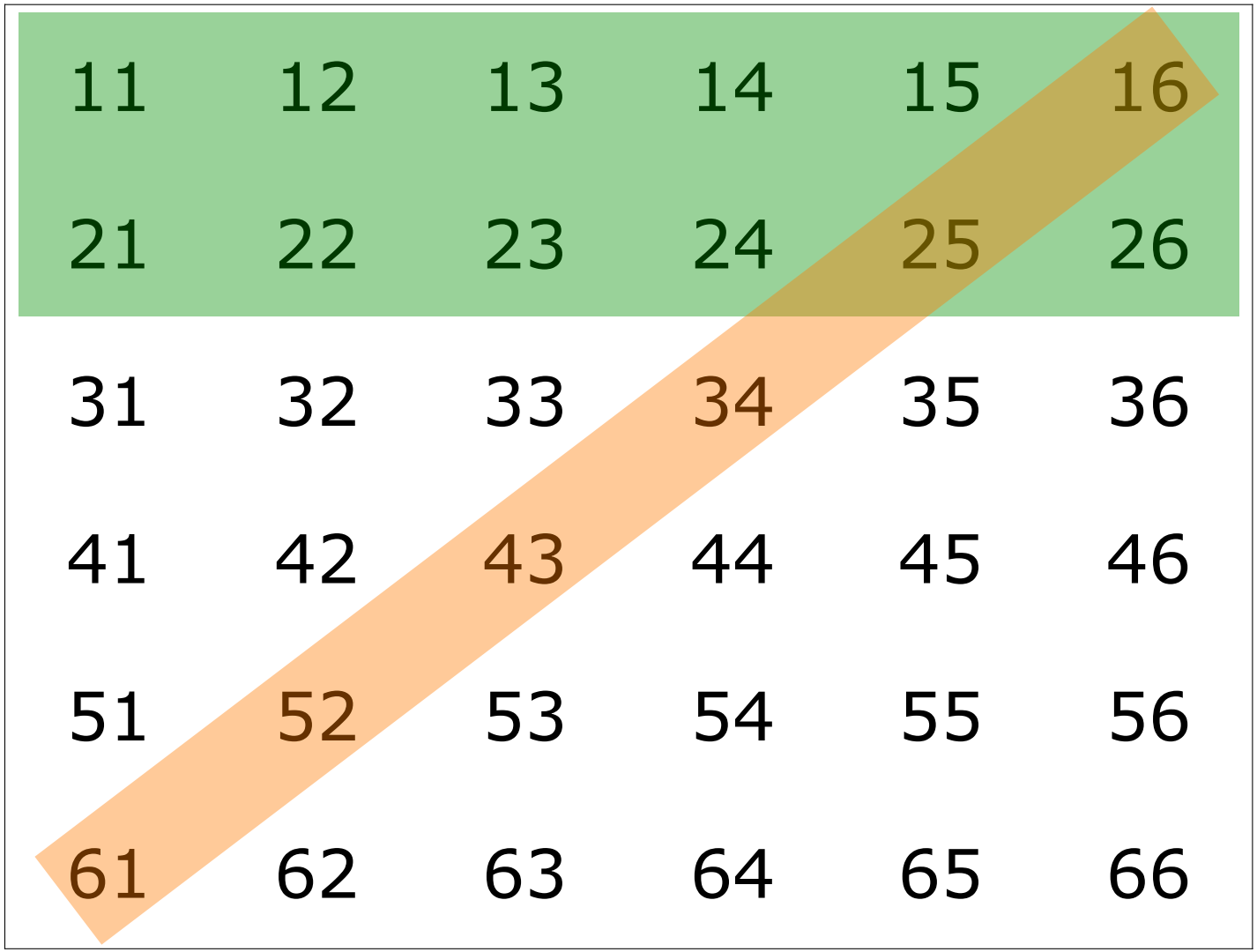
Un relazione tra eventi molto importante è quella di indipendenza tra eventi.

definizione

Due eventi sono detti indipendenti se il verificarsi di uno non altera la probabilità del verificarsi dell'altro.
Se due eventi A e B sono indipendenti, allora la probabilità che entrambi si verifichino è uguale al prodotto delle loro separate probabilità.
Se A e B sono indipendenti, allora

$$P(A \cap B) = P(A)P(B)$$

Illustrazione



Illustrazione

La realazione di indipendenza è illustrata nel precedente diagramma di Venn.

- Se i 36 esiti possibili che costituiscono lo spazio campione sono equiprobabili, allora $P(A) = 12/36$, $P(C) = 6/36$ e

$$P(A \cap C) = 2/36$$

- Dato che $P(A \cap C) = P(A)P(C) = 12/36 * 6/36 = 2/6 * 1/6 = 2/36$, gli eventi A e C sono indipedenti

quindi

Il fatto di osservare 1 o 2 nel primo lancio non cambia la probabilità di ottenere un totale di 7 con entrambi i lanci.

Eventi dipendenti

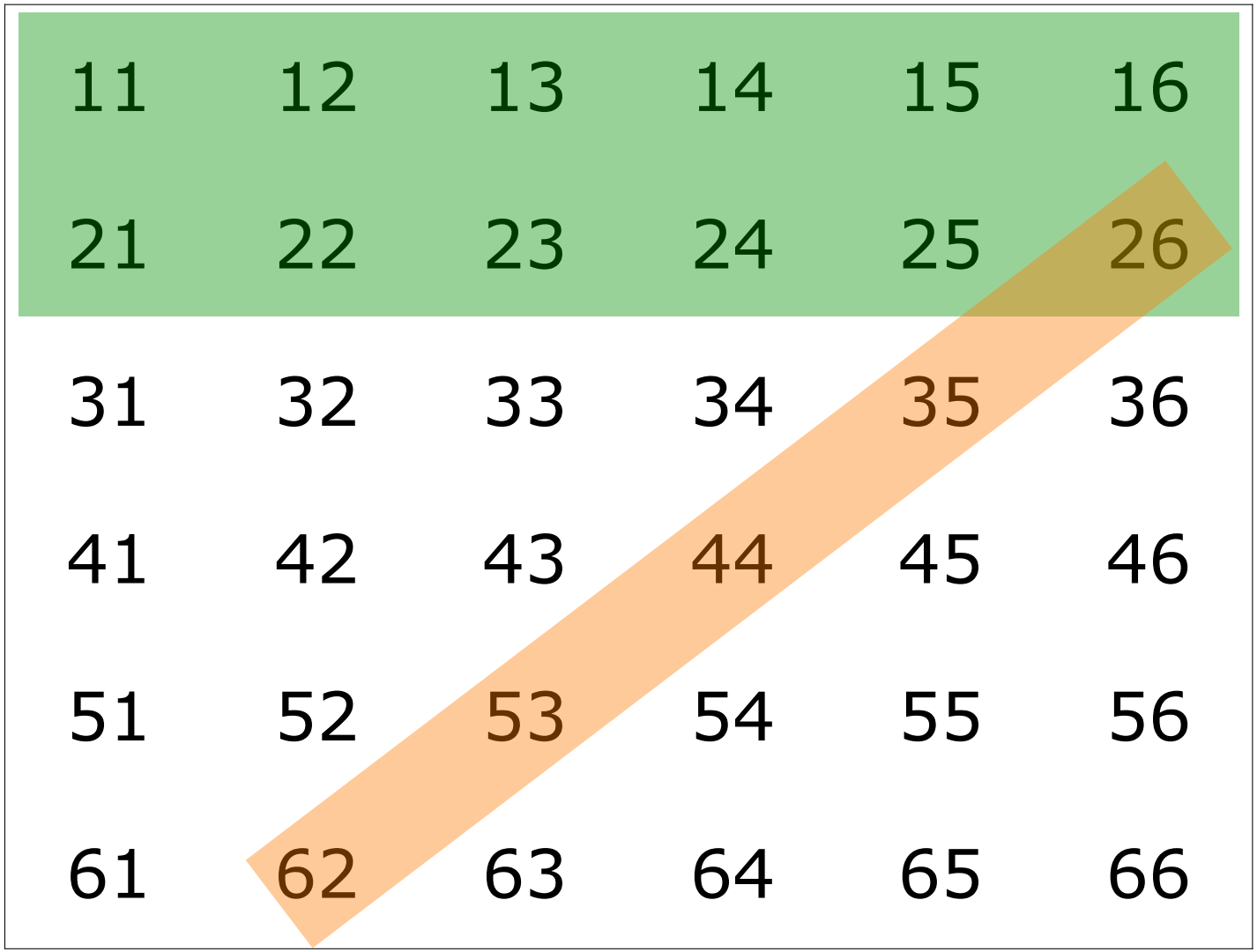
La relazione di dipendenza tra eventi è illustrata nel successivo diagramma di Venn.

- Gli eventi
  - A (osservare 1 o 2 nel primo lancio)
  - D (ottenere un totale di 8 nei due lanci)

  sono dipendenti o associati.
- Infatti $P(A) = 12/36 = 1/3$, $P(D) = 5/36$ e

$$P(A \cap D) = 1/36 = 3/108 \neq P(A)P(D) = 5/108$$

Illustrazione



Eventi dipendenti

osservazione

La relazione di indipendenza non deve essere confusa con la relazione di disgiunzione:
se due eventi A e B sono disgiunti, allora non hanno esiti in comune e

$$P(A \cap B) = 0 \neq P(A)P(B)$$

Gli eventi disgiunti sono dipendenti dato che il verificarsi del primo preclude la possibilità del verificarsi dell'altro.

Illustrazione

Un gruppo di studenti è costituito da 24 femmine e 26 maschi. Di questi, 12 studentesse e 18 studenti maschi giocano a tennis. Una persona viene scelta a caso da questo gruppo.

1. Qual è la probabilità che questa persona giochi a tennis?
2. Se sappiamo che la persona scelta è una studentessa, qual è la probabilità che giochi a tennis?

Risposta:

1. $(12 + 18)/(24 + 26) = 30/50 = 0,6$
2. $12/24 = 0,5$

Calcolo della probabilità condizionata

Le probabilità calcolate conoscendo alcune informazioni, o date talune circostanze, sono dette probabilità condizionate e sono denotate da $P(A|B)$.

definizione

La probabilità condizionata $P(A|B)$ è la probabilità del verificarsi dell'evento A all'interno dello spazio campione ridotto rappresentato dall'evento B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

In generale,

per due eventi associati A e B, $P(B|A) \neq P(B)$ e $P(A|B) \neq P(A)$;

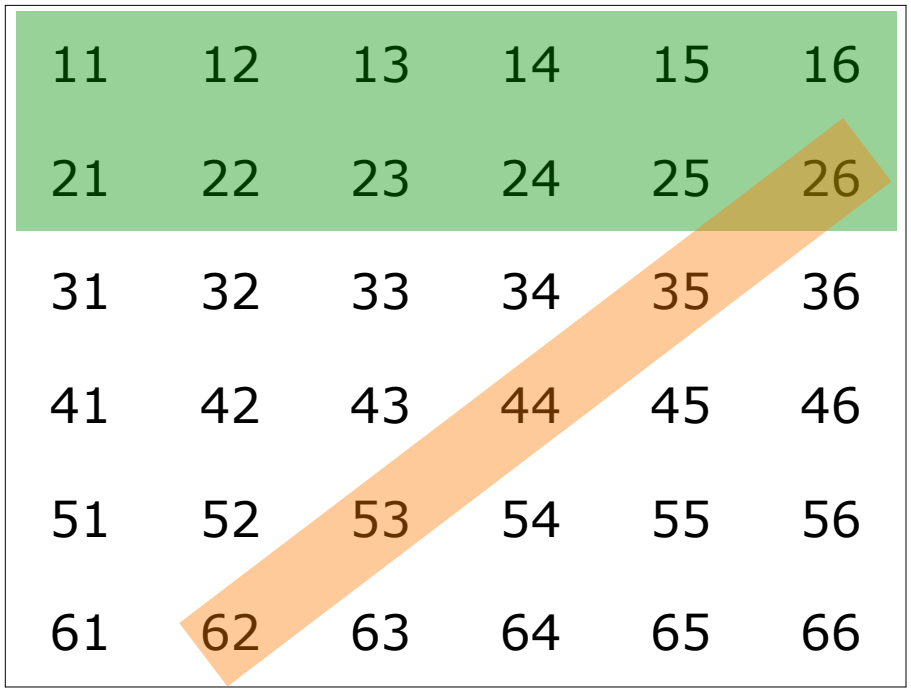
per due eventi indipendenti invece, $P(B|A) = P(B)$ e $P(A|B) = P(A)$, in altri termini, le probabilità condizionate e non condizionate sono uguali.

Calcolo della probabilità condizionata

Riprendiamo l'esempio precedente sugli eventi

- A (osservare 1 o 2 nel primo lancio)
- D (ottenere un totale di 8 nei due lanci)

che sono tra loro dipendenti (o associati).



Calcolo della probabilità condizionata

Avevamo che:

$$P(A) = 12/36 = 1/3,$$

$$P(D) = 5/36,$$

$$P(A \cap D) = 1/36 \neq P(A)P(B).$$

Le probabilità condizionate di A e D sono:

$$P(A|D) = \frac{1/36}{5/36} = 1/5,$$

$$P(D|A) = \frac{1/36}{12/36} = 1/12,$$

da cui otteniamo le equivalenze

$$P(A \cap D) = P(A)P(D|A) = \frac{1}{3}\frac{1}{12} = 1/36,$$

$$P(A \cap D) = P(D)P(A|D) = \frac{5}{36}\frac{1}{5} = 1/36.$$

Illustrazione

- Qual è la probabilità che la somma dei due lanci sia 3?
- Sapendo che il primo lancio produce 1, qual è la probabilità che la somma dei due lanci sia 3?
- **Risposta 1**: $P(E_1) \times P(E_2|E_1) = \frac{12}{36} \times \frac{2}{12} = 2/36$.
- **Risposta 2**: $1 \times P(E_2|E_1) = 1/6$.

Conclusioni

- La probabilità dell'esito di un esperimento aleatorio è la proporzione di prove in cui quell'esito viene osservato, se l'esperimento viene ripetuto molte volte.

- Gli eventi composti si verificano quando viene osservato uno degli eventi semplici che li definiscono.

- La probabilità di un evento è uguale alla somma delle probabilità degli eventi semplici che lo costituiscono.

- Due eventi di dicono disgiunti se non hanno nessun evento semplice in comune. - La probabilità condizionata dell'evento A dato l'evento B è la probabilità di osservare A nello spazio campione ristretto di B.

- Due eventi A e B si dicono indipendenti se $P(A|B) = P(A)$.

Conclusioni